MULTI-LAYER CROSS-ATTENTION IS PROVABLY OPTIMAL FOR MULTI-MODAL IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent progress has rapidly advanced our understanding of the mechanisms underlying in-context learning in modern attention-based neural networks. However, existing results focus exclusively on unimodal data; in contrast, the theoretical underpinnings of in-context learning for multi-modal data remain poorly understood. We introduce a mathematically tractable framework for studying multi-modal learning and explore when transformer-like architectures can recover Bayes-optimal performance in-context. To model multi-modal problems, we assume the observed data arises from a latent factor model. Our first result comprises a negative take on expressibility: we prove that single-layer self-attention models fail to recover the Bayes-optimal predictor uniformly over the task distribution. To address this limitation, we introduce a novel, linearized cross-attention mechanism, which we study in the regime where both the number of cross-attention layers and the context length are large. We show that this cross-attention mechanism is provably Bayes optimal when optimized using gradient flow. Our results underscore the benefits of depth for in-context learning and establish the provable utility of cross-attention for multi-modal distributions.

1 INTRODUCTION

Large language models exhibit striking "in-context learning" (ICL) behavior (Brown et al., 2020): given a sequence of input-output pairs for a new task, the model can infer the rule (the function from input to output) without any update to the model weights. This phenomenon has motivated a growing body of theoretical work that tries to understand when and how ICL succeeds in linearized attention-based models (Zhang et al., 2024a; Lu et al., 2025; Zhang et al., 2025).

Prior work on ICL has focused on unimodal data; in contrast, modern foundation models routinely process multi-modal data, such as text, images, video (Li et al., 2024), multi-omic data (Cui et al., 2025), etc. In this work, we formulate and study an ICL problem on multi-modal data. Our main contributions are as follows:

- We introduce a statistical model for ICL on multi-modal data (Section 3). The proposed data distribution is a latent factor model where the factors are dependent across the data modalities, thereby capturing cross-modal commonalities and differences.
- We show that a single layer of linear self-attention (LSA) or softmax self-attention (SA) with finitely many heads fails to achieve Bayes-optimal performance in-context over this class of data distributions, and the softmax case reduces to the linear one (Section 4).
- We propose a multi-layer model with cross-attention (CA) plus SA and skip connections to solve the ICL problem in this context (Section 5). Linearizing the attention mechanism, we show that, even with a restriction of the model parameters, gradient flow on the population loss converges to the Bayes-optimal ICL predictor (Section 6).
- We complement our theory with numerical experiments in controlled and benchmark settings (Section 7). Synthetic experiments verify the qualitative predictions from our analyses. Experiments on corrupted avMNIST provide evidence that the new mechanisms identified by our theory are useful for multi-modal learning in conventional deep-learning architectures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To highlight the technical challenge in our setting, note that prior theory for ICL in regression assumes covariates sampled from a fixed distribution (usually Gaussian), independent of the learning task (Zhang et al., 2024a; 2025; Lu et al., 2025). Indeed, (Zhang et al., 2024a; 2025) note that the invariance of the covariate distribution across prompts is crucial in establishing the success of single-layer LSA. In sharp contrast, our multi-modal model exhibits covariate-shift across prompts, and the prediction rule in a prompt is naturally coupled with the covariate distribution.

These differences are significant, and provably rule out any single layer of attention in our setup, linear or softmax (Theorem 4.1). To overcome this, we introduce a novel multi-layer model based on linear cross-attention (LCA) and LSA, and show that it achieves Bayes-optimal performance for multi-modal ICL (Theorems 6.2 and 6.3). Our analyses deviate substantially from prior work, requiring new ideas to analyze the loss landscape in multi-modal problems. Our architecture differs from standard ones in the empirical literature in critical ways, namely in how we employ the CA mechanism and additional skip connections (Section 5). Our work is the first to provide theoretical guarantees on versions of attention-based models that achieve Bayes-optimal performance in multi-modal ICL.

2 Prior work

The first observation of ICL in large language models (Brown et al., 2020) motivated significant research on its foundations. Early work studied the expressivity of transformers (Bai et al., 2023; Akyürek et al., 2023; Garg et al., 2022), showing that they can implement common statistical algorithms in context. Connections with meta-learning and variants of gradient descent were noted in (Von Oswald et al., 2023; Ahn et al., 2023; Zhang et al., 2024b), generalization and stability were studied in (Li et al., 2023), and a Bayesian interpretation was offered in (Xie et al., 2022).

Training dynamics of attention-based models and ICL performance of the trained models have also been investigated. Perhaps the closest to our work is (Zhang et al., 2024a), which studies ICL for regression and shows that a single-layer LSA model trained by gradient flow on the population loss learns the Bayes-optimal predictor. Gradient flow or descent dynamics for multi-head self-attention were later analyzed in (Chen et al., 2024; Zhang et al., 2025), and (Huang et al., 2023) goes beyond linearized attention to stylized one-layer softmax transformers trained by gradient descent.

More recently, ICL has been studied for problems ranging from Gaussian mixture classification and clustering to non-parametric regression (Shen et al., 2025; Maulen-Soto et al., 2025; Ma et al., 2025; Ching et al., 2026). Attention-based models have also been studied as Gaussian sequence multi-index models (Cui et al., 2024; Arnaboldi et al., 2025; Troiani et al., 2025) and on sparse-token classification tasks (Oymak et al., 2023; Barnfield et al., 2025), though not with a focus on ICL. ICL at infinite token dimension was studied in (Lu et al., 2025; Letey et al., 2025); see also (Wu et al., 2024) for the effect of the number of pre-training tasks at finite token dimension. All of these works, however, consider unimodal data, whereas we focus on ICL for multi-modal data.

In diverse artificial intelligence problems, multi-modal learning—combining information from varied modes, e.g., text, image, video—is known to improve prediction, reasoning and learning capabilities. Although examined extensively in empirical studies (Radford et al., 2021; Alayrac et al., 2022; Jaegle et al., 2021; Wang et al., 2024), this setting is far from rigorously understood, and bridging this theoretical gap is a primary aim of our manuscript. Multi-modal learning is useful when the modes share common information. To capture this, we use latent variable models, widely used in statistical estimation for multi-modal data (Nandy & Ma, 2024; Ding et al., 2022; Mergny & Zdeborová, 2025; Keup & Zdeborová, 2025; Deshpande et al., 2018; Yang et al., 2025; Sergazinov et al., 2025), but these works do not study ICL or attention-based models. On the theoretical end, (Liu et al., 2025; Gui et al., 2025; Cai et al., 2024) study multi-modal contrastive learning, but do not focus on ICL.

3 Problem setup

We study a multi-modal framework where each prompt (task) provides a context of input–output examples; the inputs contain features from two distinct modalities. To model a natural multi-modal setting, we assume the data is drawn from a latent variable model. Unlike prior work on ICL, we allow the distribution of covariates to differ across tasks.

Formally, we observe N prompts (tasks) with (training) context length L. For each prompt $j \in [N]$, we observe a sequence $(\bar{\boldsymbol{z}}_i^{(j)}, \tilde{\boldsymbol{z}}_i^{(j)}, y_i^{(j)})_{i=1}^L$ where $y_i^{(j)} \in \mathbb{R}$ is a response and $\bar{\boldsymbol{z}}_i^{(j)} \in \mathbb{R}^{d_1}, \tilde{\boldsymbol{z}}_i^{(j)} \in \mathbb{R}^{d_2}$ are two modalities of covariates, e.g., the modalities could represent an image and associated text annotation, respectively. In each prompt, we additionally observe a tuple $(\bar{\boldsymbol{z}}_q^{(j)}, \tilde{\boldsymbol{z}}_q^{(j)}, y_q^{(j)})$. We input the prompt and covariate pair $(\bar{\boldsymbol{z}}_q^{(j)}, \tilde{\boldsymbol{z}}_q^{(j)})$ into the model, and it outputs a prediction $\hat{y}_q^{(j)}$ for $y_q^{(j)}$. While training, the model weights are tuned to minimize the error in predicting the y_q values on the training prompts. To assess the ICL performance of the fitted model, we consider a fresh prompt, and study the prediction error of the fitted model on the y_q associated with the test prompt.

We now turn to a statistical model for the data. The covariates and response are generated via

$$\bar{\boldsymbol{z}}_i^{(j)} = u_i^{(j)} \boldsymbol{v}^{(j)} + \boldsymbol{\xi}_i^{(j)}, \quad \tilde{\boldsymbol{z}}_i^{(j)} = u_i^{(j)} \boldsymbol{r}^{(j)} + \boldsymbol{\eta}_i^{(j)}, \quad y_i^{(j)} = \zeta^{(j)} u_i^{(j)}. \tag{3.1}$$

The latent variable $u_i^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ is shared across both modalities and the response. The prompt-specific regression coefficient, which relates $u_i^{(j)}$ to $y_i^{(j)}$, is given by $\zeta^{(j)} \in \mathbb{R}$. We assume that across prompts, $\zeta^{(j)}$ are sampled i.i.d. from a distribution with mean 0 and variance 1. On the other hand, $\boldsymbol{v}^{(j)} \in \mathbb{R}^{d_1}$ and $\boldsymbol{r}^{(j)} \in \mathbb{R}^{d_2}$ are modality-specific vectors sampled i.i.d. across prompts. We defer precise assumptions about their distributions (Assumption 6.1). Finally, $\boldsymbol{\xi}_i^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{d_1}(0, \boldsymbol{I})$ and $\boldsymbol{\eta}_i^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{d_2}(0, \boldsymbol{I})$ are i.i.d. noise variables, and all unique variables on the right-hand side of (3.1) are mutually independent. The setup in (3.1) formalizes a fundamental idea underlying multi-modal machine learning: each modality provides a different (noisy) view of a common latent feature u which drives the response y.

Focusing on a single task, omitting the prompt superscript, stacking the covariates, we denote $\boldsymbol{x}_i := [\bar{\boldsymbol{z}}_i \quad \tilde{\boldsymbol{z}}_i]^\top$ and let $d = d_1 + d_2$. Then, in each prompt, the covariates and response are jointly Gaussian:

$$\begin{bmatrix} \boldsymbol{x}_i \\ y_i \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{d+1}(0, \boldsymbol{\Sigma}) \quad \text{with} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top & \zeta \boldsymbol{m} \\ \zeta \boldsymbol{m}^\top & \zeta^2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{m} := \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{r} \end{bmatrix}. \tag{3.2}$$

Note, the same construction extends directly to three or more modalities by stacking all views into a single vector $\boldsymbol{x}_i$. The joint Gaussianity of the covariates and response allows us to write

$$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + \varepsilon_i \quad \text{where} \quad \boldsymbol{w} = \frac{\zeta}{1 + \|\boldsymbol{m}\|^2} \boldsymbol{m} \quad \text{and} \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0,\ \zeta^2/(1 + \|\boldsymbol{m}\|^2)\right). \tag{3.3}$$

Our goal is to learn the task-specific Bayes-optimal predictor $\langle w, x \rangle$. The setup (3.3) is reminiscent of the regression setting considered in (Zhang et al., 2024a; Lu et al., 2025; Zhang et al., 2025), yet presents several differences that make the learning problem significantly more difficult. Upon setting

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} := \mathrm{Cov}(\boldsymbol{x} \mid \boldsymbol{m}) = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top$$

as the covariance of the covariates $\{\boldsymbol{x}_i\}_{i=1}^L \cup \{\boldsymbol{x}_q\}$ conditioned on $\boldsymbol{m}$, we note that the distribution of covariates varies across prompts, since $\boldsymbol{m}$ is sampled anew for each task. Furthermore, the task vector $\boldsymbol{w}$ and covariate distribution are coupled through $\boldsymbol{m}$. Similarly, the variance of the additive noise and the task vector are coupled through the task-specific parameter ζ. These dependencies invalidate prior ICL theory in our setup. As we will see below, this leads to the failure of single-layer attention, including the linearized models previously shown to be successful for ICL when the covariate distribution does not change across prompts (Zhang et al., 2024a; 2025).

4 ICL AND FAILURE OF SINGLE-LAYER ATTENTION

In this section, we introduce ICL formally and show that single-layer attention fails at the ICL task. To this end, we represent a single prompt by its covariate matrix $\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L] \in \mathbb{R}^{d \times L}$ and target vector $\boldsymbol{y} := [y_1, \ldots, y_L]^\top \in \mathbb{R}^L$. We define the (non-learned) joint embedding matrix

$$\boldsymbol{E}_{\boldsymbol{X}} := \begin{bmatrix} \boldsymbol{X} & \boldsymbol{x}_q \\ \boldsymbol{y}^\top & 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_L & \boldsymbol{x}_q \\ y_1 & y_2 & \cdots & y_L & 0 \end{bmatrix} \in \mathbb{R}^{(d+1) \times (L+1)}$$

A model maps $\boldsymbol{E}_{\boldsymbol{X}}$ to a scalar prediction $\hat{y}_q$ for y_q. So far, the context length L is a free variable. In the subsequent discussion, it will be helpful to assume that the context length $L = L_{\mathrm{tr}}$ for prompts in the training data and $L = L_{\mathrm{te}}$ at test time. Our model architecture is independent of the context length, and thus this distinction is mostly for mathematical convenience.

Formal ICL criterion. For a new prompt with test context length $L = L_{\text{te}}$, let $\hat{y}_q^{(\text{te})}$ denote the model prediction. We say a model *in-context learns* if

$$\hat{y}_q^{(\text{te})} \longrightarrow y_q^{\text{Bayes},(\text{te})} := \mathbb{E}[y_q^{(\text{te})} | \boldsymbol{m}^{(\text{te})}, \zeta^{(\text{te})}, \boldsymbol{x}_q^{(\text{te})}] = \langle \boldsymbol{w}^{(\text{te})}, \boldsymbol{x}_q^{(\text{te})} \rangle \tag{4.1}$$

almost surely (a.s.) as $L_{\text{te}} \to \infty$, where the a.s. statement is with respect to the randomness of the new prompt. In words, ICL in our framework is synonymous with the model recovering the *Bayes-optimal predictor* $\langle \boldsymbol{w}^{(\text{te})}, \boldsymbol{x}_q^{(\text{te})} \rangle$ a.s. in the asymptotic $L_{\text{te}} \to \infty$ limit. The asymptotic context-length criterion (4.1) matches the mathematical interpretation of ICL in prior work (Zhang et al., 2024a), yet unlike in the linear regression setup (Garg et al., 2022; Zhang et al., 2024a; Lu et al., 2025), the task vector $\boldsymbol{w}$, covariate distribution, and error variances are all coupled via latent variables $\boldsymbol{m}$, ζ. These dependencies make both the learning problem—and its analysis—substantially more difficult.

Single-layer attention baselines. As a starting point, we study two simple baselines and show that neither can in-context learn in our setting. The first is a single LSA layer acting on $\boldsymbol{E}_{\boldsymbol{X}}$ as seen in (Von Oswald et al., 2023; Ahn et al., 2023; Zhang et al., 2024a; Lu et al., 2025). We write

$$\hat{y}_q = (\mathsf{LSA}(\boldsymbol{E}_{\boldsymbol{X}}; \theta))_{d+1,L+1} \quad \text{where} \quad \mathsf{LSA}(\boldsymbol{E}_{\boldsymbol{X}}; \theta) = \boldsymbol{E}_{\boldsymbol{X}} + \boldsymbol{W}^{PV} \boldsymbol{E}_{\boldsymbol{X}} \cdot (\boldsymbol{E}_{\boldsymbol{X}}^\top \boldsymbol{W}^{KQ} \boldsymbol{E}_{\boldsymbol{X}} / L) \tag{4.2}$$

and $\theta = \{\boldsymbol{W}^{PV}, \boldsymbol{W}^{KQ}\}$ are learnable weight matrices. This model is a linearized analogue of the second baseline, the SA block

$$\mathsf{SA}(\boldsymbol{E}_{\boldsymbol{X}}; \theta) = \boldsymbol{E}_{\boldsymbol{X}} + \boldsymbol{W}^P \boldsymbol{W}^V \boldsymbol{E}_{\boldsymbol{X}} \cdot \mathsf{softmax}\left((\boldsymbol{W}^K \boldsymbol{E}_{\boldsymbol{X}})^\top \boldsymbol{W}^Q \boldsymbol{E}_{\boldsymbol{X}} \right) \tag{4.3}$$

—where now $\theta = \{\boldsymbol{W}^P, \boldsymbol{W}^V, \boldsymbol{W}^K, \boldsymbol{W}^Q\}$—without the MLP layer and layer normalization, and with tied parameters $\boldsymbol{W}^{KQ} = (\boldsymbol{W}^K)^\top \boldsymbol{W}^Q$ and $\boldsymbol{W}^{PV} = \boldsymbol{W}^P \boldsymbol{W}^V$. Collapsing the key and query, and the projection and value weight matrices into single matrices differs from a practical transformer implementation, but is mathematically identical with regard to expressibility.

In demonstrating the success of the LSA model in unimodal ICL tasks, prior work (Zhang et al., 2024a) relies on proving that the parameters θ can emulate a global, inverse, covariance $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$. However, in the multi-modal setting, as the covariance $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is random, no fixed parameter θ can perform this inversion role and hence single-layer models fail at the ICL task as shown in Theorem 4.1.

Theorem 4.1 (Single-layer attention fails at ICL)**.** *In the setting of Section 3, and assuming $\|\boldsymbol{m}\|$ and ζ are atomless [1], neither the LSA model (4.2), with $\hat{y}_q = (\mathsf{LSA}(\boldsymbol{E}_{\boldsymbol{X}}; \theta))_{d+1,L_{\text{te}}+1}$ and fixed $\theta = \{\boldsymbol{W}^{PV}, \boldsymbol{W}^{KQ}\}$, nor the SA model (4.3), with any finite number of heads, arbitrary linear key, query, value and projection maps, a residual connection and an affine readout (see Appendix D.3), can in-context learn in the sense of (4.1). In each case, for every fixed parameter configuration, $\mathbb{P}\left(\lim_{L_{\text{te}} \to \infty} \hat{y}_q^{(\text{te})} = y_q^{\text{Bayes},(\text{te})} \right) = 0$.*

The above negative result (see proof in Appendix D) motivates examining more complex and deeper architectures for multimodal problems. Zhang et al. (2024a) demonstrates, for a particular shifting-covariate distribution, that the LSA model with weights obtained via gradient flow fails to in-context learn. In contrast, Theorem 4.1 establishes a negative result in full generality, showing that no parameter configuration will yield success at the ICL task under the multi-modal setting. Under the Gaussian latent factor model of Section 3, a softmax head has exactly the same infinite-context predictor as its linear counterpart (Proposition D.4), so replacing linear attention by softmax does not enlarge the attainable class at a single layer.

5 A MULTI-LAYER CA MODEL

Motivated by Theorem 4.1, we propose a model where the embedding matrix is now learned and obtained from a deep network of layers akin to CA. Consider the model

$$\hat{y}_q = \mathsf{SA}(\boldsymbol{E}_{\boldsymbol{F}}; \theta)_{d+1,L+1} \quad \text{where} \quad \boldsymbol{E}_{\boldsymbol{F}} := \begin{bmatrix} \boldsymbol{F} & \boldsymbol{x}_q \\ \boldsymbol{y}^\top & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{F} := \mathsf{CA}_T(\boldsymbol{X}; \gamma). \tag{5.1}$$

[1] For instance, any distribution on $\boldsymbol{m}$ that is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$.

The above model (5.1) is a composition of a self-attention layer applied to an embedding $\boldsymbol{E}_{\boldsymbol{F}}$ that is in turn obtained by applying multiple layers of CA between the covariates $\boldsymbol{X}$ and an evolving distribution $\{\boldsymbol{F}_t\}_{t\geq 0}$ defined below. The final SA layer is defined analogously to (4.3) with parameters θ and where $\boldsymbol{E}_{\boldsymbol{F}}$ now replaces $\boldsymbol{E}_{\boldsymbol{X}}$. To define the CA model $\mathsf{CA}_T(\boldsymbol{X};\gamma)$, consider the recurrence

$$\boldsymbol{F}_t = \boldsymbol{F}_{t-1} + \boldsymbol{S}_{t-1} + \boldsymbol{A}_{t-1}, \text{ for } t = 1,\dots,T. \quad \text{Define} \quad \boldsymbol{F} := \boldsymbol{F}_T := \mathsf{CA}_T(\boldsymbol{X};\gamma). \tag{5.2}$$

Above, $\boldsymbol{A}_{t-1}$ is the output of an attention head,

$$\mathsf{A}_{t-1} = \mathsf{A}(\boldsymbol{Q}_{t-1},\boldsymbol{K}_{t-1},\boldsymbol{V}_{t-1}) := \boldsymbol{V}_{t-1}\cdot \mathsf{softmax}\left(\boldsymbol{K}_{t-1}^\top \boldsymbol{Q}_{t-1}\right), \tag{5.3}$$
$$\boldsymbol{V}_{t-1} = \boldsymbol{W}_{t-1}^V \boldsymbol{X},\ \ \boldsymbol{K}_{t-1} = \boldsymbol{W}_{t-1}^K \boldsymbol{X},\ \ \boldsymbol{Q}_{t-1} = \boldsymbol{W}_{t-1}^Q \boldsymbol{F}_{t-1}.$$

Here, $\boldsymbol{W}_{t-1}^V, \boldsymbol{W}_{t-1}^K, \boldsymbol{W}_{t-1}^Q$ are learnable weight matrices, and $\boldsymbol{S}_{t-1} = \boldsymbol{W}_{t-1}^S \boldsymbol{X}$ with learnable weight matrix $\boldsymbol{W}_{t-1}^S$. We take $\boldsymbol{F}_0 = 0_{d\times L}$ to begin the recurrence and set $\boldsymbol{F} = \boldsymbol{F}_T$ with T denoting the depth of the CA embedding. The weights of the embedding are $\gamma = \{\boldsymbol{W}_t^V, \boldsymbol{W}_t^K, \boldsymbol{W}_t^Q, \boldsymbol{W}_t^S\}_{t=0}^{T-1}$.

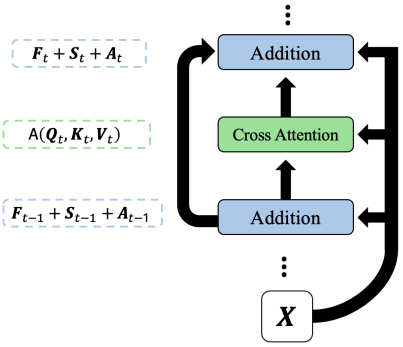

Figure 1: Re-injection of covariates $\boldsymbol{X}$ throughout the CA_T (LCA_T) embeddings. The propagation of the raw data $\boldsymbol{X}$ occurs both through $\boldsymbol{S}_t$ as well as through the CA block $\boldsymbol{A}_t$.

Notice that since the query matrix $\boldsymbol{Q}_{t-1}$ depends on $\boldsymbol{F}_{t-1}$, $\boldsymbol{A}_{t-1}$ is an application of CA between the original data matrix $\boldsymbol{X}$ and an evolving data distribution $\boldsymbol{F}_{t-1}$. For this reason, we view (5.1) as a T-layer CA model followed by a single layer of SA to obtain the final model output. The model includes the skip-connection of the previous layer input $\boldsymbol{F}_{t-1}$ in each layer (5.2) resembling standard transformers. Diverging from conventional architectures (Vaswani et al., 2017), we also include a learnable skip-connection $\boldsymbol{S}_{t-1}$ that injects the unaltered raw data $\boldsymbol{X}$ into each layer as displayed in Figure 1. Taking inspiration from unrolled neural networks (Gregor & LeCun, 2010; Monga et al., 2021), the inclusion of $\boldsymbol{S}_{t-1}$ was preceded by theoretical considerations—specifically, to yield embeddings $\boldsymbol{E}_{\boldsymbol{F}}$ with desirable properties—and later supported by ablation experiments (see Section 7 and Appendix B). As earlier, we use the bottom-right corner of the final output as our prediction $\hat{y}_q$.

Unlike conventional multimodal transformers, where cross-attention mediates interactions between different modalities (Tan & Bansal, 2019), our CA block repeatedly attends from the evolving state $\boldsymbol{F}_t$ back to the fixed multimodal sequence $\boldsymbol{X}$ itself. Together with the layer-wise re-injection term $\boldsymbol{S}_t$, this gives the model persistent access to the original input, heuristically enabling an iterative correction of the covariate representation; we examine the empirical relevance of these new architectural proposals through ablations in Section 7.

To ensure the analysis remains tractable, we replace $\mathsf{softmax}(\cdot)$ with the normalized identity $(\cdot)/L$, along the lines of Section 4 and prior theoretical literature on ICL (Von Oswald et al., 2023; Ahn et al., 2023; Zhang et al., 2024a; Lu et al., 2025). This yields the linearization

$$\hat{y}_q = \mathsf{LSA}(\boldsymbol{E}_{\boldsymbol{F}};\theta)_{d+1,L+1} \quad \text{where} \quad \boldsymbol{E}_{\boldsymbol{F}} = \begin{bmatrix} \boldsymbol{F} & \boldsymbol{x}_q \\ \boldsymbol{y}^\top & 0 \end{bmatrix} \quad \boldsymbol{F} = \mathsf{LCA}_T(\boldsymbol{X};\gamma), \tag{5.4}$$

and $\mathsf{LCA}_T(\boldsymbol{X};\gamma)$ denotes linearized CA where $\mathsf{softmax}(\cdot)$ is replaced by the normalized identity $(\cdot)/L$ in (5.3).

For mathematical simplification, we do not include two key components of conventional transformers—layer normalization (Xiong et al., 2020) and non-linear activations (MLP layers). In Section 7 (Table 1), we incorporate these components, together with other refinements, into our architecture to empirically validate the mechanisms motivated by our theory. That said, note the model (5.4) is highly nonlinear in $\boldsymbol{X}$ and the final LSA readout is itself a non-linear (cubic) transformation of the learned embedding $\boldsymbol{E}_{\boldsymbol{F}}$. Additionally, we notice qualitative similarities between the minima of the loss landscape both with and without layer normalization (Figure 4, Appendix A).

Weight simplifications. Thus far, the learnable weights have remained unconstrained. We now restrict their parameter space to ensure analytical tractability. To this end, we consider two simplifications. For both, we freeze $\boldsymbol{W}^K_{t-1} = \boldsymbol{W}^Q_{t-1} = \boldsymbol{I}_d$, and use an initialization strategy similar to (Zhang et al., 2024a) for $\boldsymbol{W}^{PV}$ and $\boldsymbol{W}^{KQ}$; see Appendix E.1.

One-parameter model. In the first case, we tie weights across layers and set $\boldsymbol{W}^S_{t-1} = -\boldsymbol{W}^V_{t-1} = \alpha \boldsymbol{I}_d$ for all $0 \le t \le T-1$ where $\alpha \in \mathbb{R}$ is a learnable scalar parameter. This gives a one-parameter model $\hat{y}_q = f(\boldsymbol{X}, \boldsymbol{x}_q, \boldsymbol{y}; \alpha)$.

Two-parameter model. For the second model, we instead set $\boldsymbol{W}^S_{t-1} = \alpha \boldsymbol{I}_d \quad \text{and} \quad \boldsymbol{W}^V_{t-1} = \beta \boldsymbol{I}_d$ for all $0 \le t \le T-1$ where $(\alpha, \beta) \in \mathbb{R}^2$ are learnable parameters, yielding a two-parameter model $\hat{y}_q = f(\boldsymbol{X}, \boldsymbol{x}_q, \boldsymbol{y}; \alpha, \beta)$.

The first model is recovered from the second by setting $\beta = -\alpha$. Since the one- and two-parameter models are subfamilies of the fully parameterized architecture (5.4), they help witness the expressibility of the unconstrained architecture. Importantly, these simplifications do not preclude optimality: we show in the following that the two restricted models already achieve Bayes-optimal prediction in the ICL setting when trained by gradient flow. Indeed, Figure 3 shows that less-restricted parametrizations exhibit qualitatively similar behavior as predicted in Section 6.

6 Training Setup and Main Results

Training setup. Herein, we consider the one- and two-parameter models $\hat{y}_q = f(\boldsymbol{X}, \boldsymbol{x}_q, y; \theta)$ of Section 5 where θ either denotes the parameters α or (α, β). Starting with a set of N training prompts $\{(\boldsymbol{x}_i^{(j)}, y_i^{(j)})_{i=1}^{L_{\text{tr}}} \cup (\boldsymbol{x}_q^{(j)}, y_q^{(j)})\}_{j=1}^N$, given real-valued, continuous targets, it is natural to consider the empirical squared loss on the target query: $\ell_{N,L_{\text{tr}}}(\theta) := \frac{1}{N}\sum_{j=1}^N (y_q^{(j)} - \hat{y}_q^{(j)})^2$. As in prior work (Zhang et al., 2024a; Ahn et al., 2023), taking the number of training prompts $N \to \infty$, we operate on the population loss $\ell_{L_{\text{tr}}}(\theta) := \lim_{N\to\infty} \ell_{N,L_{\text{tr}}}(\theta) = \mathbb{E}[(y_q - \hat{y}_q)^2]$, where the expectation is taken with respect to the joint distribution detailed in Section 3. As a final reduction, we consider the asymptotic, training context-length regime and define the loss $\ell(\theta) := \lim_{L_{\text{tr}}\to\infty} \ell_{L_{\text{tr}}}(\theta)$. This regime is qualitatively similar to prior work (Zhang et al., 2024a) whereby one establishes the ICL capacity of the model at large context length, however our presentation diverges by taking the $L_{\text{tr}} \to \infty$ limit before training on the loss function.

Having established the loss of interest, the training procedure for both variants of the CA model will simply be to run gradient flow on the limiting objective ℓ. That is, the model parameters evolve according to the ordinary differential equation $\frac{\mathrm{d}}{\mathrm{d}t}\theta_t = -\nabla \ell(\theta_t)$, where $(\theta_t)_{t\ge 0}$ is the trajectory of the parameters over the training horizon. Gradient flow can be seen as the limit of gradient descent in vanishing step size. Analyzing optimization in the gradient flow limit has enabled precise characterization of deep neural network training dynamics (Saxe et al., 2013; Chizat & Bach, 2018); this approach has been leveraged in previous studies on ICL as well (Zhang et al., 2024a; 2025).

Main results. We assume the following on the stacked modality-dependent vector $\boldsymbol{m}$ from (3.2).

Assumption 6.1 (Support of $\|\boldsymbol{m}\|$)**.** Assume $\boldsymbol{m}$ has a continuous distribution and $\underline{m} := \operatorname{ess\,inf} \|\boldsymbol{m}\|^2 < \operatorname{ess\,sup} \|\boldsymbol{m}\|^2 =: \overline{m} < \infty$ so that $\|\boldsymbol{m}\|$ is a.s. bounded and non-degenerate.

The boundedness assumption on $\boldsymbol{m}$ is mild, e.g., it holds as soon as the covariates are normalized, for instance $\boldsymbol{x}_i \in [0,1]^d$. The continuity assumption enters only through $\mathbb{P}(\boldsymbol{m} = \boldsymbol{0}_d) = 0$, which ensures that $Z := 1 + \|\boldsymbol{m}\|^2 > 1$ almost surely, and could be replaced by that condition, so in particular the results below require no density for $\|\boldsymbol{m}\|$, whose law may have atoms. We now present our main results. We show that a deep model of CA layers learns the Bayes-optimal prediction rule in-context for this multi-modal problem.

Theorem 6.2 (One-parameter model optimality)**.** *Assume the data generating process in Section 3 and Assumption 6.1. Run gradient flow on the loss $\ell(\alpha)$ for the one-parameter model. Then, we have:*

1. *Gradient flow converges. Formally, letting $(\alpha_{t,T})_{t\ge 0}$ denote the gradient flow trajectory for a CA stack of depth T, the limit $\alpha_T^* := \lim_{t\to\infty} \alpha_{t,T}$ exists.*

2. *As $T \to \infty$, $\alpha_T^* \to \frac{2}{2+\underline{m}+\overline{m}} =: \alpha^*$ and the resulting predictor is Bayes optimal as the depth diverges. Formally,* $\lim_{T\to\infty} \lim_{L_{\mathrm{te}}\to\infty} \hat{y}_q^{(\mathrm{te})} = y_q^{\mathrm{Bayes},(\mathrm{te})}$ *a.s.*

We prove this result in Appendix F. Since prior linear-attention ICL analyses study fixed-depth, often single-layer, architectures, our infinite depth analysis necessitates a different strategy: we use novel convex-analytic, epi-convergence-style arguments to show that the one-parameter loss converges to a limiting objective whose minimizer can be analyzed. We now turn to the two-parameter setting.

Theorem 6.3 (Two-parameter model optimality)**.** *Assume the data generating process in Section 3 and Assumption 6.1. Run gradient flow on the loss $\ell(\alpha, \beta)$ for the two-parameter model with initialization $\beta_0 \in (-2/(\overline{m}+1), 0)$ and $\alpha_0 = \alpha^*(\beta_0)$ where $\beta \mapsto \alpha^*(\beta)$ is the profiled minimizer (G.1) of Appendix G. Then, we have:*

1. *For large depth T, gradient flow converges. Formally, letting $(\alpha_{t,T}, \beta_{t,T})_{t\geq 0}$ denote the gradient flow trajectory for a CA stack, the limit $(\alpha_T^*, \beta_T^*) := \lim_{t\to\infty}(\alpha_{t,T}, \beta_{t,T})$ exists.*

2. *With α^* as defined in Theorem 6.2, as $T \to \infty$, $\alpha_T^* \to \alpha^*$ and $\beta_T^* \to -\alpha^*$, and the resulting predictor is Bayes optimal. Formally,* $\lim_{T\to\infty} \lim_{L_{\mathrm{te}}\to\infty} \hat{y}_q^{(\mathrm{te})} = y_q^{\mathrm{Bayes},(\mathrm{te})}$ *a.s.*

The proof can be found in Appendix G. This theorem is considerably more involved and requires new proof techniques: since the loss is non-convex and non-coercive, the fixed-depth, gradient-flow limit (α_T^*, β_T^*) can only be identified as a stationary point rather than a global minimizer. We handle this by reducing the problem to the evolution of β_T^* and showing, via a uniform derivative control, that, as $T \to \infty$, admissible stationary points collapse to the same limit. Theorem 6.3 shows that the limiting behavior of the two-parameter model mimics that of the one-parameter setting, whereby in both cases gradient flow achieves the Bayes-optimal predictor.

A careful reading of the proof of Theorem 6.2 informs the choice of initialization in Theorem 6.3. Solving the recurrence (5.2) in the one-parameter model reveals $\lim_{L_{\mathrm{te}}\to\infty} \hat{y}_q^{\mathrm{te}} = y_q^{\mathrm{Bayes},(\mathrm{te})} + e(\alpha)$ with error $e(\alpha) = O(\|\boldsymbol{I} - \alpha\boldsymbol{\Sigma}_{\boldsymbol{x}}\|^T)$. A short calculation shows that $e(\alpha) \to 0$ as $T \to \infty$ when $\alpha \in (0, 2/(\overline{m}+1))$. Given the correspondence between the two models, this lends to $\beta_0 \in (-2/(\overline{m}+1), 0)$ as a suitable initialization in the two-parameter setup. Moreover, the two-parameter loss $\ell(\alpha, \beta)$ turns out to be quadratic in α with explicit minimizer $\alpha^*(\beta)$, and hence starting at $\alpha_0 = \alpha^*(\beta_0)$ becomes an obvious choice. This intuition is supported in a visualization of the loss landscape provided in Figure 4, Appendix A.

The role of depth. Recall the embedding $\boldsymbol{F} = \mathsf{LCA}_T(\boldsymbol{X}; \gamma)$ (5.1) produced by the T-layer CA stack. For the one- and two-parameter models at fixed depth and parameters, the final prediction has the form $\hat{y}_q^{(\mathrm{te})} = \boldsymbol{w}^{(\mathrm{te})\top}(\frac{1}{L_{\mathrm{te}}}\boldsymbol{X}^{(\mathrm{te})}\boldsymbol{F}^\top)\boldsymbol{x}_q^{(\mathrm{te})} + o(1)$ almost surely as $L_{\mathrm{te}} \to \infty$, as shown in Appendix E (Lemma E.1). Thus, if the CA stack can enforce

$$\frac{1}{L_{\mathrm{te}}}\boldsymbol{X}^{(\mathrm{te})}\boldsymbol{F}^\top \longrightarrow \boldsymbol{I}_d \quad \text{as } T \to \infty, \tag{6.1}$$

then $\hat{y}_q^{(\mathrm{te})} \to y_q^{\mathrm{Bayes},(\mathrm{te})}$ as both $T \to \infty$ and $L_{\mathrm{te}} \to \infty$. An instance where (6.1) can occur is if $\boldsymbol{F} \approx \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{X}$ —that is, when $\boldsymbol{F}$ whitens the covariates. Crucially, since the embedding $\boldsymbol{F}$ is prompt-dependent, (6.1) is achievable despite $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ being random: the model is not constrained to learn a single global inverse covariance, rather it can compute $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ implicitly from the prompt itself. Here lies the significance of the CA embedding as a denoising mechanism for the covariates, where, in contrast, a single-layer attention model can only learn an average covariance, and thus fails to be Bayes-optimal.

Finite context and depth. The optimality results of Theorems 6.2 and 6.3 are asymptotic in both the test-context length and the depth. To supplement our main results, we offer non-asymptotic error bounds in Appendix H for the one-parameter model at a fixed parameter. The excess risk splits into a bias that contracts geometrically in the depth and a noise term of order d/L_{te} from the finite context. The geometric error decay rate is also evidenced numerically in Figure 2 whereby the model exhibits high accuracy at moderate depth.

7 NUMERICAL EXPERIMENTS

We verify the qualitative predictions of Theorems 4.1, 6.2, and 6.3 on synthetic data (Appendix I), test whether they persist for less-restricted parametrizations, and ablate our architectural components in the presence of standard transformer mechanisms on the multi-modal avMNIST dataset.

Validation of theoretical predictions. In Figure 2 (**Left**), we observe the failure of the single-layer LSA model proved in Theorem 4.1. In contrast, LCA-based models achieve error rates which are several orders of magnitude smaller as L_{te} grows. Figure 2 (**Right**) shows the effect of depth in the LCA_T embedding, (5.4), in aiding model performance on the multi-modal ICL task. Despite the asymptotic frameworks in Theorems 6.2 and 6.3, we observe exceptional performance at moderate depth as anticipated by the geometric-rate error decay $e(\alpha)$ discussed at the end of Section 6.

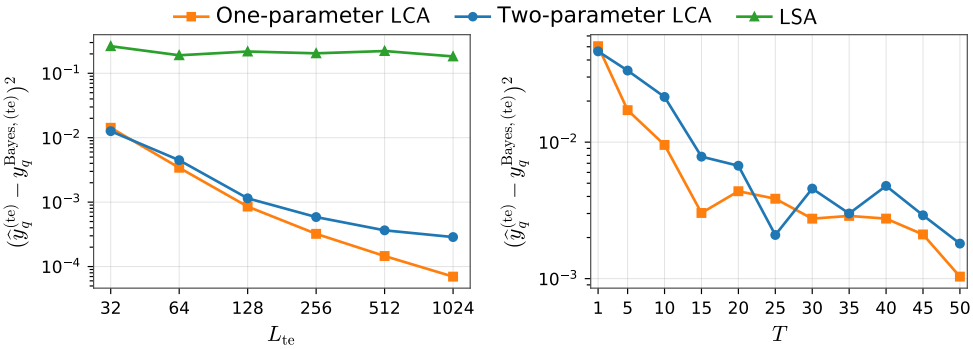

Figure 2: Prediction error for the one- and two-parameter LCA-based models and the LSA model. Left: ICL error at fixed depth for varying L_{te}. Right: ICL error at fixed L_{te} for varying depth T.

Less-restricted parametrizations. Here we test whether the predictions of Section 6 persist when the weights $(\boldsymbol{W}_t^S, \boldsymbol{W}_t^V)$ are untied across layers and taken to be scalar multiples of the identity, diagonal, or full matrices ('untied scalar', 'diagonal', and 'full-matrix' in Figure 3, see Appendix I). Figure 3 (**Left**) displays the normalized traces $\widehat{\alpha}_t = d^{-1}\operatorname{tr}(\boldsymbol{W}_t^S)$ and $\widehat{\beta}_t = d^{-1}\operatorname{tr}(\boldsymbol{W}_t^V)$ across layers, showing that they cluster around the one- and two-parameter $T \to \infty$ gradient-flow limit $(\alpha^*, -\alpha^*)$ (Theorems 6.2–6.3). The normalized traces $(\widehat{\alpha}_t, \widehat{\beta}_t)$ on their own do not confirm a full collapse of the form $(\boldsymbol{W}_t^S, \boldsymbol{W}_t^V) \approx (\alpha^* \boldsymbol{I}, -\alpha^* \boldsymbol{I})$, so Figure 3 (**Right**) displays the whitening diagnostic $\|\boldsymbol{X}\boldsymbol{F}_t^\top/L - \boldsymbol{I}\|_F/\sqrt{d}$ across layers. In line with (6.1), even these less-restricted models exhibit the same prompt-dependent covariance-whitening as proven earlier for our simpler models.

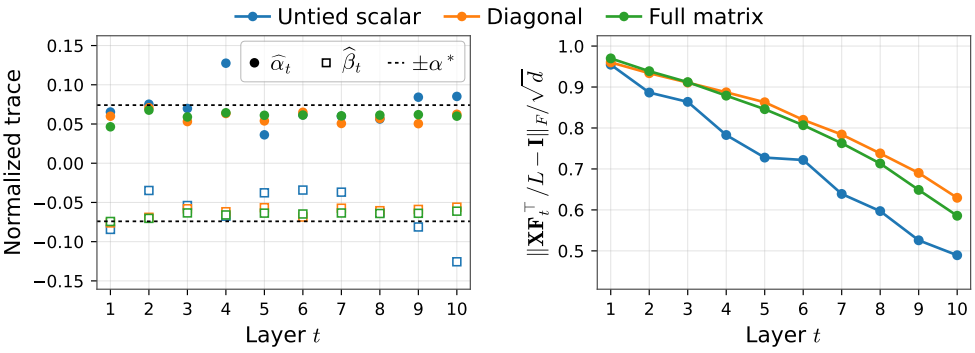

Figure 3: Diagnostics for less-restricted LCA parametrizations (setup same as in Figure 2; see Appendix I). **Left**: layerwise scalar projections; **Right**: whitening error across layers.

Performance of theory-motivated components in standard multi-modal transformers. We next ask whether the novel theory-motivated components of our CA architecture (Section 5) remain useful

after replacing the linearized model with standard transformer blocks—softmax multi-head attention, layer normalization, MLP blocks. We call this architecture—with all the standard transformer blocks on top of our innovations in Section 5—the "full paper-inspired model" (see Appendix J for details). Recall from (5.2)–(5.3) that each layer updates an evolving state $\boldsymbol{F}_t$ using two nonstandard ingredients: a CA term $\boldsymbol{A}_t$ between $\boldsymbol{F}_t$ and the data $\boldsymbol{X}$, and an additional skip term $\boldsymbol{S}_t$ which re-injects $\boldsymbol{X}$ at every layer. We evaluate on corrupted[2] avMNIST (Vielzeuf et al., 2019), a paired audio-visual digit classification benchmark. We compare the full paper-inspired model against a standard bidirectional CA transformer (Lu et al., 2019) and three ablations: (i) A no-injection ablation removes the $\boldsymbol{S}_t$ term from (5.2), while retaining CA between $\boldsymbol{F}_t$ and $\boldsymbol{X}$. (ii) A no-CA ablation retains the re-injection term but replaces the $\boldsymbol{F}_t$–$\boldsymbol{X}$ CA update by self-attention of $\boldsymbol{F}_t$ with itself. (iii) The final ablation removes both mechanisms, yielding a conventional self-attention model. Architecture and training details are deferred to Appendix J.

Model	Params	Best test acc. (%)	Best epoch
Full paper-inspired model (Inj. & cross-attn)	4.36×10^4	**63.85**	64
No injection	3.82×10^4	62.72	65
No cross-attn	4.36×10^4	62.10	79
No inj. & no cross-attn	3.82×10^4	60.79	90
Standard bi-CA transformer	6.07×10^4	62.72	62

Table 1: Classification results on avMNIST for the full paper-inspired model, three ablations thereof, and a standard bi-directional cross-attention transformer. All models use (pre) layer-norm, 4-headed multi-head attention, 10 attention layers, embedding dim. $d = 16$, hidden dim. $d_h = 32$, dropout rate of 0.1, GELU MLP blocks, and an appended CLS token. Training: cross-entropy loss over 100 epochs using AdamW optimizer with learning rate 3×10^{-4} and weight decay 10^{-4}. Model parameter counts and best test accuracy with corresponding epoch are shown.

Table 1 shows that our full paper-inspired model achieves the best test accuracy among all compared architectures. In particular, it outperforms each ablation. Additionally, it improves over the standard bidirectional cross-attention transformer, a standard fusion architecture for multimodal transformers (Lu et al., 2019), despite the latter having more trainable parameters. These results suggest that both mechanisms identified by our theory—cross-attention between the evolving state $\boldsymbol{F}_t$ and the original data $\boldsymbol{X}$, and recurrent re-injection of the covariates through $\boldsymbol{S}_t$—remain useful in more practical deep-learning implementations. While preliminary, the results are promising and motivate further investigation of our architecture in larger-scale multimodal settings.

8 Discussion

We studied a multi-modal task in which each prompt is generated by a latent factor model, inducing prompt-dependent covariate statistics which place the setting outside of standard unimodal ICL analyses. To demonstrate this added complexity, we first established the failure of single-layer attention, both the LSA models standard in the literature and their softmax counterparts. Motivated by this limitation, we introduced a novel transformer-like architecture, incorporating both depth and CA, proving not only its expressibility for the ICL task, but also demonstrating its capacity to learn the Bayes-optimal predictor via gradient flow. To our knowledge, our work is the first to tackle a multi-modal ICL task in a mathematically tractable setting and provably overcome covariate distribution shifts among prompts. While preliminary, our empirical findings (Table 1) demonstrate the promise of our new architectural innovations (Section 5) and motivate, as an important future direction, further empirical studies benchmarking these approaches on larger-scale multi-modal problems. Another important future direction would be to extend our population loss level analysis to sample-level results, as done earlier for unimodal problems (Lu et al., 2025). We expect this direction to pose intriguing mathematical challenges—requiring non-trivial developments in random matrix theory.

[2]We add substantial noise to covariates so the task is not saturated; this explains accuracies in the 60% range in Table 1.

AI USE STATEMENT

The research questions, the multi-modal ICL framework and the novel architecture were devised entirely by the authors. Prior versions of Theorem 4.1, focusing on linear attention, and of Theorem 6.3, with an added technical assumption on the limit points of $(\beta_T)_{T\geq 1}$, were proved entirely by the authors. We used generative AI assistants to extend the proof of Theorem 4.1 from linear to softmax attention and to remove the aforementioned technical assumption from Theorem 6.3. The softmax extension is a reduction to the linear case, since Gaussian exponential tilting (Lemma D.3, Proposition D.4) gives a softmax head the same infinite-context predictor as a linear-attention head. The proof is then completed by the argument of the authors' original linear-attention proof, isolated as Lemma D.1. Additionally, generative AI was used to help derive the finite-context bounds of Appendix H. Generative AI also helped implement experiment code and produce figures, and was used in polishing the prose of the paper such as in locating grammatical errors of our draft. All AI-assisted content was checked thoroughly by the authors, and we take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

The data-generating model (Section 3), the architectures (Sections 4–5) and the assumptions behind each result are stated in the main text, with complete proofs in Appendices C–H. Experimental settings for the synthetic simulations and the avMNIST ablations are given in Appendices A, B, I and J, and the code for all experiments will be released upon acceptance.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborova. Asymptotics of SGD in sequence-single index models and single-layer attention networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*, 2023.

Nicholas Barnfield, Hugo Cui, and Yue M Lu. High-dimensional analysis of single-layer attention for sparse-token classification. *arXiv preprint arXiv:2509.25153*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Tianxi Cai, Feiqing Huang, Ryumei Nakada, Linjun Zhang, and Doudou Zhou. Contrastive learning on multimodal analysis of electronic health records. *arXiv preprint arXiv:2403.14926*, 2024.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G Underwood, and Richard J Samworth. Efficient and minimax-optimal in-context nonparametric regression with transformers. *arXiv preprint arXiv:2601.15014*, 2026.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Haotian Cui, Alejandro Tejada-Lapuerta, Maria Brbić, Julio Saez-Rodriguez, Simona Cristea, Hani Goodarzi, Mohammad Lotfollahi, Fabian J Theis, and Bo Wang. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633, 2025.

Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. *Advances in Neural Information Processing Systems*, 37:36342–36389, 2024.

Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel. Contextual stochastic block models. *Advances in Neural Information Processing Systems*, 31, 2018.

Daisy Yi Ding, Shuangning Li, Balasubramanian Narasimhan, and Robert Tibshirani. Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences*, 119(38): e2202113119, 2022.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pp. 399–406, 2010.

Yu Gui, Cong Ma, and Zongming Ma. Multi-modal contrastive learning adapts to intrinsic dimensions of shared latent variables. *arXiv preprint arXiv:2505.12473*, 2025.

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

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Christian Keup and Lenka Zdeborová. Optimal thresholds and algorithms for a model of multi-modal learning in high dimensions. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(9): 093302, 2025.

Mary I Letey, Jacob A Zavatone-Veth, Yue M Lu, and Cengiz Pehlevan. Pretrain-test task alignment governs generalization in in-context learning. *arXiv preprint arXiv:2509.26551*, 2025.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International conference on machine learning*, pp. 19565–19594. PMLR, 2023.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Continual multimodal contrastive learning. *arXiv preprint arXiv:2503.14963*, 2025.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Yue M Lu, Mary Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025.

Tianyi Ma, Tengyao Wang, and Richard J Samworth. Provable test-time adaptivity and distributional robustness of in-context learning. *arXiv preprint arXiv:2510.23254*, 2025.

Rodrigo Maulen-Soto, Pierre Marion, and Claire Boyer. Attention-based clustering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Pierre Mergny and Lenka Zdeborová. Spectral thresholds in correlated spiked models and fundamental limits of partial least squares. *arXiv preprint arXiv:2510.17561*, 2025.

Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

Sagnik Nandy and Zongming Ma. Multimodal data integration and cross-modal querying via orchestrated approximate message passing. *arXiv preprint arXiv:2407.19030*, 2024.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, pp. 26724–26768. PMLR, 2023.

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

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Renat Sergazinov, Armeen Taeb, and Irina Gaynanova. A spectral method for multi-view subspace learning using the product of projections. *Biometrika*, pp. asaf088, 2025.

Wei Shen, Ruida Zhou, Jing Yang, and Cong Shen. On the training convergence of transformers for in-context classification of gaussian mixtures. In *Forty-second International Conference on Machine Learning*, 2025.

Jean-Jacques E Slotine, Weiping Li, et al. *Applied nonlinear control*, volume 199. Prentice hall Englewood Cliffs, NJ, 1991.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*, pp. 5100–5111. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514/.

Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborova. Fundamental limits of learning in sequence multi-index models and deep attention networks: high-dimensional asymptotics and sharp thresholds. In *Forty-second International Conference on Machine Learning*, 2025.

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

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: A multilayer approach for multimodal fusion. In Laura Leal-Taixé and Stefan Roth (eds.), *Computer Vision – ECCV 2018 Workshops*, pp. 575–589, Cham, 2019. Springer International Publishing. ISBN 978-3-030-11024-6.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *The Twelfth International Conference on Learning Representations*, 2024.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International conference on machine learning*, pp. 10524–10533. PMLR, 2020.

Xiaodong Yang, Buyu Lin, and Subhabrata Sen. Fundamental limits of community detection from multi-view data: multi-layer, dynamic and partially labeled block models. *The Annals of Statistics*, 53(6):2728–2756, 2025.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024a.

Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. In-context learning of a linear transformer block: Benefits of the MLP component and one-step GD initialization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.

Yedi Zhang, Aaditya K Singh, Peter E. Latham, and Andrew M Saxe. Training dynamics of in-context learning in linear attention. In *Forty-second International Conference on Machine Learning*, 2025.

A VISUALIZATION OF THE LOSS LANDSCAPE

The rationale behind the initializations given in Theorem 6.3 and discussed in Section 6 is supported in Figure 4 which illustrates a steep valley along the curve $\{(\alpha^*(\beta), \beta) : \beta \in (-2/(\overline{m}+1), 0)\}$ as $\alpha^*(\beta) \approx -\beta$ ($\alpha^*(\beta) \to -\beta$ as $T \to \infty$) in this window. To contrast our setting with a more sophisticated model, we also plot the loss landscape of the two-parameter model where pre-Layer Normalization (Xiong et al., 2020) is added between the LCA layers. Interestingly, this model reveals a similarly located and sized ravine in the loss landscape, hinting at the insights from our ablated setting to more complex models.

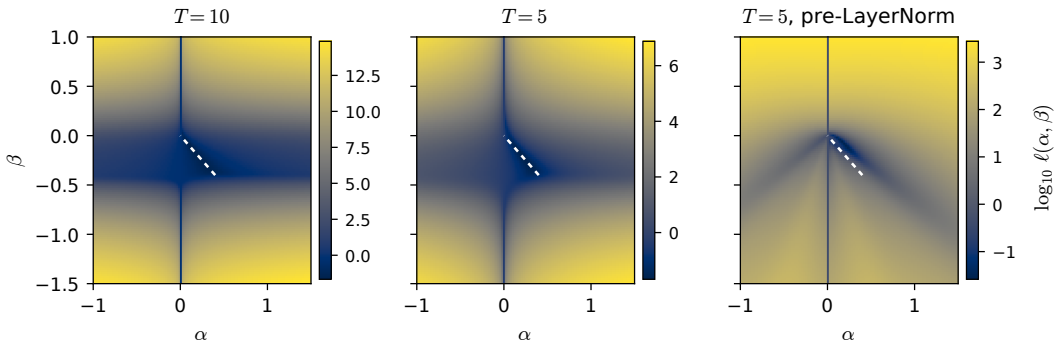

Figure 4: Log-scale heat maps of the two-parameter excess loss $\ell(\alpha, \beta)$ with $d_1 = d_2 = 16$ and $\|\boldsymbol{m}\| \sim \mathrm{Unif}(0, 2)$. (**Left**) depth $T = 10$. (**Middle**) depth $T = 5$. (**Right**) depth $T = 5$ with pre-LayerNorm added between the CA layers. The dashed line marks $\alpha = -\beta$ over $\beta \in (-2/(\overline{m}+1), 0)$. Each panel carries its own colour scale, as the three surfaces span different magnitudes.

B ABLATION EXPERIMENTS ON SYNTHETIC DATA

In this appendix, we examine the significance of the S_t skip-connection and the CA between the layer outputs $(F_t)_{t\geq 0}$ and the original raw data X—displayed in Figure 1—on the data generating distribution compatible with Section 3 and Assumption 6.1. To this end, we compare, empirically, the trained performance of the one- and two-parameter LCA models defined in Section 5 against ablated models. We introduce three new architectures in the one- and two-parameter settings. First, consider the one-parameter LCA model without S_t (equivalently with $\alpha \equiv 0$)—that is, we modify (5.2) to be

$$\boldsymbol{F}_t = \boldsymbol{F}_{t-1} + \boldsymbol{A}_{t-1}$$

and operate with one learnable parameter as in Section 5. Second, building on the preceding model, we now set $\boldsymbol{A}_t = \mathsf{A}(\boldsymbol{Q}_t, \boldsymbol{K}_t, \boldsymbol{V}_t)$ with

$$\boldsymbol{V}_t = \alpha \boldsymbol{F}_t, \quad \boldsymbol{K}_t = \boldsymbol{Q}_t = \boldsymbol{F}_t, \tag{B.1}$$

so that $\boldsymbol{A}_t$ is no longer a LCA implementation, but rather is LSA applied to $\boldsymbol{F}_t$ with a single learnable parameter. We refer to this architecture as a "one-parameter deep LSA without S_t" to delineate it from the single-layer LSA model of Section 4. Lastly, as a final trainable comparison, we consider the LSA modification given in (B.1) and re-introduce S_t (again producing two learnable parameters) in the architecture, and refer to this ablation as a "two-parameter deep LSA." As a naive baseline, we also consider the in-context sample mean:

$$\bar{y}_{L_{\mathrm{te}}} = \frac{1}{L_{\mathrm{te}}} \sum_{i=1}^{L_{\mathrm{te}}} y_i.$$

Both no-S_t ablations (the one-parameter LCA model and the one-parameter deep LSA model) are initialized with $\boldsymbol{F}_0 = \boldsymbol{X}$.

In Figure 5 (**Left**), we see that the removal of the S_t skip-connection in the embedding architecture compromises model performance. Notably, the ablated models cease to in-context learn, and their error plateaus as L_{te} grows, whereas the full models continue to improve. The one-parameter deep

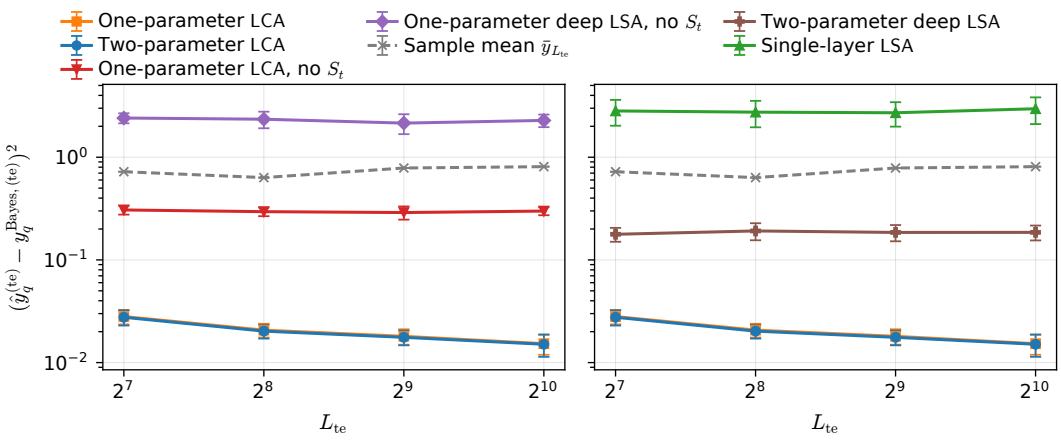

Figure 5: In-context performance at various L_{te} ($T = 10$, $d_1 = d_2 = 16$). (**Left**) one- and two-parameter LCA-based models, the two ablations without S_t, and the sample mean $\bar{y}_{L_{\text{te}}}$. (**Right**) one- and two-parameter LCA-based models, the two-parameter deep LSA model, the LSA model from (4.2), and the sample mean. Models are optimized on $\ell_{N,L_{\text{tr}}}$ ($L_{\text{tr}} = 100, N = 2000$) using gradient descent. Performance is averaged over 1000 test-prompts where error bars represent standard deviation over 10 separate training runs. The sweep starts at $L_{\text{te}} = 128$, since for L_{te} comparable to d the sample covariance fluctuates enough that the depth-T recurrence leaves its stable region on a small fraction of prompts.

LSA ablation is in addition worse than the naive in-context sample mean at every L_{te}. Moreover, in Figure 5 (**Right**), we see that swapping the LCA mechanism for LSA on F_t while keeping the S_t skip-connection deteriorates model performance, albeit improving upon the sample mean baseline and the single-layer LSA model. This further emphasizes the utility of the cross-attention and particularly of our novel S_t component.

C Notation, basic identities, and the Bayes-predictor

C.1 Notation and conventions

We work with a single prompt (task) unless otherwise stated. For two sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, we write

$$a_n = O(b_n)$$

to mean $a_n \leq C \cdot b_n$ for some constant $C > 0$ for n sufficiently large. We also use $o(1)$ to denote a sequence of terms which vanish as the ambient index grows. For example, writing $a_n = o(1)$ implies $a_n \to 0$ as $n \to \infty$.

C.2 Joint Gaussian form and Bayes-predictor

Recall the data generating process (Section 3): for $i \in [L] \cup \{q\}$,

$$\boldsymbol{x}_i = u_i \boldsymbol{m} + \boldsymbol{\mu}_i, \qquad y_i = \zeta u_i,$$

with $u_i \sim \mathcal{N}(0,1)$ and $\boldsymbol{\mu}_i = [\boldsymbol{\xi}_i; \boldsymbol{\eta}_i] \sim \mathcal{N}(0, I_d)$ independent across tokens i and independent of $(\boldsymbol{m}, \zeta)$.

Lemma C.1 (Joint covariance and Bayes coefficient)**.** *Conditional on $(\boldsymbol{m}, \zeta)$, the pair $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d+1}$ is jointly Gaussian with mean 0 and covariance*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}} & \zeta \boldsymbol{m} \\ \zeta \boldsymbol{m}^\top & \zeta^2 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top.$$

Moreover,

$$\mathbb{E}[y_i \mid \boldsymbol{m}, \zeta, \boldsymbol{x}_i] = \boldsymbol{w}^\top \boldsymbol{x}_i, \qquad \boldsymbol{w} = \frac{\zeta}{1 + \|\boldsymbol{m}\|^2} \boldsymbol{m},$$

Symbol	Meaning
$d = d_1 + d_2$	total covariate dimension
L	context length (we use L_{tr} or L_{te} when needed)
$\boldsymbol{x}_i \in \mathbb{R}^d$	stacked covariate for context token $i \in [L]$
$\boldsymbol{x}_q \in \mathbb{R}^d$	query covariate
$y_i \in \mathbb{R}$	context response for token $i \in [L]$
$y_q \in \mathbb{R}$	query response
$\boldsymbol{m} \in \mathbb{R}^d$	stacked modality vector (prompt-specific), $\boldsymbol{m} = [\boldsymbol{v}; \boldsymbol{r}]$
$\zeta \in \mathbb{R}$	prompt-specific scalar controlling the response
$u_i \sim \mathcal{N}(0,1)$	latent scalar shared across modalities and response
$\boldsymbol{\mu}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$	stacked noise in covariates
$\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top$	covariate covariance conditional on $\boldsymbol{m}$
$\boldsymbol{\Sigma}$	joint covariance of (x, y) conditioned on (m, ζ)
$Z := 1 + \|\boldsymbol{m}\|^2$	top eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ (rank-one spike)
$\boldsymbol{w} \in \mathbb{R}^d$	Bayes-optimal regression coefficient: $\boldsymbol{w} = \frac{\zeta}{1+\|\boldsymbol{m}\|^2}\boldsymbol{m}$
$\underline{m}, \overline{m}$	$\underline{m} = \operatorname{ess\,inf} \|\boldsymbol{m}\|^2$, $\overline{m} = \operatorname{ess\,sup} \|\boldsymbol{m}\|^2$
$\underline{Z}, \overline{Z}$	$\underline{Z} = 1 + \underline{m} = \operatorname{ess\,inf} Z$, $\overline{Z} = 1 + \overline{m} = \operatorname{ess\,sup} Z$

and we may write

$$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + \varepsilon_i, \quad \varepsilon_i \perp\!\!\!\perp \boldsymbol{x}_i \mid (\boldsymbol{m}, \zeta), \quad \varepsilon_i \sim \mathcal{N}\left(0, \frac{\zeta^2}{1 + \|\boldsymbol{m}\|^2}\right).$$

Proof. We compute conditional covariances given (m, ζ):

$$\boldsymbol{\Sigma}_{x,x} := \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top \mid \boldsymbol{m}, \zeta] = \mathbb{E}[(u_i \boldsymbol{m} + \boldsymbol{\mu}_i)(u_i \boldsymbol{m} + \boldsymbol{\mu}_i)^\top] = \boldsymbol{m}\boldsymbol{m}^\top + \boldsymbol{I} = \boldsymbol{\Sigma}_{\boldsymbol{x}},$$

$$\boldsymbol{\Sigma}_{x,y} := \mathbb{E}[\boldsymbol{x}_i y_i \mid \boldsymbol{m}, \zeta] = \mathbb{E}[(u_i \boldsymbol{m} + \boldsymbol{\mu}_i) \cdot \zeta u_i] = \zeta\, \mathbb{E}[u_i^2]\, \boldsymbol{m} = \zeta \boldsymbol{m},$$

$$\boldsymbol{\Sigma}_{y,y} := \mathbb{E}[y_i^2 \mid \boldsymbol{m}, \zeta] = \zeta^2 \mathbb{E}[u_i^2] = \zeta^2,$$

yielding the stated $\boldsymbol{\Sigma}$. For a jointly Gaussian vector, the conditional mean is

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i, \boldsymbol{m}, \zeta] = \boldsymbol{\Sigma}_{x,y}^\top \boldsymbol{\Sigma}_{x,x}^{-1} \boldsymbol{x}_i = (\zeta \boldsymbol{m})^\top \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{x}_i$$

and so $\boldsymbol{w} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}(\zeta \boldsymbol{m})$. Using the Sherman–Morrison formula,

$$\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} = (\boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top)^{-1} = \boldsymbol{I} - \frac{\boldsymbol{m}\boldsymbol{m}^\top}{1 + \|\boldsymbol{m}\|^2}$$

and hence,

$$\boldsymbol{w} = \left(\boldsymbol{I} - \frac{\boldsymbol{m}\boldsymbol{m}^\top}{1 + \|\boldsymbol{m}\|^2}\right) \zeta \boldsymbol{m} = \zeta \boldsymbol{m} - \zeta \frac{\|\boldsymbol{m}\|^2}{1 + \|\boldsymbol{m}\|^2} \boldsymbol{m} = \frac{\zeta}{1 + \|\boldsymbol{m}\|^2} \boldsymbol{m}.$$

Finally, the conditional variance of y_i given $\boldsymbol{x}_i$ is

$$\begin{aligned} \operatorname{Var}(y_i \mid \boldsymbol{x}_i, \boldsymbol{m}, \zeta) &= \boldsymbol{\Sigma}_{y,y} - \boldsymbol{\Sigma}_{x,y}^\top \boldsymbol{\Sigma}_{x,x}^{-1} \boldsymbol{\Sigma}_{x,y} \\ &= \zeta^2 - (\zeta \boldsymbol{m})^\top \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} (\zeta \boldsymbol{m}) \\ &= \zeta^2 \left(1 - \|\boldsymbol{m}\|^2/(1 + \|\boldsymbol{m}\|^2)\right) \\ &= \zeta^2/(1 + \|\boldsymbol{m}\|^2). \end{aligned}$$

The residual $\varepsilon_i := y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle$ is independent of $\boldsymbol{x}_i$ conditional on $(\boldsymbol{m}, \zeta)$ by Gaussianity. □

C.3 Law of large numbers for prompt statistics

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L] \in \mathbb{R}^{d \times L}$ and $\boldsymbol{y} = [y_1, \ldots, y_L]^\top \in \mathbb{R}^L$ denote the covariate matrix and target vector, and define the context sample covariance

$$\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} := \frac{1}{L} \boldsymbol{X}\boldsymbol{X}^\top.$$

Lemma C.2 (Matrix LLN at fixed dimension)**.** *Fix $(\boldsymbol{m}, \zeta)$ and thus $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top$. Then, almost surely as $L \to \infty$,*

$$\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \to \boldsymbol{\Sigma}_{\boldsymbol{x}}$$

entry-wise, and hence also in operator norm. Moreover,

$$\frac{1}{L}\boldsymbol{X}\boldsymbol{y} \to \mathbb{E}[\boldsymbol{x}_i y_i \mid \boldsymbol{m}, \zeta] = \zeta \boldsymbol{m}, \qquad \frac{1}{L}\sum_{i=1}^{L} y_i^2 \to \mathbb{E}[y_i^2 \mid \boldsymbol{m}, \zeta] = \zeta^2.$$

Finally, writing $y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + \varepsilon_i$ as in Lemma C.1 and letting $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_L)^\top$,

$$\frac{1}{L}\boldsymbol{X}\boldsymbol{\varepsilon} \to \boldsymbol{0} \quad \textit{almost surely.}$$

Proof. Conditional on $(\boldsymbol{m}, \zeta)$, the pairs $(\boldsymbol{x}_i, y_i)$ are i.i.d. Gaussian with finite second moments, so the stated convergences follow from the law of large numbers entrywise. □

D Proof of Theorem 4.1

Theorem 4.1 concerns two classes of single-layer attention models, namely the linear self-attention model (4.2) and the softmax self-attention block (4.3) with finitely many heads. Sections D.1 and D.2 give the argument for linear attention, demonstrating that its infinite-context predictor is affine in the query, with a slope that is a quadratic polynomial in the task scalar ζ whose linear coefficient is a *fixed* matrix applied to $\boldsymbol{m}$, and Lemma D.1 shows that no such slope can match the oracle coefficient $\boldsymbol{w} = \zeta\boldsymbol{m}/(1 + \|\boldsymbol{m}\|^2)$. Section D.3 extends the conclusion to softmax attention by reduction to this linear case, as Proposition D.4 demonstrates that a softmax head has the same infinite-context predictor as a linear-attention head.

Throughout this appendix we write the context tokens as

$$\boldsymbol{z}_j := \begin{bmatrix} \boldsymbol{x}_j \\ y_j \end{bmatrix} \in \mathbb{R}^{d+1}, \qquad j \in [L_{\mathrm{te}}],$$

which, conditionally on the task $(\boldsymbol{m}, \zeta)$, are i.i.d. $\mathcal{N}_{d+1}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ as in (3.2) and independent of the query $\boldsymbol{x}_q$.

D.1 The linear case I

This section and the next constitute the proof for linear attention. The lemma below isolates its concluding step, which is stated separately because Section D.3 reuses it.

Lemma D.1. *Suppose $\|\boldsymbol{m}\|$ and ζ are atomless and ζ is independent of $\boldsymbol{m}$. Let $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ be a fixed matrix, let $\boldsymbol{w}_0, \boldsymbol{w}_2 : \mathbb{R}^d \to \mathbb{R}^d$ be measurable, and suppose a predictor has the infinite-context limit $\langle \hat{\boldsymbol{w}}(\boldsymbol{m}, \zeta), \boldsymbol{x}_q \rangle + k(\boldsymbol{m}, \zeta)$, where k is measurable and*

$$\hat{\boldsymbol{w}}(\boldsymbol{m}, s) = \boldsymbol{w}_0(\boldsymbol{m}) + s\,\boldsymbol{B}^\top \boldsymbol{m} + s^2\,\boldsymbol{w}_2(\boldsymbol{m}), \qquad s \in \mathbb{R}. \tag{D.1}$$

Then

$$\mathbb{P}\Big(\langle \hat{\boldsymbol{w}}(\boldsymbol{m}, \zeta), \boldsymbol{x}_q \rangle + k(\boldsymbol{m}, \zeta) = \langle \boldsymbol{w}, \boldsymbol{x}_q \rangle\Big) = 0. \tag{D.2}$$

Proof. Fix $\boldsymbol{m}$ and subtract the oracle coefficient $\boldsymbol{w}(\boldsymbol{m}, s) = s\boldsymbol{m}/(1 + \|\boldsymbol{m}\|^2)$ of (3.3), giving the vector-valued polynomial

$$q_{\boldsymbol{m}}(s) := \hat{\boldsymbol{w}}(\boldsymbol{m}, s) - \frac{s\boldsymbol{m}}{1 + \|\boldsymbol{m}\|^2} = \boldsymbol{w}_0(\boldsymbol{m}) + \left(\boldsymbol{B}^\top \boldsymbol{m} - \frac{\boldsymbol{m}}{1 + \|\boldsymbol{m}\|^2}\right) s + \boldsymbol{w}_2(\boldsymbol{m}) s^2.$$

If $q_{\boldsymbol{m}}$ is not identically zero, at least one of its scalar components is a nonzero polynomial of degree at most two, so its zero set is finite. Since ζ is atomless and independent of $\boldsymbol{m}$,

$$\mathbb{P}\big(q_{\boldsymbol{m}}(\zeta) = \boldsymbol{0}_d \mid \boldsymbol{m}\big) = 0 \quad \text{whenever } q_{\boldsymbol{m}} \not\equiv \boldsymbol{0}_d.$$

If instead $q_{\boldsymbol{m}} \equiv \mathbf{0}_d$, all of its coefficients vanish. In particular, the linear coefficient gives

$$\boldsymbol{B}^\top \boldsymbol{m} = \frac{\boldsymbol{m}}{1 + \|\boldsymbol{m}\|^2}. \tag{D.3}$$

For $\boldsymbol{m} \neq \mathbf{0}_d$, equation (D.3) implies that $(1 + \|\boldsymbol{m}\|^2)^{-1}$ belongs to the finite spectrum $\sigma(\boldsymbol{B}^\top) = \sigma(\boldsymbol{B})$. The atomlessness of $\|\boldsymbol{m}\|$ gives both $\mathbb{P}(\boldsymbol{m} = \mathbf{0}_d) = 0$ and $\mathbb{P}((1 + \|\boldsymbol{m}\|^2)^{-1} \in \sigma(\boldsymbol{B})) = 0$, since $r \mapsto (1 + r^2)^{-1}$ is injective on $[0, \infty)$. Thus $\mathbb{P}(q_{\boldsymbol{m}} \equiv \mathbf{0}_d) = 0$. Combining the two cases and averaging over $\boldsymbol{m}$,

$$\mathbb{P}\left(\hat{\boldsymbol{w}}(\boldsymbol{m}, \zeta) = \boldsymbol{w}(\boldsymbol{m}, \zeta)\right) = 0.$$

Finally, condition on $(\boldsymbol{m}, \zeta)$ and write $\boldsymbol{\delta} := \hat{\boldsymbol{w}}(\boldsymbol{m}, \zeta) - \boldsymbol{w}(\boldsymbol{m}, \zeta)$. Whenever $\boldsymbol{\delta} \neq \mathbf{0}_d$, the scalar $\boldsymbol{\delta}^\top \boldsymbol{x}_q$ is Gaussian with mean zero and variance $\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\delta} \geq \|\boldsymbol{\delta}\|^2 > 0$, because $\boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{I}$. It therefore equals the fixed scalar $-k(\boldsymbol{m}, \zeta)$ with conditional probability zero. Averaging over the task proves (D.2). □

Only two features of the slope are used, namely that it is a polynomial of degree at most two in ζ, and that the coefficient of ζ is a fixed matrix acting on $\boldsymbol{m}$. Lemma D.1 therefore applies for any number of heads and for both linear and softmax attention.

D.2 The linear case II, failure of single-layer LSA

Recall the single-layer LSA predictor

$$\hat{y}_q = \mathsf{LSA}(\boldsymbol{E}_{\boldsymbol{X}}; \theta)_{d+1, L+1}, \quad \mathsf{LSA}(\boldsymbol{E}_{\boldsymbol{X}}; \theta) = \boldsymbol{E}_{\boldsymbol{X}} + \boldsymbol{W}^{PV} \boldsymbol{E}_{\boldsymbol{X}} \cdot \frac{\boldsymbol{E}_{\boldsymbol{X}}^\top \boldsymbol{W}^{KQ} \boldsymbol{E}_{\boldsymbol{X}}}{L}, \tag{D.4}$$

where

$$\boldsymbol{E}_{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{x}_q \\ \boldsymbol{y}^\top & 0 \end{bmatrix} \in \mathbb{R}^{(d+1) \times (L+1)}.$$

Proposition D.2 (The linear case)**.** *Suppose $\|\boldsymbol{m}\|$ and ζ are atomless and ζ is independent of $\boldsymbol{m}$. For every fixed $\theta = \{\boldsymbol{W}^{PV}, \boldsymbol{W}^{KQ}\}$, the predictor (D.4) satisfies $\mathbb{P}(\lim_{L_{\mathrm{te}} \to \infty} \hat{y}_q = \langle \boldsymbol{w}, \boldsymbol{x}_q \rangle) = 0$.*

Proof. Let us write the weight matrix $\boldsymbol{W}^{KQ}$ in block form

$$\boldsymbol{W}^{KQ} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b}' \\ \boldsymbol{b}^\top & c \end{bmatrix}, \qquad \boldsymbol{A} \in \mathbb{R}^{d \times d},\ \boldsymbol{b}, \boldsymbol{b}' \in \mathbb{R}^d,\ c \in \mathbb{R},$$

and denote the last row of $\boldsymbol{W}^{PV}$ by

$$\begin{bmatrix} \boldsymbol{u}^\top & v \end{bmatrix} := \boldsymbol{e}_{d+1}^\top \boldsymbol{W}^{PV}, \qquad \boldsymbol{u} \in \mathbb{R}^d,\ v \in \mathbb{R}.$$

Since $(\boldsymbol{E}_{\boldsymbol{X}})_{d+1, L_{\mathrm{te}}+1} = 0$, the residual term (skip-connection) in $\mathsf{LSA}(\boldsymbol{E}_{\boldsymbol{X}}; \theta)$ does not contribute to the predictor $\hat{y}_q$ and so

$$\hat{y}_q = \boldsymbol{e}_{d+1}^\top \boldsymbol{W}^{PV} \boldsymbol{E}_{\boldsymbol{X}} \cdot \frac{\boldsymbol{E}_{\boldsymbol{X}}^\top \boldsymbol{W}^{KQ} \boldsymbol{E}_{\boldsymbol{X}}}{L_{\mathrm{te}}} \boldsymbol{e}_{L_{\mathrm{te}}+1}.$$

For $i \in [L_{\mathrm{te}}]$, the i-th entry of the vector $\boldsymbol{E}_{\boldsymbol{X}}^\top \boldsymbol{W}^{KQ} \boldsymbol{E}_{\boldsymbol{X}} \boldsymbol{e}_{L_{\mathrm{te}}+1}$ is given by

$$[\boldsymbol{E}_{\boldsymbol{X}}^\top \boldsymbol{W}^{KQ} \boldsymbol{E}_{\boldsymbol{X}} \boldsymbol{e}_{L_{\mathrm{te}}+1}]_i = \begin{bmatrix} \boldsymbol{x}_i \\ y_i \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b}' \\ \boldsymbol{b}^\top & c \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_q \\ 0 \end{bmatrix} = \left(\boldsymbol{x}_i^\top \boldsymbol{A} \boldsymbol{x}_q + y_i \boldsymbol{b}^\top \boldsymbol{x}_q \right).$$

Moreover, the $(L_{\mathrm{te}} + 1)$-th entry of $\frac{1}{L_{\mathrm{te}}} \boldsymbol{E}_{\boldsymbol{X}}^\top \boldsymbol{W}^{KQ} \boldsymbol{E}_{\boldsymbol{X}} \boldsymbol{e}_{L_{\mathrm{te}}+1}$ is of order $O(\frac{1}{L_{\mathrm{te}}})$ and thus negligible as $L_{\mathrm{te}} \to \infty$. Hence, we obtain

$$\hat{y}_q = \frac{1}{L_{\mathrm{te}}} \sum_{i=1}^{L_{\mathrm{te}}} \left(\boldsymbol{u}^\top \boldsymbol{x}_i + v y_i \right) \left(\boldsymbol{x}_i^\top \boldsymbol{A} + y_i \boldsymbol{b}^\top \right) \boldsymbol{x}_q + o(1). \tag{D.5}$$

Conditioned on $(\boldsymbol{m}, \zeta)$, the context tokens $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{L_{\mathrm{te}}}$ are i.i.d., and $\boldsymbol{x}_q$ is independent of $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{L_{\mathrm{te}}}$ with the same conditional law as $\boldsymbol{x}_i$. Applying Lemma C.2 to each empirical moment that appears when expanding (D.5) yields, almost surely (for fixed $(\boldsymbol{m}, \zeta, \boldsymbol{x}_q)$),

$$\hat{y}_q \longrightarrow \langle \hat{\boldsymbol{w}}, \boldsymbol{x}_q \rangle,$$

as $L_{\mathrm{te}} \to \infty$ where the limiting regression vector $\hat{\boldsymbol{w}} = \hat{\boldsymbol{w}}(\boldsymbol{m}, \zeta) \in \mathbb{R}^d$ is

$$\hat{\boldsymbol{w}}(\boldsymbol{m}, \zeta)^\top = \boldsymbol{u}^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{A} + \Big((\boldsymbol{u}^\top \boldsymbol{m}) \boldsymbol{b}^\top + v \boldsymbol{m}^\top \boldsymbol{A} \Big) \zeta + v \boldsymbol{b}^\top \zeta^2. \tag{D.6}$$

Transposing (D.6) exhibits the form (D.1) required by Lemma D.1, with the task-independent matrix

$$\boldsymbol{B} := \boldsymbol{u}\boldsymbol{b}^\top + v\boldsymbol{A}, \qquad \boldsymbol{w}_0(\boldsymbol{m}) = \boldsymbol{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{u}, \qquad \boldsymbol{w}_2 = v\boldsymbol{b},$$

and with $k \equiv 0$. Lemma D.1 gives the claim. □

D.3 Extension to softmax attention by reduction to the linear case

We now extend Proposition D.2 to softmax attention. The extension is a reduction, as Proposition D.4 shows that a softmax head has the same infinite-context predictor as a linear-attention head, after which Lemma D.1 applies verbatim. We work in a class slightly more general than the block (4.3). For each head $h \in [H]$, with $H < \infty$ fixed, let

$$\boldsymbol{a}_h(\boldsymbol{x}_q) = \boldsymbol{K}_h \boldsymbol{x}_q + \boldsymbol{k}_h, \qquad s_{hj} = \boldsymbol{a}_h(\boldsymbol{x}_q)^\top \boldsymbol{z}_j, \qquad f_{hj} = \boldsymbol{c}_h^\top \boldsymbol{z}_j + c_{h,0}, \tag{D.7}$$

where $\boldsymbol{K}_h \in \mathbb{R}^{(d+1)\times d}$, $\boldsymbol{k}_h, \boldsymbol{c}_h \in \mathbb{R}^{d+1}$ and $c_{h,0} \in \mathbb{R}$. Let J_h be either $[L_{\mathrm{te}}]$ or $[L_{\mathrm{te}}] \cup \{q\}$, according to whether the query token participates in the attention, and if it does, its score and value may be arbitrary finite measurable functions of $\boldsymbol{x}_q$. The predictor is

$$\hat{y}_q^{(L_{\mathrm{te}})} = \langle \boldsymbol{\ell}_1, \boldsymbol{x}_q \rangle + \ell_0 + \sum_{h=1}^{H} \frac{\sum_{j \in J_h} e^{s_{hj}} f_{hj}}{\sum_{j \in J_h} e^{s_{hj}}}. \tag{D.8}$$

All parameters, and H, are fixed independently of the task and of L_{te}.

This class contains the block (4.3) with arbitrary linear key, query, value and projection maps of any head dimension, a residual connection and an affine scalar readout. Indeed, appending a fixed placeholder to the query's response slot and writing the effective key–query form of head h as $\boldsymbol{W}_h^{KQ}$, the context score $\boldsymbol{z}_j^\top \boldsymbol{W}_h^{KQ}(\boldsymbol{x}_q, p)$ is affine in $\boldsymbol{z}_j$ with coefficient of the form (D.7), and linear value and projection maps followed by a scalar readout likewise make f_{hj} affine in $\boldsymbol{z}_j$. Terms in a score that are common to all keys cancel in the softmax.

Lemma D.3 (Gaussian exponential tilt)**.** *Let $\boldsymbol{z} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succeq 0$. For every $\boldsymbol{a}, \boldsymbol{c} \in \mathbb{R}^k$ and $c_0 \in \mathbb{R}$,*

$$\mathbb{E}\big[e^{\boldsymbol{a}^\top \boldsymbol{z}}\big] = e^{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}/2}, \qquad \frac{\mathbb{E}\big[(\boldsymbol{c}^\top \boldsymbol{z} + c_0) e^{\boldsymbol{a}^\top \boldsymbol{z}}\big]}{\mathbb{E}\big[e^{\boldsymbol{a}^\top \boldsymbol{z}}\big]} = \boldsymbol{c}^\top \boldsymbol{\Sigma} \boldsymbol{a} + c_0. \tag{D.9}$$

Moreover $\mathbb{E}\big[e^{|\boldsymbol{a}^\top \boldsymbol{z}|} |\boldsymbol{c}^\top \boldsymbol{z} + c_0|\big] < \infty$.

Proof. Write $\boldsymbol{z} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{g}$ with $\boldsymbol{g} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{I})$ and set $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{a}$. Completing the square in the standard Gaussian density φ_k gives $e^{\boldsymbol{\beta}^\top \boldsymbol{g}} \varphi_k(\boldsymbol{g}) = e^{\|\boldsymbol{\beta}\|^2/2} \varphi_k(\boldsymbol{g} - \boldsymbol{\beta})$, and integrating yields

$$\mathbb{E}\big[e^{\boldsymbol{\beta}^\top \boldsymbol{g}}\big] = e^{\|\boldsymbol{\beta}\|^2/2}, \qquad \mathbb{E}\big[\boldsymbol{g}\, e^{\boldsymbol{\beta}^\top \boldsymbol{g}}\big] = \boldsymbol{\beta}\, e^{\|\boldsymbol{\beta}\|^2/2}.$$

Multiplying the second identity by $\boldsymbol{\Sigma}^{1/2}$ gives $\mathbb{E}[\boldsymbol{z} e^{\boldsymbol{a}^\top \boldsymbol{z}}] = \boldsymbol{\Sigma}\boldsymbol{a}\, e^{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}/2}$, which proves (D.9), and the argument covers singular $\boldsymbol{\Sigma}$. For integrability, use $e^{|t|} \le e^t + e^{-t}$ together with the Cauchy–Schwarz inequality, noting that a Gaussian moment generating function is finite at every real argument and that the affine value has finite second moment. □

The second identity in (D.9) is the reason softmax collapses onto the linear-attention population class here, as exponential tilting shifts a Gaussian mean by exactly $\boldsymbol{\Sigma}\boldsymbol{a}$, which is linear in the tilt.

Proposition D.4 (Population predictor and linear-attention equivalence)**.** *For every fixed parameter configuration in (D.8), almost surely*

$$\lim_{L_{\mathrm{te}} \to \infty} \hat{y}_q^{(L_{\mathrm{te}})} = \langle \hat{\boldsymbol{w}}(\boldsymbol{m}, \zeta), \boldsymbol{x}_q \rangle + k(\boldsymbol{m}, \zeta), \tag{D.10}$$

where

$$\hat{\boldsymbol{w}}(\boldsymbol{m},\zeta)^\top = \boldsymbol{\ell}_1^\top + \sum_{h=1}^{H} \boldsymbol{c}_h^\top \boldsymbol{\Sigma} \boldsymbol{K}_h, \qquad k(\boldsymbol{m},\zeta) = \ell_0 + \sum_{h=1}^{H} \big(\boldsymbol{c}_h^\top \boldsymbol{\Sigma} \boldsymbol{k}_h + c_{h,0}\big). \tag{D.11}$$

Replacing the softmax ratio of head h by its normalized linear counterpart $L_{\mathrm{te}}^{-1} \sum_{j \in J_h} s_{hj} f_{hj}$ yields the same limit, minus $\sum_h c_{h,0}$. Consequently, if all value biases $c_{h,0}$ vanish then the softmax and the normalized linear model with the same *parameters have the same infinite-context predictor almost surely, and with an unrestricted readout intercept the two families realize the same class of infinite-context predictors.*

Proof. Condition on $(\boldsymbol{m}, \zeta, \boldsymbol{x}_q)$ and fix a head, so that $\boldsymbol{a} := \boldsymbol{a}_h(\boldsymbol{x}_q)$, $\boldsymbol{c} := \boldsymbol{c}_h$ and $c_0 := c_{h,0}$ are fixed while the context tokens are i.i.d. Gaussian. By Lemma D.3 the numerator and denominator summands are integrable, so the strong law gives

$$\frac{1}{L_{\mathrm{te}}} \sum_{j=1}^{L_{\mathrm{te}}} e^{s_{hj}} f_{hj} \longrightarrow (\boldsymbol{c}^\top \boldsymbol{\Sigma} \boldsymbol{a} + c_0)\, e^{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}/2}, \qquad \frac{1}{L_{\mathrm{te}}} \sum_{j=1}^{L_{\mathrm{te}}} e^{s_{hj}} \longrightarrow e^{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}/2} > 0.$$

If the query participates, its numerator and denominator contributions are $e^{s_{hq}} f_{hq}/L_{\mathrm{te}}$ and $e^{s_{hq}}/L_{\mathrm{te}}$, which vanish. Head h therefore converges to $\boldsymbol{c}^\top \boldsymbol{\Sigma} \boldsymbol{a} + c_0$, and summing the finitely many heads and expanding $\boldsymbol{a}_h(\boldsymbol{x}_q) = \boldsymbol{K}_h \boldsymbol{x}_q + \boldsymbol{k}_h$ gives (D.10)–(D.11). For the linear counterpart, the strong law gives

$$\frac{1}{L_{\mathrm{te}}} \sum_{j=1}^{L_{\mathrm{te}}} s_{hj} f_{hj} \longrightarrow \mathbb{E}\big[(\boldsymbol{a}^\top \boldsymbol{z})(\boldsymbol{c}^\top \boldsymbol{z} + c_0)\big] = \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{c} + c_0\, \boldsymbol{a}^\top \mathbb{E}[\boldsymbol{z}] = \boldsymbol{c}^\top \boldsymbol{\Sigma} \boldsymbol{a},$$

since $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{0}$, and any query contribution again vanishes after division by L_{te}. The only difference is the constant $\sum_h c_{h,0}$, which a readout intercept absorbs. Integrating the conditional probability-one statements over $(\boldsymbol{m}, \zeta, \boldsymbol{x}_q)$ completes the proof. □

Proposition D.5 (The softmax case)**.** *Suppose $\|\boldsymbol{m}\|$ and ζ are atomless and ζ is independent of $\boldsymbol{m}$. For every fixed parameter configuration in* (D.8)*, $\mathbb{P}(\lim_{L_{\mathrm{te}} \to \infty} \hat{y}_q^{(L_{\mathrm{te}})} = \langle \boldsymbol{w}, \boldsymbol{x}_q \rangle) = 0$. The same holds for every fixed finite-head normalized linear-attention counterpart of* (D.8)*.*

Proof. Write

$$\boldsymbol{c}_h = \begin{bmatrix} \boldsymbol{u}_h \\ v_h \end{bmatrix}, \qquad \boldsymbol{K}_h = \begin{bmatrix} \boldsymbol{A}_h \\ \boldsymbol{b}_h^\top \end{bmatrix}, \qquad \boldsymbol{u}_h \in \mathbb{R}^d,\ v_h \in \mathbb{R},\ \boldsymbol{A}_h \in \mathbb{R}^{d \times d},\ \boldsymbol{b}_h \in \mathbb{R}^d.$$

Since $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}} & \zeta \boldsymbol{m} \\ \zeta \boldsymbol{m}^\top & \zeta^2 \end{bmatrix}$ by (3.2), block multiplication gives

$$\boldsymbol{c}_h^\top \boldsymbol{\Sigma} \boldsymbol{K}_h = \boldsymbol{u}_h^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{A}_h + \zeta\, \boldsymbol{m}^\top \big(v_h \boldsymbol{A}_h + \boldsymbol{u}_h \boldsymbol{b}_h^\top\big) + \zeta^2 v_h \boldsymbol{b}_h^\top. \tag{D.12}$$

Substituting (D.12) into (D.11) and transposing exhibits the form (D.1), with

$$\boldsymbol{B} := \sum_{h=1}^{H} \big(v_h \boldsymbol{A}_h + \boldsymbol{u}_h \boldsymbol{b}_h^\top\big), \qquad \boldsymbol{w}_0(\boldsymbol{m}) = \boldsymbol{\ell}_1 + \sum_{h=1}^{H} \boldsymbol{A}_h^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{u}_h, \qquad \boldsymbol{w}_2 = \sum_{h=1}^{H} v_h \boldsymbol{b}_h,$$

all independent of the task except through $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, which enters only $\boldsymbol{w}_0$. The intercept k in (D.11) is measurable, and Proposition D.4 supplies the almost-sure limit, so Lemma D.1 applies and gives the first claim. For the linear counterpart, Proposition D.4 shows that the slope is unchanged and only the task-independent intercept differs, so Lemma D.1 applies verbatim. □

Proof of Theorem 4.1. The statement for the linear model (4.2) is Proposition D.2, and the statement for the finite-head softmax class (D.8), which contains the block (4.3), is Proposition D.5. In both cases the conclusion is exactly the failure of the criterion (4.1). □

E CA ALGEBRA AND PRELIMINARIES

This appendix collects algebraic identities for the LCA stack and the final LSA readout. These identities are used in the proofs of Theorems 6.2 and 6.3.

E.1 LSA READOUT WITH FROZEN PARAMETERS

Taking inspiration from the initialization strategy of Zhang et al. (2024a), we initialize the $\boldsymbol{W}^{PV}$ and $\boldsymbol{W}^{KQ}$ by

$$\boldsymbol{W}^{PV} = \begin{bmatrix} \mathbf{0}_{d\times d} & \mathbf{0}_d \\ \mathbf{0}_d^\top & 1 \end{bmatrix}, \qquad \boldsymbol{W}^{KQ} = \begin{bmatrix} \boldsymbol{I}_d & \mathbf{0}_d \\ \mathbf{0}_d^\top & 0 \end{bmatrix}.$$

Lemma E.1 (Closed-form readout)**.** *With the above $\boldsymbol{W}^{PV}, \boldsymbol{W}^{KQ}$, the single-layer LSA readout satisfies, for any embedding $\boldsymbol{F} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_L]$,*

$$\widehat{y}_q = \mathrm{LSA}(\boldsymbol{E}_{\boldsymbol{F}})_{d+1,L+1} = \boldsymbol{w}^\top \left[\frac{1}{L}\sum_{i=1}^{L} \boldsymbol{x}_i \boldsymbol{f}_i^\top\right] \boldsymbol{x}_q + \left[\frac{1}{L}\sum_{i=1}^{L} \varepsilon_i \boldsymbol{f}_i^\top\right] \boldsymbol{x}_q.$$

For the one- and two-scalar CA models of Appendix E.2, fix d, $T \geq 1$, and real parameters α, β independent of the prompt and of L, with $\beta = -\alpha$ in the one-parameter case. Define

$$P_T(\boldsymbol{H}) := \alpha \sum_{k=0}^{T-1} (\boldsymbol{I} + \beta \boldsymbol{H})^k, \qquad \boldsymbol{H} \in \mathbb{R}^{d\times d}.$$

Then, as $L \to \infty$,

$$N_L := \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{F}_T^\top \boldsymbol{x}_q}{L} \to 0, \qquad \hat{y}_q \to \hat{y}_{q,\infty} := \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} P_T(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \boldsymbol{x}_q \quad \textit{almost surely}.$$

Under Assumption 6.1, both convergences also hold in mean square, and the corresponding expected squared prediction losses converge.

Proof. By definition,

$$\mathsf{LSA}(\boldsymbol{E}_{\boldsymbol{F}}) = \boldsymbol{E}_{\boldsymbol{F}} + \boldsymbol{W}^{PV} \boldsymbol{E}_{\boldsymbol{F}} \cdot \frac{\boldsymbol{E}_{\boldsymbol{F}}^\top \boldsymbol{W}^{KQ} \boldsymbol{E}_{\boldsymbol{F}}}{L}.$$

Since $(\boldsymbol{E}_{\boldsymbol{F}})_{d+1,L+1} = 0$, given the definitions of $\boldsymbol{W}^{PV}$ and $\boldsymbol{W}^{KQ}$, we have

$$\begin{aligned}
\hat{y}_q &= \boldsymbol{e}_{d+1}^\top \boldsymbol{W}^{PV} \boldsymbol{E}_{\boldsymbol{F}} \cdot \frac{\boldsymbol{E}_{\boldsymbol{F}}^\top \boldsymbol{W}^{KQ} \boldsymbol{E}_{\boldsymbol{F}}}{L} \boldsymbol{e}_{L+1} \\
&= \boldsymbol{e}_{d+1}^\top \frac{\boldsymbol{E}_{\boldsymbol{F}} \boldsymbol{E}_{\boldsymbol{F}}^\top}{L} \begin{bmatrix} \boldsymbol{x}_q \\ 0 \end{bmatrix} \\
&= \boldsymbol{e}_{d+1}^\top \frac{1}{L} \begin{bmatrix} \boldsymbol{F}\boldsymbol{F}^\top + \boldsymbol{x}_q \boldsymbol{x}_q^\top & \boldsymbol{F}\boldsymbol{y} \\ \boldsymbol{y}^\top \boldsymbol{F}^\top & \boldsymbol{y}^\top \boldsymbol{y} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_q \\ 0 \end{bmatrix} \\
&= \frac{\boldsymbol{y}^\top \boldsymbol{F}^\top \boldsymbol{x}_q}{L} \\
&= \boldsymbol{w}^\top \left[\frac{1}{L}\sum_{i=1}^{L} \boldsymbol{x}_i \boldsymbol{f}_i^\top\right] \boldsymbol{x}_q + \left[\frac{1}{L}\sum_{i=1}^{L} \varepsilon_i \boldsymbol{f}_i^\top\right] \boldsymbol{x}_q.
\end{aligned}$$

For the scalar CA models, the recurrence gives $\boldsymbol{F}_T = P_T(\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}})\boldsymbol{X}$, where $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} = \boldsymbol{X}\boldsymbol{X}^\top/L$, see the algebraic identity in Lemma E.2. Conditional on $(\boldsymbol{m}, \zeta)$, Lemma C.2 gives $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \to \boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{X}\boldsymbol{\varepsilon}/L \to \mathbf{0}_d$ almost surely. Polynomial continuity therefore implies

$$\frac{\boldsymbol{F}_T \boldsymbol{\varepsilon}}{L} = P_T(\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}) \frac{\boldsymbol{X}\boldsymbol{\varepsilon}}{L} \to \mathbf{0}_d, \qquad \frac{\boldsymbol{X}\boldsymbol{F}_T^\top}{L} = \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} P_T(\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}) \to \boldsymbol{\Sigma}_{\boldsymbol{x}} P_T(\boldsymbol{\Sigma}_{\boldsymbol{x}}).$$

Multiplying by the fixed query $\boldsymbol{x}_q$ and using the exact readout identity proves the almost-sure claims. Averaging over the task and query preserves the probability-one statement. This argument does not require independence among the feature columns or a bound on their norms.

For mean-square convergence, assume Assumption 6.1. All norms in the following bounds are Euclidean or operator norms, as appropriate. For every integer $p \geq 1$, Jensen's inequality and the conditional Gaussian law give

$$\sup_{L\geq 1} \mathbb{E}\Big[(\mathrm{tr}\, \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}})^p \mid \boldsymbol{m}\Big] \leq \mathbb{E}[\|\boldsymbol{x}_1\|^{2p} \mid \boldsymbol{m}] \leq C_p,$$

where C_p depends only on $p, d, \overline{m}$, since $\|\boldsymbol{\Sigma}_{\boldsymbol{x}}\| \le 1 + \overline{m}$. Conditional on $(\boldsymbol{m}, \zeta, \boldsymbol{X})$, the vector $\boldsymbol{\varepsilon}$ is centered with covariance $\zeta^2(1 + \|\boldsymbol{m}\|^2)^{-1}\boldsymbol{I}_L$ and is independent of the query. Thus,

$$\mathbb{E}[N_L^2 \mid \boldsymbol{m}, \zeta, \boldsymbol{X}] = \frac{\zeta^2}{L(1 + \|\boldsymbol{m}\|^2)} \operatorname{tr}\Big(\boldsymbol{\Sigma}_{\boldsymbol{x}} P_T(\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}) \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} P_T(\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}})\Big).$$

Because P_T has degree at most $T-1$, the trace is bounded by $C(1 + \operatorname{tr}\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}})^{2T-1}$, with C depending only on $d, T, \alpha, \beta, \overline{m}$. The moment bound above yields $\mathbb{E}[N_L^2 \mid \boldsymbol{m}, \zeta] \le C\zeta^2/L$ and hence $\mathbb{E}[N_L^2] \le C/L \to 0$, using only $\mathbb{E}[\zeta^2] = 1$.

To control the signal term, let

$$\boldsymbol{D}_L := \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} P_T(\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}) - \boldsymbol{\Sigma}_{\boldsymbol{x}} P_T(\boldsymbol{\Sigma}_{\boldsymbol{x}}).$$

For each fixed $\boldsymbol{m}$, $\boldsymbol{D}_L \to \boldsymbol{0}$ almost surely. The same moment bound with $p = 4T$ gives $\sup_L \mathbb{E}[\|\boldsymbol{D}_L\|^4 \mid \boldsymbol{m}] \le C$, uniformly in $\boldsymbol{m}$. Consequently, $\|\boldsymbol{D}_L\|^2$ is conditionally uniformly integrable, so $\mathbb{E}[\|\boldsymbol{D}_L\|^2 \mid \boldsymbol{m}] \to 0$, with these expectations uniformly bounded. Since $\|\boldsymbol{w}\| \le |\zeta|/2$ and $\boldsymbol{x}_q$ is independent of $\boldsymbol{X}$ conditional on the task,

$$\mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{D}_L \boldsymbol{x}_q)^2 \mid \boldsymbol{m}, \zeta] \le \frac{1 + \overline{m}}{4} \zeta^2 \mathbb{E}[\|\boldsymbol{D}_L\|^2 \mid \boldsymbol{m}].$$

Here the conditional law of $\boldsymbol{X}$ given $\boldsymbol{m}$ does not depend on ζ. Dominated convergence over $(\boldsymbol{m}, \zeta)$ now gives $\mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{D}_L \boldsymbol{x}_q)^2] \to 0$. Combining this with $\hat{y}_q - \hat{y}_{q,\infty} = \boldsymbol{w}^\top \boldsymbol{D}_L \boldsymbol{x}_q + N_L$ proves mean-square convergence. Finally, y_q and $\hat{y}_{q,\infty}$ have finite second moments, so convergence in L^2 implies

$$\lim_{L\to\infty} \mathbb{E}[(y_q - \hat{y}_q)^2] = \mathbb{E}[(y_q - \hat{y}_{q,\infty})^2].$$

□

For the scalar CA models, $\boldsymbol{F} = \boldsymbol{F}_T$ depends on both L and T. Parameters may depend on T but are held fixed as $L \to \infty$. From Lemma E.1, if

$$\lim_{T\to\infty} \lim_{L\to\infty} \frac{1}{L} \boldsymbol{X} \boldsymbol{F}_T^\top = \boldsymbol{I}_d$$

almost surely, then the asymptotic model predictor satisfies

$$\lim_{T\to\infty} \lim_{L\to\infty} \hat{y}_q = \langle \boldsymbol{w}, \boldsymbol{x}_q \rangle,$$

hence yielding the Bayes-optimal prediction. The fixed-depth statements hold on a common probability-one event for the countably many depths. This is the key step in establishing Theorems 6.2(2) and 6.3(2).

E.2 Solving the CA recurrence

In the one- and two-scalar simplification, we take (for all layers)

$$\boldsymbol{W}^K = \boldsymbol{W}^Q = \boldsymbol{I}_d, \qquad \boldsymbol{W}^S = \alpha \boldsymbol{I}_d, \qquad \boldsymbol{W}^V = \beta \boldsymbol{I}_d,$$

where $\alpha, \beta \in \mathbb{R}$ are scalar parameters (with the one-parameter model corresponding to $\beta = -\alpha$). Layer-wise, the model then becomes, for $t = 1, \dots, T$,

$$\boldsymbol{F}_t = \boldsymbol{F}_{t-1} + \alpha \boldsymbol{X} + \frac{\beta}{L} \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{F}_{t-1},$$

with $\boldsymbol{F}_0 = \boldsymbol{0}_{d \times L}$ and $\boldsymbol{F} = \boldsymbol{F}_T$. Define the sample covariance $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} := \frac{1}{L} \boldsymbol{X} \boldsymbol{X}^\top$. The following result yields a closed-form expression for the final output of the CA stack and establishes the parameter window for which one obtains the Bayes-predictor.

Lemma E.2. *The output of the final layer of the CA network is*

$$\boldsymbol{F} = \boldsymbol{F}_T = \alpha \sum_{k=0}^{T-1} \boldsymbol{M}^k \boldsymbol{X} \quad \text{where} \quad \boldsymbol{M} = \boldsymbol{I} + \beta \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$$

and, for $\beta \neq 0$,

$$\frac{1}{L}\boldsymbol{X}\boldsymbol{F}^\top = \frac{\alpha}{\beta}\left[(\boldsymbol{I} + \beta\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}})^T - \boldsymbol{I}\right].$$

If $\alpha = -\beta \in (0, 2/(1+\overline{m}))$, *then as context length* $L \to \infty$ *followed by the number of layers* $T \to \infty$, *we have*

$$\frac{1}{L}\boldsymbol{X}\boldsymbol{F}^\top \to \boldsymbol{I}$$

almost surely.

Proof. We prove the first statement by induction. For the base case $T = 1$, we have trivially $\boldsymbol{F}_1 = \alpha\boldsymbol{X}$ since $\boldsymbol{F}_0 = \boldsymbol{0}_{d\times L}$. Now, suppose $\boldsymbol{F}_{T-1} = \alpha\sum_{k=0}^{T-2}\boldsymbol{M}^k\boldsymbol{X}$. Then,

$$\begin{aligned}
\boldsymbol{F}_T &= \boldsymbol{F}_{T-1} + \alpha\boldsymbol{X} + \beta\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}\boldsymbol{F}_{T-1} \\
&= \alpha\boldsymbol{X} + (\boldsymbol{I} + \beta\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}})\boldsymbol{F}_{T-1} \\
&= \alpha\boldsymbol{X} + \alpha\boldsymbol{M}\sum_{k=0}^{T-2}\boldsymbol{M}^k\boldsymbol{X} \\
&= \alpha\boldsymbol{X} + \alpha\sum_{k=1}^{T-1}\boldsymbol{M}^k\boldsymbol{X} = \alpha\sum_{k=0}^{T-1}\boldsymbol{M}^k\boldsymbol{X}
\end{aligned}$$

which completes the induction. As $\boldsymbol{M}$ is symmetric and $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} = \beta^{-1}(\boldsymbol{M} - \boldsymbol{I})$, a simple computation then yields

$$\begin{aligned}
\frac{1}{L}\boldsymbol{X}\boldsymbol{F}^\top &= \alpha\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}\sum_{k=0}^{T-1}\boldsymbol{M}^k = \frac{\alpha}{\beta}\sum_{k=0}^{T-1}(\boldsymbol{M}^{k+1} - \boldsymbol{M}^k) \\
&= \frac{\alpha}{\beta}(\boldsymbol{M}^T - \boldsymbol{I}) = \frac{\alpha}{\beta}\left[(\boldsymbol{I} + \beta\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}})^T - \boldsymbol{I}\right].
\end{aligned}$$

Now, if $\alpha = -\beta$, we have

$$\frac{1}{L}\boldsymbol{X}\boldsymbol{F}^\top = \boldsymbol{I} - (\boldsymbol{I} - \alpha\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}})^T \to \boldsymbol{I} - (\boldsymbol{I} - \alpha\boldsymbol{\Sigma}_{\boldsymbol{x}})^T$$

as $L \to \infty$ by the Law of Large Numbers. As $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top$, the eigenvalues of $\boldsymbol{I} - \alpha\boldsymbol{\Sigma}_{\boldsymbol{x}}$ are $1-\alpha$ and $1-\alpha(1+\|\boldsymbol{m}\|^2)$. Hence, for the choice of $\alpha \in (0, 2/(1+\overline{m}))$, it follows that the operator norm $\|\boldsymbol{I} - \alpha\boldsymbol{\Sigma}_{\boldsymbol{x}}\| < 1$ almost surely and so $(\boldsymbol{I} - \alpha\boldsymbol{\Sigma}_{\boldsymbol{x}})^T \to \boldsymbol{0}_{d\times d}$ as $T \to \infty$ almost surely. □

Remark E.3. From the proof of Lemma E.2, we see that at asymptotic context length the error away from the Bayes-optimal prediction decays geometrically in depth T so long as the model parameters satisfy $\alpha = -\beta \in (0, 2/(1+\overline{m}))$. Hence, although the result concerns the behavior at infinite depth, in practice small to moderate depth networks should suffice for accurate in-context predictions. This intuition is supported by numerical experiments in Section 7.

F Proof of Theorem 6.2

This appendix proves Theorem 6.2 for the one-parameter model, i.e. the simplified CA stack with $\beta = -\alpha$ trained by gradient flow on the limiting population loss. Throughout, we assume Assumption 6.1: $\|\boldsymbol{m}\|^2$ is almost surely bounded and non-degenerate, with $\mathbb{P}(\boldsymbol{m} = \boldsymbol{0}_d) = 0$. Define

$$Z := 1 + \|\boldsymbol{m}\|^2, \qquad \underline{Z} := \operatorname{ess\,inf} Z, \qquad \overline{Z} := \operatorname{ess\,sup} Z,$$

so $1 \leq \underline{Z} < \overline{Z} < \infty$. We begin by determining a simplified form for the loss $\ell(\alpha)$ and collecting some of its properties.

Lemma F.1 (One-parameter loss)**.** *For each fixed integer* $T \geq 1$, *the limiting population loss* $\ell(\alpha)$ *has the form*

$$\ell(\alpha) = \mathbb{E}\left[\frac{Z-1}{Z}(1-\alpha Z)^{2T}\right]$$

up to the additive constant $\mathbb{E}[1/Z]$, *where* $Z = 1 + \|\boldsymbol{m}\|^2$. *Moreover, the loss has the following properties:*

1. *$\ell(\alpha)$ is strictly convex and coercive in the sense that*

$$\lim_{|\alpha|\to\infty} \ell(\alpha) = \infty;$$

2. *$\ell(\alpha)$ has a unique minimizer which lies in $(0, 1)$.*

Proof. Fix d, an integer $T \geq 1$, and $\alpha \in \mathbb{R}$, independently of L and the prompt. Write

$$\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} = \frac{\boldsymbol{X}\boldsymbol{X}^\top}{L}, \qquad \boldsymbol{M} = \boldsymbol{I} - \alpha\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}, \qquad \boldsymbol{Q}_T = \sum_{k=0}^{T-1} \boldsymbol{M}^k.$$

Here $\boldsymbol{M}^T$ denotes the Tth matrix power, and transposes are denoted by $\top$. The recurrence $\boldsymbol{F}_t = \boldsymbol{M}\boldsymbol{F}_{t-1} + \alpha\boldsymbol{X}$, $\boldsymbol{F}_0 = \boldsymbol{0}$, gives $\boldsymbol{F}_T = \alpha\boldsymbol{Q}_T\boldsymbol{X}$ by Lemma E.2. Since $\boldsymbol{M}$ and $\boldsymbol{Q}_T$ are symmetric polynomials in $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$, telescoping yields

$$\alpha\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}\boldsymbol{Q}_T = \alpha\boldsymbol{Q}_T\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} = \sum_{k=0}^{T-1}(\boldsymbol{M}^k - \boldsymbol{M}^{k+1}) = \boldsymbol{I} - \boldsymbol{M}^T.$$

Consequently, Lemma E.1 gives the exact readout and error

$$\hat{y}_q = \frac{\boldsymbol{y}^\top\boldsymbol{F}_T^\top\boldsymbol{x}_q}{L} = \boldsymbol{w}^\top(\boldsymbol{I} - \boldsymbol{M}^T)\boldsymbol{x}_q + \frac{\alpha}{L}\boldsymbol{\varepsilon}^\top\boldsymbol{X}^\top\boldsymbol{Q}_T\boldsymbol{x}_q,$$
$$y_q - \hat{y}_q = \boldsymbol{w}^\top\boldsymbol{M}^T\boldsymbol{x}_q + \varepsilon_q - \frac{\alpha}{L}\boldsymbol{\varepsilon}^\top\boldsymbol{X}^\top\boldsymbol{Q}_T\boldsymbol{x}_q.$$

Condition first on the full task $\tau = (\boldsymbol{m}, \zeta)$ and on $\boldsymbol{X}$. Then $\boldsymbol{w} = \zeta\boldsymbol{m}/Z$ is fixed, $\boldsymbol{x}_q$ has covariance $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top$, and the residuals are centered, independent of each other and of all covariates, with variance $\sigma_\tau^2 := \zeta^2/Z$ by Lemma C.1. All three cross terms in the squared error therefore have zero conditional expectation. Keeping the matrix order, we obtain

$$\begin{aligned}\mathbb{E}[(y_q - \hat{y}_q)^2 \mid \tau, \boldsymbol{X}] &= \boldsymbol{w}^\top\boldsymbol{M}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^T\boldsymbol{w} + \sigma_\tau^2 + \frac{\sigma_\tau^2\alpha^2}{L^2}\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{Q}_T\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Q}_T) \\ &= \boldsymbol{w}^\top\boldsymbol{M}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^T\boldsymbol{w} + \sigma_\tau^2 + \frac{\sigma_\tau^2\alpha^2}{L}\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{Q}_T\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}\boldsymbol{Q}_T).\end{aligned}$$

The last term can equivalently be written as

$$\frac{\sigma_\tau^2\alpha^2}{L}\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{Q}_T\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}\boldsymbol{Q}_T) = \frac{\sigma_\tau^2\alpha}{L}\operatorname{tr}\Big(\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{Q}_T(\boldsymbol{I} - \boldsymbol{M}^T)\Big).$$

These identities also hold at $\alpha = 0$ and for singular $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$, as no inverse is used. At finite L, $\boldsymbol{M}$ need not commute with $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, so the bias term retains $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ between the two powers of $\boldsymbol{M}$.

Let $\ell_L(\alpha) := \mathbb{E}[(y_q - \hat{y}_q)^2]$ denote the finite-context population loss. As $L \to \infty$, $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \to \boldsymbol{\Sigma}_{\boldsymbol{x}}$ almost surely. The mean-square convergence in Lemma E.1, applied with $\beta = -\alpha$ and $P_T(\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}) = \alpha\boldsymbol{Q}_T$, justifies convergence of the expected loss, not just of the predictor. In particular, its Gaussian moment and uniform-integrability argument uses the boundedness of $\boldsymbol{m}$ and only $\mathbb{E}[\zeta^2] = 1$, and it also shows that the expected context-noise contribution above vanishes. Thus, with $\boldsymbol{M}_\infty := \boldsymbol{I} - \alpha\boldsymbol{\Sigma}_{\boldsymbol{x}}$,

$$\lim_{L\to\infty} \ell_L(\alpha) = \mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{M}_\infty^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}_\infty^T\boldsymbol{w} + \sigma_\tau^2\Big].$$

This limit holds at fixed depth and parameters. Only now do we use that $\boldsymbol{M}_\infty$ commutes with $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, giving $\boldsymbol{M}_\infty^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}_\infty^T = \boldsymbol{\Sigma}_{\boldsymbol{x}}(\boldsymbol{I} - \alpha\boldsymbol{\Sigma}_{\boldsymbol{x}})^{2T}$.

To average over tasks, recall that ζ is independent of $\boldsymbol{m}$, with mean zero and second moment one. Hence,

$$\boldsymbol{\Sigma}_w := \operatorname{Cov}(\boldsymbol{w} \mid \boldsymbol{\Sigma}_{\boldsymbol{x}}) = \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top \mid \boldsymbol{\Sigma}_{\boldsymbol{x}}] = \frac{\boldsymbol{m}\boldsymbol{m}^\top}{Z^2} = \frac{1}{Z}(\boldsymbol{I} - \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}), \qquad \mathbb{E}[\sigma_\tau^2 \mid \boldsymbol{\Sigma}_{\boldsymbol{x}}] = \frac{1}{Z}.$$

Both $Z = \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) - d + 1$ and $\boldsymbol{m}\boldsymbol{m}^\top = \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{I}$ are determined by $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Write $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_d)$. When $\boldsymbol{m} \neq \boldsymbol{0}_d$, choose the first column of $\boldsymbol{U}$ along $\boldsymbol{m}$. Then $\gamma_1 = Z$, $\gamma_i = 1$ for $i \geq 2$, and $\boldsymbol{U}^\top \boldsymbol{\Sigma}_w \boldsymbol{U} = \mathrm{diag}((Z-1)/Z^2, 0, \ldots, 0)$. Therefore,

$$\mathrm{tr}\big(\boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_{\boldsymbol{x}} (\boldsymbol{I} - \alpha \boldsymbol{\Sigma}_{\boldsymbol{x}})^{2T}\big) = \frac{Z-1}{Z}(1 - \alpha Z)^{2T}.$$

The same identity holds when $\boldsymbol{m} = \boldsymbol{0}_d$, since then $Z = 1$ and $\boldsymbol{\Sigma}_w = \boldsymbol{0}$. Averaging yields

$$\lim_{L \to \infty} \ell_L(\alpha) = \mathbb{E}[1/Z] + \mathbb{E}\left[\frac{Z-1}{Z}(1 - \alpha Z)^{2T}\right].$$

As in the statement, we henceforth omit the additive constant $\mathbb{E}[1/Z]$ when writing $\ell(\alpha)$, as its derivatives and minimizers are unchanged.

1. For $Z > 1$, the function $\alpha \mapsto \frac{Z-1}{Z}(1 - \alpha Z)^{2T}$ is strictly convex, and for $Z = 1$ it is identically zero. Since $\mathbb{P}(Z > 1) > 0$ under Assumption 6.1, taking expectations preserves strict convexity. Moreover, boundedness of Z makes ℓ a polynomial of degree $2T$ with leading coefficient $\mathbb{E}[(Z-1)Z^{2T-1}] > 0$. Hence ℓ is coercive.

2. By (1), $\ell(\alpha)$ has a unique minimizer. Differentiating this polynomial gives

$$\ell'(\alpha) = -2T\mathbb{E}[(Z-1)(1 - \alpha Z)^{2T-1}].$$

On $\{Z > 1\}$, the integrand is strictly positive for $\alpha \leq 0$ and strictly negative for $\alpha \geq 1$, and it vanishes on $\{Z = 1\}$. Thus $\ell'(\alpha) < 0$ for $\alpha \leq 0$ and $\ell'(\alpha) > 0$ for $\alpha \geq 1$. The unique minimizer lies in $(0, 1)$.

□

F.1 Convergence of gradient flow

Given the above established properties of $\ell(\alpha)$, by standard dynamical systems theory results, we can derive the global convergence of gradient flow on the one-parameter loss. We include a detailed proof to initiate the unfamiliar reader.

Lemma F.2 (One-parameter gradient flow convergence)**.** *Suppose Assumption 6.1 holds. Then, gradient flow on $\ell(\alpha)$ converges globally to the unique minimum of ℓ whose existence is guaranteed by Lemma F.1.*

Proof. Let $(\alpha_t)_{t \geq 0}$ be the gradient flow trajectory initialized at $\alpha_0 \in \mathbb{R}$ and let α^* denote the unique minimizer of $\ell(\alpha)$. By chain rule and the definition of gradient flow, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell(\alpha_t) = \nabla_\alpha \ell(\alpha_t)\dot{\alpha}_t = -\|\nabla_\alpha \ell(\alpha_t)\|^2 \leq 0.$$

Hence, the trajectory of the loss function $\ell(\alpha_t)$ is decreasing and since $\ell(\alpha) \geq 0$, it follows that $\ell(\alpha_t)$ converges as $t \to \infty$. Moreover, by coercivity of ℓ, $\alpha_t \in \{\alpha : \ell(\alpha) \leq \ell(\alpha_0)\}$ for all t which is compact. Thus, since ℓ' is Lipschitz on this compact set and $|\dot{\alpha}_t| = |\ell'(\alpha_t)|$ is bounded there, $t \mapsto \ell'(\alpha_t)$ is uniformly continuous. Then, by Barbalat's Lemma (Slotine et al., 1991), $\ell'(\alpha_t) \to 0$ as $t \to \infty$. As $\alpha_t \in \{\alpha : \ell(\alpha) \leq \ell(\alpha_0)\}$ which is compact, moving to a convergent subsequence, $\alpha_{t_k} \to \bar{\alpha}$ as $k \to \infty$. By continuity of ℓ', $\ell'(\alpha_{t_k}) \to \ell'(\bar{\alpha}) = 0$. Hence, $\bar{\alpha} = \alpha^*$ which is the unique minimizer of ℓ. Hence, as every convergent subsequence of $(\alpha_t)_{t \geq 0}$ has limit α^* and the trajectory lies in a compact set, it follows that $\alpha_t \to \alpha^*$ as $t \to \infty$. □

F.2 Asymptotic depth behavior

Having thus far established the convergence of gradient flow at finite depth T, we now complete the proof of Theorem 6.2 with the following result which shows that as $T \to \infty$, we recover the optimal prediction of the model. Formally, let α_T^* denote the limit of the gradient flow on the T-layer model which, from Lemma F.2, exists and coincides with the global minimizer of $\ell(\alpha)$.

Theorem F.3. *Suppose Assumption 6.1 holds. Let α_T^* denote the global minimizer of $\ell(\alpha)$ for the T-layer CA model which exists uniquely by Lemma F.1. Then,*

$$\lim_{T\to\infty} \alpha_T^* =: \alpha_\infty^* = \frac{2}{2+\underline{m}+\overline{m}} = \frac{2}{\underline{Z}+\overline{Z}}.$$

The resulting predictor satisfies $\lim_{T\to\infty}\lim_{L_{\mathrm{te}}\to\infty} \hat{y}_q^{(\mathrm{te})} = y_q^{\mathrm{Bayes},(\mathrm{te})}$ almost surely, with the context-length limit taken first.

Proof. By Lemma F.1, the excess loss at depth T is

$$R_T(\alpha) := \mathbb{E}\big[W(1-\alpha Z)^{2T}\big], \qquad W := \frac{Z-1}{Z}.$$

The total limiting loss is $\mathbb{E}[1/Z] + R_T(\alpha)$, so R_T and $f_T(\alpha) := R_T(\alpha)^{1/(2T)}$ have the same unique minimizer $\alpha_T^* \in (0,1)$. Introduce the probability measure

$$\mathrm{d}\mu := \frac{W}{\mathbb{E}[W]}\,\mathrm{d}\mathbb{P}.$$

The continuous law of $\boldsymbol{m}$ in Assumption 6.1 implies $\mathbb{P}(\boldsymbol{m}=\mathbf{0}_d)=0$, so $Z>1$ and $W>0$ almost surely. Thus $0<\mathbb{E}[W]<1$ and μ and $\mathbb{P}$ have the same null sets. In particular, Z has the same essential endpoints $\underline{Z}, \overline{Z}$ under both measures. This also covers $\underline{Z}=1$, since although the weight vanishes at 1, every neighborhood of this endpoint has positive μ-mass. No positive endpoint-density assumption is needed.

For each fixed $\alpha\in[0,1]$, write

$$f_T(\alpha) = \mathbb{E}[W]^{1/(2T)}\|1-\alpha Z\|_{L^{2T}(\mu)}.$$

For completeness, if h is a bounded nonnegative function with $H := \|h\|_{L^\infty(\mu)} > 0$, then for every $s\in(0,H)$,

$$s\,\mu(h>s)^{1/p} \le \|h\|_{L^p(\mu)} \le H, \qquad \mu(h>s)>0.$$

Letting $p\to\infty$ and then $s\uparrow H$ proves $\|h\|_{L^p(\mu)}\to H$, and the case $H=0$ is immediate. Applied to $h=|1-\alpha Z|$, this gives

$$\lim_{T\to\infty} f_T(\alpha) = \phi(\alpha), \qquad \phi(\alpha) := \max\{|1-\alpha\underline{Z}|, |1-\alpha\overline{Z}|\}. \tag{F.1}$$

Here convexity of $z\mapsto|1-\alpha z|$ bounds its value by the endpoint maximum, while continuity and positive mass in every endpoint neighborhood give the reverse inequality for the essential supremum.

To pass from fixed parameters to minimizers, the reverse triangle inequality for $L^{2T}(\mu)$ yields, for $\alpha, a\in[0,1]$,

$$\begin{aligned}|f_T(\alpha)-f_T(a)| &\le \mathbb{E}[W]^{1/(2T)}\|(\alpha-a)Z\|_{L^{2T}(\mu)}\\ &\le \overline{Z}|\alpha-a|.\end{aligned}$$

The constant is independent of T. Consequently, for every sequence $\alpha_T\to\alpha$ in $[0,1]$, equation (F.1) implies

$$|f_T(\alpha_T)-\phi(\alpha)| \le \overline{Z}|\alpha_T-\alpha| + |f_T(\alpha)-\phi(\alpha)| \longrightarrow 0. \tag{F.2}$$

The same argument applies along any subsequence of depths tending to infinity.

Set $\alpha_\infty^* := 2/(\underline{Z}+\overline{Z}) \in (0,1)$. Its endpoint errors have equal magnitudes and opposite signs, so

$$\phi(\alpha_\infty^*) = \frac{\overline{Z}-\underline{Z}}{\overline{Z}+\underline{Z}}.$$

If $\alpha<\alpha_\infty^*$, then $\phi(\alpha)\ge 1-\alpha\underline{Z} > \phi(\alpha_\infty^*)$, and if $\alpha>\alpha_\infty^*$, then $\phi(\alpha)\ge\alpha\overline{Z}-1>\phi(\alpha_\infty^*)$. Hence α_∞^* is the unique minimizer of ϕ.

By compactness of $[0,1]$, any subsequence of (α_T^*) has a further convergent subsequence, say $\alpha_{T_k}^*\to\bar{\alpha}$. Minimality gives

$$f_{T_k}(\alpha_{T_k}^*) \le f_{T_k}(\alpha_\infty^*).$$

Using (F.1) and (F.2), we obtain $\phi(\bar{\alpha}) \le \phi(\alpha^*_\infty)$, so $\bar{\alpha} = \alpha^*_\infty$. All accumulation points therefore coincide, proving $\alpha^*_T \to \alpha^*_\infty$.

It remains to check prediction convergence with the depth-dependent parameters. Since $0 < \alpha^*_\infty < 2/\overline{Z}$, there is a compact interval $J \subset (0, 2/\overline{Z})$ containing α^*_T for all sufficiently large T. Every eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ lies in $[1, \overline{Z}]$, including the eigenvalue 1 on the orthogonal complement of $\boldsymbol{m}$. Thus,

$$\rho := \max_{a \in J,\, \lambda \in [1,\overline{Z}]} |1 - a\lambda| < 1, \qquad \|(\boldsymbol{I} - \alpha^*_T \boldsymbol{\Sigma}_{\boldsymbol{x}})^T\| \le \rho^T$$

for all sufficiently large T, almost surely. At each fixed depth, Lemmas E.1 and E.2 give

$$\lim_{L_{\mathrm{te}} \to \infty} \hat{y}^{(\mathrm{te})}_q = \boldsymbol{w}^{(\mathrm{te})\top} \left[\boldsymbol{I} - (\boldsymbol{I} - \alpha^*_T \boldsymbol{\Sigma}^{(\mathrm{te})}_{\boldsymbol{x}})^T \right] \boldsymbol{x}^{(\mathrm{te})}_q$$

on a common probability-one event for the countably many depths. The difference from $\boldsymbol{w}^{(\mathrm{te})\top} \boldsymbol{x}^{(\mathrm{te})}_q$ is bounded in absolute value by $\|\boldsymbol{w}^{(\mathrm{te})}\| \|\boldsymbol{x}^{(\mathrm{te})}_q\| \rho^T$ and therefore tends to zero. This proves the stated context-then-depth prediction limit and completes Theorem 6.2(2). □

G Proof of Theorem 6.3

We now turn our attention to the model with the pair of learnable parameters (α, β). As before, we begin by determining a simplified form for the loss $\ell(\alpha, \beta)$ and collect some properties. Recall, $Z = 1 + \|\boldsymbol{m}\|^2$.

Throughout this appendix, the two-parameter optimization analysis concerns integer depths $T \ge 2$. At fixed depth, we suppress the dependence of S, A, B, ℓ, and the profiled minimizer $\alpha^*(\beta)$ on T, and constants in fixed-depth bounds may also depend on T.

Let

$$S := S(Z, \beta) = \frac{(1 + \beta Z)^T - 1}{\beta} \quad \text{where} \quad S(Z, 0) = \lim_{\beta \to 0} S(Z, \beta) = TZ,$$

$$W = \frac{Z - 1}{Z},$$

and

$$A(\beta) = \mathbb{E}[WS^2], \quad B(\beta) = \mathbb{E}[WS].$$

Lemma G.1 (Two-parameter loss)**.** *The loss $\ell(\alpha, \beta)$ has the form*

$$\ell(\alpha, \beta) = \mathbb{E}\left[\frac{Z - 1}{Z} \left(\frac{\alpha}{\beta} \big[(1 + \beta Z)^T - 1 \big] - 1 \right)^2 \right]$$

up to the additive constant $\mathbb{E}[1/Z]$, or equivalently $\ell(\alpha, \beta) = \mathbb{E}[W(\alpha S(Z, \beta) - 1)^2]$, which also covers $\beta = 0$ through $S(Z, 0) = TZ$. Moreover, for fixed $\beta \in \mathbb{R}$, the loss is strongly convex in α, with unique minimizer:

$$\alpha^*(\beta) := \frac{B(\beta)}{A(\beta)}. \tag{G.1}$$

Proof. We have

$$\begin{aligned} \frac{\boldsymbol{y}^\top \boldsymbol{F}^\top \boldsymbol{x}_q}{L} &= \frac{\boldsymbol{w}^\top \boldsymbol{X} \boldsymbol{F}^\top \boldsymbol{x}_q}{L} + \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{F}^\top \boldsymbol{x}_q}{L} \\ &= \frac{\alpha}{\beta} \boldsymbol{w}^\top (\boldsymbol{M}^T - \boldsymbol{I}) \boldsymbol{x}_q + \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{F}^\top \boldsymbol{x}_q}{L} \\ &= -\frac{\alpha}{\beta} \boldsymbol{w}^\top \boldsymbol{x}_q + \frac{\alpha}{\beta} \boldsymbol{w}^\top (\boldsymbol{I} + \beta \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})^T \boldsymbol{x}_q + \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{F}^\top \boldsymbol{x}_q}{L} \end{aligned}$$

and so

$$y_q - \hat{y}_q = \left(1 + \frac{\alpha}{\beta} \right) \boldsymbol{w}^\top \boldsymbol{x}_q - \frac{\alpha}{\beta} \boldsymbol{w}^\top (\boldsymbol{I} + \beta \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})^T \boldsymbol{x}_q - \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{F}^\top \boldsymbol{x}_q}{L} + \varepsilon_q.$$

Taking $L \to \infty$ at fixed T, α, β, with convergence of the noise term in mean square and of the expected loss justified by Lemma E.1, and omitting $\mathbb{E}[\varepsilon_q^2|\boldsymbol{\Sigma}_{\boldsymbol{x}}]$ which does not depend on (α, β), we have, up to the additive constant $\mathbb{E}[1/Z]$,

$$\begin{aligned}\ell(\alpha,\beta|\boldsymbol{\Sigma}_{\boldsymbol{x}}) \propto \left(1+\frac{\alpha}{\beta}\right)^2 \mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x}_q)^2|\boldsymbol{\Sigma}_{\boldsymbol{x}}] + \left(\frac{\alpha}{\beta}\right)^2 \mathbb{E}[(\boldsymbol{w}^\top(\boldsymbol{I}+\beta\boldsymbol{\Sigma}_{\boldsymbol{x}})^T\boldsymbol{x}_q)^2|\boldsymbol{\Sigma}_{\boldsymbol{x}}] \\ - \frac{2\alpha}{\beta}\left(1+\frac{\alpha}{\beta}\right)\mathbb{E}[\boldsymbol{w}^\top\boldsymbol{x}_q\boldsymbol{w}^\top(\boldsymbol{I}+\beta\boldsymbol{\Sigma}_{\boldsymbol{x}})^T\boldsymbol{x}_q|\boldsymbol{\Sigma}_{\boldsymbol{x}}] \\ = \left(1+\frac{\alpha}{\beta}\right)^2 \mathrm{tr}(\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_{\boldsymbol{x}}) + \left(\frac{\alpha}{\beta}\right)^2 \mathrm{tr}(\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_{\boldsymbol{x}}(\boldsymbol{I}+\beta\boldsymbol{\Sigma}_{\boldsymbol{x}})^{2T}) \\ - \frac{2\alpha}{\beta}\left(1+\frac{\alpha}{\beta}\right)\mathrm{tr}(\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_{\boldsymbol{x}}(\boldsymbol{I}+\beta\boldsymbol{\Sigma}_{\boldsymbol{x}})^T).\end{aligned}$$

In the first line we used that $\mathbb{E}[\varepsilon_q] = 0$ to get rid of cross terms and in the second, we use that $\boldsymbol{\Sigma}_w$ commutes with $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Recalling that

$$\boldsymbol{\Sigma}_w = \frac{\boldsymbol{m}\boldsymbol{m}^\top}{(1+\|\boldsymbol{m}\|^2)^2} = \frac{1}{Z}(\boldsymbol{I} - \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}),$$

the loss simplifies to

$$\begin{aligned}\ell(\alpha,\beta|\boldsymbol{\Sigma}_{\boldsymbol{x}}) &= \left(1+\frac{\alpha}{\beta}\right)^2 \frac{1}{Z}\sum_{i=1}^d(\gamma_i-1) + \left(\frac{\alpha}{\beta}\right)^2 \frac{1}{Z}\sum_{i=1}^d(\gamma_i-1)(1+\beta\gamma_i)^{2T} \\ &\quad - 2\frac{\alpha}{\beta}\left(1+\frac{\alpha}{\beta}\right)\frac{1}{Z}\sum_{i=1}^d(\gamma_i-1)(1+\beta\gamma_i)^T \\ &= \left(1+\frac{\alpha}{\beta}\right)^2 \frac{Z-1}{Z} + \left(\frac{\alpha}{\beta}\right)^2 \frac{Z-1}{Z}(1+\beta Z)^{2T} \\ &\quad - 2\frac{\alpha}{\beta}\left(1+\frac{\alpha}{\beta}\right)\frac{Z-1}{Z}(1+\beta Z)^T \\ &= \frac{Z-1}{Z}\left(\frac{\alpha}{\beta}\left[(1+\beta Z)^T - 1\right] - 1\right)^2.\end{aligned}$$

For the unconditional loss, one simply takes expectation:

$$\ell(\alpha,\beta) = \mathbb{E}\left[\frac{Z-1}{Z}\left(\frac{\alpha}{\beta}\left[(1+\beta Z)^T - 1\right] - 1\right)^2\right].$$

Expanding the above square, we see

$$\ell(\alpha,\beta) = \alpha^2\mathbb{E}[WS^2] - 2\alpha\mathbb{E}[WS] + \mathbb{E}[W] = \alpha^2 A(\beta) - 2\alpha B(\beta) + \mathbb{E}[W]$$

is quadratic in α with leading coefficient $A(\beta) > 0$ (Lemma G.2 below)—hence strongly convex—and thus has a unique minimizer at $B(\beta)/A(\beta)$. □

In light of Lemma G.1, profiling out α, we can write a reduced loss

$$F_T(\beta) := \min_{\alpha\in\mathbb{R}} \ell(\alpha,\beta) = \ell\big(\alpha^*(\beta),\beta\big) = \mathbb{E}[W] - \frac{B(\beta)^2}{A(\beta)}.$$

At this point, we notice that neither the loss $\ell(\alpha, \beta)$ nor the reduced loss $F_T(\beta)$ are convex. Hence, the subsequent analysis establishing the convergence of gradient flow will require a different and more involved route than in the one-parameter setup.

G.1 CONVERGENCE OF GRADIENT FLOW

In this section, we establish that at sufficiently large finite depth T, gradient flow on the two-parameter loss converges under suitable initialization. We first show that, regardless of initialization, the iterates $(\alpha_t)_{t\geq 0}$ remain bounded. Then, for a fixed initialization $\beta_0 \in (-2/\overline{Z}, 0)$ and $\alpha_0 = \alpha^*(\beta_0)$, we show that the iterates $(\beta_t)_{t\geq 0}$ remain in a compact subinterval of $(-2/\overline{Z}, 0)$, which does not depend on the depth, once $T \geq T'(\beta_0)$, as specified in Lemma G.6. Under this setup where the iterates $(\alpha_t, \beta_t)_{t\geq 0}$ remain bounded, we conclude by showing they must converge as $t \to \infty$. No property of the law of Z beyond Assumption 6.1 is used in this appendix.

We first establish some preliminary properties of the function $A(\beta)$.

Lemma G.2. *Suppose Assumption 6.1 holds. Then,*

1. *$A(\beta)$ is continuous and $\lim_{|\beta|\to\infty} A(\beta) = +\infty$.*
2. *$A^* := \inf_{\beta\in\mathbb{R}} A(\beta) > 0$.*

Proof. 1. The binomial expansion gives the polynomial

$$S(Z,\beta) = \sum_{j=1}^{T} \binom{T}{j} \beta^{j-1} Z^j.$$

Since Z is bounded, A is a polynomial and hence continuous. For each fixed $T \geq 2$,

$$A(\beta) = \mathbb{E}[(Z-1)Z^{2T-1}]\,\beta^{2T-2} + O(|\beta|^{2T-3}) \qquad (|\beta| \to \infty).$$

Its leading coefficient is strictly positive because $\mathbb{P}(Z > 1) > 0$. The degree $2T-2$ is positive and even, so $A(\beta) \to \infty$ as $|\beta| \to \infty$.

2. By part 1, we may choose $C > 0$ so large that $A(\beta) \geq A(0)$ for all $|\beta| > C$, and then $A^* = \min_{|\beta|\leq C} A(\beta) =: A_{\min}$, where the minimum is attained by continuity of A. Assume for contradiction that $A_{\min} = 0$. Then, $S(Z,\beta^*) = 0$ almost surely on $\{Z > 1\}$ for some β^*, and $Z > 1$ almost surely under Assumption 6.1, since the continuous law of $\boldsymbol{m}$ gives $\mathbb{P}(\boldsymbol{m} = \mathbf{0}_d) = 0$. If $\beta^* = 0$, this is impossible because $S(Z,0) = TZ > 0$. If $\beta^* \neq 0$, it forces $(1+\beta^* Z)^T = 1$ almost surely. For odd T, the only real solution of $r^T = 1$ is $r = 1$, which would give $Z = 0$, and for even T, the real solutions are $r = \pm 1$, where $r = 1$ is again impossible and $r = -1$ would make $Z = -2/\beta^*$ almost surely constant. Both cases contradict $1 \leq \underline{Z} < \overline{Z}$. Hence $A^* = A_{\min} > 0$.

□

We can show that the iterates $(\alpha_t)_{t\geq 0}$ remain bounded regardless of initialization.

Lemma G.3. *Suppose Assumption 6.1 holds. Then,*

$$\sup_{\beta\in\mathbb{R}} |\alpha^*(\beta)| \leq \sqrt{\frac{\mathbb{E}[W]}{A^*}} =: M^* < \infty.$$

Moreover, letting $(\alpha_t, \beta_t)_{0\leq t<\tau}$ be any gradient flow trajectory on ℓ, defined on an interval $[0,\tau)$ with $0 < \tau \leq \infty$, we have

$$\sup_{0\leq t<\tau} |\alpha_t| \leq M^* + \sqrt{\frac{\ell(\alpha_0,\beta_0)}{A^*}} =: M_{\alpha_0,\beta_0} < \infty.$$

Proof. By the Cauchy-Schwarz inequality,

$$B(\beta)^2 = \mathbb{E}[W^{1/2} \cdot W^{1/2} S(\beta)]^2 \leq \mathbb{E}[W]\mathbb{E}[WS(\beta)^2] = \mathbb{E}[W]A(\beta).$$

By Lemma G.2,

$$\sup_{\beta\in\mathbb{R}} |\alpha^*(\beta)| = \sup_{\beta\in\mathbb{R}} \frac{|B(\beta)|}{A(\beta)} \leq \sqrt{\frac{\mathbb{E}[W]}{A^*}} = M^* < \infty.$$

Now, setting $\Delta_t = \alpha_t - \alpha^*(\beta_t)$, notice that we may write

$$\ell(\alpha, \beta) = A(\beta)\Delta^2 + \underbrace{\mathbb{E}[W] - \frac{B(\beta)^2}{A(\beta)}}_{:=C(\beta)\geq 0}$$

where the non-negativity of $C(\beta)$ follows from the inequality $B(\beta)^2 \leq \mathbb{E}[W]A(\beta)$. Recall that the loss ℓ is non-increasing along gradient flow on $[0, \tau)$, which follows from $\frac{\mathrm{d}}{\mathrm{d}t}\ell(\alpha_t, \beta_t) = -\|\nabla\ell(\alpha_t, \beta_t)\|^2 \leq 0$. Hence, $\ell(\alpha_t, \beta_t) \leq \ell(\alpha_0, \beta_0)$ for all $t \in [0, \tau)$ and so

$$A(\beta_t)\Delta_t^2 = \ell(\alpha_t, \beta_t) - C(\beta_t) \leq \ell(\alpha_0, \beta_0).$$

Since $A(\beta_t) \geq A^*$ by Lemma G.2, we obtain $\sup_{0\leq t<\tau} |\Delta_t| \leq \sqrt{\ell(\alpha_0, \beta_0)/A^*}$ and so

$$\sup_{0\leq t<\tau} |\alpha_t| \leq \sup_{0\leq t<\tau} |\alpha^*(\beta_t)| + \sup_{0\leq t<\tau} |\Delta_t| \leq M^* + \sqrt{\frac{\ell(\alpha_0, \beta_0)}{A^*}} < \infty$$

as desired. □

Now, focusing on the iterates $(\beta_t)_{t\geq 0}$, we show that, for all sufficiently large depths, they remain in a single compact subinterval of $(-2/\overline{Z}, 0)$ which does not depend on the depth. We use the same weighted probability measure as in Appendix F, namely

$$\mathrm{d}\mu = \frac{W}{\mathbb{E}[W]}\,\mathrm{d}\mathbb{P}$$

so that μ is obtained by reweighting $\mathrm{d}\mathbb{P}$ by W and $\mathbb{E}_\mu[Y] = \mathbb{E}[WY]/\mathbb{E}[W]$ for a random variable Y. As noted in the proof of Theorem F.3, Assumption 6.1 gives $Z > 1$ and $W \geq 0$ almost surely, so μ and $\mathbb{P}$ have the same null sets and Z has the same essential endpoints $\underline{Z} < \overline{Z}$ under both measures. In particular,

$$\mu(Z \leq x) > 0 \quad \text{and} \quad \mu(Z \geq y) > 0 \qquad \text{for all } \underline{Z} < x < y < \overline{Z}. \tag{G.2}$$

Beyond $Z > 1$ almost surely and $1 \leq \underline{Z} < \overline{Z} < \infty$, the positivity of the μ-mass near the two essential endpoints is the only property of the law of Z used in the remainder of this appendix, so in particular Z need not have a density and may have atoms.

Writing

$$u_T(\beta, Z) := (1 + \beta Z)^T,$$

a short calculation shows

$$F_T(\beta) = \mathbb{E}[W]\frac{\mathrm{Var}_\mu(u_T(\beta, Z))}{\mathbb{E}_\mu\left[(1 - u_T(\beta, Z))^2\right]} \tag{G.3}$$

for any $\beta \neq 0$ which will provide a convenient representation of the loss in subsequent proofs. For $\beta \in (-2/\overline{Z}, 0)$ we also define

$$\phi(\beta) := \max\{|1 + \beta\underline{Z}|, |1 + \beta\overline{Z}|\}. \tag{G.4}$$

For such β and $z \in [\underline{Z}, \overline{Z}]$ we have $1 + \beta z \in (-1, 1)$, so $\phi(\beta) < 1$, and also $\phi(\beta) > 0$, since $1 + \beta\underline{Z}$ and $1 + \beta\overline{Z}$ cannot both vanish. By convexity of $z \mapsto |1 + \beta z|$,

$$|u_T(\beta, z)| \leq \phi(\beta)^T \qquad \text{for all } z \in [\underline{Z}, \overline{Z}], \tag{G.5}$$

and, for every $\beta \in [-2/\overline{Z}, 0]$, $|u_T(\beta, z)| \leq 1$ on $[\underline{Z}, \overline{Z}]$, so the denominator in (G.3) is at most 4. Finally, we record an elementary lower bound for variances. If Z' is an independent copy of Z under μ, then $\mathrm{Var}_\mu(h(Z)) = \frac{1}{2}\mathbb{E}_\mu[(h(Z) - h(Z'))^2]$, and hence, if $\mathcal{A}, \mathcal{B} \subset \mathbb{R}$ are disjoint Borel sets with $|h(z) - h(z')| \geq \vartheta$ for all $z \in \mathcal{A}$ and $z' \in \mathcal{B}$, then, restricting the integration to $\{Z \in \mathcal{A},\, Z' \in \mathcal{B}\} \cup \{Z \in \mathcal{B},\, Z' \in \mathcal{A}\}$,

$$\mathrm{Var}_\mu(h(Z)) \geq \mu(Z \in \mathcal{A})\,\mu(Z \in \mathcal{B})\,\vartheta^2. \tag{G.6}$$

Lemma G.4 (Loss near the boundary of the stability interval)**.** *Suppose Assumption 6.1 holds and fix $\rho \in (0, 1)$. There exist constants $\delta_L, \delta_R \in (0, 1/(4\overline{Z})]$, $c_L, c_R > 0$, $r_L, r_R \in (\rho, 1)$ and an integer $N_L \geq 1$, depending only on ρ and on the law of Z, such that*

1. *$F_T(\beta) \geq c_R\, r_R^{2T}$ for every integer $T \geq 1$ and every $\beta \in [-\delta_R, 0]$, and*
2. *$F_T(\beta) \geq c_L\, r_L^{2T}$ for every integer $T \geq N_L$ and every $\beta \in [-2/\overline{Z}, -2/\overline{Z} + \delta_L]$.*

Proof. Write $a := \underline{Z}$, $b := \overline{Z}$, and fix $x := a + (b-a)/3$ and $y := a + 2(b-a)/3$, so that $a < x < y < b$. Set $\Delta := y - x > 0$ and

$$p := \mu(Z \leq x) > 0, \qquad q := \mu(Z \geq y) > 0,$$

where the positivity follows from (G.2).

The boundary at zero. Define

$$\delta_R := \min\left\{\frac{1}{4b}, \frac{1-\rho}{2x}\right\}, \qquad r_R := 1 - \delta_R x, \qquad c_R := \mathbb{E}[W]\, p\, q\, (1 - e^{-\Delta/b})^2.$$

Since $\delta_R x \leq (1-\rho)/2$, we have $r_R \geq (1+\rho)/2 > \rho$, and clearly $r_R < 1$. Fix $T \geq 1$ and $\beta = -s$ with $0 < s \leq \delta_R$, and put $h := Ts$. Since $sz \leq sb \leq 1/4$ for $z \in [a, b]$, the function $z \mapsto u_T(-s, z) = (1 - sz)^T$ is positive and decreasing on $[a, b]$. Hence, for $z \leq x$ and $z' \geq y$,

$$\begin{aligned} u_T(-s,z) - u_T(-s,z') &\geq (1-sx)^T - (1-sy)^T = (1-sx)^T \left[1 - \left(1 - \frac{s\Delta}{1-sx}\right)^T\right] \\ &\geq (1-sx)^T \left(1 - e^{-h\Delta}\right), \end{aligned}$$

where we used $0 < s\Delta/(1-sx) < 1$, the inequality $1 - v \leq e^{-v}$, and $(1-sx)^{-1} \geq 1$. By (G.6) with $\mathcal{A} = (-\infty, x]$ and $\mathcal{B} = [y, \infty)$,

$$\mathrm{Var}_\mu(u_T(-s, Z)) \geq p\, q\, (1-sx)^{2T} \left(1 - e^{-h\Delta}\right)^2. \tag{G.7}$$

At the same time, Bernoulli's inequality $(1-v)^T \geq 1 - Tv$ for $v \in [0,1]$ gives, pointwise on $[a, b]$,

$$0 \leq 1 - (1 - sZ)^T \leq \min\{1, TsZ\} \leq \min\{1, hb\},$$

and hence

$$\mathbb{E}_\mu\left[(1 - u_T(-s, Z))^2\right] \leq \min\{1, hb\}^2.$$

Finally, for every $h > 0$,

$$1 - e^{-h\Delta} \geq (1 - e^{-\Delta/b}) \min\{1, hb\} : \tag{G.8}$$

if $hb \leq 1$, this follows from the concavity of $v \mapsto 1 - e^{-v\Delta/b}$ on $[0,1]$, evaluated at $v = hb$, and if $hb \geq 1$, from monotonicity. Combining (G.3), (G.7), the denominator bound and (G.8), the factors $\min\{1, hb\}^2$ cancel and

$$F_T(-s) \geq \mathbb{E}[W]\, p\, q\, (1 - e^{-\Delta/b})^2 (1-sx)^{2T} \geq c_R\, r_R^{2T}.$$

The bound is uniform in $h = Ts$, so in particular it covers the regime $T|\beta| \to 0$, in which both the variance and the denominator in (G.3) are of order $\min\{1, hb\}^2$. Since $F_T = \mathbb{E}[W] - B^2/A$ is continuous on $\mathbb{R}$ by Lemma G.2, the bound extends to $s = 0$, proving (1).

The boundary at $-2/\overline{Z}$. Let

$$q_0 := \max\{|1 - 2a/b|, |1 - 2x/b|\} < 1,$$

where the strict inequality holds because $0 < a < x < b$. Choose

$$\eta := \min\left\{\frac{b-x}{2}, \frac{b\,(1 - \max\{\rho, q_0\})}{8}\right\}, \qquad r_0 := 1 - \frac{2\eta}{b},$$

so that $b - \eta > x$ and $r_0 \geq 1 - (1 - \max\{\rho, q_0\})/4 > \max\{\rho, q_0\}$. Define

$$\delta_L := \frac{\min\{1,\, r_0 - \rho,\, r_0 - q_0\}}{4b}, \qquad r_L := r_0 - b\delta_L, \qquad Q_L := q_0 + b\delta_L,$$

$$q' := \mu(Z \geq b - \eta) > 0, \qquad c_L := \frac{\mathbb{E}[W]\, p\, q'}{16}.$$

Then $r_L \geq (3r_0+\rho)/4 > \rho$, $r_L < 1$, and $r_L - Q_L \geq (r_0 - q_0)/2 > 0$. Fix $\beta = -2/b + s$ with $0 \leq s \leq \delta_L$. For $z \in [a, x]$, since the affine function $z \mapsto 1 - 2z/b$ attains the maximum of its absolute value over $[a, x]$ at an endpoint,

$$|1 + \beta z| \leq |1 - 2z/b| + sz \leq q_0 + b\delta_L = Q_L,$$

where the absolute value in the definition of q_0 is needed because $1 - 2a/b < 0$ when $a > b/2$. For $z \in [b - \eta, b]$,

$$1 + \beta z = 1 - \frac{2z}{b} + sz \leq 1 - \frac{2(b-\eta)}{b} + \delta_L b = -r_L < 0.$$

Consequently $|u_T(\beta, z)| \leq Q_L^T$ for $z \leq x$ and $|u_T(\beta, z')| \geq r_L^T$ for $z' \geq b - \eta$, so that, by the reverse triangle inequality and for either parity of T,

$$|u_T(\beta, z) - u_T(\beta, z')| \geq r_L^T - Q_L^T \qquad (z \leq x,\ z' \geq b - \eta).$$

Since the denominator in (G.3) is at most 4 on $[-2/b, 0]$, (G.6) with $\mathcal{A} = (-\infty, x]$ and $\mathcal{B} = [b-\eta, \infty)$ yields

$$F_T(\beta) \geq \frac{\mathbb{E}[W]}{4}\, p\, q'\, \big(r_L^T - Q_L^T\big)^2.$$

Let $N_L \geq 1$ be such that $(Q_L/r_L)^T \leq 1/2$ for all $T \geq N_L$, which is possible because $Q_L < r_L$. For $T \geq N_L$, $(r_L^T - Q_L^T)^2 \geq r_L^{2T}/4$, and (2) follows. All constants depend only on ρ and on the law of Z. □

Proposition G.5 (Depth-uniform confinement of small sublevel sets)**.** *Suppose Assumption 6.1 holds, and fix $\rho \in (0, 1)$ and $C > 0$. Let δ_L, δ_R be the constants of Lemma G.4 for this ρ, and set*

$$K_0 := \big[-2/\overline{Z} + \delta_L,\ -\delta_R\big].$$

Then K_0 is a nonempty compact subinterval of $(-2/\overline{Z}, 0)$, and there exists an integer $T_C \geq 2$, depending only on ρ, C and the law of Z, such that for every $T \geq T_C$,

$$\big\{\beta \in [-2/\overline{Z}, 0] : F_T(\beta) \leq C\rho^{2T}\big\} \subset K_0. \tag{G.9}$$

Proof. Since $\delta_L + \delta_R \leq 1/(2\overline{Z}) < 2/\overline{Z}$, the interval K_0 is nonempty, and it lies inside $(-2/\overline{Z}, 0)$ because $\delta_L, \delta_R > 0$. As $r_L, r_R > \rho$ and $c_L, c_R > 0$, there is an integer $T_C \geq \max\{2, N_L\}$ such that $C\rho^{2T} < \min\{c_L r_L^{2T}, c_R r_R^{2T}\}$ for all $T \geq T_C$, and for instance any integer $T_C \geq \max\{2, N_L\}$ with $T_C > [\log(C/c_j)]_+/(2\log(r_j/\rho))$ for $j \in \{L, R\}$ works, where $[v]_+ := \max\{v, 0\}$. For such T, Lemma G.4 gives $F_T(\beta) > C\rho^{2T}$ on $[-2/\overline{Z}, -2/\overline{Z} + \delta_L] \cup [-\delta_R, 0]$, which is the complement of the interior of K_0 in $[-2/\overline{Z}, 0]$. This proves (G.9), and in fact the sublevel set lies in the interior of K_0. □

Lemma G.6 (Initialization below the boundary barriers)**.** *Suppose Assumption 6.1 holds, and that one initializes gradient flow at $\alpha_0 = \alpha^*(\beta_0)$ with $\beta_0 \in (-2/\overline{Z}, 0)$. Set $\rho_0 := \phi(\beta_0) \in (0, 1)$, with ϕ as in (G.4), and $C_0 := \mathbb{E}[W]/(1-\rho_0)^2$. Then, for every integer $T \geq 1$,*

$$\ell(\alpha_0, \beta_0) = F_T(\beta_0) \leq \frac{\mathbb{E}[W]\,\rho_0^{2T}}{(1-\rho_0^T)^2} \leq C_0\,\rho_0^{2T}. \tag{G.10}$$

Let $K_0 = K_0(\beta_0)$ and $T' = T'(\beta_0) := T_{C_0} \geq 2$ be the compact interval and the threshold provided by Proposition G.5 for $(\rho, C) = (\rho_0, C_0)$, and both depend on β_0 and on the law of Z, but not on the depth. Then, for all $T \geq T'$,

$$\ell(\alpha_0, \beta_0) < \min\{\ell(\alpha, -2/\overline{Z}), \ell(\alpha, 0)\}$$

for all $\alpha \in \mathbb{R}$, and, since gradient flow is non-increasing on the loss ℓ, one has $\beta_t \in K_0 \subset (-2/\overline{Z}, 0)$ for every t in the interval of existence of the trajectory.

Proof. Since $\alpha_0 = \alpha^*(\beta_0)$, we have $\ell(\alpha_0, \beta_0) = F_T(\beta_0)$. By (G.5), $|u_T(\beta_0, Z)| \leq \rho_0^T$ almost surely, so $\mathrm{Var}_\mu(u_T(\beta_0, Z)) \leq \mathbb{E}_\mu[u_T(\beta_0, Z)^2] \leq \rho_0^{2T}$, while $\mathbb{E}_\mu[(1-u_T(\beta_0, Z))^2] \geq (1-\rho_0^T)^2 \geq (1-\rho_0)^2$. Inserting these bounds into (G.3) proves (G.10).

Now let $T \geq T'$. Recall from the proof of Lemma G.3 that $\ell(\alpha, \beta) = A(\beta)(\alpha - \alpha^*(\beta))^2 + F_T(\beta) \geq F_T(\beta)$ for all $\alpha \in \mathbb{R}$. The two boundary points $-2/\overline{Z}$ and 0 lie outside K_0, so Proposition G.5 gives

$F_T(-2/\overline{Z}) > C_0\rho_0^{2T}$ and $F_T(0) > C_0\rho_0^{2T}$, and the strict inequality follows from (G.10). For the last claim, note that along the trajectory, for every $t \geq 0$ at which it is defined,

$$F_T(\beta_t) \leq \ell(\alpha_t, \beta_t) \leq \ell(\alpha_0, \beta_0) \leq C_0\rho_0^{2T},$$

so (G.9) places β_t in K_0 whenever $\beta_t \in [-2/\overline{Z}, 0]$. In particular $\beta_0 \in K_0$. Since $t \mapsto \beta_t$ is continuous, leaving K_0 would require first visiting $[-2/\overline{Z}, 0] \setminus K_0$, which is impossible. Hence $\beta_t \in K_0$ for every t in the interval of existence of the trajectory. □

Theorem G.7. *Suppose Assumption 6.1 holds, and that one initializes gradient flow at $\alpha_0 = \alpha^*(\beta_0)$ with $\beta_0 \in (-2/\overline{Z}, 0)$. Let $T' = T'(\beta_0) \geq 2$ and $K_0 = K_0(\beta_0) \subset (-2/\overline{Z}, 0)$ be the threshold and the compact interval from Lemma G.6. Then, for every integer depth $T \geq T'$, the gradient flow trajectory $(\alpha_{t,T}, \beta_{t,T})_{t\geq 0}$ exists for all $t \geq 0$ and converges, with*

$$\beta_T^* := \lim_{t\to\infty} \beta_{t,T} \in K_0$$

and

$$\alpha_T^* := \lim_{t\to\infty} \alpha_{t,T} = \alpha^*(\beta_T^*).$$

Moreover, $F_T(\beta_T^*) \leq \ell(\alpha_0, \beta_0) \leq C_0\rho_0^{2T}$, with C_0, ρ_0 as in Lemma G.6.

Proof. Fix $T \geq T'(\beta_0)$ as in the statement. The gradient $\nabla\ell$ is a polynomial in (α, β), hence locally Lipschitz, so the gradient flow has a unique maximal solution. On its interval of existence, ℓ is non-increasing along the flow, so Lemma G.6 gives $\beta_t \in K_0$, and Lemma G.3 gives $|\alpha_t| \leq M_{\alpha_0,\beta_0}$. The trajectory therefore remains in the compact rectangle $[-M_{\alpha_0,\beta_0}, M_{\alpha_0,\beta_0}] \times K_0$, on which the vector field is bounded, and the standard continuation criterion for ordinary differential equations shows that the solution exists for all $t \geq 0$. In particular, the gradient flow trajectory $(\alpha_t, \beta_t)_{t\geq 0}$ is bounded. Since ℓ is real-analytic on the open set $\mathbb{R} \times (-2/\overline{Z}, 0)$, the bounded trajectory $(\alpha_t, \beta_t)_{t\geq 0}$ converges to a single critical point by the Łojasiewicz gradient inequality. At a critical point, $\partial_\alpha \ell = 2A(\beta)\alpha - 2B(\beta) = 0$, so $\alpha_T^* = \alpha^*(\beta_T^*)$, and since $\beta_t \in K_0$ for all $t \geq 0$ and K_0 is closed, $\beta_T^* \in K_0$. Finally, by continuity of ℓ and its monotonicity along the flow, $F_T(\beta_T^*) = \ell(\alpha_T^*, \beta_T^*) = \lim_{t\to\infty} \ell(\alpha_t, \beta_t) \leq \ell(\alpha_0, \beta_0)$. □

G.2 Asymptotic depth behavior

Fix $\beta_0 \in (-2/\overline{Z}, 0)$ independently of depth, with $\alpha_0 = \alpha^*(\beta_0)$ at each depth, and let $T' = T'(\beta_0)$ be the threshold above. Having established that gradient flow converges to a stationary point at each fixed depth $T \geq T'$, we now study how these stationary points behave as $T \to \infty$. Our goal is to understand the asymptotic depth behavior of the limits of gradient flow and to show that, as in the one-parameter case, the limiting predictor is Bayes optimal. To make explicit the dependence on depth, let $\ell^{(T)}$ denote the loss ℓ where the CA model is at depth T.

In the two-parameter setting, the loss is not jointly convex, so we cannot identify the gradient-flow limit (α_T^*, β_T^*) as a global minimizer of ℓ. However, for each $T \geq T'$, what is guaranteed by Theorem G.7 is that

$$(\alpha_T^*, \beta_T^*) \text{ is a stationary point of } \ell^{(T)}, \qquad \alpha_T^* = \alpha^*(\beta_T^*), \qquad \text{and} \qquad \beta_T^* \in K_0,$$

where the compact interval $K_0 = K_0(\beta_0) \subset (-2/\overline{Z}, 0)$ does not depend on T, so in particular

$$F_T'(\beta_T^*) = 0.$$

Thus, the large-depth behavior of the gradient-flow limits is governed by the stationary points of the one-dimensional function $\beta \mapsto F_T(\beta)$. It will be convenient to work with a logarithmically rescaled version of F_T:

$$G_T(\beta) := \frac{1}{2T} \log\left(\frac{F_T(\beta)}{\mathbb{E}[W]}\right). \tag{G.11}$$

The function G_T is well-defined and smooth on any interval where $F_T > 0$, and Lemma G.8 below shows that this is the case, for all sufficiently large T, on every compact interval $K \subset I \setminus \{-\alpha^*\}$, where $I := (-2/\overline{Z}, 0)$ and $\alpha^* = 2/(\underline{Z} + \overline{Z})$. On such an interval, G_T is a strictly monotone transform of F_T, so $F_T'(\beta) = 0$ if and only if $G_T'(\beta) = 0$. Our task is therefore to understand the asymptotic behavior of G_T and G_T' as $T \to \infty$, and to show that stationary points of F_T (and G_T) concentrate near a single value of β.

Lemma G.8 (Uniform endpoint localization)**.** *Suppose Assumption 6.1 holds, and fix any compact interval $K \subset I \setminus \{-\alpha^*\}$, where $I := (-2/\overline{Z}, 0)$ and $\alpha^* := 2/(\underline{Z} + \overline{Z})$. Let $z^* \in \{\underline{Z}, \overline{Z}\}$ denote the endpoint at which $\phi(\beta) = |1 + \beta z^*|$ is attained, which is the same for every $\beta \in K$. Write $q_\beta(z) := 1 + \beta z$ and, with all expectations under μ,*

$$m_T := \mathbb{E}_\mu[u_T(\beta, Z)], \qquad Q_T := \mathbb{E}_\mu[u_T(\beta, Z)^2],$$
$$V_T := Q_T - m_T^2, \qquad D_T := \mathbb{E}_\mu\big[(1 - u_T(\beta, Z))^2\big],$$

so that $F_T = \mathbb{E}[W]\, V_T / D_T$ by (G.3), where $Q_T(\beta) > 0$, since $q_\beta(Z) = 0$ μ-almost surely would force $Z = -1/\beta$ almost surely, contradicting $\underline{Z} < \overline{Z}$. For each $T \in \mathbb{N}$ and $\beta \in K$ define the probability measure

$$\mathrm{d}\nu_{T,\beta}(z) := \frac{|1 + \beta z|^{2T}}{Q_T(\beta)}\, \mathrm{d}\mu(z).$$

Then the following hold uniformly for $\beta \in K$ as $T \to \infty$:

1. *$\nu_{T,\beta}$ converges weakly to the point mass at the active endpoint z^*,*
2. *the localization functional*

$$\mathcal{L}_T(\beta) := \frac{\mathbb{E}_\mu\big[(Z - z^*)^2\, |q_\beta(Z)|^{2T-2}\big]}{Q_T(\beta)}$$

   satisfies $\sup_{\beta \in K} \mathcal{L}_T(\beta) \longrightarrow 0$,
3. *there exist $\kappa_K \in (0, 1)$ and $N_K < \infty$ such that $V_T(\beta) \geq \kappa_K Q_T(\beta) > 0$ for all $\beta \in K$ and $T \geq N_K$, so in particular $F_T > 0$ on K for all $T \geq N_K$, and*
4. *the reduced loss satisfies $F_T(\beta)^{1/(2T)} \longrightarrow \phi(\beta)$.*

Proof. Write $a := \underline{Z}$, $b := \overline{Z}$. Since K is an interval that omits $-\alpha^*$, the active endpoint is constant on K, with $z^* = a$ when $\beta > -\alpha^*$ and $z^* = b$ when $\beta < -\alpha^*$. In the first case $q_\beta(a) > 0$, and in the second $q_\beta(b) < 0$. By compactness,

$$s_+ := \max_{\beta \in K} |\beta| < \infty, \qquad 0 < \phi_- := \min_{\beta \in K} \phi(\beta) \leq \phi(\beta) \leq \rho_K := \max_{\beta \in K} \phi(\beta) < 1.$$

All constants below may depend on K and μ, but not on $\beta \in K$ or on T.

Two basic estimates. By (G.5), $|q_\beta(z)| \leq \phi(\beta)$ on $[a, b]$, hence

$$Q_T(\beta) \leq \phi(\beta)^{2T}. \tag{G.12}$$

For the reverse direction, fix $\varepsilon \in (0, b-a)$. On the compact set $\{(\beta, z) : \beta \in K,\ z \in [a, b],\ |z - z^*| \geq \varepsilon\}$ the continuous ratio $|q_\beta(z)|/\phi(\beta)$ is strictly less than one, because z^* is the unique maximizer of $z \mapsto |q_\beta(z)|$ on $[a, b]$. Hence there is $r_\varepsilon \in [0, 1)$ with

$$|q_\beta(z)| \leq r_\varepsilon\, \phi(\beta) \qquad \text{whenever } \beta \in K,\ z \in [a, b],\ |z - z^*| \geq \varepsilon. \tag{G.13}$$

Choose $\delta_\varepsilon \in (0, \varepsilon)$ so small that $\theta_\varepsilon := 1 - s_+ \delta_\varepsilon / \phi_-$ satisfies $r_\varepsilon < \theta_\varepsilon < 1$. If $|z - z^*| < \delta_\varepsilon$, the reverse triangle inequality gives $|q_\beta(z)| \geq \phi(\beta) - s_+ \delta_\varepsilon \geq \theta_\varepsilon \phi(\beta)$, using $\phi(\beta) \geq \phi_-$. Since z^* is an essential endpoint of Z under μ, (G.2) gives $p_\varepsilon := \mu(|Z - z^*| < \delta_\varepsilon) > 0$, and therefore

$$Q_T(\beta) \geq p_\varepsilon\, \theta_\varepsilon^{2T}\, \phi(\beta)^{2T} \qquad (\beta \in K,\ T \geq 1). \tag{G.14}$$

Concentration at the active endpoint. By (G.13) and (G.14),

$$\nu_{T,\beta}(|z - z^*| \geq \varepsilon) \leq \frac{(r_\varepsilon \phi(\beta))^{2T}}{Q_T(\beta)} \leq \frac{1}{p_\varepsilon} \left(\frac{r_\varepsilon}{\theta_\varepsilon}\right)^{2T} \longrightarrow 0 \tag{G.15}$$

uniformly on K. For any continuous $f : [a, b] \to \mathbb{R}$, let $\omega_f(\varepsilon) := \sup_{|z - s| \leq \varepsilon} |f(z) - f(s)|$. Then, by (G.15),

$$\sup_{\beta \in K} \left| \int f\, \mathrm{d}\nu_{T,\beta} - f(z^*) \right| \leq \omega_f(\varepsilon) + \frac{2\|f\|_\infty}{p_\varepsilon} \left(\frac{r_\varepsilon}{\theta_\varepsilon}\right)^{2T}.$$

First let $T \to \infty$ and then $\varepsilon \downarrow 0$. Uniform continuity of f proves the uniform weak convergence in part (1).

Localization of the second moment. Take ε so small that, in addition, $s_+\varepsilon \le \phi_-/2$. On $\{|Z - z^*| < \varepsilon\}$ we then have $|q_\beta(Z)| \ge \phi(\beta) - s_+\varepsilon \ge \phi_-/2$, hence

$$\mathbb{E}_\mu\big[(Z-z^*)^2|q_\beta(Z)|^{2T-2}\,\mathbf{1}_{\{|Z-z^*|<\varepsilon\}}\big] \le \frac{4\varepsilon^2}{\phi_-^2}\,\mathbb{E}_\mu\big[|q_\beta(Z)|^{2T}\,\mathbf{1}_{\{|Z-z^*|<\varepsilon\}}\big] \le \frac{4\varepsilon^2}{\phi_-^2}\,Q_T(\beta).$$

On the complementary event, (G.13) and (G.14) give, for $T \ge 1$,

$$\frac{\mathbb{E}_\mu\big[(Z-z^*)^2|q_\beta(Z)|^{2T-2}\,\mathbf{1}_{\{|Z-z^*|\ge\varepsilon\}}\big]}{Q_T(\beta)} \le \frac{(b-a)^2\,(r_\varepsilon\phi(\beta))^{2T-2}}{p_\varepsilon\,\theta_\varepsilon^{2T}\,\phi(\beta)^{2T}} \le \frac{(b-a)^2}{p_\varepsilon\,\theta_\varepsilon^2\,\phi_-^2}\left(\frac{r_\varepsilon}{\theta_\varepsilon}\right)^{2T-2},$$

which tends to zero uniformly on K (if $r_\varepsilon = 0$ the right-hand side is interpreted as zero). Consequently $\limsup_{T\to\infty}\sup_{\beta\in K}\mathcal{L}_T(\beta) \le 4\varepsilon^2/\phi_-^2$, and letting $\varepsilon \downarrow 0$ proves part (2).

Variance comparability. Let $\mathcal{A} := \{Z \le (a+b)/2\}$ if $z^* = b$ and $\mathcal{A} := \{Z \ge (a+b)/2\}$ if $z^* = a$, so that $p := \mu(\mathcal{A}) > 0$ by (G.2), and let $r := r_{(b-a)/2}$, $\theta := \theta_{(b-a)/2}$ and $p' := p_{(b-a)/2}$ be the constants of (G.13) and (G.14) for $\varepsilon = (b-a)/2$. On $\mathcal{A}$ we have $|Z - z^*| \ge (b-a)/2$, so $|u_T(\beta, Z)| \le (r\phi(\beta))^T$ there, and (G.14) yields

$$\frac{|\mathbb{E}_\mu[u_T(\beta,Z)\mathbf{1}_\mathcal{A}]|}{\sqrt{Q_T(\beta)}} \le \frac{(r\phi(\beta))^T}{\sqrt{p'}\,\theta^T\phi(\beta)^T} = \frac{1}{\sqrt{p'}}\left(\frac{r}{\theta}\right)^T \longrightarrow 0$$

uniformly on K, while the Cauchy–Schwarz inequality gives

$$|\mathbb{E}_\mu[u_T(\beta,Z)\mathbf{1}_{\mathcal{A}^c}]| \le \sqrt{\mu(\mathcal{A}^c)}\,\sqrt{Q_T(\beta)} = \sqrt{1-p}\,\sqrt{Q_T(\beta)}.$$

Hence $\limsup_{T\to\infty}\sup_{\beta\in K}|m_T|/\sqrt{Q_T} \le \sqrt{1-p} < 1$. Fix $\gamma \in (\sqrt{1-p}, 1)$, and then there is N_K such that $m_T^2 \le \gamma^2 Q_T$ on K for all $T \ge N_K$, so $V_T = Q_T - m_T^2 \ge (1-\gamma^2)Q_T$, and part (3) holds with $\kappa_K := 1-\gamma^2$. Since $D_T \ge (1-\rho_K^T)^2 > 0$ by (G.5), it follows that $F_T = \mathbb{E}[W]V_T/D_T > 0$ on K for $T \ge N_K$.

The reduced loss. By (G.5), $|u_T| \le \rho_K^T$ on $K \times [a,b]$, so $(1-\rho_K^T)^2 \le D_T \le (1+\rho_K^T)^2$, and these bounds hold for both parities of T. For $T \ge N_K$, combining them with $\kappa_K Q_T \le V_T \le Q_T$ and with (G.12)–(G.14) gives

$$\left(\frac{\mathbb{E}[W]\,\kappa_K\,p_\varepsilon}{(1+\rho_K^T)^2}\right)^{1/(2T)}\theta_\varepsilon\,\phi(\beta) \le F_T(\beta)^{1/(2T)} \le \left(\frac{\mathbb{E}[W]}{(1-\rho_K^T)^2}\right)^{1/(2T)}\phi(\beta).$$

Both prefactors tend to one independently of β, so $\limsup_{T\to\infty}\sup_{\beta\in K}|F_T(\beta)^{1/(2T)} - \phi(\beta)| \le (1-\theta_\varepsilon)\rho_K$. Since $\theta_\varepsilon \ge 1 - s_+\varepsilon/\phi_-$, letting $\varepsilon \downarrow 0$ proves part (4). □

The above result will be used to prove the following lemma establishing the uniform convergence of the derivatives of G_T over a compact interval omitting the point $-\alpha^*$. This will then yield a contradiction as the limiting derivative is strictly nonzero away from $-\alpha^*$. Recall from Appendix F, we defined

$$\alpha^* = \frac{2}{\underline{Z} + \overline{Z}}.$$

Lemma G.9 (Asymptotics of the derivative)**.** *Suppose Assumption 6.1 holds. Fix a compact interval $K \subset I \setminus \{-\alpha^*\}$, where $I := (-2/\overline{Z}, 0)$, and let N_K be as in Lemma G.8, so that G_T is well-defined on K for $T \ge N_K$. Then:*

1. *The derivatives $G_T'(\beta)$ converge uniformly on K to the derivative of $\log\phi(\beta)$:*

$$G_T'(\beta) \longrightarrow \frac{d}{d\beta}\log\phi(\beta)$$

   uniformly for $\beta \in K$ as $T \to \infty$.

2. *In particular, since $-\alpha^* \notin K$, there exist constants $T_K < \infty$ and $c_K > 0$ such that*

$$|G_T'(\beta)| \ge c_K > 0, \qquad \forall\,\beta \in K,\ T \ge T_K.$$

Proof. Use the notation and the uniform estimates from Lemma G.8 and its proof, and let $T \geq N_K$, so that $V_T \geq \kappa_K Q_T > 0$ and $D_T \geq (1-\rho_K^T)^2 > 0$ on K. For fixed T, m_T and Q_T are polynomials in β with finite coefficients, since Z is bounded. Differentiating these polynomials gives

$$m_T' = T \int z q_\beta(z)^{T-1} \, \mathrm{d}\mu(z), \qquad Q_T' = 2T \int z q_\beta(z)^{2T-1} \, \mathrm{d}\mu(z).$$

Set

$$\lambda(\beta) := \frac{z^*}{q_\beta(z^*)}.$$

Only the endpoint value appears in this denominator, and $|q_\beta(z^*)| = \phi(\beta) \geq \phi_- > 0$, so λ is continuous and bounded on K. The identity

$$z - \lambda(\beta)\, q_\beta(z) = \frac{z - z^*}{q_\beta(z^*)}$$

therefore gives the exact representations

$$\begin{aligned} Q_T' - 2T\lambda Q_T &= \frac{2T}{q_\beta(z^*)} \int (z - z^*) q_\beta(z)^{2T-1} \, \mathrm{d}\mu(z), \\ m_T' - T\lambda m_T &= \frac{T}{q_\beta(z^*)} \int (z - z^*) q_\beta(z)^{T-1} \, \mathrm{d}\mu(z). \end{aligned}$$

By the Cauchy–Schwarz inequality and the definition of $\mathcal{L}_T$,

$$\begin{aligned} \frac{|Q_T' - 2T\lambda Q_T|}{2TQ_T} &\leq \frac{\mathbb{E}_\mu\big[|Z - z^*|\,|q_\beta(Z)|^{T-1}\,|q_\beta(Z)|^T\big]}{\phi_- \, Q_T} \leq \frac{\sqrt{\mathcal{L}_T(\beta)}}{\phi_-}, \\ \frac{|m_T' - T\lambda m_T|}{T\sqrt{Q_T}} &\leq \frac{\mathbb{E}_\mu\big[|Z - z^*|\,|q_\beta(Z)|^{T-1}\big]}{\phi_- \sqrt{Q_T}} \leq \frac{\sqrt{\mathcal{L}_T(\beta)}}{\phi_-}, \end{aligned} \tag{G.16}$$

and both right-hand sides tend to zero uniformly on K by Lemma G.8(2). Neither bound requires q_β to be nonvanishing on $[a, b]$.

Since $V_T = Q_T - m_T^2$ and $V_T' = Q_T' - 2m_T m_T'$, we have the exact decomposition

$$V_T' - 2T\lambda V_T = \big(Q_T' - 2T\lambda Q_T\big) - 2m_T\big(m_T' - T\lambda m_T\big),$$

and $|m_T| \leq \sqrt{Q_T}$. Dividing by $2TV_T \geq 2T\kappa_K Q_T$ and using (G.16),

$$\sup_{\beta \in K} \left| \frac{V_T'}{2TV_T} - \lambda(\beta) \right| \leq \frac{2}{\kappa_K \phi_-} \sup_{\beta \in K} \sqrt{\mathcal{L}_T(\beta)} \longrightarrow 0. \tag{G.17}$$

It remains to check that the denominator contributes no logarithmic derivative. Since $|q_\beta(z)| \leq \rho_K$ on $K \times [a, b]$, the derivative formulas above give the crude bounds $|m_T'| \leq Tb\rho_K^{T-1}$ and $|Q_T'| \leq 2Tb\rho_K^{2T-1}$. Because $D_T' = -2m_T' + Q_T'$ and $D_T \geq (1-\rho_K^T)^2$,

$$\sup_{\beta \in K} \left| \frac{D_T'}{2TD_T} \right| \leq \frac{b\,(\rho_K^{T-1} + \rho_K^{2T-1})}{(1-\rho_K^T)^2} \longrightarrow 0. \tag{G.18}$$

By (G.11) and (G.3), $G_T = \frac{1}{2T}\big(\log V_T - \log D_T\big)$, so

$$G_T' = \frac{1}{2T}\frac{V_T'}{V_T} - \frac{1}{2T}\frac{D_T'}{D_T},$$

and (G.17)–(G.18) give $G_T' \to \lambda$ uniformly on K. Finally, z^* is constant on K and $\phi(\beta) = |1 + \beta z^*|$ there, so

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \phi(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log |1 + \beta z^*| = \lambda(\beta),$$

which proves part (1). The function λ is continuous and strictly positive when $z^* = a$, and strictly negative when $z^* = b$. Hence $c_K := \frac{1}{2}\min_{\beta \in K} |\lambda(\beta)| > 0$. Uniform convergence supplies $T_K \geq N_K$ such that $|G_T'(\beta) - \lambda(\beta)| \leq c_K$ for all $\beta \in K$ and $T \geq T_K$, proving part (2). □

The large-depth conclusion of Theorem 6.3 will now follow by a proof by contradiction, since if the trained parameters β_T^* did not converge to $-\alpha^*$, a subsequence would accumulate at a point of the depth-independent compact interval K_0 of Theorem G.7 different from $-\alpha^*$, where Lemma G.9 forbids stationary points at large depth.

Theorem G.10. *Suppose the assumptions of Theorem G.7 hold and, for $T \geq T'(\beta_0)$, let (α_T^*, β_T^*) denote the limit of gradient flow on the two-parameter, T-layer, CA model whose existence is guaranteed by Theorem G.7. Then,*

$$\lim_{T\to\infty} \alpha_T^* = \alpha^* = \frac{2}{\underline{Z} + \overline{Z}}$$

and

$$\lim_{T\to\infty} \beta_T^* = -\alpha^*.$$

The resulting predictor is Bayes optimal in the context-then-depth limit.

Proof. By Theorem G.7, $\beta_T^* \in K_0$ for every $T \geq T'$, where $K_0 = K_0(\beta_0) \subset (-2/\overline{Z}, 0)$ is compact and does not depend on T, and $F_T'(\beta_T^*) = 0$ since (α_T^*, β_T^*) is a stationary point of $\ell^{(T)}$ with $\alpha_T^* = \alpha^*(\beta_T^*)$. We assume for contradiction that $\beta_T^* \not\to -\alpha^*$. Then there exist $\varepsilon > 0$ and depths $T_k \to \infty$ with $|\beta_{T_k}^* + \alpha^*| \geq \varepsilon$ for all k. By compactness of K_0, passing to a further subsequence we may assume that

$$\bar{\beta} := \lim_{k\to\infty} \beta_{T_k}^* \in K_0 \subset (-2/\overline{Z}, 0), \qquad |\bar{\beta} + \alpha^*| \geq \varepsilon.$$

Choose a compact interval

$$K \subset (-2/\overline{Z}, 0) \setminus \{-\alpha^*\}$$

such that $\bar{\beta} \in \mathrm{int}(K)$. Then $\beta_{T_k}^* \in K$ for all sufficiently large k. By Lemma G.8(3), $F_{T_k} > 0$ on K once $T_k \geq N_K$, so G_{T_k} is well-defined at $\beta_{T_k}^*$ and stationarity gives

$$G_{T_k}'(\beta_{T_k}^*) = \frac{F_{T_k}'(\beta_{T_k}^*)}{2T_k\, F_{T_k}(\beta_{T_k}^*)} = 0.$$

By Lemma G.9, however, $|G_{T_k}'(\beta_{T_k}^*)| \geq c_K > 0$ for all sufficiently large k—a contradiction. Therefore $\beta_T^* \to -\alpha^*$:

$$\lim_{T\to\infty} \beta_T^* = -\alpha^*. \tag{G.19}$$

The existence of the limit (G.19) is thus a consequence of the depth-uniform confinement $\beta_T^* \in K_0$ established in Theorem G.7, rather than a hypothesis. To account for the changing depth, we restore the depth in the notation of (G.1), writing $S_T(Z, \beta)$, $A_T(\beta) := \mathbb{E}[W S_T(Z, \beta)^2]$ and $B_T(\beta) := \mathbb{E}[W S_T(Z, \beta)]$ for S, A and B at depth T, so that the profiled minimizer is $\alpha_T^*(\beta) = B_T(\beta)/A_T(\beta)$. Since $S_T(Z, \beta) = -(1 - u_T(\beta, Z))/\beta$, reweighting by $\mathrm{d}\mu = W\ \mathrm{d}\mathbb{P}/\mathbb{E}[W]$ gives, for $\beta \neq 0$,

$$\alpha_T^*(\beta) = -\beta\, \frac{\mathbb{E}_\mu[1 - u_T(\beta, Z)]}{\mathbb{E}_\mu[(1 - u_T(\beta, Z))^2]}.$$

Recall from Theorem G.7 that $\beta_T^* \in K_0$ for all $T \geq T'$, where $K_0 \subset (-2/\overline{Z}, 0)$ is compact and independent of the depth. Since $\underline{Z} \geq 1$, compactness gives

$$\widetilde{\rho} := \max_{\beta\in K_0,\, \sigma\in[1,\overline{Z}]} |1 + \beta\sigma| < 1, \qquad \sup_{\beta\in K_0,\, z\in[\underline{Z},\overline{Z}]} |u_T(\beta, z)| \leq \widetilde{\rho}^{\,T}.$$

Writing $N_T(\beta) = \mathbb{E}_\mu[1 - u_T(\beta, Z)]$ and $D_T(\beta) = \mathbb{E}_\mu[(1 - u_T(\beta, Z))^2]$, uniformly on K_0 we have

$$|N_T - 1| \leq \widetilde{\rho}^{\,T}, \qquad |D_T - 1| \leq 2\widetilde{\rho}^{\,T} + \widetilde{\rho}^{\,2T}, \qquad D_T \geq (1 - \widetilde{\rho}^{\,T})^2.$$

Thus, using $|D_T - N_T| \leq \widetilde{\rho}^{\,T} + \widetilde{\rho}^{\,2T}$,

$$\sup_{\beta\in K_0} |\alpha_T^*(\beta) + \beta| \leq \max_{\beta\in K_0} |\beta|\, \frac{\widetilde{\rho}^{\,T} + \widetilde{\rho}^{\,2T}}{(1 - \widetilde{\rho}^{\,T})^2} \longrightarrow 0.$$

Evaluating at β_T^* yields $\alpha_T^* = \alpha_T^*(\beta_T^*) = -\beta_T^* + o(1) \to \alpha^*$, and therefore $\alpha_T^*/\beta_T^* \to -1$.

It remains to verify prediction convergence for this changing pair of parameters. At each fixed $T \geq T'$, Lemma E.2 and the covariance law of large numbers give

$$H_T := \lim_{L\to\infty} \frac{1}{L} \boldsymbol{X}\boldsymbol{F}_T^\top = \frac{\alpha_T^*}{\beta_T^*}\left[(\boldsymbol{I} + \beta_T^* \boldsymbol{\Sigma}_{\boldsymbol{x}})^T - \boldsymbol{I}\right]$$

almost surely. Every eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top$ lies in $[1, \overline{Z}]$, including the eigenvalue 1 on the orthogonal complement of $\boldsymbol{m}$. The same $\widetilde{\rho}$ therefore gives $\|(\boldsymbol{I} + \beta_T^* \boldsymbol{\Sigma}_{\boldsymbol{x}})^T\| \leq \widetilde{\rho}^T$ for all $T \geq T'$, so

$$\|H_T - \boldsymbol{I}\| \leq \left|1 + \frac{\alpha_T^*}{\beta_T^*}\right| + \left|\frac{\alpha_T^*}{\beta_T^*}\right| \widetilde{\rho}^T \longrightarrow 0.$$

Lemma E.1 gives $\lim_{L\to\infty} \hat{y}_q = \boldsymbol{w}^\top H_T \boldsymbol{x}_q$ on a common probability-one event for the countably many depths. Its difference from $\boldsymbol{w}^\top \boldsymbol{x}_q$ is bounded by $\|\boldsymbol{w}\|\|\boldsymbol{x}_q\|\|H_T - \boldsymbol{I}\|$ and hence tends to zero. Applied to the test prompt, this proves the claimed context-then-depth Bayes-prediction limit. □

Theorems G.7 and G.10 together prove Theorem 6.3, with $T'(\beta_0)$ the depth threshold of Lemma G.6. Of Assumption 6.1, the arguments of this appendix use only that $Z > 1$ almost surely, that $1 \leq \underline{Z} < \overline{Z} < \infty$, and that every neighborhood of each essential endpoint $\underline{Z}$, $\overline{Z}$ carries positive mass, so in particular no density of Z is required.

H Non-asymptotic guarantees

Theorems 6.2 and 6.3 send the test-context length to infinity at each fixed depth. This appendix records what survives at finite context length L and finite depth T for the one-parameter model of Section 5, at a parameter α fixed independently of the test prompt, where the excess risk splits exactly into a bias that contracts geometrically in the depth and a noise term of order d/L from the finite context. These are prediction guarantees for an already-chosen parameter, not statements about training at finite context or from finitely many prompts.

We use the notation of the proof of Lemma F.1, namely $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} = \boldsymbol{X}\boldsymbol{X}^\top/L$, $\boldsymbol{M} = \boldsymbol{I} - \alpha\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$ and $\boldsymbol{Q}_T = \sum_{k=0}^{T-1} \boldsymbol{M}^k$, so that $\boldsymbol{F}_T = \alpha \boldsymbol{Q}_T \boldsymbol{X}$ and $\alpha\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}\boldsymbol{Q}_T = \boldsymbol{I} - \boldsymbol{M}^T$. The error is measured against the task-conditional oracle $\langle \boldsymbol{w}, \boldsymbol{x}_q \rangle$ of (3.3) rather than against y_q, so that the irreducible noise $\sigma_\tau^2 = \zeta^2/Z$ does not appear

$$\mathcal{E}_{L,T}(\alpha; \tau, \boldsymbol{X}) := \mathbb{E}\left[(\hat{y}_q - \langle \boldsymbol{w}, \boldsymbol{x}_q \rangle)^2 \mid \tau, \boldsymbol{X}\right],$$

where the expectation averages over the context residuals and the query with the task $\tau = (\boldsymbol{m}, \zeta)$ and the context covariates $\boldsymbol{X}$ held fixed. Conditionally on the task, write $\boldsymbol{X} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2} \boldsymbol{G}$, where $\boldsymbol{G} \in \mathbb{R}^{d\times L}$ has independent standard Gaussian entries and is independent of the task, and for $\delta \in (0, 1)$ set

$$\mathsf{S}_\delta := \left\{ (1-\delta)^2 \boldsymbol{I} \preceq \frac{\boldsymbol{G}\boldsymbol{G}^\top}{L} \preceq (1+\delta)^2 \boldsymbol{I} \right\}. \tag{H.1}$$

Theorem H.1 (Finite-context excess risk)**.** *Suppose Assumption 6.1 holds.*

1. *For every real α and every integer $T \geq 1$,*

$$\mathcal{E}_{L,T}(\alpha; \tau, \boldsymbol{X}) = \boldsymbol{w}^\top \boldsymbol{M}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^T \boldsymbol{w} + \frac{\sigma_\tau^2 \alpha^2}{L} \operatorname{tr}\Big(\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{Q}_T \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \boldsymbol{Q}_T\Big). \tag{H.2}$$

2. *Let $L > d$, let $\delta \in (\sqrt{d/L}, 1)$, and let*

$$0 < \alpha < \frac{2}{(1+\delta)^2 \overline{Z}}, \qquad \rho_\delta(\alpha) := \max\left\{\left|1 - \alpha(1-\delta)^2\right|, \left|1 - \alpha(1+\delta)^2 \overline{Z}\right|\right\} < 1. \tag{H.3}$$

Then $\mathbb{P}(\mathsf{S}_\delta^{\mathsf{c}}) \leq 2\exp[-(\delta\sqrt{L} - \sqrt{d})^2/2]$, and on S_δ

$$\mathcal{E}_{L,T}(\alpha; \tau, \boldsymbol{X}) \leq \zeta^2 \left[W \rho_\delta(\alpha)^{2T} + \frac{(d + Z - 1)\left(1 + \rho_\delta(\alpha)^T\right)^2}{ZL(1-\delta)^2} \right] \tag{H.4}$$

simultaneously for every integer $T \geq 1$ and every α satisfying (H.3).

Proof. (1) This is the computation in the proof of Lemma F.1, which gives

$$\langle \boldsymbol{w}, \boldsymbol{x}_q \rangle - \hat{y}_q = \boldsymbol{w}^\top \boldsymbol{M}^T \boldsymbol{x}_q - \frac{\alpha}{L} \boldsymbol{\varepsilon}^\top \boldsymbol{X}^\top \boldsymbol{Q}_T \boldsymbol{x}_q.$$

Conditionally on $(\tau, \boldsymbol{X})$ the query is centered with covariance $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, the residual vector $\boldsymbol{\varepsilon}$ is centered with covariance $\sigma_\tau^2 \boldsymbol{I}_L$, and the two are independent, so the cross term has zero conditional expectation, while the two squares contribute the two terms of (H.2). Compared with the display in that proof, only the term σ_τ^2 is absent, because the error is measured against $\langle \boldsymbol{w}, \boldsymbol{x}_q \rangle$ rather than y_q. No inverse is used, so the identity holds when $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$ is singular and when $\alpha = 0$.

(2) The Gaussian singular-value bound (Vershynin, 2012, Corollary 5.35), applied to $\boldsymbol{G}^\top \in \mathbb{R}^{L \times d}$ with $t = \delta\sqrt{L} - \sqrt{d} > 0$, gives $(1-\delta)\sqrt{L} \le s_{\min}(\boldsymbol{G}^\top) \le s_{\max}(\boldsymbol{G}^\top) \le (1+\delta)\sqrt{L}$ with probability at least $1 - 2e^{-t^2/2}$, which is the stated bound on $\mathbb{P}(\mathsf{S}_\delta^{\mathsf{c}})$. Now work on S_δ. Conjugating (H.1) by $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}$ gives $(1-\delta)^2 \boldsymbol{\Sigma}_{\boldsymbol{x}} \preceq \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \preceq (1+\delta)^2 \boldsymbol{\Sigma}_{\boldsymbol{x}}$, and $\boldsymbol{I} \preceq \boldsymbol{\Sigma}_{\boldsymbol{x}} \preceq \overline{Z}\boldsymbol{I}$ places every eigenvalue of $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$ in $[(1-\delta)^2, (1+\delta)^2 \overline{Z}]$. Since $\lambda \mapsto |1 - \alpha\lambda|$ is convex, its maximum over that interval is attained at an endpoint, so $\|\boldsymbol{M}\| \le \rho_\delta(\alpha) < 1$. With $\|\boldsymbol{\Sigma}_{\boldsymbol{x}}\| = Z$ and $\|\boldsymbol{w}\|^2 = \zeta^2(Z-1)/Z^2$, the bias term of (H.2) is at most $Z\rho_\delta^{2T}\zeta^2(Z-1)/Z^2 = \zeta^2 W \rho_\delta^{2T}$. For the noise term, $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$ is positive definite on S_δ, and since all the matrices involved are polynomials in $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$, the identity $\alpha \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \boldsymbol{Q}_T = \boldsymbol{I} - \boldsymbol{M}^T$ gives

$$\boldsymbol{R}_T := \alpha^2 \boldsymbol{Q}_T \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \boldsymbol{Q}_T = (\boldsymbol{I} - \boldsymbol{M}^T)^2 \widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{-1} \succeq 0, \qquad \|\boldsymbol{R}_T\| \le \frac{(1+\rho_\delta^T)^2}{(1-\delta)^2},$$

the norm bound using $\|\boldsymbol{M}\| \le \rho_\delta$ and the lower bound $(1-\delta)^2$ on the spectrum of $\widehat{\boldsymbol{\Sigma}_{\boldsymbol{x}}}$. As $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{m}^\top$, we have $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{R}_T) = \mathrm{tr}(\boldsymbol{R}_T) + \boldsymbol{m}^\top \boldsymbol{R}_T \boldsymbol{m} \le (d + Z - 1)\|\boldsymbol{R}_T\|$. Substituting this and $\sigma_\tau^2 = \zeta^2/Z$ into (H.2) proves (H.4). The event S_δ is a property of $\boldsymbol{G}$ alone and involves neither T nor α, so the same realization serves every depth and every admissible parameter at once. □

Interpretation of the bound. The first term of (H.4) is the finite-depth bias and is geometric in T, and the second is the finite-context noise term and is of order d/L once ρ_δ^T is small. The contraction factor ρ_δ controls the sample spectrum and tends to $\max\{|1-\alpha|, |1-\alpha\overline{Z}|\}$ as $\delta \downarrow 0$, but it need not match the faster rate along the task direction that governs the infinite-context population loss of Appendix F. The population-trained parameter is admissible, since by Theorem F.3, $\alpha_T^* \to \alpha^* = 2/(\underline{Z} + \overline{Z})$, and choosing $\delta > 0$ small enough that $(1+\delta)^2\overline{Z} < \underline{Z} + \overline{Z}$ places α^* strictly inside the window (H.3), so the trained parameters are admissible for all sufficiently large depths.

I FIGURES 2 AND 3 EXPERIMENT DETAILS

For Figure 2, models are trained using standard gradient descent (in place of gradient flow) on the non-asymptotic objective $\ell_{N,L_{\mathrm{tr}}}$ with $N = 2000$ and $L_{\mathrm{tr}} = 100$, and the protocol for Figure 3 is stated separately below. Data is generated according to Section 3 with $\zeta \sim \mathcal{N}(0,1)$, $d_1 = d_2 = 16$, and $\boldsymbol{m}$ drawn from a spherical Gaussian with norm mapped uniformly on $[0, 5]$.

Further details for less-restricted LCA parametrizations (Figure 3). We use the synthetic data-generating model discussed above. Throughout these experiments, we keep $\boldsymbol{W}_t^K = \boldsymbol{W}_t^Q = \boldsymbol{I}_d$ fixed and train only the layerwise injection and value matrices $(\boldsymbol{W}_t^S, \boldsymbol{W}_t^V)_{t=0}^{T-1}$. We consider three increasingly flexible parametrizations where in all cases the weights are untied across layers unlike in the one- and two-parameter simplifications on which we base our theoretical results in Section 6. In the untied scalar model,

$$\boldsymbol{W}_t^S = \alpha_t \boldsymbol{I}_d, \qquad \boldsymbol{W}_t^V = \beta_t \boldsymbol{I}_d,$$

where each layer has its own learnable pair (α_t, β_t). In the diagonal model,

$$\boldsymbol{W}_t^S = \mathrm{diag}(\boldsymbol{a}_t), \qquad \boldsymbol{W}_t^V = \mathrm{diag}(\boldsymbol{b}_t),$$

where $\boldsymbol{a}_t, \boldsymbol{b}_t \in \mathbb{R}^d$. In the full-matrix model, $\boldsymbol{W}_t^S, \boldsymbol{W}_t^V \in \mathbb{R}^{d \times d}$ are unconstrained trainable matrices.

All three models are initialized near the theory-predicted limiting parameters $(\alpha^*, -\alpha^*)$ from Theorems 6.2–6.3. Specifically, with initial noise variance $\sigma_{\text{init}} = 0.1$, for the untied scalar model we initialize

$$\alpha_t = \alpha^* + \sigma_{\text{init}} g_t^S, \qquad \beta_t = -\alpha^* + \sigma_{\text{init}} g_t^V,$$

with independent standard Gaussian variables g_t^S, g_t^V. For the diagonal model, we initialize

$$\boldsymbol{a}_t = \alpha^* \mathbf{1}_d + \sigma_{\text{init}} \mathbf{g}_t^S, \qquad \boldsymbol{b}_t = -\alpha^* \mathbf{1}_d + \sigma_{\text{init}} \mathbf{g}_t^V,$$

where $\mathbf{g}_t^S$ and $\mathbf{g}_t^V$ have i.i.d. standard Gaussian entries and $\mathbf{1}_d = [1 \cdots 1]^\top \in \mathbb{R}^d$. For the full-matrix model, we initialize

$$\boldsymbol{W}_t^S = \alpha^* \boldsymbol{I}_d + \sigma_{\text{init}} \mathbf{G}_t^S, \qquad \boldsymbol{W}_t^V = -\alpha^* \boldsymbol{I}_d + \sigma_{\text{init}} \mathbf{G}_t^V,$$

where the entries of $\mathbf{G}_t^S, \mathbf{G}_t^V$ are i.i.d. standard Gaussian. Thus, each model has initialization biased near the theoretically optimal tied-scalar solution obtained in the one- and two-parameter models. However, the noise level σ_{init} is large with respect to the magnitude $|\alpha^*|$ as the distribution on $\boldsymbol{m}$ yields $\alpha^* = 2/27 \approx 0.074$.

We train each model by gradient descent, but—unlike the protocol for Figure 2—on freshly sampled batches of 512 prompts for 500 steps rather than on a fixed pool of N prompts, at training context length $L_{\text{tr}} = 500$, with 2000 held-out prompts used for validation and 2000 for the diagnostics reported below. After training, in Figure 3, we display the scalar projections

$$\widehat{\alpha}_t = d^{-1} \operatorname{tr}(\boldsymbol{W}_t^S), \qquad \widehat{\beta}_t = d^{-1} \operatorname{tr}(\boldsymbol{W}_t^V),$$

as well as the whitening diagnostic $\|\boldsymbol{X}\boldsymbol{F}_t^\top/L - \boldsymbol{I}\|_F/\sqrt{d}$ following the heuristic laid out at the end of Section 6. The points/curves in Figure 3 are averaged over ten independent runs.

J TABLE 1 EXPERIMENT DETAILS

This appendix gives the implementation details for the corrupted avMNIST experiments in Table 1. The goal is to test whether the two architectural mechanisms isolated by the theory—cross-attention from an evolving state back to a fixed covariate sequence, and repeated re-injection of that fixed sequence—remain useful after replacing the linearized model (5.4) by conventional deep-learning components: convolutional token encoders, softmax multi-head attention, pre-layer normalization, MLP blocks, dropout, and a classification head.

Throughout this appendix, we use the same notation as Section 5 whenever possible. In particular, $\boldsymbol{X}$ denotes the fixed encoded multimodal token sequence for one example, $\boldsymbol{F}_t$ denotes the evolving state after t layers, $\boldsymbol{A}_t$ denotes the attention update, and $\boldsymbol{S}_t$ denotes the re-injection term. The only notational difference is that tokens are written as rows. Thus, in this appendix $\boldsymbol{X}, \boldsymbol{F}_t \in \mathbb{R}^{L \times d}$, whereas the theoretical sections write tokens as columns.

J.1 DATASET, NORMALIZATION, AND TASK

We use a paired audio-visual MNIST dataset (Vielzeuf et al., 2019). Each example is a triple

$$(I_i, A_i, c_i), \qquad I_i \in \mathbb{R}^{1\times 28\times 28}, \quad A_i \in \mathbb{R}^{1\times 28\times 28}, \quad c_i \in \{0, 1, \ldots, 9\},$$

where I_i is an image, A_i is a paired audio spectrogram, and c_i is the digit class. The experiments use the standard split stored in the repository: 60,000 training examples and 10,000 test examples. The image and audio arrays have spatial resolution 28×28; a singleton channel dimension is added before passing them to the convolutional encoders. The data (I_i, A_i) is corrupted entry-wise by Gaussian noise to enforce difficulty in the learning task. For each modality separately, we add i.i.d. $\mathcal{N}(0, \sigma^2)$ noise with $\sigma = 5\hat{s}$, where $\hat{s}$ is the empirical standard deviation of that modality over the dataset, and clip the result back to the original value range. This noise is large relative to the signal it corrupts, which is what prevents the task from saturating and accounts for the accuracies reported in Table 1. Thereafter, the data is normalized to $[0, 1]$.

Each model outputs logits $z_i \in \mathbb{R}^{10}$ and is trained with the cross-entropy loss

$$\ell_{\text{CE}}(z_i, c_i) = -\log\left(\frac{\exp(z_{i,c_i})}{\sum_{k=0}^{9} \exp(z_{i,k})}\right).$$

J.2 Common convolutional tokenization

For token dimension d, each modality encoder maps a 28×28 input to 16 tokens in $\mathbb{R}^d$:

$$\Phi_I : \mathbb{R}^{1\times 28\times 28} \to \mathbb{R}^{16\times d}, \qquad \Phi_A : \mathbb{R}^{1\times 28\times 28} \to \mathbb{R}^{16\times d}.$$

The two encoders have the same architecture but do not share weights. Each encoder consists of two convolution–ReLU layers, a 2×2 max-pooling layer, two additional convolution–ReLU layers, a second 2×2 max-pooling layer, adaptive average pooling to a 4×4 grid, flattening into 16 tokens, and a final layer normalization. The hidden convolutional width is taken to be 16.

We add learned positional embeddings and learned modality embeddings:

$$\boldsymbol{X}_I^{(0)} = \Phi_I(I) + \boldsymbol{P}_I + \boldsymbol{M}_I, \qquad \boldsymbol{X}_A^{(0)} = \Phi_A(A) + \boldsymbol{P}_A + \boldsymbol{M}_A,$$

where $\boldsymbol{P}_I, \boldsymbol{P}_A \in \mathbb{R}^{16\times d}$ are learned positional embeddings and $\boldsymbol{M}_I, \boldsymbol{M}_A \in \mathbb{R}^{1\times d}$ are learned modality embeddings broadcast over tokens. The fixed multimodal sequence is

$$\boldsymbol{X} = \mathrm{concat}(\boldsymbol{c}, \boldsymbol{X}_I^{(0)}, \boldsymbol{X}_A^{(0)}) \in \mathbb{R}^{L\times d},$$

where $\boldsymbol{c} \in \mathbb{R}^d$ is a learned CLS token.

J.3 Paper-inspired multimodal architecture

The four paper-inspired models are empirical analogues of the CA recurrence (5.2). They keep a fixed encoded sequence $\boldsymbol{X}$ and update an evolving state

$$\boldsymbol{F}_t \in \mathbb{R}^{L\times d}, \qquad t = 0, 1, \ldots, T.$$

The state is initialized from the fixed sequence by a learned affine map

$$\boldsymbol{F}_0 = \boldsymbol{X}\boldsymbol{W}_0^\top + \mathbf{1}b_0^\top.$$

This differs from the simplified theoretical initialization $\boldsymbol{F}_0 = 0$.

Each fusion layer has three possible components: an attention update, an optional re-injection of $\boldsymbol{X}$, and a feed-forward residual block. We use pre-layer normalization $\mathsf{LN}(\cdot)$ throughout. In the full model, the attention update is

$$\boldsymbol{A}_{t-1} = \mathsf{MHA}_t\big(\mathsf{LN}(\boldsymbol{F}_{t-1}), \mathsf{LN}(\boldsymbol{X}), \mathsf{LN}(\boldsymbol{X})\big),$$

Thus, as in (5.3), the evolving state attends back to the fixed encoded covariates.

The empirical re-injection term is a gated version of the theoretical term $\boldsymbol{S}_{t-1} = \boldsymbol{W}_{t-1}^S \boldsymbol{X}$ which is used to provide stability during training. After the attention residual,

$$\widetilde{\boldsymbol{F}}_t = \boldsymbol{F}_{t-1} + \boldsymbol{A}_{t-1},$$

we set

$$\boldsymbol{S}_{t-1} = \boldsymbol{G}_{t-1}(\widetilde{\boldsymbol{F}}_t) \odot \left(\boldsymbol{X}(\boldsymbol{W}_{t-1}^S)^\top + \mathbf{1}(b_{t-1}^S)^\top\right),$$

where $\odot$ denotes entrywise multiplication and

$$\boldsymbol{G}_{t-1}(\widetilde{\boldsymbol{F}}_t) = \sigma\left(\widetilde{\boldsymbol{F}}_t(\boldsymbol{W}_{t-1}^G)^\top + \mathbf{1}(b_{t-1}^G)^\top\right)$$

is a learned sigmoid gate. The full update is therefore

$$\overline{\boldsymbol{F}}_t = \widetilde{\boldsymbol{F}}_t + \boldsymbol{S}_{t-1}, \qquad \boldsymbol{F}_t = \overline{\boldsymbol{F}}_t + \mathsf{FFN}_t\big(\mathsf{LN}(\overline{\boldsymbol{F}}_t)\big),$$

where FFN_t is a two-layer pointwise MLP with GELU activation, dropout, and hidden width d_h.

The classifier uses the final CLS representation:

$$h = \boldsymbol{F}_T[\mathsf{CLS}],$$

and maps h to logits in $\mathbb{R}^{10}$ using layer normalization followed by a two-layer MLP head with GELU activation and dropout.

Ablation axes. The four paper-inspired models differ only in whether the attention update uses $\boldsymbol{F}_t$–$\boldsymbol{X}$ cross-attention and whether the re-injection term $\boldsymbol{S}_t$ is present:

- **Full paper-inspired model (Inj. & cross-attn)**: uses the cross-attention update above and includes $\boldsymbol{S}_t$.
- **No injection**: uses the same $\boldsymbol{F}_t$–$\boldsymbol{X}$ cross-attention update but sets $\boldsymbol{S}_t \equiv 0$.
- **No cross-attn**: keeps the re-injection term $\boldsymbol{S}_t$, but replaces the cross-attention update by self-attention on the evolving state,

$$\boldsymbol{A}_{t-1} = \mathsf{MHA}_t\big(\mathsf{LN}(\boldsymbol{F}_{t-1}), \mathsf{LN}(\boldsymbol{F}_{t-1}), \mathsf{LN}(\boldsymbol{F}_{t-1})\big).$$

- **No inj. & no cross-attn**: uses the self-attention update above and sets $\boldsymbol{S}_t \equiv 0$.

These ablations isolate the two mechanisms appearing in (5.2)–(5.3): access to the fixed sequence $\boldsymbol{X}$ through cross-attention, and layer-wise re-injection of $\boldsymbol{X}$ through $\boldsymbol{S}_t$.

J.4 Bidirectional cross-attention baseline

The standard bidirectional baseline does not maintain a single evolving state $\boldsymbol{F}_t$ and does not re-inject a fixed sequence. Instead, it maintains separate image and audio streams

$$\boldsymbol{U}_t, \boldsymbol{V}_t \in \mathbb{R}^{16\times d}, \qquad \boldsymbol{U}_0 = \Phi_I(I) + \boldsymbol{P}_I, \qquad \boldsymbol{V}_0 = \Phi_A(A) + \boldsymbol{P}_A.$$

At layer t, the image stream attends to the audio stream and the audio stream attends to the image stream:

$$\widetilde{\boldsymbol{U}}_t = \boldsymbol{U}_{t-1} + \mathsf{MHA}_t^I\big(\mathsf{LN}(\boldsymbol{U}_{t-1}), \mathsf{LN}(\boldsymbol{V}_{t-1}), \mathsf{LN}(\boldsymbol{V}_{t-1})\big),$$
$$\widetilde{\boldsymbol{V}}_t = \boldsymbol{V}_{t-1} + \mathsf{MHA}_t^A\big(\mathsf{LN}(\boldsymbol{V}_{t-1}), \mathsf{LN}(\boldsymbol{U}_{t-1}), \mathsf{LN}(\boldsymbol{U}_{t-1})\big).$$

Each stream is then updated by its own feed-forward residual block:

$$\boldsymbol{U}_t = \widetilde{\boldsymbol{U}}_t + \mathsf{FFN}_t^I(\mathsf{LN}(\widetilde{\boldsymbol{U}}_t)), \qquad \boldsymbol{V}_t = \widetilde{\boldsymbol{V}}_t + \mathsf{FFN}_t^A(\mathsf{LN}(\widetilde{\boldsymbol{V}}_t)).$$

The final representation is obtained by mean-pooling each stream and concatenating the two pooled vectors:

$$h = \mathrm{concat}\left(\frac{1}{16}\sum_{j=1}^{16}\boldsymbol{U}_T[j], \frac{1}{16}\sum_{j=1}^{16}\boldsymbol{V}_T[j]\right) \in \mathbb{R}^{2d}.$$

A layer-normalized two-layer MLP head maps h to logits in $\mathbb{R}^{10}$. This baseline lacks the two paper-specific components: a single evolving state repeatedly attending to a fixed $\boldsymbol{X}$, and recurrent re-injection of $\boldsymbol{X}$.

J.5 Training and reporting

All models in Table 1 are trained under the same optimization budget. The hyperparameters are

$$d = 16, \qquad T = 10, \qquad \text{number of heads} = 4, \qquad d_h = 32, \qquad g = 4, \qquad \text{dropout} = 0.1.$$

We train for 100 epochs using AdamW with standard specifications: learning rate 3×10^{-4} and weight decay 10^{-4}. The batch size is 256. Obtaining the results displayed in Table 1 totaled approximately five hours of runtime on a single NVIDIA H200 GPU.