SOFTMAX ATTENTION AS A COVARIANCE READOUT: A UNIFIED VIEW OF IN-CONTEXT LEARNING AND REPETITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) exhibit two seemingly unrelated behaviors: in-context learning (ICL) and repetitive generation. In both cases, the model appears to summarize the context into a population-level statistic to guide subsequent generation while discarding token-level detail. We investigate whether this "summarization and forgetting" can be derived from the attention mechanism itself. For stationary inputs with elliptical marginals, the softmax attention outputs converge almost surely to $V\mu + \gamma V\Sigma K^\top Qx$ in the long-context regime, where μ and Σ denote the input mean and covariance, respectively, and γ is a scalar. Beyond the mean contribution, this limit provides a readout of the input's second-order statistics. Its dependence on the current input token alone leads to two consequences. (i) For in-context linear regression tasks, a residual softmax attention layer can emulate one gradient-descent step. (ii) When propagated across multiple transformer layers, this readout drives the terminal hidden state toward a deterministic function of the current token. Autoregressive generation thus reduces asymptotically to a first-order Markov chain, whose periodic orbits provide a structural explanation for repetitive generation. These two phenomena emerge as facets of a common covariance-readout principle.

1 INTRODUCTION

Transformers have become a central architecture in modern machine learning, achieving strong performance across a wide range of tasks (Vaswani et al., 2017). A particularly striking capability of large Transformer-based language models is in-context learning (ICL): the ability to adapt their predictions using demonstrations supplied in the context, without updating their parameters (Brown et al., 2020). The same pretrained model can therefore perform different tasks by conditioning on different examples, making the context a means of adaptation at inference time.

LLMs can also degenerate into repetitive loops, persistently generating the same sequence of words (Holtzman et al., 2019). The model appears to follow an established pattern within the regime implied by the context while losing sensitivity to its specific content. In a loose analogy with human automatic behavior, the pattern persists while the contextual details that should guide its continuation recede from influence.

These two seemingly unrelated behaviors suggest a common mechanism: *summarize and forget*. The model may extract statistical information from the context while losing sensitivity to the individual tokens that convey it. This perspective motivates our central question: can such a mechanism be derived from softmax self-attention itself?

To address this, we analyze softmax attention in the long-context regime. For stationary inputs with elliptical marginals, we establish that the softmax attention output $f_T(u)$ for a query token u and a context of length T satisfies

$$f_T(u) \xrightarrow[T\to\infty]{\text{a.s.}} V\mu + \gamma V\Sigma K^\top Qu,$$

where K, Q, and V are the key, query, and value matrices, μ and Σ denote the input mean and covariance, respectively, and γ is a scalar. Beyond the mean contribution, this limit provides a

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.

readout of the input's second-order statistics. Its dependence on the current query alone gives a mathematical interpretation of *summarize and forget*: distributional structure remains available for prediction, while explicit dependence on the individual context tokens vanishes.

Recent analyses of softmax attention in the long-context regime establish convergence under i.i.d. sampling and obtain closed-form affine limits for Gaussian inputs (Boursier & Boyer, 2025; Bohbot et al., 2025). We relax these assumptions within a time-series framework, modeling the context as a stationary ergodic process and extending the Gaussian characterization to covariance-directed alignment under elliptical marginals. This framework accommodates temporal dependence in the convergence analysis.

We summarize our contributions as follows:

- **Broader analysis of softmax attention.** We establish the long-context limit for stationary inputs and characterize covariance-directed alignment under elliptical marginals, relaxing independence and Gaussianity.
- **Architectural account of ICL.** We construct an explicit, task-independent parameterization of a single residual softmax attention head that asymptotically emulates one gradient-descent step for linear regression tasks, and empirically verify this correspondence.
- **Structural explanation for repetition.** We extend the result to multilayer Transformers with autoregressive decoding, deriving an asymptotically first-order Markov description of generation and using it to explain repetition.
- **Unified view of ICL and repetition.** We identify a common mechanism underlying these two seemingly unrelated phenomena.

2 Related Work

Softmax attention in the long-context regime. Most closely related, Boursier & Boyer (2025) analyze large-prompt softmax attention through the same three steps, but under different probabilistic frameworks. First, they represent an i.i.d. prompt by its empirical measure and use the strong law of large numbers to identify the infinite-prompt limit, whereas we model the ordered prompt as a stationary process allowing for temporal dependence. Second, for Gaussian inputs, they derive a closed-form affine map from the current input to the attention output. We extend the analysis to elliptical inputs, which include Gaussians as a special case. Third, they establish finite-prompt concentration bounds for i.i.d. sub-Gaussian tokens. This concentration perspective is also developed by Bohbot et al. (2025), who study the token sample complexity of attention and obtain bounds uniform over bounded query sets under the same sampling assumption. Our finite-prompt analysis instead accommodates sub-Gaussian token sequences satisfying a strong-mixing condition. Thus, we retain the sub-Gaussian tail assumption while extending the analysis to temporally dependent contexts within a time-series framework.

Algorithmic accounts of ICL. Prior work can be organized around three complementary questions: whether Transformers can represent an in-context learning algorithm, whether the pretraining objective theoretically favors such solutions, and whether these algorithms emerge empirically through pretraining. Under different architectural assumptions, existing studies address one or more of these questions.

Under *linear self-attention*, Von Oswald et al. (2023) construct one head that performs a gradient-descent step in linear regression tasks and observe corresponding behavior in pretrained models. Ahn et al. (2023) and Mahankali et al. (2024) analyze pretraining losses for Gaussian linear-regression tasks, showing that global optima of one-layer linear-attention models implement one step of GD; under structured parameterizations, Ahn et al. (2023) further identify multilayer critical points realizing multiple preconditioned steps. Under *normalized ReLU attention*, Bai et al. (2023) provide constructions and ERM-based pretraining guarantees for a family of statistical procedures, including in-context algorithm selection, and validate them experimentally. Under *standard softmax Transformers*, Garg et al. (2022) provide purely empirical evidence that Transformers can be pretrained to in-context learn dense and sparse linear functions, two-layer neural networks, and decision trees, with performance matching or exceeding task-specific learners. Akyürek et al. (2022) construct copying (`mov`), matrix multiplication (`mul`), division (`div`), and affine-transformation

(aff) primitives within Transformer decoder layers, compose them to simulate single-example gradient-descent and Sherman–Morrison ridge updates, and empirically relate pretrained models to GD, ridge regression, and least squares. In contrast to operator-level construction spanning multiple Transformer layers, we provide a direct closed-form parameterization of a single softmax attention head that emulates one gradient-descent step for linear regression tasks in the long-context limit, and show empirically that this structure emerges during pretraining.

Repetition in language models. Mahaut & Franzon (2025) distinguish natural repetition, which emerges spontaneously during text generation, from ICL-induced repetition, which exploits repetitive patterns in context. The Markov analysis of Fu et al. (2021) offers a theoretical perspective on natural repetition through the high-inflow problem: many tokens favor the same successor, facilitating repeated returns. Consistent with the copying mechanism underlying ICL-induced repetition, Hiraoka & Inui (2025) identify repetition neurons and interpret their activity as supporting repeated copying of previous context. Complementary studies examine repetition from a data perspective: Li et al. (2023) link degeneration to repetitive patterns in training data, while Xu et al. (2022) use synthetic repetitive examples to penalize and disrupt self-reinforcing sentence repetition. Our work examines the Markov assumption underlying the analysis of Fu et al. (2021), deriving conditions under which autoregressive Transformer generation becomes asymptotically first-order Markovian.

3 Softmax Attention as a Covariance Readout

In this section, we characterize the long-context alignment of softmax attention and quantify its finite-context rate. Section 3.1 introduces the softmax-weighted mean representation of attention and establishes covariance-directed alignment under elliptical symmetry. Section 3.2 derives mean-square error bounds, showing how context length and dependence among tokens affect the alignment rate. Section 3.3 examines the covariance-readout empirically on natural-language embeddings.

Notation. We write $\langle u, v\rangle = u^\top v$ for the Euclidean inner product, $\|u\|$ for the Euclidean norm, and $\|A\|_{\mathrm{op}}$ for the induced operator norm. $A \succeq 0$ means that A is symmetric positive semidefinite. The notation $\overset{d}{=}$ denotes equality in distribution, and $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence.

3.1 Softmax Attention Alignment in the Long-Context Regime

Softmax-weighted mean representation of attention. Let $K, Q \in \mathbb{R}^{d_k \times d}$ and $V \in \mathbb{R}^{d_v \times d}$ denote the key, query, and value matrices. For a context $(x_1, \dots, x_T) \in \mathbb{R}^{d \times T}$ consisting of T tokens, the single-head self-attention function $f_T : \mathbb{R}^d \to \mathbb{R}^{d_v}$ maps each token x_i to

$$f_T(x_i) := \frac{\sum_{j=1}^T e^{\langle Qx_i, Kx_j\rangle/\sqrt{d_k}} V x_j}{\sum_{j=1}^T e^{\langle Qx_i, Kx_j\rangle/\sqrt{d_k}}}, \qquad i = 1, \dots, T.$$

Define the softmax-weighted mean $\bar{X}_T : \mathbb{R}^d \to \mathbb{R}^d$ by

$$\bar{X}_T(z) := \frac{\sum_{j=1}^T x_j e^{z^\top x_j/\sqrt{d_k}}}{\sum_{j=1}^T e^{z^\top x_j/\sqrt{d_k}}}.$$

The attention output can then be written as

$$f_T(x_i) = V \bar{X}_T(K^\top Q x_i), \qquad i = 1, \dots, T.$$

Both f_T and $\bar{X}_T$ depend on the entire context $(x_1, \dots, x_T)$. For notational simplicity, we indicate this dependence only through the context length T. This representation reduces the analysis of self-attention to that of the softmax-weighted mean.

For the probabilistic analysis, we model $(x_j)_{j\geq 1}$ as a stochastic process on $\mathbb{R}^d$. The softmax-weighted mean $\bar{X}_T(z)$ averages the context tokens using exponential weights determined by z, giving greater weight to tokens with higher query–key similarity scores. If the tokens are stationary, this weighted average may converge to a distribution-dependent readout acting on the query as the context grows. The following lemma formalizes this intuition.

Lemma 3.1 (Limit of the softmax-weighted mean)**.** *Assume that $(x_j)_{j\geq 1}$ is stationary and ergodic on $\mathbb{R}^d$, and let x have their common marginal distribution. For $z \in \mathbb{R}^d$ satisfying $\mathbb{E}e^{z^\top x/\sqrt{d_k}} < \infty$ and $\mathbb{E}\left[\|x\| e^{z^\top x/\sqrt{d_k}}\right] < \infty$,*

$$\bar{X}_T(z) \xrightarrow[T\to\infty]{\text{a.s.}} m(z) := \frac{\mathbb{E}[xe^{z^\top x/\sqrt{d_k}}]}{\mathbb{E}[e^{z^\top x/\sqrt{d_k}}]}. \tag{1}$$

The proof of Lemma 3.1 is deferred to §A.1. Additional structure in the token distribution can yield a more explicit expression for the limit $m(z)$. The following lemma shows how elliptical symmetry leads to such a characterization.

Lemma 3.2 (Alignment under elliptical symmetry)**.** *Let $\xi \in \mathbb{R}^r$ be rotationally invariant with $\mathbb{P}(\xi \neq 0) > 0$, meaning that $R\xi \overset{d}{=} \xi$ for every orthogonal matrix $R \in \mathbb{R}^{r\times r}$. Under the hypotheses of Lemma 3.1, assume that the common marginal law satisfies $x \overset{d}{=} \mu + B\xi$, where $\mu \in \mathbb{R}^d$ and $B \in \mathbb{R}^{d\times r}$ are deterministic. Set $\Sigma := BB^\top \succeq 0$. Then*

$$m(z) = \mu + \gamma_\Sigma(z)\Sigma z, \tag{2}$$

where $\gamma_\Sigma(z)$ is a scalar that depends on Σ and z only through the quadratic form $z^\top\Sigma z$, and $\gamma_\Sigma(z) > 0$ whenever $z^\top\Sigma z > 0$.

The matrix Σ is a scatter matrix. If $\mathbb{E}\|\xi\|^2 < \infty$, then $\mathrm{Cov}(x) = \kappa\Sigma$, where $\kappa := \mathbb{E}\|\xi\|^2/r > 0$; in particular, scatter and covariance coincide under the normalization $\mathrm{Cov}(\xi) = I_r$. Since positive rescaling does not affect alignment, we use covariance-based terminology throughout this paper. The proof of Lemma 3.2 is deferred to §A.2.

Covariance-directed attention alignment. We now return to self-attention. For a query token $u \in \mathbb{R}^d$, set $z = K^\top Qu$. Under the hypotheses of Lemma 3.2, abbreviate $\gamma_\Sigma(K^\top Qu)$ as γ. Then

$$f_T(u) = V\bar{X}_T(K^\top Qu) \xrightarrow[T\to\infty]{\text{a.s.}} V\mu + \gamma V\Sigma K^\top Qu. \tag{3}$$

Beyond the mean contribution, this limit provides a readout of the input's covariance.

3.2 Finite-Context Concentration of Softmax Attention

The preceding results characterize the alignment of softmax attention in the long-context regime. We now quantify its finite-context approximation error.

Assumption 3.3 (Elliptical sub-Gaussian inputs with strong mixing)**.** *Let $(x_j)_{j\geq 1}$ be strictly stationary and ergodic on $\mathbb{R}^d$, and let x have their common marginal distribution. Assume the following:*

(a) *(Elliptical sub-Gaussian marginal) $x \overset{d}{=} \mu + B\xi$, where $\mu \in \mathbb{R}^d$ and $B \in \mathbb{R}^{d\times r}$ are deterministic and $\xi \in \mathbb{R}^r$ is rotationally invariant with $\mathbb{P}(\xi \neq 0) > 0$ and $\mathbb{E}e^{u^\top\xi} \leq e^{\sigma^2\|u\|^2/2}$ for every $u \in \mathbb{R}^r$ and some $\sigma > 0$.*

(b) *(Strong mixing) Let $\alpha(h)$ denote the Rosenblatt α-mixing coefficient at integer lag $h \geq 1$. These coefficients satisfy $\tau_\alpha := 1 + 2\sum_{h=1}^{\infty}\sqrt{\alpha(h)} < \infty$.*

Proposition 3.4 (Concentration of the softmax-weighted mean)**.** *Under Assumption 3.3, set $\Sigma := BB^\top \succeq 0$. For $z \in \mathbb{R}^d$ satisfying $z^\top\Sigma z \leq d_k$,*

$$\mathbb{E}\left\|\bar{X}_T(z) - m(z)\right\|^2 \leq C\tau_\alpha\frac{\mathrm{tr}(\Sigma)}{T}, \qquad T \geq 1, \tag{4}$$

where C is a constant depending only on σ.

The mean-square error is therefore $O(T^{-1})$, with a bound proportional to τ_α and $\mathrm{tr}(\Sigma)$. The proof of Proposition 3.4 is deferred to §B.1. We next combine this bound with Lemma 3.2 to obtain a mean-square approximation bound for the attention output.

Attention alignment rate. Under Assumption 3.3, for a query token $u \in \mathbb{R}^d$, set $z = K^\top Qu$. If $z^\top \Sigma z \le d_k$, then

$$\mathbb{E}\,\|f_T(u) - V\mu - \gamma_\Sigma(z)V\Sigma z\|^2 \le \frac{C\tau_\alpha\,\|V\|_{\mathrm{op}}^2\,\mathrm{tr}(\Sigma)}{T}, \qquad T \ge 1, \tag{5}$$

where C is a constant depending only on σ. The bound quantifies how the covariance-directed alignment emerges as the context grows. The proof of equation 5 is deferred to §B.2.

Remark 3.5 (Query inclusion in self-attention)**.** *The preceding limit and concentration results are stated for fixed z. In self-attention, the query token also contributes to the keys and values. To account for its inclusion, for a query token $u \in \mathbb{R}^d$, set $z = K^\top Qu$, and append u to the context. The resulting attention output is $V\bar{X}^+_{T+1}(z)$, where*

$$\bar{X}^+_{T+1}(z) := \frac{\sum_{j=1}^T x_j e^{z^\top x_j/\sqrt{d_k}} + ue^{z^\top u/\sqrt{d_k}}}{\sum_{j=1}^T e^{z^\top x_j/\sqrt{d_k}} + e^{z^\top u/\sqrt{d_k}}}.$$

Under the hypotheses of Lemma 3.1 at this z, the correction satisfies $\left\|\bar{X}^+_{T+1}(z) - \bar{X}_T(z)\right\| = O(T^{-1})$ almost surely, so the limit is unchanged. Under Assumption 3.3 and $z^\top \Sigma z \le d_k$, we further have $\mathbb{E}\left\|\bar{X}^+_{T+1}(z) - \bar{X}_T(z)\right\|^2 = O(T^{-2})$. Thus including a fixed query preserves the $O(T^{-1})$ mean-square approximation bound. The proof is deferred to §B.3.

3.3 Alignment on Natural Language Data

We next examine whether the covariance-readout description remains informative for natural-language inputs beyond the distributional assumptions. We sample texts from MiniPile (Kaddour, 2023) and encode them using the pretrained tokenizers and input embeddings of TinyStories-3M, TinyStories-8M, TinyStories-28M, TinyStories-33M, TinyStories-1Layer-21M, and TinyStories-2Layers-33M (Eldan & Li, 2023). We apply one randomly initialized causal softmax attention layer for each model.

For each text and model, we estimate μ by averaging the full sequence of token embeddings and Σ by averaging their centered outer products. These estimates remain fixed as the causal prefix grows. At each position, we measure the cosine similarity between the attention output and the readout $V\mu + \gamma V\Sigma K^\top Qu$ in equation 3. Figure 1 shows that the mean cosine similarity rises toward one across all six embedding sources as the prefix length increases.

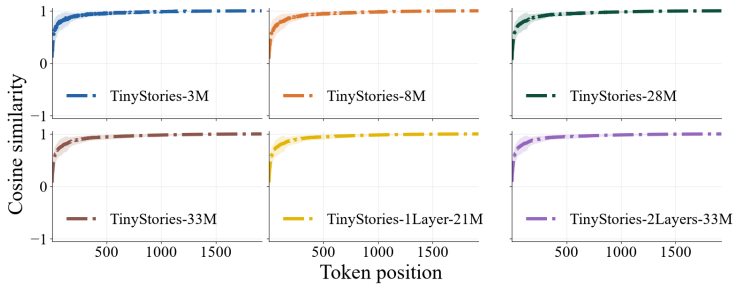


Figure 1: Alignment on natural-language data. We use the same 32 MiniPile training documents for all models, selecting documents with more than 1,920 tokens under every tokenizer, and retaining the first 1,920 tokens per tokenizer. Curves show mean cosine similarity across documents at each token position; shaded bands indicate one standard deviation.

4 Application I: Unveiling In-Context Learning

In this section, we apply our covariance-readout analysis to explain in-context learning in linear regression tasks. Section 4.1 shows that a residual softmax attention layer with certain task-independent parameters can emulate a gradient-descent step in the long-context regime. Section 4.2 examines whether such attention parameters can be learned during pretraining.

4.1 SOFTMAX ATTENTION CAN EMULATE GRADIENT DESCENT

Gradient descent as a data transformation. Consider a training dataset $\mathcal{D} = \{(u_i, v_i)\}_{i=1}^T$ comprising covariates $u_i \in \mathbb{R}^{N_u}$ and response $v_i \in \mathbb{R}^{N_v}$, and a linear model $v = Wu$ parameterized by the weight matrix $W \in \mathbb{R}^{N_v \times N_u}$. The goal is to minimize loss function $\mathcal{L}(W) := \frac{1}{T}\sum_{i=1}^T \|Wu_i - v_i\|^2$, and a gradient-descent step with learning rate η updates the weights by $\Delta W = -\eta \nabla_W \mathcal{L}(W)$. Following Von Oswald et al. (2023), this update can equivalently be expressed as a transformation of the targets. Setting $\Delta v_i := \Delta W u_i$, we have

$$\mathcal{L}(W + \Delta W) = \frac{1}{T}\sum_{i=1}^T \|(W + \Delta W)u_i - v_i\|^2 = \frac{1}{T}\sum_{i=1}^T \|Wu_i - (v_i - \Delta v_i)\|^2 .$$

The updated weights on the original data and the unchanged weights on the transformed data therefore yield identical residuals and weight gradients. Consequently, one gradient-descent step can be implemented by computing the gradient from the transformed data and the initial weights, then using it to transform the response while keeping the weights fixed.

In-context learning task. We now return to self-attention and consider an in-context learning task. The dataset $\mathcal{D}$ is represented by T demonstration tokens $e_i = (u_i^\top, v_i^\top)^\top \in \mathbb{R}^{N_u+N_v}$, each containing a covariate and its corresponding response. An additional query token $e_{T+1} = (u_{\text{test}}^\top, 0^\top)^\top$ contains the test covariate with its response masked. The task is to use the labeled demonstrations in the context to produce a prediction $\hat{v}_{\text{test}}$ for the unknown response v_{test}.

Softmax attention as a gradient-descent step. The data-transformation view of linear regression suggests that softmax attention could implement a gradient step if it can reproduce the updates to the response component of each token. We now give a construction that realizes this correspondence in the long-context regime.

Proposition 4.1 (Gradient-descent transformations through softmax attention)**.** *Assume that the demonstration tokens $(e_i)_{i\geq 1}$ are independent and identically distributed according to a centered Gaussian law. Initialize the regression weights at $W_0 = 0$ and choose a learning rate $\eta > 0$. Set $N = N_u + N_v$ and $d = d_k = d_v = N$, and define the key, query, and value matrices by*

$$K = Q = \begin{pmatrix} I_{N_u} & 0 \\ 0 & 0 \end{pmatrix}, \qquad V = -2\eta\sqrt{N}\begin{pmatrix} 0 & 0 \\ 0 & I_{N_v} \end{pmatrix}.$$

Let f_T denote the attention function evaluated against the T demonstration tokens, and write $\Delta W_0 := -\eta \nabla_W \mathcal{L}(W_0)$ for the gradient-descent update computed from these demonstrations. For each demonstration index i, the residual attention update $e_i^+ := e_i + f_T(e_i)$ satisfies

$$e_i^+ - \begin{pmatrix} u_i \\ v_i - \Delta W_0 u_i \end{pmatrix} \xrightarrow[T\to\infty]{\text{a.s.}} 0.$$

For the masked query $e_{\text{test}} = (u_{\text{test}}^\top, 0^\top)^\top$, define $e_{\text{test}}^+ := e_{\text{test}} + f_T(e_{\text{test}})$. The prediction $\hat{v}_{\text{test}} := -[0 \;\; I_{N_v}]e_{\text{test}}^+$ satisfies

$$\hat{v}_{\text{test}} - \Delta W_0 u_{\text{test}} \xrightarrow[T\to\infty]{\text{a.s.}} 0.$$

The gradient is evaluated at the initial weights W_0, and the residual attention layer outputs the transformed data, consistent with the data-transformation view of linear regression. These transformed tokens can be passed to the next layer, where the gradient is again evaluated at W_0 using the updated data, thus enabling multi-step gradient descent. The proof of Proposition 4.1 is deferred to §C.1.

The query prediction is recovered through the fixed projection $-[0 \;\; I_{N_v}]$, which reverses the sign of its updated response component. The construction also permits full self-attention by including the masked query among the keys and values, without changing the limit (Remark 3.5). And it is not unique: different choices of K and Q with the same product $K^\top Q$, together with the same V, produce identical attention updates.

4.2 Trained Transformers Do Emulate Gradient Descent

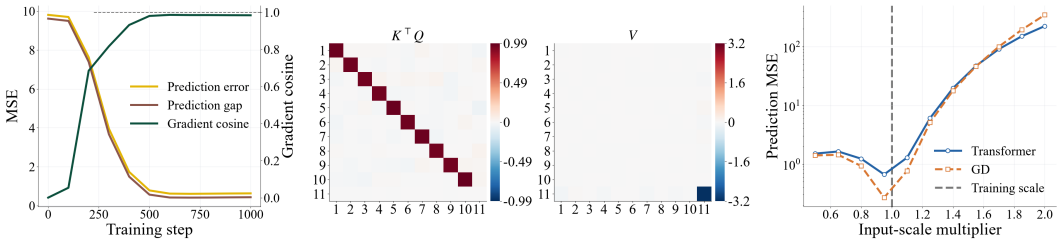
We now investigate experimentally whether training recovers the gradient-descent structure described in Proposition 4.1. For each task τ, we independently sample teacher weights $W_\tau^\star$ and i.i.d. covariates $\{u_{\tau,i}\}_{i=1}^T$, then generate responses $v_{\tau,i} = W_\tau^\star u_{\tau,i}$ for $i = 1, \dots, T$. The demonstrations are encoded as tokens $e_{\tau,i} = (u_{\tau,i}^\top, v_{\tau,i}^\top)^\top$, and the query token is $e_{\tau,T+1} = (u_{\tau,\text{test}}^\top, 0^\top)^\top$. Let $\hat{v}_\theta(\{e_{\tau,1}, \dots, e_{\tau,T}\}, e_{\tau,T+1})$ denote the prediction of a residual softmax attention model with parameters θ. We train these parameters to minimize

$$L(\theta) := \mathbb{E}_\tau \left[\left\| \hat{v}_\theta(\{e_{\tau,1}, \dots, e_{\tau,T}\}, e_{\tau,T+1}) - v_{\tau,\text{test}} \right\|^2 \right].$$

At every optimization step, we use Adam (Kingma & Ba, 2014) with a fresh batch of tasks.

One gradient-descent step and a single attention layer. We compare the single-layer prediction with $\hat{v}_{\text{GD}} := \Delta W_0 u_{\text{test}}$, obtained by taking one GD step from $W_0 = 0$ under the regression loss $\mathcal{L}(W)$. The step size is selected to minimize GD's prediction error on separate validation tasks. On held-out tasks, we average the following quantities through the training process: the prediction error $\|\hat{v}_\theta - v_{\text{test}}\|^2$, the prediction gap $\|\hat{v}_\theta - \hat{v}_{\text{GD}}\|^2$, and the gradient cosine $\cos(\nabla_{u_{\text{test}}} \hat{v}_\theta, \nabla_{u_{\text{test}}} \hat{v}_{\text{GD}})$.

Figure 2a shows decreasing prediction error and discrepancy from GD, with gradient cosine approaching one. To investigate whether this behavioral agreement reflects the proposed computation, we inspect the learned matrices in Figure 2b. The product $K^\top Q$ is approximately diagonal on the covariate block, while V is concentrated in the negative response–response entry, which resembles the construction in Proposition 4.1. To assess whether our in-context learner has found a generalizable update rule, we compare the trained attention layer with GD after rescaling covariates and responses as $(u, v) \mapsto (su, sv)$ in Figure 2c. With the trained parameters held fixed, the two predictors exhibit broadly similar error profiles across the tested scales.



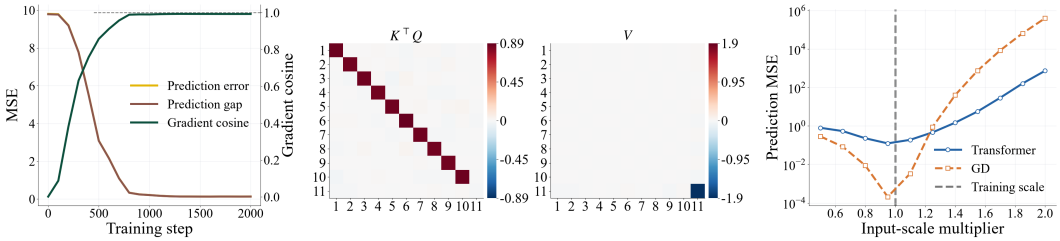
(a) Training dynamics. (b) Learned attention matrices. (c) Input scaling.

Figure 2: Comparing a pretrained single-layer softmax attention model with one GD step. We use $T = 512$, $N_u = 10$, and $N_v = 1$. (a) Query prediction error, prediction gap relative to GD, and cosine similarity during training. (b) Learned $K^\top Q$ and V. (c) Prediction error under input scaling.

Multiple gradient-descent steps and attention layers. We next train a four-layer looped softmax attention model and compare it with four GD steps, see Figure 3. The shared learned matrices retain the block structure of the proposed construction, see Figure 3b. Figure 3c shows that GD is more accurate near the training distribution, while the attention model deteriorates more slowly at larger scales. This is because the learned updates use a smaller effective step size, which could make them less susceptible to instability as input scaling increases the curvature of the regression loss.

5 Application II: Explaining LLM Repetition

Repetition in LLMs arises from distinct mechanisms. Mahaut & Franzon (2025) distinguish natural repetition, where repetitive loops emerge spontaneously after ordinary prompts, from ICL-induced repetition, where the model continues repetitive patterns supplied in the context. In this section, we focus on natural repetition and investigate how it arises from the structure of softmax attention. Section 5.1 extends the covariance-readout analysis to multiple layers and develops a Markovian explanation of natural repetition. Section 5.2 examines whether pretrained models exhibit the predicted reduction in context dependence during repetitive generation.



(a) Training dynamics. (b) Learned attention matrices. (c) Input scaling.

Figure 3: Comparing a pretrained four-layer looped softmax attention model with four GD steps. We use $T = 512$, $N_u = 10$, and $N_v = 1$. (a) Query prediction error, prediction gap relative to GD, and cosine similarity during training. (b) Learned $K^\top Q$ and V. (c) Prediction error under input scaling.

5.1 Multi-layer Alignment and Markovian Repetition

We first introduce the Transformer architecture and its context-free counterpart, then show that their outputs approach one another as the context grows.

Transformer architecture. We consider a depth-L GPT-2-style architecture. For a context of length T, let $x_i^\ell \in \mathbb{R}^d$ denote the input to layer ℓ at position i, where $\ell = 0, \dots, L-1$ and $i = 1, \dots, T$. The initial embedding x_i^0 includes any positional component. Each layer updates the representations according to

$$\tilde{x}_i^\ell = \mathrm{LN}_1^\ell(x_i^\ell), \qquad y_i^\ell = f_i^\ell(\tilde{x}_i^\ell), \qquad x_i^{\ell+1} = \Phi^\ell(x_i^\ell, y_i^\ell).$$

Here $\tilde{x}_i^\ell$ is the normalized attention input, and y_i^ℓ is the attention output before residual addition. The function f_i^ℓ is the self-attention function from Section 3.1, evaluated against the normalized prefix $(\tilde{x}_1^\ell, \dots, \tilde{x}_i^\ell)$ with matrices $K^\ell, Q^\ell \in \mathbb{R}^{d_k \times d}$ and $V^\ell \in \mathbb{R}^{d \times d}$. We set $d_v = d$, absorb the attention output projection into V^ℓ, and omit attention projection biases. The remaining operations are collected in the function $\Phi^\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, defined by

$$\Phi^\ell(x, y) := x + y + \mathrm{MLP}^\ell\Big(\mathrm{LN}_2^\ell(x + y)\Big).$$

The maps LN_1^ℓ and LN_2^ℓ are layer normalizations with separate learned affine parameters and a positive variance stabilizer, and MLP^ℓ is the positionwise feed-forward network. Thus Φ^ℓ incorporates both residual additions, the normalization before the MLP, and the MLP itself.

Context-free architecture. For a context of length T, let $\widehat{x}_i^\ell \in \mathbb{R}^d$ denote the input to layer ℓ at position i, where $\ell = 0, \dots, L-1$ and $i = 1, \dots, T$. The initial embedding $\widehat{x}_i^0 = x_i^0$ includes any positional component. At each layer, suppose the normalized attention inputs have a common elliptical marginal law represented by

$$\mathrm{LN}_1^\ell(\widehat{x}^\ell) \overset{d}{=} \mu_\ell + B_\ell \xi_\ell, \qquad \Sigma_\ell := B_\ell B_\ell^\top \succeq 0,$$

where $\mu_\ell \in \mathbb{R}^d$ and $B_\ell \in \mathbb{R}^{d \times r_\ell}$ are deterministic, and $\xi_\ell \in \mathbb{R}^{r_\ell}$ is rotationally invariant with $\mathbb{P}(\xi_\ell \neq 0) > 0$. Each layer updates the representations according to

$$\widehat{y}_i^\ell = V^\ell\big[\mu_\ell + \gamma_{\Sigma_\ell}(z_i^\ell)\Sigma_\ell z_i^\ell\big], \quad \text{where } z_i^\ell := (K^\ell)^\top Q^\ell \mathrm{LN}_1^\ell(\widehat{x}_i^\ell), \qquad \widehat{x}_i^{\ell+1} = \Phi^\ell(\widehat{x}_i^\ell, \widehat{y}_i^\ell).$$

Here $\widehat{x}_i^\ell$ and $\widehat{y}_i^\ell$ are the context-free counterparts of the layer input and attention output, respectively. This construction replaces the empirical attention output with its covariance readout.

The following corollary shows that the original transformer approaches its context-free counterpart as the context grows and quantifies the approximation error.

Corollary 5.1 (Multi-layer alignment). *For each $\ell = 0, \dots, L-1$, suppose the normalized context-free sequence $(\mathrm{LN}_1^\ell(\widehat{x}_i^\ell))_{i \geq 1}$ satisfies Assumption 3.3, with mixing factor τ_ℓ. Assume that each Φ^ℓ is globally Lipschitz. Then the terminal representations satisfy*

$$x_T^L - \widehat{x}_T^L \xrightarrow[T \to \infty]{\text{a.s.}} 0.$$

Moreover, for every $T \geq 1$,

$$\mathbb{E}\left\|x_T^L - \widehat{x}_T^L\right\|^2 \leq \frac{C_L}{T} \sum_{\ell=0}^{L-1} \tau_\ell, \tag{6}$$

where C_L is a constant depending on depth, dimensions, network parameters (including layer-normalization parameters), and Lipschitz bounds, but independent of T.

The proof of Corollary 5.1 is deferred to §D.2.

Asymptotically Markovian generation. Under autoregressive softmax sampling, the context-free architecture induces a first-order Markov process: the next-token distribution depends only on the current token and, when positional encodings are included, the current position, making the process generally time-inhomogeneous. Under the assumptions of Corollary 5.1, the transformer's representations approach their context-free counterparts, and so do its next-token distributions. Generation is therefore asymptotically first-order Markovian, with dependence on earlier individual tokens vanishing as the context grows.

In this Markovian regime, repetition can arise from the high-inflow problem identified by Fu et al. (2021): many tokens assign substantial transition probability to the same successor, favoring repeated returns to particular tokens. When these transitions form loops, generation can repeatedly traverse the same token sequences.

5.2 Empirical Analysis

We next empirically examine whether repetitive generation exhibits the reduced dependence on context suggested by our analysis. Using the same MiniPile prompts and six TinyStories models as in Figure 1, we generate continuations by greedy decoding with the full pretrained models. For continuations ending in repetitive loops, we measure the similarity between terminal representations associated with different occurrences of the same token, separately within the prompt tail and the generated continuation. Within each window, we average pairwise cosine similarities for each repeated token type, then average over token types and documents.

Table 1 shows high similarity between representations of the same token across different positions in all six models, which is already present in the prompt tails and persists during repetitive generation. The terminal representation is largely determined by the current token, with limited sensitivity to the preceding context. The similarities tend to decrease with embedding dimension and depth, suggesting that this degeneration is less pronounced in larger models.

Table 1: Same-token representation similarity during repetitive generation. Each model receives the same 32 MiniPile documents, truncated to 1,920 tokens, and generates 128 tokens by greedy decoding with the end-of-sequence token suppressed. The two rows report mean cosine similarity between occurrences of the same token within the final 128 prompt tokens and the 128 generated tokens, respectively, using only documents whose continuations are classified as repetitive.

	TinyStories-**3M**	TinyStories-**8M**	TinyStories-**28M**	TinyStories-**33M**	TinyStories-**1Layer-21M**	TinyStories-**2Layers-33M**
Prompt Tail	0.9820	0.9647	0.9608	0.9346	0.9884	0.9725
Generation	0.9815	0.9580	0.9534	0.9231	0.9810	0.9785

6 Conclusion

We have developed a covariance-readout description of softmax attention that connects ICL and natural repetition through a common *summarize and forget* mechanism. Our analysis characterizes the long-context limit and its finite-context error, showing how attention retains distributional information while losing explicit dependence on individual context tokens. This yields a task-independent construction that asymptotically emulates one gradient-descent step for linear regression and a multilayer extension supporting a conditional Markovian account of repetition. Experiments provide evidence consistent with the theoretical analysis.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis and interpreting results. Generating synthetic data sets, implementing methods, assisting with translation and cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for sourcing/searching for information and editing a research paper to improve readability. We have reviewed all AI-assisted work. All AI-assisted proof writing was carefully checked and revised by the authors for mathematical correctness and logical consistency. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our code is available at `https://anonymous.4open.science/r/Softmax-Attention-as-a-Covariance-Readout-9247`.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and François-Xavier Vialard. Token sample complexity of attention. *arXiv preprint arXiv:2512.10656*, 2025.

Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: a measure-based perspective. *arXiv preprint arXiv:2512.11784*, 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.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12848–12856, 2021.

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.

Tatsuya Hiraoka and Kentaro Inui. Repetition neurons: How do language models produce repetitions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 483–495, 2025.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Jean Kaddour. The minipile challenge for data-efficient language models. *arXiv preprint arXiv:2304.08442*, 2023.

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

Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective. *Advances in Neural Information Processing Systems*, 36:72888–72903, 2023.

Arvind Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations*, volume 2024, pp. 32077–32096, 2024.

Matéo Mahaut and Francesca Franzon. Repetitions are not all alike: distinct mechanisms sustain repetition in language models. *arXiv preprint arXiv:2504.01100*, 2025.

Emmanuel Rio et al. *Asymptotic theory of weakly dependent random processes*, volume 80. Springer, 2017.

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.

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.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Advances in Neural Information Processing Systems*, 35:3082–3095, 2022.

Appendix of "Softmax Attention as a Covariance Readout"

A Proofs of Section 3.1

A.1 Proof of Lemma 3.1

Proof of Lemma 3.1. Fix $z \in \mathbb{R}^d$ satisfying the stated moment assumptions, and define

$$D(z) := \mathbb{E}[e^{z^\top x/\sqrt{d_k}}], \qquad N(z) := \mathbb{E}[x e^{z^\top x/\sqrt{d_k}}].$$

The assumptions ensure that $D(z) < \infty$ and every coordinate of the vector defining $N(z)$ is integrable. Moreover, $D(z) > 0$, since the exponential is strictly positive. By stationarity, ergodicity, and Birkhoff's ergodic theorem,

$$\frac{1}{T}\sum_{j=1}^{T} e^{z^\top x_j/\sqrt{d_k}} \xrightarrow[T\to\infty]{\text{a.s.}} D(z), \qquad \frac{1}{T}\sum_{j=1}^{T} x_j e^{z^\top x_j/\sqrt{d_k}} \xrightarrow[T\to\infty]{\text{a.s.}} N(z),$$

where the second convergence follows coordinatewise. Taking the ratio gives

$$\bar{X}_T(z) \xrightarrow[T\to\infty]{\text{a.s.}} \frac{N(z)}{D(z)} = m(z).$$

□

A.2 Proof of Lemma 3.2

Proof of Lemma 3.2. Expectations under the common marginal law can be evaluated using the representation $x = \mu + B\xi$. Write

$$w := B^\top z, \qquad q := \|w\|^2 = z^\top \Sigma z, \qquad \vartheta := w^\top \xi = z^\top B\xi.$$

Canceling the factor $e^{z^\top \mu/\sqrt{d_k}}$ in the numerator and denominator yields

$$m(z) - \mu = \frac{N}{D}, \qquad N := \mathbb{E}[B\xi e^{\vartheta/\sqrt{d_k}}], \qquad D := \mathbb{E}[e^{\vartheta/\sqrt{d_k}}].$$

The pointwise moment assumptions imply $0 < D < \infty$ and $\mathbb{E}[\|B\xi\| e^{\vartheta/\sqrt{d_k}}] < \infty$, because $\|B\xi\| \le \|x\| + \|\mu\|$. They also imply $\mathbb{E}[|\vartheta| e^{\vartheta/\sqrt{d_k}}] < \infty$, since $|\vartheta| \le \|z\|\,\|B\xi\|$. Thus all expectations used below are well defined, without requiring a tilted first moment of the full vector ξ.

If $q = 0$, then $w = 0$ and $\vartheta = 0$. The moment assumptions give $\mathbb{E}\,\|B\xi\| < \infty$, and central symmetry yields $\mathbb{E}[B\xi] = 0$. Hence $m(z) = \mu$. We may therefore take $\gamma_\Sigma(z) = 0$ in this case.

Suppose now that $q > 0$, and define

$$R := \frac{2ww^\top}{q} - I_r.$$

This matrix is orthogonal and satisfies $Rw = w$. Consequently,

$$w^\top R\xi = \vartheta, \qquad BR\xi = \frac{2\Sigma z}{q}\vartheta - B\xi.$$

Rotational invariance gives $R\xi \overset{d}{=} \xi$, so

$$N = \mathbb{E}[BR\xi e^{\vartheta/\sqrt{d_k}}] = \frac{2\Sigma z}{q}\mathbb{E}[\vartheta e^{\vartheta/\sqrt{d_k}}] - N.$$

Rearranging gives

$$m(z) - \mu = \gamma_\Sigma(z)\Sigma z, \qquad \gamma_\Sigma(z) := \frac{\mathbb{E}[\vartheta e^{\vartheta/\sqrt{d_k}}]}{q\,\mathbb{E}[e^{\vartheta/\sqrt{d_k}}]}.$$

The argument applies to arbitrary $B \in \mathbb{R}^{d\times r}$, including singular and rectangular matrices.

To establish positivity, observe that $\xi \overset{d}{=} -\xi$, so ϑ is symmetric. Symmetry also implies $\mathbb{E}[|\vartheta| e^{-\vartheta/\sqrt{d_k}}] < \infty$. Therefore

$$\mathbb{E}[\vartheta e^{\vartheta/\sqrt{d_k}}] = \frac{1}{2}\mathbb{E}\Big[\vartheta\left(e^{\vartheta/\sqrt{d_k}} - e^{-\vartheta/\sqrt{d_k}}\right)\Big] = \mathbb{E}[\vartheta \sinh(\vartheta/\sqrt{d_k})].$$

The integrand is nonnegative and is strictly positive when $\vartheta \neq 0$. Since $w \neq 0$, rotational invariance and $\mathbb{P}(\xi \neq 0) > 0$ imply $\mathbb{P}(\vartheta \neq 0) > 0$: otherwise, rotating $w/\|w\|$ to each coordinate direction would show that every coordinate of ξ vanishes almost surely. Thus $\gamma_\Sigma(z) > 0$.

Finally, rotational invariance gives $\vartheta \overset{d}{=} \sqrt{q}\,\xi_1$, where ξ_1 denotes the first coordinate of ξ. Hence

$$\gamma_\Sigma(z) = \frac{\mathbb{E}[\xi_1 e^{\sqrt{q}\,\xi_1/\sqrt{d_k}}]}{\sqrt{q}\,\mathbb{E}[e^{\sqrt{q}\,\xi_1/\sqrt{d_k}}]}, \qquad q > 0.$$

For a fixed law of ξ and fixed d_k, this expression depends on Σ and z only through $q = z^\top \Sigma z$.

For a deterministic query token u, set $z = K^\top Q u$. Applying Lemma 3.1 and the linear map V gives $f_T(u) \xrightarrow[T\to\infty]{\text{a.s.}} V\mu + \gamma_\Sigma(z) V\Sigma z$, proving equation 3. □

B Proofs of Section 3.2

B.1 Proof of Proposition 3.4

We first introduce the mixing coefficients and an auxiliary moment bound. For $1 \le a \le b \le \infty$, write $\mathcal{F}_a^b := \sigma(x_j : a \le j \le b)$ and define

$$\alpha(h) := \sup_{t\ge 1} \sup_{A\in\mathcal{F}_1^t,\, C\in\mathcal{F}_{t+h}^\infty} |\mathbb{P}(A\cap C) - \mathbb{P}(A)\mathbb{P}(C)|, \qquad h \ge 1.$$

For random variables U and W measurable with respect to $\mathcal{F}_1^t$ and $\mathcal{F}_{t+h}^\infty$, respectively, the standard covariance inequalities give

$$|\operatorname{Cov}(U,W)| \le C\sqrt{\alpha(h)}\,\|U\|_{L^4}\,\|W\|_{L^4}, \qquad |\operatorname{Cov}(U,W)| \le 4\alpha(h)\,\|U\|_{L^\infty}\,\|W\|_{L^\infty},$$

whenever the corresponding norms are finite; see Rio et al. (2017, Eq. (1.12b)) for the first inequality. Here C is an absolute constant. Applying a measurable function to each token does not increase the mixing coefficients.

Lemma B.1 (Fourth moment of bounded mixing sums)**.** *Let $(Y_j)_{j\ge1}$ be centered real random variables with $|Y_j| \le 1$ almost surely and strong-mixing coefficients bounded by $\alpha(h)$. Assume that $\tau_\alpha := 1 + 2\sum_{h=1}^\infty \sqrt{\alpha(h)} < \infty$. Then*

$$\mathbb{E}\left|\sum_{j=1}^T Y_j\right|^4 \le CT^2\tau_\alpha^2, \qquad T \ge 1,$$

for an absolute constant C.

Proof of Lemma B.1. Set $\beta(0) := 1$ and $\beta(h) := \alpha(h)$ for $h \ge 1$. For indices $i \le j \le k \le \ell$, centering and the bounded covariance inequality give

$$|\mathbb{E}[Y_iY_jY_kY_\ell]| \le 4\min\{\beta(j-i), \beta(\ell-k)\} \le 4\sqrt{\beta(j-i)\beta(\ell-k)}.$$

For a positive outer gap, apply the covariance inequality with the corresponding endpoint as one factor; for a zero gap, use $|Y_j| \le 1$. Expanding the fourth power and ordering the indices introduces a factor of at most 24. For fixed outer gaps $j-i$ and $\ell-k$, there are at most T^2 choices of i and $k-j$. Hence

$$\mathbb{E}\left|\sum_{j=1}^T Y_j\right|^4 \le 96T^2\left(\sum_{h=0}^{T-1}\sqrt{\beta(h)}\right)^2 \le 96T^2\tau_\alpha^2.$$

□

Proof of Proposition 3.4. Fix a deterministic $z \in \mathbb{R}^d$ with $z^\top \Sigma z \le d_k$. Throughout the proof, C_σ denotes a finite constant depending only on σ, whose value may change between occurrences. Introduce the proof auxiliaries $\zeta_j := x_j - \mu$ and $\zeta := x - \mu \overset{d}{=} B\xi$, and define

$$W_j := e^{z^\top \zeta_j/\sqrt{d_k}}, \qquad D_T := \frac{1}{T}\sum_{j=1}^{T} W_j, \qquad N_T := \frac{1}{T}\sum_{j=1}^{T} \zeta_j W_j, \qquad W := e^{z^\top \zeta/\sqrt{d_k}}.$$

The common factor $e^{z^\top \mu/\sqrt{d_k}}$ cancels from the original empirical and population softmax-weighted means, so

$$\bar{X}_T(z) = \mu + \frac{N_T}{D_T}, \qquad m(z) = \mu + m_c, \qquad m_c := \frac{\mathbb{E}[\zeta W]}{\mathbb{E}W}.$$

This identity holds for every deterministic μ and leaves the attention inputs and the query unchanged. Set $U_j := (\zeta_j - m_c)W_j$ and $R_T := T^{-1}\sum_{j=1}^{T} U_j$. Then $\mathbb{E}U_j = 0$ and

$$\bar{X}_T(z) - m(z) = \frac{R_T}{D_T}.$$

The sub-Gaussian assumption on ξ and the restriction on z give

$$\mathbb{E}e^{v^\top \zeta} \le e^{\sigma^2 v^\top \Sigma v/2}, \qquad \mathbb{E}W^a \le e^{\sigma^2 a^2/2}, \qquad v \in \mathbb{R}^d, \quad a \in \mathbb{R}.$$

For every fixed $p \ge 2$, the resulting scalar moment bounds and Minkowski's inequality yield

$$\left\|v^\top \zeta\right\|_{L^p} \le C_{p,\sigma}\sqrt{v^\top \Sigma v}, \qquad \mathbb{E}\left\|\zeta\right\|^p \le C_{p,\sigma}\big(\mathrm{tr}(\Sigma)\big)^{p/2}.$$

Indeed, for an orthonormal basis $(e_a)_{a=1}^d$, one has $\left\|\|\zeta\|^2\right\|_{L^{p/2}} \le \sum_{a=1}^d \left\|e_a^\top \zeta\right\|_{L^p}^2$. Since $\mathbb{E}\zeta = 0$, Jensen's inequality gives $\mathbb{E}W \ge 1$. Combining these estimates with Hölder's inequality yields

$$\left\|(v^\top \zeta)W\right\|_{L^4} \le \left\|v^\top \zeta\right\|_{L^8} \|W\|_{L^8} \le C_\sigma\sqrt{v^\top \Sigma v}, \qquad |v^\top m_c| \le C_\sigma\sqrt{v^\top \Sigma v}.$$

It follows that

$$\left\|v^\top U_1\right\|_{L^4}^2 \le C_\sigma v^\top \Sigma v, \qquad \|m_c\|^2 \le C_\sigma \,\mathrm{tr}(\Sigma). \tag{7}$$

The sequence $(U_j)_{j\ge 1}$ is centered and stationary, with mixing coefficients bounded by those of $(x_j)_{j\ge 1}$. Applying the L^4 covariance inequality and summing over lags gives

$$\mathbb{E}|v^\top R_T|^2 \le \frac{C\tau_\alpha}{T}\left\|v^\top U_1\right\|_{L^4}^2 \le \frac{C_\sigma \tau_\alpha}{T} v^\top \Sigma v.$$

Summing this bound over an orthonormal basis yields

$$\mathbb{E}\left\|R_T\right\|^2 \le C_\sigma \tau_\alpha \frac{\mathrm{tr}(\Sigma)}{T}. \tag{8}$$

To control small values of the denominator, set $p_0 := \mathbb{P}(z^\top \zeta \ge 0)$, $I_j := \mathbf{1}_{\{z^\top \zeta_j \ge 0\}}$, and $Y_j := I_j - p_0$. Central symmetry of ζ gives $p_0 \ge 1/2$, and $W_j \ge I_j$. Thus $D_T < 1/4$ implies $|\sum_{j=1}^T Y_j| > T/4$. Lemma B.1 and Markov's inequality yield

$$\mathbb{P}(D_T < 1/4) \le \frac{4^4}{T^4}\mathbb{E}\left|\sum_{j=1}^{T} Y_j\right|^4 \le \frac{C\tau_\alpha^2}{T^2}. \tag{9}$$

We next bound the fourth moment of the softmax-weighted mean relative to μ, uniformly in T. By convexity,

$$D_T^{-8} \le \frac{1}{T}\sum_{j=1}^{T} W_j^{-8}, \qquad \|N_T\|^8 \le \frac{1}{T}\sum_{j=1}^{T} \left\|\zeta_j W_j\right\|^8.$$

Stationarity and the marginal moment bounds give $\mathbb{E}D_T^{-8} \le \mathbb{E}W^{-8} \le C_\sigma$ and

$$\mathbb{E}\left\|N_T\right\|^8 \le \mathbb{E}[\|\zeta\|^8 W^8] \le \big(\mathbb{E}\left\|\zeta\right\|^{16}\big)^{1/2}\big(\mathbb{E}W^{16}\big)^{1/2} \le C_\sigma\big(\mathrm{tr}(\Sigma)\big)^4.$$

Consequently, Cauchy–Schwarz yields

$$\mathbb{E}\left\|\bar{X}_T(z)-\mu\right\|^4 \le \left(\mathbb{E}\left\|N_T\right\|^8\right)^{1/2}\left(\mathbb{E}D_T^{-8}\right)^{1/2} \le C_\sigma\big(\mathrm{tr}(\Sigma)\big)^2.$$

Together with equation 7, this gives

$$\mathbb{E}\left\|\bar{X}_T(z)-m(z)\right\|^4 \le C_\sigma\big(\mathrm{tr}(\Sigma)\big)^2. \tag{10}$$

Finally, let $A_T := \{D_T \ge 1/4\}$. On this event, $\left\|\bar{X}_T(z)-m(z)\right\|^2 \le 16\left\|R_T\right\|^2$, so equation 8 gives

$$\mathbb{E}\Big[\left\|\bar{X}_T(z)-m(z)\right\|^2 \mathbf{1}_{A_T}\Big] \le C_\sigma\tau_\alpha\frac{\mathrm{tr}(\Sigma)}{T}.$$

On the complementary event, Cauchy–Schwarz, equation 9, and equation 10 give

$$\mathbb{E}\Big[\left\|\bar{X}_T(z)-m(z)\right\|^2 \mathbf{1}_{A_T^c}\Big] \le \left(\mathbb{E}\left\|\bar{X}_T(z)-m(z)\right\|^4\right)^{1/2}\mathbb{P}(A_T^c)^{1/2} \le C_\sigma\tau_\alpha\frac{\mathrm{tr}(\Sigma)}{T}.$$

Adding the two estimates proves equation 4. Every bound involves only the law of $x-\mu$, so the constant is independent of μ. □

B.2 Proof of equation 5

Proof of equation 5. Fix a deterministic query token u and set $z = K^\top Qu$. Under Assumption 3.3, Lemma 3.2 gives $m(z) = \mu + \gamma_\Sigma(z)\Sigma z$, with the convention $\gamma_\Sigma(z) = 0$ when $z^\top\Sigma z = 0$. Using $f_T(u) = V\bar{X}_T(z)$, we obtain

$$\begin{aligned}\mathbb{E}\Big[\left\|f_T(u)-V[\mu+\gamma_\Sigma(z)\Sigma z]\right\|^2\Big] &= \mathbb{E}\Big[\left\|V(\bar{X}_T(z)-m(z))\right\|^2\Big]\\ &\le \left\|V\right\|_{\mathrm{op}}^2\,\mathbb{E}\left\|\bar{X}_T(z)-m(z)\right\|^2\\ &\le \frac{C\tau_\alpha\left\|V\right\|_{\mathrm{op}}^2\,\mathrm{tr}(\Sigma)}{T},\end{aligned}$$

where the last inequality follows from Proposition 3.4. □

B.3 Proof of Remark 3.5

Proof of Remark 3.5. Fix a deterministic query token u and set $z = K^\top Qu$. Define

$$b := e^{z^\top u/\sqrt{d_k}}, \qquad W_j := e^{z^\top x_j/\sqrt{d_k}}, \qquad D_T := \frac{1}{T}\sum_{j=1}^T W_j.$$

Writing $\Delta_T := \bar{X}_{T+1}^+(z) - \bar{X}_T(z)$, direct subtraction gives

$$\Delta_T = \frac{b}{TD_T+b}\big(u-\bar{X}_T(z)\big).$$

Under the hypotheses of Lemma 3.1 at z, Birkhoff's ergodic theorem gives $D_T \to D := \mathbb{E}W_1 \in (0,\infty)$ almost surely, while $\bar{X}_T(z) \to m(z)$ almost surely. Hence

$$T\Delta_T \xrightarrow[T\to\infty]{\text{a.s.}} \frac{b}{D}\big(u-m(z)\big).$$

The limit is finite, so $\left\|\Delta_T\right\| = O(T^{-1})$ almost surely. This is the rate of the self-inclusion correction, not a convergence rate for $\bar{X}_T(z)-m(z)$. In particular, $V\bar{X}_{T+1}^+(z) \xrightarrow[T\to\infty]{\text{a.s.}} Vm(z)$.

Now impose Assumption 3.3 and $z^\top\Sigma z \le d_k$. Let C_σ denote a finite constant depending only on σ, whose value may change between occurrences. Define

$$\widetilde{W}_j := e^{z^\top(x_j-\mu)/\sqrt{d_k}}, \qquad \widetilde{D}_T := \frac{1}{T}\sum_{j=1}^T \widetilde{W}_j = e^{-z^\top\mu/\sqrt{d_k}}D_T, \qquad b_\mu := e^{z^\top(u-\mu)/\sqrt{d_k}}.$$

Canceling the same common factor in the exact correction gives

$$\Delta_T = \frac{b_\mu}{T\widetilde{D}_T + b_\mu}\big((u-\mu) - (\bar{X}_T(z)-\mu)\big).$$

The sub-Gaussian bound gives $\mathbb{E}\widetilde{W}_1^a \le e^{\sigma^2 a^2/2}$ for every $a \in \mathbb{R}$. Convexity and stationarity therefore imply

$$\mathbb{E}\widetilde{D}_T^{-4} \le \frac{1}{T}\sum_{j=1}^{T}\mathbb{E}\widetilde{W}_j^{-4} = \mathbb{E}\widetilde{W}_1^{-4} \le e^{8\sigma^2}.$$

The moment estimates in the proof of Proposition 3.4 give $\mathbb{E}\left\|\bar{X}_T(z)-\mu\right\|^4 \le C_\sigma(\mathrm{tr}(\Sigma))^2$, uniformly in T. Consequently,

$$\left(\mathbb{E}\left\|u-\bar{X}_T(z)\right\|^4\right)^{1/2} \le C_\sigma\big(\|u-\mu\|^2 + \mathrm{tr}(\Sigma)\big).$$

Using $T\widetilde{D}_T + b_\mu \ge T\widetilde{D}_T$ and applying Cauchy–Schwarz yields

$$\mathbb{E}\left\|\Delta_T\right\|^2 \le \frac{b_\mu^2}{T^2}\big(\mathbb{E}\widetilde{D}_T^{-4}\big)^{1/2}\left(\mathbb{E}\left\|u-\bar{X}_T(z)\right\|^4\right)^{1/2} \le \frac{C_\sigma b_\mu^2}{T^2}\big(\|u-\mu\|^2 + \mathrm{tr}(\Sigma)\big). \tag{11}$$

This proves an $O(T^{-2})$ mean-square bound for the self-inclusion correction, with a constant depending on the fixed query through b_μ and $\|u-\mu\|$. Since $\bar{X}^+_{T+1}(z) - m(z) = \bar{X}_T(z) - m(z) + \Delta_T$, combining equation 11 with Proposition 3.4 gives

$$\mathbb{E}\left\|\bar{X}^+_{T+1}(z) - m(z)\right\|^2 \le C_\sigma\tau_\alpha\frac{\mathrm{tr}(\Sigma)}{T} + \frac{C_\sigma b_\mu^2}{T^2}\big(\|u-\mu\|^2 + \mathrm{tr}(\Sigma)\big) = O(T^{-1}).$$

Thus the $O(T^{-2})$ term measures the additional error from including the query, while the population approximation remains $O(T^{-1})$ in mean square. Applying V preserves these rates, with an additional factor $\|V\|_{\mathrm{op}}^2$ and population target $V[\mu + \gamma_\Sigma(z)\Sigma z]$. The $C_\sigma\tau_\alpha\,\mathrm{tr}(\Sigma)/T$ bound for the original context is independent of μ; the additional self-inclusion bound depends on the query relative to μ. □

C Proofs of Section 4

C.1 Proof of Proposition 4.1

Proof of Proposition 4.1. Let $e = (u^\top, v^\top)^\top$ have the common Gaussian distribution of the demonstration tokens, and write

$$C := \mathbb{E}[ee^\top] = \begin{pmatrix} C_{uu} & C_{uv} \\ C_{vu} & C_{vv} \end{pmatrix}, \qquad C_{vu} = \mathbb{E}[vu^\top].$$

Since $W_0 = 0$, the gradient-descent update computed from the first T demonstrations is

$$\Delta W_0 = -\eta\nabla_W\mathcal{L}(W_0) = \frac{2\eta}{T}\sum_{j=1}^{T} v_j u_j^\top \xrightarrow[T\to\infty]{\text{a.s.}} 2\eta C_{vu},$$

where the convergence follows entrywise from the strong law of large numbers. In particular, ΔW_0 depends on the demonstrations and on T.

We next identify the attention limit. The Gaussian moment-generating function satisfies $\mathbb{E}e^{t^\top e} = e^{t^\top Ct/2}$ for every $t \in \mathbb{R}^N$. Differentiating this identity and setting $t = z/\sqrt{N}$ gives

$$m(z) = \frac{\mathbb{E}[e\, e^{z^\top e/\sqrt{N}}]}{\mathbb{E}[e^{z^\top e/\sqrt{N}}]} = \frac{Cz}{\sqrt{N}}.$$

All required exponential moments are finite, including when C is singular. Applying Lemma 3.1 to the demonstration context, for any deterministic query token $a = (a_u^\top, a_v^\top)^\top$, we obtain

$$f_T(a) \xrightarrow[T\to\infty]{\text{a.s.}} \frac{1}{\sqrt{N}}VCK^\top Qa = \begin{pmatrix} 0 \\ -2\eta C_{vu}a_u \end{pmatrix}.$$

Here we used $K^\top Q = \mathrm{diag}(I_{N_u}, 0)$ and $V = -2\eta\sqrt{N}\ \mathrm{diag}(0, I_{N_v})$.

To apply this limit to a demonstration token e_i, fix its index i and condition on e_i. The remaining demonstration tokens are still independent with the original Gaussian law, while the query e_i is fixed under this conditioning. The strong law therefore applies to the attention numerator and denominator with the ith term omitted. Restoring that term adds a finite vector or scalar divided by T to the corresponding empirical average, leaving both limits unchanged. Consequently,

$$f_T(e_i) \xrightarrow[T\to\infty]{\text{a.s.}} \begin{pmatrix} 0 \\ -2\eta C_{vu} u_i \end{pmatrix}.$$

Taking a countable intersection of probability-one events establishes this convergence for every fixed demonstration index i. Combining it with the convergence of ΔW_0 yields

$$e_i^+ - \begin{pmatrix} u_i \\ v_i - \Delta W_0 u_i \end{pmatrix} = f_T(e_i) + \begin{pmatrix} 0 \\ \Delta W_0 u_i \end{pmatrix} \xrightarrow[T\to\infty]{\text{a.s.}} 0,$$

which proves the first assertion.

Finally, hold the test covariate u_{test} fixed as T grows. Since the masked query has zero response component, its prediction satisfies

$$\hat{v}_{\text{test}} = -[0\ \ I_{N_v}] e_{\text{test}}^+ = -[0\ \ I_{N_v}] f_T(e_{\text{test}}) \xrightarrow[T\to\infty]{\text{a.s.}} 2\eta C_{vu} u_{\text{test}}.$$

The empirical gradient update also satisfies $\Delta W_0 u_{\text{test}} \xrightarrow[T\to\infty]{\text{a.s.}} 2\eta C_{vu} u_{\text{test}}$. Subtracting the two convergent quantities gives

$$\hat{v}_{\text{test}} - \Delta W_0 u_{\text{test}} \xrightarrow[T\to\infty]{\text{a.s.}} 0,$$

completing the proof. □

D Proofs of Section 5

D.1 Auxiliary estimates

We first establish a uniform approximation bound for bounded attention inputs and a stability estimate for causal self-attention. Throughout this section, write $s := \sqrt{d_k}$.

Lemma D.1 (Uniform approximation of bounded softmax-weighted means)**.** *Let $(U_j)_{j\ge 1}$ be strictly stationary and ergodic on $\mathbb{R}^d$, with $\|U_j\| \le R$ almost surely and strong-mixing coefficients satisfying $\tau := 1 + 2\sum_{h=1}^{\infty} \sqrt{\alpha(h)} < \infty$. Let U have their common marginal distribution, and define*

$$\bar{U}_T(z) := \frac{\sum_{j=1}^T U_j e^{z^\top U_j/s}}{\sum_{j=1}^T e^{z^\top U_j/s}}, \qquad m_U(z) := \frac{\mathbb{E}[U e^{z^\top U/s}]}{\mathbb{E}[e^{z^\top U/s}]}.$$

For every finite $S \ge 0$, set $E_T := \sup_{\|z\|\le S} \big\|\bar{U}_T(z) - m_U(z)\big\|$. Then

$$E_T \xrightarrow[T\to\infty]{\text{a.s.}} 0, \qquad \mathbb{E}E_T^2 \le C_* R^2 e^{4RS/s} \frac{\tau}{T},$$

where $C_* > 0$ is a universal constant.

Proof of Lemma D.1. We first record a vector version of the strong-mixing covariance inequality. If random vectors A and B have finite fourth moments and are measurable with respect to sigma-fields separated by lag h, then

$$|\mathbb{E}\langle A - \mathbb{E}A, B - \mathbb{E}B\rangle| \le C\sqrt{\alpha(h)} \big(\mathbb{E}\,\|A\|^4\big)^{1/4} \big(\mathbb{E}\,\|B\|^4\big)^{1/4},$$

with a constant independent of the vector dimension. Indeed, let G be an auxiliary independent standard Gaussian vector. Apply the scalar L^4 covariance inequality to $\langle G, A\rangle$ and $\langle G, B\rangle$ for each fixed G, and then average over G. The identities $\mathbb{E}_G[\langle G, a\rangle\langle G, b\rangle] = \langle a, b\rangle$ and $\mathbb{E}_G|\langle G, a\rangle|^4 = 3\,\|a\|^4$, together with Cauchy–Schwarz, give the stated bound.

For $k \geq 1$, define the empirical tensor-moment error

$$\Delta_{T,k} := \frac{1}{T}\sum_{j=1}^{T} U_j^{\otimes k} - \mathbb{E}[U^{\otimes k}],$$

where the tensor space carries its Euclidean norm, so that $\left\|U^{\otimes k}\right\| = \|U\|^k$. Applying the preceding covariance inequality and summing over lags yields

$$\mathbb{E}\left\|\Delta_{T,k}\right\|^2 \leq CR^{2k}\frac{\tau}{T}.$$

Here and below, C denotes a universal constant whose value may change between occurrences.

Write

$$D_T(z) := \frac{1}{T}\sum_{j=1}^{T} e^{z^\top U_j/s}, \qquad N_T(z) := \frac{1}{T}\sum_{j=1}^{T} U_j e^{z^\top U_j/s},$$

and let $D(z) := \mathbb{E}[e^{z^\top U/s}]$ and $N(z) := \mathbb{E}[Ue^{z^\top U/s}]$. Expanding the exponential and bounding tensor contractions gives

$$\sup_{\|z\|\leq S} |D_T(z) - D(z)| \leq \sum_{k=1}^{\infty} \frac{(S/s)^k}{k!}\left\|\Delta_{T,k}\right\|,$$

$$\sup_{\|z\|\leq S} \|N_T(z) - N(z)\| \leq \sum_{k=0}^{\infty} \frac{(S/s)^k}{k!}\left\|\Delta_{T,k+1}\right\|.$$

The series converge absolutely because the inputs are bounded. Setting $q := RS/s$ and applying Minkowski's inequality, we obtain

$$\left(\mathbb{E}\sup_{\|z\|\leq S} |D_T(z) - D(z)|^2\right)^{1/2} \leq C\big(e^q - 1\big)\sqrt{\frac{\tau}{T}},$$

$$\left(\mathbb{E}\sup_{\|z\|\leq S} \|N_T(z) - N(z)\|^2\right)^{1/2} \leq CRe^q\sqrt{\frac{\tau}{T}}.$$

For every $\|z\| \leq S$, we have $D_T(z) \geq e^{-q}$ and $\|m_U(z)\| \leq R$. Consequently,

$$\bar{U}_T(z) - m_U(z) = \frac{N_T(z) - N(z) - m_U(z)\big(D_T(z) - D(z)\big)}{D_T(z)},$$

and therefore

$$\big(\mathbb{E}E_T^2\big)^{1/2} \leq CRe^{2q}\sqrt{\frac{\tau}{T}}.$$

Squaring proves the asserted bound for a suitable universal constant C_*.

For almost-sure convergence, apply Birkhoff's theorem simultaneously to all coordinates of $U^{\otimes k}$ for every integer $k \geq 1$. On the resulting probability-one event, $\Delta_{T,k} \to 0$ for every k. Moreover, $\|\Delta_{T,k}\| \leq 2R^k$, so the tails of the two exponential series above are uniformly bounded by convergent factorial series. Their finite partial sums converge to zero, which proves uniform convergence of both numerator and denominator. The deterministic denominator bound then gives $E_T \to 0$ almost surely. All suprema are measurable because the functions are continuous on the compact query ball. □

Lemma D.2 (Stability of causal self-attention)**.** *For a prefix $a = (a_1, \ldots, a_T)$, define*

$$\mathcal{A}_T(a) := V\frac{\sum_{j=1}^{T} a_j e^{\langle Qa_T, Ka_j\rangle/s}}{\sum_{j=1}^{T} e^{\langle Qa_T, Ka_j\rangle/s}}.$$

Suppose $\|a_j\|, \|b_j\| \leq R$ for two prefixes a and b. Set $\eta_j := \|a_j - b_j\|$ and $M := \left\|K^\top Q\right\|_{\mathrm{op}} R^2/s$. Then

$$\|\mathcal{A}_T(a) - \mathcal{A}_T(b)\| \leq \|V\|_{\mathrm{op}}\left[2M\eta_T + \frac{e^{2M}(1+2M)}{T}\sum_{j=1}^{T}\eta_j\right].$$

Proof of Lemma D.2. Interpolate between the prefixes by setting $c_j(t) := b_j + t(a_j - b_j)$ for $t \in [0,1]$. Define

$$s_j(t) := \frac{\langle Qc_T(t), Kc_j(t)\rangle}{s}, \qquad p_j(t) := \frac{e^{s_j(t)}}{\sum_{k=1}^{T} e^{s_k(t)}}, \qquad \bar{c}(t) := \sum_{j=1}^{T} p_j(t)c_j(t).$$

Since $\|c_j(t)\| \le R$, all scores lie in $[-M, M]$, and hence $p_j(t) \le e^{2M}/T$. Differentiating the weighted mean gives

$$\dot{\bar{c}}(t) = \sum_{j=1}^{T} p_j(t)\dot{c}_j(t) + \sum_{j=1}^{T} p_j(t)\big(c_j(t) - \bar{c}(t)\big)\dot{s}_j(t).$$

Moreover,

$$\|c_j(t) - \bar{c}(t)\| \le 2R, \qquad |\dot{s}_j(t)| \le \frac{\|K^\top Q\|_{\mathrm{op}} R}{s}\big(\eta_T + \eta_j\big).$$

It follows that

$$\|\dot{\bar{c}}(t)\| \le 2M\eta_T + (1+2M)\sum_{j=1}^{T} p_j(t)\eta_j \le 2M\eta_T + \frac{e^{2M}(1+2M)}{T}\sum_{j=1}^{T}\eta_j.$$

Integrating over $t \in [0,1]$ and applying V proves the result. □

D.2 Proof of Corollary 5.1

Proof of Corollary 5.1. Layer normalization with fixed affine parameters and a positive variance stabilizer is globally bounded and globally Lipschitz. Thus, for each layer, there are finite constants R_ℓ and β_ℓ satisfying

$$\left\|\mathrm{LN}_1^\ell(x)\right\| \le R_\ell, \qquad \left\|\mathrm{LN}_1^\ell(x) - \mathrm{LN}_1^\ell(x')\right\| \le \beta_\ell \|x - x'\|.$$

For learned gain g_ℓ, bias h_ℓ, and variance stabilizer $\varepsilon_\ell > 0$, one may take

$$R_\ell := \sqrt{d}\,\|g_\ell\|_\infty + \|h_\ell\|, \qquad \beta_\ell := \frac{\|g_\ell\|_\infty}{\sqrt{\varepsilon_\ell}}.$$

Define

$$U_i^\ell := \mathrm{LN}_1^\ell(\widehat{x}_i^\ell), \qquad S_\ell := \left\|(K^\ell)^\top Q^\ell\right\|_{\mathrm{op}} R_\ell, \qquad M_\ell := \frac{R_\ell S_\ell}{s}, \qquad v_\ell := \left\|V^\ell\right\|_{\mathrm{op}}.$$

Then the context-free tilt $z_i^\ell = (K^\ell)^\top Q^\ell U_i^\ell$ satisfies $\left\|z_i^\ell\right\| \le S_\ell$.

By the layerwise instance of Assumption 3.3, $(U_i^\ell)_{i\ge 1}$ is strictly stationary and ergodic, has marginal $\mu_\ell + B_\ell\xi_\ell$, and has mixing factor τ_ℓ. The layer-normalization bound supplies the boundedness required by Lemma D.1.

Let $\bar{U}_i^\ell(z)$ denote the softmax-weighted mean of $(U_1^\ell, \dots, U_i^\ell)$, and let $m_\ell(z)$ be its marginal tilted mean. Boundedness ensures all required exponential moments are finite. Lemma 3.2 therefore gives

$$m_\ell(z) = \mu_\ell + \gamma_{\Sigma_\ell}(z)\Sigma_\ell z.$$

In particular, $\widehat{y}_i^\ell = V^\ell m_\ell(z_i^\ell)$. When $z^\top\Sigma_\ell z = 0$, this expression is interpreted as $m_\ell(z) = \mu_\ell$.

Set

$$E_i^\ell := \sup_{\|z\| \le S_\ell} \left\|\bar{U}_i^\ell(z) - m_\ell(z)\right\|.$$

By Lemma D.1,

$$E_i^\ell \xrightarrow[i\to\infty]{\text{a.s.}} 0, \qquad \mathbb{E}(E_i^\ell)^2 \le C_* R_\ell^2 e^{4M_\ell}\frac{\tau_\ell}{i}. \tag{12}$$

The supremum permits evaluation at the context-dependent tilt z_i^ℓ without any independence assumption.

To compare the two architectures, introduce the empirical attention output on the context-free prefix,

$$\check{y}_i^\ell := V^\ell \bar{U}_i^\ell(z_i^\ell).$$

Then

$$\|\check{y}_i^\ell - \widehat{y}_i^\ell\| \le v_\ell E_i^\ell.$$

Write $\delta_i^\ell := \|x_i^\ell - \widehat{x}_i^\ell\|$. Applying Lemma D.2 to the actual and context-free normalized prefixes gives

$$\|y_i^\ell - \check{y}_i^\ell\| \le v_\ell \beta_\ell \left[2M_\ell \delta_i^\ell + \frac{e^{2M_\ell}(1+2M_\ell)}{i} \sum_{j=1}^{i} \delta_j^\ell \right].$$

Choose $\lambda_\ell < \infty$ so that

$$\|\Phi^\ell(x,y) - \Phi^\ell(x',y')\| \le \lambda_\ell\big(\|x-x'\| + \|y-y'\|\big).$$

Combining the preceding estimates yields

$$\delta_i^{\ell+1} \le a_\ell \delta_i^\ell + \frac{b_\ell}{i} \sum_{j=1}^{i} \delta_j^\ell + \lambda_\ell v_\ell E_i^\ell, \tag{13}$$

where

$$a_\ell := \lambda_\ell\big(1 + 2M_\ell v_\ell \beta_\ell\big), \qquad b_\ell := \lambda_\ell v_\ell \beta_\ell e^{2M_\ell}(1+2M_\ell).$$

We first prove almost-sure convergence. The two architectures have the same initial embeddings, so $\delta_i^0 = 0$. Suppose $\delta_i^\ell \to 0$ almost surely. Its Cesàro averages also converge to zero, and equation 12 gives $E_i^\ell \to 0$ almost surely. Equation equation 13 therefore implies $\delta_i^{\ell+1} \to 0$ almost surely. Induction over the fixed number of layers proves

$$x_T^L - \widehat{x}_T^L \xrightarrow[T\to\infty]{\text{a.s.}} 0.$$

For the mean-square bound, define

$$e_i^\ell := \big(\mathbb{E}(\delta_i^\ell)^2\big)^{1/2}, \qquad \rho_\ell := a_\ell + 2b_\ell, \qquad \nu_\ell := \sqrt{C_*}\lambda_\ell v_\ell R_\ell e^{2M_\ell}.$$

Minkowski's inequality, equation 13, and equation 12 give

$$e_i^{\ell+1} \le a_\ell e_i^\ell + \frac{b_\ell}{i} \sum_{j=1}^{i} e_j^\ell + \nu_\ell \sqrt{\frac{\tau_\ell}{i}}.$$

Since $\sum_{j=1}^{i} j^{-1/2} \le 2\sqrt{i}$, induction shows that

$$e_i^\ell \le \frac{H_\ell}{\sqrt{i}}, \qquad H_0 := 0, \qquad H_{\ell+1} := \rho_\ell H_\ell + \nu_\ell \sqrt{\tau_\ell}.$$

Expanding this recursion yields

$$H_L = \sum_{\ell=0}^{L-1} \nu_\ell \sqrt{\tau_\ell} \prod_{k=\ell+1}^{L-1} \rho_k.$$

Consequently, Cauchy–Schwarz gives

$$\mathbb{E}\,\|x_T^L - \widehat{x}_T^L\|^2 \le \frac{H_L^2}{T} \le \frac{C_L}{T} \sum_{\ell=0}^{L-1} \tau_\ell,$$

where an admissible constant is

$$C_L := \sum_{\ell=0}^{L-1} \nu_\ell^2 \prod_{k=\ell+1}^{L-1} \rho_k^2, \tag{14}$$

with empty products equal to one. All factors depend only on the fixed architecture, normalization parameters, and Lipschitz bounds, and are independent of T. This proves equation 6. □

The expression equation 14 makes the depth dependence explicit. If $\rho_\ell \leq A$ and $\nu_\ell \leq B$ uniformly across layers, with $A \geq 1$, then

$$C_L \leq B^2 \sum_{j=0}^{L-1} A^{2j}.$$

Thus this bound is linear in depth when $A = 1$ and geometric when $A > 1$. Dimension enters through the normalization bounds R_ℓ, the score bounds M_ℓ, and the layerwise operator and Lipschitz bounds.