TRANSFORMERS AS RANDOM-LAG AUTOREGRESSIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we consider a probabilistic generative model approach to explain the Transformer architecture and leads to improved new architectures. We make a minor modification to the classical autoregressive model by replacing its fixed unit lag with a *random lag*, and prove that the exact posterior mean under the resulting random-lag autoregressive model (RLA) is precisely causal Softmax attention. As a conditional expectation, this attention operator is the unique mean-square-optimal estimator of the latent value readout given the observed prefix. In RLA, all the standard transformer components are endowed with a generative role, such as RoPE, ALiBi positional encodings, and query-key normalization (QK-Norm). By replacing the Gaussian assumption of RLA with a linear directional family on the sphere and again computing the exact posterior mean, we obtain a *spherical linear attention* (SLA). Similar to softmax attention, SLA has normalized and non-negative attention weights. The experiments show the competitive performance of SLA.

1 INTRODUCTION

Understanding and designing the Transformer architecture has been a central problem (Vaswani et al., 2017). There are three main research lines toward a principled modeling of its softmax attention operator: (1) the **kernel-based view** interprets softmax attention as normalized kernel smoothing over input similarities, and its kernel reformulation enabled linear-time attention through feature maps (Tsai et al., 2019; Katharopoulos et al., 2020); (2) the **energy-based view** identifies attention with the update rule that minimizes the energy of a continuous modern Hopfield network, establishing attention as associative-memory retrieval with exponential storage capacity (Ramsauer et al., 2021); (3) the **compression-based view**, represented by CRATE, derives attention and the feed-forward network as gradient-descent steps that compress and sparsify token representations under a sparse rate reduction objective, yielding the first white-box explanation of the full Transformer block (Yu et al., 2023). However, despite diversified developments in this area, there is still no optimality proof for the *exact* softmax attention formulation. In this paper, we address this problem from a probabilistic generative model approach.

This work is mainly motivated by CRATE. Starting from the marginal distribution of an individual token, it models unordered token set as Gaussian mixtures on low-dimensional subspaces and characterizes attention and feed-forward neural network (FFN) as compression and sparse-coding steps for minimizing a sparse rate reduction objective, respectively. In contrast, we model an ordered token sequence through the conditional distribution $p(\mathbf{z}_n \mid \mathbf{z}_{<n})$ and specify its joint distribution through the well-known autoregressive factorization in LLM research:

$$p(\mathbf{z}_{1:N}) = p(\mathbf{z}_1) \prod_{n=2}^{N} p(\mathbf{z}_n \mid \mathbf{z}_{<n}). \tag{1}$$

In (1), we assume $p(\mathbf{z}_n \mid \mathbf{z}_{<n}) \sim \mathcal{N}(\boldsymbol{\mu}_n, \mathbf{D})$. Then we get our main results by following the three steps: First, we specify the conditional dependence by a *random parent position*: each token is generated from an earlier position drawn at random rather than from a fixed predecessor. Second, the posterior weights over candidate parents are computed from Bayes' rule, which corresponds to attention weights. Third, the formulation of softmax attention is given by the exact posterior mean of the value under the generative model. In words, the generative story is simple: every token is

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.

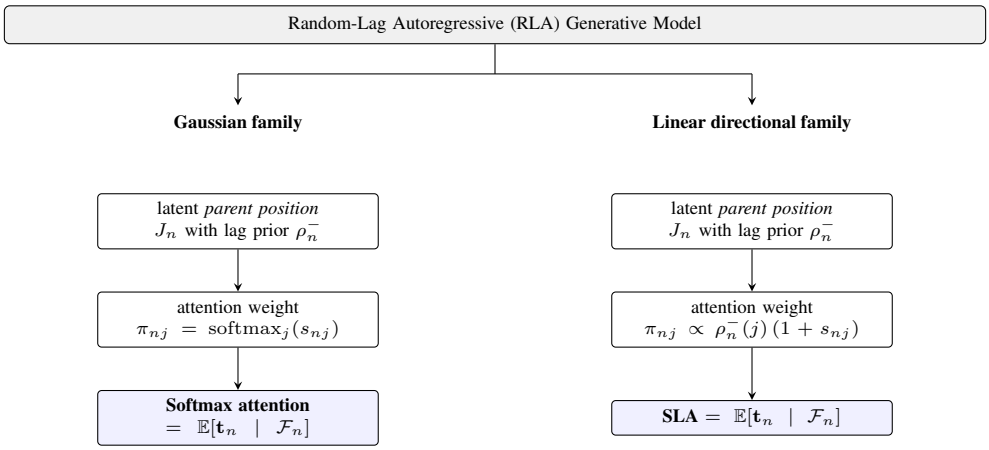


Figure 1: Overview of the paper. A single generative recipe (top) is used twice. With the Gaussian family (§2), the exact posterior mean over a latent parent position is softmax attention, and the standard components arise as parts of the posterior score. Replacing only the distribution with the linear directional family (§3) gives SLA.

a random-lag autoregression of one earlier token in history and then attention is the Bayes-optimal inference of the value. Figure 1 shows how this single recipe is used twice in this paper—with a Gaussian family it yields softmax attention, and with a *linear directional family* it yields *spherical linear attention*. Our contributions can be summarized as follows:

First, we propose a **random-lag autoregressive model (RLA)** for token sequence. Relative to its classical ancestor, the mixture transition distribution (Raftery, 1985; Le et al., 1996), RLA introduces lag-dependent transitions and history-dependent mixing weights, and shifts the object of interest from the model itself to the exact posterior over its latent parent position (§2.1). We prove that all the components of softmax attention—the query–key cross term, the key energy term, and the RoPE and ALiBi positional encodings—arise from the posterior mean of RLA, with RoPE characterized as the canonical norm-preserving transition. Since this posterior mean is a conditional expectation—the Bayes estimator under squared loss—the identification is also an MSE-optimality statement for the exact softmax attention formulation (Remark 2.6).

Second, we design a novel **spherical linear attention (SLA)** by simply changing the Gaussian assumption of RLA to the linear directional family. Both softmax attention and SLA are exact posterior means over parent positions, while SLA is computed under the linear directional distribution. As result, nonnegativity, normalization, and exact relative position are therefore consequences of the posterior rather than properties engineered into a feature map. Meanwhile, a by-product that may be of independent interest: the well-known gates in linear attention become learnable instantiations of free factors in RLA.

Third, we **pretrain and evaluate SLA** at the 340M-parameter scale on 15B SlimPajama tokens. Under the complete GLA evaluation protocol, our model SLA achieves the competitive six-task commonsense average and the best Wiki./LMB. perplexity among linear attention baselines, and outperforms GLA on two of three recall-intensive tasks, while retaining a constant-memory recurrent form.

How to read this paper. RLA is an explanatory model: its parameters are not estimated by likelihood but instantiated as the projection matrices of the layer. Section 2 develops the Gaussian branch: §2.1 defines the model, Theorem 2.3 derives the score decomposition (6), Theorem 2.4 shows that norm preservation forces the RoPE form of the transition, and Theorem 2.5 with Remark 2.6 identifies standard softmax attention as the exact posterior mean, which is the minimum mean-squared-error estimate. Section 3 changes only the distributional family to obtain SLA and its constant-memory recurrence, and Section 4 reports experiments of SLA. Proofs are collected in the appendices; readers interested only in the architecture may read Theorem 2.5 and §4.

1.1 Other Related Work

Positional structure and normalization. RoPE encodes relative position by rotating queries and keys(Su et al., 2021), while ALiBi introduces position encoding through an additive bias proportional to distance(Press et al., 2022). QK-Norm further normalize queries and keys and learning the score scale (Henry et al., 2020). In this paper, we show that RoPE is a natural result under the condition of keeping key-norm, while the ALiBi position encoding corresponds to a geometric prior in RLA. Then the normalization of key norm is necessary for getting the exact softmax attention formulation

Linear Attention and Recurrent Inference. Linear attention drops the softmax normalizer so that the value average becomes a running state, giving $O(N)$ computation with $O(1)$ memory at inference (Katharopoulos et al., 2020). Qin et al. (2022) secured normalized nonnegative weights from a ReLU map and added a cosine reweighting that decomposes exactly by angle difference, bringing locality inside the linear form. Sun et al. (2023) build relative position into the state transition As comparison, our SLA obtains these properties from a single source rather than by assembling them: its weights are the exact posterior mean of the linear directional observation model on the sphere, where nonnegativity is the legality of that density, normalization is its partition function, and relative position is the state transition. The recurrence property is therefore not from a designed approximation of softmax attention but from the properties of RLA under the linear directional family.

Gating Mechanism of Linear Attention. Hochreiter & Schmidhuber (1997) introduced the input and output gates, controlling how much of a new observation enters the memory cell and how much of the cell is exposed; Gers et al. (2000) added the forget gate once it emerged that an ungated cell saturates on long streams, giving the state a multiplicative per-step decay; Cho et al. (2014) then coupled writing and forgetting into a single update gate. Softmax attention dispensed with all of them. Sun et al. (2023) reinstate forgetting as a fixed geometric decay, identical at every token, while Gu & Dao (2023) and Yang et al. (2024a) make it data-dependent and per-channel. Dao & Gu (2024) then show that a recurrence with scalar decay and a masked linear attention are the same object. Schlag et al. (2021) gate the write instead: a delta rule erases the association a key already carries before storing a new one; Yang et al. (2025) combine the two, decay for forgetting at large and a delta step for targeted replacement. In this paper, we show that the free factors such as the lag prior ρ_n^- and the self weight η_n provides a natural entry for imposing data dependent gates.

2 Random-Lag Sequence Generation and Attention as a Posterior Mean

The observed information is $\mathcal{F}_n = \sigma(\mathbf{z}_{\le n})$. Throughout, $\mathcal{F}_n$ simply means "everything observed up to position n", and $\mathbb{E}[\cdot \mid \mathcal{F}_n]$ averages over the hidden parent choices given the sequence so far; the measure-theoretic details are collected in Appendix A. In the autoregressive joint density, call $p(c_n, \mathbf{z}_n \mid \mathbf{z}_{<n})$ the factor for row n. In this paper, a latent variable c_n is *local*—a per-token local latent variable in the standard Bayesian sense—when the joint density has the form $\prod_n p(c_n, \mathbf{z}_n \mid \mathbf{z}_{<n})$: no factor for row $m \neq n$ depends on c_n. Dependence through observed states is allowed. Appendix A gives the proof for this section.

2.1 The Generative Model

We model the sequence through observed states $\mathbf{z}_n$. For a fixed head dimension $d_h \le d$ and a fixed lag $\ell > 0$, the classical fixed-lag linear Gaussian autoregressive model:

$$\text{Fixed lag:} \quad \mathbf{z}_n = \mathbf{A}_\ell \mathbf{z}_{n-\ell} + \boldsymbol{\varepsilon}_n, \quad n > \ell, \tag{2}$$

A fixed lag ℓ requires initial states $\mathbf{z}_1, \dots, \mathbf{z}_\ell$; the usual first-order model takes $\ell = 1$. By comparison, our random-lag Gaussian autoregressive model (RLA) belongs to the family of mixture transition distribution (MTD) models (Raftery, 1985; Le et al., 1996) with random lag transition:

$$\text{Random lag:} \quad \ell_n \sim \rho_n(\cdot \mid \mathbf{z}_{<n}), \qquad \mathbf{z}_n = \mathbf{A}_{\ell_n} \mathbf{z}_{n-\ell_n} + \boldsymbol{\varepsilon}_n, \quad n \ge 2. \tag{3}$$

Below, we give the complete definition of RLA.

Assumption 2.1 (Definition of RLA)**.** On a common probability space, let $\mathcal{F}_0 = \{\varnothing, \Omega\}$ and $\mathcal{F}_n = \sigma(\mathbf{z}_1, \ldots, \mathbf{z}_n)$. The model is specified by four components:

(A1) **Initial distribution.** Take $\mathbf{D} \succ 0$, set $p_1 = \mathcal{N}(0, \mathbf{D})$, and draw $\mathbf{z}_1 \sim p_1$.

(A2) **Random-lag transition.** Fix $d_h \leq d$; take full-column-rank matrices $\mathbf{P}_q, \mathbf{P}_k \in \mathbb{R}^{d \times d_h}$ and $\mathbf{\Theta} \in \mathbb{R}^{d_h \times d_h}$, and define $\mathbf{A}_\ell = \mathbf{P}_q \mathbf{\Theta}^\ell \mathbf{P}_k^\top \quad (\ell \geq 1)$. For $n \geq 2$, given past observations, sample from $\mathcal{L}_n = \{1, \ldots, n-1\}$ according to $\ell_n \sim \rho_n(\cdot \mid \mathbf{z}_{<n})$, set $J_n = n - \ell_n$, and specify the conditional law of $\mathbf{z}_n$ in two parts:

  (a) *Distribution family*: the conditional Gaussian family with fixed covariance $\mathbf{D}$, $p(\mathbf{z}_n \mid J_n = j, \mathcal{F}_{n-1}) = \mathcal{N}(\mathbf{z}_n; \boldsymbol{\mu}_{nj}, \mathbf{D})$;

  (b) *Mean–parent relation*: the mean is the low-rank prediction of the parent state, $\boldsymbol{\mu}_{nj} = \mathbf{A}_{n-j}\mathbf{z}_j = \mathbf{P}_q \mathbf{\Theta}^{n-j} \mathbf{k}_j$, with $\mathbf{k}_j = \mathbf{P}_k^\top \mathbf{z}_j$.

(A3) **Value projection.** Given $\mathbf{P}_v \in \mathbb{R}^{d \times d_h}$, define $\mathbf{v}_j = \mathbf{P}_v^\top \mathbf{z}_j$, $j = 1, \ldots, N$.

(A4) **Value readout.** For any $\mathcal{F}_n$-measurable $\eta_n \in [0, 1]$, define

$$\mathbf{t}_n = (1 - \eta_n)\mathbf{v}_{J_n} + \eta_n \mathbf{v}_n \quad (n \geq 2), \qquad \mathbf{t}_1 = \mathbf{v}_1.$$

Why autoregressive, and how RLA differs from the classical MTD. The transition equation $\mathbf{z}_n = \mathbf{A}_{\ell_n}\mathbf{z}_{n-\ell_n} + \boldsymbol{\varepsilon}_n$ carries no hidden state: the states $\mathbf{z}_n$ are observed, and the only latent variable is the discrete parent position J_n. We therefore say *autoregressive* rather than *state-space*: the process is an observed, innovation-driven autoregression whose order is drawn at random at every step, not a hidden-state model in the sense of Kalman-type filters or of structured state-space layers (Gu & Dao, 2023). With the states observed, RLA is a variable-order Markov chain, and its observation conditional (4) is exactly a mixture transition distribution (Raftery, 1985; Le et al., 1996). It departs from the classical MTD in three substantive ways: (i) *Lag-dependent transitions.* The classical MTD mixes a single transition kernel Q over lags, $p(\mathbf{z}_n \mid \mathbf{z}_{<n}) = \sum_\ell \lambda_\ell Q(\mathbf{z}_n \mid \mathbf{z}_{n-\ell})$, so the lag only selects which past state enters. RLA instead gives each lag its own kernel through the powers of a single transition matrix, $\mathbf{A}_\ell = \mathbf{P}_q \mathbf{\Theta}^\ell \mathbf{P}_k^\top$: the lag acts *inside* the kernel, which is what makes relative-position geometry (rotations) identifiable from the model at all. With $\mathbf{\Theta} = \mathbf{I}$ the kernels coincide and the Gaussian MTD is recovered exactly. (ii) *History-dependent mixing.* The mixing weights $\rho_n(\cdot \mid \mathbf{z}_{<n})$ may depend on the observed past, whereas the MTD proportions λ_ℓ are global constants; the forget gate of Table 1 is exactly such a history-dependent mixing distribution. (iii) *Object of inference.* The MTD is used as a stand-alone forecasting model: it is fitted to an observed series, and its output is the predictive law of $\mathbf{z}_n$. Here the object of interest is the exact posterior of the mixture index J_n and the posterior mean of the value readout $\mathbf{t}_n$ in (A4), which equals casual softmax attention; the self-inclusive extension of Theorem 2.5 has no MTD counterpart.

Measurability and posterior mean. The information set $\mathcal{F}_{n-1}$ contains the observations before the current one is generated, and $\mathcal{F}_n$ additionally contains the current observation. Thus, for all $j \leq n$, the values $\mathbf{v}_j$ in (A3) are $\mathcal{F}_n$-measurable. Since J_n in (A4) is unobserved, $\mathbf{t}_n$ is generally not $\mathcal{F}_n$-measurable; its conditional mean $\mathbb{E}[\mathbf{t}_n \mid \mathcal{F}_n]$ is determined by the observed prefix. Components (A1)–(A2) specify the generative distribution; (A3)–(A4) define the value readout whose posterior mean is used by Attention.

Remark 2.2 (Time additivity and matrix powers). For lag ℓ, let the induced transformation on the head coordinates be T_ℓ. If addition of time intervals corresponds to composition, namely $T_0 = \mathbf{I}$ and $T_{a+b} = T_a T_b$ $(a, b \geq 0)$, the discrete-time semigroup property gives $T_\ell = T_1^\ell$. Setting $T_1 = \mathbf{\Theta}$ yields the parameterization $\mathbf{\Theta}^\ell$ in (A2).

The observation conditional distribution is

$$p(\mathbf{z}_n \mid \mathbf{z}_{<n}) = \sum_{\ell=1}^{n-1} \rho_n(\ell \mid \mathbf{z}_{<n}) \mathcal{N}(\mathbf{z}_n; \mathbf{A}_\ell \mathbf{z}_{n-\ell}, \mathbf{D}). \tag{4}$$

Together with p_1, this determines the observed joint distribution through Equation (1).

2.2 Exact Parent Posterior

Theorem 2.3 (RLA parent-position posterior)**.** *Under Assumption 2.1, let $J_n = n - \ell_n$ and define*

$$\mathbf{q}_n = \mathbf{P}_q^\top \mathbf{D}^{-1} \mathbf{z}_n, \qquad \mathbf{k}_j = \mathbf{P}_k^\top \mathbf{z}_j, \qquad \rho_n^-(j) = \rho_n(n - j \mid \mathbf{z}_{<n}). \tag{5}$$

Then, for $j < n$, the exact posterior over the parent position is

$$\begin{aligned} s_{nj} &= \log \rho_n^-(j) + \langle \mathbf{q}_n, \mathbf{\Theta}^{n-j} \mathbf{k}_j \rangle - \tfrac{1}{2} \|\mathbf{\Theta}^{n-j} \mathbf{k}_j\|_{\mathbf{H}}^2, \\ \pi_{nj} &:= \mathbb{P}(J_n = j \mid \mathcal{F}_n) = \mathrm{softmax}_{j<n}(s_{nj}). \end{aligned} \tag{6}$$

where $\mathbf{H} = \mathbf{P}_q^\top \mathbf{D}^{-1} \mathbf{P}_q$. If $\mathbf{\Theta}^\top \mathbf{H} \mathbf{\Theta} = \mathbf{H}$ and $\|\mathbf{k}_j\|_{\mathbf{H}} = c_K$, the energy term is the common constant $c_K^2/2$, and the score reduces to

$$s_{nj} = \log \rho_n^-(j) + \langle \mathbf{q}_n, \mathbf{\Theta}^{n-j} \mathbf{k}_j \rangle - \tfrac{1}{2} c_K^2. \tag{7}$$

Thus, the lag prior gives an additive relative bias, transition powers give relative positional transformations, and the query–key inner product measures content matching.

Proof: Appendix B.

2.3 RoPE and key normalization

Theorem 2.4 ($\mathbf{H}$-orthogonality of the transition and the RoPE parameterization)**.** *Let $\mathbf{P}_q \in \mathbb{R}^{d \times d_h}$ have full column rank and let $\nu > 0$. Choose $\mathbf{D} = \mathbf{P}_q \mathbf{P}_q^\top + \nu \mathbf{I}$, write $\mathbf{G} = \mathbf{P}_q^\top \mathbf{P}_q$, and define $\mathbf{H} = \mathbf{G}(\mathbf{G} + \nu \mathbf{I})^{-1} = \mathbf{P}_q^\top \mathbf{D}^{-1} \mathbf{P}_q \succ 0$. Let the head dimension be $d_h = 2r$. For an invertible matrix $\mathbf{\Theta} \in \mathbb{R}^{d_h \times d_h}$, assume $\det \mathbf{\Theta} > 0$ and the adjacent-lag energy identity $\|\mathbf{\Theta}^2 \mathbf{u}\|_{\mathbf{H}}^2 = \|\mathbf{\Theta} \mathbf{u}\|_{\mathbf{H}}^2$ for all $\mathbf{u}$. Because $\mathbf{\Theta}$ is invertible, this identity implies the isometry condition under the metric $\mathbf{H}$, $\mathbf{\Theta}^\top \mathbf{H} \mathbf{\Theta} = \mathbf{H}$, i.e., $\mathbf{\Theta}$ is $\mathbf{H}$-orthogonal. Then there exist an orthogonal matrix $\mathbf{S}$ and angles $\theta_1, \ldots, \theta_r$ such that*

$$\begin{aligned} \mathbf{\Theta} &= \mathbf{H}^{-1/2} \mathbf{S} \mathbf{R} \mathbf{S}^\top \mathbf{H}^{1/2}, \qquad \mathbf{R} = \mathrm{blkdiag}(R(\theta_1), \ldots, R(\theta_r)), \\ R(\theta) &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \end{aligned} \tag{8}$$

Moreover, for every $\mathbf{u} \in \mathbb{R}^{d_h}$,

$$\|\mathbf{\Theta}^\ell \mathbf{u}\|_{\mathbf{H}}^2 = \|\mathbf{u}\|_{\mathbf{H}}^2, \qquad \ell = 1, 2, \ldots.$$

Conversely, this parameterization satisfies the assumptions for arbitrary $\theta_i \in \mathbb{R}$ and preserves energy at every integer lag. Thus, $\theta_1, \ldots, \theta_{d_h/2}$ are free parameters that may be fixed in advance or learned (modulo 2π).

For a specified transition, $\mathbf{S}$ is an orthonormal basis reducing $\mathbf{H}^{1/2} \mathbf{\Theta} \mathbf{H}^{-1/2}$ to rotation blocks, and need not be unique. When constructing a model through Equation (8), any orthogonal $\mathbf{S}$ is admissible. The proof and equivalent expressions for the rotary inner product appear in Appendix C.

Projection parameterization in rotary coordinates. Let $\mathbf{W}_k = \mathbf{P}_k^\top$ and $\mathbf{W}_v = \mathbf{P}_v^\top$. Define $\widetilde{\mathbf{W}}_q = \mathbf{S}^\top \mathbf{H}^{-1/2} \mathbf{W}_q$ and $\widetilde{\mathbf{W}}_k = \mathbf{S}^\top \mathbf{H}^{1/2} \mathbf{W}_k$, leaving $\mathbf{W}_v$ unchanged. Then

$$\begin{aligned} \mathbf{\Theta}^\ell &= \mathbf{H}^{-1/2} \mathbf{S} \mathbf{R}^\ell \mathbf{S}^\top \mathbf{H}^{1/2}, \\ \mathbf{q}_n^\top \mathbf{\Theta}^{n-j} \mathbf{k}_j &= (\mathbf{R}^{-n} \widetilde{\mathbf{W}}_q \mathbf{z}_n)^\top (\mathbf{R}^{-j} \widetilde{\mathbf{W}}_k \mathbf{z}_j). \end{aligned} \tag{9}$$

The powers follow by cancellation of adjacent inverse factors, with no commutativity assumption on $\mathbf{H}$ and $\mathbf{R}$. Thus ordinary RoPE acts on transformed projections, while $\mathbf{\Theta}$ need not itself be a rotation in the original coordinates. For a fixed model these transformations can be absorbed into projection weights.

Small-noise regime and key normalization. As ν in $\mathbf{D} = \mathbf{P}_q\mathbf{P}_q^\top + \nu\mathbf{I}$ is a free hyperparameter to be specified, we can assume $\nu \to 0^+$, *i.e.*, the small-noise regime. In this regime, assume that $\mathbf{P}_q \in \mathbb{R}^{d\times d_h}$ has full column rank ($\mathrm{rank}(\mathbf{P}_q) = d_h$), and keep $\mathbf{P}_q$ fixed as $\nu \to 0^+$. Define

$$\mathbf{G} = \mathbf{P}_q^\top\mathbf{P}_q \succ 0, \qquad g_{\min} := \lambda_{\min}(\mathbf{G}) = \sigma_{\min}(\mathbf{P}_q)^2 > 0. \tag{10}$$

The small-noise condition is $\nu/g_{\min} \ll 1$. Since $\mathbf{H} = \mathbf{G}(\mathbf{G} + \nu\mathbf{I})^{-1}$, its eigenvalues give

$$(1 + \nu/g_{\min})^{-1}\mathbf{I} \preceq \mathbf{H} \preceq \mathbf{I}, \qquad \mathbf{H} \to \mathbf{I} \quad (\nu \to 0^+). \tag{11}$$

As result, in this regime, normalizing the key under the metric $\mathbf{H}$ in (6) reduces to Euclidean key normalization.

2.4 $\mathcal{F}$-Measurable Algebra and Standard Attention

During generation, parents come only from the past. After observing the current token, its value and self-weight are $\mathcal{F}_n$-measurable. Standard self-inclusive attention can therefore be expressed as a posterior mean under the same model, while the generative transition in Assumption 2.1 remains conditioned on the strict past.

Theorem 2.5 (Standard attention is an exact posterior mean of the generative model)**.** *Under the local-latent structure in Assumption 2.1, take $\mathbf{\Theta}^\top\mathbf{H}\mathbf{\Theta} = \mathbf{H}$ and $\|\widehat{\mathbf{k}}_j\|_{\mathbf{H}} = c_K$, and let*

$$\rho_n^-(j) \propto e^{b_{nj}}, \qquad \mathbf{z}_n \mid J_n = j, \mathcal{F}_{n-1} \sim \mathcal{N}(\mathbf{P}_q\mathbf{\Theta}^{n-j}\widehat{\mathbf{k}}_j, \mathbf{D}), \quad j < n. \tag{12}$$

Then the strict-past Softmax row computed from $\ell_{nj} = \langle \mathbf{q}_n, \mathbf{\Theta}^{n-j}\widehat{\mathbf{k}}_j\rangle + b_{nj}$ is exactly the posterior of J_n. At the first position, set $\alpha_{11} = 1$ and $\mathbf{t}_1 = \mathbf{v}_1$. Then

$$\mathrm{Attention}(\mathbf{z}_{\le n}) = \mathbb{E}[\mathbf{t}_n \mid \mathcal{F}_n] = (1 - \eta_n)\sum_{j<n}\pi_{nj}\mathbf{v}_j + \eta_n\mathbf{v}_n. \tag{13}$$

Proof: Appendix B.

Remark 2.6 (MSE optimality: posterior mean as conditional expectation). The posterior mean in Theorem 2.5 is the conditional expectation $\mathbb{E}[\mathbf{t}_n \mid \mathcal{F}_n]$, and the conditional expectation is the classical Bayes estimator under squared loss: for every square-integrable $\mathcal{F}_n$-measurable estimator $\widehat{\mathbf{t}}(\mathbf{z}_{\le n})$ of the value readout,

$$\mathbb{E}\big\|\mathbf{t}_n - \mathbb{E}[\mathbf{t}_n \mid \mathcal{F}_n]\big\|^2 \;\le\; \mathbb{E}\big\|\mathbf{t}_n - \widehat{\mathbf{t}}(\mathbf{z}_{\le n})\big\|^2,$$

with equality only at the conditional expectation itself; equivalently, the error is orthogonal to every $\mathcal{F}_n$-measurable readout, $\mathbb{E}\big[\big(\mathbf{t}_n - \mathbb{E}[\mathbf{t}_n \mid \mathcal{F}_n]\big)^\top\mathbf{g}(\mathbf{z}_{\le n})\big] = 0$. The softmax attention operator is therefore not one plausible aggregation of the values but the unique mean-square-optimal readout of $\mathbf{t}_n$ under the RLA generative model—an optimality statement for the *exact* softmax attention formulation. The same statement holds verbatim for SLA under Assumption 3.1.

In Theorem 2.5, we let self weight $\eta_n > 0$ to be a determined value. In the standard softmax attention, η_n is uniquely defined by extrapolating (7) to $j = n$. However, in practice, η_n can be parametrized by a learnable scalar function (called *update gate* in this paper) with value in $[0, 1]$. Meanwhile, although we set $\rho_n^-(j) \propto e^{b_{nj}}$, our derivation does not make any assumption that ρ_n should be independent from the past data. In fact, it can also be by a data-dependent learnable function.

Multi-head output. Write the single-head function for head h as $\mathrm{SHA}(\mathbf{z}_n, \mathbf{z}_{<n}; h) = \sum_{j\le n}\alpha_{njh}\mathbf{v}_{jh}$. The multi-head output is simply $\mathbf{o}_n = \mathbf{W}_O\,\mathrm{concat}_{h=1}^H\,\mathrm{SHA}(\mathbf{z}_n, \mathbf{z}_{<n}; h)$, where $\mathbf{W}_O \in \mathbb{R}^{d\times\sum_h d_h}$. This is the deterministic linear output projection applied after computing each head.

3 Spherical Linear Attention from RLA

We work throughout in the small-noise regime $\nu \to 0^+$ of (11): $\mathbf{H} = \mathbf{P}_q^\top\mathbf{D}^{-1}\mathbf{P}_q = \mathbf{I}_{d_h}$, and keys are Euclidean-normalized, $\widehat{\mathbf{k}}_j = \mathbf{P}_k^\top\mathbf{z}_j/\|\mathbf{P}_k^\top\mathbf{z}_j\|$; the presentation below is the uniform small-noise

limits of densities defined at $\nu > 0$. The latent-parent skeleton of Assumption 2.1 supports a second branch: replace only the distribution family of (A2), keeping the mean–parent relation and every other component unchanged. Motivated by the well-known QK-Norm and the derivation of key normalization in Section 2.1, we generalize the conditional Gaussian assumption in Assumption 2.1 to the linear directional family as Assumption 3.1:

Assumption 3.1 (Spherical linear-attention generative model: linear modification of Assumption 2.1)**.** Keep components (A1), (A3), and (A4) of Assumption 2.1, the latent parent J_n, and its local-latent structure, in the small-noise regime declared above. Replace (A2) as follows.

(A2) Linear directional family. Specify the conditional law of $\mathbf{z}_n$ in the same two parts as (A2) of Assumption 2.1, changing only the distribution family:

(a) *Distribution family*: the linear directional family with parent-dependent mean $\boldsymbol{\mu}_{nj}$ and global parameters $(\boldsymbol{\Sigma}, \mathbf{D})$,

$$p(\mathbf{z}_n \mid J_n = j, \mathcal{F}_{n-1}) = \left(1 + \beta_0 \frac{\mathbf{z}_n^\top \boldsymbol{\Sigma} \mu_{nj}}{\sqrt{\mathbf{z}_n^\top \boldsymbol{\Sigma} \mathbf{z}_n}}\right) \mathcal{N}(\mathbf{z}_n; \mathbf{0}, \mathbf{D}), \tag{14}$$

with the measurement metric $\boldsymbol{\Sigma} = \mathbf{D}^{-1}\mathbf{P}_q\mathbf{P}_q^\top\mathbf{D}^{-1}$ and the dimension constant $\beta_0 := d_h\,\Gamma(d_h/2)/\big(\sqrt{2}\,\Gamma\big((d_h+1)/2\big)\big)$, both shared by all (n, j). On the parameter domain $\boldsymbol{\mu}_{nj} \in \mathrm{range}(\mathbf{P}_q)$ with $\beta_0^2\,\boldsymbol{\mu}_{nj}^\top\boldsymbol{\Sigma}\boldsymbol{\mu}_{nj} \le 1$, this is a legal density with mean $\boldsymbol{\mu}_{nj}$ and uncentered second moment $\mathbf{D}$ (Appendix C.1);

(b) *Mean–parent relation*: the mean is the normalized parent key embedded into the observation space with transition $\boldsymbol{\Theta}^{n-j}$ and rescaled onto the parameter-domain boundary, in two steps,

$$\widetilde{\boldsymbol{\mu}}_{nj} = \mathbf{P}_q\boldsymbol{\Theta}^{n-j}\frac{\mathbf{P}_k^\top\mathbf{z}_j}{\|\mathbf{P}_k^\top\mathbf{z}_j\|}, \qquad \boldsymbol{\mu}_{nj} = \frac{\widetilde{\boldsymbol{\mu}}_{nj}}{\beta_0\,\big\|\mathbf{P}_q^\top\mathbf{D}^{-1}\widetilde{\boldsymbol{\mu}}_{nj}\big\|}, \tag{15}$$

where $\boldsymbol{\mu}_{nj}$ is normalized to satisfy the the normalized-key of the low-rank prediction in Assumption 2.1: $\beta_0^2\boldsymbol{\mu}_{nj}^\top\boldsymbol{\Sigma}\boldsymbol{\mu}_{nj} = 1$, since $\|\mathbf{P}_q^\top\mathbf{D}^{-1}\widetilde{\boldsymbol{\mu}}_{nj}\| = 1$ in the small-noise regime.

Then under the same proof procedure for Theorem 2.3, we obtain Theorem 3.2 for Spherical linear attention (SLA).

Theorem 3.2 (SLA is the exact posterior mean under the linear directional family)**.** *Under Assumption 3.1, adopt the reparameterization*

$$\mathbf{W}_q = \mathbf{P}_q^\top\mathbf{D}^{-1}, \qquad \mathbf{W}_k = \mathbf{P}_k^\top, \qquad \mathbf{W}_v = \mathbf{P}_v^\top, \tag{16}$$

write $\mathbf{q}_n = \mathbf{W}_q\mathbf{z}_n$, $\mathbf{k}_j = \mathbf{W}_k\mathbf{z}_j$, $\mathbf{v}_j = \mathbf{W}_v\mathbf{z}_j$, *and define the unit directions* $\widehat{\mathbf{q}}_n = \mathbf{q}_n/\|\mathbf{q}_n\|$, $\widehat{\mathbf{k}}_j = \mathbf{k}_j/\|\mathbf{k}_j\|$ *and the spherical linear-attention score*

$$s_{nj} = \langle\widehat{\mathbf{q}}_n, \boldsymbol{\Theta}^{n-j}\widehat{\mathbf{k}}_j\rangle \in [-1, 1]. \tag{17}$$

Then the exact parent posterior over the strict past is

$$\pi_{nj} = \mathbb{P}(J_n = j \mid \mathcal{F}_n) = \frac{\rho_n^-(j)\big(1 + s_{nj}\big)}{\sum_{j'<n}\rho_n^-(j')\big(1 + s_{nj'}\big)}, \qquad j < n, \tag{18}$$

with unnormalized weights in $[0, 2\rho_n^-(j)]$. *And then the SLA operator can be defined as the exact posterior mean:*

$$\mathrm{SLA}(\mathbf{z}_{\le n}) := \mathbb{E}[\mathbf{t}_n \mid \mathcal{F}_n] = (1 - \eta_n)\sum_{j<n}\pi_{nj}\mathbf{v}_j + \eta_n\mathbf{v}_n. \tag{19}$$

Algorithmically, $(\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \boldsymbol{\Theta}, \lambda)$ *are free parameters optimized directly, and the identities* (16) *serve as interpretation rather than constraint.*

Proof: Appendix B.

Relation to softmax attention. Beyond enforcing QK-Norm in the generative model, the only difference between the SLA posterior weight (18) and the softmax weight (7) is the replacement of $\exp(s_{nj})$ by $1 + s_{nj}$: the SLA attention weight is exactly the first-order truncation of the softmax attention weight, since $\exp(c) = 1 + c + O(c^2)$.

Learnable instantiations: forget gate and update gate. Theorem 3.2 leaves two quantities to be determined: the lag prior ρ_n^- and the self weight, denoted η_n in (19). We instantiate ρ_n^- as a learnable distribution driven by the token at each position—the *forget gate*—and η_n as a learnable scalar driven by the current data $\mathbf{z}_n$—the *update gate*. Both taking values in $(0, 1]$ through the parameterization $g(x) = \exp(-\operatorname{softplus}(x))$. The forget gate realizes the lag prior as a product decay $\rho_n^-(j) \propto \prod_{i=j}^{n-1} \gamma_i$ in which every factor is $\mathcal{F}_{n-1}$-measurable, so no current observation leaks into the prior. Table 1 gives the precise definitions: each gate is a linear score of a concrete observation plus a learnable bias, initialized at $\mathbf{u} = \mathbf{0}$, $b = 0$ so that $g = \frac{1}{2}$. Notably, both gates are model-agnostic: the same instantiation also improves softmax attention—where η_n is the update gate of Section 2.1 and ρ_n^- acts as the additive log-prior b_{nj} of Theorem 2.5; while in linear attention, the data-dependent decay coincides with the gate convention of gated linear attention. Notably, to the best of our knowledge, the self weight η_n has not been used for training transformer with softmax attention.

Gate	Definition	Data-dependent score	Bias
Forget gate γ_n	$\gamma_n = \exp\big(-\operatorname{softplus}(\mathbf{u}_\gamma^\top \mathbf{z}_n + b_\gamma)\big)$	$\mathbf{u}_\gamma^\top \mathbf{z}_n$ (token $\mathbf{z}_n$)	b_γ
Update gate η_n	$\eta_n = \exp\big(-\operatorname{softplus}(\mathbf{u}_a^\top \mathbf{z}_n + b_a)\big)$	$\mathbf{u}_a^\top \mathbf{z}_n$ (current token $\mathbf{z}_n$)	b_a

Table 1: Precise definitions of the two learnable gates. Both satisfy $g(x) = \exp(-\operatorname{softplus}(x)) = \operatorname{sigmoid}(-x) \in (0, 1)$ and are initialized at $\mathbf{u} = \mathbf{0}$, $b = 0$, i.e. $g = \frac{1}{2}$.

Recurrent inference. With the gates of Table 1, the SLA operator retains an exact constant-memory recurrent form. Maintaining the state $(\mathbf{M}_n, \mathbf{s}_n^v, \mathbf{s}_n^k, z_n^{\mathrm{sum}})$, initialized at zero, the state is updated by

$$\begin{aligned} \mathbf{M}_{n+1} &= \gamma_n \, \boldsymbol{\Theta}\big(\mathbf{M}_n + \widehat{\mathbf{k}}_n \mathbf{v}_n^\top\big), \qquad \mathbf{s}_{n+1}^v = \gamma_n \big(\mathbf{s}_n^v + \mathbf{v}_n\big), \\ \mathbf{s}_{n+1}^k &= \gamma_n \boldsymbol{\Theta}\big(\mathbf{s}_n^k + \widehat{\mathbf{k}}_n\big), \qquad z_{n+1}^{\mathrm{sum}} = \gamma_n \big(z_n^{\mathrm{sum}} + 1\big), \end{aligned} \tag{20}$$

and the self-inclusive output collapses to the single fraction

$$\mathrm{SLA}(\mathbf{z}_{\le n}) = \frac{\eta_n \big(1 + c_{nn}\big)\mathbf{v}_n + \mathbf{s}_n^v + \mathbf{M}_n^\top \widehat{\mathbf{q}}_n}{\eta_n \big(1 + c_{nn}\big) + z_n^{\mathrm{sum}} + \big\langle \widehat{\mathbf{q}}_n, \mathbf{s}_n^k \big\rangle}. \tag{21}$$

Thus the model admits both the parallel form (18) and a constant-memory recurrent form; with $\boldsymbol{\Theta} = \mathbf{I}$, it is a gated linear attention whose every quantity—the decay, the self weight, the score, and the normalizer—retains its exact posterior meaning, computed with an $O(d_h^2)$ state and an $O(1)$ update per token. Signed Retention is recovered as the degenerate special case obtained by dropping the shift and the normalizer, and the weights form a nonnegative, exactly relative-position kernel linear attention (Appendix C.1).

Remark 3.3 (Almost-sure positivity of the normalizer). The normalizer $\sum_{j<n} \rho_n^-(j)(1 + s_{nj}) = 1 + \big\langle \widehat{\mathbf{q}}_n, \sum_{j<n} \rho_n^-(j) \, \boldsymbol{\Theta}^{n-j} \widehat{\mathbf{k}}_j \big\rangle$ is nonnegative, and it vanishes only if all $s_{nj} = -1$, which is unlikely happen. In practice, we can add some small tolerance $\epsilon = 10^{-8}$ on the normalizer to prevent zero normalizer.

4 Experiments

Setup. We pretrain SLA, a 340M-parameter SLA language model, on 15B tokens of SlimPajama (Soboleva et al., 2023) (28611 steps of 524288 tokens, learning rate 10^{-2} cosine-annealed to 10^{-3}). We evaluate zero-shot downstream performance with the lm-eval harness (Gao et al., 2021) under the GLA evaluation protocol (Yang et al., 2024a), and compare against Transformer++ (Touvron et al., 2023), RetNet (Sun et al., 2023), Mamba (Gu & Dao, 2023), and GLA (Yang et al., 2024a) at the same 340M / 15B-token scale; baseline numbers are taken from the GLA paper.

Table 2: General downstream tasks and language modeling (340M params / 15B SlimPajama tokens, zero-shot). Numbers for Transformer++/RetNet/Mamba/GLA are taken from the GLA paper (Table 2). SLA is evaluated with the GLA-aligned protocol (lm-eval zero-shot; Wiki./LMB. report ppl; Hella./ARC-c report acc_{norm}).

Model	Wiki. ppl ↓	LMB. ppl ↓	LMB. acc ↑	PIQA acc ↑	Hella. acc_{norm} ↑	Wino. acc ↑	ARC-e acc ↑	ARC-c acc_{norm} ↑	Avg. ↑
Transformer++	28.39	42.69	31.0	63.3	34.0	50.4	44.5	24.2	41.2
RetNet	32.33	49.19	28.6	63.5	33.5	52.5	44.5	23.4	41.0
Mamba	28.39	39.66	30.6	65.0	35.4	50.1	46.3	23.6	41.8
GLA	28.65	43.35	30.3	64.8	34.5	51.4	45.1	22.7	41.5
SLA	**26.05**	**41.15**	30.6	**66.0**	**37.7**	48.8	**47.1**	24.1	**42.4**

Table 3: Recall-intensive tasks (340M params / 15B tokens, zero-shot, inputs truncated to 2K tokens). Numbers for Transformer++/RetNet/Mamba/GLA are taken from the GLA paper (Table 3). Avg. is the three-task mean.

Model	FDA ↑	SWDE ↑	SQuAD ↑	Avg. ↑
Transformer++	21.4	42.2	22.1	28.6
RetNet	2.9	13.3	27.6	14.6
Mamba	2.1	12.4	23.0	12.5
GLA	8.1	18.6	27.2	18.0
SLA	6.8	**22.9**	**30.3**	**20.0**

Remarks. (1) SLA achieves the best 6-task commonsense average (42.4) among subquadratic models, surpassing GLA (41.5) and Mamba (41.8), with the best Wiki./LMB. perplexity and the best PIQA/Hella./ARC-e accuracy among all subquadratic baselines. Its Wino. score (48.8) trails GLA (51.4); at the 340M scale all models sit near the random baseline ($\approx$50) on this task, so we attribute the gap to task noise. (2) On recall-intensive tasks, SLA outperforms GLA on SWDE (22.9 vs. 18.6) and SQuAD (30.3 vs. 27.2), but lags on FDA (6.8 vs. 8.1, and 21.4 for Transformer++), consistent with the known weakness of kernel-smoothed linear attention on exact associative recall.

5 Conclusion

We developed a random-lag autoregressive model from the conditional distribution of token sequences. Under locality, $\mathbf{H}$-orthogonality of the transition, and equal key energy, standard attention is an exact posterior mean of this model; RoPE, relative positional biases, and key-energy constraints receive interpretations in terms of transition geometry, lag priors, and emission noise, respectively. In the small-noise regime, key normalization under the metric $\mathbf{H}$ converges to Euclidean key normalization and the $\mathbf{H}$-orthogonal transition converges to a Euclidean block rotation, the canonical instance of which is RoPE. Replacing the Gaussian emission family by the linear directional family turns the same posterior mean into spherical linear attention, with nonnegative normalized weights, exact relative positions, and a constant-memory recurrence, and recovers signed Retention as a degenerate special case. Future work will complete the remaining recall-intensive evaluations and ablations, probe history-dependent lag priors as a generative selectivity mechanism, and evaluate larger-scale behavior and computational efficiency.

AI USE STATEMENT

Generative AI tools assisted with drafting, English translation, restructuring, and the exposition of mathematical derivations, and were used to numerically verify algebraic identities. All experiments were designed and run by the authors; AI tools did not generate, alter, or select any experimental results. The authors take full responsibility for all content.

REPRODUCIBILITY STATEMENT

The main text specifies the generative assumptions, posterior scores, the finite-noise key normalization under the metric $\mathbf{H}$ and its small-noise limit, and the spherical linear-attention recurrences. The appendices provide the supporting proofs. Section 4 reports the model and training configuration, the training and validation losses, and zero-shot downstream evaluations under the GLA protocol. Baseline results are reported from Yang et al. (2024a); our models use the same training corpus, token budget, tokenizer, context length, and evaluation harness, so that the numbers are directly comparable.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, 2014.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, et al. Rethinking attention with Performers. In *International Conference on Learning Representations*, 2021.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, 2024.

Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B*, 55(4):757–779, 1993.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for Transformers. In *Findings of EMNLP*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive Transformers with linear attention. In *ICML*, 2020.

Nhu D. Le, R. Douglas Martin, and Adrian E. Raftery. Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association*, 91(436):1504–1515, 1996.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, et al. Eagle and Finch: RWKV with matrix-valued states and dynamic recurrence. In *Conference on Language Modeling (COLM)*, 2024.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosFormer: Rethinking Softmax in attention. In *International Conference on Learning Representations*, 2022.

Zhen Qin, Songlin Yang, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. HGRN2: Gated linear RNNs with state expansion. In *Conference on Language Modeling (COLM)*, 2024.

Adrian E. Raftery. A model for high-order Markov chains. *Journal of the Royal Statistical Society: Series B*, 47(3):528–541, 1985.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, et al. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear Transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, 2021.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Yutao Sun, Li Dong, Shaohan Huang, et al. Retentive network: A successor to Transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for Transformer's attention via the lens of kernel. In *EMNLP*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the Transformer architecture. In *International Conference on Machine Learning*, 2020.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention Transformers with hardware-efficient training. In *International Conference on Machine Learning*, 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear Transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, 2024b.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Hao Bai, Yuexiang Zhai, Benjamin D. Haeffele, and Yi Ma. White-box Transformers via sparse rate reduction: Compression is all there is? *arXiv preprint arXiv:2311.13110*, 2023.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R. Steeves, Joel Hestness, and Salem Haykal. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. *https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpjama*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *NeurIPS*, 2019.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *https://github.com/EleutherAI/lm-evaluation-harness*, 2021.

A Autoregressive Background and Joint-Distribution Details

Observed information and notation. Write $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$, $\mathbf{z}_n \in \mathbb{R}^d$, $\mathbf{z}_{<n} = (\mathbf{z}_1, \dots, \mathbf{z}_{n-1})$, $\mathbf{z}_{\le n} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$. On a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$, define $\mathcal{F}_0 = \{\varnothing, \Omega\}$, $\mathcal{F}_n = \sigma(\mathbf{z}_1, \dots, \mathbf{z}_n)$. Here $\sigma(\cdot)$ denotes the σ-algebra generated by the random variables: $\mathcal{F}_{n-1}$ records all observations before the current position is generated, and $\mathcal{F}_n$ additionally includes the current observation. A quantity is $\mathcal{F}_n$-measurable if it is determined by observations up to position n. The parent position $J_n < n$ is sampled given $\mathcal{F}_{n-1}$, and its posterior is conditioned on $\mathcal{F}_n$; value conditional means are also computed under $\mathcal{F}_n$ and can therefore use deterministic functions of the current observation.

Throughout, lowercase bold symbols denote vectors (e.g., $\mathbf{x}, \mathbf{g}$) and uppercase bold symbols denote matrices (e.g., $\mathbf{W}, \mathbf{H}$); vector components, scalar random variables, and dimensions use ordinary type (e.g., g_i, G_i, d). Calligraphic letters denote sets, information families, or families of latent variables; random vectors also use lowercase bold symbols.

We write $\mathcal{N}(\mathbf{m}, \mathbf{C})$ for a Gaussian distribution with mean $\mathbf{m}$ and covariance $\mathbf{C}$, and $\varphi_{\mathbf{C}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{C})$ for its zero-mean density; $\mathbf{C} \succ 0$ denotes positive definiteness. $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$, $\|\mathbf{u}\|^2 = \mathbf{u}^\top \mathbf{u}$, and $\|\mathbf{u}\|^2_{\mathbf{H}} = \mathbf{u}^\top \mathbf{H} \mathbf{u}$ denote the inner product, squared Euclidean norm, and squared norm under the metric $\mathbf{H} \succ 0$, respectively. The operators $\mathrm{blkdiag}$ and $\odot$ denote block-diagonal concatenation and elementwise multiplication, respectively, and $\mathrm{Cat}(\rho)$ denotes a categorical distribution with mass function ρ. For a finite candidate set I, define $\mathrm{softmax}_{j \in I}\, s_j = \exp(s_j) / \sum_{i \in I} \exp(s_i)$. The scalar function $\sigma(x) = (1 + e^{-x})^{-1}$ denotes the sigmoid, and $\mathrm{logit}(p) = \log[p/(1-p)]$; its arguments distinguish the sigmoid from the operator generating a σ-algebra.

A.1 Autoregressive Coordinates and Posterior Factorization with Local Latent Variables

Definition A.1 (Gaussian autoregressive coordinates)**.** Let $\mathbf{z} \sim \mathcal{N}(0, \mathbf{\Lambda}^{-1})$ with $\mathbf{\Lambda} \succ 0$. For $n \ge 2$, define

$$\begin{aligned} \mathbf{B}_n &= \mathrm{Cov}(\mathbf{z}_n, \mathbf{z}_{<n}) \, \mathrm{Cov}(\mathbf{z}_{<n})^{-1}, \\ \mathbf{D}_n &= \mathrm{Var}(\mathbf{z}_n \mid \mathbf{z}_{<n}); \qquad \mathbf{B}_1 \mathbf{z}_{<1} := 0, \quad \mathbf{D}_1 = \mathrm{Var}(\mathbf{z}_1). \end{aligned} \tag{22}$$

Place $\mathbf{B}_n$ in the strictly block-lower-triangular matrix $\mathbf{B}$, in its block row n, and let $\mathbf{D}_{\mathrm{all}} = \mathrm{blkdiag}(\mathbf{D}_1, \dots, \mathbf{D}_N)$. We call $(\mathbf{B}, \mathbf{D}_{\mathrm{all}})$ the autoregressive coordinates of this Gaussian distribution—equivalently, the modified Cholesky (LDL) factors of the precision $\mathbf{\Lambda}$.

Define the fresh component (innovation) as the residual obtained by subtracting the historical conditional mean from the current state.

Lemma A.2 (Innovation form and determinant invariance)**.** *The blocks of $\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{B})\mathbf{z}$ are mutually independent, with $\boldsymbol{\varepsilon}_n \sim \mathcal{N}(0, \mathbf{D}_n)$. Moreover,*

$$\begin{aligned} p(\mathbf{z}) &= \prod_{n=1}^{N} \mathcal{N}(\mathbf{z}_n; \mathbf{B}_n \mathbf{z}_{<n}, \mathbf{D}_n), \\ \mathbf{\Lambda} &= (\mathbf{I} - \mathbf{B})^\top \mathbf{D}_{\mathrm{all}}^{-1} (\mathbf{I} - \mathbf{B}), \qquad \log |\mathbf{\Lambda}| = -\sum_{n=1}^{N} \log |\mathbf{D}_n|. \end{aligned} \tag{23}$$

For every positive-definite $\mathbf{\Lambda}$, the factorization exists and is unique under the unit block-lower-triangular and positive-definite block-diagonal constraints above.

Proof. Gaussian conditional densities give $\mathbb{E}[\mathbf{z}_n \mid \mathbf{z}_{<n}] = \mathbf{B}_n \mathbf{z}_{<n}$ and conditional covariance $\mathbf{D}_n$. By the definition of the regression coefficients, $\mathrm{Cov}(\boldsymbol{\varepsilon}_n, \mathbf{z}_{<n}) = 0$; past innovations $\boldsymbol{\varepsilon}_j$ are linear functions of $\mathbf{z}_{<n}$, so the innovation blocks are uncorrelated. They are jointly Gaussian and therefore independent. Positive definiteness ensures that each conditional covariance is positive definite. Let $\mathbf{T} = \mathbf{I} - \mathbf{B}$, so $|\mathbf{T}| = 1$. Applying the change of variables $\boldsymbol{\varepsilon} = \mathbf{T}\mathbf{z}$, the exponent of the independent innovation density is $-\mathbf{z}^\top \mathbf{T}^\top \mathbf{D}_{\mathrm{all}}^{-1} \mathbf{T} \mathbf{z} / 2$, which gives the precision formula; taking determinants gives the final identity. Any factorization with the stated triangular structure gives the same sequential conditional means and covariances, which are unique for a nondegenerate Gaussian distribution; hence the factorization is unique. □

We call the nth conditional factor of the autoregressive joint density row n. A latent variable c_n is *local*—a per-token local latent variable in the standard Bayesian sense, as opposed to a global one—if it occurs only in row n; each row may depend on the observed history.

Lemma A.3 (A sufficient condition for posterior factorization with local latent variables)**.** *Let $c_n \in \mathcal{C}_n$ be discrete, let $\rho_n(\cdot \mid \mathbf{z}_{<n})$ be a probability mass function, and let $f_n(\cdot \mid \mathbf{z}_{<n}, c_n)$ be a probability density. Then*

$$p(\mathbf{z}, \mathbf{c}) = \prod_{n=1}^{N} \rho_n(c_n \mid \mathbf{z}_{<n}) f_n(\mathbf{z}_n \mid \mathbf{z}_{<n}, c_n) \tag{24}$$

defines a normalized joint distribution. Let $m_n(\mathbf{z}_{\le n}) := \sum_{c \in \mathcal{C}_n} \rho_n(c \mid \mathbf{z}_{<n}) f_n(\mathbf{z}_n \mid \mathbf{z}_{<n}, c)$. At observations where all $0 < m_n(\mathbf{z}_{\le n}) < \infty$, the posterior satisfies

$$\begin{aligned} p(\mathbf{c} \mid \mathbf{z}) &= \underbrace{\prod_{n=1}^{N} p(c_n \mid \mathbf{z}_{\le n})}_{\textit{row-wise factorization}}, \\ p(c_n = c \mid \mathbf{z}_{\le n}) &= \frac{\rho_n(c \mid \mathbf{z}_{<n}) f_n(\mathbf{z}_n \mid \mathbf{z}_{<n}, c)}{m_n(\mathbf{z}_{\le n})}. \end{aligned} \tag{25}$$

Proof. Integrating observations and summing latent variables backward from the final row gives a factor of 1 per row, proving normalization. With observations fixed, each factor contains only its own c_n, so summing over $\mathbf{c}$ gives $\prod_n m_n$; Bayes' rule yields the posterior factorization, and the same calculation for each observed prefix yields the row posteriors. □

Proposition A.4 (Innovation-coordinate description of the joint Gaussian)**.** *Under Assumption 2.1, additionally suppose that each $\rho_n(\ell)$ is independent of the observations. Then, for every $\boldsymbol{\ell}$ of positive probability,*

$$\begin{aligned} p(\mathbf{z} \mid \boldsymbol{\ell}) &= \mathcal{N}(0, \boldsymbol{\Lambda}(\boldsymbol{\ell})^{-1}), \\ \boldsymbol{\Lambda}(\boldsymbol{\ell}) &= (\mathbf{I} - \mathbf{B}(\boldsymbol{\ell}))^{\top} \operatorname{blkdiag}(\mathbf{D}, \dots, \mathbf{D})^{-1} (\mathbf{I} - \mathbf{B}(\boldsymbol{\ell})), \\ \log |\boldsymbol{\Lambda}(\boldsymbol{\ell})| &= -N \log |\mathbf{D}|, \end{aligned} \tag{26}$$

where $c_n = n - \ell_n < n$ is the parent position, $\mathbf{B}_{n,c_n} = \mathbf{A}_{\ell_n}$, and all other blocks vanish. Figure 2 shows the correspondence between random parents and nonzero matrix blocks. The precision graph has positions as nodes and nonzero off-diagonal precision blocks as edges, so its edges connect only parents and children.

Proof. When priors are observation-independent, $\prod_n \rho_n(\ell_n)$ factors out of the joint density; with $\boldsymbol{\ell}$ fixed, the product of emission factors is a normalized triangular Gaussian density, so $p(\boldsymbol{\ell}) = \prod_n \rho_n(\ell_n)$. Dividing and applying Lemma A.2 gives the result. More directly, each squared residual contains only $\mathbf{z}_n$ and $\mathbf{z}_{c_n}$; therefore,

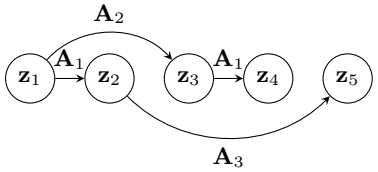
$$\boldsymbol{\Lambda}_{nn} = \mathbf{D}^{-1} + \sum_{t : c_t = n} \mathbf{A}_{\ell_t}^{\top} \mathbf{D}^{-1} \mathbf{A}_{\ell_t}, \qquad \boldsymbol{\Lambda}_{n,c_n} = -\mathbf{D}^{-1} \mathbf{A}_{\ell_n}.$$

Symmetry gives the transposed blocks, and all other off-diagonal blocks vanish. Each child has exactly one parent, so a single residual creates no cross term between distinct parents. If the undirected graph had a cycle, its latest node would have two parents, a contradiction; hence the graph is acyclic. The determinant identity follows from the unit block-lower-triangular transformation. □

Remark A.5 (Connection between noise covariance and the CRATE generative model)*.* Conditioned on subspace branch k, CRATE's low-dimensional Gaussian mixture model (Yu et al., 2023) is $\mathbf{x} = \mathbf{U}_k \mathbf{a} + \sigma \mathbf{w}$, where $\mathbf{a} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_k)$ and $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$ are independent. Therefore,

$$\operatorname{Cov}(\mathbf{x} \mid k) = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{U}_k^{\top} + \sigma^2 \mathbf{I}.$$

Taking $\mathbf{P}_q = \mathbf{U}_k \boldsymbol{\Sigma}_k^{1/2}$ and $\nu = \sigma^2$ gives our choice $\mathbf{D} = \mathbf{P}_q \mathbf{P}_q^{\top} + \nu \mathbf{I}$; when $\boldsymbol{\Sigma}_k = \mathbf{I}$, the low-rank term is the subspace projection $\mathbf{U}_k \mathbf{U}_k^{\top}$. Both constructions share a low-rank-plus-isotropic covariance structure. CRATE uses it for the token distribution conditioned on a subspace; we use it for innovation noise in a random-lag transition, thereby determining the query projection and the noise-induced metric $\mathbf{H}$.



$$\mathbf{B}(\boldsymbol{\ell}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \mathbf{A}_1 & 0 & 0 & 0 & 0 \\ \mathbf{A}_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{A}_1 & 0 & 0 \\ 0 & \mathbf{A}_3 & 0 & 0 & 0 \end{pmatrix}$$

$\boldsymbol{\ell} = (1, 2, 1, 3)$, $c = (1, 1, 3, 2)$.

$$c_n \longrightarrow n \quad \Longleftrightarrow \quad \mathbf{B}_{n,c_n} = \mathbf{A}_{n-c_n} \quad \Longleftrightarrow \quad \boldsymbol{\varepsilon}_n = \mathbf{z}_n - \mathbf{B}_{n,c_n}\mathbf{z}_{c_n}.$$

Figure 2: Correspondence between random-lag transitions and the full block-lower-triangular matrix ($N = 5$). Arrows point from parent positions to current states; matrix rows and columns correspond to child and parent positions, respectively. Each 0 denotes a $d \times d$ zero block. Row 1 is zero, and only the parent column can be nonzero in each subsequent row; stacking the rows yields the innovation coordinates $\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{B})\mathbf{z}$.

B Proofs of the Main Results

B.1 Proof of Result 2.3

Proof. Fix row n and candidate parent $j < n$. By locality (Lemma A.3) with $c_n = J_n$,

$$\pi_{nj} = \mathbb{P}(J_n = j \mid \mathcal{F}_n) = \frac{\rho_n^-(j)\,\mathcal{N}(\mathbf{z}_n; \boldsymbol{\mu}_{nj}, \mathbf{D})}{\sum_{j'<n} \rho_n^-(j')\,\mathcal{N}(\mathbf{z}_n; \boldsymbol{\mu}_{nj'}, \mathbf{D})}, \qquad \boldsymbol{\mu}_{nj} = \mathbf{P}_q \boldsymbol{\Theta}^{n-j}\mathbf{k}_j. \tag{27}$$

Step 1: expand the Gaussian quadratic form. Discarding all terms independent of j,

$$\begin{aligned} -\log \mathcal{N}(\mathbf{z}_n; \boldsymbol{\mu}_{nj}, \mathbf{D}) &= \tfrac{1}{2}(\mathbf{z}_n - \boldsymbol{\mu}_{nj})^\top \mathbf{D}^{-1}(\mathbf{z}_n - \boldsymbol{\mu}_{nj}) + \text{const} \\ &= \underbrace{\tfrac{1}{2}\mathbf{z}_n^\top \mathbf{D}^{-1}\mathbf{z}_n}_{\text{indep. of } j} - \underbrace{\mathbf{z}_n^\top \mathbf{D}^{-1}\mathbf{P}_q\boldsymbol{\Theta}^{n-j}\mathbf{k}_j}_{\text{cross term}} + \underbrace{\tfrac{1}{2}\boldsymbol{\mu}_{nj}^\top \mathbf{D}^{-1}\boldsymbol{\mu}_{nj}}_{\text{energy}} + \text{const}. \end{aligned} \tag{28}$$

The cross term is the query–key inner product: with $\mathbf{W}_q = \mathbf{P}_q^\top \mathbf{D}^{-1}$ and $\mathbf{q}_n = \mathbf{W}_q \mathbf{z}_n$,

$$\mathbf{z}_n^\top \mathbf{D}^{-1}\mathbf{P}_q\boldsymbol{\Theta}^{n-j}\mathbf{k}_j = \left(\mathbf{P}_q^\top \mathbf{D}^{-1}\mathbf{z}_n\right)^\top \boldsymbol{\Theta}^{n-j}\mathbf{k}_j = \langle \mathbf{q}_n, \boldsymbol{\Theta}^{n-j}\mathbf{k}_j \rangle. \tag{29}$$

The energy term is the key norm under the noise-induced metric $\mathbf{H} = \mathbf{P}_q^\top \mathbf{D}^{-1}\mathbf{P}_q$:

$$\boldsymbol{\mu}_{nj}^\top \mathbf{D}^{-1}\boldsymbol{\mu}_{nj} = \mathbf{k}_j^\top (\boldsymbol{\Theta}^{n-j})^\top \underbrace{\mathbf{P}_q^\top \mathbf{D}^{-1}\mathbf{P}_q}_{=\mathbf{H}} \boldsymbol{\Theta}^{n-j}\mathbf{k}_j = \|\boldsymbol{\Theta}^{n-j}\mathbf{k}_j\|_{\mathbf{H}}^2. \tag{30}$$

Step 2: read off the score. Substituting (29) and (30) into (28), all j-independent factors cancel between numerator and denominator of (27):

$$\begin{aligned} s_{nj} &= \log \rho_n^-(j) + \langle \mathbf{q}_n, \boldsymbol{\Theta}^{n-j}\mathbf{k}_j \rangle - \tfrac{1}{2}\|\boldsymbol{\Theta}^{n-j}\mathbf{k}_j\|_{\mathbf{H}}^2, \\ \pi_{nj} &= \mathrm{softmax}_{j<n}(s_{nj}), \end{aligned} \tag{31}$$

which is (6).

Step 3: isometry reduction. Assume $\boldsymbol{\Theta}^\top \mathbf{H}\boldsymbol{\Theta} = \mathbf{H}$. Then inductively $(\boldsymbol{\Theta}^\ell)^\top \mathbf{H}\boldsymbol{\Theta}^\ell = \mathbf{H}$ for every $\ell \geq 1$, so

$$\|\boldsymbol{\Theta}^{n-j}\mathbf{k}_j\|_{\mathbf{H}}^2 = \|\mathbf{k}_j\|_{\mathbf{H}}^2 = c_K^2 \qquad \text{for all } j < n, \tag{32}$$

and the energy term in (31) is a j-independent additive constant that cancels in the softmax:

$$s_{nj} = \log \rho_n^-(j) + \langle \mathbf{q}_n, \boldsymbol{\Theta}^{n-j}\mathbf{k}_j \rangle - \tfrac{1}{2}c_K^2, \tag{33}$$

which is (7).

□

B.2 Proof of the small-noise metric bound (11)

Proof. **Step 1: spectrum of H.** The matrices $\mathbf{G}$ and $(\mathbf{G}+\nu\mathbf{I})^{-1}$ commute, so $\mathbf{H} = \mathbf{G}(\mathbf{G}+\nu\mathbf{I})^{-1}$ is symmetric positive definite with the same eigenvectors as $\mathbf{G}$. Each eigenvalue $g_i \geq g_{\min} > 0$ of $\mathbf{G}$ contributes

$$\lambda_i(\mathbf{H}) = \frac{g_i}{g_i+\nu} \in \Big[\frac{g_{\min}}{g_{\min}+\nu},\, 1\Big) = \Big[(1+\nu/g_{\min})^{-1},\, 1\Big). \tag{34}$$

Step 2: Loewner bounds and limit. Bounding each eigenvalue in (34) gives the Loewner bounds

$$(1+\nu/g_{\min})^{-1}\mathbf{I} \preceq \mathbf{H} \preceq \mathbf{I}. \tag{35}$$

Moreover,

$$\lambda_i(\mathbf{I}-\mathbf{H}) = \frac{\nu}{g_i+\nu} \leq \frac{\nu}{g_{\min}} \longrightarrow 0 \qquad (\nu \to 0^+), \tag{36}$$

so $\mathbf{H} \to \mathbf{I}_{d_h}$ in spectral norm. Equations (35) and (36) give (11). □

B.3 Proof of Result 2.5

Proof. **Step 1: strict-past posterior.** The model (12) is the RLA of Assumption 2.1 with $\mathbf{k}_j$ replaced by the $\mathbf{H}$-normalized key $\widehat{\mathbf{k}}_j$. Result 2.3 therefore applies and gives, for $j < n$,

$$s_{nj} = \log \rho_n^-(j) + \langle \mathbf{q}_n, \mathbf{\Theta}^{n-j}\widehat{\mathbf{k}}_j\rangle - \tfrac{1}{2}\|\mathbf{\Theta}^{n-j}\widehat{\mathbf{k}}_j\|_{\mathbf{H}}^2. \tag{37}$$

Step 2: cancellation of the energy. By $\mathbf{\Theta}^\top\mathbf{H}\mathbf{\Theta} = \mathbf{H}$ and $\|\widehat{\mathbf{k}}_j\|_{\mathbf{H}} = c_K$,

$$\|\mathbf{\Theta}^{n-j}\widehat{\mathbf{k}}_j\|_{\mathbf{H}}^2 = \|\widehat{\mathbf{k}}_j\|_{\mathbf{H}}^2 = c_K^2 \qquad \text{for all } j < n. \tag{38}$$

The energy is a common additive constant and cancels in the softmax:

$$\begin{aligned}\pi_{nj} &= \mathrm{softmax}_{j<n}\Big(\langle \mathbf{q}_n, \mathbf{\Theta}^{n-j}\widehat{\mathbf{k}}_j\rangle + \log \rho_n^-(j)\Big) \\ &= \mathrm{softmax}_{j<n}(\ell_{nj}),\end{aligned} \tag{39}$$

where the second line uses $\rho_n^-(j) \propto e^{b_{nj}}$: multiplicative constants on the prior are absorbed by the softmax normalizer.

Step 3: self-weight extension. Extend the same score to $j = n$ and define

$$\alpha_{nj} = \frac{\exp(\ell_{nj})}{\sum_{j'\leq n}\exp(\ell_{nj'})} \quad (j \leq n), \qquad \eta_n := \alpha_{nn}. \tag{40}$$

For $j < n$, both π_{nj} in (39) and α_{nj} in (40) are proportional to the same $\exp(\ell_{nj})$, hence

$$\mathbb{P}(J_n = j \mid \mathcal{F}_n) = \pi_{nj} = \frac{\alpha_{nj}}{\sum_{j'<n}\alpha_{nj'}} = \frac{\alpha_{nj}}{1-\alpha_{nn}}. \tag{41}$$

Step 4: posterior mean. The values $\mathbf{v}_j$ ($j \leq n$) and the self weight η_n are $\mathcal{F}_n$-measurable, so they move outside the conditional expectation:

$$\begin{aligned}\mathbb{E}[\mathbf{t}_n \mid \mathcal{F}_n] &= (1-\eta_n)\,\mathbb{E}[\mathbf{v}_{J_n} \mid \mathcal{F}_n] + \eta_n\mathbf{v}_n \\ &= (1-\eta_n)\sum_{j<n}\pi_{nj}\mathbf{v}_j + \eta_n\mathbf{v}_n,\end{aligned} \tag{42}$$

with $\alpha_{11} = 1$ and $\mathbf{t}_1 = \mathbf{v}_1$ at the first position. This is (13). □

C Proof of the RoPE Parameterization

Proof. **Step 1: from the energy identity to H-orthogonality.** Let $\mathbf{A} := \mathbf{\Theta}^\top\mathbf{H}\mathbf{\Theta} - \mathbf{H}$, which is symmetric. The adjacent-lag identity $\|\mathbf{\Theta}^2\mathbf{u}\|_{\mathbf{H}}^2 = \|\mathbf{\Theta}\mathbf{u}\|_{\mathbf{H}}^2$ for all $\mathbf{u}$ gives

$$\mathbf{u}^\top\mathbf{\Theta}^\top\mathbf{A}\mathbf{\Theta}\mathbf{u} = 0 \quad \forall\mathbf{u} \quad\Longrightarrow\quad \mathbf{\Theta}^\top\mathbf{A}\mathbf{\Theta} = \mathbf{0} \quad\Longrightarrow\quad \mathbf{A} = \mathbf{0}, \tag{43}$$

where the last implication left-multiplies by $(\mathbf{\Theta}^\top)^{-1}$ and right-multiplies by $\mathbf{\Theta}^{-1}$, using invertibility of $\mathbf{\Theta}$. Hence

$$\mathbf{\Theta}^\top \mathbf{H} \mathbf{\Theta} = \mathbf{H}. \tag{44}$$

Step 2: orthogonalization. Define $\mathbf{O} := \mathbf{H}^{1/2} \mathbf{\Theta} \mathbf{H}^{-1/2}$. Then (44) gives

$$\mathbf{O}^\top \mathbf{O} = \mathbf{H}^{-1/2} \mathbf{\Theta}^\top \mathbf{H} \mathbf{\Theta} \mathbf{H}^{-1/2} = \mathbf{H}^{-1/2} \mathbf{H} \mathbf{H}^{-1/2} = \mathbf{I}_{d_h}, \tag{45}$$

so $\mathbf{O}$ is orthogonal; and

$$\det \mathbf{O} = \det \mathbf{H}^{1/2} \cdot \det \mathbf{\Theta} \cdot \det \mathbf{H}^{-1/2} = \det \mathbf{\Theta} > 0 \quad \Longrightarrow \quad \det \mathbf{O} = 1. \tag{46}$$

Step 3: rotation-block decomposition. For a real orthogonal matrix:

- nonreal eigenvalues occur in conjugate pairs $e^{\pm \mathrm{i}\theta}$ on the unit circle, and their real two-dimensional invariant subspaces admit orthonormal bases in which the restriction of $\mathbf{O}$ is $R(\theta)$;
- the remaining eigenvalues are $+1$ and -1: $\det \mathbf{O} = 1$ forces the multiplicity of -1 to be even, and the even total dimension $d_h = 2r$ forces the multiplicity of $+1$ to be even; pairs of equal real eigenvalues act as $R(\pi)$ (two -1's) or $R(0)$ (two $+1$'s).

Concatenating these orthonormal bases into $\mathbf{S}$ gives

$$\mathbf{S}^\top \mathbf{O} \mathbf{S} = \mathbf{R} = \mathrm{blkdiag}\big(R(\theta_1), \ldots, R(\theta_r)\big), \qquad R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{47}$$

Solving (47) for $\mathbf{\Theta} = \mathbf{H}^{-1/2} \mathbf{O} \mathbf{H}^{1/2}$ with $\mathbf{O} = \mathbf{S}\mathbf{R}\mathbf{S}^\top$ yields

$$\mathbf{\Theta} = \mathbf{H}^{-1/2} \mathbf{S} \mathbf{R} \mathbf{S}^\top \mathbf{H}^{1/2}, \tag{48}$$

which is (8).

Step 4: converse. For arbitrary angles $\theta_1, \ldots, \theta_r$, the matrix (48) satisfies, by (45) with $\mathbf{O} = \mathbf{S}\mathbf{R}\mathbf{S}^\top$ orthogonal,

$$\mathbf{\Theta}^\top \mathbf{H} \mathbf{\Theta} = \mathbf{H}, \qquad \det \mathbf{\Theta} = \det \mathbf{R} = \prod_{i=1}^{r} \big(\cos^2 \theta_i + \sin^2 \theta_i\big) = 1, \tag{49}$$

and iterating (44) gives $\|\mathbf{\Theta}^\ell \mathbf{u}\|_{\mathbf{H}}^2 = \|\mathbf{u}\|_{\mathbf{H}}^2$ for every $\ell \geq 1$. So the rotation frequencies are free parameters, fixed in advance or learned modulo 2π.

Step 5: explicit construction of $\mathbf{D}$ and $\mathbf{H}$ from the query factor. For any full-column-rank $\mathbf{P}_q$ in Assumption 2.1, let $\mathbf{G} := \mathbf{P}_q^\top \mathbf{P}_q$, choose a positive noise scale ν, and define

$$\boxed{\mathbf{D}(\mathbf{P}_q) = \mathbf{P}_q \mathbf{P}_q^\top + \nu \mathbf{I}_d, \qquad \mathbf{H}(\mathbf{P}_q) = \mathbf{G}(\mathbf{G} + \nu \mathbf{I}_{d_h})^{-1}.} \tag{50}$$

Taking $\nu = 1$ makes $\mathbf{P}_q$ fully specify a pair $\mathbf{D}, \mathbf{H}$. They satisfy the previously defined relations because

$$\begin{aligned} \mathbf{D}\mathbf{P}_q = \mathbf{P}_q(\mathbf{G} + \nu \mathbf{I}), \qquad \mathbf{D}^{-1}\mathbf{P}_q = \mathbf{P}_q(\mathbf{G} + \nu \mathbf{I})^{-1}, \\ \mathbf{P}_q^\top \mathbf{D}^{-1} \mathbf{P}_q = \mathbf{G}(\mathbf{G} + \nu \mathbf{I})^{-1} = \mathbf{H}. \end{aligned} \tag{51}$$

Positive definiteness $\mathbf{D} \succ 0$ follows directly from $\nu > 0$. If $s_i > 0$ are the singular values of $\mathbf{P}_q$, then the eigenvalues of $\mathbf{H}$ are

$$\lambda_i(\mathbf{H}) = \frac{s_i^2}{s_i^2 + \nu} \in (0, 1). \tag{52}$$

Thus the construction satisfies $0 \prec \mathbf{H} \prec \mathbf{I}$, with the metric in each direction determined by the singular values of $\mathbf{P}_q$. Symmetry of $\mathbf{H}$ also follows from commutativity of $\mathbf{G}$ and $(\mathbf{G} + \nu \mathbf{I})^{-1}$.

Step 6: compatibility with RoPE through parameterization. Given $\mathbf{H}(\mathbf{P}_q)$, choose the free rotation parameters obtained above and define

$$\mathbf{\Theta}(\mathbf{P}_q) = \mathbf{H}(\mathbf{P}_q)^{-1/2} \mathbf{S} \mathbf{R} \mathbf{S}^\top \mathbf{H}(\mathbf{P}_q)^{1/2}. \tag{53}$$

Direct substitution gives $\mathbf{\Theta}^\top \mathbf{H} \mathbf{\Theta} = \mathbf{H}$, and the inner-product representation is given in (54) below. This transition satisfies isometry for any full-column-rank $\mathbf{P}_q$; one may also take $\mathbf{S} = \mathbf{I}$. If instead the Euclidean transition $\mathbf{\Theta} = \mathbf{R}$ is fixed in the original coordinates, one additionally requires $\mathbf{HR} = \mathbf{RH}$. Thus, when only full column rank of the query factor is assumed, the transition parameterization should vary with the induced metric as above. □

The rotary representation above uses an energy identity on the entire head space, invertibility of the transition, and a positive determinant. Invertibility allows $\mathbf{H}$-orthogonality to follow from equal energies at adjacent lags, while a positive determinant ensures decomposition of the orthogonal representation into rotation blocks.

Let $\widetilde{\mathbf{q}}_n^* = \mathbf{S}^\top \mathbf{H}^{-1/2} \mathbf{q}_n$, $\widetilde{\mathbf{k}}_j^* = \mathbf{S}^\top \mathbf{H}^{1/2} \mathbf{k}_j$. Then

$$\langle \mathbf{q}_n, \mathbf{\Theta}^{n-j} \mathbf{k}_j \rangle = \langle \widetilde{\mathbf{q}}_n^*, \mathbf{R}^{n-j} \widetilde{\mathbf{k}}_j^* \rangle = \langle \mathbf{R}^{-n} \widetilde{\mathbf{q}}_n^*, \mathbf{R}^{-j} \widetilde{\mathbf{k}}_j^* \rangle, \qquad \|\mathbf{k}_j\|_{\mathbf{H}}^2 = \|\widetilde{\mathbf{k}}_j^*\|^2. \tag{54}$$

If an implementation uses forward powers $\mathbf{R}^n$ instead of negative powers, take the generative rotation to be the inverse of the implementation's rotation.

C.1 Proof of Result 3.2

Proof. We work in the small-noise regime $\mathbf{H} = \mathbf{I}_{d_h}$ with orthogonal $\mathbf{\Theta}$, abbreviate the transitioned unit key and its direction by

$$\mathbf{w}_{nj} := \mathbf{\Theta}^{n-j} \widehat{\mathbf{k}}_j \in S^{d_h - 1}, \qquad \|\mathbf{w}_{nj}\| = 1, \qquad \mathbf{q}_n = \mathbf{W}_q \mathbf{z}_n, \quad \mathbf{W}_q = \mathbf{P}_q^\top \mathbf{D}^{-1}, \tag{55}$$

and write $\boldsymbol{\sigma}_n = \widehat{\mathbf{q}}_n = \mathbf{q}_n / \|\mathbf{q}_n\|$.

Step 1: two identities for the mean. Applying $\mathbf{W}_q$ to (15),

$$\mathbf{W}_q \widetilde{\boldsymbol{\mu}}_{nj} = \underbrace{\mathbf{P}_q^\top \mathbf{D}^{-1} \mathbf{P}_q}_{=\mathbf{H}=\mathbf{I}_{d_h}} \mathbf{\Theta}^{n-j} \widehat{\mathbf{k}}_j = \mathbf{w}_{nj}, \qquad \|\mathbf{W}_q \widetilde{\boldsymbol{\mu}}_{nj}\| = \|\mathbf{w}_{nj}\| = 1. \tag{56}$$

The boundary rescaling in (15) therefore gives

$$\boldsymbol{\mu}_{nj} = \beta_0^{-1} \mathbf{P}_q \mathbf{w}_{nj}, \qquad \mathbf{W}_q \boldsymbol{\mu}_{nj} = \beta_0^{-1} \mathbf{w}_{nj}, \qquad \beta_0^2 \boldsymbol{\mu}_{nj}^\top \mathbf{\Sigma} \boldsymbol{\mu}_{nj} = \beta_0^2 \|\mathbf{W}_q \boldsymbol{\mu}_{nj}\|^2 = 1, \tag{57}$$

i.e. $\boldsymbol{\mu}_{nj}$ lies on the parameter-domain boundary, as claimed in Assumption 3.1.

Step 2: query disintegration under the base measure. Under $\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$,

$$\mathbf{W}_q \mathbf{D} \mathbf{W}_q^\top = \mathbf{P}_q^\top \mathbf{D}^{-1} \mathbf{P}_q = \mathbf{I}_{d_h} \quad \Longrightarrow \quad \mathbf{q}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_h}), \tag{58}$$

and Gaussian regression gives the orthogonal decomposition

$$\mathbf{z}_n = \mathbf{P}_q \mathbf{q}_n + \boldsymbol{\varepsilon}_n^\perp, \qquad \boldsymbol{\varepsilon}_n^\perp \sim \mathcal{N}(\mathbf{0}, \mathbf{D} - \mathbf{P}_q \mathbf{P}_q^\top), \qquad \boldsymbol{\varepsilon}_n^\perp \perp \mathbf{q}_n. \tag{59}$$

The polar decomposition $\mathbf{q}_n = r_n \boldsymbol{\sigma}_n$ satisfies

$$r_n = \|\mathbf{q}_n\| \sim \chi_{d_h}, \qquad \boldsymbol{\sigma}_n \sim \mathrm{Unif}(S^{d_h - 1}), \qquad r_n \perp \boldsymbol{\sigma}_n, \tag{60}$$

and the tilt function in (14) factorizes through $(r_n, \boldsymbol{\sigma}_n)$:

$$\mathbf{z}_n^\top \mathbf{\Sigma} \boldsymbol{\mu}_{nj} = (\mathbf{W}_q \mathbf{z}_n)^\top (\mathbf{W}_q \boldsymbol{\mu}_{nj}) = \beta_0^{-1} r_n \langle \boldsymbol{\sigma}_n, \mathbf{w}_{nj} \rangle, \tag{61}$$

$$\sqrt{\mathbf{z}_n^\top \mathbf{\Sigma} \mathbf{z}_n} = \|\mathbf{W}_q \mathbf{z}_n\| = r_n, \tag{62}$$

where (61) uses $(\mathbf{W}_q \mathbf{z}_n)^\top = \mathbf{z}_n^\top \mathbf{D}^{-1} \mathbf{P}_q$ and (57).

Step 3: legality of the tilted density. Substituting (61) and (62), the tilted density (14) reads

$$p(\mathbf{z}_n \mid J_n = j, \mathcal{F}_{n-1}) = \Big(1 + \big\langle \boldsymbol{\sigma}_n, \mathbf{w}_{nj} \big\rangle\Big) \mathcal{N}(\mathbf{z}_n; \mathbf{0}, \mathbf{D}), \qquad \boldsymbol{\sigma}_n = \mathbf{W}_q \mathbf{z}_n / \|\mathbf{W}_q \mathbf{z}_n\|. \tag{63}$$

Write $t(\boldsymbol{\sigma}) = \langle \boldsymbol{\sigma}, \mathbf{w}_{nj} \rangle$. We verify the four properties.

- *Nonnegativity.* By Cauchy–Schwarz and $\|\mathbf{w}_{nj}\| = 1$,

$$1 + t(\boldsymbol{\sigma}) \in [1 - \|\boldsymbol{\sigma}\|\,\|\mathbf{w}_{nj}\|,\ 1 + \|\boldsymbol{\sigma}\|\,\|\mathbf{w}_{nj}\|] = [0, 2] \qquad \forall \boldsymbol{\sigma} \in S^{d_h - 1}. \tag{64}$$

- *Partition function.* Sign symmetry of the uniform sphere gives $\mathbb{E}[t(\boldsymbol{\sigma}_n)] = \langle \mathbb{E}\boldsymbol{\sigma}_n, \mathbf{w}_{nj} \rangle = 0$, hence

$$\int \Big(1 + t(\boldsymbol{\sigma}_n)\Big)\, \mathcal{N}(\mathbf{z}_n; \mathbf{0}, \mathbf{D})\, d\mathbf{z}_n = 1. \tag{65}$$

- *Mean.* Using (59), $\mathbb{E}[\boldsymbol{\varepsilon}_n^{\perp}] = \mathbf{0}$, $r_n \perp \boldsymbol{\sigma}_n$, and the uniform second moment $\mathbb{E}[\boldsymbol{\sigma}_n \boldsymbol{\sigma}_n^{\top}] = \mathbf{I}_{d_h}/d_h$,

$$\mathbb{E}\big[\mathbf{z}_n\big(1 + t(\boldsymbol{\sigma}_n)\big)\big] = \mathbf{P}_q\, \mathbb{E}[r_n]\, \mathbb{E}[\boldsymbol{\sigma}_n \boldsymbol{\sigma}_n^{\top}]\, \mathbf{w}_{nj} = \frac{\mathbb{E}[r_n]}{d_h}\, \mathbf{P}_q \mathbf{w}_{nj} = \frac{\mathbb{E}[r_n]}{d_h}\, \widetilde{\boldsymbol{\mu}}_{nj}. \tag{66}$$

  Matching (66) to $\boldsymbol{\mu}_{nj} = \beta_0^{-1} \mathbf{P}_q \mathbf{w}_{nj}$ pins the dimension constant:

$$\beta_0 = \frac{d_h}{\mathbb{E}[r_n]} = \frac{d_h\, \Gamma(d_h/2)}{\sqrt{2}\, \Gamma\big((d_h + 1)/2\big)}, \qquad \mathbb{E}[r_n] = \frac{\sqrt{2}\, \Gamma\big((d_h + 1)/2\big)}{\Gamma(d_h/2)}, \tag{67}$$

  the mean of the χ_{d_h} distribution.

- *Second moment.* The residual contributes nothing: the terms containing $\boldsymbol{\varepsilon}_n^{\perp}$ carry a factor $\mathbb{E}[\boldsymbol{\varepsilon}_n^{\perp}] = \mathbf{0}$ or $\mathbb{E}[t(\boldsymbol{\sigma}_n)] = 0$, and the third moment of the uniform sphere vanishes by sign symmetry, $\mathbb{E}[\boldsymbol{\sigma}_n \boldsymbol{\sigma}_n^{\top} \langle \boldsymbol{\sigma}_n, \mathbf{w}_{nj} \rangle] = \mathbf{0}$. Hence

$$\mathbb{E}\big[\mathbf{z}_n \mathbf{z}_n^{\top}\big(1 + t(\boldsymbol{\sigma}_n)\big)\big] = \mathbb{E}[\mathbf{z}_n \mathbf{z}_n^{\top}] = \mathbf{D}, \qquad \mathrm{Cov}(\mathbf{z}_n \mid J_n = j) = \mathbf{D} - \boldsymbol{\mu}_{nj} \boldsymbol{\mu}_{nj}^{\top}. \tag{68}$$

Step 4: exact posterior. The likelihood ratio against the base measure is, by (63) and (55),

$$\frac{p(\mathbf{z}_n \mid J_n = j, \mathcal{F}_{n-1})}{\mathcal{N}(\mathbf{z}_n; \mathbf{0}, \mathbf{D})} = 1 + \big\langle \boldsymbol{\sigma}_n, \mathbf{w}_{nj} \big\rangle = 1 + s_{nj}, \qquad s_{nj} = \big\langle \widehat{\mathbf{q}}_n, \boldsymbol{\Theta}^{n-j} \widehat{\mathbf{k}}_j \big\rangle \in [-1, 1], \tag{69}$$

where s_{nj} is $\mathcal{F}_n$-measurable given j. Bayes' rule together with locality (Lemma A.3) cancels the common Gaussian factor and gives

$$\pi_{nj} = \mathbb{P}(J_n = j \mid \mathcal{F}_n) = \frac{\rho_n^-(j)\big(1 + s_{nj}\big)}{\sum_{j' < n} \rho_n^-(j')\big(1 + s_{nj'}\big)}, \tag{70}$$

which is (18); $|s_{nj}| \le 1$ gives the weight bound $2\rho_n^-(j)$.

Step 5: posterior mean and joint normalization. Taking the conditional expectation of the value readout $\mathbf{t}_n$ in (A4),

$$\mathrm{SLA}(\mathbf{z}_{\le n}) := \mathbb{E}[\mathbf{t}_n \mid \mathcal{F}_n] = (1 - \eta_n) \sum_{j < n} \pi_{nj} \mathbf{v}_j + \eta_n \mathbf{v}_n, \tag{71}$$

since $\mathbf{v}_j$ ($j \le n$) and η_n are $\mathcal{F}_n$-measurable; this is (19), and the self-inclusive weights follow by the self-weight extension of Step 3 of Result 2.5 with ℓ_{nj} replaced by $\log \rho_n^-(j) + \log(1 + s_{nj})$. For the joint density, each row density (63) is normalized for every j by (65), so integrating $\mathbf{z}_N$ and summing J_N contributes a factor of one; iterating the same operation backward from $n = N$ to $n = 2$ (together with p_1) proves normalization of the joint.

Step 6: recurrent form. Instantiate the lag prior by the product decay

$$\rho_n^-(j) = \frac{\prod_{i=j}^{n-1} \gamma_i}{z_n^{\mathrm{sum}}}, \qquad z_n^{\mathrm{sum}} = \sum_{j < n} \prod_{i=j}^{n-1} \gamma_i, \qquad \gamma_i = \gamma(\mathbf{z}_i) \in (0, 1), \tag{72}$$

where each γ_i is $\sigma(\mathbf{z}_i)$-measurable, so ρ_n^- is $\mathcal{F}_{n-1}$-measurable as Assumption 2.1 requires. Define the accumulators

$$\mathbf{M}_n = \sum_{j < n} \Big(\prod_{i=j}^{n-1} \gamma_i\Big) \boldsymbol{\Theta}^{n-j} \widehat{\mathbf{k}}_j \mathbf{v}_j^{\top}, \qquad \mathbf{s}_n^v = \sum_{j < n} \Big(\prod_{i=j}^{n-1} \gamma_i\Big) \mathbf{v}_j, \qquad \mathbf{s}_n^k = \sum_{j < n} \Big(\prod_{i=j}^{n-1} \gamma_i\Big) \boldsymbol{\Theta}^{n-j} \widehat{\mathbf{k}}_j. \tag{73}$$

Factorizing the product and the transition per lag,

$$\Big(\prod_{i=j}^{n}\gamma_i\Big)\boldsymbol{\Theta}^{n+1-j} = \gamma_n\,\boldsymbol{\Theta}\Big(\Big(\prod_{i=j}^{n-1}\gamma_i\Big)\boldsymbol{\Theta}^{n-j}\Big) \qquad (j<n), \tag{74}$$

and isolating the $j=n$ summand of (73) gives the fixed-state updates

$$\begin{aligned}
\mathbf{M}_{n+1} &= \gamma_n\boldsymbol{\Theta}\big(\mathbf{M}_n + \widehat{\mathbf{k}}_n\mathbf{v}_n^\top\big),\\
\mathbf{s}^v_{n+1} &= \gamma_n\big(\mathbf{s}^v_n + \mathbf{v}_n\big),\\
\mathbf{s}^k_{n+1} &= \gamma_n\boldsymbol{\Theta}\big(\mathbf{s}^k_n + \widehat{\mathbf{k}}_n\big),\\
z^{\mathrm{sum}}_{n+1} &= \gamma_n\big(z^{\mathrm{sum}}_n + 1\big),
\end{aligned}$$

which is (20). Finally, splitting $1+s_{nj} = 1+\langle\widehat{\mathbf{q}}_n,\boldsymbol{\Theta}^{n-j}\widehat{\mathbf{k}}_j\rangle$,

$$\sum_{j<n}\rho_n^-(j)\big(1+s_{nj}\big)\mathbf{v}_j = \frac{\mathbf{s}^v_n + \mathbf{M}_n^\top\widehat{\mathbf{q}}_n}{z^{\mathrm{sum}}_n}, \tag{75}$$

$$\sum_{j<n}\rho_n^-(j)\big(1+s_{nj}\big) = \frac{z^{\mathrm{sum}}_n + \big\langle\widehat{\mathbf{q}}_n,\mathbf{s}^k_n\big\rangle}{z^{\mathrm{sum}}_n}, \tag{76}$$

so the ratio of (75) and (76) gives the strict-past readout

$$\mathbb{E}[\mathbf{v}_{J_n}\mid\mathcal{F}_n] = \frac{\mathbf{s}^v_n + \mathbf{M}_n^\top\widehat{\mathbf{q}}_n}{z^{\mathrm{sum}}_n + \big\langle\widehat{\mathbf{q}}_n,\mathbf{s}^k_n\big\rangle}, \tag{77}$$

and attaching the self weight η_n gives the self-inclusive readout (21). The geometric prior $\gamma_i\equiv e^{-\lambda}$ is the special case of (20)–(21) with a constant forget gate. □