Sessa: Selective State Space Attention

Liubomyr Horbatko
liubomir.horbatko@gmail.com

Abstract

Modern sequence models are dominated by Transformers, where self-attention mixes information from the visible context in an input-dependent way. However, when retrieval is not sharp and attention remains diffuse over an effective support $S_{\text{eff}}(t)$, the influence of any individual token is diluted, typically scaling as $O(1/S_{\text{eff}}(t))$ and reaching $O(1/\ell)$ for old tokens in full-prefix settings. Structured state-space models process sequences recurrently through an explicit feedback path; selective variants such as Mamba make this feedback input-dependent, yet when freeze time cannot be sustained over long intervals, their long-range sensitivity decays exponentially with lag. Existing architectures therefore either retrieve from the past in a single read or propagate information through a single feedback chain. We introduce Sessa, a decoder that places attention inside a feedback path, enabling recurrent many-path aggregation within a layer. Under stated assumptions, Sessa admits regimes with a power-law memory tail in lag ℓ of order $O(\ell^{-\beta})$ for $0 < \beta < 1$, which is asymptotically slower than $1/\ell$; moreover, this rate is tight in an explicit diffuse uniform-routing setting where the influence is $\Theta(\ell^{-\beta})$. Under the same conditions, only Sessa among the compared model classes realizes flexible selective retrieval, including non-decaying profiles. Empirically, under matched architectures and training budgets, Sessa achieves the strongest performance on our long-context benchmarks while remaining competitive with Transformer and Mamba style baselines on short-context language modeling.

1 Introduction

Long-context sequence modeling is central to modern foundation models across language, vision, speech, time series, and genomics (Bommasani et al., 2021; Brown et al., 2020; Dosovitskiy et al., 2021; Baevski et al., 2020; Ansari et al., 2024; Dalla-Torre et al., 2025). Despite the architectural flexibility of the foundation-model paradigm, state-of-the-art systems are still overwhelmingly based on the Transformer and its self-attention mechanism (Vaswani et al., 2017).

A useful lens is to describe modern sequence mixers by how they route information from the past and how they maintain memory over time. In many modern architectures, routing decisions are input-dependent: the model uses the current token and its context to decide which parts of the visible history to consult. Under this view, self-attention implements an input-dependent *direct-read* mechanism: at each position, it computes a query-dependent pattern of relevance over the visible context and uses it to read out information from selected past positions. This framing highlights attention's key strength, a selection mechanism over variable support length, but also a structural limitation: the retrieval is performed in a single pass, without an internal feedback loop that would repeatedly incorporate past readouts into an evolving state. Separately, standard implementations are also computationally expensive at long contexts due to quadratic time/memory scaling (Vaswani et al., 2017; Rabe and Staats, 2021).

In parallel, structured recurrent sequence models, especially state space models (SSMs), which realize long-range dynamics through a latent state and an explicit feedback path, have re-emerged as a compelling alternative for long-context modeling (Gu et al., 2022a,b). SSMs can be interpreted as modern descendants of classical dynamical systems (Kalman, 1960) and admit linear (or near-linear) scaling in sequence length. However, for information-dense discrete data, a persistent challenge is that stable feedback dynamics often exhibit rapid attenuation of distant information (commonly exponential forgetting (Huang et al., 2025)), which can hinder integrating multiple far-apart evidence snippets under heavy distractors. Selective SSMs (e.g., Mamba) can conditionally slow this

attenuation by modulating the effective transition (Gu and Dao, 2024; Dao and Gu, 2024) (e.g., $A_{\text{ssm},t} \approx I$ on selected steps, "freeze time" (Huang et al., 2025)), but this mechanism is input-dependent and can fail when relevant and irrelevant positions induce similar local representations, leading to preserving or overwriting the wrong content.

These perspectives suggest complementary long-context failure modes. Stable feedback dynamics can suffer from exponential forgetting. Attention, while input-dependent, can suffer from dilution: when attention mass is spread across a large effective support of competing tokens (e.g., many near-tied logits), individual weights, and thus per-token contributions and sensitivities, decrease roughly inversely with that support (often behaving like $O(1/S_{\text{eff}}(t))$, and in the worst case like $O(1/T)$ when the effective support grows proportionally with context length T)(Mudarisov et al., 2025). In practice, both effects can limit reliable long-range evidence integration.

We introduce **Sessa**, a decoder architecture that injects input-dependent attention into a feedback (recurrent) path, combining direct-read input-dependent routing with stateful aggregation through the feedback channel. Viewed through a temporal routing lens, for a fixed source token τ and target position t (lag $\ell = t - \tau$), a single self-attention layer routes influence via a *single routing step* (a direct edge $\tau \to t$), while chain-structured state-space recurrences propagate along the unique length-ℓ temporal chain. Sessa introduces route diversity within a single layer: its attention-induced feedback operator aggregates contributions over multiple internal routing depths (and, in dense patterns, many temporal paths), which can help sustain long-range sensitivity when routing is diffuse (formalized in Section 4.2). Concretely, while self-attention corresponds to an input-dependent direct-read system (in the values), Sessa realizes an input-dependent feedback system: it maintains a latent state over unbounded horizons, while the feedback dynamics remain input-dependent via attention-based routing inside the loop (potentially over variable-support patterns). Intuitively, Sessa retains the representational benefits of recurrence for long-range accumulation while leveraging attention as an input-dependent mechanism within the feedback pathway.

Related architectural ideas have introduced recurrence or feedback into sequence modeling (Dai et al., 2019; Fan et al., 2020; Bulatov et al., 2022; Hutchins et al., 2022; Hwang et al., 2024). These approaches span a variety of feedback constructions and are typically presented in architecture-specific terms. Our contribution is complementary but mathematically different: we propose a routing-induced systems perspective that separates how context produces routing/mixing coefficients from how those coefficients are composed over time, and we use this lens to relate input-dependent routing directly to long-context sensitivity and memory-decay behavior.

Our contributions are:

- **Architecture.** We propose the Sessa sequence mixer, integrating attention into the recurrent feedback pathway under an otherwise standard decoder macro-architecture.
- **Memory.** We characterize long-range sensitivity of Sessa and identify a heavy-tail memory regime in which the feedback solve induces a **power-law influence tail in the lag ℓ of order $O(\ell^{-\beta_{\text{tail}}})$ with $0 < \beta_{\text{tail}} < 1$.** In this diffuse, low-separation routing regime, attenuation is asymptotically slower than the exponential forgetting exhibited by many stable or contractive SSM regimes, and it mitigates inverse-support dilution effects under the stated assumptions (Section 4.2; Theorem 8).
- **Selective retrieval.** In the matched theoretical regime, we show that deep Sessa realizes flexible selective retrieval profiles, including non-decaying ones, whereas diffuse fixed-depth Transformers and failed-freeze-time fixed-depth Mamba do not (Section 4.2.8; Theorem 12; Proposition 13).
- **Empirics.** Under matched architectures and training budgets, Sessa achieves the strongest performance on our long-context benchmarks while remaining competitive on short-context language modeling.

We additionally prove a universal approximation result for a broad class of causal sequence mappings in Appendix I (Theorem 14).

2 Background

We separate two largely independent aspects of causal mixers:

(i) how routing/mixing coefficients are produced from context, and
(ii) whether information is accessed via a single read or accumulated through feedback.

Terminology We use *system* to refer to the memory mechanism (direct-read or feedback). We use *routing* to refer to the coefficients that specify how information flows over time for example attention weights α^{fwd}, the induced feedback matrix B_{fb}, or the transition operators in a recurrence. Routing is the collection of coefficients, meaning weights or operators, that determine information flow over time. The system determines whether routing is applied once (direct-read) or repeatedly composed via feedback.

2.1 Direct-read and feedback systems

We model a broad class of sequence mixers by expressing each output as a mixture of a chosen stream u_t with coefficients that may depend on the available context $x_{0:t}$.

Definition 1 (Direct-read variable-support system)**.** We say that $\mathcal{F}$ is a direct-read system with respect to a chosen stream u_t if, for every t,

$$y_t \;=\; \sum_{\tau \in S_t} K_{t,\tau}(x_{0:t})\, u_\tau, \qquad S_t \subseteq \{0, \dots, t\}, \tag{1}$$

so each y_t is produced by a single input-addressed read, i.e., a mixture over the visible index set S_t. If $|S_t|$ varies with t, we call the system *variable-support*. If there exists $W \geq 1$ such that $K_{t,\tau} \equiv 0$ whenever $t - \tau \geq W$, equivalently, $S_t \subseteq \{\max(0, t - W + 1), \dots, t\}$, we call it *bounded-support direct-read*.

Remark 2.1 (Kernel representations alone do not distinguish direct-read or feedback). On any finite horizon T, any causal linear map admits a lower-triangular kernel representation (Kalman, 1960; Antsaklis and Michel, 2006). $y_t = \sum_{\tau \leq t} K_{t,\tau} u_\tau$, so kernel form alone does not identify whether influence is produced by a single read or by an internal recurrence. Here, direct-read refers to the computation graph: y_t is formed by one read/mix over a visible set, without repeated composition of the same mixing primitive inside the layer.

Dimensions. $u_\tau \in \mathbb{R}^D$, $y_t \in \mathbb{R}^D$, and $K_{t,\tau}(x_{0:t})$ is a linear map of the appropriate shape.

In contrast, models with an explicit state and feedback naturally take a feedback form.

Definition 2 (Feedback system: state-space or operator form)**.** We say that $\mathcal{G}$ is a feedback system with respect to a chosen stream u_t if there exist states h_t in a possibly time-varying state space $\mathcal{H}_t$ such that, for each $t \geq 0$,

$$\text{with, e.g., } h_{-1} = 0, \qquad h_t \;=\; A_{\text{ssm},t}(x_{0:t})\, h_{t-1} \;+\; B_{\text{ssm},t}(x_{0:t})\, u_t, \qquad y_t \;=\; C_{\text{ssm},t}(x_{0:t})\, h_t \;+\; D_{\text{ssm},t}(x_{0:t})\, u_t. \tag{2}$$

The recurrence composes the routing over time, so y_t can depend on arbitrarily old inputs even when each update is local in h_{t-1}.

Remark 2.2 (One-hop and multi-hop routing). We view routing as propagation on a directed acyclic graph (DAG) over time indices induced by the mixing coefficients. Fix a horizon T and nodes $\{0, \dots, T-1\}$.

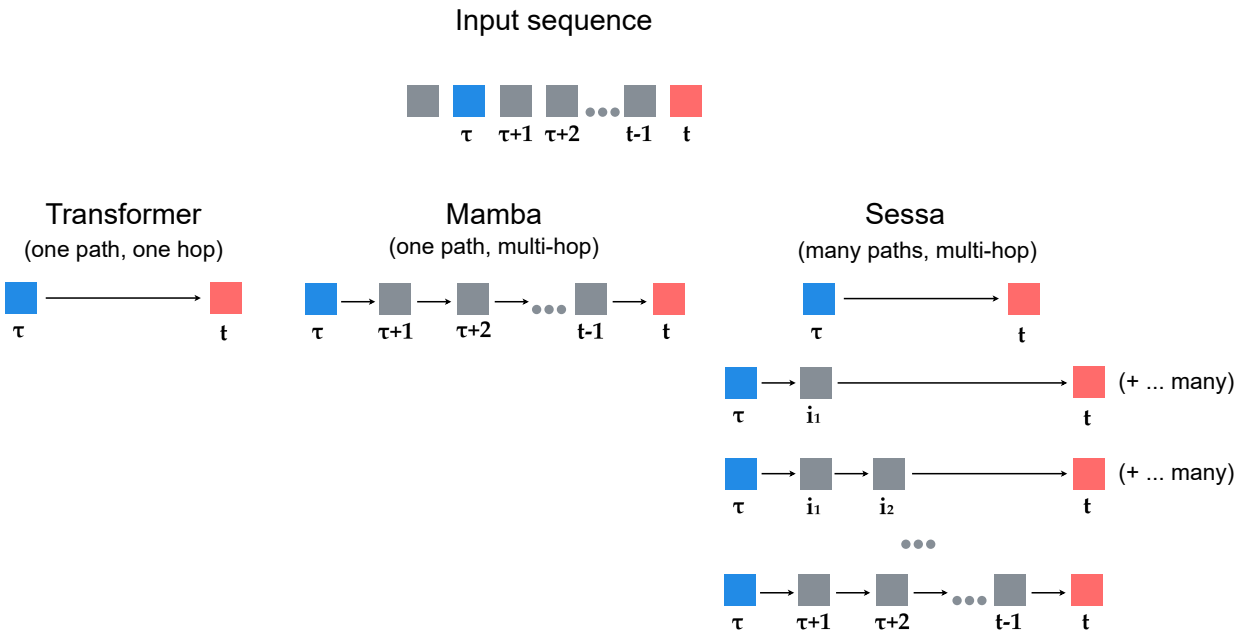

Figure 1: One-hop and multi-hop temporal routing within a single mixer layer.
Transformer: influence from τ to t follows a single direct edge (one-hop).
Mamba: influence from τ to t follows the chain $\tau \to \cdots \to t$ (multi-hop along a single path).
Sessa: influence from τ to t aggregates over many paths with varying hop counts (multi-hop over many paths).

Direct-read (one-hop). A direct-read system forms y_t by a single read from a visible set S_t using coefficients $K_{t,\tau}$: in the routing graph, this corresponds to using only direct edges $\tau \to t$. Influence from τ reaches t in one routing step.

Feedback (multi-hop). A feedback mechanism can apply routing repeatedly through an internal state or solve, allowing influence from τ to reach t through paths with intermediate nodes. This repeated composition is what we call multi-hop routing.

The classical finite-dimensional state-space case corresponds to $\mathcal{H}_t = \mathbb{R}^{\mathsf{N}}$ with fixed N for all t. Structured SSM layers (e.g., S4/S4D and Mamba) are instances of this special case.

Hop counts in the solve Sessa's mixer output s is defined by a causal lower-triangular solve

$$(I - B_{\mathrm{fb}})\, s = f, \qquad [B_{\mathrm{fb}}]_{t,j} = 0 \text{ for } j \geq t, \tag{3}$$

On any finite horizon T, B_{fb} is strictly lower-triangular and hence nilpotent ($B_{\mathrm{fb}}^T = 0$) (Horn and Johnson, 2012). Hence,

$$(I - B_{\mathrm{fb}})^{-1} = \sum_{k=0}^{T-1} B_{\mathrm{fb}}^k, \qquad \text{and} \qquad s = \sum_{k=0}^{T-1} B_{\mathrm{fb}}^k f. \tag{4}$$

Each term $B_{\mathrm{fb}}^k f$ corresponds to routing through k feedback steps, a k-hop contribution. Equivalently, for indices $\tau \leq t$,

$$(B_{\mathrm{fb}}^k)_{t,\tau} = \sum_{\tau = i_0 < i_1 < \cdots < i_k = t} \prod_{r=1}^{k} [B_{\mathrm{fb}}]_{i_r, i_{r-1}}, \qquad k \geq 1, \tag{5}$$

which is a sum over all length-k directed paths from τ to t in the feedback-induced routing graph. This explicit path expansion is the mechanism behind heavy-tail regimes analyzed later: even if individual edges are small under diffuse routing, the number of admissible paths grows with lag, and the solve aggregates contributions across all hop counts.

2.2 Self-attention as direct-read

Standard causal self-attention fits Definition 1 when the mixed stream is the sequence of value vectors. At position t, over a visible index set $\mathcal{W}_t \subseteq \{0, \dots, t\}$:

$$y_t = \sum_{j \in \mathcal{W}_t} \alpha_{t,j}^{\mathrm{fwd}}\, v_j, \qquad \alpha_{t,j}^{\mathrm{fwd}} = \frac{\exp\big(\sigma_k\, q_t^\top k_j\big)}{\sum_{i \in \mathcal{W}_t} \exp\big(\sigma_k\, q_t^\top k_i\big)}, \tag{6}$$

with $q_t = W_Q x_t$, $k_j = W_K x_j$, and $v_j = W_V x_j$.

Lemma 2.3 (Self-attention is a direct-read system in V)**.** *At each position t, self-attention computes y_t by a single input-addressed read from the visible set $\mathcal{W}_t$, mixing the value vectors $(v_j)_{j \in \mathcal{W}_t}$ with context-dependent weights $\alpha_{t,j}^{\mathrm{fwd}}$.*

Full-prefix, windowed, and sparse attention all fit the same direct-read template through the choice of visible set $\mathcal{W}_t$ (Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020; Ding et al., 2023).

2.3 State-space models as feedback

Structured state-space models (SSMs) implement sequence mixing through a latent state and a (possibly selective) recurrence. A standard form is

$$h_t = A_{\mathrm{ssm}}\, h_{t-1} + B_{\mathrm{ssm}}\, x_t, \qquad y_t = C_{\mathrm{ssm}}\, h_t, \tag{7}$$

where $A_{\mathrm{ssm}} \in \mathbb{R}^{\mathsf{N} \times \mathsf{N}}$ encodes temporal dynamics and is typically constrained (diagonal/structured/low-rank) for efficiency.

Modern language-oriented SSMs such as Mamba often employ input-dependent recurrences that fit Definition 2:

$$h_t \;=\; A_{\mathrm{ssm},t}(x_{0:t})\, h_{t-1} \;+\; B_{\mathrm{ssm},t}(x_{0:t})\, x_t, \qquad y_t \;=\; C_{\mathrm{ssm},t}(x_{0:t})\, h_t. \tag{8}$$

In Mamba, the discrete transition commonly takes the form

$$A_{\mathrm{ssm},t} = \mathrm{diag}(\exp(-\lambda_n \Delta_t)),$$

so a lag-ℓ memory factor contains terms of the form

$$\exp\Big(-\lambda_n \sum_{r=t-\ell+1}^{t} \Delta_r\Big).$$

Accordingly, long-range memory is preserved only when the model can create a long *preserve corridor* of steps with $\Delta_r \approx 0$.

This suggests the matched comparison principle used later in the paper. For attention, broken sharp selection means that softmax mass cannot concentrate on a small set of indices. For Mamba, the analogous failure mode is *failed freeze time*: the model cannot sustain a long preserve corridor on the relevant interval. For the three-way comparison in this paper, we say that a Mamba layer is in a *failed freeze-time regime* on an input set of interest

if there exists $c_\Delta > 0$ such that for every relevant pair $\tau < t$,

$$\sum_{r=\tau+1}^{t} \Delta_r \;\geq\; c_\Delta (t-\tau).$$

Equivalently, the average discretization step along every relevant interval is bounded below by a positive constant. In Mamba this implies

$$\exp\Big(-\lambda_n \sum_{r=\tau+1}^{t} \Delta_r\Big) \;\leq\; e^{-\lambda_n c_\Delta (t-\tau)},$$

so long-range influence is exponentially small in the lag. This is the Mamba counterpart of diffuse attention used in the matched comparisons below: in attention, the selector cannot concentrate mass on a few indices; in Mamba, the model cannot maintain $\Delta_r \approx 0$ on a long relevant corridor.

3 Model Architecture

We instantiate the one-hop and multi-hop routing viewpoint of Section 2.1 with a concrete layer, Sessa. Sessa uses a single gated-MLP-style block that wraps a recurrent mixer, rather than alternating separate attention and MLP blocks. The mixer itself combines (i) a standard causal forward-attention signal and (ii) a feedback term that mixes past mixer outputs.

The official implementation is available at https://github.com/LibratioAI/sessa.

Notation. Inputs and outputs have shape $x, y \in \mathbb{R}^{B_{\text{batch}} \times T \times D}$ with $t \in \{0, \dots, T-1\}$. We use an internal key and query width d_k and scale $\sigma_k = d_k^{-1/2}$. All definitions apply per batch element; we omit the batch index when clear.

3.1 Sessa block

Given $x \in \mathbb{R}^{B_{\text{batch}} \times T \times D}$, the block applies pre-norm, a gated projection, the mixer, and a residual connection:

$$\tilde{x} = \mathrm{LN}(x), \tag{9}$$

$$(a, g) = \mathrm{split}(\tilde{x} W^{\text{in}} + b^{\text{in}}), \qquad a, g \in \mathbb{R}^{B_{\text{batch}} \times T \times D}, \tag{10}$$

$$\bar{a} = \mathrm{GELU}(a), \tag{11}$$

$$s = \mathrm{Mixer}(\bar{a}) \in \mathbb{R}^{B_{\text{batch}} \times T \times D}, \tag{12}$$

$$y = x + ((s \odot g) W^{\text{out}} + b^{\text{out}}). \tag{13}$$

We use Layer Normalization (Ba et al., 2016) and the GELU nonlinearity (Hendrycks and Gimpel, 2016). Here $W^{\text{in}} \in \mathbb{R}^{D \times 2D}$ and $W^{\text{out}} \in \mathbb{R}^{D \times D}$. The elementwise gate g plays the usual role of gated MLP variants (Hua et al., 2022; Shazeer, 2020): it modulates the mixer output before the residual add.

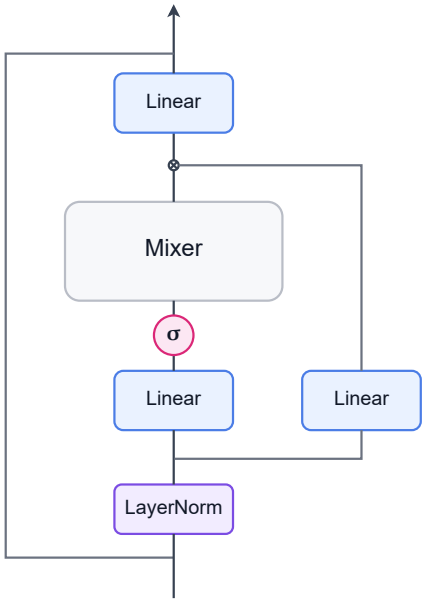

Figure 2: Sessa Layer.

3.2 Sessa mixer

The mixer maps $\bar{a} \in \mathbb{R}^{B_{\text{batch}} \times T \times D}$ to $s \in \mathbb{R}^{B_{\text{batch}} \times T \times D}$. It uses two causal attention mechanisms: (i) a forward causal attention that produces a forward signal $f_t \in \mathbb{R}^D$, and (ii) a feedback attention that produces weights over the strict past, used inside a causal feedback solve.

Projections. At each time t, we form forward queries, keys, and values, as well as feedback queries and keys, using standard linear projections:

$$q_t^f = \bar{a}_t W_{Qf}, \quad k_t^f = \bar{a}_t W_{Kf}, \quad v_t = \bar{a}_t W_V, \qquad q_t^b = \bar{a}_t W_{Qb}, \quad k_t^b = \bar{a}_t W_{Kb}, \tag{14}$$

where $q^f, k^f, q^b, k^b \in \mathbb{R}^{d_k}$ and $v_t \in \mathbb{R}^D$. We apply RoPE to the forward pair (q^f, k^f). We use rotary position embeddings in the forward branch (Su et al., 2021).

Forward attention. Define causal weights over $j \le t$:

$$\alpha_{t,j}^{\text{fwd}} = \text{softmax}_{0 \le j \le t}\Big(\sigma_k \langle \text{RoPE}(q_t^f), \text{RoPE}(k_j^f) \rangle\Big), \tag{15}$$

and the forward signal

$$f_t = \sum_{j=0}^{t} \alpha_{t,j}^{\text{fwd}} \, v_j \in \mathbb{R}^D. \tag{16}$$

This is a one-hop mixture of values $(v_j)_{j \le t}$ over a finite visible set.

Feedback attention. Define feedback weights over the strict past $j < t$:

$$\alpha_{t,j}^{\text{fb}} = \begin{cases} \text{softmax}_{0 \le j \le t-1}\Big(\sigma_k \langle q_t^b, k_j^b \rangle\Big), & t \ge 1, \ j < t, \\ 0, & j \ge t, \end{cases} \qquad \alpha_{0,j}^{\text{fb}} = 0 \ \ \forall j. \tag{17}$$

Feedback gain. We modulate the feedback with a scalar gain $\gamma_t \in (-1, 1)$:

$$\gamma_t = \tanh(\langle \bar{a}_t, w^\gamma \rangle + b^\gamma). \tag{18}$$

The bound controls feedback magnitude: since $\alpha^{\text{fb}}_{t,\cdot}$ is a convex distribution over $j < t$, the feedback term is a convex combination of past states scaled by $|\gamma_t| < 1$.

Feedback routing matrix.

$$[B_{\text{fb}}]_{t,j} = \gamma_t \, \alpha^{\text{fb}}_{t,j}, \qquad [B_{\text{fb}}]_{t,j} = 0 \text{ for } j \geq t. \tag{19}$$

Scalar routing and feature-wise solve. Here B_{fb} is a scalar strictly lower-triangular routing matrix (each $[B_{\text{fb}}]_{t,j} \in \mathbb{R}$). The solve $(I - B_{\text{fb}})s = f$ is applied independently to each feature dimension of $s, f \in \mathbb{R}^{T \times D}$: for every $d \in \{1, \ldots, D\}$,

$$(I - B_{\text{fb}}) \, s_{:,d} = f_{:,d},$$

In vectorized form,

$$\big(I_D \otimes (I - B_{\text{fb}})\big) \, \text{vec}(s) = \text{vec}(f).$$

The resulting recurrence (22) therefore uses scalar–vector multiplication ($[B_{\text{fb}}]_{t,j} \, s_j$ with $[B_{\text{fb}}]_{t,j} \in \mathbb{R}$ and $s_j \in \mathbb{R}^D$).

Lower-triangular solve. The mixer output $s \in \mathbb{R}^{T \times D}$ is the unique solution of

$$(I - B_{\text{fb}})s = f \tag{20}$$

which is a unit-lower-triangular solve with D right-hand sides. This can be implemented with optimized triangular-solve routines (e.g., batched `solve_triangular`/TRSM kernels), avoiding explicit formation of $(I - B_{\text{fb}})^{-1}$. Thus, in the dense full-prefix formulation, the mixer remains quadratic in T. Equivalently, forward substitution gives the explicit recurrence

$$s_0 = f_0, \tag{21}$$

$$s_t = f_t + \sum_{j=0}^{t-1} [B_{\text{fb}}]_{t,j} \, s_j \;=\; f_t + \gamma_t \sum_{j=0}^{t-1} \alpha^{\text{fb}}_{t,j} \, s_j, \qquad t \geq 1. \tag{22}$$

Remark 3.1 (Multi-hop routing view: exact on finite horizons). Since B_{fb} is strictly lower-triangular on a finite horizon T, it is nilpotent ($B^T_{\text{fb}} = 0$) and therefore

$$(I - B_{\text{fb}})^{-1} = \sum_{k=0}^{T-1} B^k_{\text{fb}} \quad \text{and hence} \quad s = \sum_{k=0}^{T-1} B^k_{\text{fb}} f.$$

The term $B^k_{\text{fb}} f$ aggregates contributions that traverse k internal routing steps through the feedback operator. Thus, unlike self-attention's one-hop read, the solve realizes multi-hop routing, which can produce the heavy-tail influence regimes analyzed in Section 4.2.

3.3 Positional encoding

RoPE in the forward path. In the forward attention (16) we apply RoPE to (q^f, k^f), following common practice in decoder-only Transformers (Touvron et al., 2023; Black et al., 2022). This injects relative positional information into the attention logits while preserving causal masking.

No positional encoding in feedback. We do not apply RoPE, or any other positional encoding, to the feedback attention (17). The feedback path already induces an absolute time direction: the strictly lower-triangular

feedback operator (19) and the causal solve (20) correspond to a forward substitution recurrence (22), whose output at time t depends on an iterated aggregation of the strict past. This temporal asymmetry can generate position-dependent signals even when the mixer input is time-constant.

Corollary I.8, proved in Appendix I.5, shows that a single Sessa block can produce a deterministic, position-dependent additive offset: there exist parameters and vectors $(p_t)_{t=0}^{T-1} \subset \mathbb{R}^D$ such that for all inputs x in any fixed compact set $\mathcal{D} \subset \mathbb{R}^{B_{\text{batch}} \times T \times D}$,

$$y_t = x_t + p_t, \qquad t = 0, \dots, T-1.$$

Moreover, these offsets can be chosen separated on $\mathcal{D}$ in the following sense: there exist a unit direction $u \in \mathbb{R}^D$ and a scale $\lambda > 0$ such that $p_t = c_t(\lambda u)$ with c_t pairwise distinct and the scalar ranges $\{\langle x_t + p_t, u\rangle : x \in \mathcal{D}\}$ are pairwise disjoint over t. By Corollary 4.13, the position index t is recoverable by a continuous token-wise map on the set of shifted tokens, so the feedback mechanism can supply an absolute positional signal internally.

4 Theory

This section establishes four properties of Sessa:

(i) stability of the feedback solve,
(ii) long-range memory, including flexible selective retrieval,
(iii) internal positional encoding,
(iv) universal approximation.

Remark 4.1 (LayerNorm). All stability and Jacobian statements in this section are stated for the formulation with Norm = Id. For the pre-norm LayerNorm extension relevant to universal approximation, we assume an explicit $\varepsilon > 0$ and use the corresponding Lipschitz bounds for the normalization map; see Appendix J.

4.1 Stability of the feedback solve

We isolate the operation in Sessa that induces multi-hop behavior: the causal lower-triangular solve

$$(I - B_{\text{fb}}(x))\, s \;=\; f(x), \qquad [B_{\text{fb}}]_{t,j}(x) = \gamma_t(x)\, \alpha^{\text{fb}}_{t,j}(x), \qquad [B_{\text{fb}}]_{t,j}(x) = 0 \text{ for } j \geq t, \tag{23}$$

where $\alpha^{\text{fb}}_{t,\cdot}(x)$ is a convex distribution over the strict past, $j < t$, produced by the feedback attention, and $\gamma_t(x) \in (-1, 1)$ is a bounded scalar gain. The quantity $f(x)$ is the forward aggregation defined in Section 3.

Scalar feedback matrix Throughout the stability analysis, $B_{\text{fb}}(x) \in \mathbb{R}^{T \times T}$ is scalar-valued: each entry $[B_{\text{fb}}]_{t,j}(x) \in \mathbb{R}$. The solve acts feature-wise on $s, f \in \mathbb{R}^{T \times r}$. In vectorized form, $(I_r \otimes (I - B_{\text{fb}}))\text{vec}(s) = \text{vec}(f)$.

Norms For a finite or infinite token sequence $u = (u_t)$ with $u_t \in \mathbb{R}^r$, define

$$\|u\|_{\infty,2} := \sup_t \|u_t\|_2,$$

and for a finite tensor $U \in \mathbb{R}^{T \times r}$, define $\|U\|_{\infty,2} := \max_{0 \leq t \leq T-1} \|U_t\|_2$.

Assumption 1 (Uniform row contraction on the feedback margin)**.** *For every radius $R \geq 0$ there exists $\rho(R) \in [0, 1)$ such that for all inputs x with $\|x\|_{\infty,2} \leq R$,*

$$\sup_t |\gamma_t(x)| \;\leq\; \rho(R) \;<\; 1. \tag{24}$$

Since each $\alpha^{\mathrm{fb}}_{t,\cdot}(x)$ is a convex distribution over $j < t$, Assumption 1 implies the row-sum bound

$$\sup_{t\geq 1} \sum_{j<t} |[B_{\mathrm{fb}}]_{t,j}(x)| \;\leq\; \rho(R) \;<\; 1. \tag{25}$$

Lemma 4.2 (Causal lower-triangular solve is bounded on ℓ_∞)**.** *Let B_{fb} be strictly lower-triangular, possibly on an infinite horizon, and define $(B_{\mathrm{fb}}s)_t := \sum_{j<t} [B_{\mathrm{fb}}]_{t,j} s_j$. If $\sup_t \sum_{j<t} |[B_{\mathrm{fb}}]_{t,j}| \leq \rho < 1$, then for every $f \in \ell_\infty(\mathbb{N}, \mathbb{R}^r)$ there exists a unique $s \in \ell_\infty(\mathbb{N}, \mathbb{R}^r)$ solving $(I - B_{\mathrm{fb}})s = f$, and*

$$\|s\|_{\infty,2} \;\leq\; \frac{1}{1-\rho}\, \|f\|_{\infty,2}.$$

Proof sketch. Forward substitution gives existence and uniqueness. The bound follows by a standard induction on the partial maxima $\max_{k\leq t} \|s_k\|_2$ using the row-sum estimate. See Appendix D.4. □

Proposition 2 (One-block stability bound)**.** *Fix a Sessa block G acting on finite or infinite sequences with the feedback solve (23). Assume moreover that all tokenwise affine maps appearing in the block (in particular, the output projection and the residual affine terms) are fixed and have finite operator norms and finite bias magnitudes. Assume that for every $R \geq 0$ there exist finite constants $F_R, G_R < \infty$ such that on the ball $\|x\|_{\infty,2} \leq R$,*

$$\|f(x)\|_{\infty,2} \leq F_R, \qquad \|g(x)\|_{\infty,2} \leq G_R, \qquad \sup_t |\gamma_t(x)| \leq \rho(R) < 1,$$

Here $g(x)$ denotes the tokenwise gate, the Hadamard multiplier applied to s before the output projection. Then there exists $C_R < \infty$ such that $\|G(x)\|_{\infty,2} \leq C_R$ for all $\|x\|_{\infty,2} \leq R$. In particular, G is BIBO-stable on $\ell_\infty(\mathbb{N}, \mathbb{R}^D)$.

Proof sketch. By Lemma 4.2 and (25), $\|s\|_{\infty,2} \leq (1-\rho(R))^{-1}\|f\|_{\infty,2}$. Then $\|s \odot g\|_{\infty,2} \leq \|s\|_{\infty,2}\|g\|_{\infty,2}$. Since bounded tokenwise affine maps send bounded sets to bounded sets, the output projection together with the residual affine terms yields a ball-to-ball bound for G. Appendix Proposition 25 strengthens this by giving an explicit ball-to-ball constant in terms of matrix/operator norms and bias magnitudes; see Appendix D. □

4.2 Long-range memory

We compare long-range memory through Jacobian-based diagnostics that separate the memory mechanism from routing adaptation. Let $y = G(x)$ denote the output of a causal mixer or block applied to an input token sequence $x = (x_0, \ldots, x_{T-1})$, and fix a source position $\tau \leq t$ with lag

$$\ell := t - \tau.$$

Our analysis uses three related diagnostics.

Diagnostics.

(i) *Fixed-routing influence Jacobians.* We first freeze a realized routing pattern and differentiate only the induced linear map from an injected stream to the output. This yields, for example,

$$J^{\mathrm{attn}} = \frac{\partial y}{\partial v}\Big|_{\alpha^{\mathrm{fwd}}}, \qquad J^{\mathrm{sessa}} = \frac{\partial s}{\partial f}\Big|_{B_{\mathrm{fb}}},$$

and the corresponding SSM impulse Jacobian J^{ssm} induced by a realized sequence $(A_{\mathrm{ssm},t}, B_{\mathrm{ssm},t}, C_{\mathrm{ssm},t})$. These quantities isolate the memory mechanism under a common realized routing regime.

(ii) *End-to-end block Jacobians.* We then return to the full input-dependent block and measure the actual

sensitivity of output token y_t to a past input token x_τ:

$$J^{\mathrm{e2e}}_{t,\tau}(x) := \frac{\partial y_t(x)}{\partial x_\tau}.$$

Unlike the fixed-routing Jacobians, these derivatives include both transport through the memory mechanism and the dependence of the routing coefficients on the input. They are the relevant one-block quantities for comparing diffuse attention, failed-freeze-time Mamba, and Sessa under smooth-routing assumptions.

(iii) *Scalar transport scores for deep retrieval.* For selective retrieval we extract scalar scores from deep end-to-end Jacobians. For a depth-N_{layer} stack with hidden states

$$h^{(0)} = x, \qquad h^{(1)}, \dots, h^{(N_{\mathrm{layer}})},$$

we write

$$J^{\mathrm{e2e},(N_{\mathrm{layer}})}_{t,\tau}(x) := \frac{\partial h_t^{(N_{\mathrm{layer}})}(x)}{\partial h_\tau^{(0)}(x)}.$$

Later we evaluate these blocks against source and target probes to obtain scalar transport scores, written generically as S, which are the quantities used in the selective-retrieval theorem.

These diagnostics play complementary roles. Fixed-routing Jacobians expose the structural difference between one-hop direct read, chain-structured feedback, and Sessa's many-path feedback solve. End-to-end block Jacobians capture the actual behavior of the nonlinear input-dependent block. Scalar transport scores are needed for the positive retrieval statements, since they let us compare source and distractor influence after composing end-to-end Jacobians across layers.

All decay statements in this subsection are expressed in the lag $\ell = t - \tau$, not in the context length T.

The key structural difference is that, for Sessa, the fixed-routing solve

$$(I - B_{\mathrm{fb}})^{-1}$$

aggregates contributions over multiple hop counts and, in dense regimes, over many temporal paths. This accumulation across hop counts and paths is the mechanism behind the polynomial tail analyzed below.

4.2.1 Fixed-routing Jacobians

We begin with realized routing patterns and isolate the induced memory operators. Worst-case comparisons over all inputs and parameters are uninformative, since any model can suppress a token. Instead, we compare the architectures within common diffuse-weight regimes by studying the corresponding fixed-routing influence operators.

Attention value Jacobian For causal self-attention, for a given set of attention weights $\alpha^{\mathrm{fwd}}_{t,\tau}$, the map from values to output is linear:

$$y_t = \sum_{\tau \le t} \alpha^{\mathrm{fwd}}_{t,\tau} v_\tau$$

We define the value influence Jacobian

$$J^{\mathrm{attn}}_{t,\tau} := \frac{\partial y_t}{\partial v_\tau}\Big|_{\alpha^{\mathrm{fwd}}} = \alpha^{\mathrm{fwd}}_{t,\tau}\, I_D. \tag{26}$$

Solve Jacobian In Sessa, for a given feedback matrix B_{fb}, i.e., a given routing pattern inside the loop, the lower-triangular solve

$$(I - B_{\mathrm{fb}})s = f$$

is linear in f. We define the solve influence Jacobian

$$J^{\text{sessa}} := \frac{\partial s}{\partial f}\Big|_{B_{\text{fb}}} = (I - B_{\text{fb}})^{-1}, \qquad J^{\text{sessa}}_{t,\tau} = [(I - B_{\text{fb}})^{-1}]_{t,\tau}. \tag{27}$$

Because B_{fb} is scalar-valued, the solve acts identically on each feature dimension; equivalently, if $f_t, s_t \in \mathbb{R}^{d_f}$, the full feature-block Jacobian is

$$J^{\text{sessa}}_{t,\tau} I_{d_f}.$$

SSM impulse Jacobian For a feedback recurrence $h_t = A_{\text{ssm},t} h_{t-1} + B_{\text{ssm},t} u_t$, $y_t = C_{\text{ssm},t} h_t$, given a realized sequence of transitions $(A_{\text{ssm},t}, B_{\text{ssm},t}, C_{\text{ssm},t})$, the impulse influence from u_τ to y_t is

$$J^{\text{ssm}}_{t,\tau} := C_{\text{ssm},t}\Big(\prod_{r=\tau+1}^{t} A_{\text{ssm},r}\Big) B_{\text{ssm},\tau}, \qquad 0 \le \tau \le t. \tag{28}$$

Convention: time-ordered product. We interpret the matrix product in (28) as the left-to-right time-unrolling consistent with the recurrence $h_t = A_{\text{ssm},t} h_{t-1} + \cdots$:

$$\prod_{r=\tau+1}^{t} A_{\text{ssm},r} \;:=\; A_{\text{ssm},t} A_{\text{ssm},t-1} \cdots A_{\text{ssm},\tau+1}.$$

Equivalently, the product is time-ordered with later-time factors on the left. For the empty product we use

$$\prod_{r=t+1}^{t} (\cdot) \;:=\; I,$$

so that the definition also covers the case $t = \tau$.

These Jacobians isolate the memory mechanism under a common routing regime.

4.2.2 End-to-end Jacobians

Definition 3 (End-to-end block Jacobian)**.** Let $y = G(x)$ denote the output of a single mixer/block G applied to an input token sequence $x \in (\mathbb{R}^D)^T$. We define the end-to-end Jacobian blocks by

$$J^{\text{e2e}}_{t,\tau}(x) := \frac{\partial y_t(x)}{\partial x_\tau} \in \mathbb{R}^{D \times D}.$$

For $\tau < t$, $J^{\text{e2e}}_{t,\tau}(x)$ measures long-range influence without freezing routing.

Definition 4 (Diffuse attention regime)**.** We say that an attention mechanism is in a diffuse, low-separation regime on a horizon T if, for each t, its pre-softmax logits $\beth_{t,j}$ over the visible set satisfy a bounded spread

$$\max_{j \in \mathcal{W}_t} \beth_{t,j} - \min_{j \in \mathcal{W}_t} \beth_{t,j} \le \Delta \quad \text{for some finite } \Delta,$$

uniformly over the inputs under consideration. In this regime, softmax weights are near-uniform: Appendix Lemma C.1 implies that for full-prefix attention with $|\mathcal{W}_t| = t + 1$,

$$\alpha^{\text{fwd}}_{t,j} = \Theta(1/|\mathcal{W}_t|).$$

In particular, for full-prefix causal attention one has

$$\mathcal{W}_t = \{0, \dots, t\}, \qquad |\mathcal{W}_t| = t + 1,$$

whereas for strictly-lower attention one has

$$\mathcal{W}_t = \{0, \dots, t-1\}, \qquad |\mathcal{W}_t| = t \quad \text{for } t \geq 1.$$

We state diffuse bounds in terms of the visible-set size $|\mathcal{W}_t|$ to cover full-prefix and strict-past attention uniformly.

We assume diffuse attention rows $\alpha_{t,j}^{\mathrm{fwd}} \leq c_2/|\mathcal{W}_t|$ (Definition 4), together with the following smooth-routing bound on the input set of interest:

$$\sum_{j \in \mathcal{W}_t} \Big\| \frac{\partial \alpha_{t,j}^{\mathrm{fwd}}(x)}{\partial x_\tau} \Big\|_2 \;\leq\; \frac{L_\alpha}{|\mathcal{W}_t|}, \qquad \tau < t. \tag{29}$$

Appendix B derives this from standard softmax calculus under mild logit-sensitivity control.

Lemma 4.3 (Smooth-routing for standard causal attention)**.** *Assume a single-head causal attention row is* $\alpha_{t,\cdot}^{\mathrm{fwd}}(x) = \mathrm{softmax}(\beth_{t,0}(x), \dots, \beth_{t,t}(x))$ *with logits* $\beth_{t,j}(x) = \langle q(x_t), k(x_j) \rangle$ *where* q, k *are tokenwise maps. Then for every* $\tau < t$*,*

$$\sum_{j \leq t} \Big\| \frac{\partial \alpha_{t,j}^{\mathrm{fwd}}(x)}{\partial x_\tau} \Big\|_2 \;\leq\; 2\, \alpha_{t,\tau}^{\mathrm{fwd}}(x) \Big\| \frac{\partial \beth_{t,\tau}(x)}{\partial x_\tau} \Big\|_2.$$

In particular, if $\|\partial \beth_{t,\tau}/\partial x_\tau\|_2 \leq L_\beth$ *on* $\mathcal{X}_R$*, then*

$$\sum_{j \leq t} \Big\| \frac{\partial \alpha_{t,j}^{\mathrm{fwd}}(x)}{\partial x_\tau} \Big\|_2 \;\leq\; 2L_\beth\, \alpha_{t,\tau}^{\mathrm{fwd}}(x) \;\lesssim\; \frac{1}{|\mathcal{W}_t|}$$

in the diffuse regime of Definition 4. For full-prefix attention one has $|\mathcal{W}_t| = t + 1$*. Full proof in Appendix C.1.*

4.2.3 Exponential forgetting in LTI systems

Consider a finite-dimensional linear time-invariant feedback system in state-space form:

$$h_t = A_{\mathrm{ssm}} h_{t-1} + B_{\mathrm{ssm}} u_t, \qquad y_t = C_{\mathrm{ssm}} h_t, \tag{30}$$

with constant matrices $(A_{\mathrm{ssm}}, B_{\mathrm{ssm}}, C_{\mathrm{ssm}})$. Under an impulse input at time τ, i.e. $u_\tau \neq 0$ and $u_t = 0$ for $t \neq \tau$, the contribution to y_t is mediated by $A_{\mathrm{ssm}}^{t-\tau} = A_{\mathrm{ssm}}^{\ell}$.

Proposition 3 (Exponential decay in BIBO-stable LTI feedback systems)**.** *Assume (30) is BIBO-stable. Then there exist constants* $c > 0$ *and* $\kappa \in (0, 1)$ *such that for all lags* $\ell \geq 0$*,*

$$\|C_{\mathrm{ssm}} A_{\mathrm{ssm}}^{\ell} B_{\mathrm{ssm}}\| \;\leq\; c\, \kappa^{\ell}.$$

Equivalently, the impulse response and long-range influence mediated by the state transition decay exponentially in the lag ℓ*.*

Proof sketch. BIBO stability implies internal stability of any minimal controllable and observable realization, hence $\rho_{\mathrm{spec}}(A_{\mathrm{ssm,co}}) < 1$ (Dahleh et al., 2011c). Therefore $\|A_{\mathrm{ssm,co}}^{\ell}\| \leq c\kappa^{\ell}$ and $\|C_{\mathrm{ssm}} A_{\mathrm{ssm}}^{\ell} B_{\mathrm{ssm}}\| = \|C_{\mathrm{ssm,co}} A_{\mathrm{ssm,co}}^{\ell} B_{\mathrm{ssm,co}}\| \leq c' \kappa^{\ell}$. Proof in Appendix C.3. □

4.2.4 Exponential forgetting in Mamba

Mamba-style layers fit Definition 2 as feedback systems. Their update maps $A_{\mathrm{ssm},t}(x_{0:t}), B_{\mathrm{ssm},t}(x_{0:t}), C_{\mathrm{ssm},t}(x_{0:t})$ depend on the input.

Convention: discrete scan coefficients In what follows, $A_{\mathrm{ssm},t}, B_{\mathrm{ssm},t}, C_{\mathrm{ssm},t}$ denote the discrete-time scan coefficients actually used in the recurrence $h_t = A_{\mathrm{ssm},t} h_{t-1} + B_{\mathrm{ssm},t} u_t$ after discretization, such as ZOH, unless

stated otherwise.

Exponential forgetting is not automatic for general input-dependent feedback systems. Section 4.2.6 gives a counterexample in a diffuse feedback-routing regime. For Mamba, the relevant condition is *failed freeze time*: the model cannot sustain a long interval with $\Delta_t \approx 0$.

Accumulated discretization time In Mamba's standard ZOH-diagonal parameterization, long-range influence is controlled by the accumulated discretization time

$$\sum_{r=\tau+1}^{t} \Delta_r,$$

since the transition product contains factors of the form

$$\exp\Big(-a_n \sum_{r=\tau+1}^{t} \Delta_r\Big).$$

Accordingly, failed freeze time converts control in accumulated discretization time into exponential decay in the lag.

Proposition 4 (Mamba end-to-end Jacobian bound)**.** *Consider a Mamba block with state $h_t \in \mathbb{R}^{d_{\text{state}}}$ and output $y_t \in \mathbb{R}^D$:*

$$h_{-1} = 0, \qquad h_t = A_{\text{ssm},t}(x_t)\, h_{t-1} + G_{\text{ssm},t}(x_t)\, \widetilde{B_{\text{ssm},t}}(x_t)\, u_t(x_t), \qquad y_t = C_{\text{ssm},t}(x_t)\, h_t,$$

where the parametrization is local and ZOH-diagonal: for each mode n,

$$[A_{\text{ssm},t}(x_t)]_n = \exp(-a_n \Delta_t(x_t)), \qquad [G_{\text{ssm},t}(x_t)]_n = \frac{1-\exp(-a_n \Delta_t(x_t))}{a_n},$$

with input-independent rates satisfying

$$a_n \geq \lambda > 0 \qquad \text{for all modes } n.$$

Assume there exist constants $U_R, G_{\max}, C_R, L_A, L_B, L_u < \infty$ such that for all $x \in \mathcal{X}_R$ and all t,

$$\|u_t(x_t)\| \leq U_R, \qquad \|\widetilde{B_{\text{ssm},t}}(x_t)\| \leq G_{\max}, \qquad \|C_{\text{ssm},t}(x_t)\| \leq C_R,$$

$$\Big\|\frac{\partial A_{\text{ssm},t}(x_t)}{\partial x_t}\Big\| \leq L_A, \qquad \Big\|\frac{\partial \widetilde{B_{\text{ssm},t}}(x_t)}{\partial x_t}\Big\| \leq L_B, \qquad \Big\|\frac{\partial u_t(x_t)}{\partial x_t}\Big\| \leq L_u.$$

For $\tau < t$ with lag $\ell = t - \tau$, define

$$\Pi_{t,\ell}(x) := \exp\Big(-\lambda \sum_{r=\tau+1}^{t} \Delta_r(x)\Big).$$

Then for every $x \in \mathcal{X}_R$ and every $\tau < t$,

$$\Big\|\frac{\partial y_t(x)}{\partial x_\tau}\Big\| \leq C(R)\, \Pi_{t,\ell}(x),$$

where one may take

$$C(R) := C_R\, J_R,$$

with

$$J_R := L_A\, H_R \;+\; \frac{L_A}{\lambda}\, G_{\max} U_R \;+\; \frac{1}{\lambda}(L_B\, U_R + G_{\max}\, L_u), \qquad H_R := \sqrt{d_{\text{state}}}\, \frac{G_{\max}\, U_R}{\lambda}.$$

Proof sketch. Differentiate the ZOH recurrence. By locality, for $t > \tau$ one has

$$\frac{\partial h_t}{\partial x_\tau} = A_{\mathrm{ssm},t}(x_t)\frac{\partial h_{t-1}}{\partial x_\tau}.$$

Thus the long-range dependence is controlled by the transition product. Lemma 4.4 yields the uniform state bound H_R, which controls the source-time injection derivative $\partial h_\tau/\partial x_\tau$. Since each diagonal transition satisfies

$$\Big\| \prod_{r=\tau+1}^{t} A_{\mathrm{ssm},r}(x_r) \Big\| \le \exp\Big(-\lambda \sum_{r=\tau+1}^{t} \Delta_r(x) \Big) = \Pi_{t,\ell}(x),$$

the displayed bound follows. Proof in Appendix C.4. □

ZOH discretization under freezing In Mamba, the discrete-time coefficients arise from a stable continuous-time diagonal kernel via ZOH (Gu and Dao, 2024). For each mode with continuous parameter $A = -a$ with $a > 0$ and step size $\Delta_t \ge 0$,

$$\bar{A}_t = e^{-a\Delta_t} \in [0,1], \qquad \bar{B}_t = \frac{1 - e^{-a\Delta_t}}{a}\,\widetilde{B_{\mathrm{ssm},t}}.$$

Here $A_{\mathrm{ssm},t} = \bar{A}_t$ and $\bar{B}_t u_t = G_{\mathrm{ssm},t}\,\widetilde{B_{\mathrm{ssm},t}} u_t$. In particular, when "freezing time" with $\Delta_t = 0$ one has $\bar{A}_t = 1$ and $\bar{B}_t = 0$, so the update injects no new input while holding the state.

Lemma 4.4 (Bounded state for ZOH-diagonal Mamba channels)**.** *Consider the scalar ZOH recurrence*

$$h_{-1} = 0, \qquad h_t = e^{-a\Delta_t} h_{t-1} + \frac{1 - e^{-a\Delta_t}}{a} b_t, \qquad a \ge a_{\min} > 0,\ \Delta_t \ge 0.$$

If $|b_t| \le M$ for all t, then $\sup_t |h_t| \le M/a_{\min}$. More generally, $\sup_t |h_t| \le \max\{|h_{-1}|,\ \sup_s |b_s|/a_{\min}\}$.

Proof sketch. Write $h_t = \theta_t h_{t-1} + (1-\theta_t)\frac{b_t}{a}$ with $\theta_t := e^{-a\Delta_t} \in [0,1]$. Thus h_t is a convex combination of h_{t-1} and $\frac{b_t}{a}$, yielding $|h_t| \le \max\{|h_{t-1}|,\ |b_t|/a\}$. Since $a \ge a_{\min}$, we have $|b_t|/a \le |b_t|/a_{\min}$, and the claim follows by induction. Proof in Appendix C.5. □

Failure of freeze time Mamba may slow decay by keeping $\lambda_n \Delta_t \approx 0$ over selected steps. We rule out this behavior by assuming that accumulated discretization time grows linearly on every relevant interval.

Proposition 5 (Failed freeze time yields exponential forgetting)**.** *Consider a single-mode diagonal selective SSM channel with memory factor*

$$\Pi_{t,\ell} := \prod_{r=t-\ell+1}^{t} \exp(-\lambda\,\Delta_r) = \exp\Big(-\lambda \sum_{r=t-\ell+1}^{t} \Delta_r \Big), \qquad \lambda > 0.$$

Assume there exists $c_\Delta > 0$ such that for every relevant pair $\tau < t$,

$$\sum_{r=\tau+1}^{t} \Delta_r \ge c_\Delta (t - \tau).$$

Then

$$\Pi_{t,\ell} \le \exp(-\lambda c_\Delta \ell).$$

Equivalently, once freeze time cannot be maintained over a long interval, the memory factor is exponentially small in the lag.

Proof sketch. This is immediate from

$$\Pi_{t,\ell} = \exp\Big(-\lambda \sum_{r=\tau+1}^{t} \Delta_r\Big)$$

and the assumed linear lower bound on the accumulated discretization time. Proof in Appendix C.7. □

4.2.5 Attention dilution

For causal self-attention, the direct contribution of token τ to y_t is the one-hop weight $\alpha_{t,\tau}^{\mathrm{fwd}}$. In diffuse regimes this is $O(1/|\mathcal{W}_t|)$, hence $O(1/(t+1))$ for full-prefix attention. For very old tokens with $\tau = O(1)$ and $t \asymp \ell$, this becomes $O(1/\ell)$. This is a dilution phenomenon controlled primarily by the query time t, rather than a multi-hop forgetting mechanism.

4.2.6 Polynomial decay in Sessa

We formalize a regime in which the Sessa feedback solve yields polynomial decay in the lag ℓ.

Scalar recursion Let $(\gamma_t)_{t\geq 0}$ be scalars and let $(\alpha_{t,j}^{\mathrm{fb}})_{t\geq 1,\, 0\leq j<t}$ satisfy $\alpha_{t,j}^{\mathrm{fb}} \geq 0$ and $\sum_{j<t} \alpha_{t,j}^{\mathrm{fb}} \leq 1$. Given a forward sequence (f_t), define

$$y_0 = f_0, \qquad y_t = f_t + \gamma_t \sum_{j=0}^{t-1} \alpha_{t,j}^{\mathrm{fb}} y_j, \quad t \geq 1. \tag{31}$$

For an impulse input at time τ, set $f_\tau = 1$ and $f_t = 0$ for $t \neq \tau$. This yields an influence profile y_t supported on $t \geq \tau$; the relevant memory variable is the lag $\ell = t - \tau$.

Assumption 6 (Diffuse feedback routing envelope)**.** *There exists $c_2 \in (0, \infty)$ such that for all $t \geq 1$ and all $0 \leq j < t$,*

$$\alpha_{t,j}^{\mathrm{fb}} \leq \frac{c_2}{t}. \tag{32}$$

Assumption 7 (Bounded feedback gain)**.** *There exists $\gamma_{\max} \in [0, 1)$ such that $|\gamma_t| \leq \gamma_{\max}$ for all $t \geq 0$.*

Define $\beta_{\mathrm{tail}} := 1 - \gamma_{\max} c_2$ and assume $\gamma_{\max} c_2 < 1$, so $\beta_{\mathrm{tail}} \in (0, 1]$.

Theorem 8 (Polynomial decay of impulse influence)**.** *Under Assumptions 6–7 and $\beta_{\mathrm{tail}} := 1 - \gamma_{\max} c_2 \in (0, 1]$, the impulse influence induced by (31) satisfies, for all lags $\ell \geq 1$,*

$$|y_{\tau+\ell}| \;\leq\; C\, \ell^{-\beta_{\mathrm{tail}}}, \qquad \textit{for instance} \qquad C = (1 - \beta_{\mathrm{tail}})\, e^{\,1-\beta_{\mathrm{tail}}}.$$

uniformly over the impulse time τ (when the same constants apply).

Proof sketch. Shift the recursion to start at τ and apply a comparison argument controlling partial sums by a harmonic-growth recursion, yielding $\ell^{-\beta_{\mathrm{tail}}}$. For $0 < \beta_{\mathrm{tail}} < 1$, the full proof appears in Appendix E, Corollary E.4 with $j = \tau$. The endpoint case $\beta_{\mathrm{tail}} = 1$ corresponds to $\eta = \gamma_{\max} c_2 = 0$, hence $\gamma_t = 0$ for all t and therefore $y_{\tau+\ell} = 0$ for all $\ell \geq 1$; see also Remark E.2. □

Remark 4.5 (Subcriticality). Whenever we refer in prose to a polynomial tail induced by diffuse feedback routing, this always means the subcritical regime

$$\alpha_{t,j}^{\mathrm{fb}} \leq \frac{c_2}{t}, \qquad |\gamma_t| \leq \gamma_{\max}, \qquad \gamma_{\max} c_2 < 1.$$

Equivalently,

$$\beta_{\mathrm{tail}} := 1 - \gamma_{\max} c_2 \in (0, 1].$$

The nontrivial heavy-tail case is $0 < \beta_{\text{tail}} < 1$. The endpoint $\beta_{\text{tail}} = 1$ corresponds to $\gamma_{\max} c_2 = 0$, in which case the post-source impulse is identically zero; see Remark E.2. Thus bounded gains alone do not suffice: the strict subcriticality condition $\gamma_{\max} c_2 < 1$ is essential in every use of Theorem 8.

Comparison and sharpness. Under the subcritical diffuse-routing assumptions above, Sessa yields a polynomial tail $\ell^{-\beta_{\text{tail}}}$, unlike the exponential forgetting of stable LTI feedback systems (Proposition 3) and failed-freeze-time Mamba (Section 4.2.4). The exponent is sharp: in the explicit uniform-routing regime $\alpha^{\text{fb}}_{t,j} = \frac{1}{t}\mathbf{1}[j < t]$ with constant $\gamma \in (0, 1)$, Appendix Corollary F.2 gives the closed form

$$y_{\tau+\ell} = \gamma \, \frac{\Gamma(\tau + 1)}{\Gamma(\tau + 1 + \gamma)} \cdot \frac{\Gamma(\tau + \ell + \gamma)}{\Gamma(\tau + \ell + 1)},$$

and hence $y_{\tau+\ell} = \Theta_\tau(\ell^{-\beta_{\text{tail}}})$ with $\beta_{\text{tail}} = 1 - \gamma$ for every fixed τ. Appendix Corollary F.3 further gives a uniform two-sided envelope on every bounded source family for a single layer. These one-layer statements are distinct from the deep selective-retrieval theorem below, which uses a different multi-layer construction.

Connection to attention dilution Diffuse attention in a one-hop mixer yields per-token weights of order $O(1/t)$ and, for very old tokens, $O(1/\ell)$. In contrast, under the diffuse-routing assumptions of Theorem 8, Sessa yields a tail $O(\ell^{-\beta_{\text{tail}}})$ with $\beta_{\text{tail}} < 1$, which is asymptotically slower than $1/\ell$ and therefore can mitigate dilution by sustaining longer-range influence through the stateful feedback channel while remaining BIBO-stable under Section 4.1.

Proposition 9 (Decay envelopes in the diffuse regime)**.** *Fix a horizon T and consider the fixed-routing influence Jacobians of Section 4.2.1. The three items below are stated under the mechanism-specific assumptions introduced above.*

(i) ***Transformer.*** *In the diffuse regime with full-prefix visibility, the value Jacobian satisfies*

$$\|J^{\text{attn}}_{t,\tau}\| = \alpha^{\text{fwd}}_{t,\tau} = \Theta\Big(\frac{1}{t+1}\Big) \qquad (\tau \le t),$$

and in particular for a fixed old source $\tau = O(1)$ and lag $\ell = t - \tau$,

$$\|J^{\text{attn}}_{\tau+\ell,\tau}\| = \Theta(1/\ell).$$

(ii) ***Mamba.*** *Assume the realized recurrence has diagonal transitions*

$$A_{\text{ssm},r} = \operatorname{diag}(\exp(-a_n \Delta_r)), \qquad a_n \ge \lambda > 0,$$

and bounded input/output factors $\sup_r \|B_{\text{ssm},r}\|, \sup_r \|C_{\text{ssm},r}\| < \infty$. If, on the region of interest,

$$\sum_{r=\tau+1}^{t} \Delta_r \ge c_\Delta (t - \tau),$$

then the impulse Jacobian obeys

$$\|J^{\text{ssm}}_{t,\tau}\| \le c \, \exp(-\lambda c_\Delta (t - \tau)) = c \, e^{-\lambda c_\Delta \ell}.$$

This expresses exponential forgetting under failed freeze time: the model cannot maintain a long preserve corridor, so accumulated discretization time grows linearly in the lag.

(iii) ***Sessa.*** *Under the hypotheses of Theorem 8, the solve Jacobian column corresponding to an impulse in f obeys the polynomial envelope*

$$|J^{\text{sessa}}_{\tau+\ell,\tau}| \le C \, \ell^{-\beta_{\text{tail}}}, \qquad \beta_{\text{tail}} \in (0, 1],$$

as in Theorem 8. Moreover, in the explicit uniform-routing regime $[B_{\mathrm{fb}}]_{t,j} = \begin{cases} 0, & t = 0, \\ \frac{\gamma}{t}\mathbf{1}[j < t], & t \geq 1, \end{cases}$ *with* $\gamma \in (0,1)$ *and* $\beta_{\mathrm{tail}} = 1 - \gamma$*, this envelope is tight in the following qualified sense: for every fixed source position* τ*,*

$$|J^{\mathrm{sessa}}_{\tau+\ell,\tau}| = \Theta_\tau(\ell^{-\beta_{\mathrm{tail}}}),$$

by Corollary F.2. Moreover, for every bounded source family $0 \leq \tau \leq \tau_{\max}$ *there exist constants* $c^-_{\tau_{\max}}, c^+_{\tau_{\max}} > 0$ *such that*

$$c^-_{\tau_{\max}}\, \ell^{-\beta_{\mathrm{tail}}} \leq |J^{\mathrm{sessa}}_{\tau+\ell,\tau}| \leq c^+_{\tau_{\max}}\, \ell^{-\beta_{\mathrm{tail}}}$$

for all $0 \leq \tau \leq \tau_{\max}$ *and all* $\ell \geq 1$*, by Corollary F.3. In particular, the same two-sided bound holds uniformly on every fixed finite horizon.*

Proof in Appendix C.2.

Proposition 10 (End-to-end decay envelopes)**.** *Fix a horizon* T *and consider one-block end-to-end Jacobians. In item (i) we assume the diffuse smooth-routing regime of Section 4.2.2. Assume additionally that tokenwise maps are bounded and Lipschitz on the input set:* $\|v(x_t)\| \leq V_R$ *and* $\|\partial v(x_t)/\partial x_t\| \leq L_v$*.*

(i) ***Transformer.*** *For* $y_t = \sum_{j \leq t} \alpha^{\mathrm{fwd}}_{t,j}(x)\, v(x_j)$ *and any* $\tau < t$*,*

$$\left\|\frac{\partial y_t}{\partial x_\tau}\right\| \;\leq\; \alpha^{\mathrm{fwd}}_{t,\tau} L_v \;+\; V_R \sum_{j \leq t} \left\|\frac{\partial \alpha^{\mathrm{fwd}}_{t,j}}{\partial x_\tau}\right\| \;\lesssim\; \frac{1}{t+1}.$$

In particular, for a fixed old source $\tau = O(1)$ *and lag* $\ell = t - \tau$*, one gets* $\|J^{\mathrm{e2e}}_{\tau+\ell,\tau}\| = O(1/\ell)$*.*

(ii) ***Mamba.*** *Assume the block admits a local ZOH-diagonal parametrization as in Proposition 4. If, on the input set of interest, there exists* $c_\Delta > 0$ *such that for every* $\tau < t$*,*

$$\sum_{r=\tau+1}^{t} \Delta_r(x) \geq c_\Delta(t - \tau),$$

then Corollary 4.6 yields

$$\left\|\frac{\partial y_t}{\partial x_\tau}\right\| \leq C(R)\exp(-\lambda c_\Delta(t-\tau)) = C(R)e^{-\lambda c_\Delta \ell}.$$

(iii) ***Sessa.*** *Assume additionally the hypotheses of Corollary B.7. Under the diffuse feedback routing assumptions of Appendix B,*

$$\left\|\frac{\partial y_t}{\partial x_\tau}\right\| \;\leq\; C\,\ell^{-\beta_{\mathrm{tail}}}(1 + \log(1+\ell)), \qquad \beta_{\mathrm{tail}} \in (0,1),$$

via Corollary B.7.

Proof sketch. (i) Differentiate $y_t = \sum_{j \leq t} \alpha^{\mathrm{fwd}}_{t,j}(x) v(x_j)$: one term is controlled by $\alpha^{\mathrm{fwd}}_{t,\tau} L_v$ and the other by $V_R \sum_{j \in \mathcal{W}_t} \|\partial \alpha^{\mathrm{fwd}}_{t,j}/\partial x_\tau\|$. Under the diffuse smooth-routing regime both are $O(1/|\mathcal{W}_t|)$, hence $O(1/(t+1))$ for full-prefix attention.

(ii) Combine Proposition 4 with the deterministic failed-freeze-time condition

$$\sum_{r=\tau+1}^{t} \Delta_r(x) \geq c_\Delta(t - \tau),$$

or equivalently use Corollary 4.6.

(iii) This follows from Corollary B.7 under the additional Sessa assumptions stated in item (iii).

□

Corollary 4.6 (Failed freeze time implies exponential decay of Mamba end-to-end Jacobians)**.** *Under the hypotheses of Proposition 4, assume additionally the failed freeze-time condition of Proposition 5, namely that there exists $c_\Delta > 0$ such that*

$$\sum_{r=\tau+1}^{t} \Delta_r(x) \geq c_\Delta (t-\tau)$$

for every relevant pair $\tau < t$ and every $x \in \mathcal{X}_R$. Then

$$\left\| \frac{\partial y_t(x)}{\partial x_\tau} \right\| \leq C(R) \exp(-\lambda c_\Delta (t-\tau)).$$

Proof sketch. Combine Proposition 4 with Proposition 5. □

4.2.7 Deep end-to-end bounds

The fixed-routing Jacobians remain useful as mechanism diagnostics, but deep architectural statements must be made for the *end-to-end Jacobians*

$$J_{t,\tau}^{\mathrm{e2e},(N_{\mathrm{layer}})}(x) := \frac{\partial h_t^{(N_{\mathrm{layer}})}(x)}{\partial h_\tau^{(0)}(x)} \in \mathbb{R}^{D \times D},$$

since these are the quantities that compose across layers by the chain rule. The next theorem gives the corresponding deep path-sum expansion.

Theorem 11 (Deep end-to-end aggregation)**.** *Fix a depth $N_{\mathrm{layer}} \geq 1$, a finite horizon T, and a compact input set $\mathcal{X}_0$. Let*

$$h^{(0)} = x \in \mathcal{X}_0, \qquad h^{(n_{\mathrm{layer}})} = F_{n_{\mathrm{layer}}}(h^{(n_{\mathrm{layer}}-1)}), \qquad n_{\mathrm{layer}} = 1, \dots, N_{\mathrm{layer}},$$

where each block $F_{n_{\mathrm{layer}}}$ is causal and continuously differentiable on the relevant compact set

$$\mathcal{X}_{n_{\mathrm{layer}}-1} := F_{n_{\mathrm{layer}}-1} \circ \cdots \circ F_1(\mathcal{X}_0).$$

Assume that for each layer n_{layer} there exist constants

$$d_{n_{\mathrm{layer}}} \geq 0, \qquad \lambda_{n_{\mathrm{layer}}} \geq 0,$$

and a scalar lower-triangular kernel

$$K_{n_{\mathrm{layer}}} : \{(t,\tau) : 0 \leq \tau < t \leq T-1\} \to [0,\infty)$$

such that for every $u \in \mathcal{X}_{n_{\mathrm{layer}}-1}$ and every $0 \leq \tau \leq t \leq T-1$,

$$\left\| \frac{\partial F_{n_{\mathrm{layer}},t}(u)}{\partial u_\tau} \right\| \leq d_{n_{\mathrm{layer}}} \mathbf{1}[t=\tau] + \lambda_{n_{\mathrm{layer}}} K_{n_{\mathrm{layer}}}(t,\tau) \mathbf{1}[\tau < t]. \tag{33}$$

Then for every $x \in \mathcal{X}_0$ and every $0 \leq \tau < t \leq T-1$,

$$\left\| J_{t,\tau}^{\mathrm{e2e},(N_{\mathrm{layer}})}(x) \right\| \leq \sum_{k=1}^{N_{\mathrm{layer}}} \sum_{1 \leq n_{\mathrm{layer},1} < \cdots < n_{\mathrm{layer},k} \leq N_{\mathrm{layer}}} \left(\prod_{m \notin \{n_{\mathrm{layer},1}, \dots, n_{\mathrm{layer},k}\}} d_m \right) \cdot \sum_{\tau = i_0 < i_1 < \cdots < i_k = t} \prod_{r=1}^{k} \lambda_{n_{\mathrm{layer},r}} K_{n_{\mathrm{layer},r}}(i_r, i_{r-1}). \tag{34}$$

The same expansion also gives the diagonal bound

$$\left\|J_{t,t}^{\mathrm{e2e},(N_{\mathrm{layer}})}(x)\right\| \le \prod_{n_{\mathrm{layer}}=1}^{N_{\mathrm{layer}}} d_{n_{\mathrm{layer}}}.$$

Proof sketch. This is a direct chain-rule expansion for the full block Jacobian. Proof in Appendix H. □

Thus deep long-range memory is controlled by the path sum induced by the one-block end-to-end Jacobian envelopes.

For the family-over-horizon comparison used below, one needs a horizon-uniform version of this calculus, i.e., bounds whose constants are independent of the context length T. The fixed-horizon model-class estimates and the abstract horizon-uniform lifting are recorded in Appendix H–H.5. Here we state only the resulting horizon-uniform decay envelopes needed for the comparison-class impossibility argument.

Corollary 4.7 (Horizon-uniform deep decay envelopes)**.** *Assume the hypotheses of Appendix Theorem 35.*

(i) ***Transformer.*** *Assume that for each layer n_{layer} there exists $a_{n_{\mathrm{layer}}} > 0$ such that*

$$K_{n_{\mathrm{layer}}}(t,\tau) \le \frac{a_{n_{\mathrm{layer}}}}{t+1}, \qquad \tau < t.$$

Fix a bounded source family $0 \le \tau \le \tau_{\max}$. Then for every $\ell \ge 1$,

$$\sup_{T \ge \tau_{\max}+\ell+1} \sup_{0\le\tau\le\tau_{\max}} \sup_{x \in \mathcal{X}_0^{(T)}} \left\|J_{\tau+\ell,\tau}^{\mathrm{e2e},(N_{\mathrm{layer}})}(x;T)\right\| \lesssim_{\tau_{\max},N_{\mathrm{layer}}} \frac{(\log(1+\ell))^{N_{\mathrm{layer}}-1}}{1+\ell}.$$

In particular, the right-hand side tends to 0 as $\ell \to \infty$, so this is a genuine horizon-uniform asymptotic dilution law on bounded-source families.

(ii) ***Mamba.*** *Assume that for each layer n_{layer} there exist $a_{n_{\mathrm{layer}}} > 0$ and $c_{n_{\mathrm{layer}}} > 0$ such that*

$$K_{n_{\mathrm{layer}}}(t,\tau) \le a_{n_{\mathrm{layer}}} e^{-c_{n_{\mathrm{layer}}}(t-\tau)}, \qquad \tau < t.$$

Set $c_* := \min_{n_{\mathrm{layer}}} c_{n_{\mathrm{layer}}}$. Then for every $\ell \ge 1$,

$$\sup_{T \ge \ell+1} \sup_{0\le\tau\le T-\ell-1} \sup_{x \in \mathcal{X}_0^{(T)}} \left\|J_{\tau+\ell,\tau}^{\mathrm{e2e},(N_{\mathrm{layer}})}(x;T)\right\| \lesssim_{N_{\mathrm{layer}}} (1+\ell)^{N_{\mathrm{layer}}-1} e^{-c_*\ell}.$$

In particular, this yields a genuine horizon-uniform exponential forgetting law in the lag ℓ.

(iii) ***Sessa.*** *Assume that for each layer n_{layer} there exist $a_{n_{\mathrm{layer}}} > 0$ and a common exponent $\beta_{\mathrm{tail}} \in (0,1)$ such that*

$$K_{n_{\mathrm{layer}}}(t,\tau) \le a_{n_{\mathrm{layer}}} (t-\tau)^{-\beta_{\mathrm{tail}}} (1+\log(1+t-\tau)), \qquad \tau < t.$$

Then for every $\ell \ge 1$,

$$\sup_{T \ge \ell+1} \sup_{0\le\tau\le T-\ell-1} \sup_{x \in \mathcal{X}_0^{(T)}} \left\|J_{\tau+\ell,\tau}^{\mathrm{e2e},(N_{\mathrm{layer}})}(x;T)\right\| \lesssim_{N_{\mathrm{layer}},\beta_{\mathrm{tail}}} \sum_{k=1}^{N_{\mathrm{layer}}} \ell^{k(1-\beta_{\mathrm{tail}})-1} (1+\log(1+\ell))^k.$$

In particular, if

$$N_{\mathrm{layer}}(1-\beta_{\mathrm{tail}}) < 1,$$

then the right-hand side tends to 0 as $\ell \to \infty$, yielding a genuine horizon-uniform asymptotic decay law in the lag. Outside this subcritical regime, one still retains a controlled horizon-uniform upper envelope.

Proof sketch. Apply the horizon-uniform residual calculus in Appendix Theorem 35. The Transformer, Mamba, and Sessa kernel-class estimates are proved in Appendix Propositions 32, 33, and 34, respectively. Combining those bounds yields the stated horizon-uniform envelopes. □

Consequence The fixed-horizon deep bounds are recorded in Appendix H, whereas Corollary 4.7 gives lag laws uniform in T. Thus diffuse Transformers dilute like $(\log \ell)^{N_{\text{layer}}-1}/\ell$ on bounded-source families, failed-freeze-time Mamba attenuates exponentially, and Sessa retains the stated heavy-tail upper envelope. These are upper-envelope results. They are the right tool for the impossibility statements for the comparison classes, but they do not yet yield a positive retrieval theorem for Sessa. The next subsection does.

4.2.8 Flexible finite-horizon selective retrieval

We now state the main positive memory theorem of the section. The point is not merely that Sessa admits a heavy-tail upper envelope, but that on each finite-horizon family it can realize prescribed retrieval exponents $\nu_k(\beta) = k(1-\beta) - 1$, with constants uniform in both the horizon H and the source index τ_*. For each H and τ_*, the realizing network may depend on (H, τ_*), while the retrieval-profile constants remain uniform in both parameters.

Definition 5 (Flexible finite-horizon profile realization)**.** Fix an integer $\tau_{\max} \geq 0$, an exponent $\nu \in \mathbb{R}$, and for each $H \geq 1$ a horizon

$$T_H := \tau_{\max} + H + 1.$$

Let $\mathcal{X}_0^{(H)} \subset (\mathbb{R}^D)^{T_H}$ be compact input sets satisfying the uniform bound

$$\sup_{H \geq 1} \sup_{x \in \mathcal{X}_0^{(H)}} \|x\|_{\infty,2} \leq R < \infty.$$

Let $\mathfrak{C}$ be an architecture class. We say that $\mathfrak{C}$ realizes the profile ν on the bounded source family

$$0 \leq \tau_* \leq \tau_{\max}$$

if there exist constants

$$m_- > 0, \qquad m_+ < \infty, \qquad c_- > 0,$$

independent of H and τ_*, such that for every $H \geq 1$ and every source index $\tau_* \in \{0, \dots, \tau_{\max}\}$, there exist

(i) a network $G_{H,\tau_*} \in \mathfrak{C}$ acting on $(\mathbb{R}^D)^{T_H}$,

(ii) a source probe

$$c^{(H,\tau_*)} \in \mathbb{R}^D \qquad \text{and target probes} \qquad \rho_t^{(H,\tau_*)} \in \mathbb{R}^D, \quad 0 \leq t \leq T_H - 1,$$

satisfying the normalization bounds

$$\|c^{(H,\tau_*)}\|_2 \leq 1, \qquad \|\rho_t^{(H,\tau_*)}\|_2 \leq 1 \quad (0 \leq t \leq T_H - 1),$$

(iii) the full end-to-end Jacobian blocks

$$J_{t,\tau}^{G_{H,\tau_*}}(x) := \frac{\partial G_{H,\tau_*,t}(x)}{\partial x_\tau} \in \mathbb{R}^{D \times D},$$

(iv) the scalar transport score

$$\mathsf{S}_{t,\tau}^{(H,\tau_*)}(x) := (\rho_t^{(H,\tau_*)})^\top J_{t,\tau}^{G_{H,\tau_*}}(x)\, c^{(H,\tau_*)},$$

(v) and the corresponding selective margin

$$\mathsf{M}^{(H,\tau_*)}_{t,\tau_*}(x) := \mathsf{S}^{(H,\tau_*)}_{t,\tau_*}(x) - \sum_{\substack{0\le\tau<t\\ \tau\neq\tau_*}} \bigl|\mathsf{S}^{(H,\tau_*)}_{t,\tau}(x)\bigr|.$$

These data are required to satisfy, for every $x \in \mathcal{X}_0^{(H)}$,

$$m_- \le \mathsf{M}^{(H,\tau_*)}_{\tau_*+1,\tau_*}(x) \le m_+,$$

and

$$\mathsf{M}^{(H,\tau_*)}_{\tau_*+\ell,\tau_*}(x) \ge c_-(1+\ell)^\nu, \qquad 1 \le \ell \le H.$$

Theorem 12 (Flexible finite-horizon selective retrieval for deep Sessa)**.** *Work in the identity-normalized formulation with the exact GELU activation*

$$\mathrm{GELU}(z) = z\,\Phi(z),$$

and assume

$$D \ge 7.$$

Fix

$$\beta \in (0,1), \qquad k \ge 1, \qquad \tau_{\max} \ge 0,$$

and define

$$\nu_k(\beta) := k(1-\beta) - 1.$$

Let $\{\mathcal{X}_0^{(H)}\}_{H\ge1}$ be a uniformly bounded family of compact sets as in Definition 5. Then the class of LN-free Sessa networks realizes the profile $\nu_k(\beta)$ on the bounded source family $0 \le \tau_* \le \tau_{\max}$ in the sense of Definition 5.

More precisely, there exist constants

$$m_- > 0, \qquad m_+ < \infty, \qquad c_- > 0,$$

depending only on $(k, \beta, \tau_{\max}, R)$, but independent of H and τ_*, such that for every $H \ge 1$ and every $\tau_* \in \{0, \dots, \tau_{\max}\}$, there exist a finite-depth LN-free Sessa network

$$G_{H,\tau_*} : (\mathbb{R}^D)^{T_H} \to (\mathbb{R}^D)^{T_H}$$

and a scalar channel score $\mathsf{S}^{(H,\tau_*)}$ with selective margin $\mathsf{M}^{(H,\tau_*)}$ such that for every $x \in \mathcal{X}_0^{(H)}$,

$$m_- \le \mathsf{M}^{(H,\tau_*)}_{\tau_*+1,\tau_*}(x) \le m_+,$$

and

$$\mathsf{M}^{(H,\tau_*)}_{\tau_*+\ell,\tau_*}(x) \ge c_-(1+\ell)^{\nu_k(\beta)}, \qquad 1 \le \ell \le H.$$

Consequently: if $\nu_k(\beta) < 0$, deep Sessa realizes a decaying profile; if $\nu_k(\beta) = 0$, it realizes a frozen profile; and if $\nu_k(\beta) > 0$, it realizes an increasing profile.

Proof sketch. Composite architecture. Fix $H \ge 1$ and $0 \le \tau_* \le \tau_{\max}$. Set

$$L_H := \tau_{\max} + H, \qquad T_H := L_H + 1.$$

We construct

$$G_{H,\tau_*} = M_{H,k} \circ \cdots \circ M_{H,1} \circ S_{H,\tau_*,\varepsilon_H} \circ Q_H \circ P_H.$$

Here P_H writes a strictly ordered positional code, Q_H is a signal-transparent preparatory network producing the

power profile, $S_{H,\tau_*,\varepsilon_H}$ selects the source τ_*, and $M_{H,1},\dots,M_{H,k}$ are diffuse profile-compensated macro-layers.

By Corollaries 4.11 and 4.12, P_H writes a strictly ordered positional code on e_{pos} while remaining transparent to perturbations along e_{sig}. Corollary K.21 yields a constant-depth network Q_H that preserves the signal and positional channels and writes a profile

$$r_t \asymp (t+1)^{1-\beta}.$$

Lemma K.12 yields a selector with gain $\asymp 1$ at τ_* and off-target suppression $\varepsilon_H \asymp (H+1)^{-1}$. Lemma K.22 yields macro-layers whose selected-channel transport has kernel size $\asymp (i+1)^{-\beta}$.

Appendix Lemma K.9 identifies the selected-channel transport of the post-preparatory stack with the actual Jacobian score. The desired lower bound follows by restricting to balanced k-jump paths and applying Lemma K.25, while the competitor contribution is controlled by Lemma K.26. Choosing the construction constants appropriately makes the competitor mass absorbable for all $1 \le \ell \le H$, yielding the stated anchored bounds. □

Corollary 4.8 (Flexible frozen and increasing profiles require depth)**.** *Under Theorem 12:*

(i) for $k = 1$, one has

$$\nu_1(\beta) = -\beta < 0,$$

so only decaying profiles occur;

(ii) for $k \ge 2$ and

$$\beta = 1 - \frac{1}{k},$$

one gets the frozen profile $\nu_k(\beta) = 0$;

(iii) for $k \ge 2$ and

$$0 < \beta < 1 - \frac{1}{k},$$

one gets the increasing profile $\nu_k(\beta) > 0$.

4.2.9 Impossibility for the comparison classes in the same flexible finite-horizon regime

This is the matching negative statement in the same family-over-H regime. By the horizon-uniform end-to-end envelopes from Section 4.2.7, diffuse fixed-depth Transformers and failed-freeze-time fixed-depth Mamba admit only decaying upper bounds, so they cannot realize frozen or increasing retrieval profiles.

Proposition 13 (Comparison-class impossibility for flexible selective retrieval)**.** *Fix $\tau_{\max} \ge 0$, and let*

$$T_H = \tau_{\max} + H + 1.$$

Assume we are given, for every $H \ge 1$ and every $\tau_* \in \{0,\dots,\tau_{\max}\}$, a network

$$G^{\mathrm{comp}}_{H,\tau_*}$$

from one of the following two comparison classes: a depth-L causal Transformer in the diffuse smooth-routing regime, or a depth-L causal Mamba stack in the failed-freeze-time regime.

Assume moreover that, in the Transformer case, the family satisfies the hypotheses of Corollary 4.7, item (i), with constants independent of H and τ_*, and that, in the Mamba case, the family satisfies the hypotheses of Corollary 4.7, item (ii), with constants independent of H and τ_*.

Then no such comparison-class family can realize a frozen or increasing profile in the sense of Definition 5. More precisely:

(i) ***Transformer.*** *There do not exist constants $m_- > 0$, $m_+ < \infty$, $c_- > 0$, and $\nu \ge 0$, independent of H and*

τ_*, *such that*

$$m_- \le \mathsf{M}^{(H,\tau_*)}_{\tau_*+1,\tau_*}(x) \le m_+,$$

and

$$\mathsf{M}^{(H,\tau_*)}_{\tau_*+\ell,\tau_*}(x) \ge c_-(1+\ell)^\nu, \qquad 1 \le \ell \le H,$$

hold uniformly for all H, τ_*, x.

(ii) ***Mamba.*** *The same impossibility holds for failed-freeze-time Mamba families.*

Proof. Assume toward a contradiction that such a realization exists. By Definition 5, the probes satisfy

$$\|c^{(H,\tau_*)}\|_2 \le 1, \qquad \|\rho_t^{(H,\tau_*)}\|_2 \le 1.$$

Hence for every admissible H, τ_*, x, t, τ,

$$\big|\mathsf{S}^{(H,\tau_*)}_{t,\tau}(x)\big| = \big|(\rho_t^{(H,\tau_*)})^\top J^{G^{\mathrm{comp}}_{H,\tau_*}}_{t,\tau}(x)\, c^{(H,\tau_*)}\big| \le \big\|J^{G^{\mathrm{comp}}_{H,\tau_*}}_{t,\tau}(x)\big\|.$$

Therefore

$$\mathsf{M}^{(H,\tau_*)}_{t,\tau_*}(x) \le \big|\mathsf{S}^{(H,\tau_*)}_{t,\tau_*}(x)\big| \le \big\|J^{G^{\mathrm{comp}}_{H,\tau_*}}_{t,\tau_*}(x)\big\|.$$

For Transformers, Corollary 4.7, item (i), applied to the family $G^{\mathrm{comp}}_{H,\tau_*}$, gives the horizon-uniform bounded-source-family envelope

$$\big\|J^{G^{\mathrm{comp}}_{H,\tau_*}}_{\tau+\ell,\tau}(x)\big\| \lesssim \frac{(\log(1+\ell))^{L-1}}{1+\ell},$$

uniformly over all admissible H, τ_*, x and all $0 \le \tau \le \tau_{\max}$. This tends to 0 as $\ell \to \infty$.

For Mamba, item (ii) gives

$$\big\|J^{G^{\mathrm{comp}}_{H,\tau_*}}_{\tau+\ell,\tau}(x)\big\| \lesssim (1+\ell)^{L-1} e^{-c\ell},$$

uniformly over all admissible H, τ_*, x, τ. This also tends to 0.

Since a frozen or increasing profile would require

$$\mathsf{M}^{(H,\tau_*)}_{\tau_*+\ell,\tau_*}(x) \ge c_-(1+\ell)^\nu \qquad (\nu \ge 0),$$

uniformly in all admissible H, τ_*, x, ℓ, this is impossible in either comparison class. □

Corollary 4.9 (Flexible selective retrieval separates Sessa from the comparison classes)**.** *In the regime of Definition 5:*

(i) deep identity-normalized Sessa realizes the full exponent family

$$\nu_k(\beta) = k(1-\beta) - 1;$$

(ii) diffuse fixed-depth Transformers and failed-freeze-time fixed-depth Mamba do not realize frozen or increasing profiles.

Thus, in this uniform finite-horizon family-over-H regime, deep Sessa supports flexible selective retrieval, whereas the two comparison classes do not.

4.3 Internal positional encoding

Sessa does not require an explicit absolute positional embedding in the feedback branch. The key point is that the feedback solve can itself write a separated absolute positional signal. The main lemma gives this positional writer, and the corollaries record the two refinements used later: one-directional writing with signal transparency, and continuous recovery of the position index.

Lemma 4.10 (Feedback generates ordered separated positional codes)**.** *Fix $T \geq 2$ and model width $m \geq 1$. There exists a single width-m Sessa block $G^{(1)}$ and vectors $p_0, \dots, p_{T-1} \in \mathbb{R}^m$ such that for all token sequences $h \in \mathbb{R}^{T \times m}$,*

$$G^{(1)}(h)_t = h_t + p_t, \qquad t = 0, \dots, T-1.$$

Moreover, for any compact $\mathcal{K}\_\text{set} \subset \mathbb{R}^{T \times m}$ the offsets can be chosen so that there exist a unit direction $u \in \mathbb{R}^m$ and pairwise disjoint compact intervals

$$J_0 < J_1 < \cdots < J_{T-1} \subset (0, \infty)$$

with

$$\langle h_t + p_t, u \rangle \in J_t \qquad \text{for all } h \in \mathcal{K}\_\text{set},\ t = 0, \dots, T-1.$$

Proof sketch. Choose parameters so that the mixer input is constant, the forward branch produces a constant forward signal, and the feedback routing is chosen so that the induced scalar solve generates a deterministic strictly increasing sequence on the finite prefix. Project that scalar sequence onto a chosen direction, then shift and rescale it so that the resulting compact scalar ranges are pairwise disjoint, strictly ordered, and contained in $(0, \infty)$. See Appendix I.5. □

Corollary 4.11 (One-directional internal positional writer)**.** *Under the hypotheses of Lemma 4.10, the block can be chosen so that there exists a unit direction $e_{\text{pos}} \in \mathbb{R}^m$ and scalars $\lambda_0, \dots, \lambda_{T-1}$ with*

$$G^{(1)}(h)_t = h_t + \lambda_t e_{\text{pos}}, \qquad t = 0, \dots, T-1,$$

for all token sequences $h \in \mathbb{R}^{T \times m}$. Moreover, for any compact $\mathcal{K}\_\text{set} \subset \mathbb{R}^{T \times m}$, the same block can be chosen so that there exist pairwise disjoint compact intervals

$$J_0 < J_1 < \cdots < J_{T-1} \subset (0, \infty)$$

with

$$\langle G^{(1)}(h)_t, e_{\text{pos}} \rangle \in J_t \qquad \text{for all } h \in \mathcal{K}\_\text{set},\ t = 0, \dots, T-1.$$

Proof. In the construction underlying Lemma 4.10, the deterministic scalar sequence generated by the feedback solve is written onto a chosen output direction. Choosing that output direction to be e_{pos} and writing no offset on the orthogonal complement yields the form

$$G^{(1)}(h)_t = h_t + \lambda_t e_{\text{pos}}.$$

The interval-separation conclusion is exactly the same as in Lemma 4.10. □

Corollary 4.12 (Signal transparency of the one-directional positional writer)**.** *Under the hypotheses of Corollary 4.11, let $e_{\text{sig}} \in \mathbb{R}^m$ be any unit vector with*

$$e_{\text{sig}} \perp e_{\text{pos}}.$$

Then for every token sequence $h \in \mathbb{R}^{T \times m}$, every source index $\tau \in \{0, \dots, T-1\}$, and every scalar $a \in \mathbb{R}$,

$$G^{(1)}\big(h + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau]\big)_t = G^{(1)}(h)_t + a\, e_{\text{sig}} \mathbf{1}[t = \tau], \qquad t = 0, \dots, T-1.$$

In particular,

$$\langle G^{(1)}(h + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau])_t, e_{\text{pos}} \rangle = \langle G^{(1)}(h)_t, e_{\text{pos}} \rangle \qquad \forall\, t,$$

so perturbations along e_{sig} leave the internally written positional coordinate unchanged.

Proof. By Corollary 4.11,

$$G^{(1)}(h)_t = h_t + \lambda_t e_{\text{pos}}.$$

Therefore

$$G^{(1)}(h + a\, e_{\text{sig}}\mathbf{1}[\cdot = \tau])_t = h_t + a\, e_{\text{sig}}\mathbf{1}[t = \tau] + \lambda_t e_{\text{pos}} = G^{(1)}(h)_t + a\, e_{\text{sig}}\mathbf{1}[t = \tau].$$

Since $e_{\text{sig}} \perp e_{\text{pos}}$, taking the e_{pos}-coordinate gives the second claim. □

Corollary 4.13 (Continuous recovery of the position index)**.** *Under the hypotheses of Corollary 4.11, fix a compact set*

$$\mathcal{K}\_\text{set} \subset \mathbb{R}^{T\times m},$$

and choose the block so that there exist pairwise disjoint compact intervals

$$J_0 < J_1 < \cdots < J_{T-1} \subset (0, \infty)$$

with

$$\langle G^{(1)}(h)_t, e_{\text{pos}}\rangle \in J_t \qquad \forall\, h \in \mathcal{K}\_\text{set},\ \forall\, t = 0, \dots, T-1.$$

Then there exists a continuous map

$$\psi : \mathbb{R}^m \to \mathbb{R}$$

such that

$$\psi(G^{(1)}(h)_t) = t \qquad \forall\, h \in \mathcal{K}\_\text{set},\ \forall\, t = 0, \dots, T-1.$$

In particular, the position index t is recoverable by a continuous tokenwise map on the shifted-token set

$$\bigcup_{t=0}^{T-1} \{G^{(1)}(h)_t :\ h \in \mathcal{K}\_\text{set}\}.$$

Proof. Write each compact interval as

$$J_t = [a_t, b_t].$$

Since the intervals are pairwise disjoint and ordered, one has

$$b_t < a_{t+1} \qquad (t = 0, \dots, T-2).$$

Define a continuous function $g : \mathbb{R} \to \mathbb{R}$ by requiring

$$g(s) = t \qquad \text{for all } s \in J_t,$$

interpolating linearly on each gap $[b_t, a_{t+1}]$, and extending constantly on $(-\infty, a_0]$ and $[b_{T-1}, \infty)$. Then g is continuous on $\mathbb{R}$ and satisfies $g|_{J_t} \equiv t$ for every t.

Now define

$$\psi(z) := g(\langle z, e_{\text{pos}}\rangle), \qquad z \in \mathbb{R}^m.$$

Since $z \mapsto \langle z, e_{\text{pos}}\rangle$ is continuous, ψ is continuous. Moreover, for every $h \in \mathcal{K}\_\text{set}$ and every t,

$$\langle G^{(1)}(h)_t, e_{\text{pos}}\rangle \in J_t,$$

hence

$$\psi(G^{(1)}(h)_t) = g(\langle G^{(1)}(h)_t, e_{\text{pos}}\rangle) = t.$$

□

Consequence Sessa can internally generate an absolute positional code through feedback, even when the forward branch uses only relative-position-aware routing such as RoPE.

4.4 Universal approximation of causal maps

We state a universal approximation result for Sessa networks on compact domains, in the standard causal decoder setting. Since intermediate constructions may require an internal width $m \geq D$, we state the result for Sessa with tokenwise linear adapters $D \to m \to D$.

Definition 6 (Causality)**.** A map $F : \mathcal{D} \to \mathbb{R}^{T\times D}$ is causal if for every t and all $x, x' \in \mathcal{D}$, $x_{0:t} = x'_{0:t}$ implies $F(x)_t = F(x')_t$.

Theorem 14 (Universal approximation by concrete Sessa with adapters)**.** *Let $\mathcal{D} \subset \mathbb{R}^{T\times D}$ be compact and let $F : \mathcal{D} \to \mathbb{R}^{T\times D}$ be continuous and causal. Then for any $\varepsilon > 0$ there exist an even query/key width $d_k \geq 2$, a model width $m \geq D$, tokenwise adapters*

$$\mathrm{Embed} : \mathbb{R}^D \to \mathbb{R}^m, \qquad \mathrm{Unembed} : \mathbb{R}^m \to \mathbb{R}^D,$$

and a finite-depth width-m concrete Sessa network G such that

$$\sup_{x\in\mathcal{D}} \Big\| F(x) - \mathrm{Unembed}(G(\mathrm{Embed}(x))) \Big\|_F < \varepsilon.$$

Proof sketch. (i) Use a single Sessa block to write an internal positional code.
(ii) Use a finite stack of concrete Sessa blocks to encode each relevant causal prefix into dedicated internal coordinates.
(iii) Apply a finite tokenwise readout stack, again implemented by concrete Sessa blocks, to approximate the desired causal output on the resulting compact encoded-state set.

Details appear in Appendix I, in the proof of Theorem 14. □

5 Experiments

We compare three model variants that share the same decoder macro-architecture and training setup and differ only in the sequence mixer. The mixers are Sessa mixer, multi-head self-attention, and Mamba2 mixer. We match parameter count, use the same optimizer and training schedule, and train all models for the same number of optimization steps.

We do not report aggregate results on the full Long Range Arena (LRA) suite (Tay et al., 2021). Although LRA was originally proposed as a testbed for long-range dependencies, subsequent analyses have highlighted several issues suggesting that strong performance on LRA can be confounded by factors unrelated to robust long-context reasoning. (Tay et al., 2021; Miralles-González et al., 2025) We evaluate long-context behavior on SymbolSoup and Diffuse MQAR, and short-context language modeling on SimpleStories. (Finke et al., 2025; SimpleStories Project, 2025)

5.1 Synthetic long-range tasks

5.1.1 Datasets and tasks

SymbolSoup. SymbolSoup is a long-range classification dataset with two informative stylized blocks separated by label-independent noise. Each example contains three noise blocks and two stylized blocks, one from each style family. The order of the two stylized blocks is randomized.

> noise `<sep1>` first/second stylized part `<sep2>` noise `<sep1>` second/first stylized part `<sep2>` noise `<sep>` `<label>`.

The label is the pair of styles used in the two stylized blocks. Stylized blocks are generated by a Markov-like process with unigram and bigram preferences and occasional motif insertion plus small symbol noise.

Diffuse MQAR. We additionally evaluate on a modified multi-query associative recall benchmark based on MQAR (Arora et al., 2024). Relative to the original formulation, our variant uses multi-token keys, structured distractors with shared prefixes and mismatched suffixes, and explicit control of the source–query lag. Each example contains a prefix memory block of key–value pairs, a noise block populated with distractor key–value-like patterns, and a terminal query block. The test split includes retrieval lags up to 4× larger than those seen during training.

Table 1: Long-context test results (mean ± std over 2 seeds). For SymbolSoup we report classification accuracy; for Diffuse MQAR we report token accuracy.

Model	SymbolSoup Acc ↑	Diffuse MQAR Token Acc ↑
Sessa	0.8601 ± 0.0016	0.1541 ± 0.0071
Transformer	0.7921 ± 0.0070	0.1222 ± 0.0003
Mamba2	0.0500 ± 0.0000	0.0021 ± 0.0000

Mamba-2 did not converge on SymbolSoup or Diffuse MQAR. We view this as qualitatively consistent with our selective-SSM theory: when noise makes the selection signal weakly separable, the resulting non-vanishing freeze-time errors restore exponential attenuation of long-range influence, as formalized in Proposition 5 and Corollary 4.6. This interpretation is relevant to Mamba-2 because it is itself a selective SSM, specifically a scalar-identity restricted variant in the SSD framework (Dao and Gu, 2024).

5.2 SimpleStories language modeling

5.2.1 Dataset and task

For the short-context regime we use a SimpleStories corpus of short, synthetically generated stories. Each story is written in simplified English with a small vocabulary and constrained syntax.

We treat this corpus as a causal language modeling benchmark. The text is tokenized with a subword tokenizer shared across all architectures, and training sequences are formed by concatenating stories and splitting them into fixed-length segments. The model predicts the next token at each position using a left-to-right mask. We report validation perplexity.

Table 2: SimpleStories test results (mean ± std over 2 seeds).

Model	Perplexity ↓	Top-1 acc ↑	Top-5 acc ↑
Transformer	7.6701 ± 0.0313	$50.441 \pm 0.059\%$	$78.497 \pm 0.062\%$
Mamba2	7.7229 ± 0.0207	$50.299 \pm 0.046\%$	$78.302 \pm 0.043\%$
Sessa	8.3700 ± 0.0482	$49.144 \pm 0.081\%$	$77.119 \pm 0.090\%$

We hypothesize that the small performance drop of Sessa in the short-context regime is due to the feedback mechanism being less necessary for this task. Under matched parameter count, a portion of Sessa's capacity is allocated to the feedback branch, which may be weakly utilized on short-context. To test this interpretation, we ran a control experiment with the feedback branch removed while keeping the remainder of the architecture unchanged. The ablated model improves over full Sessa on SimpleStories, reducing test perplexity from 8.3700 ± 0.0482 to 8.0902 ± 0.0192 and increasing top-1 accuracy from $49.144 \pm 0.081\%$ to $49.648 \pm 0.026\%$. This supports the view that feedback is less beneficial in the short-context regime, while remaining consistent with Sessa's stronger results on long-context tasks, where feedback appears to be more useful.

6 Discussion

The main comparison in this paper is not between favorable operating regimes of Transformers, Mamba, and Sessa, but between matched regimes in which sharp retrieval is unavailable. For attention, this appears as diffuse, low-separation routing, so the selector cannot concentrate mass on a small set of relevant indices. For Mamba,

the analogous failure is failed freeze time, so the model cannot maintain a long preserve corridor on the relevant interval. These are natural failure regimes for the respective architectures, and they provide a common basis for comparison.

In this matched setting, the difference comes from the memory mechanism rather than from access to sharp routing. Diffuse attention remains one-hop and therefore suffers dilution. Failed-freeze-time Mamba remains chain-structured and therefore exhibits exponential attenuation. Sessa is also studied in a diffuse regime, but its feedback solve aggregates influence over multiple hop counts and, in dense settings, over many temporal paths. This is the structural source of its slower long-range decay.

The main separation is not only in the polynomial tail, but in the selective-retrieval result. In the same family-over-H regime, deep Sessa realizes flexible selective retrieval profiles, whereas diffuse fixed-depth Transformers and failed-freeze-time fixed-depth Mamba do not realize frozen or increasing profiles. Thus the separation is not merely quantitative at the level of decay rates; it is qualitative at the level of what retrieval behavior the architectures can realize under the same matched breakdown of sharp retrieval.

The broader point is that long-context behavior depends not only on how routing coefficients are produced, but also on how they are composed over time. When sharp retrieval fails, as can become increasingly likely as context length grows, this distinction becomes decisive. In that regime, Sessa can still support flexible selective retrieval through its multi-hop feedback structure.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. doi: 10.48550/arXiv.2403.07815. URL https://openreview.net/forum?id=gerNCVqqtR. Accepted by TMLR (OpenReview); arXiv:2403.07815.

Panos J. Antsaklis and Anthony N. Michel. *Linear Systems*. Birkhäuser, Boston, 1 edition, 2006.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2312.04927. URL https://openreview.net/forum?id=LY3ukUANko. ICLR 2024 poster; arXiv:2312.04927.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. 2016. doi: 10.48550/arXiv.1607.06450. URL https://arxiv.org/abs/1607.06450.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2006.11477. arXiv:2006.11477.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. doi: 10.48550/arXiv.2004.05150. URL https://arxiv.org/abs/2004.05150.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9/.

Rishi Bommasani et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. doi: 10.48550/arXiv.2108.07258. URL https://arxiv.org/abs/2108.07258. Stanford CRFM report.

Tom B. Brown et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

Aydar Bulatov, Yuri Kuratov, and Mikhail Burtsev. Recurrent memory transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.48550/arXiv.2207.06881. URL https://arxiv.org/abs/2207.06881. NeurIPS 2022; arXiv:2207.06881.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. doi: 10.48550/arXiv.1904.10509. URL https://arxiv.org/abs/1904.10509.

Mohammed Dahleh, Munther A. Dahleh, and George Verghese. Lectures on dynamic systems and control, chapter 15: External input-output stability. MIT OpenCourseWare (6.241J/16.338J), course notes, 2011a. URL https://ocw.mit.edu/courses/6-241j-dynamic-systems-and-control-spring-2011/5b744a33f5db9b0cc70dbc04a9de5706_MIT6_241JS11_chap15.pdf.

Mohammed Dahleh, Munther A. Dahleh, and George Verghese. Lectures on dynamic systems and control, chapter 27: Poles and zeros of mimo systems. MIT OpenCourseWare (6.241J/16.338J), course notes, 2011b. URL https://ocw.mit.edu/courses/6-241j-dynamic-systems-and-control-spring-2011/8a8013268491f54fd65614f299a05290_MIT6_241JS11_chap27.pdf.

Mohammed Dahleh, Munther A. Dahleh, and George Verghese. Lectures on dynamic systems and control, chapter 30: Minimality and stability of interconnected systems. MIT OpenCourseWare (6.241J/16.338J), course notes, 2011c. URL https://ocw.mit.edu/courses/6-241j-dynamic-systems-and-control-spring-2011/5e41c2e287bde74f5326d258e89c951c_MIT6_241JS11_chap30.pdf.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285/. arXiv:1901.02860; doi:10.48550/arXiv.1901.02860.

Hugo Dalla-Torre et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025. doi: 10.1038/s41592-024-02523-z. URL https://www.nature.com/articles/s41592-024-02523-z. Version of record published online 28 Nov 2024; issue date Feb 2025.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10041–10071. PMLR, 2024. doi: 10.48550/arXiv.2405.21060. URL https://proceedings.mlr.press/v235/dao24a.html. ICML 2024; introduces Mamba-2 via the SSD framework; arXiv:2405.21060.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023. doi: 10.48550/arXiv.2307.02486. URL https://arxiv.org/abs/2307.02486.

Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

Angela Fan, Thibaut Lavril, Edouard Grave, Armand Joulin, and Sainbayar Sukhbaatar. Addressing some limitations of transformers with feedback memory. *arXiv preprint arXiv:2002.09402*, 2020. doi: 10.48550/arXiv.2002.09402. URL https://arxiv.org/abs/2002.09402. OpenReview submission notes it was under review for ICLR 2021.

Lennart Finke, Chandan Sreedhara, Thomas Dooms, Mat Allen, Emerald Zhang, Juan Diego Rodriguez, Noa Nabeshima, Thomas Marshall, and Dan Braun. Parameterized synthetic text generation with simplestories. In *NeurIPS 2025 Datasets and Benchmarks Track*, 2025. doi: 10.48550/arXiv.2504.09184. URL https://openreview.net/forum?id=sVh3eQ642W. NeurIPS 2025 Datasets and Benchmarks Track poster (OpenReview); arXiv:2504.09184.

Walter Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *Journal of Mathematics and Physics*, 38:77–81, 1959. doi: 10.1002/sapm195938177.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. doi: 10.48550/arXiv.2312.00752. URL https://openreview.net/forum?id=tEYskw1VY2. COLM 2024 (OpenReview); arXiv:2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022a. doi: 10.48550/arXiv.2111.00396. URL https://arxiv.org/abs/2111.00396. arXiv:2111.00396.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b. doi: 10.48550/arXiv.2206.11893. URL https://arxiv.org/abs/2206.11893. arXiv:2206.11893; introduces S4D.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). 2016. doi: 10.48550/arXiv.1606.08415. URL https://arxiv.org/abs/1606.08415.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. ISBN 9780521839402. doi: 10.1017/CBO9781139020411.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc V. Le. Transformer quality in linear time. In *Proceedings of the*

39th International Conference on Machine Learning (ICML), volume 162 of *Proceedings of Machine Learning Research*, pages 9099–9117. PMLR, 2022. URL https://proceedings.mlr.press/v162/hua22a.html.

Ningyuan Huang, Miguel Sarabia, Abhinav Moudgil, Pau Rodriguez, Luca Zappella, and Federico Danieli. Understanding input selectivity in mamba: Impact on approximation power, memorization, and associative recall capacity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 25693–25727. PMLR, 2025. doi: 10.48550/arXiv.2506.11891. URL https://proceedings.mlr.press/v267/huang25ab.html. ICML 2025; arXiv:2506.11891.

DeLesley Scott Hutchins, Imanol Schlag, Yuhuai Wu, Ethan Dyer, and Behnam Neyshabur. Block-recurrent transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.48550/arXiv.2203.07852. URL https://arxiv.org/abs/2203.07852. NeurIPS 2022; arXiv:2203.07852.

Dongseong Hwang, Weiran Wang, Zhuoyuan Huo, Khe Chai Sim, and Pedro Moreno Mengibar. Transformerfam: Feedback attention is working memory. *arXiv preprint arXiv:2404.09173*, 2024. doi: 10.48550/arXiv.2404.09173. URL https://arxiv.org/abs/2404.09173.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82 (1):35–45, 1960. doi: 10.1115/1.3662552.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. doi: 10.1016/S0893-6080(05)80131-5.

Pablo Miralles-González, Javier Huertas-Tato, Alejandro Martín, and David Camacho. On the locality bias and results in the long range arena. *arXiv preprint arXiv:2501.14850*, 2025. doi: 10.48550/arXiv.2501.14850. URL https://arxiv.org/abs/2501.14850.

Timur Mudarisov, Mikhail Burtsev, Tatiana Petrova, and Radu State. Limitations of normalization in attention mechanism. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025. doi: 10.48550/arXiv.2508.17821. URL https://arxiv.org/abs/2508.17821. NeurIPS 2025 poster (OpenReview id: 16kX08MCav); arXiv:2508.17821v2 (revised 20 Oct 2025).

Markus N. Rabe and Charles Staats. Self-attention does not need $O(n^2)$ memory. *arXiv preprint arXiv:2112.05682*, 2021. doi: 10.48550/arXiv.2112.05682. URL https://arxiv.org/abs/2112.05682.

Noam Shazeer. Glu variants improve transformer. 2020. doi: 10.48550/arXiv.2002.05202. URL https://arxiv.org/abs/2002.05202.

SimpleStories Project. SimpleStories/SimpleStories. Hugging Face Datasets, 2025. URL https://huggingface.co/datasets/SimpleStories/SimpleStories. Accessed: 2026-01-29.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. 2021. doi: 10.48550/arXiv.2104.09864. URL https://arxiv.org/abs/2104.09864.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k. arXiv:2011.04006.

Heinrich Tietze. über funktionen, die auf einer abgeschlossenen menge stetig sind. *Journal für die reine und angewandte Mathematik*, 145:9–14, 1915. doi: 10.1515/crll.1915.145.9.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. 2023. doi: 10.48550/arXiv.2302.13971. URL https://arxiv.org/abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5998–6008, 2017. doi: 10.48550/arXiv.1706.03762. URL https://arxiv.org/abs/1706.03762.

Ruibin Xiong, Yunchang Yang, Di He, et al. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. URL https://proceedings.mlr.press/v119/xiong20b.html.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. doi: 10.48550/arXiv.2007.14062. URL https://arxiv.org/abs/2007.14062. arXiv:2007.14062.

Appendix

A Definitions and notation

A.1 Sequence norms and bounded-input sets

Definition 7 (Sup–ℓ_2 norm and bounded-input balls)**.** Fix a horizon $T \in \mathbb{N}^*$ and token width $D \in \mathbb{N}^*$. For a finite sequence $x = (x_0, \dots, x_{T-1}) \in (\mathbb{R}^D)^T$ define

$$\|x\|_{\infty,2} := \max_{0 \le t \le T-1} \|x_t\|_2.$$

For $R \ge 0$ define the ball

$$\mathcal{X}_R := \{x \in (\mathbb{R}^D)^T : \ \|x\|_{\infty,2} \le R\}.$$

For infinite sequences $(x_t)_{t \ge 0}$ we use the analogous norm $\|x\|_{\infty,2} := \sup_{t \ge 0} \|x_t\|_2 \in [0, \infty]$.

$$\|X\|_{\infty,2} \le \|X\|_F \le \sqrt{T}\, \|X\|_{\infty,2} \qquad \text{for } X \in \mathbb{R}^{T \times D}. \tag{35}$$

A.2 BIBO stability on ℓ_∞

Definition 8 (BIBO stability on ℓ_∞)**.** A map $\mathcal{N} : \ell_\infty(\mathbb{N}, \mathbb{R}^D) \to \ell_\infty(\mathbb{N}, \mathbb{R}^D)$ is BIBO-stable with respect to $\|\cdot\|_{\infty,2}$ if for every $B \ge 0$ there exists $C_B < \infty$ such that

$$\|x\|_{\infty,2} \le B \quad \Longrightarrow \quad \|\mathcal{N}(x)\|_{\infty,2} \le C_B.$$

B Jacobian tails under diffuse feedback routing

B.1 Sessa feedback solve as a parametric linear system

Fix a horizon $T \in \mathbb{N}^*$ and token width $D \in \mathbb{N}^*$. Let $x = (x_0, \dots, x_{T-1}) \in (\mathbb{R}^D)^T$ be the input token sequence. Let $f(x) = (f_0(x), \dots, f_{T-1}(x)) \in (\mathbb{R}^r)^T$ be the forward sequence, where r is the value space dimension, and let $\alpha^{\mathrm{fb}}(x) = (\alpha^{\mathrm{fb}}_{t,j}(x))_{0 \le j < t \le T-1}$ be the strictly-lower attention weights. Let $\gamma(x) = (\gamma_0(x), \dots, \gamma_{T-1}(x))$ be the feedback gains.

Define the strictly lower-triangular matrix $B_{\mathrm{fb}}(x) \in \mathbb{R}^{T \times T}$ by

$$[B_{\mathrm{fb}}]_{t,j}(x) = \begin{cases} \gamma_t(x)\, \alpha^{\mathrm{fb}}_{t,j}(x), & j < t, \\ 0, & j \ge t. \end{cases}$$

The mixer output $s(x) = (s_0(x), \dots, s_{T-1}(x)) \in (\mathbb{R}^r)^T$ is defined as the unique solution to the causal solve

$$(I - B_{\mathrm{fb}}(x))\, s(x) = f(x). \tag{36}$$

Equivalently, by forward substitution,

$$s_0 = f_0, \qquad s_t = f_t + \gamma_t \sum_{j=0}^{t-1} \alpha^{\mathrm{fb}}_{t,j} s_j, \quad t \ge 1. \tag{37}$$

We measure long-range sensitivity by the Jacobian blocks

$$J_{t,\tau}(x) := \frac{\partial s_t(x)}{\partial x_\tau} \in \mathbb{R}^{r \times D}, \qquad 0 \le \tau \le t \le T-1.$$

Throughout this appendix we focus on the long-range case $\tau < t$ and lag $\ell := t - \tau \geq 1$.

B.2 Assumptions for diffuse routing and smoothness

Fix a radius $R \geq 0$ and work on the ball $\mathcal{X}_R$ from Definition 7.

Remark B.1 (On the use of $t+1$ and t in dilution bounds). In this appendix the feedback attention is strictly-lower, meaning that $j < t$, so $|\mathcal{W}_t| = t$ for $t \geq 1$. We write $O(1/(t+1))$ to avoid a special case at $t = 0$ and to match harmonic-series bounds; for $t \geq 1$ this is equivalent to $O(1/t)$ up to absolute constants.

Assumption 15 (Row-stochasticity and diffuse envelope of feedback attention). *For every $x \in \mathcal{X}_R$ and every $t \geq 1$,*

$$\alpha^{\mathrm{fb}}_{t,j}(x) \geq 0, \qquad \sum_{j=0}^{t-1} \alpha^{\mathrm{fb}}_{t,j}(x) = 1, \qquad \alpha^{\mathrm{fb}}_{t,j}(x) \leq \frac{c_2}{t} \quad \forall j < t,$$

for some constant $c_2 = c_2(R) \in (0, \infty)$. We set $\alpha^{\mathrm{fb}}_{0,\cdot} \equiv 0$.

Assumption 16 (Bounded feedback gain and nontrivial diffuse regime). *For every $x \in \mathcal{X}_R$ and every t,*

$$|\gamma_t(x)| \leq \gamma_{\max} < 1,$$

and the diffuse feedback mass satisfies

$$\eta := \gamma_{\max} c_2 < 1, \qquad \beta_{\mathrm{tail}} := 1 - \eta \in (0, 1).$$

Assumption 17 (Token-wise local feedback gain). *On $\mathcal{X}_R$, the feedback gain is token-wise: for each t one has $\gamma_t(x) = \gamma(x_t)$. In particular, for $\tau < t$,*

$$\frac{\partial \gamma_t(x)}{\partial x_\tau} = 0.$$

Assume additionally the token-wise Jacobian is bounded:

$$\Big\| \frac{\partial \gamma(x_t)}{\partial x_t} \Big\|_2 \leq L_\gamma \qquad \text{for all } \|x_t\|_2 \leq R.$$

Assumption 18 (Causality of forward branch and routing). *For each time k, the quantities $f_k(x)$, $\alpha^{\mathrm{fb}}_{k,\cdot}(x)$, and $\gamma_k(x)$ depend only on the prefix $x_{0:k}$. Equivalently, for any $\tau > k$,*

$$\frac{\partial f_k(x)}{\partial x_\tau} = 0, \qquad \frac{\partial \alpha^{\mathrm{fb}}_{k,j}(x)}{\partial x_\tau} = 0 \quad (\forall j < k), \qquad \frac{\partial \gamma_k(x)}{\partial x_\tau} = 0.$$

Assumption 19 (Local, same-token smoothness bounds). *There exist finite constants $L_{f,0} = L_{f,0}(R)$ and $L_{\alpha,0} = L_{\alpha,0}(R)$ such that for all $x \in \mathcal{X}_R$ and all t,*

$$\Big\| \frac{\partial f_t(x)}{\partial x_t} \Big\|_2 \leq L_{f,0}, \qquad \sum_{j=0}^{t-1} \Big\| \frac{\partial \alpha^{\mathrm{fb}}_{t,j}(x)}{\partial x_t} \Big\|_2 \leq L_{\alpha,0}.$$

Assumption 20 (Bounded forward sequence). *There exists $F_R < \infty$ such that*

$$\|f(x)\|_{\infty,2} \leq F_R \qquad \forall x \in \mathcal{X}_R.$$

Assumption 21 (Forward-branch dilution of cross-token Jacobians). *There exists $L_f = L_f(R) < \infty$ such that*

for all $x \in \mathcal{X}_R$, all $t \geq \tau$, and all $\tau < t$,

$$\Big\|\frac{\partial f_t(x)}{\partial x_\tau}\Big\|_2 \leq \frac{L_f}{t+1}.$$

Here $\|\cdot\|_2$ is the operator norm of the matrix $\mathbb{R}^D \to \mathbb{R}^r$.

Assumption 22 (Smooth routing: α-weighted logit sensitivity)**.** *Let $\alpha^{\mathrm{fb}}_{t,\cdot}(x) = \mathrm{softmax}(\beth_{t,0}(x), \dots, \beth_{t,t-1}(x))$ denote the feedback-attention row at time t, over $j < t$, with pre-softmax logits $\beth_{t,i}(x)$ that may depend on the full prefix $x_{0:t}$. There exists $L_{\mathrm{route}} = L_{\mathrm{route}}(R) < \infty$ such that for all $x \in \mathcal{X}_R$ and all $t > \tau \geq 0$,*

$$\sum_{i=0}^{t-1} \alpha^{\mathrm{fb}}_{t,i}(x) \Big\|\frac{\partial \beth_{t,i}(x)}{\partial x_\tau}\Big\|_2 \;\leq\; \frac{L_{\mathrm{route}}}{t+1}.$$

Consequently, by Lemma B.4,

$$\sum_{j=0}^{t-1} \Big\|\frac{\partial \alpha^{\mathrm{fb}}_{t,j}(x)}{\partial x_\tau}\Big\|_2 \;\leq\; \frac{2L_{\mathrm{route}}}{t+1}.$$

Remark B.2 (When Assumption 22 holds). If the feedback query is token-wise, $q_t = q(x_t)$, then for $\tau < t$ the dependence of $\alpha^{\mathrm{fb}}_{t,\cdot}$ on x_τ typically enters only through key-side logits involving k_τ, so only a small subset of logits have nonzero $\partial \beth_{t,i}/\partial x_\tau$. In that case, Assumption 22 reduces to the corresponding localized logit-sensitivity bound. More generally, if q_t, or other components upstream of logits, has cross-token sensitivity, Assumption 22 requires that the resulting α^{fb}-weighted logit sensitivities still dilute as $O(1/(t+1))$ on $\mathcal{X}_R$.

B.3 Auxiliary lemmas

Lemma B.3 (Bound on the mixer state)**.** *Under Assumption 16–20, for all $x \in \mathcal{X}_R$,*

$$\|s(x)\|_{\infty,2} \leq S_R := \frac{F_R}{1-\gamma_{\max}}.$$

Proof. Since each $\alpha^{\mathrm{fb}}_{t,\cdot}$ is a convex distribution and $|\gamma_t| \leq \gamma_{\max}$,

$$\|s_t\|_2 \leq \|f_t\|_2 + \gamma_{\max} \max_{j<t} \|s_j\|_2.$$

A standard induction on $\max_{k\leq t} \|s_k\|_2$ yields $\|s\|_{\infty,2} \leq (1-\gamma_{\max})^{-1}\|f\|_{\infty,2} \leq (1-\gamma_{\max})^{-1}F_R$. □

Lemma B.4 (Softmax row derivative: total variation bound)**.** *Let $\alpha = \mathrm{softmax}(\beth) \in \mathbb{R}^n$ with logits $\beth \in \mathbb{R}^n$ depending on a parameter z. Then*

$$\sum_j \Big\|\frac{\partial \alpha_j}{\partial z}\Big\| \;\leq\; 2\sum_i \alpha_i \Big\|\frac{\partial \beth_i}{\partial z}\Big\|_2.$$

Proof. The softmax Jacobian satisfies $\partial \alpha_j/\partial \beth_i = \alpha_j(\mathbf{1}[j=i] - \alpha_i)$. Thus

$$\sum_{j=1}^{n} \Big|\frac{\partial \alpha_j}{\partial \beth_i}\Big| = 2\alpha_i(1-\alpha_i) \leq 2\alpha_i.$$

By the chain rule, $\sum_j \|\partial \alpha_j/\partial z\| \leq \sum_i (\sum_j |\partial \alpha_j/\partial \beth_i|)\, \|\partial \beth_i/\partial z\| \leq 2\sum_i \alpha_i \|\partial \beth_i/\partial z\|$. □

Lemma B.5 (Polynomial tail of the inverse kernel entries)**.** *Fix $x \in \mathcal{X}_R$ and let $K(x) := (I - B_{\mathrm{fb}}(x))^{-1}$. Under Assumptions 15–16, there exists a constant*

$$C_K := \eta e^{\eta} = (1-\beta_{\mathrm{tail}})e^{1-\beta_{\mathrm{tail}}}$$

such that for all $0 \le k < t \le T-1$,

$$|K_{t,k}(x)| \le C_K\,(t-k)^{-\beta_{\text{tail}}}, \qquad \textit{and} \qquad K_{t,t}(x) = 1.$$

Proof. Fix $x \in \mathcal{X}_R$, and abbreviate

$$B_{\text{fb}} := B_{\text{fb}}(x), \qquad \alpha_{t,j} := \alpha^{\text{fb}}_{t,j}(x), \qquad K := K(x) = (I - B_{\text{fb}})^{-1}.$$

Since B_{fb} is strictly lower-triangular on the finite horizon $\{0, \dots, T-1\}$, one has $B_{\text{fb}}^T = 0$, hence

$$K = (I - B_{\text{fb}})^{-1} = \sum_{m=0}^{T-1} B_{\text{fb}}^m.$$

Therefore K is lower-triangular with unit diagonal:

$$K_{t,t} = 1, \qquad K_{t,k} = 0 \text{ for } t < k.$$

It remains to prove the off-diagonal estimate.

Fix a source index $k \in \{0, \dots, T-1\}$, and define

$$u_t := |K_{t,k}| \qquad (t \ge k).$$

Then $u_k = |K_{k,k}| = 1$. Also, since $(I - B_{\text{fb}})K = I$, equivalently $K = I + B_{\text{fb}}K$, for every $t > k$ we have

$$K_{t,k} = \sum_{j<t} \big[B_{\text{fb}}\big]_{t,j}\, K_{j,k}.$$

Because $K_{j,k} = 0$ for $j < k$, this reduces to

$$K_{t,k} = \sum_{j=k}^{t-1} \big[B_{\text{fb}}\big]_{t,j}\, K_{j,k} = \gamma_t(x) \sum_{j=k}^{t-1} \alpha_{t,j}\, K_{j,k}.$$

Taking absolute values and using Assumption 16,

$$u_t \le |\gamma_t(x)| \sum_{j=k}^{t-1} \alpha_{t,j}\, u_j \le \gamma_{\max} \sum_{j=k}^{t-1} \alpha_{t,j}\, u_j, \qquad t > k. \tag{1}$$

We now compare u to an explicit impulse-response sequence. Define $(v_t^{(k)})_{t \ge 0}$ by

$$v_t^{(k)} := \begin{cases} 0, & t < k, \\ 1, & t = k, \\ \gamma_{\max} \displaystyle\sum_{j=0}^{t-1} \tilde{\alpha}_{t,j}\, v_j^{(k)}, & t > k, \end{cases}$$

where the coefficients $\tilde{\alpha}_{t,j}$ are the following extension of the finite-horizon row weights:

$$\tilde{\alpha}_{t,j} := \begin{cases} \alpha_{t,j}, & 0 \le j < t \le T-1, \\ 0, & t \ge T,\ 0 \le j < t. \end{cases}$$

Then $\tilde{\alpha}_{t,j} \geq 0$, $\sum_{j<t} \tilde{\alpha}_{t,j} \leq 1$ for every $t \geq 1$, and by Assumption 15,

$$\tilde{\alpha}_{t,j} \leq \frac{c_2}{t} \qquad (t \geq 1,\ 0 \leq j < t).$$

Thus the scalar recursion defining $v^{(k)}$ satisfies the hypotheses of Corollary E.4 with impulse position $j = k$, attention envelope constant c_2, and feedback bound $\gamma_{\max}$. In particular, with

$$\eta := \gamma_{\max} c_2, \qquad \beta_{\text{tail}} := 1 - \eta \in (0, 1),$$

that corollary yields

$$v_t^{(k)} \leq \eta e^{\eta} (t-k)^{-\beta_{\text{tail}}} \qquad \text{for all } t > k. \tag{2}$$

It remains to show that $u_t \leq v_t^{(k)}$ for all $t \in \{k, \dots, T-1\}$. We prove this by induction on t.

For $t = k$, one has $u_k = 1 = v_k^{(k)}$.

Now let $t > k$, and assume $u_j \leq v_j^{(k)}$ for every $j \in \{k, \dots, t-1\}$. Using (1), the nonnegativity of the coefficients $\alpha_{t,j}$, and the induction hypothesis, we obtain

$$u_t \leq \gamma_{\max} \sum_{j=k}^{t-1} \alpha_{t,j}\, u_j \leq \gamma_{\max} \sum_{j=k}^{t-1} \alpha_{t,j}\, v_j^{(k)}.$$

Since $v_j^{(k)} = 0$ for $j < k$ and $\tilde{\alpha}_{t,j} = \alpha_{t,j}$ for $t \leq T-1$, this is exactly

$$u_t \leq \gamma_{\max} \sum_{j=0}^{t-1} \tilde{\alpha}_{t,j}\, v_j^{(k)} = v_t^{(k)}.$$

This closes the induction.

Combining the comparison $u_t \leq v_t^{(k)}$ with (2), we conclude that for every $0 \leq k < t \leq T-1$,

$$|K_{t,k}(x)| = u_t \leq v_t^{(k)} \leq \eta e^{\eta} (t-k)^{-\beta_{\text{tail}}}.$$

Thus the claim holds with

$$C_K := \eta e^{\eta} = (1 - \beta_{\text{tail}}) e^{1-\beta_{\text{tail}}}.$$

Together with $K_{t,t} = 1$, this proves the lemma. □

Lemma B.6 (A convolution bound)**.** *Let $\beta_{\text{tail}} \in (0,1)$. There exists $C_{\beta_{\text{tail}}} < \infty$ such that for all integers $\ell \geq 1$ and all $\tau \geq 0$,*

$$\sum_{k=\tau}^{\tau+\ell-1} \frac{1}{(\tau+\ell-k)^{\beta_{\text{tail}}}} \cdot \frac{1}{k+1} \ \leq\ C_{\beta_{\text{tail}}}\, \ell^{-\beta_{\text{tail}}} (1 + \log(1+\ell)).$$

One may take, for instance,

$$C_{\beta_{\text{tail}}} := 2^{\beta_{\text{tail}}} + \frac{2^{\beta_{\text{tail}}}}{1 - \beta_{\text{tail}}}.$$

Proof. Write $k = \tau + m$ where $m = 0, \dots, \ell - 1$:

$$\sum_{m=0}^{\ell-1} \frac{1}{(\ell-m)^{\beta_{\text{tail}}}} \cdot \frac{1}{\tau+m+1}.$$

Split into $m \leq \lfloor \ell/2 \rfloor$ and $m > \lfloor \ell/2 \rfloor$.

If $m \le \ell/2$, then $(\ell - m)^{-\beta_{\text{tail}}} \le (\ell/2)^{-\beta_{\text{tail}}} = 2^{\beta_{\text{tail}}} \ell^{-\beta_{\text{tail}}}$ and

$$\sum_{m=0}^{\lfloor \ell/2 \rfloor} \frac{1}{\tau + m + 1} \le 1 + \int_0^{\ell/2} \frac{dm}{\tau + m + 1} \le 1 + \log(1 + \ell).$$

Thus this part is $\le 2^{\beta_{\text{tail}}} \ell^{-\beta_{\text{tail}}} (1 + \log(1 + \ell))$.

If $m > \ell/2$, then $\tau + m + 1 \ge \ell/2$, so $(\tau + m + 1)^{-1} \le 2/\ell$, hence

$$\sum_{m > \ell/2} \frac{1}{(\ell - m)^{\beta_{\text{tail}}}} \cdot \frac{1}{\tau + m + 1} \le \frac{2}{\ell} \sum_{r=1}^{\lfloor \ell/2 \rfloor} \frac{1}{r^{\beta_{\text{tail}}}} \le \frac{2}{\ell} \Big(1 + \int_1^{\ell/2} r^{-\beta_{\text{tail}}} \, dr \Big) \le \frac{2}{\ell} \cdot \frac{1}{1 - \beta_{\text{tail}}} \left(\frac{\ell}{2} \right)^{1 - \beta_{\text{tail}}} = \frac{2^{\beta_{\text{tail}}}}{1 - \beta_{\text{tail}}} \, \ell^{-\beta_{\text{tail}}}.$$

Combine the two bounds. □

B.4 Polynomial Jacobian tail

Theorem 23 (Polynomial Jacobian tail under diffuse routing)**.** *Assume Assumptions 15–22, 17, 18, and 19 hold on $\mathcal{X}_R$, and let $\beta_{\text{tail}} := 1 - \gamma_{\max} c_2 \in (0, 1)$ as in Assumption 16. Then there exists a constant $C(R) < \infty$ such that for every $x \in \mathcal{X}_R$ and every pair $\tau < t$ with lag $\ell = t - \tau \ge 1$,*

$$\Big\| \frac{\partial s_t(x)}{\partial x_\tau} \Big\|_2 \;\le\; C(R) \, \ell^{-\beta_{\text{tail}}} (1 + \log(1 + \ell)).$$

In particular, long-range sensitivity decays at least polynomially in the lag, up to a logarithmic factor.

One may take explicitly

$$C(R) := \widetilde{C}_K \Big(A_0(R) + (1 + C_{\beta_{\text{tail}}}) \, A_1(R) \Big), \quad \widetilde{C}_K := \max\{1, C_K\}, \quad C_K = \eta e^{\eta}, \quad \eta = \gamma_{\max} c_2,$$

where $C_{\beta_{\text{tail}}}$ is as in Lemma B.6 and

$$A_1(R) := L_f + 2\gamma_{\max} \, S_R \, L_{\text{route}}, \qquad A_0(R) := L_{f,0} + L_\gamma \, S_R + \gamma_{\max} \, S_R \, L_{\alpha,0}, \qquad S_R = \frac{F_R}{1 - \gamma_{\max}}.$$

Proof. Fix $x \in \mathcal{X}_R$ and a source index τ. Differentiate the solve (36) with respect to x_τ:

$$(I - B_{\text{fb}}) \, \frac{\partial s}{\partial x_\tau} - \frac{\partial B_{\text{fb}}}{\partial x_\tau} \, s = \frac{\partial f}{\partial x_\tau}.$$

Multiplying by $K = (I - B_{\text{fb}})^{-1}$ gives

$$\frac{\partial s}{\partial x_\tau} = K \Big(\frac{\partial f}{\partial x_\tau} + \frac{\partial B_{\text{fb}}}{\partial x_\tau} \, s \Big). \tag{38}$$

Taking the t-th row and operator norms yields

$$\Big\| \frac{\partial s_t}{\partial x_\tau} \Big\|_2 \le \sum_{k=0}^{t} |K_{t,k}| \cdot \Big\| \frac{\partial f_k}{\partial x_\tau} + \big(\frac{\partial B_{\text{fb}}}{\partial x_\tau} s \big)_k \Big\|_2. \tag{39}$$

By Assumption 18, if $k < \tau$ then $\partial f_k / \partial x_\tau = 0$ and $\partial B_{\text{fb},k,\cdot} / \partial x_\tau = 0$, hence the sum starts at $k = \tau$.

Bounding the forcing term. We treat the single index $k = \tau$ separately from the range $k > \tau$.

Case 1: $k > \tau$. For $k > \tau$, Assumption 21 gives

$$\Big\|\frac{\partial f_k}{\partial x_\tau}\Big\|_2 \le \frac{L_f}{k+1}.$$

It remains to bound $\|(\partial B_{\text{fb}}/\partial x_\tau)s\|$. For $k > \tau$ we use the full decomposition

$$\frac{\partial [B_{\text{fb}}]_{k,j}}{\partial x_\tau} = \frac{\partial \gamma_k}{\partial x_\tau}\,\alpha^{\text{fb}}_{k,j} + \gamma_k\,\frac{\partial \alpha^{\text{fb}}_{k,j}}{\partial x_\tau}.$$

By Assumption 17, $\partial\gamma_k/\partial x_\tau = 0$ for $k > \tau$, so only the second term remains. Therefore, using Lemma B.3 and Assumption 22,

$$\Big\|\big(\frac{\partial B_{\text{fb}}}{\partial x_\tau}s\big)_k\Big\|_2 \le |\gamma_k| \sum_{j<k}\Big\|\frac{\partial \alpha^{\text{fb}}_{k,j}}{\partial x_\tau}\Big\|_2 \cdot \|s_j\|_2 \le \gamma_{\max}\,S_R \sum_{j<k}\Big\|\frac{\partial \alpha^{\text{fb}}_{k,j}}{\partial x_\tau}\Big\|_2 \le \gamma_{\max}\,S_R \cdot \frac{2L_{\text{route}}}{k+1}.$$

Thus for all $k > \tau$,

$$\Big\|\frac{\partial f_k}{\partial x_\tau} + \big(\frac{\partial B_{\text{fb}}}{\partial x_\tau}s\big)_k\Big\|_2 \le \frac{A_1(R)}{k+1}, \qquad A_1(R) := L_f + 2\gamma_{\max}\,S_R\,L_{\text{route}}.$$

Case 2: $k = \tau$. Using Assumption 19 and Lemma B.3, we bound

$$\Big\|\frac{\partial f_\tau}{\partial x_\tau}\Big\|_2 \le L_{f,0}.$$

Moreover, since $[B_{\text{fb}}]_{\tau,j} = \gamma_\tau\,\alpha^{\text{fb}}_{\tau,j}$ for $j < \tau$,

$$\Big\|\big(\tfrac{\partial B_{\text{fb}}}{\partial x_\tau}s\big)_\tau\Big\|_2 \le \Big\|\tfrac{\partial \gamma_\tau}{\partial x_\tau}\Big\|_2 \cdot \sum_{j<\tau}\alpha^{\text{fb}}_{\tau,j}\|s_j\|_2 \;+\; |\gamma_\tau| \sum_{j<\tau}\Big\|\tfrac{\partial \alpha^{\text{fb}}_{\tau,j}}{\partial x_\tau}\Big\|_2 \cdot \|s_j\|_2 \le L_\gamma\,S_R + \gamma_{\max}\,L_{\alpha,0}\,S_R.$$

Hence

$$\Big\|\frac{\partial f_\tau}{\partial x_\tau} + \big(\frac{\partial B_{\text{fb}}}{\partial x_\tau}s\big)_\tau\Big\|_2 \le A_0(R), \qquad A_0(R) := L_{f,0} + L_\gamma\,S_R + \gamma_{\max}\,S_R\,L_{\alpha,0}.$$

Kernel tail and convolution. Plugging the forcing bound into (39) and using Lemma B.5 yields

$$\Big\|\frac{\partial s_t}{\partial x_\tau}\Big\|_2 \le\; |K_{t,\tau}|\,A_0(R) \;+\; \sum_{k=\tau+1}^{t}|K_{t,k}| \cdot \frac{A_1(R)}{k+1} \le |K_{t,\tau}|\,A_0(R) \;+\; A_1(R)\Big(\frac{1}{t+1} + \sum_{k=\tau+1}^{t-1} C_K\,(t-k)^{-\beta_{\text{tail}}} \cdot \frac{1}{k+1}\Big).$$

Let $\ell = t - \tau \ge 1$.

We keep the $k = t$ term explicit and show it can be absorbed into the final tail factor:

$$\frac{1}{t+1} \le \frac{1}{\tau+\ell+1} \le \frac{1}{\ell+1} \le \ell^{-1}.$$

Since $\beta_{\text{tail}} \in (0,1)$ and $\ell \ge 1$, we have $\ell^{1-\beta_{\text{tail}}} \ge 1$, hence

$$\ell^{-\beta_{\text{tail}}} = \ell^{1-\beta_{\text{tail}}}\,\ell^{-1} \;\ge\; \ell^{-1}.$$

Therefore,

$$\frac{1}{t+1} \;\le\; \ell^{-1} \;\le\; \ell^{-\beta_{\text{tail}}} \;\le\; \ell^{-\beta_{\text{tail}}}(1+\log(1+\ell)), \tag{40}$$

so the $k = t$ contribution $\frac{A_1(R)}{t+1}$ is dominated by the same $\ell^{-\beta_{\text{tail}}}(1+\log(1+\ell))$ envelope, with constant 1.

For the isolated term, Lemma B.5 gives $|K_{t,\tau}| \le C_K\, \ell^{-\beta_{\text{tail}}}$. For the remaining sum, apply Lemma B.6 Note that $\sum_{k=\tau+1}^{t-1} \le \sum_{k=\tau}^{t-1}$:

$$\sum_{k=\tau}^{t-1} (t-k)^{-\beta_{\text{tail}}} \cdot \frac{1}{k+1} \le C_{\beta_{\text{tail}}}\, \ell^{-\beta_{\text{tail}}} (1 + \log(1+\ell)).$$

Therefore

$$\left\| \frac{\partial s_t}{\partial x_\tau} \right\|_2 \le C_K\, A_0(R)\, \ell^{-\beta_{\text{tail}}} \;+\; A_1(R)\, \ell^{-\beta_{\text{tail}}} (1 + \log(1+\ell)) \;+\; C_K\, C_{\beta_{\text{tail}}}\, A_1(R)\, \ell^{-\beta_{\text{tail}}} (1 + \log(1+\ell)).$$

Since $\ell^{-\beta_{\text{tail}}} \le \ell^{-\beta_{\text{tail}}} (1 + \log(1+\ell))$ for $\ell \ge 1$ and $\widetilde{C}_K = \max\{1, C_K\} \ge 1$ and $\widetilde{C}_K \ge C_K$, we obtain

$$\left\| \frac{\partial s_t}{\partial x_\tau} \right\|_2 \le \widetilde{C}_K \Big(A_0(R) + (1 + C_{\beta_{\text{tail}}}) A_1(R) \Big)\, \ell^{-\beta_{\text{tail}}} (1 + \log(1+\ell)),$$

which is the claim with the stated $C(R)$. □

B.5 Jacobian tail for block outputs

Consider the simplified block output of the form

$$y_t = x_t + W^{\text{out}}\, (s_t \odot g_t) + b^{\text{out}},$$

where $g_t = g_t(x_t)$ is token-wise and serves as a gate, and W^{out} is a fixed matrix.

Corollary B.7 (Jacobian tail for block outputs)**.** *Under the assumptions of Theorem 23, suppose additionally that $\|g(x)\|_{\infty,2} \le G_R$ for all $x \in \mathcal{X}_R$. Then for every $\tau < t$ with lag $\ell = t - \tau \ge 1$,*

$$\left\| \frac{\partial y_t(x)}{\partial x_\tau} \right\|_2 \;\le\; \|W^{\text{out}}\|_2\, G_R \cdot C(R)\, \ell^{-\beta_{\text{tail}}} (1 + \log(1+\ell)), \qquad \forall x \in \mathcal{X}_R.$$

Proof. For $\tau < t$, $\partial x_t / \partial x_\tau = 0$, and since g_t is token-wise, $\partial g_t / \partial x_\tau = 0$. Thus

$$\frac{\partial y_t}{\partial x_\tau} = W^{\text{out}}\, \mathrm{Diag}(g_t)\, \frac{\partial s_t}{\partial x_\tau}.$$

Taking operator norms and using $\|\mathrm{Diag}(g_t)\|_2 \le \|g_t\|_2 \le G_R$ plus Theorem 23 gives the result. □

C Proofs for Section 4.2

Lemma C.1 (Bounded logit spread implies near-uniform softmax weights)**.** *Let $\mathcal{J}$ be a finite index set with $n := |\mathcal{J}|$, and let $(\beth_j)_{j \in \mathcal{J}} \subset \mathbb{R}$ be logits. Define the softmax weights*

$$\alpha_j \;=\; \frac{e^{\beth_j}}{\sum_{i \in \mathcal{J}} e^{\beth_i}}, \qquad j \in \mathcal{J}.$$

If the logit spread is bounded by

$$\Delta \;:=\; \max_{i \in \mathcal{J}} \beth_i \;-\; \min_{i \in \mathcal{J}} \beth_i \;\le\; \Delta_0 < \infty,$$

then for every $j \in \mathcal{J}$,

$$\frac{e^{-\Delta_0}}{n} \;\le\; \alpha_j \;\le\; \frac{e^{\Delta_0}}{n}. \tag{41}$$

Equivalently, for all $i, j \in \mathcal{J}$ one has $e^{-\Delta_0} \le \alpha_i / \alpha_j \le e^{\Delta_0}$. In particular, if Δ_0 is uniformly bounded while n grows, then $\alpha_j = \Theta(1/n)$ uniformly over $j \in \mathcal{J}$.

Proof. Let $ב_{\min} := \min_{i\in\mathcal{J}} ב_i$. Then $ב_{\min} \le ב_j \le ב_{\min} + \Delta_0$ for all $j \in \mathcal{J}$, hence $e^{ב_{\min}} \le e^{ב_j} \le e^{ב_{\min}+\Delta_0}$ and

$$n\, e^{ב_{\min}} \;\le\; \sum_{i\in\mathcal{J}} e^{ב_i} \;\le\; n\, e^{ב_{\min}+\Delta_0}.$$

Dividing $e^{ב_j}$ by these bounds yields (41). □

C.1 Proof of Lemma 4.3

Proof of Lemma 4.3. Fix a time t and an index $\tau < t$. Write

$$\alpha^{\mathrm{fwd}}_{t,\cdot}(x) = \mathrm{softmax}(ב_{t,0}(x), \dots, ב_{t,t}(x)), \qquad \alpha_j := \alpha^{\mathrm{fwd}}_{t,j}(x), \quad \beta_j := ב_{t,j}(x), \qquad 0 \le j \le t.$$

Thus $\alpha = \mathrm{softmax}(\beta) \in \mathbb{R}^{t+1}$ and $\sum_{j\le t} \alpha_j = 1$.

Recall the standard softmax Jacobian identity: for all $j, i \in \{0, \dots, t\}$, the softmax partial derivatives satisfy

$$\frac{\partial \alpha_j}{\partial \beta_i} = \alpha_j(\mathbf{1}[j = i] - \alpha_i). \tag{42}$$

By assumption, for each $j \le t$,

$$\beta_j = ב_{t,j}(x) = \langle q(x_t),\, k(x_j)\rangle,$$

where q, k are token-wise maps. Since $\tau < t$, the quantity $q(x_t)$ depends only on x_t, hence $\partial q(x_t)/\partial x_\tau = 0$. Similarly, $k(x_j)$ depends only on x_j, hence $\partial k(x_j)/\partial x_\tau = 0$ unless $j = \tau$. Therefore,

$$\frac{\partial \beta_i}{\partial x_\tau} = 0 \quad \text{for all } i \ne \tau, \qquad \text{and potentially} \qquad \frac{\partial \beta_\tau}{\partial x_\tau} \ne 0. \tag{43}$$

Consequently, by the chain rule and (43),

$$\frac{\partial \alpha_j}{\partial x_\tau} = \sum_{i\le t} \frac{\partial \alpha_j}{\partial \beta_i}\frac{\partial \beta_i}{\partial x_\tau} = \frac{\partial \alpha_j}{\partial \beta_\tau}\frac{\partial \beta_\tau}{\partial x_\tau} = \alpha_j(\mathbf{1}[j = \tau] - \alpha_\tau)\,\frac{\partial \beta_\tau}{\partial x_\tau},$$

where we used (42) in the last step. Taking operator norms gives

$$\Big\|\frac{\partial \alpha_j}{\partial x_\tau}\Big\|_2 = |\alpha_j(\mathbf{1}[j = \tau] - \alpha_\tau)| \cdot \Big\|\frac{\partial \beta_\tau}{\partial x_\tau}\Big\|_2. \tag{44}$$

Summing (44) over $j \le t$ yields

$$\sum_{j\le t}\Big\|\frac{\partial \alpha_j}{\partial x_\tau}\Big\|_2 = \Big(\sum_{j\le t} |\alpha_j(\mathbf{1}[j = \tau] - \alpha_\tau)|\Big)\Big\|\frac{\partial \beta_\tau}{\partial x_\tau}\Big\|_2.$$

To evaluate the scalar sum, note that

$$\sum_{j\le t} |\alpha_j(\mathbf{1}[j = \tau] - \alpha_\tau)| = \underbrace{\alpha_\tau(1 - \alpha_\tau)}_{j=\tau} + \underbrace{\sum_{j\ne\tau} \alpha_j\alpha_\tau}_{j\ne\tau} = \alpha_\tau(1 - \alpha_\tau) + \alpha_\tau \sum_{j\ne\tau} \alpha_j = 2\alpha_\tau(1 - \alpha_\tau) \le 2\alpha_\tau,$$

since $\sum_{j\ne\tau} \alpha_j = 1 - \alpha_\tau$ and $1 - \alpha_\tau \le 1$. Therefore,

$$\sum_{j\le t}\Big\|\frac{\partial \alpha^{\mathrm{fwd}}_{t,j}(x)}{\partial x_\tau}\Big\|_2 \le 2\,\alpha^{\mathrm{fwd}}_{t,\tau}(x)\Big\|\frac{\partial ב_{t,\tau}(x)}{\partial x_\tau}\Big\|_2,$$

which is the first claim.

In particular. If $\|\partial \beth_{t,\tau}(x)/\partial x_\tau\|_2 \le L_\beth$ on $\mathcal{X}_R$, then

$$\sum_{j\le t} \Big\| \frac{\partial \alpha^{\mathrm{fwd}}_{t,j}(x)}{\partial x_\tau} \Big\|_2 \le 2L_\beth\, \alpha^{\mathrm{fwd}}_{t,\tau}(x).$$

In the diffuse regime of Definition 4, Lemma C.1 implies $\alpha^{\mathrm{fwd}}_{t,\tau}(x) = \Theta(1/|\mathcal{W}_t|)$ uniformly over $\tau \in \mathcal{W}_t$, hence the right-hand side is $\lesssim 1/|\mathcal{W}_t|$. For full-prefix attention $|\mathcal{W}_t| = t+1$. □

C.2 Proof of Proposition 9

Proof of Proposition 9. Fix a horizon T and work with the fixed-routing Jacobians from Section 4.2.1.

(1) Transformer: attention one-hop dilution. By definition of the value influence Jacobian under realized attention weights, by Eq. (26),

$$J^{\mathrm{attn}}_{t,\tau} = \frac{\partial y_t}{\partial v_\tau}\Big|_{\alpha^{\mathrm{fwd}}} = \alpha^{\mathrm{fwd}}_{t,\tau}\, I.$$

Taking operator norms and using $\|I\| = 1$ gives

$$\|J^{\mathrm{attn}}_{t,\tau}\| = \|\alpha^{\mathrm{fwd}}_{t,\tau} I\| = \alpha^{\mathrm{fwd}}_{t,\tau}.$$

Assume the shared diffuse (low-separation) regime of Definition 4 with full-prefix visibility $\mathcal{W}_t = \{0,\dots,t\}$, so $|\mathcal{W}_t| = t+1$. The bounded logit spread over $\mathcal{W}_t$ implies, by Lemma C.1, that for every $\tau \le t$,

$$\frac{e^{-\Delta}}{t+1} \le \alpha^{\mathrm{fwd}}_{t,\tau} \le \frac{e^{\Delta}}{t+1},$$

hence $\alpha^{\mathrm{fwd}}_{t,\tau} = \Theta(1/(t+1))$ and therefore

$$\|J^{\mathrm{attn}}_{t,\tau}\| = \Theta\Big(\frac{1}{t+1}\Big) \qquad (\tau \le t).$$

For a fixed old source $\tau = O(1)$ and lag $\ell = t - \tau$, we have

$$\|J^{\mathrm{attn}}_{\tau+\ell,\tau}\| = \alpha^{\mathrm{fwd}}_{\tau+\ell,\tau} = \Theta\Big(\frac{1}{\tau+\ell+1}\Big) = \Theta(1/\ell),$$

since τ is fixed and $\ell \to \infty$.

(2) Mamba under failed freeze time. By definition of the fixed-routing impulse Jacobian for an SSM, by Eq. (28),

$$J^{\mathrm{ssm}}_{t,\tau} = C_{\mathrm{ssm},t}\Big(\prod_{r=\tau+1}^{t} A_{\mathrm{ssm},r}\Big) B_{\mathrm{ssm},\tau}, \qquad 0 \le \tau \le t.$$

Assume the realized recurrence has diagonal transitions

$$A_{\mathrm{ssm},r} = \mathrm{diag}(\exp(-a_n \Delta_r)), \qquad a_n \ge \lambda > 0,$$

and bounded input/output factors

$$\sup_r \|B_{\mathrm{ssm},r}\| \le B_{\max}, \qquad \sup_r \|C_{\mathrm{ssm},r}\| \le C_{\max}.$$

Then

$$\Big\| \prod_{r=\tau+1}^{t} A_{\mathrm{ssm},r} \Big\| = \max_n \exp\Big(- a_n \sum_{r=\tau+1}^{t} \Delta_r \Big) \le \exp\Big(- \lambda \sum_{r=\tau+1}^{t} \Delta_r \Big).$$

Under the failed-freeze-time condition

$$\sum_{r=\tau+1}^{t} \Delta_r \ge c_\Delta (t-\tau),$$

it follows that

$$\|J^{\mathrm{ssm}}_{t,\tau}\| \le C_{\max} B_{\max} \exp(- \lambda c_\Delta (t-\tau)).$$

Setting $c := C_{\max} B_{\max}$ and $\ell := t - \tau$ gives

$$\|J^{\mathrm{ssm}}_{t,\tau}\| \le c\, e^{-\lambda c_\Delta \ell}.$$

(3) Sessa: diffuse feedback routing. For a realized feedback matrix B_{fb}, the solve Jacobian is the resolvent given by Eq. (27)

$$J^{\mathrm{sessa}} = (I - B_{\mathrm{fb}})^{-1}, \qquad J^{\mathrm{sessa}}_{t,\tau} = [(I - B_{\mathrm{fb}})^{-1}]_{t,\tau}.$$

Since B_{fb} is scalar-valued, $J^{\mathrm{sessa}}_{t,\tau} \in \mathbb{R}$ is a scalar coefficient shared across features.

Fix τ and consider the impulse in the forward stream f at time τ: $f_\tau = 1$ and $f_t = 0$ for $t \neq \tau$. Let s be the solution to $(I - B_{\mathrm{fb}})s = f$. By linearity, $s_t = J^{\mathrm{sessa}}_{t,\tau}$ for all t. Moreover, by forward substitution (equivalently (31)), $s_\tau = 1$ and for $t > \tau$,

$$s_t = f_t + \gamma_t \sum_{j=0}^{t-1} \alpha^{\mathrm{fb}}_{t,j} s_j = \gamma_t \sum_{j=\tau}^{t-1} \alpha^{\mathrm{fb}}_{t,j} s_j,$$

since $f_t = 0$ for $t \neq \tau$ and $s_j = 0$ for $j < \tau$ in a strictly causal solve.

Under Assumptions 6–7 we have $\alpha^{\mathrm{fb}}_{t,j} \le c_2/t$ for all $j < t$ and $|\gamma_t| \le \gamma_{\max} < 1$, and defining $\beta_{\mathrm{tail}} := 1 - \gamma_{\max} c_2 \in (0, 1]$, with $\gamma_{\max} c_2 < 1$, Theorem 8 applies to this impulse recursion, shifted to start at τ, and yields that for all lags $\ell \ge 1$,

$$|J^{\mathrm{sessa}}_{\tau+\ell,\tau}| = |s_{\tau+\ell}| \le C\, \ell^{-\beta_{\mathrm{tail}}},$$

for an explicit constant C, e.g. $C = (1 - \beta_{\mathrm{tail}}) e^{1-\beta_{\mathrm{tail}}}$.

Tightness. In the explicit uniform-routing regime

$$[B_{\mathrm{fb}}]_{t,j} = \begin{cases} 0, & t = 0, \\ \dfrac{\gamma}{t} \mathbf{1}[j < t], & t \ge 1, \end{cases} \qquad \gamma \in (0, 1),$$

one has $\alpha^{\mathrm{fb}}_{t,j} = t^{-1} \mathbf{1}[j < t]$ and constant gain $\gamma_t \equiv \gamma$, hence $\beta_{\mathrm{tail}} = 1 - \gamma$. Appendix Corollary F.2 gives, for every fixed source position τ,

$$|J^{\mathrm{sessa}}_{\tau+\ell,\tau}| = \Theta_\tau(\ell^{-\beta_{\mathrm{tail}}}).$$

Moreover, Appendix Corollary F.3 yields the stronger uniform statement that for every $\tau_{\max} < \infty$ there exist constants $c^-_{\tau_{\max}}, c^+_{\tau_{\max}} > 0$ such that

$$c^-_{\tau_{\max}}\, \ell^{-\beta_{\mathrm{tail}}} \le |J^{\mathrm{sessa}}_{\tau+\ell,\tau}| \le c^+_{\tau_{\max}}\, \ell^{-\beta_{\mathrm{tail}}}$$

for all $0 \le \tau \le \tau_{\max}$ and all $\ell \ge 1$. Thus the one-layer envelope is tight for each fixed source and uniformly on every bounded source family, in particular on every fixed finite horizon. □

C.3 Proof of Proposition 3

Proof. The claim is about the input–output map and is independent of the chosen realization. By the controllable and observable decomposition, also known as the Kalman decomposition (Antsaklis and Michel, 2006), there exists

a similarity transform that isolates the controllable and observable subsystem $(A_{\text{ssm,co}}, B_{\text{ssm,co}}, C_{\text{ssm,co}})$ such that for all $\ell \geq 0$,

$$C_{\text{ssm}} A_{\text{ssm}}^{\ell} B_{\text{ssm}} \;=\; C_{\text{ssm,co}}\, A_{\text{ssm,co}}^{\ell}\, B_{\text{ssm,co}}.$$

Moreover, $(A_{\text{ssm,co}}, B_{\text{ssm,co}}, C_{\text{ssm,co}})$ is a minimal realization of the same transfer function, so it admits no pole–zero cancellations and its poles coincide with the reachable and observable eigenvalues of $A_{\text{ssm,co}}$ (Dahleh et al., 2011b). Since the transfer function is BIBO stable, all its poles lie strictly inside the unit disk (DT case) (Dahleh et al., 2011a); hence $\rho_{\text{spec}}\big(A_{\text{ssm,co}}\big) < 1$. It follows from standard finite-dimensional matrix power bounds that there exist $c > 0$ and $\kappa \in (0, 1)$ such that $\|A_{\text{ssm,co}}^{\ell}\| \leq c\,\kappa^{\ell}$ for all ℓ, and therefore $\|C_{\text{ssm}} A_{\text{ssm}}^{\ell} B_{\text{ssm}}\| = \|C_{\text{ssm,co}} A_{\text{ssm,co}}^{\ell} B_{\text{ssm,co}}\| \leq c'\kappa^{\ell}$. □

C.4 Proof of Proposition 4

The key point is that, under ZOH discretization, the state-transition product is controlled by the accumulated discretization time

$$\sum_{r=\tau+1}^{t} \Delta_r(x),$$

since each channel contributes a factor $\exp(-a_n \Delta_r(x))$. Accordingly, the proof first obtains an end-to-end Jacobian bound in terms of

$$\Pi_{t,\ell}(x) = \exp\Big(-\lambda \sum_{r=\tau+1}^{t} \Delta_r(x)\Big),$$

and only then converts this into exponential-in-lag decay under failed freeze time.

Proof. Fix $x \in \mathcal{X}_R$ and indices $\tau < t$, and set $\ell := t - \tau \geq 1$. Write $J_{t,\tau}^{h} := \partial h_t(x)/\partial x_\tau$ and $J_{t,\tau}^{\text{e2e}} := \partial y_t(x)/\partial x_\tau$. We use the product convention

$$\prod_{r=\tau+1}^{t} A_{\text{ssm},r} := A_{\text{ssm},t} A_{\text{ssm},t-1} \cdots A_{\text{ssm},\tau+1}, \qquad \prod_{r=t+1}^{t} (\cdot) := I.$$

State bound via ZOH convexity. In a ZOH-diagonal channel, each mode n evolves as the scalar recursion

$$(h_t)_n = e^{-a_n \Delta_t}\,(h_{t-1})_n + \frac{1 - e^{-a_n \Delta_t}}{a_n}\,(b_t)_n, \qquad a_n \geq \lambda,\ \Delta_t \geq 0,$$

where we take

$$b_t := \widetilde{B_{\text{ssm},t}}(x_t) u_t(x_t).$$

By the bounds on $\widetilde{B_{\text{ssm},t}}$ and u_t on $\mathcal{X}_R$, we have

$$\|b_t\| \leq G_{\max} U_R,$$

and hence $|(b_t)_n| \leq G_{\max} U_R$ for each mode. Since $h_{-1} = 0$, Lemma 4.4 applied componentwise with $a_{\min} = \lambda$ gives

$$\sup_t |(h_t)_n| \leq \frac{G_{\max} U_R}{\lambda} \qquad \text{for every mode } n.$$

Therefore

$$\|h_t\|_2 \leq \sqrt{d_{\text{state}}}\, \|h_t\|_\infty \leq \sqrt{d_{\text{state}}}\, \frac{G_{\max} U_R}{\lambda} =: H_R.$$

Jacobian recursion for $t > \tau$. For $t > \tau$, locality implies

$$\frac{\partial A_{\mathrm{ssm},t}(x_t)}{\partial x_\tau} = \frac{\partial \widetilde{B_{\mathrm{ssm},t}}(x_t)}{\partial x_\tau} = \frac{\partial u_t(x_t)}{\partial x_\tau} = \frac{\partial G_{\mathrm{ssm},t}(x_t)}{\partial x_\tau} = 0.$$

Differentiating

$$h_t = A_{\mathrm{ssm},t}(x_t)\, h_{t-1} + G_{\mathrm{ssm},t}(x_t)\, \widetilde{B_{\mathrm{ssm},t}}(x_t)\, u_t(x_t)$$

with respect to x_τ yields

$$J^h_{t,\tau} = A_{\mathrm{ssm},t}(x_t)\, J^h_{t-1,\tau}, \qquad t > \tau.$$

Iterating gives

$$J^h_{t,\tau} = \Big(\prod_{r=\tau+1}^{t} A_{\mathrm{ssm},r}(x_r) \Big)\, J^h_{\tau,\tau}.$$

Source-time derivative bound. At $t = \tau$, write $b_\tau := \widetilde{B_{\mathrm{ssm},\tau}}(x_\tau) u_\tau(x_\tau)$ and differentiate the ZOH update:

$$J^h_{\tau,\tau} = \Big(\frac{\partial A_{\mathrm{ssm},\tau}(x_\tau)}{\partial x_\tau}\Big) h_{\tau-1} + \Big(\frac{\partial G_{\mathrm{ssm},\tau}(x_\tau)}{\partial x_\tau}\Big) b_\tau + G_{\mathrm{ssm},\tau}(x_\tau)\, \frac{\partial b_\tau}{\partial x_\tau}.$$

Moreover,

$$\frac{\partial b_\tau}{\partial x_\tau} = \Big(\frac{\partial \widetilde{B_{\mathrm{ssm},\tau}}(x_\tau)}{\partial x_\tau}\Big) u_\tau + \widetilde{B_{\mathrm{ssm},\tau}}(x_\tau) \Big(\frac{\partial u_\tau(x_\tau)}{\partial x_\tau}\Big).$$

Since

$$G_{\mathrm{ssm},\tau}(x_\tau) = \mathrm{diag}\Big(\frac{1 - [A_{\mathrm{ssm},\tau}(x_\tau)]_n}{a_n}\Big)_n,$$

we have the operator bounds

$$\|G_{\mathrm{ssm},\tau}(x_\tau)\| \le \frac{1}{\lambda}, \qquad \Big\|\frac{\partial G_{\mathrm{ssm},\tau}(x_\tau)}{\partial x_\tau}\Big\| \le \frac{1}{\lambda}\Big\|\frac{\partial A_{\mathrm{ssm},\tau}(x_\tau)}{\partial x_\tau}\Big\|.$$

Using

$$\|h_{\tau-1}\| \le H_R, \qquad \|b_\tau\| \le G_{\max} U_R,$$

together with the derivative bounds gives

$$\|J^h_{\tau,\tau}\| \le L_A\, H_R \;+\; \frac{L_A}{\lambda}\, G_{\max} U_R \;+\; \frac{1}{\lambda}(L_B\, U_R + G_{\max}\, L_u) =: J_R.$$

Transition product bound by accumulated discretization time. Since each $A_{\mathrm{ssm},r}$ is diagonal with entries $\exp(-a_n \Delta_r)$ and $a_n \ge \lambda$,

$$\Big\| \prod_{r=\tau+1}^{t} A_{\mathrm{ssm},r}(x_r) \Big\| = \max_n \exp\Big(- a_n \sum_{r=\tau+1}^{t} \Delta_r(x) \Big) \le \exp\Big(- \lambda \sum_{r=\tau+1}^{t} \Delta_r(x) \Big) =: \Pi_{t,\ell}(x).$$

Therefore

$$\|J^h_{t,\tau}\| \le \Pi_{t,\ell}(x)\, \|J^h_{\tau,\tau}\| \le J_R\, \Pi_{t,\ell}(x).$$

Output Jacobian. For $\tau < t$, locality implies $\partial C_{\mathrm{ssm},t}(x_t)/\partial x_\tau = 0$, so

$$\frac{\partial y_t}{\partial x_\tau} = C_{\mathrm{ssm},t}(x_t)\, J^h_{t,\tau}.$$

Hence

$$\Big\|\frac{\partial y_t(x)}{\partial x_\tau}\Big\| \le \|C_{\mathrm{ssm},t}(x_t)\|\,\|J^h_{t,\tau}\| \le C_R\,J_R\,\Pi_{t,\ell}(x).$$

Thus the claim holds with

$$C(R) := C_R J_R.$$

□

C.5 Proof of Lemma 4.4

Proof of Lemma 4.4. Fix $t \ge 0$ and define $\theta_t := e^{-a\Delta_t} \in [0,1]$, since $a > 0$ and $\Delta_t \ge 0$. Then $1-\theta_t = 1-e^{-a\Delta_t} \in [0,1]$, and the update can be rewritten as

$$h_t = \theta_t\, h_{t-1} + (1-\theta_t)\,\frac{b_t}{a}.$$

Taking absolute values and using the triangle inequality yields

$$|h_t| \le \theta_t|h_{t-1}| + (1-\theta_t)\frac{|b_t|}{a}.$$

Since $\theta_t \in [0,1]$, for any $u, v \ge 0$ one has $\theta_t u + (1-\theta_t)v \le \max\{u,v\}$, hence

$$|h_t| \le \max\Big\{|h_{t-1}|,\ \frac{|b_t|}{a}\Big\} \le \max\Big\{|h_{t-1}|,\ \frac{|b_t|}{a_{\min}}\Big\},$$

using $a \ge a_{\min}$.

Define

$$B_t := \max\Big\{|h_{-1}|,\ \max_{0\le s\le t}\frac{|b_s|}{a_{\min}}\Big\}.$$

We claim by induction that $|h_t| \le B_t$ for all $t \ge 0$. For $t = 0$ this follows from the previous inequality. If $|h_{t-1}| \le B_{t-1}$, then

$$|h_t| \le \max\Big\{|h_{t-1}|,\ \frac{|b_t|}{a_{\min}}\Big\} \le \max\Big\{B_{t-1},\ \frac{|b_t|}{a_{\min}}\Big\} = B_t,$$

proving the induction. Taking $\sup_{t\ge 0}$ gives

$$\sup_{t\ge 0}|h_t| \le \max\Big\{|h_{-1}|,\ \sup_{s\ge 0}\frac{|b_s|}{a_{\min}}\Big\},$$

which is the general bound.

If additionally $|b_t| \le M$ for all t and $h_{-1} = 0$, then the right-hand side is at most $M/a_{\min}$, proving $\sup_t |h_t| \le M/a_{\min}$. □

Remark C.2 (Vector and diagonal case). For diagonal $A = -\mathrm{diag}(a_n)$ with $\min_n a_n \ge a_{\min}$, the bound holds componentwise for each mode and channel, and hence yields the uniform bound $\|h_t\|_\infty \le \sup_s \|b_s\|_\infty / a_{\min}$. More generally, for any monotone norm $\|\cdot\|$ one has $\|h_t\| \le \|\mathbf{1}\|\,\sup_s \|b_s\|_\infty / a_{\min}$.

C.6 Proof of Corollary 4.6

Proof. Proposition 4 gives

$$\Big\|\frac{\partial y_t(x)}{\partial x_\tau}\Big\| \le C(R)\,\Pi_{t,\ell}(x), \qquad \Pi_{t,\ell}(x) = \exp\Big(-\lambda\sum_{r=\tau+1}^{t}\Delta_r(x)\Big).$$

Under failed freeze time,

$$\sum_{r=\tau+1}^{t} \Delta_r(x) \geq c_\Delta(t-\tau).$$

Applying Proposition 5 yields

$$\Pi_{t,\ell}(x) \leq \exp(-\lambda c_\Delta(t-\tau)),$$

and therefore

$$\Big\|\frac{\partial y_t(x)}{\partial x_\tau}\Big\| \leq C(R)\exp(-\lambda c_\Delta(t-\tau)).$$

□

Remark C.3 (Local windows). If $A_{\text{ssm},t}, \widetilde{B_{\text{ssm},t}}, C_{\text{ssm},t}, u_t$ depend on a fixed window $x_{t-K:t}$, the same argument yields

$$\Big\|\frac{\partial y_t}{\partial x_\tau}\Big\| \leq C(R)\,\exp\Big(-\lambda \sum_{r=\tau+K+1}^{t} \Delta_r(x)\Big) \qquad (t > \tau + K),$$

so the same failed-freeze-time conclusion holds up to a finite-window slack.

C.7 Proof of Proposition 5

Proof. By definition,

$$\Pi_{t,\ell} = \exp\Big(-\lambda \sum_{r=\tau+1}^{t} \Delta_r\Big).$$

Under the failed-freeze-time condition

$$\sum_{r=\tau+1}^{t} \Delta_r \geq c_\Delta(t-\tau) = c_\Delta \ell,$$

we obtain

$$\Pi_{t,\ell} \leq \exp(-\lambda c_\Delta \ell).$$

This is exactly the claim. □

C.8 Details for Proposition 10

Proof. **(1) Transformer attention in the no-freeze setting.** Let $y_t(x) = \sum_{j\in\mathcal{W}_t} \alpha_{t,j}(x)\, v(x_j)$. For $\tau < t$, differentiate:

$$\frac{\partial y_t}{\partial x_\tau} = \alpha_{t,\tau}\frac{\partial v(x_\tau)}{\partial x_\tau} + \sum_{j\in\mathcal{W}_t} \frac{\partial \alpha_{t,j}(x)}{\partial x_\tau}\, v(x_j).$$

Taking operator norms and using $\|\partial v(x_\tau)/\partial x_\tau\| \leq L_v$ and $\|v(x_j)\| \leq V_R$ yields

$$\Big\|\frac{\partial y_t}{\partial x_\tau}\Big\| \leq \alpha_{t,\tau} L_v + V_R \sum_{j\in\mathcal{W}_t} \Big\|\frac{\partial \alpha_{t,j}}{\partial x_\tau}\Big\|.$$

Under the shared regime in Section 4.2.2, $\alpha_{t,\tau} \leq c_2/|\mathcal{W}_t|$ and $\sum_{j\in\mathcal{W}_t} \|\partial\alpha_{t,j}/\partial x_\tau\| \leq L_\alpha/|\mathcal{W}_t|$, hence $\|\partial y_t/\partial x_\tau\| \lesssim 1/|\mathcal{W}_t|$. For full-prefix attention $|\mathcal{W}_t| = t+1$, recovering $\|\partial y_t/\partial x_\tau\| \lesssim 1/(t+1)$.

(2) Mamba under failed freeze time. Item (2) follows by combining Proposition 4 with failed freeze time, namely

$$\sum_{r=\tau+1}^{t} \Delta_r(x) \geq c_\Delta(t-\tau),$$

that is, by Corollary 4.6. □

D BIBO stability on infinite horizons and uniform-in-T bounds

We extend the finite-horizon BIBO statement to infinite sequences under an explicit row-contraction condition, and to uniform-in-T bounds for truncated length-T networks without appealing to compactness.

D.1 Sequence norms and stability definition

We use the norm $\|\cdot\|_{\infty,2}$ and balls from Definition 7. For finite tensors we also use the comparison (35).

D.2 Feedback matrix and row-contraction condition

Fix a causal width-m Sessa block G as in Section 3.1, but now acting on infinite sequences in $\ell_\infty(\mathbb{N},\mathbb{R}^m)$. We emphasize that the block input and output live in $\mathbb{R}^m$, while the triangular solve $(I - B_{\mathrm{fb}})s = f$ is performed in a value space $\mathbb{R}^r$: in our definition, $s_t \in \mathbb{R}^r$, $f_t \in \mathbb{R}^r$, $g_t \in \mathbb{R}^r$, and $z_t = s_t \odot g_t \in \mathbb{R}^r$, and the output projection is token-wise affine $o : \mathbb{R}^r \to \mathbb{R}^m$.

Causal feedback-attention weights. For each input x, the masked softmax in the feedback branch defines strictly lower-triangular weights $(\alpha^{\mathrm{fb}}_{t\tau}(x))_{t,\tau\geq 0}$ with

$$\alpha^{\mathrm{fb}}_{t\tau}(x) \geq 0, \qquad \alpha^{\mathrm{fb}}_{t\tau}(x) = 0 \text{ for } \tau \geq t, \qquad \sum_{\tau<t} \alpha^{\mathrm{fb}}_{t\tau}(x) = 1 \text{ for } t \geq 1, \tag{45}$$

with the empty sum $= 0$ for $t = 0$. These properties hold as follows: for $t \geq 1$ each row t is a softmax over the finite set $\{0,\dots,t-1\}$, hence $\alpha^{\mathrm{fb}}_{t\tau} \geq 0$ and $\sum_{\tau<t} \alpha^{\mathrm{fb}}_{t\tau} = 1$; for $t = 0$ we set $\alpha^{\mathrm{fb}}_{0\tau} = 0$ for all τ, i.e. the context is empty, so the empty sum equals 0.

Feedback attention matrix. Define $\mathrm{A}_{\mathrm{fb}}(x) := (\alpha^{\mathrm{fb}}_{t\tau}(x))_{t,\tau\geq 0}$.

Feedback coefficient and the Sessa matrix B_{fb}. By definition of the Sessa block, the feedback coefficient is

$$\gamma_t(x) = \tanh(u_t(x)) \in (-1,1),$$

computed token-wise from the block input, via affine maps and element-wise nonlinearities. Define the diagonal operator $\Gamma_{\mathrm{fb}}(x) := \mathrm{diag}(\gamma_t(x))_{t\geq 0}$ and the strictly lower-triangular matrix

$$B_{\mathrm{fb}}(x) := \Gamma_{\mathrm{fb}}(x)\ \mathrm{A}_{\mathrm{fb}}(x) \qquad \Longleftrightarrow \qquad [B_{\mathrm{fb}}]_{t,\tau}(x) = \gamma_t(x)\,\alpha^{\mathrm{fb}}_{t\tau}(x). \tag{46}$$

Assumption 24 (Uniform feedback margin and row contraction)**.** *For every radius $R \geq 0$ there exists $\rho(R) \in [0,1)$ such that for all inputs $x \in \ell_\infty(\mathbb{N},\mathbb{R}^m)$ with $\|x\|_{\infty,2} \leq R$,*

$$\sup_{t\geq 0} |\gamma_t(x)| \leq \rho(R). \tag{47}$$

In particular, using (45)–(46), for every x,

$$\sup_{t\geq 0} \sum_{\tau<t} |[B_{\mathrm{fb}}]_{t,\tau}(x)| = \sup_{t\geq 1} \sum_{\tau<t} |[B_{\mathrm{fb}}]_{t,\tau}(x)| = \sup_{t\geq 1} |\gamma_t(x)| \leq \sup_{t\geq 0} |\gamma_t(x)| \leq \rho(R) < 1. \tag{$\star$}$$

Remark D.1 (An explicit choice of (R)). If $u_t(x)$ is produced by a token-wise feedforward stack of affine maps and element-wise nonlinearities σ satisfying $|\sigma(z)| \leq |z|$ coordinate-wise; this holds for GELU. Affine and linear maps

are handled separately via spectral norms as in Lemma D.2. Then for some explicit constants $c_\gamma \geq 0$, $L_{\gamma,\mathrm{pre}} \geq 0$ depending only on the block parameters,

$$\sup_{t\geq 0} |u_t(x)| \;\leq\; c_\gamma + L_{\gamma,\mathrm{pre}} \|x\|_{\infty,2}. \tag{48}$$

Hence on the ball $\|x\|_{\infty,2} \leq R$ one can take

$$\rho(R) := \tanh(c_\gamma + L_{\gamma,\mathrm{pre}} R) \;<\; 1. \tag{49}$$

The strict inequality holds since $c_\gamma + L_{\gamma,\mathrm{pre}} R < \infty$ and $\tanh(\cdot) < 1$ for finite arguments.

D.3 Causal triangular solve on ℓ_∞

The only operation that truly changes nature at $T = \infty$ is the lower-triangular solve. We treat it as a causal linear system.

D.4 Proof of Lemma 4.2

Proof. Let $B_{\mathrm{fb}} = ([B_{\mathrm{fb}}]_{t,\tau})_{t,\tau\geq 0}$ be strictly lower-triangular and define the causal operator $(B_{\mathrm{fb}} s)_t := \sum_{\tau<t} [B_{\mathrm{fb}}]_{t,\tau} s_\tau$, a finite sum for each fixed t, acting on $\mathbb{R}^r$-valued sequences. Here $[B_{\mathrm{fb}}]_{t,\tau} \in \mathbb{R}$ is scalar and multiplies $s_\tau \in \mathbb{R}^r$, i.e. scalar–vector multiplication. Assume

$$\sup_{t\geq 0} \sum_{\tau<t} |[B_{\mathrm{fb}}]_{t,\tau}| \;\leq\; \rho \;<\; 1.$$

Then for every bounded input $f \in \ell_\infty(\mathbb{N}, \mathbb{R}^r)$ there exists a unique bounded solution $s \in \ell_\infty(\mathbb{N}, \mathbb{R}^r)$ to

$$s = f + B_{\mathrm{fb}} s \qquad \text{equivalently, } (I - B_{\mathrm{fb}})s = f,$$

and it satisfies the explicit bound

$$\|s\|_{\infty,2} \;\leq\; \frac{1}{1-\rho} \|f\|_{\infty,2}. \tag{50}$$

Existence and uniqueness follow by forward substitution: for $t = 0$, $s_0 = f_0$; for $t \geq 1$,

$$s_t = f_t + \sum_{\tau<t} [B_{\mathrm{fb}}]_{t,\tau} s_\tau$$

depends only on previously defined $(s_\tau)_{\tau<t}$. Thus a unique sequence s exists.

For the bound, define the partial maxima

$$M_t := \max_{0\leq k\leq t} \|s_k\|_2 \qquad (t \geq 0).$$

For $t = 0$ we have $s_0 = f_0$, hence $M_0 = \|s_0\|_2 \leq \|f\|_{\infty,2}$. For $t \geq 1$, using the row-sum estimate and $M_{t-1} \geq \|s_\tau\|_2$ for all $\tau < t$,

$$\|s_t\|_2 \leq \|f_t\|_2 + \sum_{\tau<t} |[B_{\mathrm{fb}}]_{t,\tau}| \|s_\tau\|_2 \leq \|f\|_{\infty,2} + \rho\, M_{t-1}.$$

We now prove by induction that for all $t \geq 0$,

$$M_t \leq \frac{1}{1-\rho} \|f\|_{\infty,2}.$$

The base case $t = 0$ holds since $M_0 \leq \|f\|_{\infty,2} \leq \frac{1}{1-\rho}\|f\|_{\infty,2}$. Assume the claim holds for $t-1$ with some $t \geq 1$.

Then the previous estimate gives

$$\|s_t\|_2 \le \|f\|_{\infty,2} + \rho\, M_{t-1} \le \|f\|_{\infty,2} + \rho\, \frac{1}{1-\rho}\|f\|_{\infty,2} = \frac{1}{1-\rho}\|f\|_{\infty,2}.$$

Hence $M_t = \max\{M_{t-1}, \|s_t\|_2\} \le \frac{1}{1-\rho}\|f\|_{\infty,2}$, completing the induction. Taking $\sup_{t\ge 0}$ gives $\|s\|_{\infty,2} = \sup_t \|s_t\|_2 = \sup_t M_t \le \frac{1}{1-\rho}\|f\|_{\infty,2}$, which is (50). □

D.5 Explicit one-block bound without compactness

We now bound one Sessa block on ℓ_∞ balls by tracking constants explicitly.

Lemma D.2 (Token-wise affine bound). *Let $y_t = x_t W + b$ with $W \in \mathbb{R}^{d\times d'}$ and $b \in \mathbb{R}^{d'}$, where the same W and b are used for all tokens. Then for any sequence x, finite or infinite,*

$$\|y\|_{\infty,2} \le \|W\|_2\, \|x\|_{\infty,2} + \|b\|_2,$$

where $\|\cdot\|_2$ is the spectral norm for matrices and Euclidean norm for vectors.

Lemma D.3 (Causal attention is ℓ_∞-nonexpansive). *Let $\mathrm{A}_{\mathrm{fb}} = (\alpha^{\mathrm{fb}}_{t\tau})$ satisfy (45). Then for any value sequence v, the sequence y defined by $y_t := \sum_{\tau<t} \alpha^{\mathrm{fb}}_{t\tau} v_\tau$ satisfies*

$$\|y\|_{\infty,2} \le \|v\|_{\infty,2}.$$

Proof. For $t \ge 1$, y_t is a convex combination of $\{v_\tau\}_{\tau<t}$, hence

$$\|y_t\|_2 \le \sup_{\tau<t} \|v_\tau\|_2 \le \|v\|_{\infty,2}.$$

For $t = 0$ the sum is empty, hence $y_0 = 0$ and $\|y_0\|_2 \le \|v\|_{\infty,2}$ as well. Taking the supremum over $t \ge 0$ gives $\|y\|_{\infty,2} \le \|v\|_{\infty,2}$. □

Proposition 25 (One Sessa block: explicit ball-to-ball bound). *Consider one width-m Sessa block $G : \ell_\infty(\mathbb{N}, \mathbb{R}^m) \to \ell_\infty(\mathbb{N}, \mathbb{R}^m)$. Assume:*

- *the feedback matrix is $B_{\mathrm{fb}}(x) = \Gamma_{\mathrm{fb}}(x)\, \mathrm{A}_{\mathrm{fb}}(x)$ with $\mathrm{A}_{\mathrm{fb}}(x)$ satisfying (45) and $\gamma_t(x) = \tanh(u_t(x))$ as above;*
- *the block produces sequences $f(x), g(x) \in \ell_\infty(\mathbb{N}, \mathbb{R}^r)$ and an output projection $o : \mathbb{R}^r \to \mathbb{R}^m$ given token-wise by*

$$o(z)_t = z_t W^{\mathrm{out}} + b^{\mathrm{out}}, \qquad W^{\mathrm{out}} \in \mathbb{R}^{r\times m}, \quad b^{\mathrm{out}} \in \mathbb{R}^m;$$

- *the block output is $G(x) = x + o(z)$ with $z_t = s_t \odot g_t \in \mathbb{R}^r$ and the solve is in value space:*

$$z_t = s_t \odot g_t \in \mathbb{R}^r, \qquad (I - B_{\mathrm{fb}}(x))s = f(x), \qquad s \in \ell_\infty(\mathbb{N}, \mathbb{R}^r).$$

Suppose there exist explicit constants $c_f, c_g, c_\gamma \ge 0$ and $L_f, L_g, L_{\gamma,\mathrm{pre}} \ge 0$, depending only on the block parameters, such that for all inputs x,

$$\|f(x)\|_{\infty,2} \le c_f + L_f\|x\|_{\infty,2}, \quad \|g(x)\|_{\infty,2} \le c_g + L_g\|x\|_{\infty,2}, \quad \sup_t |u_t(x)| \le c_\gamma + L_{\gamma,\mathrm{pre}}\|x\|_{\infty,2}. \tag{51}$$

Define, for $R \ge 0$,

$$\rho_R := \tanh(c_\gamma + L_{\gamma,\mathrm{pre}} R) \in [0,1), \qquad F_R := c_f + L_f R, \qquad G_R := c_g + L_g R.$$

Then for all x with $\|x\|_{\infty,2} \le R$, the block output satisfies the explicit bound

$$\|G(x)\|_{\infty,2} \;\le\; R \;+\; \|W^{\text{out}}\|_2 \,\frac{F_R\, G_R}{1-\rho_R} \;+\; \|b^{\text{out}}\|_2. \tag{52}$$

Proof. On $\|x\|_{\infty,2} \le R$, (51) gives $\|f\|_{\infty,2} \le F_R$ and $\|g\|_{\infty,2} \le G_R$. Also $\sup_t |u_t(x)| \le c_\gamma + L_{\gamma,\text{pre}} R$, hence $\sup_t |\gamma_t(x)| \le \rho_R$. Using ($\star$), we get $\sup_t \sum_{\tau<t} |[B_{\text{fb}}]_{t,\tau}(x)| \le \rho_R < 1$. Lemma 4.2 then yields

$$\|s\|_{\infty,2} \le \frac{1}{1-\rho_R}\|f\|_{\infty,2} \le \frac{F_R}{1-\rho_R}.$$

For the element-wise product in $\mathbb{R}^r$, for each t,

$$\|z_t\|_2 = \|s_t \odot g_t\|_2 \le \|s_t\|_2\, \|g_t\|_2,$$

since

$$\|s_t \odot g_t\|_2^2 = \sum_i s_{ti}^2 g_{ti}^2 \le \sum_i s_{ti}^2 \Big(\sum_j g_{tj}^2 \Big) = \|s_t\|_2^2 \|g_t\|_2^2.$$

Hence

$$\|z\|_{\infty,2} \le \|s\|_{\infty,2}\, \|g\|_{\infty,2} \le \frac{F_R}{1-\rho_R}\, G_R.$$

Finally, by Lemma D.2 for $o(z) = zW_o + b_o$ and the residual $G(x) = x + o(z)$,

$$\|G(x)\|_{\infty,2} \le \|x\|_{\infty,2} + \|o(z)\|_{\infty,2} \le R + \|W^{\text{out}}\|_2 \|z\|_{\infty,2} + \|b^{\text{out}}\|_2,$$

which gives (52). □

Remark D.4 (Explicit dependence of the constants in (51)). Each branch, including the query, key, and value maps and the MLPs producing f, g, and u and related components, is a finite composition of token-wise affine maps, RoPE_t rotations that are orthogonal and norm-preserving, masked softmax attention as in Lemma D.3, and element-wise nonlinearities whose growth is at most linear on bounded sets. The solve $(I - B_{\text{fb}})s = f$ and the Hadamard product $z = s \odot g$ take place in the value space $\mathbb{R}^r$, while the output projection $o : \mathbb{R}^r \to \mathbb{R}^m$ is token-wise affine. Thus one can always choose $c_\bullet$ and $L_\bullet$ explicitly from the operator norms of the weight matrices involved and the norms of the biases, by repeated use of Lemma D.2 and the inequality $\|\text{GELU}(v)\|_2 \le \|v\|_2$.

E Polynomial decay of token influence in the feedback recursion

E.1 Scalar recursion and impulse response

We work on discrete time $t \in \mathbb{N} = \{0, 1, 2, \dots\}$. Let $(\gamma_t)_{t\ge 0}$ be a sequence in $\mathbb{R}$, and let $\{\alpha_{t,j}^{\text{fb}}\}_{t\ge 1,\ 0\le j<t}$ be nonnegative weights such that, for every $t \ge 1$,

$$\alpha_{t,j}^{\text{fb}} \ge 0, \qquad \sum_{j=0}^{t-1} \alpha_{t,j}^{\text{fb}} \le 1. \tag{53}$$

Given an input sequence $(f_t)_{t\ge 0}$, consider the recursion

$$y_0 = f_0, \qquad y_t = f_t + \gamma_t \sum_{j=0}^{t-1} \alpha_{t,j}^{\text{fb}} y_j, \quad t \ge 1. \tag{54}$$

To isolate the influence of a single token, we consider the impulse input at time 0:

$$f_0 = 1, \qquad f_t = 0 \text{ for } t \geq 1,$$

so that (54) reduces to the impulse response recursion

$$\begin{cases} y_0 = 1, \\ y_t = \gamma_t \displaystyle\sum_{j=0}^{t-1} \alpha_{t,j}^{\mathrm{fb}} y_j, \quad t \geq 1. \end{cases} \tag{55}$$

In the full vector model, y_t can be interpreted as a scalar influence coefficient, e.g. an entry of $(I - B_{\mathrm{fb}})^{-1}$.

E.2 Assumptions

Assumption 26 (Upper envelope on attention)**.** *There exists a constant $c_2 \in (0, \infty)$ such that for all $t \geq 1$ and all $0 \leq j < t$,*

$$\alpha_{t,j}^{\mathrm{fb}} \leq \frac{c_2}{t}, \qquad \textit{and (53) holds.} \tag{56}$$

Remark E.1 (On the size of c_2). Under (53) with $\sum_{j=0}^{t-1} \alpha_{t,j}^{\mathrm{fb}} \leq 1$, the conclusion $c_2 \geq 1$ no longer follows. If one additionally has $\sum_{j=0}^{t-1} \alpha_{t,j}^{\mathrm{fb}} = 1$ for all t, then $c_2 \geq 1$ is necessary.

Assumption 27 (Bounded feedback)**.** *There exists $\gamma_{\max} \in [0, 1)$ such that for all $t \geq 0$,*

$$|\gamma_t| \leq \gamma_{\max}. \tag{57}$$

Define the feedback mass parameter

$$\eta := \gamma_{\max} c_2, \tag{58}$$

and assume the nontrivial feedback regime

$$0 < \eta < 1. \tag{59}$$

Equivalently, define the tail exponent

$$\beta_{\mathrm{tail}} := 1 - \eta = 1 - \gamma_{\max} c_2 \in (0, 1], \tag{60}$$

so that $\eta = 1 - \beta_{\mathrm{tail}}$.

Remark E.2 (Degenerate case $\eta = 0$). If $\eta = 0$ then $\gamma_{\max} = 0$ and hence $\gamma_t = 0$ for all t. The recursion (55) has no feedback and the impulse response is trivial: $y_0 = 1$ and $y_t = 0$ for all $t \geq 1$. We therefore focus on $0 < \eta < 1$ when stating a genuine power-law tail.

E.3 Bounded logits imply near-uniform softmax weights

Bounded logits imply near-uniform softmax weights. This is an immediate specialization of Lemma C.1. Indeed, fix $t \geq 1$ and take the index set $\mathcal{J} = \{0, \dots, t-1\}$ with $n = |\mathcal{J}| = t$. If the logits satisfy $\beth_{\min} \leq \beth_{t,j} \leq \beth_{\max}$ for all $j \in \mathcal{J}$, then the spread is $\Delta_0 = \beth_{\max} - \beth_{\min}$, and Lemma C.1 gives, for all $j < t$,

$$\frac{e^{\beth_{\min} - \beth_{\max}}}{t} \;\leq\; \alpha_{t,j}^{\mathrm{fb}} \;\leq\; \frac{e^{\beth_{\max} - \beth_{\min}}}{t}. \tag{61}$$

In particular, Assumption 26 holds with $c_2 = e^{\beth_{\max} - \beth_{\min}}$.

E.4 Polynomial decay theorem

Theorem 28 (Polynomial decay of the impulse response)**.** *Consider the impulse recursion (55). Suppose Assumptions 26 and 27 hold and $0 < \eta = \gamma_{\max} c_2 < 1$; equivalently $\beta_{\text{tail}} = 1 - \eta \in (0, 1)$. Then for all $t \geq 1$,*

$$|y_t| \leq C\, t^{-\beta_{\text{tail}}}, \qquad \text{where one may take} \qquad C := (1 - \beta_{\text{tail}})\, e^{\,1-\beta_{\text{tail}}} = \eta e^{\eta}. \tag{62}$$

In particular, since $\beta_{\text{tail}} > 0$, we have $\lim_{t\to\infty} y_t = 0$.

Proof. Assume $0 < \eta < 1$. The degenerate case $\eta = 0$ is covered by Remark E.2. Let $z_t := |y_t|$. From (55) and Assumptions 26–27, for $t \geq 1$,

$$z_t = \Big|\gamma_t \sum_{j=0}^{t-1} \alpha_{t,j}^{\text{fb}} y_j\Big| \leq |\gamma_t| \sum_{j=0}^{t-1} \alpha_{t,j}^{\text{fb}} |y_j| \leq \gamma_{\max} \sum_{j=0}^{t-1} \alpha_{t,j}^{\text{fb}} z_j.$$

Define the comparison sequence $(\tilde{y}_t)_{t\geq 0}$ by

$$\tilde{y}_0 = 1, \qquad \tilde{y}_t = \gamma_{\max} \sum_{j=0}^{t-1} \alpha_{t,j}^{\text{fb}} \tilde{y}_j, \quad t \geq 1. \tag{63}$$

By induction on t, using $\alpha_{t,j}^{\text{fb}} \geq 0$, we have $z_t \leq \tilde{y}_t$ for all t, hence

$$|y_t| = z_t \leq \tilde{y}_t \qquad \forall t. \tag{64}$$

Let $s_t := \sum_{k=0}^{t} \tilde{y}_k$. Since $\tilde{y}_k \geq 0$, the sequence s_t is increasing and $s_t \geq 1$. Using (63) and $\alpha_{t,j}^{\text{fb}} \leq c_2/t$ we obtain, for $t \geq 1$,

$$\tilde{y}_t \;= \gamma_{\max} \sum_{j=0}^{t-1} \alpha_{t,j}^{\text{fb}} \tilde{y}_j \leq \gamma_{\max} \sum_{j=0}^{t-1} \frac{c_2}{t} \tilde{y}_j = \frac{\eta}{t}\, s_{t-1}.$$

Therefore,

$$s_t = s_{t-1} + \tilde{y}_t \leq s_{t-1} + \frac{\eta}{t} s_{t-1} = s_{t-1}\Big(1 + \frac{\eta}{t}\Big), \qquad t \geq 1. \tag{65}$$

Taking logarithms and using $\log(1 + x) \leq x$ for $x > -1$,

$$\log s_n \leq \log s_0 + \sum_{t=1}^{n} \log\Big(1 + \frac{\eta}{t}\Big) \leq \sum_{t=1}^{n} \frac{\eta}{t} = \eta H_n,$$

where $H_n = \sum_{t=1}^{n} \frac{1}{t}$ is the n-th harmonic number. Using $H_n \leq 1 + \log n$ for $n \geq 1$ gives

$$s_n \leq e^{\eta} n^{\eta} \qquad \forall n \geq 1. \tag{66}$$

Finally, for $t \geq 1$ we use $s_{t-1} \leq s_t$ and (66):

$$\tilde{y}_t \leq \frac{\eta}{t} s_{t-1} \leq \frac{\eta}{t} s_t \leq \frac{\eta}{t}\, e^{\eta} t^{\eta} = \eta e^{\eta}\, t^{\eta-1}.$$

Since $\eta - 1 = -(1 - \eta) = -\beta_{\text{tail}}$, we obtain $\tilde{y}_t \leq \eta e^{\eta}\, t^{-\beta_{\text{tail}}}$. Combining with (64) yields (62) with $C = \eta e^{\eta}$. □

E.5 Finite-horizon formulation

Corollary E.3 (Finite-horizon bound)**.** *Fix $T \in \mathbb{N}^*$ and consider (55) only for $t \in \{0, 1, \dots, T-1\}$. Assume that Assumptions 26 and 27 hold for all $1 \leq t \leq T-1$ with the same constants c_2 and $\gamma_{\max}$, and $0 < \eta = \gamma_{\max} c_2 < 1$; equivalently $\beta_{\text{tail}} = 1 - \eta \in (0, 1)$. Then (62) holds for all $t \in \{1, \dots, T-1\}$ with the same constant $C = \eta e^{\eta} =$*

$(1-\beta_{\text{tail}})e^{1-\beta_{\text{tail}}}$.

Proof. This is an immediate restriction of Theorem 28 to $1 \le t \le T-1$. □

E.6 Impulse at an arbitrary position j

Corollary E.4 (Decay from an impulse at position j)**.** *Fix an index $j \ge 0$. Consider (54) with the impulse input at j:*

$$f_j = 1, \qquad f_t = 0 \text{ for } t \ne j,$$

and with $y_t = 0$ for $t < j$. Equivalently, $y_0 = 0$ if $j > 0$ and the recursion is started from $t = j$. Assume Assumptions 26–27 and $0 < \eta = \gamma_{\max} c_2 < 1$; equivalently $\beta_{\text{tail}} = 1 - \eta \in (0,1)$. Then for all $t > j$,

$$|y_t| \le C\,(t-j)^{-\beta_{\text{tail}}}, \qquad \text{where one may take} \qquad C := \eta e^{\eta} = (1-\beta_{\text{tail}})e^{1-\beta_{\text{tail}}}.$$

Proof. Define $u_n := |y_{j+n}|$ for $n \ge 0$. Then $u_0 = |y_j| = 1$. For $n \ge 1$, since $y_k = 0$ for $k < j$ and $\alpha^{\text{fb}}_{j+n,k} \ge 0$,

$$u_n = |y_{j+n}| = \Big|\gamma_{j+n} \sum_{k=0}^{j+n-1} \alpha^{\text{fb}}_{j+n,k} y_k\Big| \le |\gamma_{j+n}| \sum_{k=j}^{j+n-1} \alpha^{\text{fb}}_{j+n,k} |y_k| \le \gamma_{\max} \sum_{r=0}^{n-1} \alpha^{\text{fb}}_{j+n,j+r}\, u_r.$$

Moreover, by Assumption 26,

$$\alpha^{\text{fb}}_{j+n,j+r} \le \frac{c_2}{j+n} \le \frac{c_2}{n} \qquad (n \ge 1),$$

since $j + n \ge n$. Thus the sequence u_n satisfies the same comparison inequality as in the proof of Theorem 28, with the same $\gamma_{\max}$ and the envelope c_2/n, so repeating that argument yields

$$u_n \le \eta e^{\eta}\, n^{\eta-1} = (\eta e^{\eta})\, n^{-\beta_{\text{tail}}}.$$

Substituting $n = t - j$ yields the claim. □

F Tightness of the polynomial tail in a realizable regime

This section complements Theorem 28 with the upper bound $O(\ell^{-\beta_{\text{tail}}})$ by exhibiting a concrete diffuse routing regime in which the impulse influence is exactly polynomial, that is, $\Theta(\ell^{-\beta_{\text{tail}}})$. This eliminates the semantic ambiguity that an upper bound alone does not preclude faster decay, for instance exponential decay.

F.1 Gamma-ratio inequality of Gautschi

Lemma F.1 (Gautschi inequality for $0 < \gamma < 1$)**.** *Let $\gamma \in (0,1)$ and $t \ge 1$ be an integer. Then*

$$(t+1)^{\gamma-1} \;\le\; \frac{\Gamma(t+\gamma)}{\Gamma(t+1)} \;\le\; t^{\gamma-1}. \tag{67}$$

Equivalently, with $\beta_{\text{tail}} := 1 - \gamma \in (0,1)$,

$$(t+1)^{-\beta_{\text{tail}}} \le \frac{\Gamma(t+\gamma)}{\Gamma(t+1)} \le t^{-\beta_{\text{tail}}}.$$

Proof. By Gautschi's inequality (Gautschi, 1959), for $x > 0$ and $0 < \gamma < 1$,

$$x^{1-\gamma} < \frac{\Gamma(x+1)}{\Gamma(x+\gamma)} < (x+1)^{1-\gamma}.$$

Setting $x = t$ and taking reciprocals yields

$$(t+1)^{\gamma-1} \le \frac{\Gamma(t+\gamma)}{\Gamma(t+1)} \le t^{\gamma-1},$$

which is (67). □

F.2 Uniform routing yields a $\Theta(\ell^{-\beta_{\text{tail}}})$ tail

We consider the scalar impulse recursion from Section E:

$$y_0 = f_0, \qquad y_t = f_t + \gamma_t \sum_{j=0}^{t-1} \alpha_{t,j}^{\text{fb}} y_j, \quad t \ge 1. \tag{68}$$

Proposition 29 (Tightness under uniform routing)**.** *Assume uniform routing, which is maximally diffuse, and constant positive feedback:*

$$\alpha_{t,j}^{\text{fb}} = \frac{1}{t}\mathbf{1}[j < t], \qquad \gamma_t \equiv \gamma \in (0,1).$$

Consider an impulse at time 0: $f_0 = 1$ and $f_t = 0$ for all $t \ge 1$. Then for every $t \ge 1$ the impulse influence admits the closed form

$$y_t = \frac{\gamma}{\Gamma(1+\gamma)} \cdot \frac{\Gamma(t+\gamma)}{\Gamma(t+1)}. \tag{69}$$

Consequently, letting $\beta_{\text{tail}} := 1 - \gamma \in (0,1)$, one has the two-sided bound

$$\frac{\gamma}{\Gamma(1+\gamma)}\,(t+1)^{-\beta_{\text{tail}}} \;\le\; y_t \;\le\; \frac{\gamma}{\Gamma(1+\gamma)}\,t^{-\beta_{\text{tail}}}, \qquad t \ge 1, \tag{70}$$

and in particular

$$y_t = \Theta(t^{-\beta_{\text{tail}}}) \quad \textit{and hence} \quad y_t = \Omega(t^{-\beta_{\text{tail}}}).$$

Proof. Define partial sums $S_t := \sum_{k=0}^{t} y_k$. Under the stated assumptions and for $t \ge 1$,

$$y_t = \frac{\gamma}{t}\sum_{j=0}^{t-1} y_j = \frac{\gamma}{t} S_{t-1}, \qquad S_t = S_{t-1} + y_t = S_{t-1}\Big(1 + \frac{\gamma}{t}\Big),$$

with $S_0 = y_0 = f_0 = 1$. Thus

$$S_t = \prod_{i=1}^{t}\Big(1 + \frac{\gamma}{i}\Big) = \prod_{i=1}^{t}\frac{i+\gamma}{i} = \frac{\Gamma(t+1+\gamma)}{\Gamma(1+\gamma)\Gamma(t+1)}.$$

Using $y_t = \frac{\gamma}{t} S_{t-1}$ and $\Gamma(t+1) = t\,\Gamma(t)$ gives

$$y_t = \frac{\gamma}{t} \cdot \frac{\Gamma(t+\gamma)}{\Gamma(1+\gamma)\Gamma(t)} = \frac{\gamma}{\Gamma(1+\gamma)} \cdot \frac{\Gamma(t+\gamma)}{\Gamma(t+1)},$$

which is (69). The two-sided bound (70) follows directly from Lemma F.1 with $\beta_{\text{tail}} = 1 - \gamma$. □

Corollary F.2 (Uniform routing with an impulse at an arbitrary source position)**.** *Assume the same explicit uniform-routing regime as in Proposition 29:*

$$\alpha_{t,j}^{\text{fb}} = \frac{1}{t}\mathbf{1}[j < t], \qquad \gamma_t \equiv \gamma \in (0,1).$$

Consider an impulse at time $\tau \geq 0$, i.e.

$$f_\tau = 1, \qquad f_t = 0 \text{ for } t \neq \tau,$$

with $y_t = 0$ for $t < \tau$. Then for every $\ell \geq 1$,

$$y_{\tau+\ell} = \gamma \, \frac{\Gamma(\tau+1)}{\Gamma(\tau+1+\gamma)} \cdot \frac{\Gamma(\tau+\ell+\gamma)}{\Gamma(\tau+\ell+1)}. \tag{71}$$

Consequently, with $\beta_{\text{tail}} := 1-\gamma \in (0,1)$, for every fixed source position τ,

$$y_{\tau+\ell} = \Theta_\tau(\ell^{-\beta_{\text{tail}}}) \qquad (\ell \to \infty).$$

Moreover, the prefactor depends on τ and satisfies

$$\gamma \, \frac{\Gamma(\tau+1)}{\Gamma(\tau+1+\gamma)} \asymp \tau^{-\gamma} \qquad (\tau \to \infty).$$

In particular, there is no positive lower constant $c_- > 0$ such that

$$y_{\tau+\ell} \geq c_- \, \ell^{-\beta_{\text{tail}}}$$

for all source positions τ and all $\ell \geq 1$ on an unbounded horizon.

Proof. Define partial sums

$$S_t := \sum_{k=\tau}^{t} y_k, \qquad t \geq \tau.$$

Then $S_\tau = y_\tau = 1$. For $t \geq \tau+1$, the recursion gives

$$y_t = \frac{\gamma}{t} \sum_{j=\tau}^{t-1} y_j = \frac{\gamma}{t} S_{t-1}, \qquad S_t = S_{t-1} + y_t = S_{t-1}\Big(1 + \frac{\gamma}{t}\Big).$$

Hence, for every $t \geq \tau+1$,

$$S_t = \prod_{i=\tau+1}^{t} \Big(1 + \frac{\gamma}{i}\Big) = \prod_{i=\tau+1}^{t} \frac{i+\gamma}{i} = \frac{\Gamma(t+1+\gamma)\Gamma(\tau+1)}{\Gamma(\tau+1+\gamma)\Gamma(t+1)}.$$

Using $y_t = \frac{\gamma}{t} S_{t-1}$ and $\Gamma(t+1) = t\,\Gamma(t)$ yields

$$y_t = \frac{\gamma}{t} \cdot \frac{\Gamma(t+\gamma)\Gamma(\tau+1)}{\Gamma(\tau+1+\gamma)\Gamma(t)} = \gamma \, \frac{\Gamma(\tau+1)}{\Gamma(\tau+1+\gamma)} \cdot \frac{\Gamma(t+\gamma)}{\Gamma(t+1)}.$$

Setting $t = \tau + \ell$ gives (71).

For fixed τ, the factor

$$\gamma \, \frac{\Gamma(\tau+1)}{\Gamma(\tau+1+\gamma)}$$

is a positive constant depending only on τ, while Lemma F.1 gives

$$(\tau+\ell+1)^{-\beta_{\text{tail}}} \leq \frac{\Gamma(\tau+\ell+\gamma)}{\Gamma(\tau+\ell+1)} \leq (\tau+\ell)^{-\beta_{\text{tail}}}.$$

Since τ is fixed, this implies

$$\frac{\Gamma(\tau+\ell+\gamma)}{\Gamma(\tau+\ell+1)} = \Theta_\tau(\ell^{-\beta_{\text{tail}}}),$$

hence $y_{\tau+\ell} = \Theta_\tau(\ell^{-\beta_{\text{tail}}})$.

The Gamma-ratio asymptotic gives

$$\frac{\Gamma(\tau+1)}{\Gamma(\tau+1+\gamma)} \asymp (\tau+1)^{-\gamma} \qquad (\tau \to \infty),$$

so the source-dependent prefactor decays polynomially with τ.

Moreover, taking $\ell = 1$ in (71) gives

$$y_{\tau+1} = \gamma \frac{\Gamma(\tau+1)}{\Gamma(\tau+1+\gamma)} \cdot \frac{\Gamma(\tau+1+\gamma)}{\Gamma(\tau+2)} = \frac{\gamma}{\tau+1}.$$

Hence $y_{\tau+1} \to 0$ as $\tau \to \infty$. Therefore no positive lower constant independent of τ can satisfy

$$y_{\tau+\ell} \geq c_- \, \ell^{-\beta_{\text{tail}}}$$

for all source positions τ and all $\ell \geq 1$ on an unbounded horizon. □

Corollary F.3 (Uniform two-sided heavy-tail envelope on a bounded source family)**.** *Fix $\tau_{\max} \in \mathbb{N}$. Under the regime of Corollary F.2, there exist constants $c^-_{\tau_{\max},\gamma}, c^+_{\tau_{\max},\gamma} > 0$ such that for every source position $0 \leq \tau \leq \tau_{\max}$ and every $\ell \geq 1$,*

$$c^-_{\tau_{\max},\gamma} \, \ell^{-\beta_{\text{tail}}} \leq y_{\tau+\ell} \leq c^+_{\tau_{\max},\gamma} \, \ell^{-\beta_{\text{tail}}}, \qquad \beta_{\text{tail}} := 1-\gamma.$$

In particular, the explicit uniform-routing regime realizes a uniform two-sided heavy-tail envelope on every bounded source family, and hence on every fixed finite horizon.

Proof. Write

$$a_\tau := \gamma \frac{\Gamma(\tau+1)}{\Gamma(\tau+1+\gamma)}.$$

Since the set $\{0, \dots, \tau_{\max}\}$ is finite and each a_τ is positive,

$$m_{\tau_{\max},\gamma} := \min_{0 \leq \tau \leq \tau_{\max}} a_\tau > 0, \qquad M_{\tau_{\max},\gamma} := \max_{0 \leq \tau \leq \tau_{\max}} a_\tau < \infty.$$

By Corollary F.2,

$$y_{\tau+\ell} = a_\tau \frac{\Gamma(\tau+\ell+\gamma)}{\Gamma(\tau+\ell+1)}.$$

Lemma F.1 yields

$$(\tau+\ell+1)^{-\beta_{\text{tail}}} \leq \frac{\Gamma(\tau+\ell+\gamma)}{\Gamma(\tau+\ell+1)} \leq (\tau+\ell)^{-\beta_{\text{tail}}}.$$

Therefore

$$y_{\tau+\ell} \leq M_{\tau_{\max},\gamma} \, (\tau+\ell)^{-\beta_{\text{tail}}} \leq M_{\tau_{\max},\gamma} \, \ell^{-\beta_{\text{tail}}}.$$

Also, since $0 \leq \tau \leq \tau_{\max}$ and $\ell \geq 1$,

$$\tau+\ell+1 \leq \tau_{\max}+\ell+1 \leq (\tau_{\max}+2)\ell,$$

hence

$$(\tau+\ell+1)^{-\beta_{\text{tail}}} \geq (\tau_{\max}+2)^{-\beta_{\text{tail}}} \, \ell^{-\beta_{\text{tail}}}.$$

Thus

$$y_{\tau+\ell} \geq m_{\tau_{\max},\gamma}\,(\tau+\ell+1)^{-\beta_{\text{tail}}} \geq m_{\tau_{\max},\gamma}\,(\tau_{\max}+2)^{-\beta_{\text{tail}}}\,\ell^{-\beta_{\text{tail}}}.$$

So one may take

$$c^{-}_{\tau_{\max},\gamma} := m_{\tau_{\max},\gamma}\,(\tau_{\max}+2)^{-\beta_{\text{tail}}}, \qquad c^{+}_{\tau_{\max},\gamma} := M_{\tau_{\max},\gamma}.$$

□

Consequence for the influence kernel In the lower-triangular solve $s = Kf$ with $K = (I - B_{\text{fb}})^{-1}$, choosing

$$[B_{\text{fb}}]_{t,j} = \gamma\,\alpha^{\text{fb}}_{t,j} = \begin{cases} 0, & t = 0, \\ \dfrac{\gamma}{t}\,\mathbf{1}[j < t], & t \geq 1, \end{cases}$$

yields that the column $K_{\cdot,0}$ is precisely the impulse response $(y_t)_{t\geq 0}$ above. Hence,

$$|K_{t,0}| = \Theta(t^{-\beta_{\text{tail}}}),$$

so the polynomial envelope in Theorem 8 is sharp, and the rate is attained by a concrete heavy-tailed memory mode.

Remark F.4 (Impulse at time τ). Assume $\gamma \in (0,1)$. The same computation applies to an impulse at time τ. If $f_\tau = 1$, $f_t = 0$ for $t \neq \tau$, and $y_t = 0$ for $t < \tau$, then for $t \geq \tau + 1$

$$y_t = \gamma\,\frac{\Gamma(t+\gamma)\Gamma(\tau+1)}{\Gamma(t+1)\Gamma(\tau+1+\gamma)} = C(\tau,\gamma)\cdot\frac{\Gamma(t+\gamma)}{\Gamma(t+1)},$$

with $C(\tau,\gamma) := \gamma\,\Gamma(\tau+1)/\Gamma(\tau+1+\gamma) > 0$. Hence, for $\ell = t - \tau$, the lag-ℓ tail is again $\Theta(\ell^{-\beta_{\text{tail}}})$ by Lemma F.1, in agreement with Corollary E.4.

G Heavy-tail convolution estimates

Definition 9 (Discrete convolution on positive lags)**.** For nonnegative sequences $a, b : \mathbb{N}^* \to [0,\infty)$, define

$$(a * b)(n) := \sum_{m=1}^{n-1} a(n-m)b(m), \qquad n \geq 2,$$

and $(a * b)(1) := 0$. Inductively define $a^{(*1)} := a$ and $a^{(*k)} := a^{(*(k-1))} * a$ for $k \geq 2$.

Lemma G.1 (Discrete power convolution)**.** *Let $\sigma, \rho > 0$, and define*

$$u_\sigma(n) := n^{\sigma-1}, \qquad u_\rho(n) := n^{\rho-1}, \qquad n \in \mathbb{N}^*.$$

Then there exist constants $c_{\sigma,\rho}, C_{\sigma,\rho} \in (0,\infty)$ such that

$$c_{\sigma,\rho}\,n^{\sigma+\rho-1} \leq (u_\sigma * u_\rho)(n) \leq C_{\sigma,\rho}\,n^{\sigma+\rho-1}, \qquad n \geq 2.$$

Proof. Fix $n \geq 2$.

For the upper bound, split the sum into the two regions

$$1 \leq m \leq \left\lfloor \frac{n}{2} \right\rfloor \qquad \text{and} \qquad \left\lfloor \frac{n}{2} \right\rfloor + 1 \leq m \leq n-1.$$

If $1 \le m \le n/2$, then $n-m \in [n/2, n-1]$, hence

$$(n-m)^{\sigma-1} \le C_\sigma\, n^{\sigma-1}, \qquad C_\sigma := \max\{1, 2^{1-\sigma}\}.$$

Therefore

$$\sum_{m=1}^{\lfloor n/2 \rfloor} (n-m)^{\sigma-1} m^{\rho-1} \le C_\sigma n^{\sigma-1} \sum_{m=1}^{\lfloor n/2 \rfloor} m^{\rho-1}.$$

Since $\rho > 0$, the standard integral comparison gives

$$\sum_{m=1}^{\lfloor n/2 \rfloor} m^{\rho-1} \le 1 + \int_1^{n/2} x^{\rho-1}\, dx \le C'_\rho\, n^\rho$$

for some constant C'_ρ depending only on ρ. Hence

$$\sum_{m=1}^{\lfloor n/2 \rfloor} (n-m)^{\sigma-1} m^{\rho-1} \le C_\sigma C'_\rho\, n^{\sigma+\rho-1}.$$

If $\lfloor n/2 \rfloor + 1 \le m \le n-1$, then $m \in [n/2, n-1]$, hence

$$m^{\rho-1} \le C_\rho\, n^{\rho-1}, \qquad C_\rho := \max\{1, 2^{1-\rho}\}.$$

Therefore

$$\sum_{m=\lfloor n/2 \rfloor+1}^{n-1} (n-m)^{\sigma-1} m^{\rho-1} \le C_\rho n^{\rho-1} \sum_{m=\lfloor n/2 \rfloor+1}^{n-1} (n-m)^{\sigma-1}.$$

After the change of variable $r = n-m$, the inner sum becomes

$$\sum_{r=1}^{\lceil n/2 \rceil - 1} r^{\sigma-1} \le C'_\sigma n^\sigma$$

for some constant C'_σ depending only on σ. Hence

$$\sum_{m=\lfloor n/2 \rfloor+1}^{n-1} (n-m)^{\sigma-1} m^{\rho-1} \le C_\rho C'_\sigma\, n^{\sigma+\rho-1}.$$

Adding the two estimates proves the upper bound.

For the lower bound, restrict the sum to the central block

$$\left\lfloor \frac{n}{4} \right\rfloor \le m \le \left\lfloor \frac{3n}{4} \right\rfloor.$$

For every such m and every $n \ge 4$ one has

$$\frac{n}{4} \le m \le \frac{3n}{4}, \qquad \frac{n}{4} \le n-m \le \frac{3n}{4}.$$

Hence

$$m^{\rho-1} \ge c_\rho\, n^{\rho-1}, \qquad (n-m)^{\sigma-1} \ge c_\sigma\, n^{\sigma-1},$$

where one may take

$$c_\rho := \min\{1, 4^{1-\rho}\}, \qquad c_\sigma := \min\{1, 4^{1-\sigma}\}.$$

Indeed, if $\rho \le 1$, then $m \le n$ implies $m^{\rho-1} \ge n^{\rho-1}$; if $\rho \ge 1$, then $m \ge n/4$ implies $m^{\rho-1} \ge 4^{1-\rho} n^{\rho-1}$. The same

argument applies to $(n-m)^{\sigma-1}$.

Therefore every summand in the central block is bounded below by

$$c_\sigma c_\rho \, n^{\sigma+\rho-2}.$$

The number of integers in the central block is at least $n/2-2$. Consequently, for all $n \geq 8$,

$$(u_\sigma * u_\rho)(n) \geq \left(\frac{n}{2}-2\right) c_\sigma c_\rho \, n^{\sigma+\rho-2} \geq \frac{c_\sigma c_\rho}{4} \, n^{\sigma+\rho-1}.$$

Since only finitely many values $2 \leq n < 8$ remain, their minimum ratio to $n^{\sigma+\rho-1}$ is positive. Adjusting the constant completes the proof. □

Theorem 30 (Heavy-tail convolution class)**.** *Fix $\beta_{\text{tail}} \in (0,1)$ and define*

$$f_{\beta_{\text{tail}}}(n) := n^{-\beta_{\text{tail}}}, \qquad n \in \mathbb{N}^*.$$

Then, for every fixed $k \geq 1$, there exist constants $c_{k,\beta_{\text{tail}}}, C_{k,\beta_{\text{tail}}} \in (0,\infty)$ such that

$$c_{k,\beta_{\text{tail}}} \, n^{k(1-\beta_{\text{tail}})-1} \leq f_{\beta_{\text{tail}}}^{(*k)}(n) \leq C_{k,\beta_{\text{tail}}} \, n^{k(1-\beta_{\text{tail}})-1}, \qquad n \geq k. \tag{72}$$

Proof. Set

$$\sigma := 1-\beta_{\text{tail}} \in (0,1).$$

Then

$$f_{\beta_{\text{tail}}}(n) = n^{-\beta_{\text{tail}}} = n^{\sigma-1} = u_\sigma(n).$$

We prove by induction on k that there exist constants $a_k, b_k > 0$ such that

$$a_k \, n^{k\sigma-1} \leq u_\sigma^{(*k)}(n) \leq b_k \, n^{k\sigma-1}, \qquad n \geq k. \tag{73}$$

For $k=1$, this is exactly

$$u_\sigma(n) = n^{\sigma-1}.$$

Assume now that (73) holds for some $k \geq 1$.

Fix $n \geq k+1$. By definition,

$$u_\sigma^{(*(k+1))}(n) = \sum_{m=1}^{n-1} u_\sigma^{(*k)}(n-m) u_\sigma(m).$$

For the upper bound, note that $u_\sigma^{(*k)}(r) = 0$ for $r < k$, since it is a k-fold convolution of positive-lag sequences. Hence, after enlarging b_k if necessary, we may write

$$u_\sigma^{(*k)}(r) \leq b_k \, r^{k\sigma-1} \qquad \text{for every } r \geq 1.$$

Therefore

$$u_\sigma^{(*(k+1))}(n) \leq b_k \sum_{m=1}^{n-1} (n-m)^{k\sigma-1} m^{\sigma-1}.$$

Applying Lemma G.1 with exponents $k\sigma$ and σ yields

$$u_\sigma^{(*(k+1))}(n) \leq b_{k+1} \, n^{(k+1)\sigma-1}$$

for some constant $b_{k+1} > 0$.

For the lower bound, rewrite the sum using $r := n - m$:

$$u_\sigma^{(*(k+1))}(n) = \sum_{r=1}^{n-1} u_\sigma^{(*k)}(r)\, u_\sigma(n-r).$$

Since $u_\sigma^{(*k)}(r) = 0$ for $r < k$, this becomes

$$u_\sigma^{(*(k+1))}(n) = \sum_{r=k}^{n-1} u_\sigma^{(*k)}(r)\, (n-r)^{\sigma-1}.$$

Applying the lower induction hypothesis on the range $r \geq k$ gives

$$u_\sigma^{(*(k+1))}(n) \geq a_k \sum_{r=k}^{n-1} r^{k\sigma-1}(n-r)^{\sigma-1}.$$

Now write

$$\sum_{r=k}^{n-1} r^{k\sigma-1}(n-r)^{\sigma-1} = \sum_{r=1}^{n-1} r^{k\sigma-1}(n-r)^{\sigma-1} - \sum_{r=1}^{k-1} r^{k\sigma-1}(n-r)^{\sigma-1}.$$

By Lemma G.1, the full sum is bounded below by

$$c\, n^{(k+1)\sigma-1}$$

for some constant $c > 0$ depending only on k and σ.

On the other hand, since $k-1$ is fixed,

$$\sum_{r=1}^{k-1} r^{k\sigma-1}(n-r)^{\sigma-1} \leq C\, n^{\sigma-1}$$

for some constant $C > 0$ depending only on k and σ. Because $k\sigma > 0$, one has

$$n^{\sigma-1} = o\big(n^{(k+1)\sigma-1}\big) \qquad \text{as } n \to \infty.$$

Hence there exist constants $c' > 0$ and N_k such that, for all $n \geq N_k$,

$$\sum_{r=k}^{n-1} r^{k\sigma-1}(n-r)^{\sigma-1} \geq c'\, n^{(k+1)\sigma-1}.$$

Therefore, for all $n \geq N_k$,

$$u_\sigma^{(*(k+1))}(n) \geq a_k c'\, n^{(k+1)\sigma-1}.$$

It remains to treat the finitely many values $k+1 \leq n < N_k$. For each such n, one has $u_\sigma^{(*(k+1))}(n) > 0$ because n can be written as a sum of $k+1$ positive integers. Hence the ratio

$$\frac{u_\sigma^{(*(k+1))}(n)}{n^{(k+1)\sigma-1}}$$

is positive for each of those finitely many n. Taking the minimum of these finitely many positive ratios and $a_k c'$ gives a constant $a_{k+1} > 0$ such that

$$u_\sigma^{(*(k+1))}(n) \geq a_{k+1}\, n^{(k+1)\sigma-1} \qquad \text{for all } n \geq k+1.$$

This closes the induction.

Since $f_{\beta_{\text{tail}}} = u_\sigma$ with $\sigma = 1 - \beta_{\text{tail}}$, we obtain

$$f^{(*k)}_{\beta_{\text{tail}}}(n) \asymp n^{k(1-\beta_{\text{tail}})-1}, \qquad n \geq k.$$

This is (72) □

H Deep Jacobian estimates

H.1 Setup

Fix a depth $N_{\text{layer}} \geq 1$, a finite horizon T, and a compact input set $\mathcal{X}_0$. Let

$$h^{(0)} = x \in \mathcal{X}_0, \qquad h^{(n_{\text{layer}})} = F_{n_{\text{layer}}}(h^{(n_{\text{layer}}-1)}), \qquad n_{\text{layer}} = 1, \dots, N_{\text{layer}},$$

where each $F_{n_{\text{layer}}}$ is causal and continuously differentiable on the relevant compact set

$$\mathcal{X}_{n_{\text{layer}}-1} := F_{n_{\text{layer}}-1} \circ \cdots \circ F_1(\mathcal{X}_0).$$

For each layer n_{layer} and each $0 \leq \tau \leq t \leq T-1$, define the one-block Jacobian block

$$J^{(n_{\text{layer}})}_{t,\tau}(u) := \frac{\partial F_{n_{\text{layer}},t}(u)}{\partial u_\tau} \in \mathbb{R}^{D\times D}, \qquad u \in \mathcal{X}_{n_{\text{layer}}-1}.$$

Define also the full end-to-end Jacobian blocks

$$J^{\text{e2e},(N_{\text{layer}})}_{t,\tau}(x) := \frac{\partial h_t^{(N_{\text{layer}})}(x)}{\partial h_\tau^{(0)}(x)} \in \mathbb{R}^{D\times D}.$$

For scalar lower-triangular kernels $\mathcal{A}, \mathcal{B}$ on

$$\{(t,\tau) : 0 \leq \tau \leq t \leq T-1\},$$

we use the standard kernel product

$$(\mathcal{A}\mathcal{B})(t,\tau) := \sum_{j=\tau}^{t} \mathcal{A}(t,j)\mathcal{B}(j,\tau).$$

H.2 Residual calculus

Theorem 31 (Residual calculus)**.** *Assume that for each layer n_{layer} there exist constants*

$$d_{n_{\text{layer}}} \geq 0, \qquad \lambda_{n_{\text{layer}}} \geq 0,$$

and a scalar lower-triangular kernel

$$K_{n_{\text{layer}}} : \{(t,\tau) : 0 \leq \tau < t \leq T-1\} \to [0,\infty)$$

such that for every $u \in \mathcal{X}_{n_{\text{layer}}-1}$ and every $0 \leq \tau \leq t \leq T-1$,

$$\|J^{(n_{\text{layer}})}_{t,\tau}(u)\| \leq d_{n_{\text{layer}}}\, \mathbf{1}[t=\tau] + \lambda_{n_{\text{layer}}}\, K_{n_{\text{layer}}}(t,\tau)\, \mathbf{1}[\tau < t]. \tag{74}$$

Then, for every $x \in \mathcal{X}_0$, every $0 \le \tau < t \le T-1$, and every depth $N_{\mathrm{layer}} \ge 1$,

$$\|J_{t,\tau}^{\mathrm{e2e},(N_{\mathrm{layer}})}(x)\| \le \sum_{k=1}^{N_{\mathrm{layer}}} \sum_{1 \le n_{\mathrm{layer},1} < \cdots < n_{\mathrm{layer},k} \le N_{\mathrm{layer}}} \left(\prod_{m \notin \{n_{\mathrm{layer},1}, \ldots, n_{\mathrm{layer},k}\}} d_m \right) \cdot \sum_{\tau = i_0 < i_1 < \cdots < i_k = t} \prod_{r=1}^{k} \lambda_{n_{\mathrm{layer},r}} K_{n_{\mathrm{layer},r}}(i_r, i_{r-1}). \tag{75}$$

Moreover, for the diagonal blocks one has

$$\|J_{t,t}^{\mathrm{e2e},(N_{\mathrm{layer}})}(x)\| \le \prod_{n_{\mathrm{layer}}=1}^{N_{\mathrm{layer}}} d_{n_{\mathrm{layer}}}.$$

Proof. For each layer n_{layer}, define the scalar diagonal kernel

$$\mathcal{D}_{n_{\mathrm{layer}}}(t,\tau) := d_{n_{\mathrm{layer}}} \mathbf{1}[t=\tau],$$

and the scalar strictly lower-triangular kernel

$$\mathcal{G}_{n_{\mathrm{layer}}}(t,\tau) := \lambda_{n_{\mathrm{layer}}} K_{n_{\mathrm{layer}}}(t,\tau) \mathbf{1}[\tau < t].$$

Then (74) says precisely that

$$\|J_{t,\tau}^{(n_{\mathrm{layer}})}(u)\| \le \mathcal{D}_{n_{\mathrm{layer}}}(t,\tau) + \mathcal{G}_{n_{\mathrm{layer}}}(t,\tau) \qquad \forall u \in \mathcal{X}_{n_{\mathrm{layer}}-1}.$$

We prove by induction on the depth $p \in \{1, \ldots, N_{\mathrm{layer}}\}$ that

$$\left\| \frac{\partial h_t^{(p)}(x)}{\partial h_\tau^{(0)}(x)} \right\| \le [(\mathcal{D}_p + \mathcal{G}_p) \cdots (\mathcal{D}_1 + \mathcal{G}_1)](t,\tau) \qquad (0 \le \tau \le t \le T-1). \tag{76}$$

For $p=1$, (76) is exactly (74) evaluated at $u = x \in \mathcal{X}_0$.

Assume now that (76) holds for some $p-1 \ge 1$. By the chain rule,

$$\frac{\partial h_t^{(p)}(x)}{\partial h_\tau^{(0)}(x)} = \sum_{j=\tau}^{t} \frac{\partial F_{p,t}(h^{(p-1)}(x))}{\partial h_j^{(p-1)}(x)} \cdot \frac{\partial h_j^{(p-1)}(x)}{\partial h_\tau^{(0)}(x)}.$$

Taking operator norms and using submultiplicativity gives

$$\left\| \frac{\partial h_t^{(p)}(x)}{\partial h_\tau^{(0)}(x)} \right\| \le \sum_{j=\tau}^{t} \left\| \frac{\partial F_{p,t}(h^{(p-1)}(x))}{\partial h_j^{(p-1)}(x)} \right\| \cdot \left\| \frac{\partial h_j^{(p-1)}(x)}{\partial h_\tau^{(0)}(x)} \right\|.$$

Since $h^{(p-1)}(x) \in \mathcal{X}_{p-1}$, the one-block bound (74) applies:

$$\left\| \frac{\partial F_{p,t}(h^{(p-1)}(x))}{\partial h_j^{(p-1)}(x)} \right\| \le \mathcal{D}_p(t,j) + \mathcal{G}_p(t,j).$$

Using the induction hypothesis for the second factor, we get

$$\left\|\frac{\partial h_t^{(p)}(x)}{\partial h_\tau^{(0)}(x)}\right\| \le \sum_{j=\tau}^{t} (\mathcal{D}_p + \mathcal{G}_p)(t,j)\,[(\mathcal{D}_{p-1} + \mathcal{G}_{p-1}) \cdots (\mathcal{D}_1 + \mathcal{G}_1)](j,\tau).$$

This is exactly

$$[(\mathcal{D}_p + \mathcal{G}_p) \cdots (\mathcal{D}_1 + \mathcal{G}_1)](t,\tau),$$

which proves (76) for depth p.

Taking $p = N_{\text{layer}}$ yields

$$\|J_{t,\tau}^{\text{e2e},(N_{\text{layer}})}(x)\| \le [(\mathcal{D}_{N_{\text{layer}}} + \mathcal{G}_{N_{\text{layer}}}) \cdots (\mathcal{D}_1 + \mathcal{G}_1)](t,\tau).$$

We now expand the right-hand side. Since each $\mathcal{D}_{n_{\text{layer}}}$ is diagonal and equals $d_{n_{\text{layer}}} I$ as a kernel, one has the exact product expansion

$$(\mathcal{D}_{N_{\text{layer}}} + \mathcal{G}_{N_{\text{layer}}}) \cdots (\mathcal{D}_1 + \mathcal{G}_1) = \sum_{S \subseteq \{1,\dots,N_{\text{layer}}\}} \left(\prod_{m \notin S} d_m\right) \overrightarrow{\prod_{n_{\text{layer}} \in S}} \mathcal{G}_{n_{\text{layer}}},$$

where the ordered product is taken in increasing layer order. For $\tau < t$, the empty-set term vanishes because it is purely diagonal. Thus

$$\|J_{t,\tau}^{\text{e2e},(N_{\text{layer}})}(x)\| \le \sum_{k=1}^{N_{\text{layer}}} \sum_{1 \le n_{\text{layer},1} < \cdots < n_{\text{layer},k} \le N_{\text{layer}}} \left(\prod_{m \notin \{n_{\text{layer},1},\dots,n_{\text{layer},k}\}} d_m\right) (\mathcal{G}_{n_{\text{layer},k}} \cdots \mathcal{G}_{n_{\text{layer},1}})(t,\tau).$$

Finally, by repeated expansion of the kernel product,

$$(\mathcal{G}_{n_{\text{layer},k}} \cdots \mathcal{G}_{n_{\text{layer},1}})(t,\tau) = \sum_{\tau = i_0 < i_1 < \cdots < i_k = t} \prod_{r=1}^{k} \lambda_{n_{\text{layer},r}}\, K_{n_{\text{layer},r}}(i_r, i_{r-1}),$$

which gives (75).

For the diagonal blocks $\tau = t$, only the empty-set term survives, hence

$$\|J_{t,t}^{\text{e2e},(N_{\text{layer}})}(x)\| \le \prod_{n_{\text{layer}}=1}^{N_{\text{layer}}} d_{n_{\text{layer}}}.$$

□

H.3 A harmonic-kernel bound

For diffuse Transformer blocks the one-block kernel depends on the query time t. The next lemma gives the corresponding convolution bound for

$$\mathcal{H}(t,\tau) := \frac{1}{t+1}\mathbf{1}[\tau < t].$$

Lemma H.1 (Nested harmonic bound)**.** *Fix $k \ge 1$ and define*

$$\mathcal{H}(t,\tau) := \frac{1}{t+1}\mathbf{1}[\tau < t].$$

Then for every $0 \le \tau < t \le T-1$,

$$(\mathcal{H}^k)(t,\tau) \le \frac{1}{t+1} \cdot \frac{H_t^{k-1}}{(k-1)!}, \tag{77}$$

where

$$H_t := \sum_{m=1}^{t} \frac{1}{m}$$

is the t*-th harmonic number, with the convention* $H_0 := 0$*. Consequently, for every fixed* k*,*

$$(\mathcal{H}^k)(t,\tau) \lesssim_k \frac{(\log(1+t))^{k-1}}{t+1}.$$

Proof. For $k = 1$ the claim is immediate:

$$\mathcal{H}(t,\tau) = \frac{1}{t+1}\mathbf{1}[\tau < t] \le \frac{1}{t+1}.$$

Assume now $k \ge 2$. By the kernel-product expansion,

$$(\mathcal{H}^k)(t,\tau) = \sum_{\tau = i_0 < i_1 < \cdots < i_k = t} \prod_{r=1}^{k} \frac{1}{i_r + 1}.$$

Since $i_k = t$, the last factor is exactly $\frac{1}{t+1}$, hence

$$(\mathcal{H}^k)(t,\tau) = \frac{1}{t+1} \sum_{\tau < i_1 < \cdots < i_{k-1} < t} \prod_{r=1}^{k-1} \frac{1}{i_r + 1}.$$

Dropping the lower bound τ only enlarges the sum, so

$$(\mathcal{H}^k)(t,\tau) \le \frac{1}{t+1} \sum_{0 < i_1 < \cdots < i_{k-1} < t} \prod_{r=1}^{k-1} \frac{1}{i_r + 1}.$$

Now expand

$$\left(\sum_{m=1}^{t-1} \frac{1}{m+1} \right)^{k-1}.$$

Every strictly increasing $(k-1)$-tuple

$$0 < i_1 < \cdots < i_{k-1} < t$$

appears exactly $(k-1)!$ times among the ordered monomials in this expansion. Therefore

$$\sum_{0 < i_1 < \cdots < i_{k-1} < t} \prod_{r=1}^{k-1} \frac{1}{i_r + 1} \le \frac{1}{(k-1)!} \left(\sum_{m=1}^{t-1} \frac{1}{m+1} \right)^{k-1} \le \frac{H_t^{k-1}}{(k-1)!}.$$

Substituting this into the previous display gives

$$(\mathcal{H}^k)(t,\tau) \le \frac{1}{t+1} \cdot \frac{H_t^{k-1}}{(k-1)!},$$

which is (77).

Since $H_t \lesssim \log(1+t)$, the logarithmic form follows. □

H.4 Model-specific bounds

Proposition 32 (Deep Transformer bound)**.** *Assume the hypotheses of Theorem 31. Assume in addition that for each layer n_{layer} there exists $a_{n_{\mathrm{layer}}} > 0$ such that*

$$K_{n_{\mathrm{layer}}}(t,\tau) \le \frac{a_{n_{\mathrm{layer}}}}{t+1}, \qquad \tau < t.$$

Fix a bounded source family $0 \le \tau \le \tau_{\max}$. Then for every $x \in \mathcal{X}_0$ and every $\ell \ge 1$ with $\tau + \ell \le T-1$,

$$\left\| J^{\mathrm{e2e},(N_{\mathrm{layer}})}_{\tau+\ell,\tau}(x) \right\| \lesssim_{\tau_{\max},N_{\mathrm{layer}}} \frac{(\log(1+\ell))^{N_{\mathrm{layer}}-1}}{1+\ell}.$$

Proof. Fix an ordered layer subset

$$1 \le n_{\mathrm{layer},1} < \cdots < n_{\mathrm{layer},k} \le N_{\mathrm{layer}}.$$

Define

$$\mathcal{H}(t,\tau) := \frac{1}{t+1}\mathbf{1}[\tau < t].$$

By the assumption on $K_{n_{\mathrm{layer}}}$,

$$K_{n_{\mathrm{layer},r}}(i_r, i_{r-1}) \le a_{n_{\mathrm{layer},r}} \mathcal{H}(i_r, i_{r-1}) \qquad \forall r.$$

Therefore

$$\sum_{\tau = i_0 < \cdots < i_k = t} \prod_{r=1}^{k} \lambda_{n_{\mathrm{layer},r}} K_{n_{\mathrm{layer},r}}(i_r, i_{r-1}) \le \left(\prod_{r=1}^{k} \lambda_{n_{\mathrm{layer},r}} a_{n_{\mathrm{layer},r}} \right) (\mathcal{H}^k)(t,\tau).$$

By Lemma H.1,

$$(\mathcal{H}^k)(t,\tau) \lesssim_k \frac{(\log(1+t))^{k-1}}{t+1}.$$

Insert this estimate into Theorem 31:

$$\|J^{\mathrm{e2e},(N_{\mathrm{layer}})}_{t,\tau}(x)\| \lesssim_{N_{\mathrm{layer}}} \sum_{k=1}^{N_{\mathrm{layer}}} \sum_{1 \le n_{\mathrm{layer},1} < \cdots < n_{\mathrm{layer},k} \le N_{\mathrm{layer}}} \left(\prod_{m \notin \{n_{\mathrm{layer},1}, \ldots, n_{\mathrm{layer},k}\}} d_m \right) \left(\prod_{r=1}^{k} \lambda_{n_{\mathrm{layer},r}} a_{n_{\mathrm{layer},r}} \right) \frac{(\log(1+t))^{k-1}}{t+1}.$$

Since N_{layer} is fixed, the finite sum is bounded by

$$C_{N_{\mathrm{layer}}} \frac{(\log(1+t))^{N_{\mathrm{layer}}-1}}{t+1}.$$

Now restrict to the bounded source family $0 \le \tau \le \tau_{\max}$ and set $t = \tau + \ell$. Then

$$t + 1 = \tau + \ell + 1 \asymp_{\tau_{\max}} 1 + \ell, \qquad \log(1+t) \asymp_{\tau_{\max}} \log(1+\ell),$$

uniformly for $0 \le \tau \le \tau_{\max}$. Hence

$$\left\| J^{\mathrm{e2e},(N_{\mathrm{layer}})}_{\tau+\ell,\tau}(x) \right\| \lesssim_{\tau_{\max},N_{\mathrm{layer}}} \frac{(\log(1+\ell))^{N_{\mathrm{layer}}-1}}{1+\ell}.$$

□

Proposition 33 (Deep Mamba bound under failed freeze time)**.** *Assume the hypotheses of Theorem 31. Assume in addition that for each layer n_{layer} there exist $a_{n_{\mathrm{layer}}} > 0$ and $c_{n_{\mathrm{layer}}} > 0$ such that*

$$K_{n_{\mathrm{layer}}}(t,\tau) \le a_{n_{\mathrm{layer}}} e^{-c_{n_{\mathrm{layer}}}(t-\tau)}, \qquad \tau < t.$$

Set

$$c_* := \min_{1\le n_{\text{layer}}\le N_{\text{layer}}} c_{n_{\text{layer}}}.$$

Then for every $x \in \mathcal{X}_0$ and every $\tau < t$,

$$\left\| J_{t,\tau}^{\text{e2e},(N_{\text{layer}})}(x) \right\| \lesssim_{N_{\text{layer}}} (1+t-\tau)^{N_{\text{layer}}-1} e^{-c_*(t-\tau)}.$$

Proof. Fix an ordered layer subset

$$1 \le n_{\text{layer},1} < \cdots < n_{\text{layer},k} \le N_{\text{layer}}$$

and write $\ell := t-\tau$. For every temporal path $\tau = i_0 < \cdots < i_k = t$, one has

$$\prod_{r=1}^{k} K_{n_{\text{layer},r}}(i_r, i_{r-1}) \le \left(\prod_{r=1}^{k} a_{n_{\text{layer},r}} \right) \exp\left(-\sum_{r=1}^{k} c_{n_{\text{layer},r}} (i_r - i_{r-1}) \right) \le \left(\prod_{r=1}^{k} a_{n_{\text{layer},r}} \right) e^{-c_*\ell}.$$

The number of strictly increasing temporal paths

$$\tau = i_0 < i_1 < \cdots < i_k = t$$

is the number of compositions of ℓ into k positive integers, namely

$$\binom{\ell-1}{k-1},$$

with the convention that this is 0 if $\ell < k$. Therefore

$$\sum_{\tau=i_0<\cdots<i_k=t} \prod_{r=1}^{k} \lambda_{n_{\text{layer},r}} K_{n_{\text{layer},r}}(i_r, i_{r-1}) \le \left(\prod_{r=1}^{k} \lambda_{n_{\text{layer},r}} a_{n_{\text{layer},r}} \right) \binom{\ell-1}{k-1} e^{-c_*\ell}.$$

Insert this estimate into Theorem 31:

$$\| J_{t,\tau}^{\text{e2e},(N_{\text{layer}})}(x) \| \le \sum_{k=1}^{N_{\text{layer}}} \sum_{1\le n_{\text{layer},1}<\cdots<n_{\text{layer},k}\le N_{\text{layer}}} \left(\prod_{m\notin\{n_{\text{layer},1},\ldots,n_{\text{layer},k}\}} d_m \right) \left(\prod_{r=1}^{k} \lambda_{n_{\text{layer},r}} a_{n_{\text{layer},r}} \right) \binom{\ell-1}{k-1} e^{-c_*\ell}.$$

Since N_{layer} is fixed and

$$\binom{\ell-1}{k-1} \lesssim_k (1+\ell)^{k-1},$$

the finite sum is bounded by a constant multiple of

$$(1+\ell)^{N_{\text{layer}}-1} e^{-c_*\ell}.$$

□

Proposition 34 (Deep Sessa bound)**.** *Assume the hypotheses of Theorem 31. Assume in addition that for each layer n_{layer} there exist $a_{n_{\text{layer}}} > 0$ and a common exponent $\beta_{\text{tail}} \in (0,1)$ such that*

$$K_{n_{\text{layer}}}(t,\tau) \le a_{n_{\text{layer}}} (t-\tau)^{-\beta_{\text{tail}}} (1+\log(1+t-\tau)), \qquad \tau < t.$$

Then for every $x \in \mathcal{X}_0$ and every $\tau < t$,

$$\left\| J_{t,\tau}^{\text{e2e},(N_{\text{layer}})}(x) \right\| \lesssim_{N_{\text{layer}},\beta_{\text{tail}}} \sum_{k=1}^{N_{\text{layer}}} (t-\tau)^{k(1-\beta_{\text{tail}})-1} (1+\log(1+t-\tau))^k.$$

In particular, since N_{layer} is fixed,

$$\left\| J_{t,\tau}^{\text{e2e},(N_{\text{layer}})}(x) \right\| \lesssim_{N_{\text{layer}},\beta_{\text{tail}}} (t-\tau)^{N_{\text{layer}}(1-\beta_{\text{tail}})-1} \big(1+\log(1+t-\tau)\big)^{N_{\text{layer}}}.$$

Proof. Fix $\tau < t$ and write $\ell := t - \tau$. Fix an ordered layer subset

$$1 \le n_{\text{layer},1} < \cdots < n_{\text{layer},k} \le N_{\text{layer}}.$$

For every temporal path $\tau = i_0 < \cdots < i_k = t$, set

$$m_r := i_r - i_{r-1} \in \mathbb{N}^*.$$

Then

$$m_1 + \cdots + m_k = \ell.$$

Using the bound on $K_{n_{\text{layer}}}$,

$$\prod_{r=1}^{k} K_{n_{\text{layer},r}}(i_r, i_{r-1}) \le \left(\prod_{r=1}^{k} a_{n_{\text{layer},r}} \right) \prod_{r=1}^{k} m_r^{-\beta_{\text{tail}}} (1+\log(1+m_r)).$$

Since every $m_r \le \ell$, one has

$$1+\log(1+m_r) \le 1+\log(1+\ell).$$

Therefore

$$\prod_{r=1}^{k} K_{n_{\text{layer},r}}(i_r, i_{r-1}) \le \left(\prod_{r=1}^{k} a_{n_{\text{layer},r}} \right) \big(1+\log(1+\ell)\big)^k \prod_{r=1}^{k} m_r^{-\beta_{\text{tail}}}.$$

Summing over all temporal paths from τ to t gives

$$\begin{aligned} &\sum_{\tau=i_0<\cdots<i_k=t} \prod_{r=1}^{k} \lambda_{n_{\text{layer},r}} K_{n_{\text{layer},r}}(i_r, i_{r-1}) \\ &\quad \le \left(\prod_{r=1}^{k} \lambda_{n_{\text{layer},r}} a_{n_{\text{layer},r}} \right) \big(1+\log(1+\ell)\big)^k \sum_{\substack{m_1,\dots,m_k \ge 1 \\ m_1+\cdots+m_k=\ell}} m_1^{-\beta_{\text{tail}}} \cdots m_k^{-\beta_{\text{tail}}}. \end{aligned}$$

The remaining sum is exactly the k-fold positive-lag convolution

$$f_{\beta_{\text{tail}}}^{(*k)}(\ell), \qquad f_{\beta_{\text{tail}}}(n) := n^{-\beta_{\text{tail}}}.$$

By Theorem 30,

$$f_{\beta_{\text{tail}}}^{(*k)}(\ell) \lesssim_{k,\beta_{\text{tail}}} \ell^{k(1-\beta_{\text{tail}})-1}.$$

Hence

$$\sum_{\tau=i_0<\cdots<i_k=t} \prod_{r=1}^{k} \lambda_{n_{\text{layer},r}} K_{n_{\text{layer},r}}(i_r, i_{r-1}) \lesssim_{k,\beta_{\text{tail}}} \left(\prod_{r=1}^{k} \lambda_{n_{\text{layer},r}} a_{n_{\text{layer},r}} \right) \ell^{k(1-\beta_{\text{tail}})-1} \big(1+\log(1+\ell)\big)^k.$$

Insert this estimate into Theorem 31 and sum over

$$k = 1, \dots, N_{\text{layer}}.$$

Since N_{layer} is fixed, the finite sum yields the stated bound.

The final simplified estimate follows because, for $\beta_{\text{tail}} \in (0,1)$, the exponent

$$k(1-\beta_{\text{tail}})-1$$

is increasing in k, so the $k = N_{\text{layer}}$ term dominates the smaller-k terms up to a constant. □

H.5 Horizon-uniform bounds

We now state the horizon-uniform version used in Section 4.2.7.

Theorem 35 (Horizon-uniform residual calculus)**.** *Fix a depth $N_{\text{layer}} \geq 1$. For each horizon $T \geq 1$, let*

$$h^{(0,T)} = x \in \mathcal{X}_0^{(T)}, \qquad h^{(n_{\text{layer}},T)} = F_{n_{\text{layer}}}^{(T)}(h^{(n_{\text{layer}}-1,T)}), \qquad n_{\text{layer}} = 1, \dots, N_{\text{layer}},$$

where $\mathcal{X}_0^{(T)} \subset (\mathbb{R}^D)^T$ is compact and each $F_{n_{\text{layer}}}^{(T)}$ is causal and continuously differentiable on the relevant compact set

$$\mathcal{X}_{n_{\text{layer}}-1}^{(T)} := F_{n_{\text{layer}}-1}^{(T)} \circ \cdots \circ F_1^{(T)}(\mathcal{X}_0^{(T)}).$$

Define the full end-to-end Jacobian blocks by

$$J_{t,\tau}^{\text{e2e},(N_{\text{layer}})}(x;T) := \frac{\partial h_t^{(N_{\text{layer}},T)}(x)}{\partial h_\tau^{(0,T)}(x)} \in \mathbb{R}^{D\times D}, \qquad 0 \leq \tau \leq t \leq T-1.$$

Assume that for each layer n_{layer} there exist constants

$$d_{n_{\text{layer}}} \geq 0, \qquad \lambda_{n_{\text{layer}}} \geq 0,$$

independent of T, and a scalar lower-triangular kernel

$$K_{n_{\text{layer}}} : \{(t,\tau) : 0 \leq \tau < t < \infty\} \to [0,\infty)$$

independent of T, such that for every horizon $T \geq 1$, every $u \in \mathcal{X}_{n_{\text{layer}}-1}^{(T)}$, and every $0 \leq \tau \leq t \leq T-1$,

$$\left\| \frac{\partial F_{n_{\text{layer}},t}^{(T)}(u)}{\partial u_\tau} \right\| \leq d_{n_{\text{layer}}} \mathbf{1}[t=\tau] + \lambda_{n_{\text{layer}}} K_{n_{\text{layer}}}(t,\tau)\, \mathbf{1}[\tau < t].$$

Then for every horizon $T \geq 1$, every $x \in \mathcal{X}_0^{(T)}$, and every $0 \leq \tau < t \leq T-1$,

$$\left\| J_{t,\tau}^{\text{e2e},(N_{\text{layer}})}(x;T) \right\| \leq \sum_{k=1}^{N_{\text{layer}}} \sum_{1 \leq n_{\text{layer},1} < \cdots < n_{\text{layer},k} \leq N_{\text{layer}}} \left(\prod_{m \notin \{n_{\text{layer},1}, \dots, n_{\text{layer},k}\}} d_m \right) \cdot \sum_{\tau = i_0 < i_1 < \cdots < i_k = t} \prod_{r=1}^{k} \lambda_{n_{\text{layer},r}} K_{n_{\text{layer},r}}(i_r, i_{r-1}). \tag{78}$$

Moreover,

$$\left\| J_{t,t}^{\text{e2e},(N_{\text{layer}})}(x;T) \right\| \leq \prod_{n_{\text{layer}}=1}^{N_{\text{layer}}} d_{n_{\text{layer}}}.$$

In particular, the right-hand side of (78) is independent of T.

Proof. Fix a horizon $T \geq 1$. Apply Theorem 31 to the horizon-T stack

$$F_1^{(T)}, \dots, F_{N_{\text{layer}}}^{(T)}$$

on the compact input set $\mathcal{X}_0^{(T)}$. The hypotheses of Theorem 31 are satisfied with the same layerwise constants $d_{n_{\text{layer}}}, \lambda_{n_{\text{layer}}}$ and the same kernels $K_{n_{\text{layer}}}$, because these are assumed to be independent of T. Therefore, for this fixed horizon T, Theorem 31 gives exactly the path-sum bound (78) and the same diagonal estimate.

Since the displayed right-hand side contains no dependence on T, the same bound holds verbatim for every horizon $T \geq 1$. □

Corollary H.2 (Horizon-uniform decay bounds)**.** *Assume the hypotheses of Theorem 35.*

(i) ***Transformer.*** *Assume that for each layer n_{layer} there exists $a_{n_{\text{layer}}} > 0$ such that*

$$K_{n_{\text{layer}}}(t, \tau) \leq \frac{a_{n_{\text{layer}}}}{t+1}, \qquad \tau < t.$$

Fix a bounded source family $0 \leq \tau \leq \tau_{\max}$. Then

$$\sup_{T \geq \tau_{\max}+\ell+1} \sup_{0 \leq \tau \leq \tau_{\max}} \sup_{x \in \mathcal{X}_0^{(T)}} \left\| J_{\tau+\ell,\tau}^{\text{e2e},(N_{\text{layer}})}(x; T) \right\| \lesssim_{\tau_{\max}, N_{\text{layer}}} \frac{(\log(1+\ell))^{N_{\text{layer}}-1}}{1+\ell}.$$

(ii) ***Mamba.*** *Assume that for each layer n_{layer} there exist $a_{n_{\text{layer}}} > 0$ and $c_{n_{\text{layer}}} > 0$ such that*

$$K_{n_{\text{layer}}}(t, \tau) \leq a_{n_{\text{layer}}} e^{-c_{n_{\text{layer}}}(t-\tau)}, \qquad \tau < t.$$

Set $c_* := \min_{n_{\text{layer}}} c_{n_{\text{layer}}}$. Then

$$\sup_{T \geq \ell+1} \sup_{0 \leq \tau \leq T-\ell-1} \sup_{x \in \mathcal{X}_0^{(T)}} \left\| J_{\tau+\ell,\tau}^{\text{e2e},(N_{\text{layer}})}(x; T) \right\| \lesssim_{N_{\text{layer}}} (1+\ell)^{N_{\text{layer}}-1} e^{-c_*\ell}.$$

(iii) ***Sessa.*** *Assume that for each layer n_{layer} there exist $a_{n_{\text{layer}}} > 0$ and a common exponent $\beta_{\text{tail}} \in (0, 1)$ such that*

$$K_{n_{\text{layer}}}(t, \tau) \leq a_{n_{\text{layer}}} (t-\tau)^{-\beta_{\text{tail}}} (1 + \log(1 + t - \tau)), \qquad \tau < t.$$

Then

$$\sup_{T \geq \ell+1} \sup_{0 \leq \tau \leq T-\ell-1} \sup_{x \in \mathcal{X}_0^{(T)}} \left\| J_{\tau+\ell,\tau}^{\text{e2e},(N_{\text{layer}})}(x; T) \right\| \lesssim_{N_{\text{layer}}, \beta_{\text{tail}}} \sum_{k=1}^{N_{\text{layer}}} \ell^{k(1-\beta_{\text{tail}})-1} (1 + \log(1+\ell))^k.$$

In particular, if $N_{\text{layer}}(1 - \beta_{\text{tail}}) < 1$, then the right-hand side tends to 0 as $\ell \to \infty$.

Proof. Apply Theorem 35 and then repeat exactly the kernel-class estimates used in the proofs of Propositions 32, 33, and 34. Because the layerwise envelope parameters are horizon-uniform, the resulting constants are independent of T. Taking the indicated suprema over all admissible horizons therefore leaves the bounds unchanged. For the Transformer case, the passage from $t = \tau + \ell$ to $1 + \ell$ is uniform on bounded-source families $0 \leq \tau \leq \tau_{\max}$. For the Sessa case, the final asymptotic decay to 0 occurs exactly when the largest power

$$\ell^{N_{\text{layer}}(1-\beta_{\text{tail}})-1}$$

has negative exponent, i.e. when $N_{\text{layer}}(1 - \beta_{\text{tail}}) < 1$. □

I Universal approximation for Sessa with adapters

I.1 Preliminaries and notation

Fix $T \geq 3$ and $d_{\text{ext}} \in \mathbb{N}^*$. Inputs are

$$x = (x_0, \dots, x_{T-1}) \in (\mathbb{R}^{d_{\text{ext}}})^T \cong \mathbb{R}^{T \times d_{\text{ext}}},$$

and outputs are in $\mathbb{R}^{T \times d_{\text{ext}}}$. For $X \in \mathbb{R}^{T \times d_{\text{ext}}}$ define

$$\|X\|_F^2 = \sum_{t=0}^{T-1} \|X_t\|_2^2.$$

Let $\mathcal{D} \subset \mathbb{R}^{T \times d_{\text{ext}}}$ be compact and

$$M_{\mathcal{D}} := \sup_{x \in \mathcal{D}} \|x\|_F < \infty.$$

Hence $\|x_t\|_2 \leq M_{\mathcal{D}}$ for all $x \in \mathcal{D}$ and all t.

Definition 10 (Causality)**.** $F : \mathcal{D} \to \mathbb{R}^{T \times d_{\text{ext}}}$ is causal if for every t and all $x, x' \in \mathcal{D}$, $x_{0:t} = x'_{0:t}$ implies $F(x)_t = F(x')_t$.

Lemma I.1 (Prefix factorization of continuous causal maps)**.** *Let*

$$\mathcal{D} \subset \mathbb{R}^{T \times d_{\text{ext}}}$$

be compact and let

$$F : \mathcal{D} \to \mathbb{R}^{T \times d_{\text{ext}}}$$

be continuous and causal. For each $t \in \{0, \dots, T-1\}$, define

$$p_t : \mathcal{D} \to (\mathbb{R}^{d_{\text{ext}}})^{t+1}, \qquad p_t(x) := x_{0:t},$$

and

$$\mathcal{P}_t^{\text{pref}} := p_t(\mathcal{D}).$$

Then there exists a unique continuous map

$$\widehat{F}_t : \mathcal{P}_t^{\text{pref}} \to \mathbb{R}^{d_{\text{ext}}}$$

such that

$$\widehat{F}_t(x_{0:t}) = F(x)_t \qquad \forall x \in \mathcal{D}.$$

Proof. Uniqueness is immediate because p_t is surjective onto $\mathcal{P}_t^{\text{pref}}$.

Causality ensures that $\widehat{F}_t$ is well defined: if $p_t(x) = p_t(x')$, then $x_{0:t} = x'_{0:t}$, hence

$$F(x)_t = F(x')_t.$$

Let

$$\text{pr}_t : \mathbb{R}^{T \times d_{\text{ext}}} \to \mathbb{R}^{d_{\text{ext}}}, \qquad \text{pr}_t(y) := y_t,$$

and define

$$g_t := \text{pr}_t \circ F : \mathcal{D} \to \mathbb{R}^{d_{\text{ext}}}.$$

Then

$$g_t = \widehat{F}_t \circ p_t.$$

Let $C \subset \mathbb{R}^{d_{\text{ext}}}$ be closed. Since g_t is continuous, $g_t^{-1}(C)$ is closed in the compact set $\mathcal{D}$, hence compact. Applying p_t, the image

$$p_t(g_t^{-1}(C))$$

is compact in $\mathcal{P}_t^{\text{pref}}$, hence closed because $\mathcal{P}_t^{\text{pref}}$ is Hausdorff. Moreover,

$$\widehat{F}_t^{-1}(C) = p_t(g_t^{-1}(C)).$$

Therefore $\widehat{F}_t$ is continuous. □

I.2 Architecture and function classes

Sessa blocks of width m Fix an even query–key width $d_k \in 2\mathbb{N}$, a model width $m \in \mathbb{N}^*$, and a tokenwise pre-normalization map

$$\text{Norm} : \mathbb{R}^m \to \mathbb{R}^m$$

applied independently to each token. We consider two choices:

$$\text{Norm} = \text{Id} \qquad \text{and} \qquad \text{Norm} = \text{LN}_{\varepsilon_{\text{ln}}} \quad (\varepsilon_{\text{ln}} > 0).$$

A width-m Sessa block is the block of Section 3 specialized to model width m, and we use the following RoPE convention throughout this section.

Write every $z \in \mathbb{R}^{d_k}$ as

$$z = (z^{(0)}, z^{(1)}, \dots, z^{(d_k/2-1)}), \qquad z^{(r)} \in \mathbb{R}^2.$$

Fix a RoPE base $\vartheta > 1$ and define the standard pairwise frequencies

$$\omega_r := \vartheta^{-2r/d_k}, \qquad r = 0, \dots, d_k/2 - 1.$$

In particular,

$$\omega_0 = 1.$$

For every $\tau \in \mathbb{R}$ define

$$\text{RoPE}_\tau(z) := (R_{\omega_0 \tau} z^{(0)}, R_{\omega_1 \tau} z^{(1)}, \dots, R_{\omega_{d_k/2-1} \tau} z^{(d_k/2-1)}),$$

where R_θ denotes the planar rotation by angle θ. In the architecture, $\tau = t \in \{0, \dots, T-1\}$; in the constructions below we also allow shifts such as $\tau = -\ell$. All diagonalization arguments use only the first rotary pair. Hence, whenever $q, k \in \mathbb{R}^{d_k}$ are supported on that first pair,

$$\langle \text{RoPE}_t(q), \text{RoPE}_j(k) \rangle = \langle R_t q_{1:2},\, R_j k_{1:2} \rangle.$$

The comparison RoPE-Transformer class uses the same convention.

Parameters and dimensions

$$W^{\text{in}} \in \mathbb{R}^{m \times 2m}, \qquad b^{\text{in}} \in \mathbb{R}^{2m}, \qquad W^{\text{out}} \in \mathbb{R}^{m \times m}, \qquad b^{\text{out}} \in \mathbb{R}^m,$$

$$W_{Qf}, W_{Kf}, W_{Qb}, W_{Kb} \in \mathbb{R}^{m \times d_k}, \qquad W_V \in \mathbb{R}^{m \times m},$$

$$w^\gamma \in \mathbb{R}^m, \qquad b^\gamma \in \mathbb{R}.$$

Tokenwise preprocessing Given $x \in \mathbb{R}^{T\times m}$:

$$\begin{aligned}
\tilde{x}_t &= \mathrm{Norm}(x_t) \in \mathbb{R}^m,\\
u_t &= \tilde{x}_t W^{\mathrm{in}} + b^{\mathrm{in}} \in \mathbb{R}^{2m},\\
u_t &= (a_t, g_t), \qquad a_t, g_t \in \mathbb{R}^m,\\
\bar{a}_t &= \mathrm{GELU}(a_t) \in \mathbb{R}^m.
\end{aligned}$$

Attention-feedback operator We fix the attention scale to

$$\sigma_k := d_k^{-1/2}.$$

Define

$$q_t^f = \bar{a}_t W_{Qf}, \qquad k_t^f = \bar{a}_t W_{Kf}, \qquad v_t = \bar{a}_t W_V, \qquad q_t^b = \bar{a}_t W_{Qb}, \qquad k_t^b = \bar{a}_t W_{Kb},$$

with

$$q_t^f, k_t^f, q_t^b, k_t^b \in \mathbb{R}^{d_k}, \qquad v_t \in \mathbb{R}^m.$$

For the causal forward branch ($j \le t$), define

$$\tilde{q}_t^f = \mathrm{RoPE}_t(q_t^f), \qquad \tilde{k}_j^f = \mathrm{RoPE}_j(k_j^f),$$

and define

$$\alpha_{t,j}^{\mathrm{fwd}} = \frac{\exp\Big(\sigma_k \langle \tilde{q}_t^f, \tilde{k}_j^f \rangle\Big)\mathbf{1}[j \le t]}{\sum_{\tau \le t} \exp\Big(\sigma_k \langle \tilde{q}_t^f, \tilde{k}_\tau^f \rangle\Big)}, \qquad f_t = \sum_{j \le t} \alpha_{t,j}^{\mathrm{fwd}} v_j.$$

For the strictly lower feedback branch ($j < t$), define

$$\alpha_{t,j}^{\mathrm{fb}} = \frac{\exp\Big(\sigma_k \langle q_t^b, k_j^b \rangle\Big)\mathbf{1}[j < t]}{\sum_{\tau < t} \exp\Big(\sigma_k \langle q_t^b, k_\tau^b \rangle\Big)}, \qquad \alpha_{0,\cdot}^{\mathrm{fb}} = 0.$$

$$\gamma_t = \tanh(\langle \bar{a}_t, w^\gamma \rangle + b^\gamma) \in (-1, 1).$$

$$[B_{\mathrm{fb}}]_{t,j} = \gamma_t \alpha_{t,j}^{\mathrm{fb}}, \qquad [B_{\mathrm{fb}}]_{t,j} = 0 \text{ for } j \ge t.$$

The mixer output is defined by the exact solve

$$(I - B_{\mathrm{fb}})s = f.$$

Since B_{fb} is strictly lower triangular, the system has a unique solution.

Residual update

$$y_t = x_t + ((s_t \odot g_t)W^{\mathrm{out}} + b^{\mathrm{out}}).$$

Function classes Let

$$\mathrm{ConcreteSessaBlocks}_{\mathrm{Norm}}(d_k, m)$$

denote the set of all width-m concrete Sessa blocks above with the chosen pre-normalization map Norm. Define

$$\Omega^{d_k}_{\mathrm{Sessa,Norm}}(m) := \Big\{ G_{N_{\mathrm{layer}}} \circ \cdots \circ G_1 : \ G_{n_{\mathrm{layer}}} \in \mathrm{ConcreteSessaBlocks}_{\mathrm{Norm}}(d_k, m) \text{ for all } n_{\mathrm{layer}},\ N_{\mathrm{layer}} \in \mathbb{N}^* \Big\}.$$

Tokenwise input and output adapters Fix the external data dimension d_{ext} and a model width $m \geq d_{\text{ext}}$. Define tokenwise affine adapters

$$\text{Embed}(x)_t := x_t W^{\text{emb}} + b^{\text{emb}} \in \mathbb{R}^m, \qquad \text{Unembed}(h)_t := h_t W^{\text{un}} + b^{\text{un}} \in \mathbb{R}^{d_{\text{ext}}}.$$

Parameters and dimensions

$$W^{\text{emb}} \in \mathbb{R}^{d_{\text{ext}} \times m}, \quad b^{\text{emb}} \in \mathbb{R}^m, \qquad W^{\text{un}} \in \mathbb{R}^{m \times d_{\text{ext}}}, \quad b^{\text{un}} \in \mathbb{R}^{d_{\text{ext}}}.$$

$$\text{Unembed} \circ \text{Embed} = \text{Id} \qquad \text{on } \mathbb{R}^{T \times d_{\text{ext}}}.$$

We consider Sessa networks of the form

$$x \;\mapsto\; \text{Unembed}(G(\text{Embed}(x))),$$

with

$$G \in \Omega^{d_k}_{\text{Sessa,Id}}(m)$$

in the main LN-free theorem, and

$$G \in \Omega^{d_k}_{\text{Sessa,LN}_{\varepsilon_{\text{ln}}}}(m)$$

in the LayerNorm extension.

Causal RoPE-Transformer class We also define a causal decoder-only RoPE-Transformer class of functions from $\mathbb{R}^{T \times d_{\text{ext}}} \to \mathbb{R}^{T \times d_{\text{ext}}}$, with internal model width m and adapters.

A width-m RoPE-Transformer block is a standard decoder block operating on $\mathbb{R}^{T \times m}$: it consists of causal self-attention with $j \leq t$, RoPE applied to queries and keys in the logits, and fixed scale $\sigma_k = d_k^{-1/2}$, together with a tokenwise FFN of hidden width r and residual connections in $\mathbb{R}^m$. An absolute positional embedding $E \in \mathbb{R}^{T \times m}$ is added once at the network input. Let $\Omega^{H,d_k,r}_{\text{RoPETr,cau}}(m)$ be the set of finite compositions of such blocks on $\mathbb{R}^{T \times m}$.

Finally define the adapted function class

$$\Omega^{H,d_k,r}_{\text{RoPETr,cau}}(d_{\text{ext}} \to m \to d_{\text{ext}}) := \Big\{ x \mapsto \text{Unembed}(\tilde{g}(\text{Embed}(x) + E)) : \; \tilde{g} \in \Omega^{H,d_k,r}_{\text{RoPETr,cau}}(m), \; E \in \mathbb{R}^{T \times m} \Big\}.$$

I.3 Softmax lemmas

Lemma I.2 (Softmax concentration)**.** *Let $v \in \mathbb{R}^n$ and let $i^* = \arg\max_i v_i$ be unique. Let $\Delta = v_{i^*} - \max_{i \neq i^*} v_i > 0$ and fix $\delta \in (0,1)$. For $\sigma_k > 0$, define $p = \text{softmax}(\sigma_k v)$.*

$$p_{i^*} \geq 1 - \delta \quad \textit{whenever} \quad \sigma_k \Delta \geq \log \frac{n-1}{\delta}.$$

Proof.

$$1 - p_{i^*} = \frac{\sum_{i \neq i^*} e^{\sigma_k v_i}}{\sum_i e^{\sigma_k v_i}} \leq \frac{(n-1) e^{\sigma_k (v_{i^*} - \Delta)}}{e^{\sigma_k v_{i^*}}} = (n-1) e^{-\sigma_k \Delta}.$$

Thus $1 - p_{i^*} \leq \delta$ if $\sigma_k \Delta \geq \log \frac{n-1}{\delta}$. □

Corollary I.3 (Sharpening at fixed attention scale)**.** *Let $v \in \mathbb{R}^n$ and let $i^* = \arg\max_i v_i$ be unique. Let*

$\Delta = v_{i^*} - \max_{i \neq i^*} v_i > 0$, *fix* $\delta \in (0,1)$, *and fix the attention scale* $\sigma_k > 0$. *For* $c > 0$, *define*

$$p^{(c)} := \mathrm{softmax}(\sigma_k\, c^2 v).$$

Then

$$p^{(c)}_{i^*} \geq 1 - \delta \quad \textit{whenever} \quad \sigma_k\, c^2 \Delta \geq \log \frac{n-1}{\delta}.$$

Thus, in the concrete architecture where $\sigma_k = d_k^{-1/2}$ *is fixed, arbitrarily sharp softmax rows are obtained by scaling the query and key vectors by a common factor* c.

Proof. Apply Lemma I.2 to the logits $c^2 v$. □

Lemma I.4 (Error of an almost one-hot mixture)**.** *Let* $(w_j)_{j \in J} \subset \mathbb{R}^m$ *and let* $p_j \geq 0$, $\sum_{j \in J} p_j = 1$. *If* $p_{j^*} \geq 1 - \delta$ *then*

$$\Big\| \sum_{j \in J} p_j w_j - w_{j^*} \Big\|_2 \leq 2\delta \cdot V_{\max},$$

where $V_{\max} := \max_{j \in J} \|w_j\|_2$.

Proof.

$$\sum_j p_j w_j - w_{j^*} = (p_{j^*} - 1) w_{j^*} + \sum_{j \neq j^*} p_j w_j.$$

Since $\sum_{j \neq j^*} p_j = 1 - p_{j^*} \leq \delta$,

$$\Big\| \sum_j p_j w_j - w_{j^*} \Big\|_2 \leq |1 - p_{j^*}| \|w_{j^*}\|_2 + \sum_{j \neq j^*} p_j \|w_j\|_2 \leq 2\delta\, V_{\max},$$

where $V_{\max} := \max_j \|w_j\|_2$. □

I.4 RoPE diagonalization and triangular solve

Lemma I.5 (RoPE-diagonalization)**.** *Fix* $T \geq 2$ *and an even query–key width* $d_k \in 2\mathbb{N}$. *For any* $\delta \in (0,1)$ *there exists a parameter choice with one head and this* d_k *such that for all* t,

$$\alpha^{\mathrm{fwd}}_{t,t} \geq 1 - \delta, \qquad \sum_{\substack{j \leq t \\ j \neq t}} \alpha^{\mathrm{fwd}}_{t,j} \leq \delta.$$

At the architectural scale $\sigma_k = d_k^{-1/2}$, *it suffices to scale the active query/key pair by a common factor* $c_{\mathrm{diag}} > 0$ *such that*

$$\sigma_k\, c_{\mathrm{diag}}^2 \Delta_T \geq \log \frac{T-1}{\delta}, \qquad \Delta_T := 1 - \max_{s \in \{1,\ldots,T-1\}} \cos(s) \;>\; 0.$$

Proof. Under the RoPE convention above, RoPE_t acts pairwise on consecutive 2-dimensional coordinates with frequencies $(\omega_r)_{r=0}^{d_k/2-1}$ and

$$\omega_0 = 1.$$

Activate only the first 2-dimensional pair by choosing

$$q_0 = (1, 0, 0, \ldots, 0) \in \mathbb{R}^{d_k}, \qquad k_0 = (1, 0, 0, \ldots, 0) \in \mathbb{R}^{d_k},$$

and then setting

$$q = c_{\mathrm{diag}} q_0, \qquad k = c_{\mathrm{diag}} k_0.$$

With RoPE, $\tilde{q}_t = \mathrm{RoPE}_t(q)$ and $\tilde{k}_j = \mathrm{RoPE}_j(k)$ satisfy

$$\langle \tilde{q}_t, \tilde{k}_j \rangle = c_{\mathrm{diag}}^2 \cos(t-j),$$

since all coordinate pairs except the first are identically zero, and the first pair rotates with frequency $\omega_0 = 1$. For fixed t and $j \le t$, the unique maximum equals c_{diag}^2 at $j = t$. For $j \neq t$, $s = t-j \in \{1, \dots, T-1\}$ so $\cos(s) \le 1-\Delta_T$. Hence the logit gap is at least $c_{\mathrm{diag}}^2 \Delta_T$. Apply Corollary I.3. □

Lemma I.6 (Mixing error under diagonalization)**.** *Assume $\|v_j\|_2 \le V_{\max}$. If $\alpha_{t,t}^{\mathrm{fwd}} \ge 1-\delta$, then*

$$\Big\| \sum_{j \le t} \alpha_{t,j}^{\mathrm{fwd}} v_j - v_t \Big\|_2 \le 2\delta V_{\max}, \qquad \|f - v\|_F \le 2\delta V_{\max} \sqrt{T}.$$

Proof. Lemma I.4 with $j^* = t$, then sum over t. □

Lemma I.7 (Lower-triangular inversion)**.** *For every input x, $B_{\mathrm{fb}}(x) \in \mathbb{R}^{T \times T}$ is strictly lower-triangular. Hence $B_{\mathrm{fb}}(x)$ is nilpotent, with $B_{\mathrm{fb}}(x)^T = 0$.*

$$(I - B_{\mathrm{fb}}(x))^{-1} = \sum_{k=0}^{T-1} B_{\mathrm{fb}}(x)^k.$$

Proof. A strictly lower-triangular $T \times T$ matrix is nilpotent of index at most T. Hence $B_{\mathrm{fb}}^T = 0$, and the Neumann series terminates after $T-1$ terms. □

I.5 Generating positional codes via feedback

Corollary I.8 (A Sessa block can generate separated positional codes)**.** *Fix any tokenwise pre-normalization map*

$$\mathrm{Norm} : \mathbb{R}^m \to \mathbb{R}^m$$

(applied independently to each token), any even query/key width $d_k \ge 2$, and any model width $m \ge 1$. Then there exists a single width-m concrete Sessa block

$$G^{\mathrm{pos}} \in \mathrm{ConcreteSessaBlocks}_{\mathrm{Norm}}(d_k, m)$$

and vectors $p_0, \dots, p_{T-1} \in \mathbb{R}^m$ such that:

(i) for all $h \in \mathbb{R}^{T \times m}$ and all t,

$$G^{\mathrm{pos}}(h)_t = h_t + p_t;$$

(ii) for any prescribed unit vector $u \in \mathbb{R}^m$, one may choose

$$p_t = (\lambda c_t) u$$

with pairwise distinct scalars $(c_t)_{t=0}^{T-1}$ and some $\lambda > 0$, so that on any compact $\mathcal{K}\_\mathrm{set} \subset \mathbb{R}^{T \times m}$ the scalar sets

$$\mathcal{J}_t := \{ \langle h_t + p_t, u \rangle : \ h \in \mathcal{K}\_\mathrm{set} \}$$

are pairwise disjoint after choosing λ large enough.

Proof. Fix a prescribed unit vector $u \in \mathbb{R}^m$.

The construction does not depend on Norm: setting $W^{\mathrm{in}} = 0$ gives

$$u_t = \tilde{x}_t W^{\mathrm{in}} + b^{\mathrm{in}} = b^{\mathrm{in}},$$

for all t.

Choose $W^{\text{in}} = 0$ and choose b^{in} so that for every token

$$a_t \equiv a_* e_1, \qquad g_t \equiv e_1,$$

for some $a_* > 0$. Set

$$A := \text{GELU}(a_*) > 0.$$

Then

$$\bar{a}_t = A e_1 \qquad \forall t.$$

Choose

$$W_{Qf} = 0, \qquad W_{Kf} = 0.$$

Then all forward logits vanish, so each forward row is a causal probability vector. Choose W_V so that

$$v_t = e_1 \qquad \forall t.$$

Therefore

$$f_t = \sum_{j \le t} \alpha^{\text{fwd}}_{t,j} v_j = e_1 \qquad \forall t.$$

Choose

$$W_{Qb} = 0, \qquad W_{Kb} = 0.$$

Then for $t \ge 1$,

$$\alpha^{\text{fb}}_{t,j} = \frac{1}{t}\mathbf{1}[j < t], \qquad \alpha^{\text{fb}}_{0,\cdot} = 0.$$

Fix any constant $\gamma \in (0, 1)$, and choose

$$w^\gamma = 0, \qquad b^\gamma = \text{arctanh}(\gamma).$$

Then

$$\gamma_t \equiv \gamma, \qquad [B_{\text{fb}}]_{t,j} = \begin{cases} 0, & t = 0, \\ \dfrac{\gamma}{t}\mathbf{1}[j < t], & t \ge 1. \end{cases}$$

Since $f_t = e_1$, we have

$$s_t = c_t e_1,$$

where

$$c_0 = 1, \qquad c_t = 1 + \frac{\gamma}{t}\sum_{j=0}^{t-1} c_j \qquad (t \ge 1).$$

Let

$$S_t := \sum_{j=0}^{t} c_j, \qquad \mu_t := \frac{S_t}{t+1}.$$

Then

$$S_t = \Big(1 + \frac{\gamma}{t}\Big) S_{t-1} + 1,$$

hence

$$\mu_t = \frac{t+\gamma}{t+1}\mu_{t-1} + \frac{1}{t+1}, \qquad \mu_t - \mu_{t-1} = \frac{1 - (1-\gamma)\mu_{t-1}}{t+1}.$$

Since $\mu_0 = 1 < \frac{1}{1-\gamma}$, an induction gives

$$\mu_t < \frac{1}{1-\gamma} \qquad \forall t,$$

so

$$\mu_t - \mu_{t-1} > 0 \qquad \forall t \geq 1.$$

Now

$$c_1 = 1 + \gamma > 1 = c_0,$$

and for $t \geq 1$,

$$c_{t+1} - c_t = \gamma\Big(\frac{S_t}{t+1} - \frac{S_{t-1}}{t}\Big) = \gamma(\mu_t - \mu_{t-1}) > 0.$$

Therefore (c_t) is strictly increasing.

Choose W^{out} so that its first row is $\lambda u^\top$ and all other rows are zero, and set $b^{\text{out}} = 0$. Since

$$s_t \odot g_t = (c_t e_1) \odot e_1 = c_t e_1,$$

the residual update equals

$$(s_t \odot g_t) W^{\text{out}} = c_t(\lambda u) =: p_t.$$

Hence

$$G^{\text{pos}}(h)_t = h_t + p_t.$$

Let $\mathcal{K}\_\text{set} \subset \mathbb{R}^{T\times m}$ be compact and set

$$R := \sup_{h \in \mathcal{K}\_\text{set}} \max_t \|h_t\|_2 < \infty.$$

Then

$$|\langle h_t, u\rangle| \leq R \qquad \forall h \in \mathcal{K}\_\text{set},\ \forall t.$$

Since the c_t are pairwise distinct, let

$$\Delta_c := \min_{s \neq t} |c_s - c_t| > 0.$$

Choose

$$\lambda > \frac{2R}{\Delta_c}.$$

Then the shifted scalar sets

$$\mathcal{I}_t = \{\langle h_t + p_t, u\rangle :\ h \in \mathcal{K}\_\text{set}\} = \{\langle h_t, u\rangle + \lambda c_t :\ h \in \mathcal{K}\_\text{set}\}$$

are pairwise disjoint. □

I.6 Composition error control

Lemma I.9 (Composition error on thickened compacts)**.** *Let (X, d) be a metric space such that closed neighborhoods of compact sets are compact, for example, $X = \mathbb{R}^n$ with the Euclidean metric. Fix a compact $\mathcal{K}\_\text{set}_1 \subset X$ and continuous maps $f_i : X \to X$ for $i = 1, \dots, L$.*

Fix $\rho_{\text{nbhd}} > 0$ and define recursively

$$\widetilde{\mathcal{K}\_\text{set}}_1 := \mathcal{K}\_\text{set}_1, \qquad \mathcal{K}\_\text{set}_{i+1} := f_i(\widetilde{\mathcal{K}\_\text{set}}_i), \qquad \widetilde{\mathcal{K}\_\text{set}}_{i+1} := \overline{\mathcal{N}}_{\rho_{\text{nbhd}}}(\mathcal{K}\_\text{set}_{i+1}) = \{x \in X :\ d(x, \mathcal{K}\_\text{set}_{i+1}) \leq \rho_{\text{nbhd}}\}.$$

Then each $\widetilde{\mathcal{K}\_\text{set}}_i$ is compact.

For every $\varepsilon > 0$ there exist tolerances $\delta_1, \dots, \delta_L > 0$ such that: for any continuous maps $g_i : \widetilde{\mathcal{K}\_\text{set}}_i \to X$ satisfying, for each i,

$$\sup_{x \in \widetilde{\mathcal{K}\_\text{set}}_i} d(f_i(x), g_i(x)) \le \delta_i \quad \text{and} \quad \delta_i \le \rho_{\text{nbhd}},$$

the compositions $F := f_L \circ \cdots \circ f_1$ and $G := g_L \circ \cdots \circ g_1$ are well-defined on $\mathcal{K}\_\text{set}_1$ (and in fact $g_i(\widetilde{\mathcal{K}\_\text{set}}_i) \subset \widetilde{\mathcal{K}\_\text{set}}_{i+1}$), and

$$\sup_{x \in \mathcal{K}\_\text{set}_1} d(F(x), G(x)) \le \varepsilon.$$

Proof. Well-definedness is immediate. Fix i and $x \in \widetilde{\mathcal{K}\_\text{set}}_i$. By definition, $f_i(x) \in \mathcal{K}\_\text{set}_{i+1} = f_i(\widetilde{\mathcal{K}\_\text{set}}_i)$, hence $d(f_i(x), \mathcal{K}\_\text{set}_{i+1}) = 0$. Therefore

$$d(g_i(x), \mathcal{K}\_\text{set}_{i+1}) \le d(g_i(x), f_i(x)) + d(f_i(x), \mathcal{K}\_\text{set}_{i+1}) \le \delta_i \le \rho_{\text{nbhd}},$$

so $g_i(x) \in \widetilde{\mathcal{K}\_\text{set}}_{i+1}$. Thus $g_i(\widetilde{\mathcal{K}\_\text{set}}_i) \subset \widetilde{\mathcal{K}\_\text{set}}_{i+1}$ and all compositions are defined.

The remainder of the proof is by induction on L. For $L = 1$ it is immediate.

Assume the claim holds for $L - 1$. Let

$$F_{<L} := f_{L-1} \circ \cdots \circ f_1, \qquad G_{<L} := g_{L-1} \circ \cdots \circ g_1.$$

Since $\widetilde{\mathcal{K}\_\text{set}}_L$ is compact and f_L is continuous, f_L is uniformly continuous on $\widetilde{\mathcal{K}\_\text{set}}_L$. Pick $\eta > 0$ such that

$$d(u, v) \le \eta \;\Rightarrow\; d(f_L(u), f_L(v)) \le \varepsilon/2 \qquad \forall u, v \in \widetilde{\mathcal{K}\_\text{set}}_L.$$

Set $\delta_L := \min(\rho_{\text{nbhd}}, \varepsilon/2)$. By the inductive hypothesis applied with target accuracy η, choose $\delta_1, \dots, \delta_{L-1} > 0$ so that

$$\sup_{x \in \mathcal{K}\_\text{set}_1} d(F_{<L}(x), G_{<L}(x)) \le \eta.$$

Then for $x \in \mathcal{K}\_\text{set}_1$, noting that $G_{<L}(x) \in \widetilde{\mathcal{K}\_\text{set}}_L$ by well-definedness,

$$d(F(x), G(x)) \le d(f_L(F_{<L}(x)), f_L(G_{<L}(x))) + d(f_L(G_{<L}(x)), g_L(G_{<L}(x))) \le \varepsilon/2 + \delta_L \le \varepsilon.$$

□

Lemma I.10 (Tokenwise GELU approximation)**.** *Let $S \subset \mathbb{R}^m$ be compact and let $\Theta : S \to \mathbb{R}^p$ be continuous. Then for every $\eta > 0$ there exist $r \in \mathbb{N}^*$ and affine maps*

$$A : \mathbb{R}^m \to \mathbb{R}^r, \qquad B : \mathbb{R}^r \to \mathbb{R}^p$$

such that

$$\sup_{z \in S} \|B(\text{GELU}(A(z))) - \Theta(z)\|_2 \le \eta.$$

Moreover, if a larger width $r' \ge r$ is prescribed in advance, the same conclusion still holds with r' in place of r, by padding the hidden layer with unused coordinates.

Proof. Apply the standard one-hidden-layer universal approximation theorem for non-polynomial activations coordinatewise to the components of Θ, and concatenate the resulting hidden units into a single hidden layer. Since GELU is continuous and non-polynomial, the theorem applies; see, e.g., Hornik et al. (1989); Leshno et al. (1993). The padding claim is immediate by adding hidden coordinates with zero incoming and outgoing weights. □

Lemma I.11 (Tokenwise GELU approximation with zero-padding)**.** *Let $S \subset \mathbb{R}^m$ be compact, let $\Theta : S \to \mathbb{R}^{p_0}$ be continuous, let $\eta > 0$, and let $r_0 \in \mathbb{N}^*$. For each $r \geq r_0$, let*

$$E_r : \mathbb{R}^{p_0} \hookrightarrow \mathbb{R}^{p(r)}$$

be a coordinate zero-padding embedding, where $p(r)$ may depend on r. Then there exist $r \geq r_0$ and affine maps

$$A : \mathbb{R}^m \to \mathbb{R}^r, \qquad B : \mathbb{R}^r \to \mathbb{R}^{p(r)}$$

such that

$$\sup_{z \in S} \left\| B(\mathrm{GELU}(A(z))) - E_r(\Theta(z)) \right\|_2 \leq \eta.$$

Proof. By Lemma I.10, there exist $s \in \mathbb{N}^*$ and affine maps

$$\bar{A} : \mathbb{R}^m \to \mathbb{R}^s, \qquad \bar{B} : \mathbb{R}^s \to \mathbb{R}^{p_0}$$

such that

$$\sup_{z \in S} \left\| \bar{B}(\mathrm{GELU}(\bar{A}(z))) - \Theta(z) \right\|_2 \leq \eta.$$

Set

$$r := \max\{r_0, s\}.$$

Let

$$I_{s \to r} : \mathbb{R}^s \hookrightarrow \mathbb{R}^r$$

be the coordinate zero-padding inclusion into the first s coordinates, and let

$$\Pi_{r \to s} : \mathbb{R}^r \to \mathbb{R}^s$$

be the projection onto those first s coordinates. Define

$$A := I_{s \to r} \circ \bar{A}, \qquad B := E_r \circ \bar{B} \circ \Pi_{r \to s}.$$

Then A is affine and B is affine. Since $\mathrm{GELU}(0) = 0$ and GELU acts coordinatewise,

$$\Pi_{r \to s}(\mathrm{GELU}(A(z))) = \Pi_{r \to s}(\mathrm{GELU}(I_{s \to r}\bar{A}(z))) = \mathrm{GELU}(\bar{A}(z)).$$

Hence

$$B(\mathrm{GELU}(A(z))) = E_r(\bar{B}(\mathrm{GELU}(\bar{A}(z)))).$$

Because E_r is coordinate zero-padding, it is an isometric embedding for the Euclidean norm, so

$$\left\| B(\mathrm{GELU}(A(z))) - E_r(\Theta(z)) \right\|_2 = \left\| \bar{B}(\mathrm{GELU}(\bar{A}(z))) - \Theta(z) \right\|_2.$$

Taking the supremum over $z \in S$ gives the claim. □

I.7 Stability of finite-horizon RoPE attention

For fixed T, causal RoPE attention depends continuously on the query, key, and value arrays. The next two lemmas collect the continuity and near-diagonal transport estimates used below.

Lemma I.12 (Stability of finite-horizon RoPE attention)**.** *Fix a horizon $T \geq 1$, number of heads $H \geq 1$, even key/query width $d_k \geq 2$, value width $d_v \geq 1$, attention scale $\sigma_k > 0$, and an output matrix*

$$W^O \in \mathbb{R}^{Hd_v \times m}.$$

Let $\mathcal{K}\_\text{set} \subset \mathbb{R}^{T\times m}$ be compact, and define the compact token set

$$S_{\mathcal{K}\_\text{set}} := \{u_t : \ u \in \mathcal{K}\_\text{set},\ 0 \le t \le T-1\} \subset \mathbb{R}^m.$$

For each head $a = 1, \dots, H$, let

$$q^a, k^a, \hat{q}^a, \hat{k}^a : S_{\mathcal{K}\_\text{set}} \to \mathbb{R}^{d_k}, \qquad v^a, \hat{v}^a : S_{\mathcal{K}\_\text{set}} \to \mathbb{R}^{d_v}$$

be continuous. Let $A, \widehat{A} : \mathcal{K}\_\text{set} \to \mathbb{R}^{T\times m}$ be the corresponding causal RoPE-attention maps: for $u \in \mathcal{K}\_\text{set}$,

$$A(u)_t = \Big(\operatorname{concat}_{a=1}^{H} z_t^a(u)\Big) W^O, \qquad z_t^a(u) := \sum_{j\le t} \alpha_{t,j}^a(u)\, v^a(u_j),$$

where

$$\alpha_{t,j}^a(u) = \frac{\exp\Big(\sigma_k \langle \mathrm{RoPE}_t(q^a(u_t)),\, \mathrm{RoPE}_j(k^a(u_j))\rangle\Big)\mathbf{1}[j \le t]}{\sum_{\tau\le t} \exp\Big(\sigma_k \langle \mathrm{RoPE}_t(q^a(u_t)),\, \mathrm{RoPE}_\tau(k^a(u_\tau))\rangle\Big)},$$

and similarly $\widehat{A}$ is defined from $(\hat{q}^a, \hat{k}^a, \hat{v}^a)$.

Then for every $\varepsilon > 0$ there exists $\eta > 0$ such that

$$\sup_{z\in S_{\mathcal{K}\_\text{set}}} \max_{1\le a\le H} \Big(\|q^a(z) - \hat{q}^a(z)\|_2 + \|k^a(z) - \hat{k}^a(z)\|_2 + \|v^a(z) - \hat{v}^a(z)\|_2\Big) \le \eta$$

implies

$$\sup_{u\in\mathcal{K}\_\text{set}} \|A(u) - \widehat{A}(u)\|_F \le \varepsilon.$$

Proof. Define the finite-dimensional array space

$$\mathcal{X} := \Big((\mathbb{R}^{d_k})^H\Big)^T \times \Big((\mathbb{R}^{d_k})^H\Big)^T \times \Big((\mathbb{R}^{d_v})^H\Big)^T,$$

and equip it with the max norm

$$\|(Q,K,V)\|_{\max} := \max\Big\{ \max_{t,a} \|q_t^a\|_2,\ \max_{t,a} \|k_t^a\|_2,\ \max_{t,a} \|v_t^a\|_2 \Big\}.$$

Let

$$\mathcal{A} : \mathcal{X} \to \mathbb{R}^{T\times m}$$

denote the finite-horizon causal RoPE-attention operator defined by the displayed formulas above. RoPE attention is continuous as a composition of continuous finite-dimensional operations.

Now define continuous maps

$$\Xi, \widehat{\Xi} : \mathcal{K}\_\text{set} \to \mathcal{X}$$

by collecting the tokenwise arrays:

$$\Xi(u) := ((q^a(u_t))_{t,a},\ (k^a(u_t))_{t,a},\ (v^a(u_t))_{t,a}),$$

$$\widehat{\Xi}(u) := ((\hat{q}^a(u_t))_{t,a},\ (\hat{k}^a(u_t))_{t,a},\ (\hat{v}^a(u_t))_{t,a}).$$

Then

$$A = \mathcal{A} \circ \Xi, \qquad \widehat{A} = \mathcal{A} \circ \widehat{\Xi}.$$

The image $\Xi(\mathcal{K}\_\text{set}) \subset \mathcal{X}$ is compact. Fix $\eta_0 > 0$; then its closed η_0-neighborhood

$$\overline{\mathcal{N}}_{\eta_0}(\Xi(\mathcal{K}\_\text{set}))$$

is compact as well. Hence $\mathcal{A}$ is uniformly continuous on this neighborhood. Therefore, for the given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$x, x' \in \overline{\mathcal{N}}_{\eta_0}(\Xi(\mathcal{K}\_\text{set})), \qquad \|x - x'\|_{\max} \le \delta \quad \Longrightarrow \quad \|\mathcal{A}(x) - \mathcal{A}(x')\|_F \le \varepsilon.$$

Set $\eta := \min\{\eta_0, \delta\}$. If the stated tokenwise bound holds, then for every $u \in \mathcal{K}\_\text{set}$,

$$\|\Xi(u) - \widehat{\Xi}(u)\|_{\max} \le \eta,$$

because each of the three summands is individually bounded by η. In particular,

$$\widehat{\Xi}(u) \in \overline{\mathcal{N}}_{\eta_0}(\Xi(\mathcal{K}\_\text{set})).$$

Applying the uniform continuity estimate to $\Xi(u)$ and $\widehat{\Xi}(u)$ gives

$$\|A(u) - \widehat{A}(u)\|_F = \|\mathcal{A}(\Xi(u)) - \mathcal{A}(\widehat{\Xi}(u))\|_F \le \varepsilon \qquad \forall u \in \mathcal{K}\_\text{set}.$$

Taking the supremum over $u \in \mathcal{K}\_\text{set}$ proves the claim. □

Lemma I.13 (Near-diagonal attention transports values)**.** *Fix a horizon $T \ge 1$, an output width $s \ge 1$, and a compact set*

$$\mathcal{K}\_\text{set}' \subset \mathbb{R}^{T \times m}.$$

Let

$$S_{\mathcal{K}\_\text{set}'} := \{u_t : \ u \in \mathcal{K}\_\text{set}', \ 0 \le t \le T-1\} \subset \mathbb{R}^m.$$

Let $\phi, v : S_{\mathcal{K}\_\text{set}'} \to \mathbb{R}^s$ be continuous, and define

$$M_\phi := \sup_{z \in S_{\mathcal{K}\_\text{set}'}} \|\phi(z)\|_2 < \infty.$$

Suppose a one-head causal attention mechanism on $\mathcal{K}\_\text{set}'$ produces weights $\alpha_{t,j}(u)$ and outputs

$$f_t(u) := \sum_{j \le t} \alpha_{t,j}(u)\, v(u_j), \qquad u \in \mathcal{K}\_\text{set}'.$$

Assume that for some $\delta \in (0,1)$ and $\eta \ge 0$,

$$\alpha_{t,t}(u) \ge 1 - \delta \qquad \forall\, u \in \mathcal{K}\_\text{set}', \ \forall\, t \in \{0, \dots, T-1\},$$

and

$$\sup_{z \in S_{\mathcal{K}\_\text{set}'}} \|v(z) - \phi(z)\|_2 \le \eta.$$

Then

$$\sup_{u \in \mathcal{K}\_\text{set}'} \max_{0 \le t \le T-1} \|f_t(u) - \phi(u_t)\|_2 \le 2\delta(M_\phi + \eta) + \eta.$$

Proof. Fix $u \in \mathcal{K}\_\text{set}'$ and $t \in \{0, \dots, T-1\}$. Set

$$w_j := v(u_j) \in \mathbb{R}^s, \qquad 0 \le j \le t.$$

Then $(\alpha_{t,j}(u))_{j\le t}$ is a convex distribution and

$$f_t(u) = \sum_{j\le t} \alpha_{t,j}(u)\, w_j.$$

Moreover,

$$\|w_j\|_2 \le \|\phi(u_j)\|_2 + \|v(u_j) - \phi(u_j)\|_2 \le M_\phi + \eta \qquad \forall j \le t.$$

Since $\alpha_{t,t}(u) \ge 1 - \delta$, Lemma I.4 yields

$$\|f_t(u) - w_t\|_2 \le 2\delta(M_\phi + \eta).$$

Also,

$$\|w_t - \phi(u_t)\|_2 \le \eta.$$

Hence

$$\|f_t(u) - \phi(u_t)\|_2 \le \|f_t(u) - w_t\|_2 + \|w_t - \phi(u_t)\|_2 \le 2\delta(M_\phi + \eta) + \eta.$$

Since this bound is uniform in u and t, the claim follows. □

I.8 Universal approximation for causal RoPE-Transformers with adapters

Lemma I.14 (Universality of causal RoPE-Transformers with adapters)**.** *Let*

$$\mathcal{D} \subset \mathbb{R}^{T\times d_{\text{ext}}}$$

be compact and let

$$F : \mathcal{D} \to \mathbb{R}^{T\times d_{\text{ext}}}$$

be continuous and causal. Then for any $\varepsilon > 0$ there exist finite (H, d_k, r, m) and

$$g \in \Omega^{H,d_k,r}_{\text{RoPETr,cau}}(d_{\text{ext}} \to m \to d_{\text{ext}})$$

such that

$$\sup_{x\in\mathcal{D}} \|F(x) - g(x)\|_F < \varepsilon.$$

Moreover, the construction in the proof allows an arbitrary choice of distinct scalars $(c_t)_{t=0}^{T-1}$ in Paragraph 3, hence an arbitrary absolute embedding E supported on the pos-scalar coordinate of slice $h = 1$ with distinct entries.

Proof. Fix $\varepsilon > 0$.

0. Causal factorization For each $t \in \{0, \dots, T-1\}$, define the compact set of attainable prefixes

$$\mathcal{P}_t^{\text{pref}} := \{(x_0, \dots, x_t) : x \in \mathcal{D}\} \subset (\mathbb{R}^{d_{\text{ext}}})^{t+1}.$$

By Lemma I.1, there exists a unique continuous map

$$\widehat{F}_t : \mathcal{P}_t^{\text{pref}} \to \mathbb{R}^{d_{\text{ext}}}, \qquad \widehat{F}_t(x_0, \dots, x_t) := F(x)_t \quad (x \in \mathcal{D}).$$

Since $\mathcal{P}_t^{\text{pref}}$ is compact in Euclidean space, it is closed in $(\mathbb{R}^{d_{\text{ext}}})^{t+1}$. By Tietze extension applied coordinatewise (Tietze, 1915), extend $\widehat{F}_t$ to a continuous map

$$F_t : (\mathbb{R}^{d_{\text{ext}}})^{t+1} \to \mathbb{R}^{d_{\text{ext}}}$$

such that $F(x)_t = F_t(x_0, \dots, x_t)$ for all $x \in \mathcal{D}$. Let $M_{\mathcal{D}} := \sup_{x\in\mathcal{D}} \|x\|_F$.

1. Model width Set the number of heads to be

$$H := T + 1, \qquad d_k := 2,$$

and choose the per-head value width

$$d_v := d_{\text{ext}} + 2.$$

Define

$$m := H\, d_v = (T+1)(d_{\text{ext}} + 2).$$

We index coordinates of $\mathbb{R}^m$ by head-slices:

$$\mathbb{R}^m \cong \bigoplus_{h=1}^{H} \mathbb{R}^{d_v},$$

and within each slice $\mathbb{R}^{d_v}$ we separate content coordinates, the first d_{ext} coordinates, a constant coordinate with index $d_{\text{ext}} + 1$, and a pos-scalar coordinate with index $d_{\text{ext}} + 2$.

2. Adapters We now fix concrete adapters $\text{Embed}, \text{Unembed}$ of the form introduced in Paragraph I.2. This choice satisfies $\text{Unembed} \circ \text{Embed} = \text{Id}$ on $\mathbb{R}^{T \times d_{\text{ext}}}$. Define the sequence-level affine adapter

$$\text{Embed} : \mathbb{R}^{T \times d_{\text{ext}}} \to \mathbb{R}^{T \times m}$$

tokenwise by placing x_t into the content coordinates of slice $h = 1$, setting the constant coordinate to 1, and all other coordinates to 0:

$$\text{Embed}(x)_t = \Big((x_t,\ 1,\ 0)\ ;\ 0\ ;\ 0\ ;\ \cdots\ ;\ 0 \Big) \in \bigoplus_{h=1}^{H} \mathbb{R}^{d_{\text{ext}}+2}.$$

This is an affine map $x_t \mapsto x_t W^{\text{emb}} + b^{\text{emb}}$ for suitable W^{emb} and b^{emb}.

Define $\text{Unembed} : \mathbb{R}^{T \times m} \to \mathbb{R}^{T \times d_{\text{ext}}}$ tokenwise by reading out the content coordinates of slice $h = 1$:

$$\text{Unembed}(h)_t := (h_t^{(h=1)})_{1:d_{\text{ext}}} \in \mathbb{R}^{d_{\text{ext}}},$$

which is exactly a coordinate projection (equivalently, an affine map with $b^{\text{un}} = 0$) and satisfies $\text{Unembed} \circ \text{Embed} = \text{Id}$ on $\mathbb{R}^{T \times d_{\text{ext}}}$. Thus Unembed is linear and non-expansive in Frobenius norm:

$$\|\text{Unembed}(U) - \text{Unembed}(U')\|_F \le \|U - U'\|_F \qquad \forall U, U' \in \mathbb{R}^{T \times m}.$$

Let $\bar{x} := \text{Embed}(x) \in \mathbb{R}^{T \times m}$. The set $\bar{\mathcal{D}} := \text{Embed}(\mathcal{D})$ is compact.

3. Absolute positional code Choose distinct scalars $c_0, \ldots, c_{T-1} \in \mathbb{R}$ and define $E \in \mathbb{R}^{T \times m}$ by:

$$E_t \text{ is zero in all coordinates except the pos-scalar coordinate of slice } h = 1, \text{ where it equals } c_t.$$

Thus for all $x \in \mathcal{D}$ and all t,

$$(\bar{x}_t + E_t)^{(h=1)}_{d_{\text{ext}}+2} = c_t,$$

i.e. the pos-scalar is exactly c_t, independent of x.

4. Prefix encoding Fix a diagonalization tolerance $\delta \in (0, 1)$, to be chosen sufficiently small later. Under the standing RoPE convention fixed above, when $d_k = 2$ there is only one rotary pair and $\omega_0 = 1$, so

$$\text{RoPE}_t(z) = R_t z$$

with R_t the planar rotation by angle t radians (Su et al., 2021). Construct a single causal RoPE-attention sublayer whose output at time t stores

$$x_t,\ x_{t-1},\ \dots,\ x_0$$

in the content coordinates of slices $h = 2, 3, \dots, t+2$, respectively. Equivalently, lag $\ell = 0, \dots, t$ is stored in slice $h = \ell + 2$, and all active slices $h = 2, \dots, H$ are controlled uniformly via the one-hot estimates below.

Because slice $h = 1$ has a constant coordinate equal to 1, we may choose the linear maps W_h^Q, W_h^K so that for every token representation u:

$$q_t^{(h)} = (u_t^{(h=1)})_{d_{\text{ext}}+1}\, \bar{q}^{(h)} = \bar{q}^{(h)} \in \mathbb{R}^2, \qquad k_j^{(h)} = (u_j^{(h=1)})_{d_{\text{ext}}+1}\, \bar{k} = \bar{k} \in \mathbb{R}^2,$$

for fixed vectors $\bar{q}^{(h)}, \bar{k} \in \mathbb{R}^2$. Fix a scaling factor $c_{\text{pack}} > 0$. We set $\bar{k} = c_{\text{pack}}(1, 0)$ and for head $h \in \{2, \dots, H\}$ set

$$\bar{q}^{(h)} := c_{\text{pack}} \text{RoPE}_{-(h-2)}(1, 0) \in \mathbb{R}^2.$$

Under RoPE inside logits, for $j \le t$,

$$\left\langle \text{RoPE}_t(\bar{q}^{(h)}),\ \text{RoPE}_j(\bar{k}) \right\rangle = c_{\text{pack}}^2 \cos\left((t - (h-2)) - j\right).$$

Define for each (t, h) the maximizer

$$j^*(t, h) \in \arg\max_{0 \le j \le t} \cos\left((t - (h-2)) - j\right).$$

For $h \le t + 2$, the unique maximizer is $j^*(t, h) = t - (h - 2)$, since the maximum value 1 is attained only at argument 0. For $h > t + 2$, all arguments $(t - (h-2)) - j$ are distinct negative integers, and the corresponding cosine values are pairwise distinct (since $\cos(a) = \cos(b)$ implies $a = \pm b + 2\pi k$ for some $k \in \mathbb{Z}$, and for integers a, b this forces $k = 0$ because 2π is irrational, hence $a = \pm b$). Thus the maximizer is unique for every (t, h).

Let

$$v_{t,h}(j) := \cos\left((t - (h-2)) - j\right), \qquad j \in \{0, \dots, t\},$$

and for $t \ge 1$ define

$$\Delta_{t,h} := v_{t,h}(j^*(t, h)) - \max_{j \in \{0,\dots,t\} \setminus \{j^*(t,h)\}} v_{t,h}(j) > 0.$$

Since the set of pairs (t, h) is finite, the uniform gap

$$\Delta_* := \min_{\substack{t \in \{1,\dots,T-1\} \\ h \in \{2,\dots,H\}}} \Delta_{t,h}$$

is strictly positive. For $t = 0$, the row is exactly one-hot.

Choose c_{pack} such that

$$\sigma_k c_{\text{pack}}^2 \Delta_* \ge \log \frac{T-1}{\delta}.$$

Then by Corollary I.3, for every $x \in \mathcal{D}$, every $t \ge 1$, and every head $h \in \{2, \dots, H\}$,

$$\alpha_{t,\, j^*(t,h)}^{\text{fwd},(h)} \ge 1 - \delta.$$

For $t = 0$ the distribution is exactly one-hot on $j = 0$.

For heads $h = 2, \dots, H$, choose W_h^V so that the value vector copies the content coordinates of slice $h = 1$ (and has zeros in the last two coordinates of the head output):

$$v_j^{(h)} = (x_j,\ 0,\ 0) \in \mathbb{R}^{d_{\text{ext}}+2}.$$

For head $h = 1$, set $W_1^V \equiv 0$, so head 1 contributes 0.

Let $f_t \in \mathbb{R}^m$ denote the concatenation of head outputs. Choose $W^O = I_m$. Since slices $h \geq 2$ are initially zero, the residual update

$$h_t \leftarrow h_t + f_t$$

injects the head outputs directly into these slices.

Let $V_{\max} := \sup_{x \in \mathcal{D}} \max_j \|x_j\|_2 \leq M_{\mathcal{D}}$. For each t and each head $h \in \{2, \dots, H\}$, by Lemma I.4,

$$\left\| (f_t^{(h)})_{1:d_{\text{ext}}} - x_{j^*(t,h)} \right\|_2 \leq 2\delta\, V_{\max} \leq 2\delta\, M_{\mathcal{D}}.$$

In particular, for $h \leq t+2$ we have $j^*(t,h) = t-(h-2)$, hence slices $h = 2, \dots, t+2$ recover $(x_t, x_{t-1}, \dots, x_0)$ with per-slice content error at most $2\delta\, V_{\max}$.

5. Ideal encoded state and target map Fix $H := T+1$ heads indexed by $h = 1, \dots, H$, with head $h = 1$ unused as before. For each (t,h) with $t \in \{0, \dots, T-1\}$ and $h \in \{2, \dots, H\}$ define the deterministic index

$$j^*(t,h) \in \arg\max_{0 \leq j \leq t} \cos\left((t-(h-2)) - j\right).$$

With the same c_{pack} chosen in Paragraph 4 so that

$$\sigma_k\, c_{\text{pack}}^2 \Delta_* \;\geq\; \log \frac{T-1}{\delta},$$

Corollary I.3 gives, for every $x \in \mathcal{D}$, every $t \geq 1$, and every head $h \in \{2, \dots, H\}$, the causal attention distribution over $j \leq t$ satisfies

$$\alpha^{\text{fwd},(h)}_{t,\, j^*(t,h)} \;\geq\; 1 - \delta.$$

For $t = 0$ the attention is exactly one-hot.

Define $\hat{h}_t(x) \in \mathbb{R}^m$, where $m = (T+1)(d_{\text{ext}}+2)$, by letting slice $h = 1$ equal $(x_t, 1, c_t)$ in coordinates $(1{:}d_{\text{ext}},\ d_{\text{ext}}{+}1,\ d_{\text{ext}}{+}2)$ and zero elsewhere, and for each slice $h = 2, \dots, H$ placing $x_{j^*(t,h)}$ in the first d_{ext} coordinates and zeros in the last two; and set

$$\widehat{S} := \{\hat{h}_t(x) : x \in \mathcal{D},\ t \in \{0, \dots, T-1\}\} \subset \mathbb{R}^m.$$

Then $\widehat{S}$ is compact as a continuous image of a compact set.

For each fixed $t \in \{0, \dots, T-1\}$, define the affine map, in fact linear,

$$\text{Read}_t : \mathbb{R}^m \to (\mathbb{R}^{d_{\text{ext}}})^{t+1}$$

by reading the content coordinates of slices $h = 2, \dots, t+2$ in reverse order:

$$\text{Read}_t(u) := \Big((u^{(t+2)})_{1:d_{\text{ext}}},\ (u^{(t+1)})_{1:d_{\text{ext}}},\ \dots,\ (u^{(2)})_{1:d_{\text{ext}}} \Big).$$

Equivalently, for $\ell = 0, \dots, t$,

$$\big(\text{Read}_t(u)\big)_\ell = (u^{(t-\ell+2)})_{1:d_{\text{ext}}}.$$

By construction of the ideal encoded state and because $j^*(t,h) = t-(h-2)$ for $h \leq t+2$,

$$\text{Read}_t(\hat{h}_t(x)) = (x_0, \dots, x_t) \qquad \forall x \in \mathcal{D}.$$

Thus the pos-scalar coordinate identifies t, while the encoded slices determine the prefix $(x_0, \dots, x_t)$.

Decompose $\widehat{S}$ as the finite disjoint union $\widehat{S} = \bigsqcup_{t=0}^{T-1} \widehat{S}_t$ where $\widehat{S}_t := \{\hat{h}_t(x) : x \in \mathcal{D}\}$. Each $\widehat{S}_t$ is compact and

contained in the affine hyperplane $\{u \in \mathbb{R}^m : \ (u^{(h=1)})_{d_{\text{ext}}+2} = c_t\}$. Since the scalars c_t are distinct, the sets $\widehat{S}_t$ are pairwise separated. Therefore $\widehat{\Phi}$ is continuous on $\widehat{S}$ once each restriction $\widehat{\Phi}|_{\widehat{S}_t}$ is continuous. Now fix t. For every $u = \hat{h}_t(x) \in \widehat{S}_t$, by the defining property of F_t from Paragraph 0 and by the readout identity above,

$$\widehat{\Phi}(u) = F(x)_t = F_t(x_0, \dots, x_t) = F_t(\text{Read}_t(u)).$$

Therefore

$$\widehat{\Phi}|_{\widehat{S}_t} = F_t \circ \text{Read}_t|_{\widehat{S}_t}.$$

Read_t is a linear map, and $F_t : (\mathbb{R}^{d_{\text{ext}}})^{t+1} \to \mathbb{R}^{d_{\text{ext}}}$ is continuous, so $\widehat{\Phi}|_{\widehat{S}_t}$ is continuous. Thus $\widehat{\Phi}$ is continuous on $\widehat{S}$.

By Tietze extension applied coordinatewise, extend $\widehat{\Phi}$ to a continuous $\widetilde{\Phi} : \mathbb{R}^m \to \mathbb{R}^{d_{\text{ext}}}$.

6. FFN approximation Let $h_t^{\text{enc}}(x) \in \mathbb{R}^m$ denote the token state after the first RoPE-attention block, constructed in Paragraph 4, with $W^O = I_m$, head $h = 1$ set to zero, and the FFN set to zero. Slice $h = 1$ is unchanged by the residual, since the concatenated head output has zero slice $h = 1$, so $(h_t^{\text{enc}}(x))^{(h=1)} = (x_t, 1, c_t)$ exactly.

For each head slice $h \in \{2, \dots, H\}$, by the encoding construction in Paragraph 4 we have $\|v_j^{(h)}\|_2 \le V_{\max}$ and $\alpha_{t,\, j^*(t,h)}^{\text{fwd},(h)} \ge 1 - \delta$. Therefore Lemma I.4 gives, for each $x \in \mathcal{D}$, each t, each $h \in \{2, \dots, H\}$,

$$\left\| (h_t^{\text{enc}}(x))_{1:d_{\text{ext}}}^{(h)} - x_{j^*(t,h)} \right\|_2 \ \le \ 2\delta V_{\max},$$

and the last two coordinates of each slice are exactly zero on both sides. Therefore, for each (x, t),

$$\left\| h_t^{\text{enc}}(x) - \hat{h}_t(x) \right\|_2 \ \le \ \sqrt{\sum_{h=2}^{H} (2\delta V_{\max})^2} \ = \ 2\delta V_{\max}\sqrt{T}.$$

In particular,

$$\sup_{x \in \mathcal{D}} \max_t \ \left\| h_t^{\text{enc}}(x) - \hat{h}_t(x) \right\|_2 \ \le \ 2\delta V_{\max}\sqrt{T}.$$

Let

$$S_{\text{enc}} := \{h_t^{\text{enc}}(x) : x \in \mathcal{D}, \ t = 0, \dots, T-1\} \subset \mathbb{R}^m$$

(compact). Since $\widehat{S}$ is compact, for every radius $r_{\text{nbhd}} > 0$ the closed neighborhood

$$\overline{\mathcal{N}}_{r_{\text{nbhd}}}(\widehat{S}) := \{u \in \mathbb{R}^m : \ \text{dist}(u, \widehat{S}) \le r_{\text{nbhd}}\}$$

is compact. Fix such an $r_{\text{nbhd}} > 0$.

By uniform continuity of $\widetilde{\Phi}$ on the compact set $\overline{\mathcal{N}}_{r_{\text{nbhd}}}(\widehat{S})$, there exists a continuity tolerance

$$\delta_{\text{UC}} > 0$$

such that

$$u, v \in \overline{\mathcal{N}}_{r_{\text{nbhd}}}(\widehat{S}), \qquad \|u - v\|_2 \le \delta_{\text{UC}} \quad \Longrightarrow \quad \|\widetilde{\Phi}(u) - \widetilde{\Phi}(v)\|_2 \le \varepsilon/(3\sqrt{T}).$$

Now choose the diagonalization parameter $\delta \in (0, 1)$ above small enough so that

$$2\delta V_{\max}\sqrt{T} \le \min\{r_{\text{nbhd}}, \delta_{\text{UC}}\}.$$

Then $S_{\text{enc}} \subset \overline{\mathcal{N}}_{r_{\text{nbhd}}}(\widehat{S})$, and for all $x \in \mathcal{D}$ and all t,

$$\|h_t^{\text{enc}}(x) - \hat{h}_t(x)\|_2 \le \delta_{\text{UC}}.$$

Hence

$$\|\widetilde{\Phi}(h_t^{\text{enc}}(x)) - \widetilde{\Phi}(\hat{h}_t(x))\|_2 \le \varepsilon/(3\sqrt{T}).$$

Since $\widetilde{\Phi}(\hat{h}_t(x)) = \widehat{\Phi}(\hat{h}_t(x)) = F(x)_t$ by construction, it follows that

$$\|\widetilde{\Phi}(h_t^{\text{enc}}(x)) - F(x)_t\|_2 \le \varepsilon/(3\sqrt{T}).$$

Define the continuous map $\Psi : S_{\text{enc}} \to \mathbb{R}^{d_{\text{ext}}}$ by

$$\Psi(u) := \widetilde{\Phi}(u) - (u^{(h=1)})_{1:d_{\text{ext}}},$$

i.e. the increment needed (in slice $h = 1$ content) to turn the current content into $\widetilde{\Phi}(u)$. By the universal approximation theorem for tokenwise GELU FFNs (Leshno et al., 1993; Hornik et al., 1989), there exists a tokenwise FFN (hidden width r large enough) whose output $\text{FFN}(h)_t \in \mathbb{R}^m$ is supported only on slice $h = 1$ content coordinates and satisfies

$$\sup_{u \in S_{\text{enc}}} \left\| (\text{FFN}(u))_{1:d_{\text{ext}}}^{(h=1)} - \Psi(u) \right\|_2 \;\le\; \varepsilon/(3\sqrt{T}),$$

and $\text{FFN}(u)$ equals 0 on all other coordinates. Applying this tokenwise, define the sequence-level FFN by $\text{FFN}(h)_t := \text{FFN}(h_t)$. Using the residual connection in the second block (with its attention set to zero), the slice $h = 1$ content becomes

$$(h_t^{\text{enc}}(x))_{1:d_{\text{ext}}}^{(h=1)} + (\text{FFN}(h_t^{\text{enc}}(x)))_{1:d_{\text{ext}}}^{(h=1)} \approx \widetilde{\Phi}(h_t^{\text{enc}}(x)) \approx F(x)_t.$$

Combining the encoding and FFN errors yields for each t

$$\left\| (h_t^{\text{out}}(x))_{1:d_{\text{ext}}}^{(h=1)} - F(x)_t \right\|_2 \le \varepsilon/\sqrt{T},$$

hence $\|F(x) - g(x)\|_F \le \varepsilon$ uniformly on $\mathcal{D}$ after applying Unembed. □

I.9 Direct Sessa building blocks

Storage decomposition Fix a model width

$$m = (T+1)d_{\text{ext}} + 2.$$

Write $\mathbb{R}^m$ as the orthogonal direct sum of coordinate subspaces

$$\mathbb{R}^m = U_0 \oplus U_1 \oplus \cdots \oplus U_{T-1} \oplus U_{\text{out}} \oplus \text{span}\{e_{\text{const}}, e_{\text{pos}}\},$$

where each U_ℓ is a coordinate copy of $\mathbb{R}^{d_{\text{ext}}}$ and U_{out} is a coordinate copy of $\mathbb{R}^{d_{\text{ext}}}$.

Fix linear isometries

$$J_\ell : \mathbb{R}^{d_{\text{ext}}} \to U_\ell \qquad (\ell = 0, \dots, T-1), \qquad J_{\text{out}} : \mathbb{R}^{d_{\text{ext}}} \to U_{\text{out}},$$

and let

$$R_\ell := J_\ell^{-1} : U_\ell \to \mathbb{R}^{d_{\text{ext}}}, \qquad R_{\text{out}} := J_{\text{out}}^{-1} : U_{\text{out}} \to \mathbb{R}^{d_{\text{ext}}}.$$

Let $\pi_\ell : \mathbb{R}^m \to U_\ell$ denote the projection onto U_ℓ, let $\pi_{\text{out}} : \mathbb{R}^m \to U_{\text{out}}$ denote the projection onto U_{out}, and

let

$$\pi_{\mathrm{st}} : \mathbb{R}^m \to U_0 \oplus \cdots \oplus U_{T-1} \oplus \mathrm{span}\{e_{\mathrm{const}}, e_{\mathrm{pos}}\}$$

denote the projection onto the storage slice.

For each $\ell \in \{1, \dots, T-1\}$, let

$$T_{0\to\ell} := J_\ell \circ R_0 : U_0 \to U_\ell$$

denote the fixed coordinate-copy isomorphism, and let

$$T_{0\to\mathrm{out}} := J_{\mathrm{out}} \circ R_0 : U_0 \to U_{\mathrm{out}}$$

denote the corresponding copy map into the output slice.

Let

$$\iota_{\mathrm{st}} : \pi_{\mathrm{st}}(\mathbb{R}^m) \to \mathbb{R}^m$$

denote the linear lift obtained by restoring the output slice as the copy of U_0, i.e.

$$\pi_{\mathrm{st}}(\iota_{\mathrm{st}}(z)) = z, \qquad \pi_{\mathrm{out}}(\iota_{\mathrm{st}}(z)) = T_{0\to\mathrm{out}}(\pi_0(z)).$$

Lemma I.15 (Uniform small-signal linearization of GELU)**.** *Let $K \subset \mathbb{R}^q$ be compact. Then*

$$\sup_{u\in K} \left\| \frac{2}{\varepsilon} \mathrm{GELU}(\varepsilon u) - u \right\|_2 \longrightarrow 0 \qquad \text{as } \varepsilon \downarrow 0.$$

Consequently, for every compact $K \subset \mathbb{R}^p$, every linear map $L : \mathbb{R}^p \to \mathbb{R}^q$, and every $\eta > 0$, there exists $\varepsilon > 0$ such that

$$\sup_{z\in K} \left\| \frac{2}{\varepsilon} \mathrm{GELU}(\varepsilon L z) - L z \right\|_2 \le \eta.$$

Proof. GELU is C^1 and $\mathrm{GELU}'(0) = 1/2$. Hence

$$\mathrm{GELU}(u) = \frac{1}{2}u + r(u), \qquad \frac{\|r(u)\|_2}{\|u\|_2} \to 0 \quad \text{as } u \to 0.$$

Apply this uniformly on the compact set εK. The second statement follows by substituting $u = Lz$. □

Lemma I.16 (A single Sessa block copies one lag into a dedicated slice)**.** *Fix $\ell \in \{1, \dots, T-1\}$ and a compact set $\mathcal{K}\_\mathrm{set} \subset \mathbb{R}^{T\times m}$. Define the compact source-token set*

$$S_0 := \{\pi_0(h_t) : \ h \in \mathcal{K}\_\mathrm{set},\ 0 \le t \le T-1\} \subset U_0.$$

Then for every $\eta > 0$ there exists a width-m concrete Sessa block

$$G_\ell^{\mathrm{lag}} \in \mathrm{ConcreteSessaBlocks}_{\mathrm{Id}}(2, m)$$

such that:

(i) feedback is turned off identically, i.e. $\gamma_t \equiv 0$;

(ii) for every $h \in \mathcal{K}\_\mathrm{set}$ and every t, the block can be chosen so that its input projection depends only on the source slice U_0 (and fixed biases), i.e. it ignores all coordinates in U_r for $r \neq 0$, as well as U_{out}, e_{const}, and e_{pos};

$$\pi_r(G_\ell^{\mathrm{lag}}(h)_t) = \pi_r(h_t) \qquad \text{for all } r \in \{0, \dots, T-1\} \setminus \{\ell\},$$

and the coordinates in U_{out}, e_{const}, and e_{pos} are unchanged;

(iii) if

$$j^*(t,\ell) \in \arg\max_{0\le j\le t} \cos\big((t-\ell)-j\big),$$

then

$$\sup_{h\in\mathcal{K}\_\text{set}} \max_{0\le t\le T-1} \left\|\pi_\ell(G_\ell^{\text{lag}}(h)_t) - \pi_\ell(h_t) - T_{0\to\ell}(\pi_0(h_{j^*(t,\ell)}))\right\|_2 \le \eta.$$

In particular, for $t \ge \ell$ one has $j^*(t,\ell) = t-\ell$.

Proof. Reserve one coordinate of a_t for a constant bias so that the corresponding coordinate of $\bar{a}_t$ is strictly positive. Fix a diagonalization tolerance $\delta \in (0,1)$, to be chosen sufficiently small later. Choose W_{Qf}, W_{Kf} so that only the designated constant coordinate of $\bar{a}_t$ contributes to the forward queries and keys, and set

$$q_t^f \equiv q_{\text{diag}}^{(\ell)} := c_\ell\, \text{RoPE}_{-\ell}(1,0), \qquad k_t^f \equiv k_{\text{diag}} := c_\ell(1,0) \in \mathbb{R}^2,$$

for some scale $c_\ell > 0$. Then for $j \le t$,

$$\langle \text{RoPE}_t(q_t^f),\, \text{RoPE}_j(k_j^f)\rangle = c_\ell^2 \cos\big((t-\ell)-j\big).$$

For each t, the maximizer of $j \mapsto \cos((t-\ell)-j)$ on $\{0,\dots,t\}$ is unique; denote it by $j^*(t,\ell)$. Uniqueness is proved as in Lemma I.5: for $t \ge \ell$, the maximizer is $j = t-\ell$, while for $t < \ell$ the arguments are distinct negative integers and therefore yield distinct cosine values. Hence, by the proof of Lemma I.5 together with Corollary I.3, after choosing c_ℓ large enough we obtain

$$\alpha_{t,j^*(t,\ell)} \ge 1-\delta \qquad \forall t = 0,\dots,T-1.$$

Use d_{ext} further coordinates of a_t to encode the source slice via

$$a_t^{\text{src}} = \varepsilon\, \pi_0(h_t) \in U_0.$$

By Lemma I.15, after choosing $\varepsilon > 0$ small enough, these coordinates of

$$\bar{a}_t = \text{GELU}(a_t)$$

can be linearly mapped by W_V to approximate $T_{0\to\ell}(\pi_0(h_t))$ uniformly on the compact source-token set S_0. Choose W_V so that the resulting value vector lives only in the destination slice U_ℓ. Set $g \equiv \mathbf{1}$, set W^{out} to be the identity on U_ℓ and zero on all other coordinates, and set $b^{\text{out}} = 0$. Choose the feedback branch identically zero.

Define the compact token set

$$S_{\mathcal{K}\_\text{set}} := \{h_t :\ h \in \mathcal{K}\_\text{set},\ 0 \le t \le T-1\} \subset \mathbb{R}^m,$$

and let

$$\phi(z) := T_{0\to\ell}(\pi_0(z)), \qquad z \in S_{\mathcal{K}\_\text{set}}.$$

Set

$$M_\ell := \sup_{z\in S_{\mathcal{K}\_\text{set}}} \|\phi(z)\|_2 < \infty.$$

Choose the small-signal approximation so that the induced value map $v : S_{\mathcal{K}\_\text{set}} \to U_\ell$ satisfies

$$\sup_{z\in S_{\mathcal{K}\_\text{set}}} \|v(z) - \phi(z)\|_2 \le \eta_{\text{val}}.$$

Then for every $h \in \mathcal{K}\_\text{set}$ and every t, Lemma I.4 applied to

$$f_t(h) = \sum_{j \le t} \alpha_{t,j}\, v(h_j)$$

with distinguished index $j^*(t,\ell)$ gives

$$\left\| f_t(h) - v(h_{j^*(t,\ell)}) \right\|_2 \le 2\delta\,(M_\ell + \eta_{\text{val}}).$$

Therefore

$$\left\| f_t(h) - \phi(h_{j^*(t,\ell)}) \right\|_2 \le 2\delta\,(M_\ell + \eta_{\text{val}}) + \eta_{\text{val}}.$$

Since

$$\phi(h_{j^*(t,\ell)}) = T_{0\to\ell}(\pi_0(h_{j^*(t,\ell)})),$$

choosing δ and η_{val} sufficiently small makes the total error at most η. All remaining coordinates are unchanged by construction. □

Lemma I.17 (A diagonal Sessa block realizes a block of tokenwise GELU units)**.** *Let*

$$A : \pi_{\text{st}}(\mathbb{R}^m) \to \mathbb{R}^q$$

be affine, with

$$q \in \{1, \dots, m-1\},$$

and let

$$B : \mathbb{R}^q \to U_{\text{out}}$$

be linear. Fix a compact set

$$S \subset \pi_{\text{st}}(\mathbb{R}^m).$$

Then for every $\eta > 0$ there exists a width-m concrete Sessa block

$$G^{\text{batch}} \in \text{ConcreteSessaBlocks}_{\text{Id}}(2, m)$$

such that:

(i) feedback is turned off identically;
(ii) the storage coordinates are preserved exactly:

$$\pi_{\text{st}}(G^{\text{batch}}(h)_t) = \pi_{\text{st}}(h_t) \qquad \forall h, \forall t;$$

(iii) the input projection ignores the current output slice, i.e. it depends only on $\pi_{\text{st}}(h_t)$;
(iv) for every sequence h whose tokenwise storage states lie in S,

$$\sup_t \left\| \pi_{\text{out}}(G^{\text{batch}}(h)_t) - \pi_{\text{out}}(h_t) - B(\text{GELU}(A(\pi_{\text{st}}(h_t)))) \right\|_2 \le \eta.$$

Proof. Let the first q coordinates of a_t encode the affine preactivations

$$A(\pi_{\text{st}}(h_t)).$$

Reserve one additional coordinate of a_t for a constant bias so that the corresponding coordinate of $\bar{a}_t$ is strictly positive. Choose W_{Qf}, W_{Kf} so that only that coordinate contributes to the forward queries and keys, yielding constant queries and keys that make the forward attention arbitrarily close to diagonal uniformly in t by Lemma I.5.

Choose W_V so that the resulting value vector equals

$$B(\bar{a}_{1:q}) \in U_{\text{out}}$$

in the output slice and is zero on the storage slice. Choose $g \equiv \mathbf{1}$, choose W^{out} to be the identity on U_{out} and zero on the storage slice, set $b^{\text{out}} = 0$, and set the columns of the input projection corresponding to the current output slice U_{out} to zero. Choose the feedback branch identically zero.

Let

$$\phi(z) := B(\text{GELU}(A(z))), \qquad z \in S,$$

and set

$$M_\phi := \sup_{z \in S} \|\phi(z)\|_2 < \infty.$$

Because the input projection ignores the current output slice, the preactivations a_t depend only on $\pi_{\text{st}}(h_t)$, hence for every sequence h whose tokenwise storage states lie in S, the resulting value vector is exactly

$$v_t = \phi(\pi_{\text{st}}(h_t)) \in U_{\text{out}}.$$

By the diagonal forward-attention construction, after choosing the diagonalization tolerance $\delta \in (0, 1)$ sufficiently small we have

$$\alpha_{t,t} \geq 1 - \delta \qquad \forall t = 0, \dots, T-1.$$

Therefore, for every such sequence h and every t, Lemma I.4 applied to

$$f_t = \sum_{j \leq t} \alpha_{t,j} v_j$$

with distinguished index $j^* = t$ gives

$$\|f_t - v_t\|_2 \leq 2\delta M_\phi.$$

Choosing δ so that $2\delta M_\phi \leq \eta$ (trivial if $M_\phi = 0$) yields

$$\sup_t \|f_t - \phi(\pi_{\text{st}}(h_t))\|_2 \leq \eta.$$

Since the residual update is added only in U_{out}, this gives the desired conclusion. □

Corollary I.18 (Tokenwise GELU approximation by stacked Sessa blocks)**.** *Let*

$$S \subset \pi_{\text{st}}(\mathbb{R}^m)$$

be compact and let

$$\Theta : S \to U_{\text{out}}$$

be continuous. Then for every $\eta > 0$ there exists a finite composition

$$G^{\text{tok}} = G_M^{\text{batch}} \circ \cdots \circ G_1^{\text{batch}}, \qquad G_b^{\text{batch}} \in \text{ConcreteSessaBlocks}_{\text{Id}}(2, m),$$

such that:

(i) every G_b^{batch} preserves the storage slice exactly and ignores the current output slice in its input projection;
(ii) for every sequence h whose tokenwise storage states lie in S,

$$\pi_{\text{st}}(G^{\text{tok}}(h)_t) = \pi_{\text{st}}(h_t) \qquad \forall t,$$

and

$$\sup_t \left\|\pi_{\text{out}}(G^{\text{tok}}(h)_t) - \pi_{\text{out}}(h_t) - \Theta(\pi_{\text{st}}(h_t))\right\|_2 \le \eta.$$

Proof. By Lemma I.10, for every $\eta' > 0$ there exist a width $R \in \mathbb{N}^*$, an affine map

$$A_{\text{tot}} : \pi_{\text{st}}(\mathbb{R}^m) \to \mathbb{R}^R,$$

and an affine map

$$B_{\text{tot}} : \mathbb{R}^R \to U_{\text{out}}$$

such that

$$\sup_{z \in S} \|B_{\text{tot}}(\text{GELU}(A_{\text{tot}}(z))) - \Theta(z)\|_2 \le \eta'.$$

Write

$$B_{\text{tot}}(u) = L_{\text{tot}} u + b_{\text{tot}},$$

where

$$L_{\text{tot}} : \mathbb{R}^R \to U_{\text{out}}$$

is linear and

$$b_{\text{tot}} \in U_{\text{out}}.$$

Partition the R hidden units into batches of size at most $m - 1$:

$$R = q_1 + \cdots + q_M, \qquad 1 \le q_b \le m - 1.$$

Write accordingly

$$A_{\text{tot}} = (A_1, \dots, A_M),$$

with each

$$A_b : \pi_{\text{st}}(\mathbb{R}^m) \to \mathbb{R}^{q_b}$$

affine, and decompose the linear map L_{tot} as

$$L_{\text{tot}}(u^{(1)}, \dots, u^{(M)}) = \sum_{b=1}^{M} L_b u^{(b)},$$

where each

$$L_b : \mathbb{R}^{q_b} \to U_{\text{out}}$$

is linear.

Choose $\eta' > 0$ so that

$$\eta' \le \eta/2$$

and

$$\sup_{z \in S} \|B_{\text{tot}}(\text{GELU}(A_{\text{tot}}(z))) - \Theta(z)\|_2 \le \eta'.$$

Apply Lemma I.17 to each pair (A_b, L_b) with accuracy

$$\frac{\eta}{2(M+1)}.$$

This yields concrete Sessa batch blocks

$$G_b^{\text{batch}} \in \text{ConcreteSessaBlocks}_{\text{Id}}(2, m), \qquad b = 1, \dots, M,$$

such that each block preserves the storage slice exactly, ignores the current output slice in its input projection, and contributes

$$L_b(\mathrm{GELU}(A_b(\cdot)))$$

to the output slice up to error at most $\eta/(2(M+1))$.

It remains to represent the constant term b_{tot}. Choose the scalar constant hidden map

$$A_{\mathrm{const}} : \pi_{\mathrm{st}}(\mathbb{R}^m) \to \mathbb{R}, \qquad A_{\mathrm{const}}(z) \equiv 1,$$

and the linear map

$$L_{\mathrm{const}} : \mathbb{R} \to U_{\mathrm{out}}, \qquad L_{\mathrm{const}}(\xi) := \frac{\xi}{\mathrm{GELU}(1)}\, b_{\mathrm{tot}}.$$

Then

$$L_{\mathrm{const}}(\mathrm{GELU}(A_{\mathrm{const}}(z))) = b_{\mathrm{tot}} \qquad \forall z \in S.$$

Apply Lemma I.17 once more to $(A_{\mathrm{const}}, L_{\mathrm{const}})$, again with accuracy

$$\frac{\eta}{2(M+1)}.$$

Since each batch block preserves storage exactly and ignores the current output slice in its input projection, all blocks act on the same storage input and their contributions add in U_{out}. Hence the cumulative implementation error of the M linear batches together with the one constant batch is at most

$$(M+1) \cdot \frac{\eta}{2(M+1)} = \frac{\eta}{2}.$$

Combining this with the approximation error $\eta' \le \eta/2$ gives the total error bound η. □

I.10 Sessa universality for causal maps

Theorem (Universal approximation for Sessa with adapters). Let $\mathcal{D} \subset \mathbb{R}^{T \times d_{\mathrm{ext}}}$ be compact and let

$$F : \mathcal{D} \to \mathbb{R}^{T \times d_{\mathrm{ext}}}$$

be continuous and causal. Then for any $\varepsilon > 0$ there exist a model width $m \in \mathbb{N}^*$, an even key/query width d_k (in fact $d_k = 2$ suffices), tokenwise adapters

$$\mathrm{Embed} : \mathbb{R}^{d_{\mathrm{ext}}} \to \mathbb{R}^m, \qquad \mathrm{Unembed} : \mathbb{R}^m \to \mathbb{R}^{d_{\mathrm{ext}}},$$

and a finite-depth network

$$G \in \Omega^{d_k}_{\mathrm{Sessa,Id}}(m)$$

consisting only of the concrete Sessa blocks from Section 3, such that

$$\sup_{x \in \mathcal{D}} \Big\| F(x) - \mathrm{Unembed}(G(\mathrm{Embed}(x))) \Big\|_F < \varepsilon.$$

Proof of Theorem 14. Fix $\varepsilon > 0$.

Step 0: causal factorization. For each $t \in \{0, \dots, T-1\}$, define the compact set of attainable prefixes

$$\mathcal{P}^{\mathrm{pref}}_t := \{(x_0, \dots, x_t) : x \in \mathcal{D}\} \subset (\mathbb{R}^{d_{\mathrm{ext}}})^{t+1}.$$

By Lemma I.1, there exists a unique continuous map

$$\widehat{F}_t : \mathcal{P}_t^{\mathrm{pref}} \to \mathbb{R}^{d_{\mathrm{ext}}}, \qquad \widehat{F}_t(x_0, \dots, x_t) := F(x)_t \quad (x \in \mathcal{D}).$$

Since $\mathcal{P}_t^{\mathrm{pref}}$ is compact in Euclidean space, it is closed in $(\mathbb{R}^{d_{\mathrm{ext}}})^{t+1}$. By Tietze extension applied coordinatewise, extend $\widehat{F}_t$ to a continuous map

$$F_t : (\mathbb{R}^{d_{\mathrm{ext}}})^{t+1} \to \mathbb{R}^{d_{\mathrm{ext}}}$$

such that

$$F(x)_t = F_t(x_0, \dots, x_t) \qquad \forall x \in \mathcal{D}.$$

Step 1: width and adapters Set

$$m := (T+1)d_{\mathrm{ext}} + 2.$$

Use the storage decomposition introduced above.

Define the tokenwise embedding by

$$\mathrm{Embed}(x)_t = J_0(x_t) + J_{\mathrm{out}}(x_t) + e_{\mathrm{const}},$$

that is, place x_t in both U_0 and U_{out}, set the constant coordinate to 1, and set all other coordinates to 0.

Define Unembed tokenwise by

$$\mathrm{Unembed}(h)_t := R_{\mathrm{out}}(\pi_{\mathrm{out}}(h_t)) \in \mathbb{R}^{d_{\mathrm{ext}}}.$$

Then

$$\mathrm{Unembed}(\mathrm{Embed}(x)) = x \qquad \forall x \in \mathbb{R}^{T \times d_{\mathrm{ext}}},$$

$\mathrm{Embed}(\mathcal{D})$ is compact, and Unembed is linear and non-expansive in Frobenius norm.

Step 2: positional code Apply Corollary I.8 with $u = e_{\mathrm{pos}}$ to obtain a block

$$G^{\mathrm{pos}} \in \mathrm{ConcreteSessaBlocks}_{\mathrm{Id}}(2, m)$$

and pairwise distinct scalars $c_0, \dots, c_{T-1}$ such that

$$G^{\mathrm{pos}}(h)_t = h_t + c_t e_{\mathrm{pos}} \qquad \forall h, \forall t.$$

By construction, G^{pos} leaves $U_0, \dots, U_{T-1}$ and U_{out} unchanged.

Step 3: prefix encoding Fix a packing tolerance

$$\delta_{\mathrm{pack}} > 0,$$

to be specified later in Step 4. For each lag $\ell = 1, \dots, T-1$, apply Lemma I.16 successively on the compact set obtained after the previous blocks to construct a concrete Sessa block

$$G_\ell^{\mathrm{lag}} \in \mathrm{ConcreteSessaBlocks}_{\mathrm{Id}}(2, m)$$

that preserves all coordinates except U_ℓ and writes an approximation of the lag-ℓ token from U_0 into U_ℓ.

For $t \in \{0, \dots, T-1\}$ and $\ell \in \{1, \dots, T-1\}$, define

$$j^*(t, \ell) \in \arg\max_{0 \le j \le t} \cos\big((t-\ell) - j\big).$$

For $t \ge \ell$ one has $j^*(t, \ell) = t - \ell$.

Define the ideal encoded state $\widehat{h}_t(x) \in \mathbb{R}^m$ by:

$$\pi_0(\widehat{h}_t(x)) = J_0(x_t), \qquad \pi_\ell(\widehat{h}_t(x)) = J_\ell(x_{j^*(t,\ell)}) \quad (1 \le \ell \le T-1),$$

$$\pi_{\text{out}}(\widehat{h}_t(x)) = J_{\text{out}}(x_t), \qquad \langle \widehat{h}_t(x), e_{\text{const}} \rangle = 1, \qquad \langle \widehat{h}_t(x), e_{\text{pos}} \rangle = c_t.$$

Since each lag block depends only on the exact source slice U_0 and fixed biases, while writing only to its own destination slice and preserving all previously written slices, the packing errors do not propagate to later lag blocks. Hence, choosing per-lag accuracies $\eta_\ell > 0$ with

$$\sum_{\ell=1}^{T-1} \eta_\ell^2 \le \delta_{\text{pack}}^2,$$

we obtain for

$$G^{\text{pack}} := G^{\text{lag}}_{T-1} \circ \cdots \circ G^{\text{lag}}_1 \circ G^{\text{pos}}$$

that

$$\sup_{x \in \mathcal{D}} \max_{0 \le t \le T-1} \left\| G^{\text{pack}}(\text{Embed}(x))_t - \widehat{h}_t(x) \right\|_2 \le \delta_{\text{pack}}.$$

Step 4: target map For each t, let

$$\widehat{S}_t := \{\widehat{h}_t(x) : x \in \mathcal{D}\} \subset \mathbb{R}^m, \qquad \widehat{S} := \bigcup_{t=0}^{T-1} \widehat{S}_t.$$

Each $\widehat{S}_t$ is compact. Since the e_{pos}-coordinate equals c_t on $\widehat{S}_t$ and the scalars c_t are distinct, the sets $\widehat{S}_t$ are pairwise disjoint and positively separated.

Define the linear readout

$$\text{Read}_t : \mathbb{R}^m \to (\mathbb{R}^{d_{\text{ext}}})^{t+1}$$

by

$$\text{Read}_t(u) := (R_t \pi_t(u),\, R_{t-1}\pi_{t-1}(u),\, \ldots,\, R_0 \pi_0(u)).$$

For $u = \widehat{h}_t(x)$, one has

$$R_0 \pi_0(\widehat{h}_t(x)) = x_t,$$

and for $1 \le \ell \le t$,

$$R_\ell \pi_\ell(\widehat{h}_t(x)) = x_{j^*(t,\ell)}.$$

Since $j^*(t,\ell) = t - \ell$ for $1 \le \ell \le t$, it follows that

$$\text{Read}_t(\widehat{h}_t(x)) = (x_0, \ldots, x_t).$$

Define

$$\widehat{\Phi} : \widehat{S} \to U_{\text{out}}$$

by

$$\widehat{\Phi}(u) := J_{\text{out}}(F_t(\text{Read}_t(u))) \qquad \text{for } u \in \widehat{S}_t.$$

This is well defined because the index t is uniquely determined by the e_{pos}-coordinate of u, and if

$$u = \widehat{h}_t(x) = \widehat{h}_t(x'),$$

then

$$\text{Read}_t(u) = (x_0, \ldots, x_t) = (x'_0, \ldots, x'_t),$$

so the value of $J_{\text{out}}(F_t(\text{Read}_t(u)))$ does not depend on the choice of x.

Moreover, on each $\widehat{S}_t$ one has

$$\widehat{\Phi}|_{\widehat{S}_t} = J_{\text{out}} \circ F_t \circ \text{Read}_t|_{\widehat{S}_t},$$

hence $\widehat{\Phi}$ is continuous on each $\widehat{S}_t$, and therefore continuous on $\widehat{S}$.

Apply Tietze extension coordinatewise to the $\mathbb{R}^{d_{\text{ext}}}$-valued map

$$R_{\text{out}} \circ \widehat{\Phi} : \widehat{S} \to \mathbb{R}^{d_{\text{ext}}}.$$

This yields a continuous extension

$$\bar{\Phi} : \mathbb{R}^m \to \mathbb{R}^{d_{\text{ext}}}$$

of $R_{\text{out}} \circ \widehat{\Phi}$. Set

$$\widetilde{\Phi} := J_{\text{out}} \circ \bar{\Phi} : \mathbb{R}^m \to U_{\text{out}}.$$

Then $\widetilde{\Phi}$ extends $\widehat{\Phi}$.

Fix $\rho > 0$ and let

$$N := \overline{\mathcal{N}}_\rho(\widehat{S}) \subset \mathbb{R}^m.$$

Then N is compact, so $\widetilde{\Phi}$ is uniformly continuous on N. Choose $\delta_{\text{UC}} > 0$ such that

$$u, v \in N,\ \|u - v\|_2 \le \delta_{\text{UC}} \quad \Longrightarrow \quad \|\widetilde{\Phi}(u) - \widetilde{\Phi}(v)\|_2 \le \frac{\varepsilon}{2\sqrt{T}}.$$

Choose $\delta_{\text{pack}} > 0$ small enough that

$$\delta_{\text{pack}} \le \min\{\rho, \delta_{\text{UC}}\}$$

and that the encoding construction of Step 3 yields

$$\sup_{x \in \mathcal{D}} \max_{0 \le t \le T-1} \left\| G^{\text{pack}}(\text{Embed}(x))_t - \hat{h}_t(x) \right\|_2 \le \delta_{\text{pack}}.$$

Then for every $x \in \mathcal{D}$ and every t one has

$$G^{\text{pack}}(\text{Embed}(x))_t \in N,$$

and

$$\left\| \widetilde{\Phi}(G^{\text{pack}}(\text{Embed}(x))_t) - J_{\text{out}}(F(x)_t) \right\|_2 \le \frac{\varepsilon}{2\sqrt{T}}.$$

Step 5: tokenwise readout Define the compact storage-token set

$$S_{\text{st}} := \left\{ \pi_{\text{st}}(G^{\text{pack}}(\text{Embed}(x))_t) : \ x \in \mathcal{D},\ 0 \le t \le T-1 \right\}.$$

Define

$$\Theta : S_{\text{st}} \to U_{\text{out}}, \qquad \Theta(z) := \widetilde{\Phi}(\iota_{\text{st}}(z)) - T_{0 \to \text{out}}(\pi_0(z)).$$

Since ι_{st} is linear and $\widetilde{\Phi}$ is continuous, Θ is continuous. Moreover, for every $x \in \mathcal{D}$ and every t,

$$\iota_{\text{st}}(\pi_{\text{st}}(G^{\text{pack}}(\text{Embed}(x))_t)) = G^{\text{pack}}(\text{Embed}(x))_t,$$

since Embed initializes the output slice as a copy of U_0 and G^{pack} preserves U_{out}. Hence

$$\Theta(\pi_{\text{st}}(G^{\text{pack}}(\text{Embed}(x))_t)) = \widetilde{\Phi}(G^{\text{pack}}(\text{Embed}(x))_t) - \pi_{\text{out}}(G^{\text{pack}}(\text{Embed}(x))_t),$$

so Θ is exactly the tokenwise increment that must be added in U_{out}. Apply Corollary I.18 to S_{st} and Θ. This yields a finite composition

$$G^{\text{tok}} = G_M^{\text{batch}} \circ \cdots \circ G_1^{\text{batch}}$$

of concrete Sessa blocks such that every batch block preserves the storage coordinates exactly, every batch block ignores the current output slice in its input projection, and for all $x \in \mathcal{D}$ and all t,

$$\left\| \pi_{\text{out}}(G^{\text{tok}}(G^{\text{pack}}(\text{Embed}(x)))_t) - \widetilde{\Phi}(G^{\text{pack}}(\text{Embed}(x))_t) \right\|_2 \le \frac{\varepsilon}{2\sqrt{T}}.$$

Step 6: conclusion Set

$$G := G^{\text{tok}} \circ G^{\text{pack}} \in \Omega^2_{\text{Sessa,Id}}(m).$$

Since

$$\text{Unembed}(h)_t = R_{\text{out}}(\pi_{\text{out}}(h_t)),$$

combining the two error bounds and using that R_{out} is an isometry gives

$$\|\text{Unembed}(G(\text{Embed}(x)))_t - F(x)_t\|_2 = \|R_{\text{out}}(\pi_{\text{out}}(G(\text{Embed}(x))_t)) - F(x)_t\|_2 \le \frac{\varepsilon}{\sqrt{T}} \qquad \forall x \in \mathcal{D},\ \forall t.$$

Hence

$$\sup_{x \in \mathcal{D}} \left\| \text{Unembed}(G(\text{Embed}(x))) - F(x) \right\|_F < \varepsilon.$$

□

J Universal approximation in the pre-norm LayerNorm setting

We now extend Theorem 14 from Norm = Id to the pre-norm LayerNorm case Norm = $\text{LN}_{\varepsilon_{\text{ln}}}$ with $\varepsilon_{\text{ln}} > 0$ (Xiong et al., 2020), after a width expansion via a fixed scaffold.

J.1 Tokenwise LayerNorm

Fix a width $m \ge 2$ and $\varepsilon_{\text{ln}} > 0$. For $z \in \mathbb{R}^m$, define

$$\mu_{\text{ln}}(z) := \frac{1}{m}\langle z, \mathbf{1}\rangle, \qquad \sigma_{\text{ln}}(z) := \sqrt{\frac{1}{m}\|z - \mu_{\text{ln}}(z)\mathbf{1}\|_2^2 + \varepsilon_{\text{ln}}}, \qquad \text{LN}_{\varepsilon_{\text{ln}}}(z) := \frac{z - \mu_{\text{ln}}(z)\mathbf{1}}{\sigma_{\text{ln}}(z)}.$$

With $\varepsilon_{\text{ln}} > 0$, $\text{LN}_{\varepsilon_{\text{ln}}}$ is well-defined and continuous on all of $\mathbb{R}^m$, in particular, there is no singularity at nearly-constant tokens.

J.2 Zero-mean scaffold embedding

Fix a "dynamic" width $m_0 \ge 1$ and let $m_{\text{sc}} \ge 2$ be an even scaffold width. Let $m := m_0 + m_{\text{sc}}$ and define, for $c > 0$, the fixed zero-mean scaffold vector

$$s_{c,m_{\text{sc}}} := (\underbrace{c, \dots, c}_{m_{\text{sc}}/2}, \underbrace{-c, \dots, -c}_{m_{\text{sc}}/2}) \in \mathbb{R}^{m_{\text{sc}}}, \qquad \langle s_{c,m_{\text{sc}}}, \mathbf{1}_{m_{\text{sc}}}\rangle = 0.$$

Define the scaffold embedding

$$\Phi_{c,m_{\text{sc}}} : \mathbb{R}^{m_0} \to \mathbb{R}^m, \qquad \Phi_{c,m_{\text{sc}}}(u) := (u,\ s_{c,m_{\text{sc}}}).$$

Let $\pi_{\text{dyn}} : \mathbb{R}^m \to \mathbb{R}^{m_0}$ be the projection onto the first m_0 coordinates, and let $\pi_{\text{sc}} : \mathbb{R}^m \to \mathbb{R}^{m_{\text{sc}}}$ be the projection onto the last m_{sc} coordinates:

$$\pi_{\text{dyn}}(z_1, \dots, z_{m_0+m_{\text{sc}}}) = (z_1, \dots, z_{m_0}), \qquad \pi_{\text{sc}}(z_1, \dots, z_{m_0+m_{\text{sc}}}) = (z_{m_0+1}, \dots, z_{m_0+m_{\text{sc}}}).$$

Lemma J.1 (Approximate linearity of LayerNorm on scaffold sets)**.** *Fix $m_0 \geq 1$, $\varepsilon_{\text{ln}} > 0$, a compact set $\mathcal{K}\_\text{set} \subset \mathbb{R}^{m_0}$, and $\delta > 0$. Then there exist an even $m_{\text{sc}} \geq 2$, a scalar $c > 0$, and a constant $a > 0$ such that*

$$\sup_{u \in \mathcal{K}\_\text{set}} \left\| \pi_{\text{dyn}}(\, \text{LN}_{\varepsilon_{\text{ln}}}(\Phi_{c,m_{\text{sc}}}(u))) - a\, u \right\|_2 \leq \delta.$$

Moreover, $\pi_{\text{sc}}(\Phi_{c,m_{\text{sc}}}(u)) \equiv s_{c,m_{\text{sc}}}$ is constant on $\mathcal{K}_$set.

Proof. Let $R := \sup_{u \in \mathcal{K}\_\text{set}} \|u\|_2 < \infty$ and fix an even $m_{\text{sc}} \geq 2$. Set $m := m_0 + m_{\text{sc}}$. For $u \in \mathcal{K}_$set write

$$z := \Phi_{c,m_{\text{sc}}}(u) = (u, s_{c,m_{\text{sc}}}) \in \mathbb{R}^m.$$

Since $\langle s_{c,m_{\text{sc}}}, \mathbf{1}_{m_{\text{sc}}} \rangle = 0$, we have

$$\mu_{\text{ln}}(z) = \frac{1}{m} \sum_{i=1}^{m_0} u_i =: \mu_u, \qquad |\mu_u| \leq \frac{1}{m} \Big| \sum_{i=1}^{m_0} u_i \Big| \leq \frac{\sqrt{m_0}}{m} \|u\|_2 \leq \frac{\sqrt{m_0}}{m} R.$$

Define the mean-centered dynamic vector $\bar{u} := u - \mu_u \mathbf{1}_{m_0}$. Then the dynamic slice of LayerNorm equals

$$\pi_{\text{dyn}}(\text{LN}_{\varepsilon_{\text{ln}}}(z)) = \frac{\bar{u}}{\sigma_{\text{ln}}(z)}.$$

Define the reference scale

$$\sigma_0 := \sigma_{\text{ln}}(\Phi_{c,m_{\text{sc}}}(0)) = \sqrt{\frac{1}{m}\|s_{c,m_{\text{sc}}}\|_2^2 + \varepsilon_{\text{ln}}} = \sqrt{\frac{1}{m}\,(m_{\text{sc}} c^2) + \varepsilon_{\text{ln}}}, \qquad a := \frac{1}{\sigma_0}.$$

We estimate

$$\left\| \frac{\bar{u}}{\sigma_{\text{ln}}(z)} - au \right\|_2 \leq \left\| \frac{\bar{u} - u}{\sigma_{\text{ln}}(z)} \right\|_2 + \left\| u \left(\frac{1}{\sigma_{\text{ln}}(z)} - \frac{1}{\sigma_0} \right) \right\|_2 =: T_1 + T_2.$$

For the term T_1 (mean leakage), Since $\bar{u} - u = -\mu_u \mathbf{1}_{m_0}$,

$$T_1 = \frac{\|\mu_u \mathbf{1}_{m_0}\|_2}{\sigma_{\text{ln}}(z)} \leq \frac{\sqrt{m_0}|\mu_u|}{\sqrt{\varepsilon_{\text{ln}}}} \leq \frac{\sqrt{m_0}}{\sqrt{\varepsilon_{\text{ln}}}} \cdot \frac{\sqrt{m_0}}{m} R = \frac{m_0 R}{m\sqrt{\varepsilon_{\text{ln}}}}.$$

for the term T_2 (variance perturbation), Note that $\sigma_{\text{ln}}(z)^2 = \frac{1}{m}\|z - \mu_u \mathbf{1}\|_2^2 + \varepsilon_{\text{ln}}$ and, because $\langle s_{c,m_{\text{sc}}}, \mathbf{1}_{m_{\text{sc}}} \rangle = 0$, we have the exact decomposition

$$\|z - \mu_u \mathbf{1}\|_2^2 = \|u - \mu_u \mathbf{1}_{m_0}\|_2^2 + \|s_{c,m_{\text{sc}}} - \mu_u \mathbf{1}_{m_{\text{sc}}}\|_2^2 = \|\bar{u}\|_2^2 + \|s_{c,m_{\text{sc}}}\|_2^2 + m_{\text{sc}}\mu_u^2,$$

and the cross term vanishes since $\langle s_{c,m_{\text{sc}}}, \mathbf{1}_{m_{\text{sc}}} \rangle = 0$. Therefore

$$\sigma_{\text{ln}}(z)^2 - \sigma_0^2 = \frac{1}{m}(\|\bar{u}\|_2^2 + m_{\text{sc}}\mu_u^2) \leq \frac{1}{m}(\|u\|_2^2 + m_{\text{sc}}\mu_u^2) \leq \frac{1}{m}\left(R^2 + m_{\text{sc}} \cdot \frac{m_0 R^2}{m^2} \right) \leq \frac{2R^2}{m},$$

since $m_{\text{sc}} \leq m$ implies $m_{\text{sc}} m_0 / m^2 \leq m_0/m \leq 1$ for $m \geq m_0$.

Using $|\sqrt{A}-\sqrt{B}| \le |A-B|/(\sqrt{A}+\sqrt{B})$ and $\sigma_{\ln}(z), \sigma_0 \ge \sqrt{\varepsilon_{\ln}}$,

$$|\sigma_{\ln}(z) - \sigma_0| \le \frac{|\sigma_{\ln}(z)^2 - \sigma_0^2|}{\sigma_{\ln}(z) + \sigma_0} \le \frac{(2R^2/m)}{2\sqrt{\varepsilon_{\ln}}} = \frac{R^2}{m\sqrt{\varepsilon_{\ln}}}.$$

Hence

$$\left|\frac{1}{\sigma_{\ln}(z)} - \frac{1}{\sigma_0}\right| = \frac{|\sigma_{\ln}(z) - \sigma_0|}{\sigma_{\ln}(z)\sigma_0} \le \frac{R^2}{m\sqrt{\varepsilon_{\ln}}} \cdot \frac{1}{\varepsilon_{\ln}} = \frac{R^2}{m\,\varepsilon_{\ln}^{3/2}}.$$

Therefore

$$T_2 \le \|u\|_2 \left|\frac{1}{\sigma_{\ln}(z)} - \frac{1}{\sigma_0}\right| \le R \cdot \frac{R^2}{m\,\varepsilon_{\ln}^{3/2}} = \frac{R^3}{m\,\varepsilon_{\ln}^{3/2}}.$$

Combining,

$$\sup_{u \in \mathcal{K}\_\text{set}} \left\|\pi_{\text{dyn}}(\text{LN}_{\varepsilon_{\ln}}(\Phi_{c,m_{\text{sc}}}(u))) - au\right\|_2 \le \frac{m_0 R}{m\sqrt{\varepsilon_{\ln}}} + \frac{R^3}{m\,\varepsilon_{\ln}^{3/2}}.$$

Choose m_{sc} (hence $m = m_0 + m_{\text{sc}}$) large enough so that the right-hand side is $\le \delta$. This proves the claim; note that $c > 0$ can be arbitrary and only changes the scaling a. □

J.3 Simulating identity-normalized Sessa blocks with pre-norm LN-Sessa blocks

We call a pre-norm LN-Sessa block a Sessa block with Norm $= \text{LN}_{\varepsilon_{\ln}}$ in the tokenwise preprocessing stage, i.e. $\tilde{x}_t = \text{LN}_{\varepsilon_{\ln}}(x_t)$, and residual $y_t = x_t + o_t$.

Lemma J.2 (Simulation of an identity-normalized block by a pre-norm LN block on a scaffold)**.** *Let $G : \mathbb{R}^{T\times m_0} \to \mathbb{R}^{T\times m_0}$ be a width-m_0 concrete Sessa block from Section 3, with* Norm $=$ Id*. Fix a compact set $\mathcal{K}\_\text{set} \subset \mathbb{R}^{T\times m_0}$ and $\varepsilon_{\text{sim}} > 0$. Then there exist an even $m_{\text{sc}} \ge 2$, a scalar $c > 0$, and a width-m pre-norm LN-Sessa block $\widetilde{G} : \mathbb{R}^{T\times(m_0+m_{\text{sc}})} \to \mathbb{R}^{T\times(m_0+m_{\text{sc}})}$ with* Norm $= \text{LN}_{\varepsilon_{\ln}}$ *such that, with $m := m_0 + m_{\text{sc}}$,*

$$\sup_{x \in \mathcal{K}\_\text{set}} \left\|\pi_{\text{dyn}}(\widetilde{G}(\Phi_{c,m_{\text{sc}}}(x))) - G(x)\right\|_F \le \varepsilon_{\text{sim}}, \qquad \text{and} \qquad \pi_{\text{sc}}(\widetilde{G}(\Phi_{c,m_{\text{sc}}}(x))) \equiv s_{c,m_{\text{sc}}}.$$

Here $\Phi_{c,m_{\text{sc}}}(x)$ denotes the tokenwise application of $\Phi_{c,m_{\text{sc}}}$.

Proof. Define the compact set of attainable tokens

$$S_{\mathcal{K}\_\text{set}} := \{x_t : \ x \in \mathcal{K}\_\text{set},\ t = 0, \dots, T-1\} \subset \mathbb{R}^{m_0}.$$

Choose once and for all

$$a \in (0, \varepsilon_{\ln}^{-1/2}).$$

Define the continuous map

$$\Delta : \mathbb{R}^{T\times m_0} \to \mathbb{R}^{T\times m_0},$$

i.e. given $v \in \mathbb{R}^{T\times m_0}$, run the Sessa block from the stage after normalization, with the dynamic weights scaled by $1/a$, i.e. with first input projection on the dynamic slice $\widetilde{W}_{\text{dyn}}^{\text{in}} := a^{-1}W^{\text{in}}$, $\tilde{b}^{\text{in}} := b^{\text{in}}$, and all other dynamic parameters copied from G. Then, by construction,

$$G(x) = x + \Delta(ax) \qquad \forall x \in \mathbb{R}^{T\times m_0}.$$

Since $\mathcal{K}_$set is compact, so is $a\mathcal{K}_$set, and Δ is uniformly continuous on a compact neighborhood of $a\mathcal{K}_$set. Choose $\eta_{\text{UC}} > 0$ such that

$$\|v - v'\|_F \le \eta_{\text{UC}} \ \Rightarrow\ \|\Delta(v) - \Delta(v')\|_F \le \varepsilon_{\text{sim}} \quad \text{for all } v, v' \text{ in that neighborhood.}$$

$$\eta_{\mathrm{LN}} := \eta_{\mathrm{UC}}/\sqrt{T}.$$

Fix an even $m_{\mathrm{sc}} \geq 2$ (to be chosen large enough), set $m := m_0 + m_{\mathrm{sc}}$, and define

$$c := \sqrt{\frac{m}{m_{\mathrm{sc}}}(a^{-2} - \varepsilon_{\mathrm{ln}})} \; > \; 0.$$

Then the reference scale in Lemma J.1 equals exactly

$$\sigma_0 = \sqrt{\frac{m_{\mathrm{sc}}c^2}{m} + \varepsilon_{\mathrm{ln}}} = a^{-1}, \qquad \text{hence} \qquad \frac{1}{\sigma_0} = a.$$

Inspecting the proof of Lemma J.1, the approximation bound depends on $m = m_0 + m_{\mathrm{sc}}$ (and on $S_{\mathcal{K}\_\mathrm{set}}, \varepsilon_{\mathrm{ln}}$) and tends to 0 as $m \to \infty$; therefore, after increasing the even m_{sc} if needed, we obtain

$$\sup_{u \in S_{\mathcal{K}\_\mathrm{set}}} \left\| \pi_{\mathrm{dyn}}(\,\mathrm{LN}_{\varepsilon_{\mathrm{ln}}}(\Phi_{c,m_{\mathrm{sc}}}(u))) - au \right\|_2 \leq \eta_{\mathrm{LN}}.$$

Write the width-m_0 input projection of G as

$$W^{\mathrm{in}} = [W_a \;\; W_g], \qquad b^{\mathrm{in}} = (b_a, b_g),$$

with

$$W_a, W_g \in \mathbb{R}^{m_0 \times m_0}, \qquad b_a, b_g \in \mathbb{R}^{m_0}.$$

Decompose the widened coordinates as

$$\mathbb{R}^m = \mathbb{R}^{m_0} \oplus \mathbb{R}^{m_{\mathrm{sc}}},$$

where the first summand is the dynamic slice and the second is the scaffold slice.

Define

$$\widetilde{W}_a = \begin{bmatrix} a^{-1}W_a & 0 \\ 0 & 0 \end{bmatrix}, \qquad \widetilde{W}_g = \begin{bmatrix} a^{-1}W_g & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m \times m},$$

and

$$\widetilde{W}^{\mathrm{in}} = [\widetilde{W}_a \;\; \widetilde{W}_g] \in \mathbb{R}^{m \times 2m}, \qquad \tilde{b}^{\mathrm{in}} = (b_a, 0_{m_{\mathrm{sc}}}, b_g, 0_{m_{\mathrm{sc}}}) \in \mathbb{R}^{2m}.$$

For the mixer parameters define

$$\widetilde{W}_{Qf} = \begin{bmatrix} W_{Qf} \\ 0 \end{bmatrix}, \qquad \widetilde{W}_{Kf} = \begin{bmatrix} W_{Kf} \\ 0 \end{bmatrix}, \qquad \widetilde{W}_{Qb} = \begin{bmatrix} W_{Qb} \\ 0 \end{bmatrix}, \qquad \widetilde{W}_{Kb} = \begin{bmatrix} W_{Kb} \\ 0 \end{bmatrix} \in \mathbb{R}^{m \times d_k},$$

$$\widetilde{W}_V = \begin{bmatrix} W_V & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m \times m}, \qquad \widetilde{w}^{\gamma} = (w^{\gamma}, 0_{m_{\mathrm{sc}}}) \in \mathbb{R}^m, \qquad \tilde{b}^{\gamma} := b^{\gamma}.$$

For the output map define

$$\widetilde{W}^{\mathrm{out}} = \begin{bmatrix} W^{\mathrm{out}} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m \times m}, \qquad \tilde{b}^{\mathrm{out}} = (b^{\mathrm{out}}, 0_{m_{\mathrm{sc}}}) \in \mathbb{R}^m.$$

All remaining scaffold rows and columns are set to zero.

Thus, once the pre-norm token

$$z_t := \mathrm{LN}_{\varepsilon_{\mathrm{ln}}}(X_t)$$

is formed, every learned linear map in $\widetilde{G}$ reads only $\pi_{\mathrm{dyn}}(z_t)$, while the residual increment has zero scaffold coordinates.

For $X = \Phi_{c,m_{\mathrm{sc}}}(x)$, define

$$v_t := \pi_{\mathrm{dyn}}(\,\mathrm{LN}_{\varepsilon_{\mathrm{ln}}}(X_t)) \in \mathbb{R}^{m_0}.$$

Then the widened block has

$$\tilde{a}_t = (a^{-1} v_t W_a + b_a,\ 0_{m_{\mathrm{sc}}}), \qquad \tilde{g}_t = (a^{-1} v_t W_g + b_g,\ 0_{m_{\mathrm{sc}}}),$$

hence

$$\mathrm{GELU}(\tilde{a}_t) = (\mathrm{GELU}(a^{-1} v_t W_a + b_a),\, 0_{m_{\mathrm{sc}}}).$$

Therefore the forward logits, feedback logits, gains, dynamic mixer output, and dynamic residual increment of $\widetilde{G}$ coincide exactly with those of the width-m_0 block defining $\Delta(v)$, whereas the scaffold part of f, s, and of the residual increment is identically zero. Consequently

$$\pi_{\mathrm{dyn}}(\widetilde{G}(\Phi_{c,m_{\mathrm{sc}}}(x))) = x + \Delta(v), \qquad \pi_{\mathrm{sc}}(\widetilde{G}(\Phi_{c,m_{\mathrm{sc}}}(x))) = s_{c,m_{\mathrm{sc}}}.$$

For $x \in \mathcal{K}$_set, the tokenwise bound above implies

$$\left\|\pi_{\mathrm{dyn}}(\mathrm{LN}_{\varepsilon_{\mathrm{ln}}}(\Phi_{c,m_{\mathrm{sc}}}(x))) - ax\right\|_F \le \eta_{\mathrm{LN}}\sqrt{T} = \eta_{\mathrm{UC}},$$

hence

$$\left\|\pi_{\mathrm{dyn}}(\widetilde{G}(\Phi_{c,m_{\mathrm{sc}}}(x))) - G(x)\right\|_F = \left\|\Delta(\pi_{\mathrm{dyn}}(\mathrm{LN}_{\varepsilon_{\mathrm{ln}}}(\Phi_{c,m_{\mathrm{sc}}}(x)))) - \Delta(ax)\right\|_F \le \varepsilon_{\mathrm{sim}}.$$

Finally, since the increment has zero scaffold coordinates, the scaffold stays constant: $\pi_{\mathrm{sc}}(\widetilde{G}(\Phi_{c,m_{\mathrm{sc}}}(x))) \equiv s_{c,m_{\mathrm{sc}}}$. □

J.4 Universal approximation for pre-norm LN-Sessa

Corollary J.3 (Universal approximation for pre-norm LN-Sessa)**.** *Let $\mathcal{D} \subset \mathbb{R}^{T\times d_{\mathrm{ext}}}$ be compact and let*

$$F : \mathcal{D} \to \mathbb{R}^{T\times d_{\mathrm{ext}}}$$

be continuous and causal. Fix $\varepsilon_{\mathrm{ln}} > 0$ for tokenwise LayerNorm. Then for any $\varepsilon > 0$ there exist a model width $m \in \mathbb{N}^*$, an even key/query width d_k, tokenwise adapters

$$\mathrm{Embed} : \mathbb{R}^{d_{\mathrm{ext}}} \to \mathbb{R}^m, \qquad \mathrm{Unembed} : \mathbb{R}^m \to \mathbb{R}^{d_{\mathrm{ext}}},$$

and a finite-depth pre-norm LN-Sessa network

$$G_{\mathrm{ln}} \in \Omega^{d_k}_{\mathrm{Sessa},\mathrm{LN}_{\varepsilon_{\mathrm{ln}}}}(m),$$

such that

$$\sup_{x\in\mathcal{D}} \left\|F(x) - \mathrm{Unembed}(G_{\mathrm{ln}}(\mathrm{Embed}(x)))\right\|_F < \varepsilon.$$

Proof. By Theorem 14 for Norm = Id, choose adapters

$$\mathrm{Embed}_0 : \mathbb{R}^{d_{\mathrm{ext}}} \to \mathbb{R}^{m_0}, \qquad \mathrm{Unembed}_0 : \mathbb{R}^{m_0} \to \mathbb{R}^{d_{\mathrm{ext}}},$$

and a concrete Sessa network with Norm = Id

$$G^\star \in \Omega^{d_{k,0}}_{\mathrm{Sessa},\mathrm{Id}}(m_0)$$

of depth N_{layer} such that

$$\sup_{x\in\mathcal{D}}\left\|F(x)-\text{Unembed}_0(G^\star(\text{Embed}_0(x)))\right\|_F<\varepsilon/2.$$

Write

$$G^\star=G_{N_{\text{layer}}}\circ\cdots\circ G_1$$

as a composition of concrete Sessa blocks with Norm = Id on $\mathbb{R}^{T\times m_0}$.

Let $\mathcal{K}\_\text{set}_1:=\text{Embed}_0(\mathcal{D})$ (compact). Fix $\rho_{\text{nbhd}}>0$ and define the thickened compacts recursively as in Lemma I.9:

$$\widetilde{\mathcal{K}\_\text{set}}_1:=\mathcal{K}\_\text{set}_1,\qquad \mathcal{K}\_\text{set}_{n_{\text{layer}}+1}:=G_{n_{\text{layer}}}(\widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}}),\qquad \widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}+1}:=\overline{\mathcal{N}}_{\rho_{\text{nbhd}}}(\mathcal{K}\_\text{set}_{n_{\text{layer}}+1})\quad\text{for } n_{\text{layer}}=1,\dots,N_{\text{la}}$$

Since N_{layer} is finite, the union of attainable token sets

$$S:=\bigcup_{n_{\text{layer}}=1}^{N_{\text{layer}}}\{u_t:\ u\in\widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}},\ t=0,\dots,T-1\}\subset\mathbb{R}^{m_0}.$$

is a finite union of compact sets and hence compact.

By Lemma I.9, choose tolerances $\varepsilon^{\text{sim}}_{n_{\text{layer}}}>0$ such that if each block $G_{n_{\text{layer}}}$ is approximated on $\widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}}$ within $\varepsilon^{\text{sim}}_{n_{\text{layer}}}$, then the composed approximation error on $\mathcal{K}\_\text{set}_1$ is at most $\varepsilon/2$.

Moreover, by the same lemma we may (and do) choose them so that

$$\varepsilon^{\text{sim}}_{n_{\text{layer}}}\le\rho_{\text{nbhd}},\qquad n_{\text{layer}}=1,\dots,N_{\text{layer}}.$$

Fix once and for all a scale

$$a\in(0,\varepsilon_{\text{ln}}^{-1/2}).$$

For each layer n_{layer}, apply the construction from the proof of Lemma J.2 with target accuracy $\varepsilon^{\text{sim}}_{n_{\text{layer}}}$ and prescribed scale a. This yields a required tokenwise LN-approximation tolerance $\eta_{\text{LN}}^{(n_{\text{layer}})}>0$ such that the layer simulation error is $\le\varepsilon^{\text{sim}}_{n_{\text{layer}}}$ whenever

$$\sup_{u\in\{v_t:\,v\in\widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}},\ t=0,\dots,T-1\}}\left\|\pi_{\text{dyn}}(\,\text{LN}_{\varepsilon_{\text{ln}}}(\Phi_{c,m_{\text{sc}}}(u)))-au\right\|_2\le\eta_{\text{LN}}^{(n_{\text{layer}})}.$$

Set

$$\eta_{\text{LN}}:=\min_{1\le n_{\text{layer}}\le N_{\text{layer}}}\eta_{\text{LN}}^{(n_{\text{layer}})}.$$

Applying the proof of Lemma J.1 to the compact token set S, choose one even $m_{\text{sc}}\ge2$ and one $c>0$ such that:

- the induced reference scale equals the prescribed a, and
- $$\sup_{u\in S}\left\|\pi_{\text{dyn}}(\,\text{LN}_{\varepsilon_{\text{ln}}}(\Phi_{c,m_{\text{sc}}}(u)))-au\right\|_2\le\eta_{\text{LN}}.$$

Let $m:=m_0+m_{\text{sc}}$ and write $\Phi:=\Phi_{c,m_{\text{sc}}}$.

For each n_{layer}, apply the construction of Lemma J.2 with this common scaffold (m_{sc},c) to obtain a pre-norm LN concrete Sessa block

$$\widetilde{G}_{n_{\text{layer}}}\in\text{ConcreteSessaBlocks}_{\text{LN}_{\varepsilon_{\text{ln}}}}(d_{k,0},m)$$

viewed as a map

$$\widetilde{G}_{n_{\text{layer}}} : \mathbb{R}^{T\times m} \to \mathbb{R}^{T\times m}$$

such that

$$\sup_{h\in\widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}}} \left\|\pi_{\text{dyn}}(\widetilde{G}_{n_{\text{layer}}}(\Phi(h))) - G_{n_{\text{layer}}}(h)\right\|_F \le \varepsilon^{\text{sim}}_{n_{\text{layer}}}.$$

and

$$\pi_{\text{sc}}(\widetilde{G}_{n_{\text{layer}}}(\Phi(h))) \equiv s_{c,m_{\text{sc}}} \qquad \forall h \in \widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}}.$$

Define the induced dynamic maps

$$G^{\text{dyn}}_{n_{\text{layer}}} : \widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}} \to \mathbb{R}^{T\times m_0}, \qquad G^{\text{dyn}}_{n_{\text{layer}}}(h) := \pi_{\text{dyn}}(\widetilde{G}_{n_{\text{layer}}}(\Phi(h))).$$

Then

$$\sup_{h\in\widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}}} \|G^{\text{dyn}}_{n_{\text{layer}}}(h) - G_{n_{\text{layer}}}(h)\|_F \le \varepsilon^{\text{sim}}_{n_{\text{layer}}}.$$

Moreover, by scaffold invariance,

$$\widetilde{G}_{n_{\text{layer}}}(\Phi(h)) = \Phi(G^{\text{dyn}}_{n_{\text{layer}}}(h)) \qquad \forall h \in \widetilde{\mathcal{K}\_\text{set}}_{n_{\text{layer}}}.$$

Applying Lemma I.9 to the maps $G_{n_{\text{layer}}}$ and $G^{\text{dyn}}_{n_{\text{layer}}}$ on the dynamic space $\mathbb{R}^{T\times m_0}$ yields

$$\sup_{x\in\mathcal{D}} \left\|G^\star(\text{Embed}_0(x)) - (G^{\text{dyn}}_{N_{\text{layer}}} \circ \cdots \circ G^{\text{dyn}}_1)(\text{Embed}_0(x))\right\|_F \le \varepsilon/2.$$

Define

$$G_{\text{ln}} := \widetilde{G}_{N_{\text{layer}}} \circ \cdots \circ \widetilde{G}_1 \in \Omega^{d_{k,0}}_{\text{Sessa,LN}_{\varepsilon_{\text{ln}}}}(m).$$

Finally, define new adapters

$$\text{Embed}(x) := \Phi(\text{Embed}_0(x)) \in \mathbb{R}^{T\times m}, \qquad \text{Unembed}(u) := \text{Unembed}_0(\pi_{\text{dyn}}(u)).$$

Since

$$\text{Unembed}_0(h)_t = R_{\text{out}}(\pi_{\text{out}}(h_t)),$$

with π_{out} an orthogonal projection and R_{out} an isometry, Unembed_0 is non-expansive in Frobenius norm.

$$\text{Unembed}(G_{\text{ln}}(\text{Embed}(x))) = \text{Unembed}_0((G^{\text{dyn}}_{N_{\text{layer}}} \circ \cdots \circ G^{\text{dyn}}_1)(\text{Embed}_0(x))) \qquad \forall x \in \mathcal{D}.$$

Therefore,

$$\sup_{x\in\mathcal{D}} \left\|\text{Unembed}_0(G^\star(\text{Embed}_0(x))) - \text{Unembed}(G_{\text{ln}}(\text{Embed}(x)))\right\|_F \le \varepsilon/2.$$

Combining this with the approximation error $\varepsilon/2$ from the Norm = Id case gives the claim. □

K Proofs for flexible finite-horizon selective retrieval

Lemma K.1 (Predecessor focusing from ordered codes)**.** *Fix $T \ge 1$ and $\mu \in (0,1)$. Let $I_0 < I_1 < \cdots < I_T$ be pairwise disjoint compact intervals in $\mathbb{R}$, and assume all of them lie in $(0,\infty)$. Then there exist scalar linear feedback-query/key maps on a single coordinate such that for every token sequence u satisfying*

$$\langle u_t, e_{\text{pos}}\rangle \in I_t, \qquad 0 \le t \le T,$$

the resulting strict-past feedback attention row satisfies

$$\alpha^b_{t,t-1} \geq 1-\mu, \qquad \sum_{j=0}^{t-2} \alpha^b_{t,j} \leq \mu, \qquad 1 \leq t \leq T.$$

Proof. If $T = 1$, the claim is trivial, since the strict past of $t = 1$ contains only the index 0. Assume henceforth that $T \geq 2$. Let

$$z_t := \langle u_t, e_{\mathrm{pos}} \rangle, \qquad 0 \leq t \leq T.$$

By assumption,

$$z_t \in I_t, \qquad I_0 < I_1 < \cdots < I_T \subset (0, \infty).$$

To implement the focusing inside an actual LN-free Sessa block, we first realize a single dedicated post-GELU scalar coordinate carrying a strictly ordered positive code. Choose one a-branch coordinate to be

$$a^{\mathrm{pos}}_t = c\, z_t$$

with some fixed $c > 0$. Since $z_t > 0$ on all intervals and the exact GELU satisfies

$$\mathrm{GELU}'(x) = \Phi(x) + x\phi(x) > 0 \qquad (x > 0),$$

the scalar map $x \mapsto \mathrm{GELU}(cx)$ is strictly increasing on $(0, \infty)$. Hence the post-GELU coordinate

$$\xi_t := \mathrm{GELU}(cz_t)$$

ranges in compact intervals

$$J_t := \mathrm{GELU}(cI_t)$$

satisfying

$$J_0 < J_1 < \cdots < J_T \subset (0, \infty).$$

Now define scalar feedback queries and keys from that post-GELU coordinate:

$$q^b_t = \Lambda\, \xi_t, \qquad k^b_j = \Lambda\, \xi_j,$$

with $\Lambda > 0$ to be chosen. All unused heads and coordinates are set to zero.

Let

$$m_t := \inf J_t, \qquad M_t := \sup J_t.$$

For $2 \leq t \leq T$, compactness and strict ordering give

$$\Delta_t := m_{t-1} - M_{t-2} > 0.$$

Set

$$\Delta := \min_{2 \leq t \leq T} \Delta_t > 0, \qquad m_* := \min_{0 \leq t \leq T} m_t > 0.$$

For every $2 \leq t \leq T$, every $j \leq t-2$, and every admissible input u,

$$q^b_t k^b_{t-1} - q^b_t k^b_j = \Lambda^2 \xi_t (\xi_{t-1} - \xi_j) \geq \Lambda^2 m_* \Delta.$$

Hence each non-predecessor strict-past logit is smaller than the predecessor logit by at least

$$\Lambda^2 m_* \Delta.$$

Therefore

$$\sum_{j=0}^{t-2} \exp\Big(\langle q_t^b, k_j^b\rangle - \langle q_t^b, k_{t-1}^b\rangle\Big) \le T\, e^{-\Lambda^2 m_* \Delta}.$$

Choose Λ so large that

$$T\, e^{-\Lambda^2 m_* \Delta} \le \frac{\mu}{1-\mu}.$$

Then the softmax formula yields

$$\alpha_{t,t-1}^b = \frac{1}{1 + \sum_{j=0}^{t-2} e^{\langle q_t^b, k_j^b\rangle - \langle q_t^b, k_{t-1}^b\rangle}} \ge 1-\mu,$$

and consequently

$$\sum_{j=0}^{t-2} \alpha_{t,j}^b \le \mu.$$

For $t = 1$ the strict past contains only the predecessor 0, so the claim is trivial. □

Lemma K.2 (RoPE self-focusing)**.** *Fix $T \ge 0$ and $\mu \in (0,1)$. Let $I_0 < I_1 < \cdots < I_T$ be pairwise disjoint compact intervals in $(0,\infty)$. Then there exist forward query/key maps realized inside a single actual RoPE forward branch of an LN-free Sessa block such that for every token sequence u satisfying*

$$\langle u_t, e_{\mathrm{pos}}\rangle \in I_t, \qquad 0 \le t \le T,$$

the resulting full-prefix forward attention row satisfies

$$\alpha_{t,t}^f \ge 1-\mu, \qquad \sum_{j=0}^{t-1} \alpha_{t,j}^f \le \mu, \qquad 0 \le t \le T.$$

Proof. If $T = 0$, the statement is trivial. Assume henceforth that $T \ge 1$. Let

$$z_t := \langle u_t, e_{\mathrm{pos}}\rangle, \qquad z_t \in I_t.$$

As in the proof of Lemma K.1, choose one dedicated a-branch coordinate

$$a_t^{\mathrm{pos}} = c\, z_t$$

with $c > 0$, and let

$$\xi_t := \mathrm{GELU}(c z_t).$$

Because $z_t > 0$ and GELU is strictly increasing on $(0,\infty)$, the ranges

$$J_t := \mathrm{GELU}(c I_t)$$

are compact, strictly ordered, and positive:

$$J_0 < J_1 < \cdots < J_T \subset (0,\infty).$$

Let

$$m_t := \inf J_t, \qquad M_t := \sup J_t.$$

Since the intervals are strictly ordered and compact,

$$\delta_t := m_t - M_{t-1} > 0, \qquad 1 \le t \le T.$$

Set

$$\delta := \min_{1\le t\le T} \delta_t > 0, \qquad m_* := \min_{0\le t\le T} m_t > 0.$$

Now realize the forward query/key pair on a single RoPE plane by setting, before RoPE,

$$q_t^f = \Lambda\, \xi_t\, e_1, \qquad k_j^f = \Lambda\, \xi_j\, e_1$$

inside the first 2-dimensional RoPE plane, with all other coordinates and heads set to zero. Let

$$\ell_{t,j} := \sigma_k \langle \mathrm{RoPE}(q_t^f), \mathrm{RoPE}(k_j^f) \rangle.$$

Then for every $j \le t$,

$$\ell_{t,j} = \sigma_k \Lambda^2 \xi_t \xi_j \cos(\vartheta_t - \vartheta_j)$$

for the corresponding RoPE phases ϑ_t, ϑ_j on that plane. Hence for every $j < t$,

$$\begin{aligned} \ell_{t,t} - \ell_{t,j} &= \sigma_k \Lambda^2 \xi_t \Big(\xi_t - \xi_j \cos(\vartheta_t - \vartheta_j) \Big) \\ &\ge \sigma_k \Lambda^2 \xi_t (\xi_t - \xi_j) \qquad \text{since } \cos(\cdot) \le 1 \\ &\ge \sigma_k \Lambda^2 m_* \delta. \end{aligned}$$

Therefore, for every $1 \le t \le T$,

$$\sum_{j=0}^{t-1} \exp(\ell_{t,j} - \ell_{t,t}) \le T\, e^{-\sigma_k \Lambda^2 m_* \delta}.$$

Choose Λ so large that

$$T\, e^{-\sigma_k \Lambda^2 m_* \delta} \le \frac{\mu}{1-\mu}.$$

Then the softmax formula gives

$$\alpha_{t,t}^f = \frac{1}{1 + \sum_{j=0}^{t-1} e^{\ell_{t,j} - \ell_{t,t}}} \ge 1 - \mu,$$

and consequently

$$\sum_{j=0}^{t-1} \alpha_{t,j}^f \le \mu.$$

For $t = 0$ the statement is trivial. □

Lemma K.3 (Scaled GELU uniformly approximates ReLU)**.** *Assume the exact GELU activation*

$$\mathrm{GELU}(x) = x\, \Phi(x).$$

For $L > 0$, define

$$R_L(u) := \frac{1}{L}\, \mathrm{GELU}(Lu).$$

Then

$$\sup_{u\in\mathbb{R}} \big| R_L(u) - u_+ \big| \le \frac{1}{L\sqrt{2\pi}}, \qquad u_+ := \max\{u, 0\}.$$

Proof. Since $\mathrm{GELU}(x) = x\Phi(x)$,

$$R_L(u) = u\, \Phi(Lu).$$

If $u \geq 0$, then

$$R_L(u) - u_+ = u\Phi(Lu) - u = -u(1 - \Phi(Lu)).$$

By the Mills bound

$$1 - \Phi(v) \leq \frac{\phi(v)}{v} \qquad (v > 0),$$

we obtain for $u > 0$,

$$|R_L(u) - u_+| = u(1 - \Phi(Lu)) \leq \frac{\phi(Lu)}{L} \leq \frac{1}{L\sqrt{2\pi}}.$$

The same bound is trivial at $u = 0$.

If $u < 0$, then $u_+ = 0$ and

$$|R_L(u)| = |u|\,\Phi(Lu) = |u|\,(1 - \Phi(-Lu)).$$

Applying the same Mills bound with $v = -Lu > 0$ yields

$$|R_L(u)| \leq \frac{\phi(-Lu)}{L} = \frac{\phi(Lu)}{L} \leq \frac{1}{L\sqrt{2\pi}}.$$

Combining the two cases proves the claim. □

Lemma K.4 (Symmetrized scaled GELU equals the identity)**.** *Assume the exact GELU activation*

$$\mathrm{GELU}(x) = x\,\Phi(x).$$

For $L > 0$, define

$$R_L(x) := \frac{1}{L}\,\mathrm{GELU}(Lx), \qquad \mathrm{Id}_L(x) := R_L(x) - R_L(-x).$$

Then

$$\mathrm{Id}_L(x) = x \qquad \forall x \in \mathbb{R}.$$

In particular,

$$\sup_{x \in \mathbb{R}} |\,\mathrm{Id}_L(x) - x| = 0 \leq \frac{2}{L\sqrt{2\pi}}.$$

Proof. Since $\mathrm{GELU}(x) = x\Phi(x)$,

$$R_L(x) = x\Phi(Lx).$$

Hence

$$\mathrm{Id}_L(x) = x\Phi(Lx) - (-x)\Phi(-Lx) = x\big(\Phi(Lx) + \Phi(-Lx)\big) = x,$$

because $\Phi(-z) = 1 - \Phi(z)$. □

Corollary K.5 (Exact channel read on the a-branch)**.** *Fix a unit vector $e \in \mathbb{R}^m$ and $L > 0$. In an LN-free concrete Sessa block, if two a-coordinates are chosen as*

$$a_t^{(+)} = L\langle u_t, e\rangle, \qquad a_t^{(-)} = -L\langle u_t, e\rangle,$$

then the corresponding post-GELU coordinates satisfy

$$\frac{1}{L}\Big(\bar{a}_t^{(+)} - \bar{a}_t^{(-)}\Big) = \langle u_t, e\rangle \qquad \forall\, t.$$

Hence any scalar input channel can be read exactly by a linear value projection from two a-slots.

Proof. Apply Lemma K.4 pointwise with $x = \langle u_t, e\rangle$. □

Lemma K.6 (Plateau window from four scaled GELUs)**.** *Fix $T \geq 0$ and pairwise disjoint compact intervals*

$$I_0 < I_1 < \cdots < I_T \subset (0, \infty).$$

Fix a target index $\tau_* \in \{0, \ldots, T\}$ and an accuracy parameter $\eta \in (0, 1)$. Then there exist real numbers

$$a_- < a_+ < b_- < b_+$$

and a scalar function $W_\eta : \mathbb{R} \to \mathbb{R}$ of the form

$$W_\eta(x) = \frac{R_L(x - a_-) - R_L(x - a_+)}{a_+ - a_-} - \frac{R_L(x - b_-) - R_L(x - b_+)}{b_+ - b_-}$$

for some $L > 0$, such that

$$|W_\eta(x) - 1| \leq \eta \quad \text{for } x \in I_{\tau_*},$$
$$|W_\eta(x)| \leq \eta \quad \text{for } x \in \bigcup_{t \neq \tau_*} I_t,$$

and

$$\sup_{x \in \mathbb{R}} |W_\eta(x)| \leq 1 + \eta.$$

Moreover, W_η is realizable exactly as a linear combination of four a-branch GELU coordinates inside a single LN-free Sessa block.

Proof. Because the intervals are pairwise disjoint, compact, and strictly ordered, one can choose

$$a_- < a_+ < \inf I_{\tau_*} \leq \sup I_{\tau_*} < b_- < b_+$$

such that

$$I_{\tau_*} \subset [a_+, b_-], \qquad \bigcup_{t \neq \tau_*} I_t \subset (-\infty, a_-] \cup [b_+, \infty).$$

Define the exact piecewise-linear plateau window

$$w(x) := \frac{(x - a_-)_+ - (x - a_+)_+}{a_+ - a_-} - \frac{(x - b_-)_+ - (x - b_+)_+}{b_+ - b_-}.$$

By construction,

$$w(x) = 1 \quad \text{on } [a_+, b_-] \supset I_{\tau_*},$$
$$w(x) = 0 \quad \text{on } (-\infty, a_-] \cup [b_+, \infty) \supset \bigcup_{t \neq \tau_*} I_t,$$

and

$$0 \leq w(x) \leq 1 \qquad \forall x \in \mathbb{R}.$$

Now replace each ReLU ramp by the scaled-GELU ramp from Lemma K.3:

$$R_L(u) = \frac{1}{L} \operatorname{GELU}(Lu).$$

Set

$$W_L(x) := \frac{R_L(x - a_-) - R_L(x - a_+)}{a_+ - a_-} - \frac{R_L(x - b_-) - R_L(x - b_+)}{b_+ - b_-}.$$

Using Lemma K.3 on each of the four ramp terms,

$$\|W_L - w\|_\infty \le \frac{2}{L\sqrt{2\pi}} \left(\frac{1}{a_+ - a_-} + \frac{1}{b_+ - b_-} \right).$$

Choose L so large that the right-hand side is at most η. Then on I_{τ_*}, where $w \equiv 1$,

$$|W_L - 1| \le \eta,$$

and on $\bigcup_{t \neq \tau_*} I_t$, where $w \equiv 0$,

$$|W_L| \le \eta.$$

Also, since $0 \le w \le 1$,

$$|W_L(x)| \le |w(x)| + \eta \le 1 + \eta \qquad \forall x.$$

Set $W_\eta := W_L$.

Finally, W_η is realizable exactly inside one LN-free Sessa block because each term

$$R_L(x - c) = \frac{1}{L} \operatorname{GELU}(L(x - c))$$

is one a-branch GELU coordinate applied to an affine function of the tokenwise scalar x, and the displayed linear combination is absorbed into the value projection. □

Lemma K.7 (Writing a window into an auxiliary channel)**.** *Fix $T \ge 0$, $\tau_* \in \{0, \dots, T\}$, and $\varepsilon \in (0, 1)$. Let $\mathcal{K}\_\text{set} \subset (\mathbb{R}^m)^{T+1}$ be compact. Assume that for some unit vector $e_{\text{pos}} \in \mathbb{R}^m$,*

$$I_t := \{\langle u_t, e_{\text{pos}} \rangle : \ u \in \mathcal{K}\_\text{set}\}, \qquad 0 \le t \le T,$$

are compact and strictly ordered:

$$I_0 < I_1 < \cdots < I_T \subset (0, \infty).$$

Fix orthonormal directions

$$e_{\text{pos}}, \ e_{\text{sig}}, \ e_{\text{aux}}$$

and let $E_{\text{carry}} \subset \mathbb{R}^m$ be any fixed subspace orthogonal to all three. Assume moreover that $m \ge 6$. Then there exists a single LN-free Sessa block

$$W^{\text{write}}_{T,\tau_*,\varepsilon} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

such that the feedback branch is switched off, the e_{pos}-, e_{sig}-, and E_{carry}-channels are preserved exactly, and, writing

$$a_t(u) := \langle W^{\text{write}}_{T,\tau_*,\varepsilon}(u)_t, \ e_{\text{aux}} \rangle,$$

one has uniformly on $\mathcal{K}\_\text{set}$,

$$|a_{\tau_*}(u) - 1| \le \varepsilon, \qquad |a_t(u)| \le \varepsilon \quad (t \neq \tau_*),$$

and

$$\sup_{u \in \mathcal{K}\_\text{set}} \sup_{0 \le t \le T} |a_t(u)| \le 2.$$

Proof. Choose $\eta \in (0, \varepsilon)$ so small that

$$\eta + \eta(1 + \eta) \le \varepsilon.$$

Apply Lemma K.6 to obtain a scalar function W_η satisfying

$$|W_\eta(x) - 1| \le \eta \quad (x \in I_{\tau_*}), \qquad |W_\eta(x)| \le \eta \quad (x \in \bigcup_{t \neq \tau_*} I_t), \qquad \sup_x |W_\eta(x)| \le 1 + \eta.$$

Next apply Lemma K.2 with parameter $\mu := \eta$. This gives a forward branch whose full-prefix row satisfies

$$\alpha^f_{t,t} \geq 1 - \eta, \qquad \sum_{j<t} \alpha^f_{t,j} \leq \eta \qquad (0 \leq t \leq T).$$

We now build the block.

Values. Choose a positive constant c_1 such that

$$\mathrm{GELU}(c_1) = 1.$$

Realize the first value coordinate by a constant a-branch coordinate equal to c_1, so that

$$v^{(0)}_t \equiv 1.$$

Realize the second value coordinate as

$$v^{(1)}_t = W_\eta(\langle u_t, e_{\mathrm{pos}} \rangle),$$

using Lemma K.6.

Gate and output on the auxiliary channel. Choose two gate coordinates

$$g^{(0)}_t = \langle u_t, e_{\mathrm{aux}} \rangle, \qquad g^{(1)}_t \equiv 1.$$

Choose the output projection on the e_{aux}-channel with coefficients $(-1, +1)$ on the two gated coordinates and zero on all other channels. Because the row sum of attention is exactly 1,

$$s^{(0)}_t = \sum_{j \leq t} \alpha^f_{t,j} \cdot 1 = 1.$$

Hence the auxiliary output becomes

$$a_t(u) = \langle u_t, e_{\mathrm{aux}} \rangle - s^{(0)}_t \langle u_t, e_{\mathrm{aux}} \rangle + s^{(1)}_t = s^{(1)}_t,$$

where

$$s^{(1)}_t = \sum_{j \leq t} \alpha^f_{t,j} W_\eta(\langle u_j, e_{\mathrm{pos}} \rangle).$$

Thus the block overwrites the auxiliary channel by the forward average of W_η.

All other output columns are zero, so the e_{pos}-, e_{sig}-, and E_{carry}-channels are preserved exactly.

It remains to bound $a_t = s^{(1)}_t$.

Target time $t = \tau_*$. All indices $j < \tau_*$ are off-target, hence

$$|W_\eta(\langle u_j, e_{\mathrm{pos}} \rangle)| \leq \eta.$$

At the target index,

$$W_\eta(\langle u_{\tau_*}, e_{\mathrm{pos}} \rangle) \in [1 - \eta, 1 + \eta].$$

Therefore

$$a_{\tau_*}(u) \geq (1 - \eta)(1 - \eta) - \eta \cdot \eta \geq 1 - 2\eta,$$

and

$$a_{\tau_*}(u) \leq (1 - \eta)(1 + \eta) + \eta \cdot \eta \leq 1 + \eta.$$

Hence

$$|a_{\tau_*}(u) - 1| \leq 2\eta \leq \varepsilon.$$

Off-target times $t < \tau_*$. Then all visible indices $j \le t$ are off-target, so

$$|a_t(u)| \le \eta \le \varepsilon.$$

Off-target times $t > \tau_*$. Then self-mass is on an off-target index, so the self contribution is at most η in magnitude, while all nonself mass is at most η and every visible value has magnitude at most $1 + \eta$. Thus

$$|a_t(u)| \le \eta + \eta(1 + \eta) \le \varepsilon.$$

Finally, from

$$|a_t(u)| = |s_t^{(1)}| \le \sum_{j \le t} \alpha_{t,j}^f \sup_x |W_\eta(x)| \le 1 + \eta \le 2,$$

we obtain the uniform bound. □

Definition 11 (Signal-fiber saturation)**.** Fix $T \ge 0$, a unit signal direction $e_{\text{sig}} \in \mathbb{R}^m$, and a set $\mathcal{K}\_\text{set} \subset (\mathbb{R}^m)^{T+1}$. For $\delta \ge 0$, define

$$\text{Sat}_\delta^{\text{sig}}(\mathcal{K}\_\text{set}) := \left\{ u + z : \ u \in \mathcal{K}\_\text{set}, \ z_t = a_t e_{\text{sig}}, \ \max_{0 \le t \le T} |a_t| \le \delta \right\}.$$

Equivalently,

$$\text{Sat}_\delta^{\text{sig}}(\mathcal{K}\_\text{set}) = \left\{ u + \sum_{t=0}^{T} a_t e_{\text{sig}} \mathbf{1}[\cdot = t] : u \in \mathcal{K}\_\text{set}, \ \max_t |a_t| \le \delta \right\}.$$

If $\mathcal{K}_$set is compact, then $\text{Sat}_\delta^{\text{sig}}(\mathcal{K}\_\text{set})$ is compact.

Definition 12 (Exact signal transport)**.** Fix $T \ge 0$, a unit signal direction $e_{\text{sig}} \in \mathbb{R}^m$, and a control subspace $E_{\text{ctrl}} \subset \mathbb{R}^m$ with $e_{\text{sig}} \perp E_{\text{ctrl}}$. Let Π_{ctrl} denote the orthogonal projection onto E_{ctrl}, and let

$$\pi_{\text{sig}}(v) := \langle v, e_{\text{sig}} \rangle.$$

For $u = (u_t)_{t=0}^T \in (\mathbb{R}^m)^{T+1}$, write

$$c_t^u := \Pi_{\text{ctrl}} u_t, \qquad x_t^u := \pi_{\text{sig}}(u_t).$$

A causal map

$$B : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

is said to have *exact signal transport along e_{sig} over E_{ctrl}* on a set $\mathcal{K}\_\text{set} \subset (\mathbb{R}^m)^{T+1}$ if:

(i) B preserves the control channels exactly:

$$\Pi_{\text{ctrl}} B(u)_t = c_t^u \qquad \forall\, u \in \mathcal{K}\_\text{set}, \ \forall\, 0 \le t \le T;$$

(ii) there exists a scalar lower-triangular kernel

$$\mathcal{T}_B^u(i, j), \qquad 0 \le j \le i \le T,$$

depending only on the control stream $c^u = (c_t^u)_{t=0}^T$, such that

$$\pi_{\text{sig}}(B(u)_i) = \sum_{j=0}^{i} \mathcal{T}_B^u(i, j)\, x_j^u \qquad \forall\, u \in \mathcal{K}\_\text{set}, \ \forall\, 0 \le i \le T.$$

Lemma K.8 (Transport calculus on signal fibers)**.** *Fix $T \ge 0$, e_{sig}, E_{ctrl}, and a compact set $\mathcal{K}_$set $\subset (\mathbb{R}^m)^{T+1}$. Fix $\delta > 0$.*

(i) ***Jacobian extraction.*** *Assume B is continuously differentiable on a neighborhood of $\mathrm{Sat}_\delta^{\mathrm{sig}}(\mathcal{K}\_\mathrm{set})$, and that B has signal-blind exact scalar transport along e_{sig} over E_{ctrl} on $\mathrm{Sat}_\delta^{\mathrm{sig}}(\mathcal{K}\_\mathrm{set})$, with kernel $\mathcal{T}_B^u$. Then for every $u \in \mathcal{K}\_\mathrm{set}$ and every $0 \le j \le i \le T$,*

$$e_{\mathrm{sig}}^\top \frac{\partial B(u)_i}{\partial u_j} e_{\mathrm{sig}} = \mathcal{T}_B^u(i,j).$$

(ii) ***Composition.*** *Assume B_1 has signal-blind exact scalar transport along e_{sig} over E_{ctrl} on $\mathrm{Sat}_\delta^{\mathrm{sig}}(\mathcal{K}\_\mathrm{set})$, with kernel $\mathcal{T}_{B_1}^u$, and preserves the control channels exactly there. Assume B_2 has signal-blind exact scalar transport along e_{sig} over E_{ctrl} on $B_1(\mathrm{Sat}_\delta^{\mathrm{sig}}(\mathcal{K}\_\mathrm{set}))$, with kernel $\mathcal{T}_{B_2}^v$, and preserves the control channels exactly there. Then $B_2 \circ B_1$ also has signal-blind exact scalar transport along e_{sig} over E_{ctrl} on $\mathrm{Sat}_\delta^{\mathrm{sig}}(\mathcal{K}\_\mathrm{set})$, and its kernel is the lower-triangular kernel product*

$$\mathcal{T}_{B_2 \circ B_1}^u(i,j) = \sum_{r=j}^{i} \mathcal{T}_{B_2}^{B_1(u)}(i,r)\, \mathcal{T}_{B_1}^u(r,j).$$

Proof. For (i), fix $u \in \mathcal{K}\_\mathrm{set}$, $j \le i$, and define

$$u^{(h)} := u + h\, e_{\mathrm{sig}} \mathbf{1}[\cdot = j].$$

For $|h| < \delta$, one has $u^{(h)} \in \mathrm{Sat}_\delta^{\mathrm{sig}}(\mathcal{K}\_\mathrm{set})$. Because $e_{\mathrm{sig}} \perp E_{\mathrm{ctrl}}$,

$$\Pi_{\mathrm{ctrl}} u_t^{(h)} = \Pi_{\mathrm{ctrl}} u_t \qquad \forall\, t,$$

so the control stream is unchanged. Since the transport kernel depends only on the control stream, the same kernel $\mathcal{T}_B^u$ applies to both u and $u^{(h)}$. Therefore

$$\begin{aligned} \pi_{\mathrm{sig}}(B(u^{(h)})_i) - \pi_{\mathrm{sig}}(B(u)_i) &= \sum_{r=0}^{i} \mathcal{T}_B^u(i,r)(x_r^{u^{(h)}} - x_r^u) \\ &= h\, \mathcal{T}_B^u(i,j). \end{aligned}$$

Divide by h and let $h \to 0$. Since B is C^1,

$$e_{\mathrm{sig}}^\top \frac{\partial B(u)_i}{\partial u_j} e_{\mathrm{sig}} = \mathcal{T}_B^u(i,j).$$

For (ii), let $u \in \mathrm{Sat}_\delta^{\mathrm{sig}}(\mathcal{K}\_\mathrm{set})$. Because B_1 preserves the control channels exactly,

$$\Pi_{\mathrm{ctrl}} B_1(u)_t = \Pi_{\mathrm{ctrl}} u_t,$$

so the control stream of $B_1(u)$ equals that of u. Hence

$$\pi_{\mathrm{sig}}(B_1(u)_r) = \sum_{j=0}^{r} \mathcal{T}_{B_1}^u(r,j)\, x_j^u.$$

Applying B_2 and using exact control preservation again,

$$\begin{aligned}\pi_{\mathrm{sig}}(B_2(B_1(u))_i) &= \sum_{r=0}^{i} \mathcal{T}_{B_2}^{B_1(u)}(i,r)\,\pi_{\mathrm{sig}}(B_1(u)_r)\\ &= \sum_{r=0}^{i} \mathcal{T}_{B_2}^{B_1(u)}(i,r) \sum_{j=0}^{r} \mathcal{T}_{B_1}^{u}(r,j)\,x_j^u\\ &= \sum_{j=0}^{i} \left(\sum_{r=j}^{i} \mathcal{T}_{B_2}^{B_1(u)}(i,r)\,\mathcal{T}_{B_1}^{u}(r,j) \right) x_j^u.\end{aligned}$$

This is exactly the stated kernel-product formula. □

Definition 13 (Transparent preprocessing)**.** Fix $T \geq 0$, a unit signal direction $e_{\mathrm{sig}} \in \mathbb{R}^m$, and a control subspace $E_{\mathrm{ctrl}} \subset \mathbb{R}^m$ with $e_{\mathrm{sig}} \perp E_{\mathrm{ctrl}}$. Let

$$\Pi_{\mathrm{ctrl}} : \mathbb{R}^m \to E_{\mathrm{ctrl}}$$

be the orthogonal projection and

$$\pi_{\mathrm{sig}}(v) := \langle v, e_{\mathrm{sig}} \rangle.$$

A causal map

$$R : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

is said to be *signal-transparent along* e_{sig} *over* E_{ctrl} on a set $\mathcal{K}\_\mathrm{set} \subset (\mathbb{R}^m)^{T+1}$ if for every $u \in \mathcal{K}\_\mathrm{set}$, every $\tau \in \{0, \dots, T\}$, and every sufficiently small scalar a such that

$$u^{(a,\tau)} := u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau]$$

remains in the domain under consideration, one has

$$\Pi_{\mathrm{ctrl}} R(u^{(a,\tau)})_t = \Pi_{\mathrm{ctrl}} R(u)_t \qquad \forall\, t,$$

and

$$\pi_{\mathrm{sig}}(R(u^{(a,\tau)})_t) = \pi_{\mathrm{sig}}(R(u)_t) + a\, \mathbf{1}[t = \tau] \qquad \forall\, t.$$

Lemma K.9 (Transparent preprocessing and Jacobians)**.** *Fix $T \geq 0$, e_{sig}, and E_{ctrl}. Let*

$$R : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}, \qquad B : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

be continuously differentiable on neighborhoods of $\mathcal{K}\_\mathrm{set}$ and $\mathrm{Sat}_\delta^{\mathrm{sig}}(R(\mathcal{K}\_\mathrm{set}))$, respectively, for some $\delta > 0$.

Assume:

(i) R is signal-transparent along e_{sig} over E_{ctrl} on $\mathcal{K}\_\mathrm{set}$;

(ii) B has signal-blind exact scalar transport along e_{sig} over E_{ctrl} on $\mathrm{Sat}_\delta^{\mathrm{sig}}(R(\mathcal{K}\_\mathrm{set}))$, with kernel

$$\mathcal{T}_B^v(i,j), \qquad v \in \mathrm{Sat}_\delta^{\mathrm{sig}}(R(\mathcal{K}\_\mathrm{set})), \quad 0 \leq j \leq i \leq T.$$

Then for every $u \in \mathcal{K}\_\mathrm{set}$ and every $0 \leq j \leq i \leq T$,

$$e_{\mathrm{sig}}^\top \frac{\partial (B \circ R)(u)_i}{\partial u_j} e_{\mathrm{sig}} = \mathcal{T}_B^{R(u)}(i,j).$$

Proof. Fix $u \in \mathcal{K}\_\mathrm{set}$ and $0 \leq j \leq i \leq T$. For sufficiently small a, define

$$u^{(a,j)} := u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = j].$$

Set

$$v := R(u), \qquad v^{(a)} := R(u^{(a,j)}).$$

By signal-transparency of R,

$$\Pi_{\text{ctrl}} v_t^{(a)} = \Pi_{\text{ctrl}} v_t \qquad \forall\, t,$$

and

$$\pi_{\text{sig}}(v_t^{(a)}) = \pi_{\text{sig}}(v_t) + a\,\mathbf{1}[t = j] \qquad \forall\, t.$$

Hence $v^{(a)} \in \text{Sat}_\delta^{\text{sig}}(R(\mathcal{K}\_\text{set}))$ for all sufficiently small $|a|$, and $v^{(a)}$ and v have the same control stream. Therefore the same kernel $\mathcal{T}_B^v$ applies to both v and $v^{(a)}$, so

$$\begin{aligned}\pi_{\text{sig}}(B(v^{(a)})_i) - \pi_{\text{sig}}(B(v)_i) &= \sum_{r=0}^{i} \mathcal{T}_B^v(i,r)\big(\pi_{\text{sig}}(v_r^{(a)}) - \pi_{\text{sig}}(v_r)\big) \\ &= a\,\mathcal{T}_B^v(i,j).\end{aligned}$$

Divide by a and let $a \to 0$. Since $B \circ R$ is continuously differentiable,

$$e_{\text{sig}}^\top \frac{\partial (B \circ R)(u)_i}{\partial u_j} e_{\text{sig}} = \mathcal{T}_B^{R(u)}(i,j).$$

□

Corollary K.10 (Signal-fiber stability of the control-driven blocks)**.** *Fix $\delta \geq 0$. In each of Lemmas K.11, K.12, K.15, K.17, and K.20, replace the base compact set $\mathcal{K}\_\text{set}$ (or $\mathcal{K}\_\text{set}_H$) by its bounded signal-fiber saturation $\text{Sat}_\delta^{\text{sig}}(\mathcal{K}\_\text{set})$ (or $\text{Sat}_\delta^{\text{sig}}(\mathcal{K}\_\text{set}_H)$). Then the same concrete block or network satisfies the same conclusion, with the same constants.*

In particular, whenever one of these lemmas yields signal-blind exact scalar transport along e_{sig}, that exact transport statement also holds on every bounded signal-fiber saturation of the same control-side compact set.

Proof. In each listed lemma, the hypotheses and parameter choices depend only on channels orthogonal to e_{sig}: ordered positional ranges, two-sided tail/profile bounds, exact vanishing of designated scratch/profile channels, and carried control channels. These quantities are unchanged when $\mathcal{K}\_\text{set}$ is replaced by $\text{Sat}_\delta^{\text{sig}}(\mathcal{K}\_\text{set})$.

Moreover, the concrete constructions preserve the relevant control channels exactly and treat the e_{sig}-channel linearly. Therefore the original proofs apply verbatim on the saturated set, with the same constants. □

Lemma K.11 (Local multiplier)**.** *Fix $T \geq 0$ and $\delta > 0$. Let $\mathcal{K}\_\text{set} \subset (\mathbb{R}^m)^{T+1}$ be compact. Assume that for some unit vector $e_{\text{pos}} \in \mathbb{R}^m$,*

$$I_t := \{\langle u_t, e_{\text{pos}}\rangle : \ u \in \mathcal{K}\_\text{set}\}, \qquad 0 \leq t \leq T,$$

are compact and strictly ordered in $(0, \infty)$. Fix orthonormal directions

$$e_{\text{pos}},\ e_{\text{sig}},\ e_{\text{aux}}$$

and let $E_{\text{carry}} \subset \mathbb{R}^m$ be any fixed subspace orthogonal to all three. Assume moreover that $m \geq 4$. Assume moreover that the auxiliary channel is uniformly bounded:

$$\sup_{u \in \mathcal{K}\_\text{set}} \sup_{0 \leq t \leq T} \big|\langle u_t, e_{\text{aux}}\rangle\big| \leq M$$

for some finite M.

Then there exists a single LN-free Sessa block

$$M^{\mathrm{loc}}_{T,\delta} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

such that the feedback branch is switched off, the e_{pos}-, e_{aux}-, and E_{carry}-channels are preserved exactly, and $M^{\mathrm{loc}}_{T,\delta}$ has signal-blind exact scalar transport along e_{sig} over

$$E_{\mathrm{ctrl}} := \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{aux}}\} \oplus E_{\mathrm{carry}},$$

with diagonal kernel

$$\mathcal{T}^{u}_{M^{\mathrm{loc}}}(i,j) = D^{u}_{\mathrm{loc}}(i)\,\mathbf{1}[i=j];$$
$$\left|D^{u}_{\mathrm{loc}}(t) - \langle u_t, e_{\mathrm{aux}}\rangle\right| \le \delta \qquad \forall\, u \in \mathcal{K}\_\mathrm{set},\ \forall\, 0 \le t \le T.$$

In particular,

$$e_{\mathrm{sig}}^{\top} \frac{\partial M^{\mathrm{loc}}_{T,\delta}(u)_i}{\partial u_j} e_{\mathrm{sig}} = D^{u}_{\mathrm{loc}}(i)\,\mathbf{1}[i=j].$$

Proof. Choose a parameter

$$\mu \in (0,1)$$

to be fixed later, and apply Lemma K.2 with this μ.

Choose a positive constant c_1 such that

$$\mathrm{GELU}(c_1) = 1.$$

Realize one forward value coordinate by the constant 1:

$$v_t^{(0)} \equiv 1.$$

Next read the auxiliary channel exactly using Corollary K.5. Choose two a-slots

$$a_t^{(+)} = L\langle u_t, e_{\mathrm{aux}}\rangle, \qquad a_t^{(-)} = -L\langle u_t, e_{\mathrm{aux}}\rangle,$$

for any fixed $L > 0$, and choose the value projection so that

$$v_t^{(1)} = \frac{1}{L}(\bar{a}_t^{(+)} - \bar{a}_t^{(-)}) = \langle u_t, e_{\mathrm{aux}}\rangle.$$

Choose two gate coordinates, both equal to the signal:

$$g_t^{(0)} = \langle u_t, e_{\mathrm{sig}}\rangle, \qquad g_t^{(1)} = \langle u_t, e_{\mathrm{sig}}\rangle.$$

Choose the output projection on the e_{sig}-channel with coefficients $(-1,+1)$ on these two gated coordinates and zero on all other output channels.

Since the forward row sums to 1,

$$s_t^{(0)} = \sum_{j\le t} \alpha^{f}_{t,j} \cdot 1 = 1.$$

Hence the signal output equals

$$\langle M^{\mathrm{loc}}_{T,\delta}(u)_t, e_{\mathrm{sig}}\rangle = \langle u_t, e_{\mathrm{sig}}\rangle - s_t^{(0)}\langle u_t, e_{\mathrm{sig}}\rangle + s_t^{(1)}\langle u_t, e_{\mathrm{sig}}\rangle = s_t^{(1)}\langle u_t, e_{\mathrm{sig}}\rangle,$$

where

$$s_t^{(1)} = \sum_{j\le t} \alpha^{f}_{t,j}\, v_j^{(1)} = \sum_{j\le t} \alpha^{f}_{t,j}\, \langle u_j, e_{\mathrm{aux}}\rangle.$$

Define

$$D^u_{\mathrm{loc}}(t) := s^{(1)}_t.$$

Then

$$\langle M^{\mathrm{loc}}_{T,\delta}(u)_t, e_{\mathrm{sig}}\rangle = D^u_{\mathrm{loc}}(t)\,\langle u_t, e_{\mathrm{sig}}\rangle,$$

which is exactly signal-blind exact scalar transport with diagonal kernel

$$\mathcal{T}^u_{M^{\mathrm{loc}}}(i,j) = D^u_{\mathrm{loc}}(i)\mathbf{1}[i=j].$$

The coefficient $D^u_{\mathrm{loc}}(t)$ depends only on the forward weights and on the auxiliary values $\langle u_j, e_{\mathrm{aux}}\rangle$. By construction, both depend only on the e_{pos}-, e_{aux}-, and E_{carry}-channels, not on the signal channel. Thus the transport is signal-blind over

$$E_{\mathrm{ctrl}} := \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{aux}}\} \oplus E_{\mathrm{carry}}.$$

All output columns except the signal column are zero, so the e_{pos}-, e_{aux}-, and E_{carry}-channels are preserved exactly.

It remains to estimate $D^u_{\mathrm{loc}}(t)$. Since the auxiliary read is exact,

$$D^u_{\mathrm{loc}}(t) - \langle u_t, e_{\mathrm{aux}}\rangle = \sum_{j\le t} \alpha^f_{t,j}(\langle u_j, e_{\mathrm{aux}}\rangle - \langle u_t, e_{\mathrm{aux}}\rangle) = \sum_{j<t} \alpha^f_{t,j}(\langle u_j, e_{\mathrm{aux}}\rangle - \langle u_t, e_{\mathrm{aux}}\rangle).$$

Therefore,

$$|D^u_{\mathrm{loc}}(t) - \langle u_t, e_{\mathrm{aux}}\rangle| \le 2M \sum_{j<t} \alpha^f_{t,j}.$$

By self-focusing,

$$\sum_{j<t} \alpha^f_{t,j} \le \mu.$$

Hence

$$|D^u_{\mathrm{loc}}(t) - \langle u_t, e_{\mathrm{aux}}\rangle| \le 2M\mu.$$

Choose

$$\mu \le \min\Big\{\frac{1}{2},\ \frac{\delta}{2\max\{M,1\}}\Big\}.$$

Then

$$|D^u_{\mathrm{loc}}(t) - \langle u_t, e_{\mathrm{aux}}\rangle| \le \delta \qquad \forall\, u \in \mathcal{K}\_\mathrm{set},\ \forall\, 0 \le t \le T.$$

For any $\eta > 0$, replacing $\mathcal{K}\_\mathrm{set}$ by $\mathrm{Sat}^{\mathrm{sig}}_\eta(\mathcal{K}\_\mathrm{set})$ leaves the ordered positional ranges $(I_t)_{t=0}^T$ and the auxiliary bound M unchanged, since only the e_{sig}-channel is perturbed. The same concrete construction therefore yields the same exact diagonal signal-transport formula on $\mathrm{Sat}^{\mathrm{sig}}_\eta(\mathcal{K}\_\mathrm{set})$, with the same coefficients $D^u_{\mathrm{loc}}(i)$, because the forward weights depend only on the positional-control stream and the exact auxiliary read depends only on the e_{aux}-channel. Applying Lemma K.8(i) gives

$$e_{\mathrm{sig}}^\top \frac{\partial M^{\mathrm{loc}}_{T,\delta}(u)_i}{\partial u_j} e_{\mathrm{sig}} = D^u_{\mathrm{loc}}(i)\,\mathbf{1}[i=j].$$

□

Lemma K.12 (Two-block selector)**.** *Fix $T \ge 0$, $\varepsilon \in (0,1)$, and a compact set $\mathcal{K}\_\mathrm{set} \subset (\mathbb{R}^m)^{T+1}$. Assume that for some unit vector $e_{\mathrm{pos}} \in \mathbb{R}^m$ the scalar position ranges*

$$I_t := \{\langle u_t, e_{\mathrm{pos}}\rangle :\ u \in \mathcal{K}\_\mathrm{set}\}, \qquad 0 \le t \le T,$$

are compact and strictly ordered:

$$I_0 < I_1 < \cdots < I_T \subset (0, \infty).$$

Fix a source index $\tau_* \in \{0, \dots, T\}$ *and orthonormal directions*

$$e_{\text{pos}},\ e_{\text{sig}},\ e_{\text{aux}}.$$

Let $E_{\text{carry}} \subset \mathbb{R}^m$ *be any fixed subspace orthogonal to these three directions. Assume moreover that* $m \geq 6$*.*

Then there exists a depth-2 LN-free Sessa network

$$S_{T,\tau_*,\varepsilon} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

such that both constituent blocks have the feedback branch switched off, the e_{pos}*-channel and every channel in* E_{carry} *are preserved exactly, and* $S_{T,\tau_*,\varepsilon}$ *has signal-blind exact scalar transport along* e_{sig} *over*

$$E_{\text{ctrl}} := \text{span}\{e_{\text{pos}}\} \oplus E_{\text{carry}},$$

with diagonal kernel

$$\mathcal{T}_S^u(i,j) = D_{\text{sel}}^u(i)\, \mathbf{1}[i=j];$$

Uniformly for all $u \in \mathcal{K}\_\text{set}$*,*

$$\frac{1}{2} \leq D_{\text{sel}}^u(\tau_*) \leq 2, \qquad |D_{\text{sel}}^u(t)| \leq \varepsilon \quad (t \neq \tau_*).$$

In particular,

$$e_{\text{sig}}^\top \frac{\partial S_{T,\tau_*,\varepsilon}(u)_i}{\partial u_j} e_{\text{sig}} = D_{\text{sel}}^u(i)\, \mathbf{1}[i=j].$$

Proof. Set

$$\varepsilon_{\text{wr}} := \frac{\varepsilon}{4}, \qquad \delta_{\text{mul}} := \frac{\varepsilon}{4}.$$

Apply Lemma K.7 with accuracy ε_{wr}. This yields a forward-only block

$$W^{\text{write}}_{T,\tau_*,\varepsilon_{\text{wr}}}$$

which preserves the e_{pos}-, e_{sig}-, and E_{carry}-channels exactly and writes an auxiliary channel

$$a_t(u) := \langle W^{\text{write}}_{T,\tau_*,\varepsilon_{\text{wr}}}(u)_t,\ e_{\text{aux}} \rangle$$

satisfying

$$|a_{\tau_*}(u) - 1| \leq \frac{\varepsilon}{4}, \qquad |a_t(u)| \leq \frac{\varepsilon}{4} \quad (t \neq \tau_*),$$

and

$$|a_t(u)| \leq 2 \qquad \forall t.$$

Now apply Lemma K.11 to the image

$$\mathcal{K}\_\text{set}' := W^{\text{write}}_{T,\tau_*,\varepsilon_{\text{wr}}}(\mathcal{K}\_\text{set}),$$

with the same $e_{\text{pos}}, e_{\text{sig}}, e_{\text{aux}}, E_{\text{carry}}$, the bound $M = 2$, and accuracy $\delta_{\text{mul}} = \varepsilon/4$. This yields a forward-only block

$$M^{\text{loc}}_{T,\delta_{\text{mul}}}$$

whose signal transport is exact and diagonal:

$$\langle M^{\mathrm{loc}}_{T,\delta_{\mathrm{mul}}}(w)_t, e_{\mathrm{sig}}\rangle = D^{w}_{\mathrm{loc}}(t)\,\langle w_t, e_{\mathrm{sig}}\rangle \qquad (w \in \mathcal{K}\_\mathrm{set}'),$$

with

$$|D^{w}_{\mathrm{loc}}(t) - \langle w_t, e_{\mathrm{aux}}\rangle| \le \frac{\varepsilon}{4}.$$

Define

$$S_{T,\tau_*,\varepsilon} := M^{\mathrm{loc}}_{T,\delta_{\mathrm{mul}}} \circ W^{\mathrm{write}}_{T,\tau_*,\varepsilon_{\mathrm{wr}}}.$$

Since the writer preserves the signal channel exactly,

$$\langle W^{\mathrm{write}}_{T,\tau_*,\varepsilon_{\mathrm{wr}}}(u)_t, e_{\mathrm{sig}}\rangle = \langle u_t, e_{\mathrm{sig}}\rangle.$$

Therefore

$$\langle S_{T,\tau_*,\varepsilon}(u)_t, e_{\mathrm{sig}}\rangle = D^{W^{\mathrm{write}}(u)}_{\mathrm{loc}}(t)\,\langle u_t, e_{\mathrm{sig}}\rangle.$$

Set

$$D^{u}_{\mathrm{sel}}(t) := D^{W^{\mathrm{write}}(u)}_{\mathrm{loc}}(t).$$

Then

$$\langle S_{T,\tau_*,\varepsilon}(u)_t, e_{\mathrm{sig}}\rangle = D^{u}_{\mathrm{sel}}(t)\,\langle u_t, e_{\mathrm{sig}}\rangle,$$

so the signal transport is exact and diagonal.

The coefficient $D^{u}_{\mathrm{sel}}(t)$ depends only on the e_{pos}-, e_{aux}-, and E_{carry}-channels of the intermediate state $W^{\mathrm{write}}(u)$. The writer preserves e_{pos} and E_{carry} exactly, and its written auxiliary channel $a_t(u)$ is itself a deterministic function of the positional-control coordinate only. Hence $D^{u}_{\mathrm{sel}}(t)$ depends only on the original e_{pos}- and E_{carry}-channels, not on the signal channel. Thus the transport is signal-blind over E_{ctrl}.

The e_{pos}-channel and all of E_{carry} are preserved exactly by both blocks, hence by the composition.

Finally, at the selected source,

$$|D^{u}_{\mathrm{sel}}(\tau_*) - 1| \le |D^{u}_{\mathrm{sel}}(\tau_*) - a_{\tau_*}(u)| + |a_{\tau_*}(u) - 1| \le \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2},$$

so since $\varepsilon < 1$,

$$\frac{1}{2} \le D^{u}_{\mathrm{sel}}(\tau_*) \le \frac{3}{2} < 2.$$

For $t \neq \tau_*$,

$$|D^{u}_{\mathrm{sel}}(t)| \le |D^{u}_{\mathrm{sel}}(t) - a_t(u)| + |a_t(u)| \le \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2} \le \varepsilon.$$

For any $\eta > 0$, replacing $\mathcal{K}\_\mathrm{set}$ by $\mathrm{Sat}^{\mathrm{sig}}_{\eta}(\mathcal{K}\_\mathrm{set})$ leaves the ordered positional ranges $(I_t)_{t=0}^{T}$ unchanged. Moreover, in the concrete two-block construction, the writer depends only on the positional coordinate and preserves the signal channel exactly, while the local multiplier depends only on the positional and auxiliary channels and acts diagonally on the signal channel. Hence the same concrete construction yields the same exact diagonal signal-transport formula on $\mathrm{Sat}^{\mathrm{sig}}_{\eta}(\mathcal{K}\_\mathrm{set})$, with the same coefficients $D^{u}_{\mathrm{sel}}(i)$. Applying Lemma K.8(i) gives

$$e_{\mathrm{sig}}^{\top} \frac{\partial S_{T,\tau_*,\varepsilon}(u)_i}{\partial u_j} e_{\mathrm{sig}} = D^{u}_{\mathrm{sel}}(i)\,\mathbf{1}[i = j].$$

□

Remark K.13 (The selector depends only on position). In the concrete construction used in the proof of Lemma K.12,

the diagonal transport coefficient $D^u_{\mathrm{sel}}(t)$ depends only on the positional stream

$$(\langle u_s, e_{\mathrm{pos}}\rangle)_{s=0}^{T},$$

and is independent of the signal channel e_{sig} and of the carried channels E_{carry}.

Lemma K.14 (Selector preserves signal fibers)**.** *Under the hypotheses of Lemma K.12, let*

$$S_{T,\tau_*,\varepsilon} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

be the selector block constructed there. Then for every $\delta \geq 0$ there exists $\delta' = \delta'(\delta, \mathcal{K}\_\mathrm{set}) < \infty$ such that

$$S_{T,\tau_*,\varepsilon}(\mathrm{Sat}^{\mathrm{sig}}_{\delta}(\mathcal{K}\_\mathrm{set})) \subset \mathrm{Sat}^{\mathrm{sig}}_{\delta'}(S_{T,\tau_*,\varepsilon}(\mathcal{K}\_\mathrm{set})).$$

More precisely, if

$$u' = u + \sum_{t=0}^{T} a_t e_{\mathrm{sig}} \mathbf{1}[\cdot = t], \qquad u \in \mathcal{K}\_\mathrm{set}, \qquad \max_t |a_t| \leq \delta,$$

then

$$S_{T,\tau_*,\varepsilon}(u')_i = S_{T,\tau_*,\varepsilon}(u)_i + D^u_{\mathrm{sel}}(i)\, a_i\, e_{\mathrm{sig}}, \qquad 0 \leq i \leq T,$$

where $D^u_{\mathrm{sel}}(i)$ is the selector transport coefficient from Lemma K.12. In particular, one may take

$$\delta' := \delta \sup_{u \in \mathcal{K}\_\mathrm{set}} \sup_{0 \leq i \leq T} |D^u_{\mathrm{sel}}(i)| \leq 2\delta.$$

Proof. Fix $u \in \mathcal{K}\_\mathrm{set}$ and

$$u' = u + \sum_{t=0}^{T} a_t e_{\mathrm{sig}} \mathbf{1}[\cdot = t]$$

with $\max_t |a_t| \leq \delta$.

By Remark K.13, the coefficient $D^u_{\mathrm{sel}}(i)$ depends only on the positional stream

$$(\langle u_s, e_{\mathrm{pos}}\rangle)_{s=0}^{T},$$

which is unchanged under perturbations along e_{sig}. Moreover, in the concrete construction of $S_{T,\tau_*,\varepsilon}$, all non-signal output channels are independent of the input signal channel: the writer $W^{\mathrm{write}}_{T,\tau_*,\varepsilon_{\mathrm{wr}}}$ preserves e_{sig} exactly and writes only the auxiliary channel as a function of the positional coordinate, while $M^{\mathrm{loc}}_{T,\delta_{\mathrm{mul}}}$ preserves the positional and auxiliary channels exactly and modifies the output only on the signal channel.

Therefore

$$S_{T,\tau_*,\varepsilon}(u')_i = S_{T,\tau_*,\varepsilon}(u)_i + D^u_{\mathrm{sel}}(i)\, a_i\, e_{\mathrm{sig}},$$

and the claim follows. □

Lemma K.15 (Active diffusive transport)**.** *Fix $\beta \in (0,1)$ and set $\gamma := 1-\beta$. Let $T \geq 0$ and let $\mathcal{K}\_\mathrm{set} \subset (\mathbb{R}^m)^{T+1}$ be compact. Assume that for some orthonormal directions*

$$e_{\mathrm{pos}},\ e_{\mathrm{sig}},\ e_{\mathrm{src}},\ e_{\mathrm{tgt}} \in \mathbb{R}^m$$

the scalar position ranges

$$I_t := \{\langle u_t, e_{\mathrm{pos}}\rangle :\ u \in \mathcal{K}\_\mathrm{set}\}, \qquad 0 \leq t \leq T,$$

are compact and strictly ordered:

$$I_0 < I_1 < \cdots < I_T \subset (0, \infty).$$

Let $E_{\mathrm{carry}} \subset \mathbb{R}^m$ be any fixed subspace orthogonal to $e_{\mathrm{pos}}, e_{\mathrm{sig}}, e_{\mathrm{src}}, e_{\mathrm{tgt}}$.

Then there exists a depth-2 LN-free Sessa network

$$A^{\mathrm{act}}_{T,\beta} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

such that the first constituent block has the feedback branch switched off, while the second constituent block uses a strict-past uniform feedback solve with constant gain γ, the e_{pos}-channel and every channel in E_{carry} are preserved exactly, and $A^{\mathrm{act}}_{T,\beta}$ has signal-blind exact scalar transport along e_{sig} over

$$E_{\mathrm{ctrl}} := \mathrm{span}\{e_{\mathrm{pos}}\} \oplus E_{\mathrm{carry}},$$

with kernel

$$\mathcal{T}^{u}_{A^{\mathrm{act}}}(i,j) = D^{u}_{\mathrm{act}}(i)\,\mathbf{1}[i=j] + K^{u}_{\mathrm{act}}(i,j)\,\mathbf{1}[j<i];$$

There exist constants

$$0 < \underline{d}_{\mathrm{act}} \le \overline{d}_{\mathrm{act}} < \infty, \qquad 0 < a^{-}_{\mathrm{act}} \le a^{+}_{\mathrm{act}} < \infty,$$

depending only on β, but independent of T, such that

$$\underline{d}_{\mathrm{act}} \le D^{u}_{\mathrm{act}}(i) \le \overline{d}_{\mathrm{act}}, \qquad 0 \le i \le T,$$

and

$$a^{-}_{\mathrm{act}}(j+1)^{-\gamma}(i+1)^{-\beta} \le K^{u}_{\mathrm{act}}(i,j) \le a^{+}_{\mathrm{act}}(j+1)^{-\gamma}(i+1)^{-\beta}, \qquad 0 \le j < i \le T.$$

In particular,

$$e_{\mathrm{sig}}^{\top} \frac{\partial A^{\mathrm{act}}_{T,\beta}(u)_i}{\partial u_j} e_{\mathrm{sig}} = D^{u}_{\mathrm{act}}(i)\,\mathbf{1}[i=j] + K^{u}_{\mathrm{act}}(i,j)\,\mathbf{1}[j<i].$$

Proof. We construct

$$A^{\mathrm{act}}_{T,\beta} = R_{T,\beta} \circ C_T,$$

where C_T is a forward-only copy block and $R_{T,\beta}$ is a single feedback-transport block.

Step 1: copy of the signal into a scratch source channel. Build a forward-only LN-free Sessa block

$$C_T : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

such that

$$\langle C_T(u)_t, e_{\mathrm{src}} \rangle = \langle u_t, e_{\mathrm{sig}} \rangle \qquad (0 \le t \le T),$$

while the e_{pos}-, e_{sig}-, e_{tgt}-, and E_{carry}-channels are preserved exactly.

Switch off the feedback branch and choose two forward value coordinates equal to 1:

$$v_t^{(0)} \equiv 1, \qquad v_t^{(1)} \equiv 1.$$

Hence

$$s_t^{(0)} = 1, \qquad s_t^{(1)} = 1.$$

Choose the associated gate coordinates

$$g_t^{(0)} = \langle u_t, e_{\mathrm{src}} \rangle, \qquad g_t^{(1)} = \langle u_t, e_{\mathrm{sig}} \rangle,$$

and choose the output projection on the e_{src}-channel with coefficients $(-1, +1)$. Then

$$\langle C_T(u)_t, e_{\mathrm{src}} \rangle = \langle u_t, e_{\mathrm{src}} \rangle - \langle u_t, e_{\mathrm{src}} \rangle + \langle u_t, e_{\mathrm{sig}} \rangle = \langle u_t, e_{\mathrm{sig}} \rangle.$$

Let

$$w := C_T(u), \qquad x_j := \langle u_j, e_{\mathrm{sig}} \rangle.$$

Then

$$\langle w_j, e_{\mathrm{src}} \rangle = x_j, \qquad \langle w_j, e_{\mathrm{sig}} \rangle = x_j. \tag{79}$$

Step 2: the feedback-transport block. Now build a single LN-free Sessa block

$$R_{T,\beta} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}.$$

On one dedicated feedback channel, choose all feedback queries and keys identically zero. Then the strict-past feedback softmax is exactly uniform:

$$\alpha^b_{i,j} = \frac{1}{i}, \qquad 0 \le j < i, \quad 1 \le i \le T.$$

Choose the feedback gain to be the constant

$$\gamma_i \equiv \gamma = 1 - \beta.$$

Hence the scalar feedback matrix on that channel is

$$B_{i,j} = \frac{\gamma}{i} \mathbf{1}[j < i].$$

For the forward branch, fix $\mu_T \in (0, \frac{1}{2}]$, to be chosen below, and apply Lemma K.2 to the image $C_T(\mathcal{K}\_\text{set})$ on the ordered positional-control coordinate. Because C_T preserves the e_{pos}-channel exactly, the hypotheses still hold. This yields weights $\alpha^f_{i,j}(w)$ satisfying

$$\alpha^f_{i,i}(w) \ge 1 - \mu_T, \qquad \sum_{j=0}^{i-1} \alpha^f_{i,j}(w) \le \mu_T, \qquad 0 \le i \le T. \tag{80}$$

In particular, for every $j < i$,

$$\alpha^f_{i,j}(w) \le \mu_T. \tag{81}$$

To read the source scratch channel exactly, use Corollary K.5 on the input w and the direction e_{src}: choose two a-slots

$$a_j^{(+)} = L\langle w_j, e_{\mathrm{src}} \rangle, \qquad a_j^{(-)} = -L\langle w_j, e_{\mathrm{src}} \rangle.$$

Choose W_V so that one forward value coordinate is

$$v_j^{\mathrm{src}} = \frac{1}{L}(\bar{a}_j^{(+)} - \bar{a}_j^{(-)}) = \langle w_j, e_{\mathrm{src}} \rangle = x_j.$$

Let

$$f_i := \sum_{j \le i} \alpha^f_{i,j}(w)\, v_j^{\mathrm{src}} = \sum_{j \le i} \alpha^f_{i,j}(w)\, x_j$$

be the forward signal entering the scalar feedback solve, and let s_i denote the corresponding solve output:

$$s_0 = f_0, \qquad s_i = f_i + \frac{\gamma}{i} \sum_{j<i} s_j, \qquad 1 \le i \le T.$$

Choose the gate on that transport coordinate to be the constant 1, and choose the output projection so that the

signal channel receives exactly $+s_i$, while the e_{pos}- and E_{carry}-channels are untouched. Therefore

$$\langle R_{T,\beta}(w)_i, e_{\text{sig}} \rangle = \langle w_i, e_{\text{sig}} \rangle + s_i = x_i + s_i.$$

Step 3: resolvent kernel. Let

$$\Theta_{i,j} := [(I - B)^{-1}]_{i,j}, \qquad 0 \le j \le i \le T.$$

Then $\Theta_{i,i} = 1$, and for $j < i$,

$$\Theta_{i,j} = \frac{\gamma}{i} \sum_{r=j}^{i-1} \Theta_{r,j}.$$

As in the original proof, define

$$S_i^{(j)} := \sum_{r=j}^{i} \Theta_{r,j}.$$

Then $S_j^{(j)} = 1$ and

$$S_i^{(j)} = \Big(1 + \frac{\gamma}{i}\Big) S_{i-1}^{(j)},$$

hence

$$S_i^{(j)} = \frac{\Gamma(i+1+\gamma)\Gamma(j+1)}{\Gamma(j+1+\gamma)\Gamma(i+1)}.$$

Therefore, for $j < i$,

$$\Theta_{i,j} = \frac{\gamma}{i} S_{i-1}^{(j)} = \gamma \, \frac{\Gamma(j+1)}{\Gamma(j+1+\gamma)} \frac{\Gamma(i+\gamma)}{\Gamma(i+1)}.$$

Since $\gamma \in (0,1)$, standard Gamma-ratio bounds yield constants

$$0 < c_\Theta^- \le c_\Theta^+ < \infty$$

depending only on β, such that

$$c_\Theta^-(j+1)^{-\gamma}(i+1)^{-\beta} \le \Theta_{i,j} \le c_\Theta^+(j+1)^{-\gamma}(i+1)^{-\beta}, \qquad 0 \le j < i \le T. \tag{82}$$

Also, since $\gamma = 1 - \beta \in (0,1)$,

$$\sum_{r=1}^{n} r^{-\gamma} \lesssim_\beta n^\beta.$$

Combining this with (82), there exists a constant $C_\Sigma < \infty$, depending only on β, such that

$$\sum_{k=j+1}^{i} \Theta_{i,k} \le C_\Sigma \qquad (0 \le j < i \le T). \tag{83}$$

Finally, since $j + 1 \le i + 1 \le T + 1$,

$$\Theta_{i,j} \ge c_\Theta^-(i+1)^{-1} \ge \frac{c_\Theta^-}{T+1}. \tag{84}$$

Step 4: transport formula. Since $s = \Theta f$,

$$s_i = \sum_{k=0}^{i} \Theta_{i,k} f_k = \sum_{k=0}^{i} \Theta_{i,k} \sum_{j=0}^{k} \alpha_{k,j}^f(w) x_j = \sum_{j=0}^{i} \Big(\sum_{k=j}^{i} \Theta_{i,k} \alpha_{k,j}^f(w) \Big) x_j.$$

Therefore

$$\langle A^{\mathrm{act}}_{T,\beta}(u)_i, e_{\mathrm{sig}}\rangle = x_i + s_i = \Big(1 + \alpha^f_{i,i}(w)\Big)x_i + \sum_{j<i}\Big(\sum_{k=j}^{i}\Theta_{i,k}\alpha^f_{k,j}(w)\Big)x_j.$$

Define

$$D^u_{\mathrm{act}}(i) := 1 + \alpha^f_{i,i}(w), \qquad K^u_{\mathrm{act}}(i,j) := \sum_{k=j}^{i}\Theta_{i,k}\alpha^f_{k,j}(w) \quad (j<i).$$

Then

$$\langle A^{\mathrm{act}}_{T,\beta}(u)_i, e_{\mathrm{sig}}\rangle = D^u_{\mathrm{act}}(i)\, x_i + \sum_{j<i} K^u_{\mathrm{act}}(i,j)\, x_j.$$

This is exact scalar transport. The coefficients depend only on the positional stream of w, because the forward weights α^f were built from the positional-control coordinate only; and C_T preserves the positional coordinate exactly, so this is the same as the positional stream of u. The e_{pos}- and E_{carry}-channels are preserved exactly by construction. Thus the transport is signal-blind over E_{ctrl}.

Step 5: kernel bounds. From (80),

$$1 - \mu_T \le \alpha^f_{i,i}(w) \le 1,$$

so

$$2 - \mu_T \le D^u_{\mathrm{act}}(i) \le 2.$$

Since $\mu_T \le \frac{1}{2}$,

$$\frac{3}{2} \le D^u_{\mathrm{act}}(i) \le 2.$$

Thus we may take

$$\underline{d}_{\mathrm{act}} := \frac{3}{2}, \qquad \overline{d}_{\mathrm{act}} := 2.$$

For the off-diagonal coefficient, all summands are nonnegative. Hence for $j < i$,

$$K^u_{\mathrm{act}}(i,j) \ge \Theta_{i,j}\alpha^f_{j,j}(w) \ge (1-\mu_T)\Theta_{i,j} \ge \frac{1}{2}\,\Theta_{i,j}.$$

Combining with (82) gives

$$K^u_{\mathrm{act}}(i,j) \ge \frac{1}{2}\,c^-_\Theta (j+1)^{-\gamma}(i+1)^{-\beta}.$$

For the upper bound,

$$K^u_{\mathrm{act}}(i,j) = \Theta_{i,j}\alpha^f_{j,j}(w) + \sum_{k=j+1}^{i}\Theta_{i,k}\alpha^f_{k,j}(w) \le \Theta_{i,j} + \mu_T \sum_{k=j+1}^{i}\Theta_{i,k},$$

by (81). Now choose

$$\mu_T := \min\Big\{\frac{1}{2},\ \frac{c^-_\Theta}{4C_\Sigma(T+1)}\Big\}.$$

Then by (83),

$$\mu_T \sum_{k=j+1}^{i}\Theta_{i,k} \le \frac{c^-_\Theta}{4(T+1)}.$$

By (84),

$$\frac{c^-_\Theta}{4(T+1)} \le \frac{1}{4}\,\Theta_{i,j}.$$

Hence

$$K^u_{\mathrm{act}}(i,j) \le \frac{5}{4}\,\Theta_{i,j}.$$

Using (82),

$$K^u_{\mathrm{act}}(i,j) \le \frac{5}{4}\,c^+_\Theta (j+1)^{-\gamma}(i+1)^{-\beta}.$$

Thus the stated two-sided bounds hold with

$$a^-_{\mathrm{act}} := \frac{1}{2}\,c^-_\Theta, \qquad a^+_{\mathrm{act}} := \frac{5}{4}\,c^+_\Theta.$$

For any $\eta > 0$, replacing $\mathcal{K}\_\mathrm{set}$ by $\mathrm{Sat}^{\mathrm{sig}}_\eta(\mathcal{K}\_\mathrm{set})$ leaves the ordered positional ranges $(I_t)_{t=0}^T$ unchanged. In the concrete construction, the copy block writes the source scratch channel from the signal channel exactly and is independent of the incoming e_{src}-channel, while the transport block uses forward and feedback weights depending only on the positional stream and reads the copied source scratch channel exactly. Hence the same concrete construction yields the same exact scalar transport formula on $\mathrm{Sat}^{\mathrm{sig}}_\eta(\mathcal{K}\_\mathrm{set})$, with the same coefficients $D^u_{\mathrm{act}}(i)$ and $K^u_{\mathrm{act}}(i,j)$. Applying Lemma K.8(i) gives

$$e_{\mathrm{sig}}^\top \frac{\partial A^{\mathrm{act}}_{T,\beta}(u)_i}{\partial u_j} e_{\mathrm{sig}} = D^u_{\mathrm{act}}(i)\,\mathbf{1}[i=j] + K^u_{\mathrm{act}}(i,j)\,\mathbf{1}[j<i].$$

□

Remark K.16 (Active diffusive transport depends only on position). In the concrete construction used in the proof of Lemma K.15, the coefficients

$$D^u_{\mathrm{act}}(i), \qquad K^u_{\mathrm{act}}(i,j), \qquad 0 \le j < i \le T,$$

depend only on the positional stream

$$(\langle u_s, e_{\mathrm{pos}}\rangle)_{s=0}^T,$$

and are independent of the signal channel e_{sig} and of the carried channels E_{carry}.

Lemma K.17 (Transparent source-0 tail channel). *Fix $\beta \in (0,1)$, set $\gamma := 1-\beta$, fix $\tau_{\max} \ge 0$, and let*

$$L_H := \tau_{\max} + H.$$

Let $\mathcal{K}\_\mathrm{set}_H \subset (\mathbb{R}^m)^{L_H+1}$ be compact. Assume orthonormal directions

$$e_{\mathrm{sig}},\ e_{\mathrm{pos}},\ e_{\mathrm{tail}},\ e_{\mathrm{aux}},\ e_{\mathrm{src}},\ e_{\mathrm{tgt}} \in \mathbb{R}^m$$

and a subspace $E_{\mathrm{carry}} \subset \mathbb{R}^m$ orthogonal to all six, such that

$$I_t := \{\langle u_t, e_{\mathrm{pos}}\rangle :\ u \in \mathcal{K}\_\mathrm{set}_H\}, \qquad 0 \le t \le L_H,$$

are compact and strictly ordered:

$$I_0 < I_1 < \cdots < I_{L_H} \subset (0,\infty).$$

Then there exists a constant-depth LN-free Sessa network

$$T^{\mathrm{tail}}_H : (\mathbb{R}^m)^{L_H+1} \to (\mathbb{R}^m)^{L_H+1}$$

such that the e_{sig}-channel, the positional-control coordinate e_{pos}, and every channel in E_{carry} are preserved exactly

and, writing

$$g_t(u) := \langle T_H^{\mathrm{tail}}(u)_t,\ e_{\mathrm{tail}}\rangle, \qquad 0 \le t \le L_H,$$

there exist constants $c_g^-, c_g^+ > 0$, independent of H, such that

$$c_g^-(t+1)^{-\beta} \le g_t(u) \le c_g^+(t+1)^{-\beta}, \qquad 0 \le t \le L_H, \quad u \in \mathcal{K}\_\mathrm{set}_H;$$

T_H^{tail} is signal-transparent along e_{sig} with respect to the control pair

$$(e_{\mathrm{pos}}, e_{\mathrm{tail}}) :$$

for every $u \in \mathcal{K}\_\mathrm{set}_H$, every $\tau \in \{0, \dots, L_H\}$, and every scalar $a \in \mathbb{R}$,

$$\langle T_H^{\mathrm{tail}}(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t,\ e_{\mathrm{pos}}\rangle = \langle T_H^{\mathrm{tail}}(u)_t,\ e_{\mathrm{pos}}\rangle,$$

$$\langle T_H^{\mathrm{tail}}(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t,\ e_{\mathrm{tail}}\rangle = \langle T_H^{\mathrm{tail}}(u)_t,\ e_{\mathrm{tail}}\rangle,$$

$$\langle T_H^{\mathrm{tail}}(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t,\ e_{\mathrm{sig}}\rangle = \langle T_H^{\mathrm{tail}}(u)_t,\ e_{\mathrm{sig}}\rangle + a\, \mathbf{1}[t = \tau], \qquad 0 \le t \le L_H.$$

Proof. All auxiliary directions used below are part of the hypotheses; no fresh direction is chosen inside the construction. We construct

$$T_H^{\mathrm{tail}} = A_H^{\mathrm{tail}} \circ S_H^{\mathrm{tail}} \circ C_H,$$

where C_H writes a constant seed on the prescribed tail direction e_{tail}, S_H^{tail} selects source 0 on that tail channel, and A_H^{tail} transports the selected seed by the active diffusive block.

Step 1: constant seed writer on the prescribed tail direction. Build a forward-only LN-free Sessa block

$$C_H : (\mathbb{R}^m)^{L_H+1} \to (\mathbb{R}^m)^{L_H+1}$$

as follows.

Choose two forward value coordinates equal to 1:

$$v_t^{(0)} \equiv 1, \qquad v_t^{(1)} \equiv 1.$$

Hence the corresponding forward aggregates satisfy

$$s_t^{(0)} = 1, \qquad s_t^{(1)} = 1.$$

Choose two gate coordinates

$$g_t^{(0)} = \langle u_t, e_{\mathrm{tail}}\rangle, \qquad g_t^{(1)} \equiv 1,$$

and choose the output projection on the e_{tail}-channel with coefficients $(-1, +1)$ on these two gated coordinates and zero on all other output channels. Then

$$\langle C_H(u)_t, e_{\mathrm{tail}}\rangle = \langle u_t, e_{\mathrm{tail}}\rangle - s_t^{(0)}\langle u_t, e_{\mathrm{tail}}\rangle + s_t^{(1)} = 1.$$

Thus C_H overwrites the e_{tail}-channel by the constant seed 1.

Because the output projection vanishes on the e_{sig}-, e_{pos}-, and E_{carry}-channels, these channels are preserved exactly:

$$\langle C_H(u)_t, e_{\mathrm{sig}}\rangle = \langle u_t, e_{\mathrm{sig}}\rangle, \qquad \langle C_H(u)_t, e_{\mathrm{pos}}\rangle = \langle u_t, e_{\mathrm{pos}}\rangle,$$

and likewise on E_{carry}.

Moreover, since the written tail seed is constant and independent of the input, for every $a \in \mathbb{R}$,

$$\langle C_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t, e_{\mathrm{tail}} \rangle = \langle C_H(u)_t, e_{\mathrm{tail}} \rangle = 1,$$

while the e_{sig}-channel passes through exactly. So C_H is already signal-transparent along e_{sig} with respect to $(e_{\mathrm{pos}}, e_{\mathrm{tail}})$.

Step 2: positional selector on the tail channel. Let

$$\mathcal{K}\_\mathrm{set}_H^{(1)} := C_H(\mathcal{K}\_\mathrm{set}_H).$$

Apply Lemma K.12 to $\mathcal{K}\_\mathrm{set}_H^{(1)}$ with signal direction $e_{\mathrm{sig}}^{\mathrm{sel}} := e_{\mathrm{tail}}$, positional-control direction e_{pos}, auxiliary direction e_{aux}, source index $\tau_* = 0$, and carried-through subspace

$$E_{\mathrm{carry}}^{\mathrm{sel}} := \mathrm{span}\{e_{\mathrm{sig}}\} \oplus E_{\mathrm{carry}}.$$

Choose an exponent $M > \beta$ and set

$$\varepsilon_H := c_0 (H+1)^{-M},$$

where $c_0 > 0$ will be chosen later. The lemma yields a depth-2 network

$$S_H^{\mathrm{tail}} := S_{L_H, 0, \varepsilon_H}$$

which preserves e_{pos}, the original e_{sig}, and every channel in E_{carry} exactly, and whose exact diagonal transport on the tail channel is

$$\langle S_H^{\mathrm{tail}}(v)_t, e_{\mathrm{tail}} \rangle = D_{\mathrm{sel}}^{v}(t)\, \langle v_t, e_{\mathrm{tail}} \rangle.$$

Since $\langle C_H(u)_t, e_{\mathrm{tail}} \rangle \equiv 1$, the selected seed stream is

$$z_t(u) := \langle S_H^{\mathrm{tail}}(C_H(u))_t, e_{\mathrm{tail}} \rangle = D_{\mathrm{sel}}^{C_H(u)}(t).$$

By Lemma K.12,

$$\frac{1}{2} \le z_0(u) \le 2, \qquad |z_t(u)| \le \varepsilon_H \quad (t \ge 1).$$

By Remark K.13, in the concrete construction of $S_H^{\mathrm{tail}} = S_{L_H, 0, \varepsilon_H}$ the coefficient

$$D_{\mathrm{sel}}^{C_H(u)}(t)$$

depends only on the positional stream

$$(\langle C_H(u)_s, e_{\mathrm{pos}} \rangle)_{s=0}^{L_H}.$$

Since C_H preserves the positional coordinate exactly,

$$\langle C_H(u)_s, e_{\mathrm{pos}} \rangle = \langle u_s, e_{\mathrm{pos}} \rangle, \qquad 0 \le s \le L_H,$$

it follows that

$$z_t(u) = D_{\mathrm{sel}}^{C_H(u)}(t)$$

depends only on the original positional stream and not on the original signal channel.

Step 3: active diffusive transport on the same prescribed tail direction. Let

$$\mathcal{K}\_\mathrm{set}_H^{(2)} := S_H^{\mathrm{tail}}(\mathcal{K}\_\mathrm{set}_H^{(1)}).$$

Apply Lemma K.15 to $\mathcal{K}\_\mathrm{set}_H^{(2)}$ with positional direction e_{pos}, signal direction $e_{\mathrm{sig}}^{\mathrm{act}} := e_{\mathrm{tail}}$, scratch directions $e_{\mathrm{src}}, e_{\mathrm{tgt}}$, and carried-through subspace

$$E_{\mathrm{carry}}^{\mathrm{act}} := \mathrm{span}\{e_{\mathrm{sig}}\} \oplus E_{\mathrm{carry}}.$$

Denote the resulting network by

$$A_H^{\mathrm{tail}}.$$

By the lemma, A_H^{tail} preserves e_{pos}, the original e_{sig}, and E_{carry} exactly, and has exact scalar transport on the tail channel:

$$\langle A_H^{\mathrm{tail}}(w)_t, e_{\mathrm{tail}}\rangle = D_{\mathrm{act}}^w(t)\,\langle w_t, e_{\mathrm{tail}}\rangle + \sum_{j<t} K_{\mathrm{act}}^w(t,j)\,\langle w_j, e_{\mathrm{tail}}\rangle.$$

Therefore, for

$$g_t(u) := \langle T_H^{\mathrm{tail}}(u)_t, e_{\mathrm{tail}}\rangle,$$

we have

$$g_t(u) = D_{\mathrm{act}}^w(t)\,z_t(u) + \sum_{j<t} K_{\mathrm{act}}^w(t,j)\,z_j(u), \qquad w := S_H^{\mathrm{tail}}(C_H(u)).$$

By Remark K.16, in the concrete construction of A_H^{tail} the coefficients

$$D_{\mathrm{act}}^w(t), \qquad K_{\mathrm{act}}^w(t,j)$$

depend only on the positional stream

$$(\langle w_s, e_{\mathrm{pos}}\rangle)_{s=0}^{L_H}.$$

Since both C_H and S_H^{tail} preserve the positional coordinate exactly, this is the same as the original positional stream of u. Hence these coefficients are independent of the original signal channel.

Step 4: two-sided tail bounds. At $t = 0$, the sum is empty, so

$$g_0(u) = D_{\mathrm{act}}^w(0)\,z_0(u).$$

By Lemma K.15,

$$\underline{d}_{\mathrm{act}} \le D_{\mathrm{act}}^w(0) \le \overline{d}_{\mathrm{act}},$$

hence

$$\frac{1}{2}\,\underline{d}_{\mathrm{act}} \le g_0(u) \le 2\,\overline{d}_{\mathrm{act}}.$$

Now fix $t \ge 1$. Using the exact transport formula, the bounds on $z_j(u)$, and the coefficient bounds from Lemma K.15, we obtain

$$\begin{aligned} g_t(u) &\ge K_{\mathrm{act}}^w(t,0)\,z_0(u) - |D_{\mathrm{act}}^w(t) z_t(u)| - \sum_{j=1}^{t-1} K_{\mathrm{act}}^w(t,j)\,|z_j(u)| \\ &\ge \frac{1}{2}\,a_{\mathrm{act}}^-(t+1)^{-\beta} - \overline{d}_{\mathrm{act}}\,\varepsilon_H - a_{\mathrm{act}}^+\varepsilon_H \sum_{j=1}^{t-1}(j+1)^{-\gamma}(t+1)^{-\beta}. \end{aligned}$$

Since $\gamma = 1 - \beta \in (0,1)$,

$$\sum_{j=1}^{t-1}(j+1)^{-\gamma} \lesssim_\beta (t+1)^{\beta},$$

hence

$$g_t(u) \ge c_1(t+1)^{-\beta} - c_2\varepsilon_H$$

for constants $c_1, c_2 > 0$ independent of H.

Now $M > \beta$, so

$$\varepsilon_H = c_0(H+1)^{-M} \le c_0(H+1)^{-\beta}.$$

Also $0 \le t \le L_H = \tau_{\max} + H$, hence

$$(H+1)^{-\beta} \le (\tau_{\max}+1)^{\beta}(t+1)^{-\beta}.$$

Therefore

$$\varepsilon_H \lesssim_{\tau_{\max}} c_0(t+1)^{-\beta}.$$

Choosing $c_0 > 0$ sufficiently small makes the error absorbable, so

$$g_t(u) \ge c_g^-(t+1)^{-\beta}$$

for some $c_g^- > 0$ independent of H.

Similarly,

$$\begin{aligned} g_t(u) &\le |D^w_{\text{act}}(t) z_t(u)| + K^w_{\text{act}}(t,0)|z_0(u)| + \sum_{j=1}^{t-1} K^w_{\text{act}}(t,j)|z_j(u)| \\ &\le \overline{d}_{\text{act}}\varepsilon_H + 2a^+_{\text{act}}(t+1)^{-\beta} + a^+_{\text{act}}\varepsilon_H \sum_{j=1}^{t-1}(j+1)^{-\gamma}(t+1)^{-\beta}, \end{aligned}$$

hence

$$g_t(u) \le c_g^+(t+1)^{-\beta}$$

for some $c_g^+ < \infty$ independent of H.

Thus

$$c_g^-(t+1)^{-\beta} \le g_t(u) \le c_g^+(t+1)^{-\beta}, \qquad 0 \le t \le L_H.$$

Step 5: signal-transparency along e_{sig}. Let

$$u^{(a,\tau)} := u + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau].$$

Since $e_{\text{sig}} \perp e_{\text{pos}}$, we have

$$\langle u_t^{(a,\tau)}, e_{\text{pos}} \rangle = \langle u_t, e_{\text{pos}} \rangle \qquad \forall t.$$

By Step 1,

$$\langle C_H(u^{(a,\tau)})_t, e_{\text{tail}} \rangle = \langle C_H(u)_t, e_{\text{tail}} \rangle = 1,$$

and

$$\langle C_H(u^{(a,\tau)})_t, e_{\text{sig}} \rangle = \langle C_H(u)_t, e_{\text{sig}} \rangle + a\, \mathbf{1}[t = \tau].$$

By the dependence analysis in Step 2, $z_t(u)$ depends only on the positional stream, so

$$z_t(u^{(a,\tau)}) = z_t(u).$$

By the dependence analysis in Step 3, the coefficients $D^w_{\text{act}}, K^w_{\text{act}}$ also depend only on the positional stream, hence they are unchanged under the perturbation. Therefore the tail output is unchanged:

$$g_t(u^{(a,\tau)}) = g_t(u).$$

Since each constituent block preserves the original e_{sig}-channel exactly, the full composition satisfies

$$\langle T_H^{\text{tail}}(u^{(a,\tau)})_t, e_{\text{sig}}\rangle = \langle T_H^{\text{tail}}(u)_t, e_{\text{sig}}\rangle + a\,\mathbf{1}[t=\tau].$$

The e_{pos}-coordinate is preserved exactly at each stage as well. This proves signal-transparency. □

Lemma K.18 (Residual zero-writer)**.** *Fix $T \geq 0$, a compact set $\mathcal{K}\_\text{set} \subset (\mathbb{R}^m)^{T+1}$, orthonormal directions*

$$e_{\text{sig}},\ e_{\text{pos}},\ e_{\text{zero}} \in \mathbb{R}^m,$$

and a subspace $E_{\text{carry}} \subset \mathbb{R}^m$ orthogonal to all three. Then there exists a single LN-free Sessa block

$$Z_{T,e_{\text{zero}}} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

such that the feedback branch is switched off, the e_{sig}-channel, the e_{pos}-channel, and every channel in E_{carry} are preserved exactly, the prescribed channel is written to zero exactly:

$$\langle Z_{T,e_{\text{zero}}}(u)_t,\ e_{\text{zero}}\rangle = 0 \qquad \forall\, u \in \mathcal{K}\_\text{set},\ \forall\, 0 \leq t \leq T;$$

$Z_{T,e_{\text{zero}}}$ is signal-transparent along e_{sig} with respect to the control pair $(e_{\text{pos}}, e_{\text{zero}})$: for every $u \in \mathcal{K}\_\text{set}$, every $\tau \in \{0,\dots,T\}$, every scalar $a \in \mathbb{R}$, and every $0 \leq t \leq T$,

$$\langle Z_{T,e_{\text{zero}}}(u + a\,e_{\text{sig}}\mathbf{1}[\cdot=\tau])_t,\ e_{\text{pos}}\rangle = \langle Z_{T,e_{\text{zero}}}(u)_t,\ e_{\text{pos}}\rangle,$$

$$\langle Z_{T,e_{\text{zero}}}(u + a\,e_{\text{sig}}\mathbf{1}[\cdot=\tau])_t,\ e_{\text{zero}}\rangle = \langle Z_{T,e_{\text{zero}}}(u)_t,\ e_{\text{zero}}\rangle = 0,$$

and

$$\langle Z_{T,e_{\text{zero}}}(u + a\,e_{\text{sig}}\mathbf{1}[\cdot=\tau])_t,\ e_{\text{sig}}\rangle = \langle Z_{T,e_{\text{zero}}}(u)_t,\ e_{\text{sig}}\rangle + a\,\mathbf{1}[t=\tau].$$

Proof. Switch off the feedback branch.

Choose a positive constant c_1 such that

$$\text{GELU}(c_1) = 1.$$

Realize one forward value coordinate by the constant 1:

$$v_t^{(0)} \equiv 1.$$

Since every forward attention row sums to 1, the corresponding forward aggregate is

$$s_t^{(0)} = \sum_{j\leq t} \alpha_{t,j}^f \cdot 1 = 1 \qquad (0 \leq t \leq T).$$

Choose one gate coordinate equal to the prescribed channel:

$$g_t^{(0)} = \langle u_t, e_{\text{zero}}\rangle.$$

Choose the output projection so that this gated coordinate contributes

$$-\,e_{\text{zero}}$$

and all other output columns are zero. Then the residual update adds

$$-\,s_t^{(0)}\,g_t^{(0)}\,e_{\text{zero}} = -\,\langle u_t, e_{\text{zero}}\rangle e_{\text{zero}}.$$

Therefore

$$Z_{T,e_{\text{zero}}}(u)_t = u_t - \langle u_t, e_{\text{zero}}\rangle e_{\text{zero}}.$$

Taking the e_{zero}-coordinate gives

$$\langle Z_{T,e_{\text{zero}}}(u)_t, e_{\text{zero}}\rangle = \langle u_t, e_{\text{zero}}\rangle - \langle u_t, e_{\text{zero}}\rangle = 0,$$

which proves the exact zero-writing claim.

Because the update is supported only on the e_{zero}-direction, and

$$e_{\text{sig}},\ e_{\text{pos}},\ E_{\text{carry}} \perp e_{\text{zero}},$$

the e_{sig}-channel, the e_{pos}-channel, and all channels in E_{carry} are preserved exactly. This proves the exact preservation claim.

For signal-transparency, let

$$u^{(a,\tau)} := u + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau].$$

Since $e_{\text{sig}} \perp e_{\text{zero}}, e_{\text{pos}}$, one has

$$\langle u_t^{(a,\tau)}, e_{\text{zero}}\rangle = \langle u_t, e_{\text{zero}}\rangle, \qquad \langle u_t^{(a,\tau)}, e_{\text{pos}}\rangle = \langle u_t, e_{\text{pos}}\rangle.$$

Applying the explicit formula for $Z_{T,e_{\text{zero}}}$ yields

$$Z_{T,e_{\text{zero}}}(u^{(a,\tau)})_t = u_t + a\, e_{\text{sig}} \mathbf{1}[t = \tau] - \langle u_t, e_{\text{zero}}\rangle e_{\text{zero}} = Z_{T,e_{\text{zero}}}(u)_t + a\, e_{\text{sig}} \mathbf{1}[t = \tau].$$

Taking the e_{pos}-, e_{zero}-, and e_{sig}-coordinates gives the stated signal-transparency property. □

Lemma K.19 (Exact reset of finitely many scratch channels)**.** *Fix $T \geq 0$, orthonormal directions*

$$e_{\text{sig}},\ e_{z,1}, \dots, e_{z,p} \in \mathbb{R}^m,$$

and a subspace $E_{\text{keep}} \subset \mathbb{R}^m$ orthogonal to all of them. Then there exists a single forward-only concrete LN-free Sessa block

$$Z^{\text{scr}}_{T,\{e_{z,r}\}} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

such that $Z^{\text{scr}}_{T,\{e_{z,r}\}}$ preserves e_{sig} and every channel in E_{keep} exactly, and for every u and every t,

$$\langle Z^{\text{scr}}_{T,\{e_{z,r}\}}(u)_t, e_{z,r}\rangle = 0 \qquad (r = 1, \dots, p);$$

$Z^{\text{scr}}_{T,\{e_{z,r}\}}$ is signal-transparent along e_{sig} over E_{keep}.

Proof. Switch off the feedback branch and choose the forward queries and keys identically zero, so that every forward row has sum 1.

Choose a positive constant c_* with

$$\text{GELU}(c_*) = 1.$$

Activate one constant a-slot:

$$a_t^{(1)} \equiv c_*.$$

Then one post-GELU coordinate is identically 1. Choose W_V so that the first p forward value coordinates are all equal to 1. Since each forward row sums to 1, the corresponding forward aggregates satisfy

$$s_t^{(r)} = 1 \qquad (r = 1, \dots, p).$$

Choose the first p gate coordinates as

$$g_t^{(r)} = \langle u_t, e_{z,r} \rangle, \qquad r = 1, \dots, p,$$

and set all remaining gate coordinates to 0. Finally choose W^{out} so that the r-th active gated coordinate contributes $-e_{z,r}$, with all other output columns equal to 0. Then the residual update equals

$$-\sum_{r=1}^{p} \langle u_t, e_{z,r} \rangle e_{z,r},$$

so

$$Z^{\text{scr}}_{T,\{e_{z,r}\}}(u)_t = u_t - \sum_{r=1}^{p} \langle u_t, e_{z,r} \rangle e_{z,r}.$$

Hence each scratch channel is reset exactly to zero, while e_{sig} and every channel in E_{keep} are preserved exactly.

Now let

$$u^{(a,\tau)} := u + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau].$$

Because $e_{\text{sig}} \perp e_{z,r}$ for every r, the reset term is identical for $u^{(a,\tau)}$ and for u. Therefore

$$Z^{\text{scr}}_{T,\{e_{z,r}\}}(u^{(a,\tau)})_t = Z^{\text{scr}}_{T,\{e_{z,r}\}}(u)_t + a\, e_{\text{sig}} \mathbf{1}[t = \tau].$$

This is exactly signal-transparency along e_{sig} over E_{keep}. □

Lemma K.20 (Transparent damped predecessor integrator)**.** *Fix $\beta \in (0,1)$, set $\gamma := 1 - \beta$, and let*

$$L_H := \tau_{\max} + H.$$

Let $\mathcal{K}\_\text{set}_H \subset (\mathbb{R}^m)^{L_H+1}$ be compact. Assume orthonormal directions

$$e_{\text{sig}},\ e_{\text{pos}},\ e_{\text{tail}},\ e_{\text{prof}} \in \mathbb{R}^m$$

and a subspace $E_{\text{carry}} \subset \mathbb{R}^m$ orthogonal to all four, such that:

(i) the positional-control ranges

$$I_t := \{\langle u_t, e_{\text{pos}} \rangle :\ u \in \mathcal{K}\_\text{set}_H\}, \qquad 0 \le t \le L_H,$$

are compact and strictly ordered:

$$I_0 < I_1 < \cdots < I_{L_H} \subset (0, \infty);$$

(ii) the auxiliary tail input channel

$$g_t(u) := \langle u_t, e_{\text{tail}} \rangle$$

satisfies

$$c_g^-(t+1)^{-\beta} \le g_t(u) \le c_g^+(t+1)^{-\beta}, \qquad 0 \le t \le L_H, \quad u \in \mathcal{K}\_\text{set}_H;$$

(iii) the profile input channel is identically zero on $\mathcal{K}\_\text{set}_H$:

$$\langle u_t, e_{\text{prof}} \rangle = 0 \qquad \forall\, u \in \mathcal{K}\_\text{set}_H,\ \forall\, 0 \le t \le L_H.$$

Then there exists a single LN-free Sessa block

$$I_H : (\mathbb{R}^m)^{L_H+1} \to (\mathbb{R}^m)^{L_H+1}$$

such that the e_{sig}-channel, the e_{pos}-coordinate, the e_{tail}-channel, and every channel in E_{carry} are preserved exactly

and, writing

$$r_t(u) := \langle I_H(u)_t,\ e_{\mathrm{prof}} \rangle,$$

there exist constants $c_r^-, c_r^+ > 0$*, independent of* H*, such that*

$$c_r^-(t+1)^\gamma \le r_t(u) \le c_r^+(t+1)^\gamma, \qquad 0 \le t \le L_H, \quad u \in \mathcal{K}\_\mathrm{set}_H;$$

I_H *is signal-transparent along* e_{sig} *with respect to the control pair*

$$(e_{\mathrm{pos}}, e_{\mathrm{prof}}) :$$

for every $u \in \mathcal{K}\_\mathrm{set}_H$*, every* $\tau \in \{0, \dots, L_H\}$*, every scalar* $a \in \mathbb{R}$*, and every* $0 \le t \le L_H$*,*

$$\langle I_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t,\ e_{\mathrm{pos}} \rangle = \langle I_H(u)_t,\ e_{\mathrm{pos}} \rangle,$$

$$\langle I_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t,\ e_{\mathrm{prof}} \rangle = \langle I_H(u)_t,\ e_{\mathrm{prof}} \rangle,$$

$$\langle I_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t,\ e_{\mathrm{sig}} \rangle = \langle I_H(u)_t,\ e_{\mathrm{sig}} \rangle + a\, \mathbf{1}[t = \tau].$$

Proof. Fix a small constant

$$0 < \kappa_\mu \le 1$$

to be chosen later, and set

$$\lambda_H := 1 - \frac{1}{4(L_H + 1)} \in (0, 1), \qquad \mu_H := \kappa_\mu (L_H + 1)^{-3}.$$

Step 1: choose the attention patterns. Use Lemma K.1 on the positional-control coordinate e_{pos} with parameter μ_H. This yields strict-past feedback attention satisfying

$$\alpha^b_{t,t-1} \ge 1 - \mu_H, \qquad \sum_{j=0}^{t-2} \alpha^b_{t,j} \le \mu_H.$$

Use Lemma K.2 on the same positional-control coordinate, again with parameter μ_H, so that the forward row satisfies

$$\alpha^f_{t,t} \ge 1 - \mu_H, \qquad \sum_{j<t} \alpha^f_{t,j} \le \mu_H.$$

Both α^b and α^f depend only on the positional stream.

Step 2: feed the tail channel into the solve. Read the tail input channel exactly using Corollary K.5. Choose two a-slots

$$a_t^{(+)} = L\langle u_t, e_{\mathrm{tail}} \rangle, \qquad a_t^{(-)} = -L\langle u_t, e_{\mathrm{tail}} \rangle,$$

for any fixed $L > 0$, and choose one dedicated transport value coordinate

$$v_t^{\mathrm{tail}} = \frac{1}{L}(\bar{a}_t^{(+)} - \bar{a}_t^{(-)}) = \langle u_t, e_{\mathrm{tail}} \rangle = g_t(u).$$

Choose the feedback gain constant

$$\gamma_t \equiv \lambda_H.$$

Let $f_t(u)$ denote the forward signal entering the scalar solve on that dedicated coordinate:

$$f_t(u) = \sum_{j \le t} \alpha^f_{t,j}(u)\, g_j(u).$$

Let $s_t(u)$ be the corresponding solve output:

$$s_0(u) = f_0(u), \qquad s_t(u) = f_t(u) + \lambda_H \sum_{j<t} \alpha^b_{t,j}(u)\, s_j(u), \qquad t \geq 1.$$

Choose the gate on that dedicated coordinate to be the constant 1, and choose the output projection so that this solve output is written onto the prescribed profile direction e_{prof}, with all output columns on

$$e_{\text{sig}},\ e_{\text{pos}},\ e_{\text{tail}},\ E_{\text{carry}}$$

set to zero.

Because the input profile channel is identically zero on $\mathcal{K}\_\text{set}_H$, the residual formula gives

$$\langle I_H(u)_t, e_{\text{prof}} \rangle = \langle u_t, e_{\text{prof}} \rangle + s_t(u) = s_t(u).$$

Hence

$$r_t(u) := \langle I_H(u)_t, e_{\text{prof}} \rangle = s_t(u).$$

The e_{sig}-, e_{pos}-, e_{tail}-, and E_{carry}-channels are preserved exactly, because the output projection vanishes on those directions.

Step 3: compare with the ideal predecessor recursion. Define the ideal predecessor recursion

$$\tilde{r}_0(u) := g_0(u), \qquad \tilde{r}_t(u) := g_t(u) + \lambda_H\, \tilde{r}_{t-1}(u), \qquad t \geq 1,$$

so that

$$\tilde{r}_t(u) = \sum_{m=0}^{t} \lambda_H^{t-m} g_m(u).$$

Since $0 \leq m \leq t \leq L_H$ and $\lambda_H = 1 - \frac{1}{4(L_H+1)}$,

$$e^{-1/4} \leq \lambda_H^{t-m} \leq 1.$$

Therefore

$$e^{-1/4} \sum_{m=0}^{t} g_m(u) \leq \tilde{r}_t(u) \leq \sum_{m=0}^{t} g_m(u).$$

Using

$$c_g^-(m+1)^{-\beta} \leq g_m(u) \leq c_g^+(m+1)^{-\beta}$$

and

$$\sum_{m=0}^{t} (m+1)^{-\beta} \asymp (t+1)^{1-\beta} = (t+1)^{\gamma},$$

we obtain constants $\tilde{c}_r^-, \tilde{c}_r^+ > 0$, independent of H, such that

$$\tilde{c}_r^-(t+1)^{\gamma} \leq \tilde{r}_t(u) \leq \tilde{c}_r^+(t+1)^{\gamma}.$$

Step 4: control the perturbation error. Let $B_H(u)$ be the actual feedback matrix on the dedicated profile coordinate and B_H^* the ideal predecessor matrix

$$(B_H^*)_{t,t-1} = \lambda_H, \qquad (B_H^*)_{t,j} = 0 \ \ (j < t-1).$$

By the predecessor-focusing estimate,

$$\sup_t \sum_{j<t} |(B_H(u) - B_H^*)_{t,j}| \le C\,\mu_H$$

for an absolute constant C.

Also,

$$f_t(u) - g_t(u) = \sum_{j\le t} \alpha_{t,j}^f(u)(g_j(u) - g_t(u)) = \sum_{j<t} \alpha_{t,j}^f(u)(g_j(u) - g_t(u)),$$

hence

$$|f_t(u) - g_t(u)| \le 2c_g^+ \sum_{j<t} \alpha_{t,j}^f(u) \le 2c_g^+ \mu_H.$$

Therefore

$$\|f(u) - g(u)\|_\infty \le 2c_g^+ \mu_H.$$

Now

$$r(u) = (I - B_H(u))^{-1} f(u), \qquad \tilde{r}(u) = (I - B_H^*)^{-1} g(u),$$

so

$$r(u) - \tilde{r}(u) = (I - B_H(u))^{-1}\Big((f(u) - g(u)) + (B_H(u) - B_H^*)\tilde{r}(u)\Big).$$

Since the row sum of $B_H(u)$ is at most $\lambda_H < 1$,

$$\|(I - B_H(u))^{-1}\|_{\infty\to\infty} \le \frac{1}{1-\lambda_H} = 4(L_H + 1).$$

Also

$$\|\tilde{r}(u)\|_\infty \lesssim (L_H + 1)^\gamma.$$

Therefore there exists a constant $C_* > 0$, independent of H, such that

$$\|r(u) - \tilde{r}(u)\|_\infty \le C_*(L_H + 1)^{\gamma+1}\mu_H = C_*\kappa_\mu(L_H + 1)^{\gamma-2}.$$

Since $L_H = \tau_{\max} + H \ge \tau_{\max} + 1$, we have

$$(L_H + 1)^{\gamma-2} \le (\tau_{\max} + 2)^{\gamma-2}.$$

Choose $\kappa_\mu > 0$ so small that

$$C_*\kappa_\mu(\tau_{\max} + 2)^{\gamma-2} \le \frac{1}{2}\tilde{c}_r^-.$$

Then uniformly in H,

$$\|r(u) - \tilde{r}(u)\|_\infty \le \frac{1}{2}\tilde{c}_r^-.$$

Hence for every $0 \le t \le L_H$,

$$r_t(u) \ge \tilde{r}_t(u) - \frac{1}{2}\tilde{c}_r^- \ge \tilde{c}_r^-(t+1)^\gamma - \frac{1}{2}\tilde{c}_r^-.$$

Since $(t+1)^\gamma \ge 1$,

$$\tilde{c}_r^-(t+1)^\gamma - \frac{1}{2}\tilde{c}_r^- \ge \frac{1}{2}\tilde{c}_r^-(t+1)^\gamma.$$

So

$$r_t(u) \ge \frac{1}{2}\tilde{c}_r^-(t+1)^\gamma.$$

Similarly,

$$r_t(u) \le \tilde{r}_t(u) + \frac{1}{2}\tilde{c}_r^- \le \tilde{c}_r^+ (t+1)^\gamma + \frac{1}{2}\tilde{c}_r^-.$$

Again using $(t+1)^\gamma \ge 1$,

$$r_t(u) \le \Big(\tilde{c}_r^+ + \frac{1}{2}\tilde{c}_r^-\Big)(t+1)^\gamma.$$

Thus the stated two-sided profile bound holds with

$$c_r^- := \frac{1}{2}\tilde{c}_r^-, \qquad c_r^+ := \tilde{c}_r^+ + \frac{1}{2}\tilde{c}_r^-.$$

Step 5: verify signal-transparency. Let

$$u^{(a,\tau)} := u + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau].$$

Since $e_{\text{sig}} \perp e_{\text{pos}}, e_{\text{tail}}, e_{\text{prof}}$, one has

$$\langle u_t^{(a,\tau)}, e_{\text{pos}} \rangle = \langle u_t, e_{\text{pos}} \rangle, \qquad \langle u_t^{(a,\tau)}, e_{\text{tail}} \rangle = \langle u_t, e_{\text{tail}} \rangle, \qquad \langle u_t^{(a,\tau)}, e_{\text{prof}} \rangle = \langle u_t, e_{\text{prof}} \rangle = 0.$$

Therefore the feedback weights α^b are unchanged, since they depend only on the positional stream. The forward weights α^f are also unchanged for the same reason. Finally, the forward values g_t are unchanged, since they are exact reads of the tail channel. Hence the actual forward signal f_t, the actual feedback matrix B_H, and therefore the solve output r_t are all unchanged under perturbations along e_{sig}:

$$r_t(u^{(a,\tau)}) = r_t(u).$$

By construction, the output projection vanishes on the e_{sig}-channel, so that channel passes through exactly:

$$\langle I_H(u^{(a,\tau)})_t, e_{\text{sig}} \rangle = \langle I_H(u)_t, e_{\text{sig}} \rangle + a\, \mathbf{1}[t = \tau].$$

The e_{pos}-coordinate is preserved exactly as well. This proves the stated signal-transparency property. □

Corollary K.21 (Transparent power-profile block)**.** *Fix $\beta \in (0,1)$, set $\gamma := 1 - \beta$, fix $H \ge 1$, and let $L_H := \tau_{\max} + H$. Let $\mathcal{K}\_\text{set}_H \subset (\mathbb{R}^m)^{L_H+1}$ be the compact input set under consideration.*

Assume $\mathcal{K}\_\text{set}_H$ carries orthonormal directions

$$e_{\text{sig}},\ e_{\text{pos}} \in \mathbb{R}^m$$

such that:

(i) the original signal channel is

$$u \mapsto \langle u_t, e_{\text{sig}} \rangle;$$

(ii) the positional-control coordinate is

$$u \mapsto \langle u_t, e_{\text{pos}} \rangle,$$

with ordered positive ranges

$$I_0 < I_1 < \cdots < I_{L_H} \subset (0, \infty).$$

Fix additional orthonormal directions

$$e_{\text{prof}},\ e_{\text{tail}},\ e_{\text{aux}},\ e_{\text{src}},\ e_{\text{tgt}} \in \mathbb{R}^m$$

orthogonal to both e_{sig} and e_{pos}.

Then there exists a constant-depth LN-free Sessa network

$$Q_H : (\mathbb{R}^m)^{L_H+1} \to (\mathbb{R}^m)^{L_H+1}$$

such that the original signal channel is preserved exactly:

$$\langle Q_H(u)_t, e_{\mathrm{sig}} \rangle = \langle u_t, e_{\mathrm{sig}} \rangle \qquad (0 \le t \le L_H,\ u \in \mathcal{K}\_\mathrm{set}_H);$$

the positional-control coordinate is preserved exactly:

$$\langle Q_H(u)_t, e_{\mathrm{pos}} \rangle = \langle u_t, e_{\mathrm{pos}} \rangle \qquad (0 \le t \le L_H,\ u \in \mathcal{K}\_\mathrm{set}_H);$$

the profile channel on the prescribed direction e_{prof} satisfies the uniform two-sided bound

$$c_r^-(t+1)^\gamma \le \langle Q_H(u)_t, e_{\mathrm{prof}} \rangle \le c_r^+(t+1)^\gamma, \qquad 0 \le t \le L_H,\ u \in \mathcal{K}\_\mathrm{set}_H,$$

with constants independent of H; and Q_H is signal-transparent along e_{sig} with respect to the control pair $(e_{\mathrm{pos}}, e_{\mathrm{prof}})$: for every $u \in \mathcal{K}\_\mathrm{set}_H$, every $\tau \in \{0, \dots, L_H\}$, and every scalar $a \in \mathbb{R}$,

$$\langle Q_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t,\ e_{\mathrm{pos}} \rangle = \langle Q_H(u)_t,\ e_{\mathrm{pos}} \rangle, \qquad 0 \le t \le L_H,$$

$$\langle Q_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t,\ e_{\mathrm{prof}} \rangle = \langle Q_H(u)_t,\ e_{\mathrm{prof}} \rangle, \qquad 0 \le t \le L_H,$$

and

$$\langle Q_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t,\ e_{\mathrm{sig}} \rangle = \langle Q_H(u)_t,\ e_{\mathrm{sig}} \rangle + a\, \mathbf{1}[t = \tau], \qquad 0 \le t \le L_H.$$

Proof. The auxiliary orthonormal directions

$$e_{\mathrm{prof}},\ e_{\mathrm{tail}},\ e_{\mathrm{aux}},\ e_{\mathrm{src}},\ e_{\mathrm{tgt}}$$

are fixed by hypothesis and are orthogonal to both e_{sig} and e_{pos}.

Step 1: clear the profile channel. Apply Lemma K.18 with

$$e_{\mathrm{zero}} := e_{\mathrm{prof}}, \qquad E_{\mathrm{carry}} := \{0\}.$$

This yields a forward-only block

$$Z_H^{\mathrm{prof}} : (\mathbb{R}^m)^{L_H+1} \to (\mathbb{R}^m)^{L_H+1}$$

such that

$$\langle Z_H^{\mathrm{prof}}(u)_t, e_{\mathrm{sig}} \rangle = \langle u_t, e_{\mathrm{sig}} \rangle, \qquad \langle Z_H^{\mathrm{prof}}(u)_t, e_{\mathrm{pos}} \rangle = \langle u_t, e_{\mathrm{pos}} \rangle, \qquad \langle Z_H^{\mathrm{prof}}(u)_t, e_{\mathrm{prof}} \rangle = 0.$$

Moreover, Z_H^{prof} is signal-transparent along e_{sig} with respect to $(e_{\mathrm{pos}}, e_{\mathrm{prof}})$.

Let

$$\mathcal{K}\_\mathrm{set}_H^{(0)} := Z_H^{\mathrm{prof}}(\mathcal{K}\_\mathrm{set}_H).$$

Step 2: build the tail channel. Apply Lemma K.17 to $\mathcal{K}\_\mathrm{set}_H^{(0)}$, with

$$E_{\mathrm{carry}} := \mathrm{span}\{e_{\mathrm{prof}}\}.$$

This yields a constant-depth network

$$T_H^{\mathrm{tail}} : (\mathbb{R}^m)^{L_H+1} \to (\mathbb{R}^m)^{L_H+1}$$

such that

$$\langle T_H^{\mathrm{tail}}(v)_t, e_{\mathrm{sig}}\rangle = \langle v_t, e_{\mathrm{sig}}\rangle, \qquad \langle T_H^{\mathrm{tail}}(v)_t, e_{\mathrm{pos}}\rangle = \langle v_t, e_{\mathrm{pos}}\rangle, \qquad \langle T_H^{\mathrm{tail}}(v)_t, e_{\mathrm{prof}}\rangle = \langle v_t, e_{\mathrm{prof}}\rangle,$$

and the tail channel

$$g_t(v) := \langle T_H^{\mathrm{tail}}(v)_t, e_{\mathrm{tail}}\rangle$$

satisfies

$$c_g^-(t+1)^{-\beta} \le g_t(v) \le c_g^+(t+1)^{-\beta}.$$

Because the carried profile channel is identically zero on $\mathcal{K}\_\mathrm{set}_H^{(0)}$ and is preserved exactly by T_H^{tail}, one still has

$$\langle T_H^{\mathrm{tail}}(v)_t, e_{\mathrm{prof}}\rangle = 0 \qquad \forall\, v \in \mathcal{K}\_\mathrm{set}_H^{(0)}.$$

Let

$$\mathcal{K}\_\mathrm{set}_H^{(1)} := T_H^{\mathrm{tail}}(\mathcal{K}\_\mathrm{set}_H^{(0)}).$$

Step 3: clear the scratch channels. Apply Lemma K.19 to the scratch directions

$$e_{\mathrm{aux}},\ e_{\mathrm{src}},\ e_{\mathrm{tgt}},$$

with

$$E_{\mathrm{keep}} := \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{tail}}, e_{\mathrm{prof}}\}.$$

This yields a forward-only concrete block

$$Z_H^{\mathrm{scr}} : (\mathbb{R}^m)^{L_H+1} \to (\mathbb{R}^m)^{L_H+1}$$

such that it preserves

$$e_{\mathrm{sig}},\ e_{\mathrm{pos}},\ e_{\mathrm{tail}},\ e_{\mathrm{prof}}$$

exactly and writes

$$\langle Z_H^{\mathrm{scr}}(w)_t, e_{\mathrm{aux}}\rangle = \langle Z_H^{\mathrm{scr}}(w)_t, e_{\mathrm{src}}\rangle = \langle Z_H^{\mathrm{scr}}(w)_t, e_{\mathrm{tgt}}\rangle = 0.$$

Since Z_H^{scr} preserves the tail channel exactly, the same bounds

$$c_g^-(t+1)^{-\beta} \le \langle Z_H^{\mathrm{scr}}(w)_t, e_{\mathrm{tail}}\rangle \le c_g^+(t+1)^{-\beta}$$

hold on the image.

Let

$$\widetilde{\mathcal{K}\_\mathrm{set}}_H := Z_H^{\mathrm{scr}}(\mathcal{K}\_\mathrm{set}_H^{(1)}).$$

On $\widetilde{\mathcal{K}\_\mathrm{set}}_H$ we therefore retain the same ordered positional ranges as on $\mathcal{K}\_\mathrm{set}_H$, the same tail bounds $c_g^{\pm}(t+1)^{-\beta}$, an identically zero profile channel, and identically zero scratch channels $e_{\mathrm{aux}}, e_{\mathrm{src}}, e_{\mathrm{tgt}}$.

Step 4: integrate the tail channel. Apply Lemma K.20 to $\widetilde{\mathcal{K}\_\mathrm{set}}_H$, with

$$E_{\mathrm{carry}} := \mathrm{span}\{e_{\mathrm{aux}}, e_{\mathrm{src}}, e_{\mathrm{tgt}}\}.$$

Because these carried channels are already identically zero on $\widetilde{\mathcal{K}\_\mathrm{set}}_H$, this application is fully legitimate and keeps them zero. We obtain a single LN-free Sessa block

$$I_H : (\mathbb{R}^m)^{L_H+1} \to (\mathbb{R}^m)^{L_H+1}$$

such that

$$\langle I_H(w)_t, e_{\text{sig}}\rangle = \langle w_t, e_{\text{sig}}\rangle, \qquad \langle I_H(w)_t, e_{\text{pos}}\rangle = \langle w_t, e_{\text{pos}}\rangle, \qquad \langle I_H(w)_t, e_{\text{tail}}\rangle = \langle w_t, e_{\text{tail}}\rangle,$$

and

$$c_r^-(t+1)^\gamma \le \langle I_H(w)_t, e_{\text{prof}}\rangle \le c_r^+(t+1)^\gamma.$$

Step 5: define the preparatory network. Set

$$Q_H := I_H \circ Z_H^{\text{scr}} \circ T_H^{\text{tail}} \circ Z_H^{\text{prof}}.$$

The exact preservation and two-sided profile bounds follow immediately from the four stages above.

Step 6: verify signal-transparency. Fix $u \in \mathcal{K}\_\text{set}_H$, $\tau \in \{0, \dots, L_H\}$, and $a \in \mathbb{R}$. Define

$$u^{(a,\tau)} := u + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau].$$

By signal-transparency of Z_H^{prof},

$$Z_H^{\text{prof}}(u^{(a,\tau)}) = Z_H^{\text{prof}}(u) + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau]$$

on the signal channel, while the e_{pos}- and e_{prof}-channels are unchanged.

Applying signal-transparency of T_H^{tail} then gives

$$T_H^{\text{tail}}(Z_H^{\text{prof}}(u^{(a,\tau)})) = T_H^{\text{tail}}(Z_H^{\text{prof}}(u)) + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau]$$

on the signal channel, while the e_{pos}- and e_{tail}-channels are unchanged and the e_{prof}-channel remains zero.

Now Z_H^{scr} preserves $e_{\text{sig}}, e_{\text{pos}}, e_{\text{tail}}, e_{\text{prof}}$ exactly, so

$$Z_H^{\text{scr}}(T_H^{\text{tail}}(Z_H^{\text{prof}}(u^{(a,\tau)}))) = Z_H^{\text{scr}}(T_H^{\text{tail}}(Z_H^{\text{prof}}(u))) + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau]$$

on the signal channel, and the e_{pos}-, e_{tail}-, and e_{prof}-channels are unchanged.

Thus the two inputs fed into I_H differ only on the e_{sig}-channel and have the same e_{pos}-, e_{tail}-, and e_{prof}-streams. In the concrete construction of Lemma K.20, the feedback weights α^b and forward weights α^f depend only on the positional stream, while the forward values g_t are exact reads of the e_{tail}-channel. Hence the forward signals f_t, the feedback matrices B_H, and the solve outputs r_t are identical for the two inputs. Moreover, the output projection of I_H vanishes on the e_{sig}-, e_{pos}-, and e_{tail}-channels, so the e_{sig}-channel passes through exactly and the e_{pos}-coordinate is unchanged. Therefore

$$\langle Q_H(u^{(a,\tau)})_t, e_{\text{pos}}\rangle = \langle Q_H(u)_t, e_{\text{pos}}\rangle,$$

$$\langle Q_H(u^{(a,\tau)})_t, e_{\text{prof}}\rangle = \langle Q_H(u)_t, e_{\text{prof}}\rangle,$$

$$\langle Q_H(u^{(a,\tau)})_t, e_{\text{sig}}\rangle = \langle Q_H(u)_t, e_{\text{sig}}\rangle + a\, \mathbf{1}[t = \tau].$$

This proves the stated signal-transparency property. □

Lemma K.22 (Profile-compensated macro-layer)**.** *Fix $\beta \in (0,1)$, set $\gamma := 1 - \beta$, and fix $T \ge 0$. Let $\mathcal{K}\_\text{set} \subset (\mathbb{R}^m)^{T+1}$ be compact. Assume orthonormal directions*

$$e_{\text{sig}},\ e_{\text{pos}},\ e_{\text{prof}},\ e_{\text{src}} \in \mathbb{R}^m$$

and a subspace $E_{\text{carry}} \subset \mathbb{R}^m$ orthogonal to all four, such that:

(i) the positional-control ranges

$$I_t := \{\langle u_t, e_{\mathrm{pos}}\rangle : \ u \in \mathcal{K}\_\mathrm{set}\}, \qquad 0 \le t \le T,$$

are compact and strictly ordered:

$$I_0 < I_1 < \cdots < I_T \subset (0, \infty);$$

(ii) the profile channel

$$r_t(u) := \langle u_t, e_{\mathrm{prof}}\rangle$$

satisfies

$$c_r^-(t+1)^\gamma \le r_t(u) \le c_r^+(t+1)^\gamma, \qquad 0 \le t \le T, \quad u \in \mathcal{K}\_\mathrm{set}.$$

Then there exists a constant-depth LN-free Sessa macro-layer

$$M_T : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

such that the e_{pos}-channel, the e_{prof}-channel, and every channel in E_{carry} are preserved exactly, and M_T has signal-blind exact scalar transport along e_{sig} over

$$E_{\mathrm{ctrl}} := \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{prof}}\} \oplus E_{\mathrm{carry}},$$

with kernel

$$\mathcal{T}^u_{M_T}(i,j) = D^u_{\mathrm{mac}}(i)\,\mathbf{1}[i=j] + K^u_{\mathrm{mac}}(i,j)\,\mathbf{1}[j<i];$$

There exist constants

$$1 \le d^-_{\mathrm{mac}} \le d^+_{\mathrm{mac}} < \infty, \qquad 0 < a^-_{\mathrm{mac}} \le a^+_{\mathrm{mac}} < \infty,$$

depending only on (β, c_r^-, c_r^+), but independent of T, such that

$$d^-_{\mathrm{mac}} \le D^u_{\mathrm{mac}}(i) \le d^+_{\mathrm{mac}}, \qquad 0 \le i \le T,$$

and

$$a^-_{\mathrm{mac}}(i+1)^{-\beta} \le K^u_{\mathrm{mac}}(i,j) \le a^+_{\mathrm{mac}}(i+1)^{-\beta}, \qquad 0 \le j < i \le T.$$

In particular,

$$K^u_{\mathrm{mac}}(i,j) \le a^+_{\mathrm{mac}}(i-j+1)^{-\beta}.$$

Consequently,

$$e_{\mathrm{sig}}^\top \frac{\partial M_T(u)_i}{\partial u_j} e_{\mathrm{sig}} = D^u_{\mathrm{mac}}(i)\,\mathbf{1}[i=j] + K^u_{\mathrm{mac}}(i,j)\,\mathbf{1}[j<i].$$

Proof. Write

$$x_t := \langle u_t, e_{\mathrm{sig}}\rangle, \qquad r_t(u) := \langle u_t, e_{\mathrm{prof}}\rangle, \qquad 0 \le t \le T.$$

We construct

$$M_T = A_T^{\mathrm{diff}} \circ W_T^{\mathrm{src}},$$

where W_T^{src} is a local source writer and A_T^{diff} is the diffuse transport-bearing block.

Step 1: local source writer. Choose a parameter $\mu \in (0, \frac{1}{2}]$ and apply Lemma K.2 to the ordered positional-control coordinate e_{pos}. This yields a forward attention row satisfying

$$\alpha^f_{t,t} \ge 1 - \mu, \qquad \sum_{j<t} \alpha^f_{t,j} \le \mu, \qquad 0 \le t \le T.$$

We now build a forward-only LN-free Sessa block

$$W_T^{\mathrm{src}} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}.$$

Choose one forward value coordinate equal to 1:

$$v_t^{(0)} \equiv 1.$$

Hence

$$s_t^{(0)} = \sum_{j \le t} \alpha_{t,j}^f \cdot 1 = 1.$$

Next read the profile channel exactly using Corollary K.5. Choose two a-slots

$$a_t^{(+)} = L\langle u_t, e_{\mathrm{prof}}\rangle, \qquad a_t^{(-)} = -L\langle u_t, e_{\mathrm{prof}}\rangle$$

for any fixed $L > 0$, and choose the value projection so that

$$v_t^{(1)} = \frac{1}{L}(\bar{a}_t^{(+)} - \bar{a}_t^{(-)}) = \langle u_t, e_{\mathrm{prof}}\rangle = r_t(u).$$

Let

$$m_t^u := s_t^{(1)} := \sum_{j \le t} \alpha_{t,j}^f \, r_j(u).$$

Choose two gate coordinates

$$g_t^{(0)} = \langle u_t, e_{\mathrm{src}}\rangle, \qquad g_t^{(1)} = \langle u_t, e_{\mathrm{sig}}\rangle = x_t,$$

and choose the output projection on the e_{src}-channel with coefficients $(-1, +1)$. Then

$$\langle W_T^{\mathrm{src}}(u)_t, e_{\mathrm{src}}\rangle = \langle u_t, e_{\mathrm{src}}\rangle - s_t^{(0)}\langle u_t, e_{\mathrm{src}}\rangle + s_t^{(1)} x_t = m_t^u \, x_t.$$

All other output columns are zero, so the e_{sig}-, e_{pos}-, e_{prof}-, and E_{carry}-channels are preserved exactly.

It remains to bound m_t^u. Since every $r_j(u) \ge 0$,

$$m_t^u \ge \alpha_{t,t}^f \, r_t(u) \ge (1-\mu)c_r^-(t+1)^\gamma.$$

Also, for every $j \le t$,

$$r_j(u) \le c_r^+(j+1)^\gamma \le c_r^+(t+1)^\gamma,$$

so

$$m_t^u = \sum_{j \le t} \alpha_{t,j}^f r_j(u) \le c_r^+(t+1)^\gamma.$$

Therefore

$$m^-(t+1)^\gamma \le m_t^u \le m^+(t+1)^\gamma, \qquad m^- := (1-\mu)c_r^-, \qquad m^+ := c_r^+.$$

Step 2: diffuse transport block. Let

$$w := W_T^{\mathrm{src}}(u).$$

We now build a single LN-free Sessa block

$$A_T^{\mathrm{diff}} : (\mathbb{R}^m)^{T+1} \to (\mathbb{R}^m)^{T+1}$$

as follows.

Forward branch. Choose all forward queries and keys equal to zero:

$$q_k^f \equiv 0, \qquad k_j^f \equiv 0.$$

Hence the forward row is exactly uniform on the visible prefix:

$$\alpha_{k,j}^f = \frac{1}{k+1}\mathbf{1}[j \le k].$$

Read the source scratch channel exactly using Corollary K.5. Choose two a-slots

$$a_j^{(+)} = L\langle w_j, e_{\mathrm{src}}\rangle, \qquad a_j^{(-)} = -L\langle w_j, e_{\mathrm{src}}\rangle,$$

and choose the value projection so that

$$v_j^{\mathrm{src}} = \frac{1}{L}(\bar{a}_j^{(+)} - \bar{a}_j^{(-)}) = \langle w_j, e_{\mathrm{src}}\rangle = m_j^u\, x_j.$$

Thus the forward signal is

$$f_k = \sum_{j\le k} \alpha_{k,j}^f v_j^{\mathrm{src}} = \frac{1}{k+1}\sum_{j=0}^{k} m_j^u x_j.$$

Feedback branch. Choose all feedback queries and keys equal to zero and the feedback gain constant:

$$q_i^b \equiv 0, \qquad k_j^b \equiv 0, \qquad \gamma_i \equiv \gamma = 1-\beta.$$

Therefore the strict-past feedback row is exactly uniform:

$$\alpha_{i,k}^b = \frac{1}{i}\mathbf{1}[k < i], \qquad 1 \le i \le T,$$

and the scalar feedback matrix is

$$B_{i,k} = \frac{\gamma}{i}\mathbf{1}[k < i].$$

Let

$$\Theta_{i,k} := [(I-B)^{-1}]_{i,k}, \qquad 0 \le k \le i \le T.$$

Exactly as in the proof of Lemma K.15, one has

$$\Theta_{i,i} = 1,$$

and for $k < i$,

$$\Theta_{i,k} = \gamma\, \frac{\Gamma(k+1)}{\Gamma(k+1+\gamma)}\frac{\Gamma(i+\gamma)}{\Gamma(i+1)}.$$

Hence there exist constants

$$0 < c_\Theta^- \le c_\Theta^+ < \infty$$

depending only on β, such that

$$c_\Theta^-(k+1)^{-\gamma}(i+1)^{-\beta} \le \Theta_{i,k} \le c_\Theta^+(k+1)^{-\gamma}(i+1)^{-\beta}, \qquad 0 \le k < i \le T.$$

Write transport into the signal channel. Choose one gate coordinate identically 1, and choose the output projection so that the solve output adds $+s_i$ to the e_{sig}-channel and all output columns on

$$e_{\mathrm{pos}},\ e_{\mathrm{prof}},\ E_{\mathrm{carry}}$$

vanish.

Therefore

$$\langle A_T^{\mathrm{diff}}(w)_i, e_{\mathrm{sig}}\rangle = \langle w_i, e_{\mathrm{sig}}\rangle + s_i = x_i + s_i,$$

where

$$s_i = \sum_{k=0}^{i} \Theta_{i,k} f_k.$$

Since W_T^{src} preserves $e_{\mathrm{sig}}, e_{\mathrm{pos}}, e_{\mathrm{prof}}, E_{\mathrm{carry}}$ exactly, the full macro-layer $M_T = A_T^{\mathrm{diff}} \circ W_T^{\mathrm{src}}$ also preserves $e_{\mathrm{pos}}, e_{\mathrm{prof}}, E_{\mathrm{carry}}$ exactly.

Step 3: exact transport formula. Substituting the expression for f_k, we get

$$s_i = \sum_{k=0}^{i} \Theta_{i,k} \frac{1}{k+1} \sum_{j=0}^{k} m_j^u x_j = \sum_{j=0}^{i} \left(m_j^u \sum_{k=j}^{i} \frac{\Theta_{i,k}}{k+1} \right) x_j.$$

Define

$$L(i,j) := \sum_{k=j}^{i} \frac{\Theta_{i,k}}{k+1}, \qquad 0 \le j \le i \le T.$$

Then

$$\langle M_T(u)_i, e_{\mathrm{sig}}\rangle = x_i + \sum_{j=0}^{i} m_j^u \, L(i,j) \, x_j.$$

Since $\Theta_{i,i} = 1$, we have

$$L(i,i) = \frac{1}{i+1}.$$

Therefore

$$\langle M_T(u)_i, e_{\mathrm{sig}}\rangle = \left(1 + \frac{m_i^u}{i+1}\right) x_i + \sum_{j<i} m_j^u \, L(i,j) \, x_j.$$

Define

$$D_{\mathrm{mac}}^u(i) := 1 + \frac{m_i^u}{i+1}, \qquad K_{\mathrm{mac}}^u(i,j) := m_j^u \, L(i,j) \quad (j < i).$$

This yields exact scalar transport on the signal channel:

$$\langle M_T(u)_i, e_{\mathrm{sig}}\rangle = D_{\mathrm{mac}}^u(i) \, x_i + \sum_{j<i} K_{\mathrm{mac}}^u(i,j) \, x_j.$$

The coefficient m_j^u depends only on the e_{pos}- and e_{prof}-control streams, because the source writer uses positional self-focusing and an exact read of the profile channel only. The kernel $L(i,j)$ depends only on the fixed diffuse transport block. Hence $D_{\mathrm{mac}}^u(i)$ and $K_{\mathrm{mac}}^u(i,j)$ depend only on the control stream

$$(\Pi_{\mathrm{ctrl}} u_t)_{t=0}^{T}, \qquad E_{\mathrm{ctrl}} := \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{prof}}\} \oplus E_{\mathrm{carry}}.$$

Thus M_T has signal-blind exact scalar transport over E_{ctrl}.

Step 4: diagonal bounds. Since

$$m^-(i+1)^\gamma \le m_i^u \le m^+(i+1)^\gamma,$$

we obtain

$$1 \le D^u_{\mathrm{mac}}(i) = 1 + \frac{m^u_i}{i+1} \le 1 + m^+(i+1)^{\gamma-1} = 1 + m^+(i+1)^{-\beta} \le 1 + m^+.$$

Hence we may take

$$d^-_{\mathrm{mac}} := 1, \qquad d^+_{\mathrm{mac}} := 1 + m^+.$$

Step 5: off-diagonal upper bound. Fix $0 \le j < i \le T$. Using $\Theta_{i,i} = 1$ and the upper bound on $\Theta_{i,k}$ for $k < i$,

$$L(i,j) \le \frac{1}{i+1} + c^+_\Theta (i+1)^{-\beta} \sum_{k=j}^{i-1} (k+1)^{-1-\gamma}.$$

Since

$$\frac{1}{i+1} \le (j+1)^{-\gamma}(i+1)^{-\beta},$$

and

$$\sum_{k=j}^{i-1} (k+1)^{-1-\gamma} \le \sum_{k=j}^{\infty} (k+1)^{-1-\gamma} \lesssim_\gamma (j+1)^{-\gamma},$$

there exists $C^+_L < \infty$, depending only on β, such that

$$L(i,j) \le C^+_L (j+1)^{-\gamma}(i+1)^{-\beta}.$$

Therefore

$$K^u_{\mathrm{mac}}(i,j) = m^u_j \, L(i,j) \le m^+(j+1)^{\gamma} \cdot C^+_L (j+1)^{-\gamma}(i+1)^{-\beta}.$$

Hence

$$K^u_{\mathrm{mac}}(i,j) \le a^+_{\mathrm{mac}} (i+1)^{-\beta}, \qquad a^+_{\mathrm{mac}} := m^+ C^+_L.$$

Step 6: off-diagonal lower bound. Fix $0 \le j < i \le T$.

Case 0: $j = 0$. Since $\Theta_{i,0}$ appears in the sum defining $L(i,0)$, we have

$$L(i,0) \ge \Theta_{i,0}.$$

By the resolvent bound,

$$\Theta_{i,0} \ge c^-_\Theta (0+1)^{-\gamma}(i+1)^{-\beta} = c^-_\Theta (i+1)^{-\beta}.$$

Also $m^u_0 \ge m^-$. Therefore

$$K^u_{\mathrm{mac}}(i,0) = m^u_0 \, L(i,0) \ge m^- c^-_\Theta (i+1)^{-\beta}.$$

Case 1: $1 \le j \le i/2$. Then $2j \le i$, so

$$L(i,j) \ge \sum_{k=j}^{2j-1} \frac{\Theta_{i,k}}{k+1} \ge c^-_\Theta (i+1)^{-\beta} \sum_{k=j}^{2j-1} (k+1)^{-1-\gamma}.$$

Since the sum over one dyadic block is comparable to $(j+1)^{-\gamma}$, there exists $c^{(1)}_L > 0$, depending only on β, such that

$$L(i,j) \ge c^{(1)}_L (j+1)^{-\gamma}(i+1)^{-\beta}.$$

Hence

$$K^u_{\mathrm{mac}}(i,j) = m^u_j \, L(i,j) \ge m^-(j+1)^{\gamma} \cdot c^{(1)}_L (j+1)^{-\gamma}(i+1)^{-\beta} = m^- c^{(1)}_L (i+1)^{-\beta}.$$

Case 2: $j > i/2$. Then

$$L(i,j) \ge \frac{1}{i+1},$$

so

$$K^u_{\mathrm{mac}}(i,j) = m^u_j\, L(i,j) \geq \frac{m^u_j}{i+1} \geq \frac{m^-(j+1)^\gamma}{i+1}.$$

Since $j+1 > \frac{i+1}{2}$,

$$(j+1)^\gamma \geq 2^{-\gamma}(i+1)^\gamma.$$

Therefore

$$K^u_{\mathrm{mac}}(i,j) \geq m^- 2^{-\gamma}(i+1)^{\gamma-1} = m^- 2^{-\gamma}(i+1)^{-\beta}.$$

Combining the three cases gives

$$K^u_{\mathrm{mac}}(i,j) \geq a^-_{\mathrm{mac}}(i+1)^{-\beta}, \qquad a^-_{\mathrm{mac}} := \min\{m^- c^-_\Theta,\ m^- c^{(1)}_L,\ m^- 2^{-\gamma}\}.$$

For any $\eta > 0$, replacing $\mathcal{K}\_\mathrm{set}$ by $\mathrm{Sat}^{\mathrm{sig}}_\eta(\mathcal{K}\_\mathrm{set})$ leaves the ordered positional ranges and the two-sided profile bounds unchanged, since only the e_{sig}-channel is perturbed. The same source-writer plus diffuse-transport construction therefore yields the same exact scalar transport formula on $\mathrm{Sat}^{\mathrm{sig}}_\eta(\mathcal{K}\_\mathrm{set})$, with the same coefficients $D^u_{\mathrm{mac}}(i)$ and $K^u_{\mathrm{mac}}(i,j)$, because these coefficients depend only on the control stream $(e_{\mathrm{pos}}, e_{\mathrm{prof}}, E_{\mathrm{carry}})$. Applying Lemma K.8(i) gives

$$e_{\mathrm{sig}}^\top \frac{\partial M_T(u)_i}{\partial u_j} e_{\mathrm{sig}} = D^u_{\mathrm{mac}}(i)\,\mathbf{1}[i=j] + K^u_{\mathrm{mac}}(i,j)\,\mathbf{1}[j<i].$$

□

Corollary K.23 (Macro-layer transport)**.** *Under the hypotheses of Lemma K.22, let*

$$E_{\mathrm{ctrl}} := \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{prof}}\} \oplus E_{\mathrm{carry}}, \qquad \Pi_{\mathrm{ctrl}} : \mathbb{R}^m \to E_{\mathrm{ctrl}}, \qquad \pi_{\mathrm{sig}}(v) := \langle v, e_{\mathrm{sig}}\rangle,$$

and let M_T be the concrete macro-layer constructed there. Then for every $\delta \geq 0$, M_T has signal-blind exact scalar transport along e_{sig} over E_{ctrl} on $\mathrm{Sat}^{\mathrm{sig}}_\delta(\mathcal{K}\_\mathrm{set})$, with the same scalar transport kernel $\mathcal{T}^u_{M_T}(i,j)$ as on $\mathcal{K}\_\mathrm{set}$.

More precisely, if

$$v = u + \sum_{t=0}^{T} a_t e_{\mathrm{sig}} \mathbf{1}[\cdot = t], \qquad u \in \mathcal{K}\_\mathrm{set},$$

then

$$\Pi_{\mathrm{ctrl}} M_T(v)_i = \Pi_{\mathrm{ctrl}} v_i, \qquad 0 \leq i \leq T,$$

and

$$\pi_{\mathrm{sig}}(M_T(v)_i) = \sum_{j=0}^{i} \mathcal{T}^u_{M_T}(i,j)\, \pi_{\mathrm{sig}}(v_j), \qquad 0 \leq i \leq T.$$

The right-hand side depends only on the control stream of v, hence is independent of the choice of $u \in \mathcal{K}\_\mathrm{set}$ with the same control stream.

Proof. Write

$$M_T = A^{\mathrm{diff}}_T \circ W^{\mathrm{src}}_T$$

exactly as in the proof of Lemma K.22.

Fix

$$v = u + \sum_{t=0}^{T} a_t e_{\mathrm{sig}} \mathbf{1}[\cdot = t], \qquad u \in \mathcal{K}\_\mathrm{set}.$$

Since v differs from u only on the e_{sig}-channel, the e_{pos}-, e_{prof}-, and E_{carry}-streams are unchanged. Hence the

self-focused profile averages from the source-writer stage are unchanged:

$$m_t^v = m_t^u, \qquad 0 \le t \le T.$$

Therefore the explicit source-writer formula gives

$$\langle W_T^{\mathrm{src}}(v)_t, e_{\mathrm{src}} \rangle = m_t^u \, \pi_{\mathrm{sig}}(v_t), \qquad 0 \le t \le T.$$

Moreover, W_T^{src} preserves the channels in E_{ctrl} exactly, because it modifies only the e_{src}-channel.

In the diffuse stage, the forward row is the exact uniform prefix average, so the forward signal entering the fixed feedback solve is

$$f_k(v) = \frac{1}{k+1} \sum_{j=0}^{k} m_j^u \, \pi_{\mathrm{sig}}(v_j), \qquad 0 \le k \le T.$$

The feedback matrix B, its resolvent Θ, and the kernel

$$L(i,j) := \sum_{k=j}^{i} \frac{\Theta_{i,k}}{k+1}$$

depend only on β, hence are independent of v. Thus the solve output satisfies

$$s_i(v) = \sum_{k=0}^{i} \Theta_{i,k} f_k(v) = \sum_{j=0}^{i} m_j^u \, L(i,j) \, \pi_{\mathrm{sig}}(v_j).$$

Using the definitions from Lemma K.22,

$$D_{\mathrm{mac}}^u(i) := 1 + \frac{m_i^u}{i+1}, \qquad K_{\mathrm{mac}}^u(i,j) := m_j^u \, L(i,j) \quad (j < i),$$

we obtain

$$\pi_{\mathrm{sig}}(M_T(v)_i) = \pi_{\mathrm{sig}}(v_i) + s_i(v) = \sum_{j=0}^{i} \mathcal{T}_{M_T}^u(i,j) \, \pi_{\mathrm{sig}}(v_j).$$

Finally, A_T^{diff} modifies only the e_{sig}-channel and preserves $e_{\mathrm{pos}}, e_{\mathrm{prof}}, E_{\mathrm{carry}}$ exactly. Hence M_T preserves E_{ctrl} exactly on $\mathrm{Sat}_\delta^{\mathrm{sig}}(\mathcal{K}\_\mathrm{set})$. Since the coefficients m_j^u, and therefore $\mathcal{T}_{M_T}^u(i,j)$, depend only on the control stream, the displayed kernel is independent of the choice of $u \in \mathcal{K}\_\mathrm{set}$ with the same control stream. This proves the claim. □

Lemma K.24 (Projected macro-layer)**.** *Under the hypotheses of Lemma K.22, let*

$$\Pi_{\mathrm{src}}(v)_t := v_t - \langle v_t, e_{\mathrm{src}} \rangle e_{\mathrm{src}}, \qquad 0 \le t \le T,$$

be the tokenwise orthogonal projection that kills the e_{src}-channel, and define

$$\bar{M}_T := \Pi_{\mathrm{src}} \circ M_T.$$

Then:

(i) M_T is blind to the incoming e_{src}-channel:

$$M_T = M_T \circ \Pi_{\mathrm{src}}.$$

(ii) $\bar{M}_T$ preserves the e_{pos}-channel, the e_{prof}-channel, and every channel in E_{carry} exactly.

(iii) *$\bar{M}_T$ has signal-blind exact scalar transport along e_{sig} over*

$$E_{\mathrm{ctrl}} := \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{prof}}\} \oplus E_{\mathrm{carry}},$$

with exactly the same scalar transport kernel as M_T:

$$\mathcal{T}^{u}_{\bar{M}_T}(i,j) = \mathcal{T}^{u}_{M_T}(i,j), \qquad 0 \le j \le i \le T.$$

(iv) *For every $\delta \ge 0$ there exists $\delta' = \delta'(\delta, \mathcal{K}\_\mathrm{set}) < \infty$ such that*

$$\bar{M}_T(\mathrm{Sat}^{\mathrm{sig}}_{\delta}(\mathcal{K}\_\mathrm{set})) \subset \mathrm{Sat}^{\mathrm{sig}}_{\delta'}(\bar{M}_T(\mathcal{K}\_\mathrm{set})).$$

More precisely, if

$$u' = u + \sum_{t=0}^{T} a_t e_{\mathrm{sig}} \mathbf{1}[\cdot = t], \qquad u \in \mathcal{K}\_\mathrm{set}, \qquad \max_t |a_t| \le \delta,$$

then

$$\bar{M}_T(u')_i = \bar{M}_T(u)_i + \left(\sum_{j=0}^{i} \mathcal{T}^{u}_{M_T}(i,j) a_j \right) e_{\mathrm{sig}}, \qquad 0 \le i \le T.$$

(v) *For every $\delta \ge 0$, $\bar{M}_T$ has signal-blind exact scalar transport along e_{sig} over*

$$E_{\mathrm{ctrl}} := \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{prof}}\} \oplus E_{\mathrm{carry}}$$

on $\mathrm{Sat}^{\mathrm{sig}}_{\delta}(\mathcal{K}\_\mathrm{set})$, with the same scalar transport kernel as M_T. More precisely, if

$$v = u + \sum_{t=0}^{T} a_t e_{\mathrm{sig}} \mathbf{1}[\cdot = t], \qquad u \in \mathcal{K}\_\mathrm{set},$$

then

$$\Pi_{\mathrm{ctrl}} \bar{M}_T(v)_i = \Pi_{\mathrm{ctrl}} v_i, \qquad 0 \le i \le T,$$

and

$$\pi_{\mathrm{sig}}(\bar{M}_T(v)_i) = \sum_{j=0}^{i} \mathcal{T}^{u}_{M_T}(i,j)\, \pi_{\mathrm{sig}}(v_j), \qquad 0 \le i \le T.$$

The right-hand side depends only on the control stream of v, hence is independent of the choice of $u \in \mathcal{K}\_\mathrm{set}$ with the same control stream.

Proof. Write

$$M_T = A^{\mathrm{diff}}_T \circ W^{\mathrm{src}}_T$$

as in the proof of Lemma K.22.

For item (i), the explicit source-writer formula there gives

$$\langle W^{\mathrm{src}}_T(u)_t, e_{\mathrm{src}} \rangle = m^u_t \, \langle u_t, e_{\mathrm{sig}} \rangle,$$

where m^u_t depends only on the control stream $(e_{\mathrm{pos}}, e_{\mathrm{prof}}, E_{\mathrm{carry}})$, and not on the incoming e_{src}-coordinate. All other channels used by W^{src}_T are likewise independent of the incoming e_{src}-channel. Hence

$$W^{\mathrm{src}}_T(u) = W^{\mathrm{src}}_T(\Pi_{\mathrm{src}} u).$$

Applying A^{diff}_T yields

$$M_T(u) = M_T(\Pi_{\mathrm{src}} u),$$

which is item (i).

Item (ii) follows because M_T already preserves $e_{\mathrm{pos}}, e_{\mathrm{prof}}, E_{\mathrm{carry}}$ exactly by Lemma K.22, and Π_{src} acts as the identity on those channels.

For item (iii), Π_{src} acts as the identity on the e_{sig}-coordinate, so

$$\langle \bar{M}_T(u)_i, e_{\mathrm{sig}} \rangle = \langle M_T(u)_i, e_{\mathrm{sig}} \rangle.$$

Since M_T has signal-blind exact scalar transport with kernel $\mathcal{T}^u_{M_T}$, the same is true for $\bar{M}_T$, with the same kernel.

For item (iv), fix $u \in \mathcal{K}\_\mathrm{set}$ and

$$u' = u + \sum_{t=0}^{T} a_t e_{\mathrm{sig}} \mathbf{1}[\cdot = t], \qquad \max_t |a_t| \le \delta.$$

The control stream is unchanged, so the same transport kernel $\mathcal{T}^u_{M_T}$ applies to both u and u'. By item (iii),

$$\langle \bar{M}_T(u')_i - \bar{M}_T(u)_i, e_{\mathrm{sig}} \rangle = \sum_{j=0}^{i} \mathcal{T}^u_{M_T}(i,j) a_j.$$

In the concrete construction of Lemma K.22, the source writer modifies only the e_{src}-channel and the diffuse block modifies only the e_{sig}-channel; every channel orthogonal to

$$\mathrm{span}\{e_{\mathrm{sig}}, e_{\mathrm{pos}}, e_{\mathrm{prof}}, e_{\mathrm{src}}\} \oplus E_{\mathrm{carry}}$$

is preserved exactly. Thus the only possible signal-dependent non-signal output channel is e_{src}, and Π_{src} removes it. Hence

$$\bar{M}_T(u')_i - \bar{M}_T(u)_i = \left(\sum_{j=0}^{i} \mathcal{T}^u_{M_T}(i,j) a_j \right) e_{\mathrm{sig}},$$

which is exactly a bounded signal-fiber perturbation over $\bar{M}_T(u)$. Since T is finite and $\mathcal{K}\_\mathrm{set}$ is compact, the quantity

$$\sup_{u \in \mathcal{K}\_\mathrm{set}} \sup_{0 \le i \le T} \sum_{j=0}^{i} |\mathcal{T}^u_{M_T}(i,j)|$$

is finite, so one may take

$$\delta' := \delta \sup_{u \in \mathcal{K}\_\mathrm{set}} \sup_{0 \le i \le T} \sum_{j=0}^{i} |\mathcal{T}^u_{M_T}(i,j)|.$$

For item (v), fix $\delta \ge 0$ and $v \in \mathrm{Sat}^{\mathrm{sig}}_{\delta}(\mathcal{K}\_\mathrm{set})$. Write

$$v = u + \sum_{t=0}^{T} a_t e_{\mathrm{sig}} \mathbf{1}[\cdot = t] \qquad \text{with } u \in \mathcal{K}\_\mathrm{set}.$$

By item (iv),

$$\bar{M}_T(v)_i = \bar{M}_T(u)_i + \left(\sum_{j=0}^{i} \mathcal{T}^u_{M_T}(i,j) a_j \right) e_{\mathrm{sig}}.$$

Taking the e_{sig}-coordinate and using item (iii) on $u \in \mathcal{K}\_\mathrm{set}$, we obtain

$$\begin{aligned}\pi_{\mathrm{sig}}(\bar{M}_T(v)_i) &= \pi_{\mathrm{sig}}(\bar{M}_T(u)_i) + \sum_{j=0}^{i} \mathcal{T}^u_{M_T}(i,j)a_j \\ &= \sum_{j=0}^{i} \mathcal{T}^u_{M_T}(i,j)\,\pi_{\mathrm{sig}}(u_j) + \sum_{j=0}^{i} \mathcal{T}^u_{M_T}(i,j)a_j \\ &= \sum_{j=0}^{i} \mathcal{T}^u_{M_T}(i,j)\,\pi_{\mathrm{sig}}(v_j).\end{aligned}$$

Moreover, from the explicit construction, W_T^{src} modifies only the e_{src}-channel, A_T^{diff} modifies only the e_{sig}-channel, and Π_{src} kills only the e_{src}-channel. Hence $\bar{M}_T$ acts as the identity on

$$E_{\mathrm{ctrl}} = \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{prof}}\} \oplus E_{\mathrm{carry}}$$

for every input, and therefore

$$\Pi_{\mathrm{ctrl}}\bar{M}_T(v)_i = \Pi_{\mathrm{ctrl}}v_i.$$

Finally, since $\mathcal{T}^u_{M_T}$ depends only on the control stream, the displayed kernel is independent of the choice of $u \in \mathcal{K}\_\mathrm{set}$ with the same control stream as v. Thus $\bar{M}_T$ has signal-blind exact scalar transport on $\mathrm{Sat}^{\mathrm{sig}}_\delta(\mathcal{K}\_\mathrm{set})$ with the same kernel as M_T. This proves the claim. □

Lemma K.25 (Balanced path lower bound)**.** *Fix $\beta \in (0,1)$, set $\gamma := 1-\beta$, fix $k \geq 1$, and fix $\tau_{\max} \geq 0$. Then there exists a constant $c^{\mathrm{bal}}_{k,\beta,\tau_{\max}} > 0$ such that for every $0 \leq \tau_* \leq \tau_{\max}$ and every $\ell \geq k$, with $t = \tau_* + \ell$,*

$$\sum_{\substack{\tau_* = i_0 < i_1 < \cdots < i_k = t \\ \frac{\ell}{2k} \leq i_r - i_{r-1} \leq \frac{2\ell}{k}\ \forall r}} \prod_{r=1}^{k} (i_r+1)^{-\beta} \geq c^{\mathrm{bal}}_{k,\beta,\tau_{\max}} (1+\ell)^{k(1-\beta)-1}.$$

Proof. The number of balanced paths is $\gtrsim_k \ell^{k-1}$ for all $\ell \geq k$.

For every balanced path and every $r = 1, \dots, k$,

$$i_r + 1 \asymp_{k,\tau_{\max}} 1 + \ell.$$

Hence every balanced path contributes at least

$$C^{-1}_{k,\beta,\tau_{\max}} (1+\ell)^{-k\beta}.$$

Multiplying by the number of balanced paths gives

$$\gtrsim \ell^{k-1}(1+\ell)^{-k\beta} \asymp (1+\ell)^{k-1-k\beta} = (1+\ell)^{k(1-\beta)-1}.$$

□

Lemma K.26 (Competitor suppression)**.** *Fix $\beta \in (0,1)$, set $\gamma := 1-\beta$, fix $k \geq 1$, and fix $\tau_{\max} \geq 0$. Consider a depth-$(k+1)$ exact scalar transport stack on a distinguished signal channel, consisting of one selector block $S_{H,\tau_*,\varepsilon_H}$ followed by k diffuse profile-compensated macro-layers. Let*

$$\mathcal{T}^u_{\mathrm{stack}}(t,\tau)$$

denote the resulting exact scalar transport kernel on that signal channel. Assume the selector satisfies

$$\frac{1}{2} \le D^u_{\mathrm{sel}}(\tau_*) \le 2, \qquad |D^u_{\mathrm{sel}}(\tau)| \le \varepsilon_H \quad (\tau \ne \tau_*),$$

uniformly in u, and each macro-layer satisfies

$$1 \le D^u_{\mathrm{mac}}(i) \le d^+_{\mathrm{mac}}, \qquad K^u_{\mathrm{mac}}(i,j) \le a^+_{\mathrm{mac}}(i+1)^{-\beta}.$$

Then there exists $C_{\mathrm{comp}} < \infty$, independent of H, such that for every

$$t = \tau_* + \ell, \qquad 1 \le \ell \le H,$$

one has

$$\sum_{\substack{0 \le \tau < t \\ \tau \ne \tau_*}} \left|\mathcal{T}^u_{\mathrm{stack}}(t,\tau)\right| \le C_{\mathrm{comp}}\, \varepsilon_H\, (1+\ell)^{k(1-\beta)}.$$

In particular, if

$$\varepsilon_H \le c_0 (H+1)^{-1}$$

with $c_0 > 0$ small enough, then

$$\sum_{\substack{0 \le \tau < t \\ \tau \ne \tau_*}} \left|\mathcal{T}^u_{\mathrm{stack}}(t,\tau)\right| \le \frac{1}{2} c_{\mathrm{sig}} (1+\ell)^{k(1-\beta)-1}$$

for any prescribed $c_{\mathrm{sig}} > 0$ after reducing c_0.

Proof. Fix a competitor source $\tau \ne \tau_*$ with $\tau < t$. Any path from τ to t through the selector-plus-k-macro-layer stack must contain at least one genuine jump, because diagonal propagation alone cannot change the time index.

Fix a path with exactly j jump layers, where $1 \le j \le k$, and let

$$\tau = i_0 < i_1 < \cdots < i_j = t$$

be the corresponding jump times. The selector contributes at most ε_H at the source $\tau \ne \tau_*$. Each jump contributes at most

$$a^+_{\mathrm{mac}}(i_r+1)^{-\beta}, \qquad r = 1, \dots, j.$$

Each non-jump macro-layer contributes at most the diagonal bound d^+_{mac}.

Hence every such path has weight bounded by

$$C_0\, \varepsilon_H \prod_{r=1}^{j} (i_r+1)^{-\beta},$$

where C_0 depends only on k and d^+_{mac}.

Now sum over all jump times for fixed j:

$$\sum_{\tau = i_0 < i_1 < \cdots < i_j = t} \prod_{r=1}^{j} (i_r+1)^{-\beta} = (t+1)^{-\beta} \sum_{\tau < i_1 < \cdots < i_{j-1} < t} \prod_{r=1}^{j-1} (i_r+1)^{-\beta}.$$

Using the elementary symmetric-sum bound,

$$\sum_{\tau < i_1 < \cdots < i_{j-1} < t} \prod_{r=1}^{j-1} (i_r+1)^{-\beta} \le \frac{1}{(j-1)!} \left(\sum_{m=1}^{t-1} (m+1)^{-\beta} \right)^{j-1},$$

and

$$\sum_{m=1}^{t-1}(m+1)^{-\beta} \lesssim (1+t)^{1-\beta},$$

we obtain

$$\sum_{\tau=i_0<i_1<\cdots<i_j=t}\prod_{r=1}^{j}(i_r+1)^{-\beta} \le C_j(1+t)^{j(1-\beta)-1}.$$

Therefore

$$|\mathcal{T}^{u}_{\mathrm{stack}}(t,\tau)| \le C_1\,\varepsilon_H\sum_{j=1}^{k}(1+t)^{j(1-\beta)-1} \le C_2\,\varepsilon_H(1+t)^{k(1-\beta)-1},$$

since k is fixed.

Now $t=\tau_*+\ell$ with $0\le\tau_*\le\tau_{\max}$, so

$$1+t \asymp_{\tau_{\max}} 1+\ell.$$

Hence

$$|\mathcal{T}^{u}_{\mathrm{stack}}(t,\tau)| \lesssim \varepsilon_H(1+\ell)^{k(1-\beta)-1}.$$

Finally sum over all competitors $\tau<t$. There are at most $t\lesssim_{\tau_{\max}} 1+\ell$ of them, so

$$\sum_{\substack{0\le\tau<t\\ \tau\ne\tau_*}}|\mathcal{T}^{u}_{\mathrm{stack}}(t,\tau)| \lesssim \varepsilon_H(1+\ell)^{k(1-\beta)}.$$

This proves the first claim.

For the in-particular clause, use $1+\ell\le H+1$:

$$\varepsilon_H(1+\ell)^{k(1-\beta)} \le c_0(H+1)^{-1}(1+\ell)^{k(1-\beta)} \le c_0(1+\ell)^{k(1-\beta)-1}.$$

Reducing c_0 if necessary yields the desired factor $\frac{1}{2}c_{\mathrm{sig}}$. □

Remark K.27 (Width bookkeeping). After the positional writer has fixed the direction e_{pos}, choose once and for all six orthonormal directions

$$e_{\mathrm{sig}},\ e_{\mathrm{prof}},\ e_{\mathrm{tail}},\ e_{\mathrm{aux}},\ e_{\mathrm{src}},\ e_{\mathrm{tgt}},$$

all orthogonal to e_{pos}.

The preparatory network Q_H uses $e_{\mathrm{prof}}, e_{\mathrm{tail}}, e_{\mathrm{aux}}, e_{\mathrm{src}}, e_{\mathrm{tgt}}$; the selector block reuses e_{aux} and preserves e_{prof}; each diffuse profile-compensated macro-layer reuses e_{src} and preserves e_{prof}; the direction e_{tgt} remains available as an auxiliary spare scratch direction. No block requires any additional fresh ambient direction beyond these seven coordinates.

In the concrete architecture, each width-D block also provides D a-slots and D g-slots in the split

$$(a,g)=\mathrm{split}(xW^{\mathrm{in}}+b^{\mathrm{in}}).$$

The constructions below use at most six active a-slots and at most three active g-slots in any single block: the plateau window uses four a-slots, the window writer uses six a-slots and two g-slots, the local multiplier uses four a-slots and two g-slots, the repaired source writer uses four a-slots and two g-slots, the repaired diffuse transport block uses two a-slots and one g-slot, the damped predecessor integrator uses three a-slots and one g-slot, and the simultaneous scratch reset uses one a-slot and three g-slots.

Hence the same condition

$$D\ge 7$$

simultaneously provides the seven persistent ambient directions and enough concrete a-/g-slots for every primitive block.

Proof of Theorem 12. Fix $H \geq 1$ and $0 \leq \tau_* \leq \tau_{\max}$. Set

$$L_H := \tau_{\max} + H, \qquad T_H := L_H + 1.$$

Composite architecture. For each horizon parameter $H \geq 1$ and source index $0 \leq \tau_* \leq \tau_{\max}$, we construct

$$G_{H,\tau_*} = M_{H,k} \circ \cdots \circ M_{H,1} \circ S_{H,\tau_*,\varepsilon_H} \circ Q_H \circ P_H.$$

Here P_H writes a one-directional positional code, Q_H builds a signal-transparent preparatory power-profile channel, $S_{H,\tau_*,\varepsilon_H}$ is a selector that isolates the chosen source τ_*, and $M_{H,1}, \dots, M_{H,k}$ are the diffuse profile-compensated macro-layers that generate the target polynomial transport envelope.

Inside the proof we also introduce projected variants of the macro-layers in order to expose the exact signal-channel transport kernel while removing an auxiliary scratch channel. This internal projection does not change the realized map on the relevant signal fibers, so it is used only as a bookkeeping device in the kernel calculation.

Step 1: write the positional code. Apply Corollary 4.11 on the finite prefix $\{0, \dots, L_H\}$. This yields a block

$$P_H : (\mathbb{R}^D)^{T_H} \to (\mathbb{R}^D)^{T_H}$$

and a unit direction e_{pos} such that

$$P_H(h)_t = h_t + \lambda_t e_{\text{pos}}, \qquad 0 \leq t \leq L_H,$$

for some scalars λ_t, and such that on

$$\mathcal{K}\_\text{set}_H := P_H(\mathcal{X}_0^{(H)})$$

the scalar ranges

$$I_t := \{\langle u_t, e_{\text{pos}} \rangle : \ u \in \mathcal{K}\_\text{set}_H\}$$

are compact and strictly ordered:

$$I_0 < \cdots < I_{L_H} \subset (0, \infty).$$

Since $D \geq 7$, after fixing e_{pos} we may choose orthonormal directions

$$e_{\text{sig}}, \ e_{\text{prof}}, \ e_{\text{tail}}, \ e_{\text{aux}}, \ e_{\text{src}}, \ e_{\text{tgt}}$$

all orthogonal to e_{pos}; see Remark K.27.

By Corollary 4.12, for every $x \in \mathcal{X}_0^{(H)}$, every τ, and every scalar a,

$$P_H(x + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau])_t = P_H(x)_t + a\, e_{\text{sig}} \mathbf{1}[t = \tau].$$

In particular,

$$\langle P_H(x + a\, e_{\text{sig}} \mathbf{1}[\cdot = \tau])_t, e_{\text{pos}} \rangle = \langle P_H(x)_t, e_{\text{pos}} \rangle.$$

Step 2: build the preparatory power-profile network. Apply Corollary K.21 to the compact set $\mathcal{K}\_\text{set}_H$, with the fixed orthonormal directions

$$e_{\text{sig}}, \ e_{\text{pos}}, \ e_{\text{prof}}, \ e_{\text{tail}}, \ e_{\text{aux}}, \ e_{\text{src}}, \ e_{\text{tgt}},$$

which satisfy the hypotheses of that corollary. This yields a constant-depth network

$$Q_H : (\mathbb{R}^D)^{T_H} \to (\mathbb{R}^D)^{T_H}$$

with the following properties.

Signal preservation. The signal channel is preserved exactly:

$$\langle Q_H(u)_t, e_{\mathrm{sig}} \rangle = \langle u_t, e_{\mathrm{sig}} \rangle.$$

Positional preservation. The positional-control coordinate is preserved exactly:

$$\langle Q_H(u)_t, e_{\mathrm{pos}} \rangle = \langle u_t, e_{\mathrm{pos}} \rangle.$$

Profile growth. The profile channel on the prescribed direction e_{prof} satisfies

$$c_r^-(t+1)^\gamma \le \langle Q_H(u)_t, e_{\mathrm{prof}} \rangle \le c_r^+(t+1)^\gamma, \qquad \gamma = 1 - \beta.$$

Signal transparency. The map Q_H is signal-transparent relative to $(e_{\mathrm{pos}}, e_{\mathrm{prof}})$: for every u, every τ, and every scalar a,

$$\langle Q_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t, e_{\mathrm{pos}} \rangle = \langle Q_H(u)_t, e_{\mathrm{pos}} \rangle,$$
$$\langle Q_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t, e_{\mathrm{prof}} \rangle = \langle Q_H(u)_t, e_{\mathrm{prof}} \rangle,$$
$$\langle Q_H(u + a\, e_{\mathrm{sig}} \mathbf{1}[\cdot = \tau])_t, e_{\mathrm{sig}} \rangle = \langle Q_H(u)_t, e_{\mathrm{sig}} \rangle + a\, \mathbf{1}[t = \tau].$$

Write

$$R_H := Q_H \circ P_H.$$

Step 3: select the source index. Apply Lemma K.12 on the image of R_H, using the already fixed directions $e_{\mathrm{pos}}, e_{\mathrm{sig}}, e_{\mathrm{aux}}$, with

$$E_{\mathrm{carry}} := \mathrm{span}\{e_{\mathrm{prof}}\}, \qquad \varepsilon_H := c_0 (H+1)^{-1},$$

where $c_0 > 0$ will be fixed later. This yields a selector module

$$S_{H, \tau_*, \varepsilon_H}$$

which preserves the positional and profile channels and has exact diagonal signal transport

$$\mathcal{T}^u_S(i,j) = D^u_{\mathrm{sel}}(i) \mathbf{1}[i = j]$$

with

$$\frac{1}{2} \le D^u_{\mathrm{sel}}(\tau_*) \le 2, \qquad |D^u_{\mathrm{sel}}(\tau)| \le \varepsilon_H \quad (\tau \ne \tau_*).$$

Step 4: add the k macro-layers. Define

$$\mathcal{K}\_\mathrm{set}^{\mathrm{mac}}_{H,0} := S_{H, \tau_*, \varepsilon_H}(R_H(\mathcal{X}_0^{(H)})).$$

This is compact. By Step 2 and Step 3, on $\mathcal{K}\_\mathrm{set}^{\mathrm{mac}}_{H,0}$ the positional-control ranges are still

$$I_0 < \cdots < I_{L_H} \subset (0, \infty),$$

and the profile channel still satisfies

$$c_r^-(t+1)^\gamma \le \langle u_t, e_{\mathrm{prof}}\rangle \le c_r^+(t+1)^\gamma, \qquad 0 \le t \le L_H.$$

Apply Lemma K.22 with $T = L_H$ to $\mathcal{K}\_\mathrm{set}_{H,0}^{\mathrm{mac}}$, using the fixed directions

$$e_{\mathrm{sig}},\ e_{\mathrm{pos}},\ e_{\mathrm{prof}},\ e_{\mathrm{src}}, \qquad E_{\mathrm{carry}} := \{0\},$$

to obtain $M_{H,1}$. Define

$$\bar{M}_{H,1} := \Pi_{\mathrm{src}} \circ M_{H,1}.$$

If $k \ge 2$, set

$$\mathcal{K}\_\mathrm{set}_{H,1}^{\mathrm{mac}} := \bar{M}_{H,1}(\mathcal{K}\_\mathrm{set}_{H,0}^{\mathrm{mac}}).$$

Inductively, suppose that for some $1 \le r \le k-1$ we have already constructed

$$M_{H,1}, \dots, M_{H,r}, \qquad \bar{M}_{H,1}, \dots, \bar{M}_{H,r},$$

and compact sets

$$\mathcal{K}\_\mathrm{set}_{H,0}^{\mathrm{mac}}, \dots, \mathcal{K}\_\mathrm{set}_{H,r}^{\mathrm{mac}}$$

such that for each $1 \le s \le r$,

$$\mathcal{K}\_\mathrm{set}_{H,s}^{\mathrm{mac}} = \bar{M}_{H,s}(\mathcal{K}\_\mathrm{set}_{H,s-1}^{\mathrm{mac}}),$$

and on every $\mathcal{K}\_\mathrm{set}_{H,s}^{\mathrm{mac}}$ the same ordered positional ranges

$$I_0 < \cdots < I_{L_H} \subset (0,\infty)$$

and the same two-sided profile bounds

$$c_r^-(t+1)^\gamma \le \langle u_t, e_{\mathrm{prof}}\rangle \le c_r^+(t+1)^\gamma$$

hold.

Apply Lemma K.22 to $\mathcal{K}\_\mathrm{set}_{H,r}^{\mathrm{mac}}$, with the same fixed directions, to obtain $M_{H,r+1}$. Define

$$\bar{M}_{H,r+1} := \Pi_{\mathrm{src}} \circ M_{H,r+1}.$$

If $r+1 \le k-1$, set

$$\mathcal{K}\_\mathrm{set}_{H,r+1}^{\mathrm{mac}} := \bar{M}_{H,r+1}(\mathcal{K}\_\mathrm{set}_{H,r}^{\mathrm{mac}}).$$

By Lemma K.24(ii)–(iii), each $\bar{M}_{H,r}$ preserves the e_{pos}- and e_{prof}-channels exactly and has the same exact signal-channel transport kernel as $M_{H,r}$. Therefore the induction is well-posed, and after k steps we obtain macro-layers

$$M_{H,1}, \dots, M_{H,k}, \qquad \bar{M}_{H,1}, \dots, \bar{M}_{H,k-1},$$

all preserving the positional and profile channels and having exact signal transport kernels

$$\mathcal{T}^u_{M_{H,r}}(i,j) = D^u_{\mathrm{mac},r}(i)\mathbf{1}[i=j] + K^u_{\mathrm{mac},r}(i,j)\mathbf{1}[j<i],$$

with uniform bounds

$$1 \le D^u_{\mathrm{mac},r}(i) \le d^+_{\mathrm{mac}},$$
$$a^-_{\mathrm{mac}}(i+1)^{-\beta} \le K^u_{\mathrm{mac},r}(i,j) \le a^+_{\mathrm{mac}}(i+1)^{-\beta} \qquad (j<i).$$

Moreover, by Lemma K.24(i),

$$M_{H,r+1} = M_{H,r+1} \circ \Pi_{\mathrm{src}} \qquad (r = 1, \dots, k-1),$$

hence the actual network from the theorem statement satisfies

$$G_{H,\tau_*} = M_{H,k} \circ \cdots \circ M_{H,1} \circ S_{H,\tau_*,\varepsilon_H} \circ Q_H \circ P_H = \widehat{G}_{H,\tau_*} \circ R_H,$$

where

$$\widehat{G}_{H,\tau_*} := M_{H,k} \circ \bar{M}_{H,k-1} \circ \cdots \circ \bar{M}_{H,1} \circ S_{H,\tau_*,\varepsilon_H}, \qquad R_H := Q_H \circ P_H.$$

By Lemma K.24(iii), each $\bar{M}_{H,r}$ has the same signal-channel transport kernel as the corresponding $M_{H,r}$, so all of the above kernel bounds remain unchanged.

Step 5: identify the score with the transport kernel. Take the normalized probes in Definition 5 to be

$$c^{(H,\tau_*)} := e_{\mathrm{sig}}, \qquad \rho_t^{(H,\tau_*)} := e_{\mathrm{sig}} \qquad (0 \le t \le L_H).$$

These are independent of x, common to all source indices τ, and satisfy

$$\|c^{(H,\tau_*)}\|_2 = 1, \qquad \|\rho_t^{(H,\tau_*)}\|_2 = 1.$$

Set

$$R_H := Q_H \circ P_H.$$

By Step 1 and Step 2, R_H is signal-transparent along e_{sig} over

$$E_{\mathrm{ctrl}} := \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{prof}}\}$$

on $\mathcal{X}_0^{(H)}$.

Fix some $\delta_* > 0$, for example $\delta_* = 1$, and define

$$\mathcal{Y}_H := \mathrm{Sat}^{\mathrm{sig}}_{\delta_*}(R_H(\mathcal{X}_0^{(H)})).$$

This set is compact.

Define

$$\mathcal{Y}_{H,0} := S_{H,\tau_*,\varepsilon_H}(\mathcal{Y}_H).$$

By Lemma K.14, there exists a finite $\delta_{H,0}$ such that

$$\mathcal{Y}_{H,0} \subset \mathrm{Sat}^{\mathrm{sig}}_{\delta_{H,0}}(\mathcal{K}\_\mathrm{set}^{\mathrm{mac}}_{H,0}).$$

For $r = 1, \dots, k-1$, define inductively

$$\mathcal{Y}_{H,r} := \bar{M}_{H,r}(\mathcal{Y}_{H,r-1}).$$

By Lemma K.24(iv), there exists a finite $\delta_{H,r}$ such that

$$\mathcal{Y}_{H,r} \subset \mathrm{Sat}^{\mathrm{sig}}_{\delta_{H,r}}(\mathcal{K}\_\mathrm{set}^{\mathrm{mac}}_{H,r}), \qquad r = 1, \dots, k-1.$$

By Corollary K.10, the selector $S_{H,\tau_*,\varepsilon_H}$ has signal-blind exact scalar transport along e_{sig} over

$$E_{\mathrm{ctrl}} = \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{prof}}\}$$

on $\mathcal{Y}_H$. For each $r = 1, \dots, k-1$, Lemma K.24(v) shows that $\bar{M}_{H,r}$ has signal-blind exact scalar transport along e_{sig} over the same control subspace on $\mathcal{Y}_{H,r-1}$. Finally, since

$$\mathcal{Y}_{H,k-1} \subset \mathrm{Sat}^{\mathrm{sig}}_{\delta_{H,k-1}}(\mathcal{K}\_\mathrm{set}^{\mathrm{mac}}_{H,k-1}),$$

Corollary K.23 implies that the final macro-layer $M_{H,k}$ has signal-blind exact scalar transport along e_{sig} over the same control subspace on $\mathcal{Y}_{H,k-1}$, with the same kernel $\mathcal{T}^{u}_{M_{H,k}}$ as on $\mathcal{K}\_\mathrm{set}^{\mathrm{mac}}_{H,k-1}$.

Repeated application of Lemma K.8(ii) therefore yields that the full post-preparatory stack

$$\widehat{G}_{H,\tau_*} = M_{H,k} \circ \bar{M}_{H,k-1} \circ \cdots \circ \bar{M}_{H,1} \circ S_{H,\tau_*,\varepsilon_H}$$

has signal-blind exact scalar transport along e_{sig} over

$$E_{\mathrm{ctrl}} = \mathrm{span}\{e_{\mathrm{pos}}, e_{\mathrm{prof}}\}$$

on $\mathcal{Y}_H$, with transport kernel

$$\mathcal{T}^{u}_{\widehat{G}_{H,\tau_*}}(t,\tau).$$

Hence Lemma K.9 applies with

$$R = R_H, \qquad B = \widehat{G}_{H,\tau_*}, \qquad \mathcal{K}\_\mathrm{set} = \mathcal{X}^{(H)}_0.$$

Therefore, for every $x \in \mathcal{X}^{(H)}_0$ and every $0 \le \tau \le t \le L_H$,

$$e_{\mathrm{sig}}^{\top} \frac{\partial G_{H,\tau_*,t}(x)}{\partial x_\tau} e_{\mathrm{sig}} = \mathcal{T}^{R_H(x)}_{\widehat{G}_{H,\tau_*}}(t,\tau).$$

By our choice of score channels,

$$\mathsf{S}^{(H,\tau_*)}_{t,\tau}(x) = (\rho^{(H,\tau_*)}_t)^{\top} J^{G_{H,\tau_*}}_{t,\tau}(x)\, c^{(H,\tau_*)} = e_{\mathrm{sig}}^{\top} J^{G_{H,\tau_*}}_{t,\tau}(x)\, e_{\mathrm{sig}} = \mathcal{T}^{R_H(x)}_{\widehat{G}_{H,\tau_*}}(t,\tau).$$

Set

$$u := R_H(x).$$

Step 6: lower-bound the balanced paths. Fix

$$t = \tau_* + \ell, \qquad \ell \ge k.$$

Expand the kernel product along the intermediate states. Writing

$$u^{(0)} := u, \qquad u^{(r)} := \bar{M}_{H,r} \circ \cdots \circ \bar{M}_{H,1} \circ S_{H,\tau_*,\varepsilon_H}(u) \quad (1 \le r \le k-1),$$

one has

$$\mathcal{T}^{u}_{\widehat{G}_{H,\tau_*}} = \mathcal{T}^{u^{(k-1)}}_{M_{H,k}} \mathcal{T}^{u^{(k-2)}}_{\bar{M}_{H,k-1}} \cdots \mathcal{T}^{u^{(0)}}_{\bar{M}_{H,1}} \mathcal{T}^{u}_{S_{H,\tau_*,\varepsilon_H}}.$$

Since every factor preserves the control channels exactly and its kernel depends only on the control stream, all intermediate control streams equal that of u. Hence the same pathwise kernel bounds apply throughout. Moreover, by Lemma K.24,

$$\mathcal{T}^{u^{(r-1)}}_{\bar{M}_{H,r}}(i,j) = \mathcal{T}^{u^{(r-1)}}_{M_{H,r}}(i,j) \qquad (r = 1, \dots, k-1).$$

Consider the family of paths that use all k macro-layers as jumps and whose jump times are balanced:

$$\tau_* = i_0 < i_1 < \cdots < i_k = t, \qquad \frac{\ell}{2k} \le i_r - i_{r-1} \le \frac{2\ell}{k}.$$

For each such path, the selector contributes at least $\frac{1}{2}$, and each jump contributes at least

$$a_{\mathrm{mac}}^{-}(i_r+1)^{-\beta}.$$

Hence

$$\mathcal{T}^{u}_{\widehat{G}_{H,\tau_*}}(t,\tau_*) \geq \frac{1}{2}(a_{\mathrm{mac}}^{-})^k \sum_{\substack{\tau_*=i_0<\cdots<i_k=t \\ \text{balanced}}} \prod_{r=1}^{k}(i_r+1)^{-\beta}.$$

By Lemma K.25,

$$\mathcal{T}^{u}_{\widehat{G}_{H,\tau_*}}(t,\tau_*) \geq c_{\mathrm{good}}(1+\ell)^{k(1-\beta)-1}.$$

Step 7: handle small lags. There are only finitely many pairs (τ_*,ℓ) with

$$0 \leq \tau_* \leq \tau_{\max}, \qquad 1 \leq \ell < k.$$

For each such pair, choose the path that jumps in the first ℓ macro-layers and then propagates diagonally. Since all indices lie in the finite set $\{0,\dots,\tau_{\max}+k-1\}$, the corresponding exact path weight is bounded below by a positive constant depending only on $(k,\beta,\tau_{\max})$. Therefore there exists

$$c_{\mathrm{small}} > 0$$

such that

$$\mathcal{T}^{u}_{\widehat{G}_{H,\tau_*}}(\tau_*+\ell,\tau_*) \geq c_{\mathrm{small}} \qquad (1 \leq \ell < k).$$

Combining the large- and small-lag cases, there exists $c_{\mathrm{sig}} > 0$ such that for all $1 \leq \ell \leq H$,

$$\mathcal{T}^{u}_{\widehat{G}_{H,\tau_*}}(\tau_*+\ell,\tau_*) \geq c_{\mathrm{sig}}(1+\ell)^{\nu_k(\beta)}, \qquad \nu_k(\beta) = k(1-\beta)-1.$$

Step 8: suppress the competitors. Apply Lemma K.26 to the selector-plus-macro transport kernel. By Lemma K.24(iii), each projected macro-layer $\bar{M}_{H,r}$ has exactly the same signal-channel transport kernel as the corresponding macro-layer $M_{H,r}$, so the lemma applies verbatim to the post-preparatory stack

$$\widehat{G}_{H,\tau_*} = M_{H,k} \circ \bar{M}_{H,k-1} \circ \cdots \circ \bar{M}_{H,1} \circ S_{H,\tau_*,\varepsilon_H}.$$

Since the exact transport coefficient equals the Jacobian score coefficient on the signal channel,

$$\sum_{\substack{0\leq\tau<t \\ \tau\neq\tau_*}} \left|\mathsf{S}^{(H,\tau_*)}_{t,\tau}(x)\right| = \sum_{\substack{0\leq\tau<t \\ \tau\neq\tau_*}} \left|\mathcal{T}^{u}_{\widehat{G}_{H,\tau_*}}(t,\tau)\right| \leq C_{\mathrm{comp}}\varepsilon_H(1+\ell)^{k(1-\beta)}.$$

Choose $c_0 > 0$ small enough that

$$C_{\mathrm{comp}}\varepsilon_H(1+\ell)^{k(1-\beta)} \leq \frac{1}{2}c_{\mathrm{sig}}(1+\ell)^{\nu_k(\beta)} \qquad (1 \leq \ell \leq H).$$

Then

$$\mathsf{M}^{(H,\tau_*)}_{\tau_*+\ell,\tau_*}(x) \geq \frac{1}{2}c_{\mathrm{sig}}(1+\ell)^{\nu_k(\beta)}.$$

So we may take

$$c_{-} := \frac{1}{2}c_{\mathrm{sig}}.$$

Step 9: anchor bounds. At $\ell = 1$,

$$\mathsf{M}^{(H,\tau_*)}_{\tau_*+1,\tau_*}(x) \geq c_-(1+1)^{\nu_k(\beta)} = 2^{\nu_k(\beta)}c_-.$$

Hence we may take

$$m_- := 2^{\nu_k(\beta)}c_- > 0.$$

For the anchor upper bound, note first that

$$\mathsf{M}^{(H,\tau_*)}_{\tau_*+1,\tau_*}(x) \leq \left|\mathsf{S}^{(H,\tau_*)}_{\tau_*+1,\tau_*}(x)\right|.$$

By Step 5,

$$\mathsf{S}^{(H,\tau_*)}_{\tau_*+1,\tau_*}(x) = \mathcal{T}^{R_H(x)}_{\widehat{G}_{H,\tau_*}}(\tau_*+1,\tau_*).$$

Since the selector is diagonal, any path from τ_* to τ_*+1 through

$$\widehat{G}_{H,\tau_*} = M_{H,k} \circ \bar{M}_{H,k-1} \circ \cdots \circ \bar{M}_{H,1} \circ S_{H,\tau_*,\varepsilon_H}$$

must contain exactly one off-diagonal jump, and that jump must occur in one of the k macro-layers. Therefore

$$\begin{aligned}
&\mathcal{T}^{R_H(x)}_{\widehat{G}_{H,\tau_*}}(\tau_*+1,\tau_*) \\
&\quad = D^u_{\mathrm{sel}}(\tau_*)\sum_{r=1}^{k}\left(\prod_{q<r} D^u_{\mathrm{mac},q}(\tau_*)\right) K^u_{\mathrm{mac},r}(\tau_*+1,\tau_*)\left(\prod_{q>r} D^u_{\mathrm{mac},q}(\tau_*+1)\right),
\end{aligned}$$

where $u = R_H(x)$.

Using

$$D^u_{\mathrm{sel}}(\tau_*) \leq 2, \qquad D^u_{\mathrm{mac},q}(i) \leq d^+_{\mathrm{mac}}, \qquad K^u_{\mathrm{mac},r}(\tau_*+1,\tau_*) \leq a^+_{\mathrm{mac}}(\tau_*+2)^{-\beta} \leq a^+_{\mathrm{mac}},$$

we obtain

$$\left|\mathcal{T}^{R_H(x)}_{\widehat{G}_{H,\tau_*}}(\tau_*+1,\tau_*)\right| \leq 2k\,(d^+_{\mathrm{mac}})^{k-1}a^+_{\mathrm{mac}}.$$

Hence one may take

$$m_+ := 2k\,(d^+_{\mathrm{mac}})^{k-1}a^+_{\mathrm{mac}},$$

which is independent of H, τ_*, and x. Consequently,

$$\mathsf{M}^{(H,\tau_*)}_{\tau_*+1,\tau_*}(x) \leq m_+.$$

This verifies Definition 5. The sign classification follows immediately from the sign of

$$\nu_k(\beta) = k(1-\beta) - 1.$$

□