KRONECKER DELTA RULE FOR LONG CONTEXT LINEAR ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose KronLSTM, a linear attention architecture based on the Kronecker Delta Rule, a novel update rule for associative matrix memories. KronLSTM extends the xLSTM family with a memory-revision mechanism that corrects targeted associations while modifying the memory only along a single direction in matrix space. This minimally invasive memory revision reduces interference with previously stored information compared with delta-rule updates, making it particularly well suited for long-context language modeling. The resulting memory dynamics can be mathematically described as a Markov chain that converges to a steady-state distribution, while the expected norm of the memory matrix remains uniformly bounded throughout the entire input sequence. Hence, KronLSTM ensures stable memory dynamics without requiring additional normalization and stabilization mechanisms for the memory state. Experiments show that KronLSTM is competitive with, and often outperforms, existing linear attention architectures, including models of the (Gated) DeltaNet family, with particularly strong performance on long context benchmarks. These results establish the Kronecker Delta Rule as an effective and analytically tractable foundation for stable associative memory in recurrent linear attention architectures.

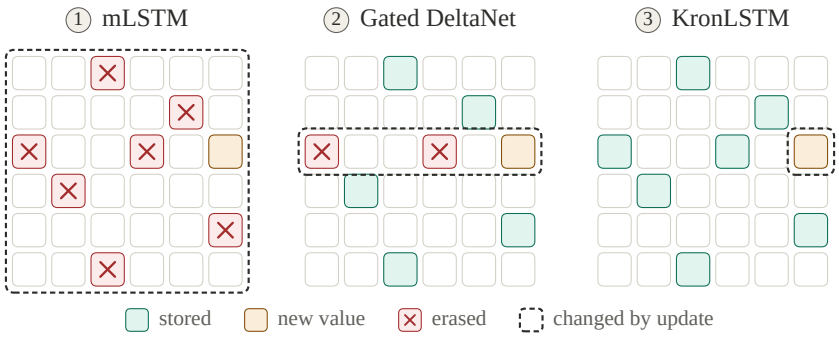

Figure 1: Representation of the memory matrix state revision for mLSTM, Gated DeltaNet, and KronLSTM. Interference with stored values: mLSTM affects all values, Gated DeltaNet affects those with overlapping keys, and KronLSTM affects only those with overlapping key and value pairs.

1 INTRODUCTION

Large language models (LLMs) predominantly use the Transformer architecture (Vaswani et al., 2017; Touvron et al., 2023a; Qwen Team, 2023; OpenAI, 2025), whose softmax attention mechanism has computational costs that grow quadratically with context length, while its key–value cache grows linearly. Consequently, inference becomes increasingly costly as contexts grow. To reduce these costs, recent work has explored replacing softmax attention with linear attention mechanisms (Katharopoulos et al., 2020; Schlag et al., 2021), which can be traced back to the idea of fast weight programmers (Schmidhuber, 1992). Linear attention compresses the context into a fixed-size memory matrix that is updated as new tokens arrive, yielding linear computational complexity and a

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

constant memory footprint. Recently, recurrent architectures based on associative memories (Kohonen, 1972; Anderson, 1972), including xLSTM (Beck et al., 2024), as well as related state-space and memory-based approaches such as Mamba (Gu & Dao, 2024; Dao & Gu, 2024; Lahoti et al., 2026), DeltaNet (Yang et al., 2024; 2025; Kimi Team et al., 2025; Hatamizadeh et al., 2026), RWKV-5 and 6 (Peng et al., 2024), and other recent architectures (Sun et al., 2024; Liu et al., 2025; Behrouz et al., 2025), have emerged as competitive alternatives to Transformers.

A crucial challenge for linear attention is controlling memory growth. If the memory of recurrent architectures grows without being bounded, its readout can become arbitrarily large, leading to numerical instabilities, particularly during learning. The original LSTM (Hochreiter, 1991; Hochreiter & Schmidhuber, 1997) uses a bounded squashing function to control the exposed cell state, while its forget gate regulates the accumulation of information in memory. RWKV-5 and RWKV-6 (Peng et al., 2024) similarly control memory growth through a forget gate. Mamba architectures (Gu & Dao, 2024; Dao & Gu, 2024; Lahoti et al., 2026) also restrain memory growth through a forget mechanism derived from a stable state-space ordinary differential equation. However, forget-based memory control becomes increasingly restrictive for long contexts. To retain information over long time spans, forget gates must remain close to one; otherwise, information from the beginning of the context is progressively attenuated. xLSTM (Beck et al., 2024) addresses this limitation by allowing forget gates close to one and introducing normalizer and stabilizer states to control the scale of its matrix memory and readout. However, normalization based on ratios of potentially large quantities can itself lead to numerical difficulties and learning instabilities, particularly in larger xLSTM networks (Beck et al., 2025b). mLSTM's numerical instabilities are illustrated by the stress test in Figure 2.

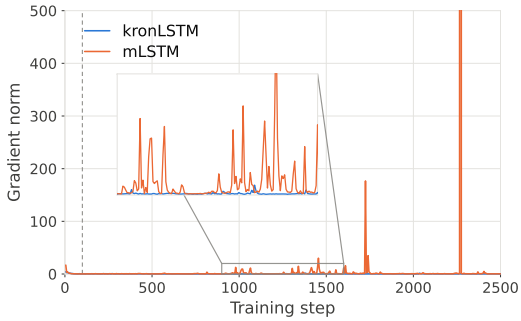

Figure 2: Stress test with high learning rates for the same mLSTM and KronLSTM architectures on language modeling. Learning becomes unstable for mLSTM as the gradient norm spikes show, while KronLSTM yields stable learning.

The DeltaNet family (Yang et al., 2024; 2025; Kimi Team et al., 2025; Hatamizadeh et al., 2026) controls memory growth through Delta Rule updates, whose dynamics converge to a steady-state distribution with bounded memory ensuring stable learning dynamics. However, revising a stored value can perturb previously stored values, leading to unwanted memory interference for long sequences. We introduce KronLSTM, which, like (Gated) DeltaNet, provides bounded memory through a delta-style update, but uses the Kronecker Delta Rule. The Kronecker Delta Rule is less invasive: it modifies the memory only along a single direction in matrix space, reducing interference with previously stored associations. This targeted memory revision results in a less noisy gradient flow over time. Our contributions are:

(i) Kronecker Delta Rule is introduced as a new update rule for associative matrix memories, which corrects the scalar associations while having minimal interference with previously stored values.
(ii) KronLSTM is introduced as a novel linear attention architecture based on the Kronecker Delta Rule equipped with a vector forget gate, a β gate, and a target gate.
(iii) Stability properties for the Kronecker Delta Rule are mathematically proved via bounds on the expected memory norm. These bounds ensure stable learning and proper gradient flow.
(iv) Experiments empirically validate the strong performance of KronLSTM at long-context language modeling and retrieval tasks. KronLSTM performs strongly across language modeling, recall, and state tracking benchmarks.

2 KronLSTM

In Subsection 2.1, we introduce the Kronecker Delta Rule and analyze its memory dynamics, and in Subsection 2.2, the Kronecker Delta Rule is embedded in the KronLSTM sequence-mixer block.

2.1 KRONECKER DELTA RULE

Linear attention update rules for matrix memory. Linear attention (Katharopoulos et al., 2020) compresses the context into a matrix memory $\boldsymbol{S}$ that is updated per token and read out per query,

$$\boldsymbol{S}_{t+1} \;=\; \boldsymbol{S}_t \;+\; \boldsymbol{v}_t \boldsymbol{k}_t^\top\,, \qquad \boldsymbol{h}_t \;=\; \boldsymbol{S}_t \boldsymbol{q}_t \;=\; \sum_{i<t} \boldsymbol{v}_i \left(\boldsymbol{k}_i^\top \boldsymbol{q}_t\right), \tag{1}$$

with $\boldsymbol{S}_t \in \mathbb{R}^{d_v \times d_k}$, keys $\boldsymbol{k}_t$ and queries $\boldsymbol{q}_t$ in $\mathbb{R}^{d_k}$, and values $\boldsymbol{v}_t$ in $\mathbb{R}^{d_v}$. Therefore, $\boldsymbol{S}$ is an associative memory (Kohonen, 1972; Anderson, 1972) with a Hebbian rule, i.e., an outer product update rule. Modern linear attention architectures like xLSTM (Beck et al., 2024), Mamba (Gu & Dao, 2024; Dao & Gu, 2024; Lahoti et al., 2026), RWKV-5 and 6 (Peng et al., 2024) rely on these associative memories equipped with gates and transformations. The memory growth is controlled by a forget gate, which stabilizes the memory dynamics.

In contrast, the Delta Rule (Widrow & Hoff, 1960; Schlag et al., 2021) corrects the memory with the difference of the retrieval to the current value when writing into the memory

$$\boldsymbol{S}_{t+1} \;=\; \boldsymbol{S}_t \;-\; \beta_t \big(\boldsymbol{S}_t \boldsymbol{k}_t - \boldsymbol{v}_t\big)\, \boldsymbol{k}_t^\top \;=\; \boldsymbol{S}_t \big(\boldsymbol{I} - \beta_t\, \boldsymbol{k}_t \boldsymbol{k}_t^\top\big) \;+\; \beta_t\, \boldsymbol{v}_t \boldsymbol{k}_t^\top, \tag{2}$$

with step size β_t. Each step is gradient descent on $\frac{1}{2}\left\|\boldsymbol{S}_t \boldsymbol{k}_t - \boldsymbol{v}_t\right\|^2$, and the transition $\boldsymbol{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top$ is a generalized Householder matrix, which admits chunkwise-parallel training (Yang et al., 2024).

The two rules differ in how an association, once stored, is changed. The Hebbian write of (1) is additive, so the memory grows with the sequence and an earlier value can be buried by later writes but never revised. The delta rule (2) revises: it removes the value currently stored at $\boldsymbol{k}_t$ before writing $\boldsymbol{v}_t$. The correction, however, proceeds along $(\boldsymbol{S}_t \boldsymbol{k}_t - \boldsymbol{v}_t)\boldsymbol{k}_t^\top$, a direction determined by the current memory content, so values stored at keys that overlap with $\boldsymbol{k}_t$ are affected as well. We aim at an error-correcting write that acts solely on the pair $(\boldsymbol{k}_t, \boldsymbol{v}_t)$.

Kronecker Delta Rule. Let $\boldsymbol{S}_t \in \mathbb{R}^{d_v \times d_k}$ be the memory matrix, $\boldsymbol{k}_t \in \mathbb{R}^{d_k}$ and $\boldsymbol{v}_t \in \mathbb{R}^{d_v}$ the key and value at step t, $c_t \in \mathbb{R}$ a target and β_t a step size. The scalar $\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t$ measures how strongly the memory associates $\boldsymbol{k}_t$ with $\boldsymbol{v}_t$. The Kronecker Delta Rule corrects this scalar towards the target along the outer product of the pair as follows

$$\boldsymbol{S}_{t+1} \;=\; \boldsymbol{S}_t \;-\; \beta_t \big(\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t \;-\; c_t\big)\, \boldsymbol{v}_t \boldsymbol{k}_t^\top \;=\; \boldsymbol{S}_t \;-\; \underbrace{\beta_t \big(\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t\big)\, \boldsymbol{v}_t \boldsymbol{k}_t^\top}_{\text{erase}} \;+\; \underbrace{\beta_t\, c_t\, \boldsymbol{v}_t \boldsymbol{k}_t^\top}_{\text{write}}. \tag{3}$$

In place of the d_v-dimensional residual $\boldsymbol{S}_t \boldsymbol{k}_t - \boldsymbol{v}_t$ of the delta rule, a scalar is corrected, and the direction $\boldsymbol{v}_t \boldsymbol{k}_t^\top$ of the correction is determined by the input pair alone without utilizing the memory content. A target $c_t \equiv 1$ stores each association with unit strength, while $c_t = 0$ erases it. The target c_t may depend on the input, i.e., serves as a gate, so that what is written becomes input-dependent. The right-hand side of (3) separates the update into an erase and a write term, both scaled by β_t. Choosing the target $\tilde{c}_t = c_t/\beta_t$ removes the step size from the write, so that β_t alone determines how much of the stored association is erased and c_t alone how much is written. KronLSTM uses this form, with both quantities computed from the input as gates (Section 2.2).

Properties of the Kronecker Delta Rule. *(i) Error correction* (see Appendix Property A2 and Property A6). An update is a step of gradient descent with step size β_t on $\frac{1}{2}\big(\boldsymbol{v}_t^\top \boldsymbol{S} \boldsymbol{k}_t - c_t\big)^2$. Representing the same pair gives $\boldsymbol{v}_t^\top \boldsymbol{S}_{t+1} \boldsymbol{k}_t - c_t \;=\; \big(\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t - c_t\big)\left(1 - \beta_t \left\|\boldsymbol{a}_t\right\|^2\right)$, so it contracts the error iff $0 < \beta_t < 2/\left\|\boldsymbol{a}_t\right\|^2$, and the Newton step $\beta_t = 1/\left\|\boldsymbol{a}_t\right\|^2$ meets the target in a single update; for $c_t = 0$ this deletes the association. *(ii) Delta Rule in vectorized form* (see Appendix Property A7). With the Kronecker product $\boldsymbol{a}_t = \boldsymbol{k}_t \otimes \boldsymbol{v}_t = \mathrm{vec}(\boldsymbol{v}_t \boldsymbol{k}_t^\top)$ and the vectorized memory $\boldsymbol{s} = \mathrm{vec}(\boldsymbol{S})$, we have $\boldsymbol{v}_t^\top \boldsymbol{S} \boldsymbol{k}_t = \boldsymbol{a}_t^\top \boldsymbol{s}$ and $\left\|\boldsymbol{a}_t\right\|^2 = \left\|\boldsymbol{k}_t\right\|^2 \left\|\boldsymbol{v}_t\right\|^2$, and (3) becomes

$$\boldsymbol{s}_{t+1} \;=\; \left(\boldsymbol{I} - \beta_t\, \boldsymbol{a}_t \boldsymbol{a}_t^\top\right) \boldsymbol{s}_t \;+\; \beta_t\, c_t\, \boldsymbol{a}_t\,, \tag{4}$$

the Widrow–Hoff Delta Rule (Widrow & Hoff, 1960) with input $\boldsymbol{a}_t$ and target c_t, which gives the rule its name. The transition factor is a generalized Householder matrix that does not depend on c_t, so the chunkwise-parallel formulation of DeltaNet (Yang et al., 2024) applies with $\boldsymbol{a}_t$ in place of $\boldsymbol{k}_t$,

the target entering only the driving term. For $\beta_t \|\boldsymbol{a}_t\|^2 \in (1,2)$ the transition factor has a negative eigenvalue, the idea behind the state-tracking capabilities of Delta Rule models (Grazzi et al., 2025).

(iii) Interference (see Appendix Property A9). For any other stored pair $(\boldsymbol{k}_i, \boldsymbol{v}_i)$, the update changes the association by

$$\Delta(\boldsymbol{v}_i^\top \boldsymbol{S}\boldsymbol{k}_i) \;=\; -\beta_t \left(\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t - c_t\right)\left(\boldsymbol{v}_i^\top \boldsymbol{v}_t\right)\left(\boldsymbol{k}_t^\top \boldsymbol{k}_i\right), \tag{5}$$

the product of value overlap and key overlap. Orthogonality in either factor protects a stored association, whereas the Hebbian and the Delta rule interference is controlled by key overlap alone.

(iv) Capacity (see Appendix Property A11). Each triple $(\boldsymbol{k}_t, \boldsymbol{v}_t, c_t)$ imposes one linear constraint $\boldsymbol{a}_t^\top \boldsymbol{s} = c_t$ on the $d_k d_v$ entries of $\boldsymbol{S}$, so up to $d_k d_v$ scalar associations can be stored exactly when the features $\{\boldsymbol{a}_i\}$ are linearly independent, against d_k value vectors for $\boldsymbol{S}\boldsymbol{k}_i = \boldsymbol{v}_i$. In particular, a single key can hold several mutually orthogonal values, which the read-out $\boldsymbol{S}\boldsymbol{k}$ returns as their sum. See Appendix A2 for more properties and more details on the above properties.

Convergence, steady state, and memory norm bounds. We will present expected bounds on the memory norm, which in turn imply stable memory dynamics that obviate numerical instabilities. Memories grow with the length of the context and the number of vectors or associations to store; therefore long contexts are most relevant when analyzing memory stability. We model a long context as a stream of i.i.d. triples $(\boldsymbol{k}_t, \boldsymbol{v}_t, c_t) \sim p$ with fixed step size β and set

$$\bar{\boldsymbol{A}} := \mathrm{E}[\boldsymbol{a}\boldsymbol{a}^\top] \succ \boldsymbol{0}, \quad \bar{\boldsymbol{b}} := \mathrm{E}[c\,\boldsymbol{a}], \quad \boldsymbol{s}^{\mathrm{LS}} := \bar{\boldsymbol{A}}^{-1}\bar{\boldsymbol{b}}, \quad \rho := c - \boldsymbol{a}^\top \boldsymbol{s}^{\mathrm{LS}}, \quad \kappa := \beta \max_i \|\boldsymbol{a}_i\|^2, \tag{6}$$

so $\boldsymbol{S}^{\mathrm{LS}} = \operatorname{argmin}_{\boldsymbol{S}} \mathrm{E}[(\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{k} - c)^2]$ is the least-squares memory and ρ the irreducible residual, with $\mathrm{E}[\rho\,\boldsymbol{a}] = \boldsymbol{0}$. Throughout, $0 < \beta \|\boldsymbol{a}_i\|^2 < 2$, and $q_\beta := 1 - \beta(2-\kappa)\lambda_{\min}(\bar{\boldsymbol{A}})$.

Theorem 1 (Unbiased mean and steady state)**.** *The expected iterate obeys the* exact *recursion* $\mathrm{E}[\boldsymbol{s}_{t+1}] = (\boldsymbol{I} - \beta\bar{\boldsymbol{A}})\mathrm{E}[\boldsymbol{s}_t] + \beta\bar{\boldsymbol{b}}$*, whose unique fixed point is* $\boldsymbol{s}^{\mathrm{LS}}$*, so* $\left\|\mathrm{E}[\boldsymbol{S}_t] - \boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}} \le r^t \left\|\boldsymbol{S}_0 - \boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}}$ *with* $r = \max_k \left|1 - \beta\lambda_k(\bar{\boldsymbol{A}})\right| < 1$*. Moreover the iterates form a time-homogeneous Markov chain on* $\mathbb{R}^{d_k d_v}$ *whose kernel contracts on laws with finite second moment; by the standard theory of Markov chains on general state spaces (Meyn & Tweedie, 1993) it has a* unique *stationary distribution* π_β*, the law of* $\boldsymbol{S}_t$ *converges to it geometrically from any initialization, and the stationary mean is* exactly $\mathrm{E}_{\pi_\beta}[\boldsymbol{S}] = \boldsymbol{S}^{\mathrm{LS}}$.

Theorem 1 combines Appendix Theorem A6 and Appendix Theorem A7 which have been proven in Appendix A2.

Theorem 2 (Memory norm bounds)**.** *With* $\lambda_{\min}, \lambda_{\max}$ *the extreme eigenvalues of* $\bar{\boldsymbol{A}}$*, the least-squares memory is sandwiched by the* explained *target energy,*

$$\frac{\mathrm{E}[c^2] - \mathrm{E}[\rho^2]}{\lambda_{\max}(\bar{\boldsymbol{A}})} \;\le\; \left\|\boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}}^2 \;\le\; \frac{\mathrm{E}[c^2] - \mathrm{E}[\rho^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})}, \tag{7}$$

and all its singular values obey the same upper bound; residual energy can only shrink it. In the non-expansive regime $\kappa \le 1$*, the steady state and — started at* $\boldsymbol{S}_0 = \boldsymbol{0}$ *— every iterate satisfies*

$$\mathrm{E}_{\pi_\beta}\left[\,\|\boldsymbol{S}\|_{\mathrm{F}}^2\,\right] \;\le\; \frac{\mathrm{E}[c^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})}, \qquad \mathrm{E}\left[\,\|\boldsymbol{S}_t\|_{\mathrm{F}}^2\,\right] \;\le\; \left(1 + q_\beta^{t/2}\right)^2 \frac{\mathrm{E}[c^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})} \;\le\; \frac{4\,\mathrm{E}[c^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})}, \tag{8}$$

the transient factor decays geometrically from 4 *to* 1.

Theorem 2 combines Appendix Theorem A8, Appendix Theorem A9, and Appendix Theorem A10 which are proven in Appendix A2. From the first step onward, the expected memory of the rule is bounded by a constant determined by the data distribution alone. More details and theoretical results are given in Appendix A2.

2.2 The KronLSTM architecture

In modern deep learning architectures for sequence modeling, the matrix update rules are stabilized and enhanced with gating mechanisms. The mLSTM (Beck et al., 2024) places the Hebbian write

(1) under a sigmoid forget gate and an exponential input gate, and Gated DeltaNet (Yang et al., 2025) decays the memory of (2) by a scalar before the correction,

$$\boldsymbol{S}_{t+1} \;=\; \mathrm{f}_t\,\boldsymbol{S}_t \;+\; \mathrm{i}_t\,\boldsymbol{v}_t\boldsymbol{k}_t^\top\,, \qquad \boldsymbol{S}_{t+1} \;=\; \mathrm{f}_t\,\boldsymbol{S}_t\left(\boldsymbol{I}-\beta_t\,\boldsymbol{k}_t\boldsymbol{k}_t^\top\right) \;+\; \beta_t\,\boldsymbol{v}_t\boldsymbol{k}_t^\top\,. \tag{9}$$

Kimi Delta Attention (Kimi Team et al., 2025) makes the decay (key) channel-wise, and Gated DeltaNet-2 (Hatamizadeh et al., 2026) replaces the single β_t by separate channel-wise erase and write gates. KronLSTM proceeds in the same way: the step size β_t and the target c_t of (3) become gates computed from the input; a forget gate decays the memory, and the rule is embedded in the projections, short convolutions, normalizations, and output gate of a standard sequence-mixer layer. All quantities below are per head, with key and value dimensions d_k and d_v; we suppress the head index and write $\boldsymbol{x}_t \in \mathbb{R}^d$ for the layer input at step t.

KronLSTM update. The forget gate is a diagonal matrix $\boldsymbol{F}_t$ with diagonal entries in $(0,1)$. It acts on the key dimension and hence decays the memory from the right, $\boldsymbol{S}_t\boldsymbol{F}_t$, retaining each key direction separately. We use target $\tilde{c}_t = c_t/\beta_t$ to decouple erase and write. The erase term of (3) is evaluated on the decayed memory $\boldsymbol{S}_t\boldsymbol{F}_t$. We obtain

$$\boldsymbol{S}_{t+1} \;=\; \underbrace{\boldsymbol{S}_t\boldsymbol{F}_t}_{\text{forget}} \;-\; \underbrace{\beta_t\big(\boldsymbol{v}_t^\top\boldsymbol{S}_t\boldsymbol{F}_t\,\boldsymbol{k}_t\big)\,\boldsymbol{v}_t\boldsymbol{k}_t^\top}_{\text{erase}} \;+\; \underbrace{c_t\,\boldsymbol{v}_t\boldsymbol{k}_t^\top}_{\text{write}}, \tag{10}$$

so that forgetting, erasing, and writing are controlled by three separate gates, $\boldsymbol{F}_t$, β_t and c_t. The standard coupled and the unit-target variant $c_t \equiv 1$ are compared in Appendix A3, Table A2.

Gates. The erase gate β_t and the write gate c_t are per-head scalars obtained from linear projections of the input followed by a sigmoid,

$$\beta_t = \sigma\big(\boldsymbol{w}_\beta^\top\boldsymbol{x}_t + b_\beta\big) \in (0,1), \qquad c_t = \sigma\big(\boldsymbol{w}_\mathrm{c}^\top\boldsymbol{x}_t + b_\mathrm{c}\big) \in (0,1), \tag{11}$$

both bounded by construction. The diagonal vector $\mathbf{f}_t$ of the forget gate $\boldsymbol{F}_t = \mathrm{diag}(\mathbf{f}_t)$ is obtained by raising a sigmoid gate to a learned positive per-head power, following the parameterization of Mamba (Gu & Dao, 2024) and Gated DeltaNet-2 (Hatamizadeh et al., 2026)

$$\mathbf{f}_t \;=\; \boldsymbol{\sigma}(-(\boldsymbol{s}_\mathrm{f})_t)^a \;=\; \big(1-\boldsymbol{\sigma}((\boldsymbol{s}_\mathrm{f})_t)\big)^a, \;\; (\boldsymbol{s}_\mathrm{f})_t = \boldsymbol{W}_\mathrm{f}\boldsymbol{x}_t + \boldsymbol{b}_\mathrm{f} \in \mathbb{R}^{d_k}, \;\; a = \exp(a_{\log}) \in \mathbb{R}_{>0}, \tag{12}$$

where $a_{\log}$ is a single learned scalar per head, broadcast over the d_k key dimensions, and the exponent is taken component-wise. The sigmoid sets retention separately for each key direction as a function of the input. At the same time, the exponent rescales the head's retention horizon by $1/a$, so that heads can operate at widely different timescales without high pre-activations. Since $a_{\log}$ is learned in log space, the baseline retention timescale spans orders of magnitude across heads, while within a head the per-direction variation comes entirely from the input. The read-out is gated by a sigmoid output gate $\mathbf{o}_t = \boldsymbol{\sigma}(\boldsymbol{W}_\mathrm{o}\boldsymbol{x}_t + \boldsymbol{b}_\mathrm{o}) \in (0,1)^{d_v}$.

KronLSTM block. Queries, keys, and values are linear projections of $\boldsymbol{x}_t$, each followed by a short causal depthwise convolution and a SiLU activation, and are then ℓ_2-normalized with $\boldsymbol{q}_t, \boldsymbol{k}_t \in \mathbb{R}^{d_k}$ and $\boldsymbol{v}_t \in \mathbb{R}^{d_v}$. The memory $\boldsymbol{S}_t \in \mathbb{R}^{d_v \times d_k}$ is updated by (10) and read with the query, $\boldsymbol{h}_t = \boldsymbol{S}_t\boldsymbol{q}_t \in \mathbb{R}^{d_v}$. The read-out passes through the sigmoid output gate and a per-head RMSNorm, and the concatenated heads are mixed by the output projection,

$$\boldsymbol{y}_t \;=\; \boldsymbol{W}_\mathrm{out}\,\mathrm{RMSNorm}\big(\mathbf{o}_t \odot \boldsymbol{h}_t\big).$$

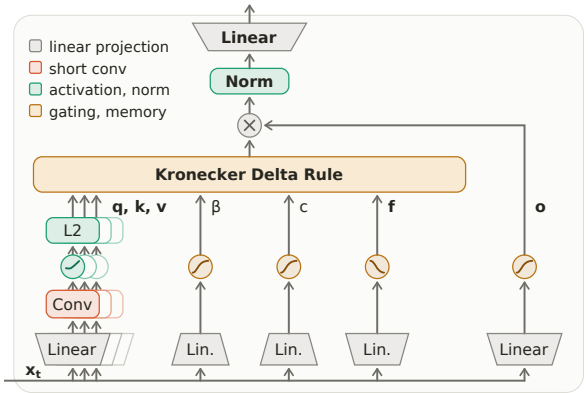

Figure 3: Architecture of the KronLSTM block.

With ℓ_2-normalized keys and values and a sigmoid β_t, we have $\beta_t\,\|\boldsymbol{k}_t\|^2\,\|\boldsymbol{v}_t\|^2 = \beta_t < 1$, so every

update is non-expansive and the memory operates in the regime of Theorem 2; the forget gate, with entries below one, only contracts it further. The architecture of the KronLSTM block is shown in Figure 3. It carries no state besides the memories $\boldsymbol{S}_t$ of its heads: there is no normalizer state and no stabilizer state as in the mLSTM. KronLSTM can serve as a drop-in replacement for the attention layer in a Transformer block, analogous to mLSTM and sLSTM in the xLSTM architecture.

3 Experimental Validation

We experimentally validate the KronLSTM architecture on different benchmarks with a focus on language modeling tasks. We compare KronLSTM with the following recurrent linear attention architectures: Gated DeltaNet (GDN) (Yang et al., 2025), Kimi Delta Attention (KDA) (Kimi Team et al., 2025), Mamba-3 (Lahoti et al., 2026), and Gated DeltaNet-2 (GDN-2) (Hatamizadeh et al., 2026). Recent deployments of linear attention in frontier models follow two hybrid-setting paradigms: either interleave the linear layers with a small number of softmax attention layers (NVIDIA, 2025; Merrill et al., 2026; Kimi Team et al., 2025) or use sliding-window attention (SWA) instead (Cabannes et al., 2026; Hatamizadeh et al., 2026). The advantage of the latter is keeping the memory footprint constant with respect to the context length. Therefore we evaluate every architecture in two settings: as a pure linear attention model, and as a hybrid with SWA layers. For the hybrid setting, we employ a window size of 2048 at a 5:1 ratio of linear attention to SWA layers. All models are pretrained on the FineWeb-Edu dataset using the Llama2 tokenizer (Touvron et al., 2023b). We train at a context of 8k, motivated by previous works (Guo et al., 2026; Beck et al., 2025a; 2026), where the throughput advantage of linear attention compared to the softmax attention emerges from 8k context onward. We evaluate the models within and extrapolating outside their context length (Section 3.2). The experiments in the following sections show 1.5B parameter models trained on 100B tokens. In Appendix A3, we provide a set of experiments at 400M model size.

3.1 Language Modeling

We assess the performance of KronLSTM compared to leading linear attention variants on standard language modeling benchmarks, concretely on five commonsense reasoning tasks (HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), ARC-easy and ARC-challenge (Clark et al., 2018), WinoGrande (Sakaguchi et al., 2021)) and on LAMBADA (Paperno et al., 2016), reporting accuracy on all six tasks, their average (Avg-6), and LAMBADA perplexity. Table 1 shows the results of both linear and hybrid models. Among all competitors, KronLSTM has the highest average accuracy over the six tasks and the lowest perplexity on LAMBADA. As a linear model, KronLSTM performs best on five out of six tasks and improves on average on the strongest competitor, KDA, by 1.6 accuracy points. The hybrid setting confirms the same performance advantage of KronLSTM, leading in four out of six tasks and showing the lowest perplexity on LAMBADA. Hybrid models with sliding-window attention change the averages by at most about one accuracy point. While Mamba-3 and GDN-2 benefit most from adding SWA, GDN and KDA change only minimally; KronLSTM remains robust, leading among all configurations.

3.2 Retrieval Capabilities

On long-context tasks, the Kronecker Delta Rule should improve the most over previous update rules, as interference grows with context length. We assess the performance of KronLSTM on two benchmarks with different context lengths.

The first benchmark is the real-world recall suite of Arora et al. (2024) (SWDE, SQuAD, FDA, TriviaQA, NQ, DROP) for 2k context. Results are given in Table 2. KronLSTM has the best average accuracy, and it is the best method in three out of six tasks in both pure linear attention models and hybrid models. In general, hybrid models using SWA gain five to nine accuracy points.

The second benchmark is the RULER needle-in-a-haystack (NIAH) tasks (Hsieh et al., 2024) where models are evaluated at 4k, 8k and 16k context length, with 16k assessing the extrapolation capabilities of the models. Results are presented in Table 3. Among all linear attention models, KronLSTM is the best or second-best method across all sequence lengths, except for task S1. KronLSTM has the highest accuracy on S3 and MQ tasks across all sequence lengths. Among all hybrid models,

Table 1: Downstream performance of 1.5B/100B models. Avg-6 is the mean accuracy over the five commonsense reasoning tasks and LAMBADA. Best per column in **bold**, second best underlined, separately for linear and hybrid models.

Model	Hella	PIQA	ARC-e	ARC-c	Wino	LMB acc	Avg-6 ↑	LMB ppl ↓
Linear models								
GDN	57.00	**73.61**	70.54	41.04	57.22	46.94	57.73	12.575
Mamba-3	57.30	73.01	70.29	40.19	57.46	45.64	57.31	12.946
KDA	57.48	73.39	71.84	41.55	57.14	47.35	58.12	12.669
GDN-2	57.45	72.96	71.13	41.89	57.30	46.92	57.94	12.542
KronLSTM	**59.19**	73.45	**74.03**	**43.60**	**60.06**	**47.97**	**59.72**	**11.716**
Hybrid models								
GDN	58.46	72.74	70.54	42.58	56.99	46.63	57.99	12.948
Mamba-3	58.14	**74.10**	72.22	43.09	57.30	46.38	58.54	13.014
KDA	57.93	72.74	71.04	41.21	56.99	47.22	57.85	12.600
GDN-2	58.35	72.91	**72.64**	42.75	58.88	47.18	58.79	11.848
KronLSTM	**59.80**	73.78	71.93	**44.11**	**58.96**	**48.77**	**59.56**	**11.178**

Table 2: Results for 1.5B models on the real-world recall suite, given in accuracies for 2k context length. Best per column in **bold**, second best underlined.

Model	SWDE	SQD	FDA	TQA	NQ	DROP	Avg-6 ↑
Linear models							
GDN	28.62	37.33	14.79	60.49	**16.47**	**21.18**	29.81
Mamba-3	27.54	35.86	11.43	59.18	15.39	21.03	28.41
KDA	25.83	37.30	12.98	**62.14**	16.31	20.46	29.17
GDN-2	29.70	36.29	14.43	60.07	15.90	19.93	29.39
KronLSTM	**30.87**	**39.28**	**15.34**	61.61	15.46	19.31	**30.31**
Hybrid models							
GDN	43.20	40.58	23.59	61.91	17.90	**24.44**	35.27
Mamba-3	42.84	**44.67**	29.22	63.92	17.26	23.86	36.96
KDA	42.30	42.79	**30.04**	61.02	18.12	21.80	36.01
GDN-2	43.38	42.19	28.31	60.84	17.36	23.05	35.85
KronLSTM	**46.89**	43.57	27.86	**64.51**	**18.82**	22.76	**37.40**

KronLSTM achieves the highest accuracy (averaged over all lengths) on all tasks and has the highest accuracy at 16k on each task, confirming its strong long context capabilities.

The improvement of KronLSTM over competitors like Gated DeltaNet is consistent with the interference analysis of Property A9. On S2 and S3, whose needles are unique sentences embedded in natural language text, KronLSTM outperforms GDN except for S2 at 16k. For hybrid models, KronLSTM outranks GDN-2 and Mamba-3 by 15.2 accuracy points on S2 at 4k. KronLSTM's margin increases on MQ and MV with sequence length. On MQ, KronLSTM's lead distance grows from 2.4 to 4.1 accuracy points when increasing the context from 4k to 16k. On MV, KronLSTM is 3.5 accuracy points behind the best method, GDN, for 4k context, while KronLSTM is the leading method for 8k and 16k, with a 1.1 and 3.8 accuracy point lead, respectively. For longer contexts, the memory updates of GDN more frequently interfere with already stored content and perturb it, resulting in more degraded performance at 16k. In contrast, the Kronecker Delta Rule is less invasive and performs better at longer contexts. The exception is S1, where KronLSTM underperforms. S1 is the most artificial task with frequently repeated noise sentences, which is unrealistic for real-world text. We attribute the drop in performance of the KronLSTM at this task to residual accumulation for recurring keys. Hybrid models with SWA have a considerably smaller S1 accuracy, except KronLSTM. The hybrid models containing the baseline linear attention layers reach at most 51.8, 26.6, and 12.6 accuracy for 4k, 8k, and 16k context length, respectively. These accuracies closely match

Table 3: Results of 1.5B models on the NIAH tasks for 8k pretraining context and 4k, 8k, and 16k evaluation context. S1–S3: S-NIAH-1/2/3; MK: MK-NIAH-1; MQ: MQ-NIAH; MV: MV-NIAH. Best per task and length in **bold**, second best underlined, separately for linear and hybrid models.

		Linear models						*Hybrid models*					
Model	Len.	S1	S2	S3	MK	MQ	MV	S1	S2	S3	MK	MQ	MV
GDN	4k	**100.0**	96.0	54.8	28.0	44.8	**46.8**	51.8	52.6	21.8	31.0	**40.6**	38.3
	8k	**100.0**	45.6	13.0	**24.0**	21.2	23.5	26.6	31.2	**22.2**	24.6	17.8	18.8
	16k	99.2	**17.8**	5.4	**17.8**	8.6	13.8	12.6	16.8	3.6	18.4	6.9	8.8
Mamba-3	4k	92.0	68.2	26.4	26.8	20.0	18.4	51.6	56.0	30.2	34.4	19.1	21.3
	8k	59.0	27.0	5.4	20.6	4.2	3.6	26.6	31.2	15.0	24.2	10.7	22.4
	16k	28.0	13.6	3.0	13.4	5.3	4.5	0.0	0.0	0.0	0.0	0.0	0.0
KDA	4k	99.8	89.4	12.0	20.8	18.4	19.1	51.8	55.8	31.4	44.8	33.0	42.5
	8k	99.0	35.4	11.6	15.8	13.3	13.3	26.6	31.2	5.0	22.8	16.3	**22.8**
	16k	93.6	14.2	1.6	11.8	3.1	5.7	12.6	16.8	4.4	17.2	3.1	11.7
GDN-2	4k	**100.0**	72.8	50.2	26.6	28.4	38.1	51.8	56.0	45.6	**51.8**	33.2	40.5
	8k	99.8	44.6	15.4	17.6	13.4	21.4	26.6	31.2	18.8	25.4	**23.2**	22.7
	16k	**99.6**	8.2	4.2	7.2	2.7	5.1	7.4	16.6	6.0	17.6	3.0	12.8
KronLSTM	4k	66.2	**98.2**	**73.4**	**28.2**	**47.2**	43.3	**64.8**	**71.2**	**51.2**	51.0	38.8	**44.5**
	8k	29.4	**60.4**	**20.4**	23.4	**24.8**	**24.6**	**33.6**	**35.0**	17.4	**28.6**	23.1	21.8
	16k	11.6	16.0	**11.8**	16.0	**12.7**	**17.6**	**16.2**	**17.4**	**7.0**	**21.0**	**12.2**	**13.2**

the fraction of the context covered by the attention window, indicating that the needle is primarily retrieved when it lies within the window, with Mamba-3 dropping to zero at 16k. The results suggest limited reliance on the recurrent architecture's memory. KronLSTM exceeds these accuracies on average across all context lengths and achieves the best performance across all six tasks at 16k. Section 3.3 examines how these methods perform after context extension.

3.3 Long Context Extension

In applications, language models do not extrapolate to longer context but are trained on them, typically by continued pretraining at increased context length (Fu et al., 2024; Gao et al., 2024). This stage unlocks retrieval far beyond the original pretraining length (Fu et al., 2024). Linear attention models have no positional encoding to adapt; rather, they must learn to keep more associations over longer time spans. We therefore branch the 1.5B linear KronLSTM and GDN-2 runs at 75B tokens and train them for the remaining 25B tokens at a sequence length of 32k on a long-document mixture derived from ProLong-64K (Gao et al., 2024), so that the context-length-extended models are token-matched to the 100B models of the previous sections. GDN-2 serves as a reference model in this experiment since it also decouples erase and write. Further details on training are described in Appendix A4.1.Table 4 shows the commonsense reasoning, LAMBADA, and recall results of the extended models. Both models benefit from length extension on the reasoning and recall tasks, and perplexity improves by about 4 perplexity points. KronLSTM keeps its lead: it has the higher average accuracy and the lower perplexity, and it is better on five of the six recall tasks.

Table 4: Long-context (32k) extension results for GDN-2 and KronLSTM models. Avg-6 is the mean accuracy of the five reasoning tasks plus LAMBADA acc; the remaining columns give accuracy on the real-world recall suite evaluated at 2048 tokens. The best per column is in **bold**.

Model	Avg-6 ↑	LMB ppl ↓	SWDE	SQD	FDA	TQA	NQ	DROP	Recall avg ↑
GDN-2	58.75	8.364	29.61	36.39	**19.60**	60.19	14.54	19.36	29.95
KronLSTM	**60.19**	**7.656**	**31.68**	**39.88**	18.15	**60.37**	**15.36**	**21.03**	**31.08**

Results for context extension at the RULER tasks are given in Figure 4 and Table A3. As expected, KronLSTM shows its advantage by extending the training context length. While both architectures perform strongly on single-needle retrieval at shorter contexts, KronLSTM increasingly outperforms GDN-2 on the more challenging NIAH tasks, ranging from 8k to 32k. At 16k, this advantage is

particularly pronounced on S-3 and MV, with KronLSTM leading on most tasks except S1, where GDN-2 remains stronger. At 32k, KronLSTM maintains its lead across all five tasks, although both architectures experience a substantial drop in retrieval accuracy compared to 16k. The gap between the two models widens with haystack size, as predicted by the interference analysis in Section 2.1: longer contexts accumulate more interference, and KronLSTM's update rule suffers less of it.

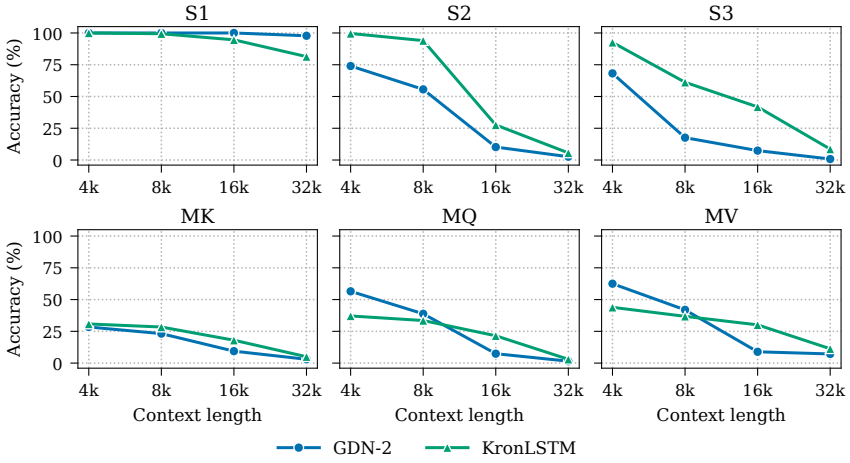

Figure 4: GDN-2 and KronLSTM results at RULER NIAH for 32k context extension.

3.4 State Tracking Capabilities

KronLSTM can perform state tracking by setting $\beta_t = 2\sigma(\boldsymbol{w}_\beta^\top \boldsymbol{x}_t + b_\beta) \in (0, 2)$, as described in Grazzi et al. (2025) for DeltaNet. We evaluate KronLSTM's state tracking capabilities on parity, S_3 permutation, and modulo arithmetic. We train up to sequence length 128 and evaluate up to 1024. Figure 5 shows that KronLSTM has similar state tracking capabilities as Gated DeltaNet on the parity and S_3 tasks, both ahead of Mamba-3, which wins on modulo arithmetic.

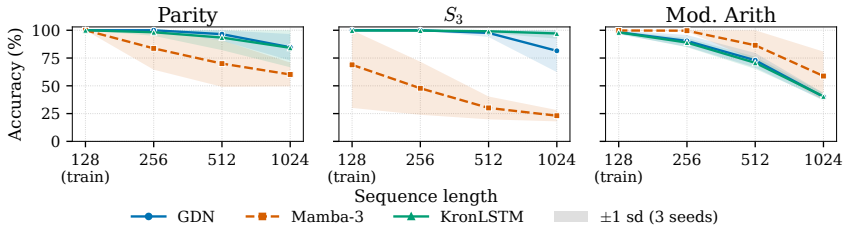

Figure 5: State tracking results for Gated DeltaNet (GDN), Mamba-3, and KronLSTM.

4 Conclusion

We have introduced the Kronecker Delta Rule, an update rule for associative matrix memories that corrects a scalar association along a single direction, and KronLSTM, which embeds this rule in a gated linear attention mixer. We proved that the memory dynamics converge to a unique steady state and that the expected memory norm is bounded at every step. KronLSTM has the best average accuracy on commonsense reasoning and real-world recall, both as a linear model and as a hybrid. Averaged over the context lengths, it is the best or second-best linear model on every natural-text RULER task, and on S1 it is the only one that retrieves needles beyond the sliding attention window in hybrid models. After extending the context to 32k, KronLSTM outperforms GDN-2 on all natural-text retrieval tasks from 16k onward. Future work includes scaling up KronLSTM and training hybrid models containing full softmax attention layers.

AI USE STATEMENT

The Kronecker Delta Rule and the KronLSTM architecture originated with the authors without AI assistance. LLMs were used to assist in developing the theoretical results and in implementing the method. Claude Fable 5 and Claude Fable 5.1 assisted in deriving the properties of the Kronecker Delta Rule and the convergence and memory norm bounds of Section 2.1 and Appendix A2; every statement and proof was checked independently by the authors, who take full responsibility for their correctness. Claude Fable 5 was further used to examine candidate formulations of the Kronecker Delta Rule, in particular the coupled, decoupled, and unit-target write variants, for their viability; the choice of the form in (10) was made by the authors on the basis of ablation studies such as in Appendix A3. Claude Fable 5 and Claude Fable 5.1 were used to optimize the chunkwise-parallel kernels of KronLSTM; the resulting kernels were numerically verified against naive reference implementations written by the authors. We did not use generative AI tools to interpret results or to design experiments. We have not used generative AI tools to generate synthetic data or to propose hypotheses, and translation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used Claude Fable 5 and Claude Opus 5 to draft parts of the manuscript and to edit it for readability. All AI-drafted text was revised by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work proposes a memory update rule and a sequence-mixer architecture for language modeling and evaluates them on public benchmarks. It involves no human subjects, no collection of new data, and no release of a dataset. All models were pretrained on FineWeb-Edu, a publicly available corpus, and inherit whatever biases it contains; we make no claims that KronLSTM mitigates or amplifies them relative to the baselines trained on the same data. The general considerations that apply to language model research, including potential misuse, environmental cost of pretraining, and encoded biases, apply to this work as well and are not specific to the proposed method.

REPRODUCIBILITY STATEMENT

The supplementary material contains the chunkwise-parallel kernels of the Kronecker Delta Rule and a KronLSTM layer, both compatible with the public `flash-linear-attention` (FLA) library (Yang & Zhang, 2026), so that KronLSTM can be imported into the FLA repository and used as a drop-in sequence mixer in existing training frameworks. All models in this work were pretrained with `lm-engine` (Mishra, 2024), a public training repository into which the supplied layer can be integrated. The complete set of hyperparameters, including architecture configuration, optimizer settings, learning-rate schedule, and hardware setup, is documented in Appendix A4, and the long-context extension recipe in Appendix A4.1. Pretraining uses the FineWeb-Edu corpus with the Llama2 tokenizer and, for context extension, a mixture derived from ProLong-64K; all evaluation suites (commonsense reasoning and LAMBADA in Section 3.1, the real-world recall suite of Arora et al. (2024) and RULER NIAH in Section 3.2) are publicly available. For the theoretical results, all assumptions are stated in Section 2.1 and complete proofs are given in Appendix A2. Together, these components provide what is needed to reproduce our training runs and evaluations.

REFERENCES

Qwen Team. Qwen Technical Report, 2023. arXiv:2309.16609 [cs].

B. Alkin, M. Beck, K. Pöppel, S. Hochreiter, and J. Brandstetter. Vision-LSTM: xLSTM as Generic Vision Backbone. In *International Conference on Learning Representations*, volume 13, Singapore, Singapore, 2025. OpenReview. URL https://openreview.net/forum?id=SiH7DwNKZZ.

J. A. Anderson. A simple neural network generating an interactive memory. *Mathematical Biosciences*, 14(3-4):197–220, 1972.

S. Arora, S. Eyuboglu, M. Zhang, A. Timalsina, S. Alberti, D. Zinsley, J. Zou, A. Rudra, and C. Ré. Simple linear attention language models balance the recall-throughput tradeoff. In

Proceedings of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 1763–1840, Vienna, Austria, 2024. PMLR. URL https://proceedings.mlr.press/v235/arora24a.html.

M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter. xLSTM: Extended Long Short-Term Memory. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107547–107603, Vancouver, BC, Canada, 2024. Curran Associates, Inc. doi: 10.52202/079017-3417. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c2ce2f2701c10a2b2f2ea0bfa43cfaa3-Abstract-Conference.html.

M. Beck, K. Pöppel, P. Lippe, and S. Hochreiter. Tiled flash linear attention: More efficient linear RNN and xLSTM kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, San Diego, CA, USA, 2025a. Curran Associates, Inc. URL https://openreview.net/forum?id=b6H64u6TqI.

M. Beck, K. Pöppel, P. Lippe, R. Kurle, P. M. Blies, G. Klambauer, S. Böck, and S. Hochreiter. xLSTM 7B: A recurrent LLM for fast and efficient inference. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3335–3357, Vancouver, BC, Canada, 2025b. PMLR. URL https://proceedings.mlr.press/v267/beck25b.html.

M. Beck, K. Schweighofer, S. Böck, S. Lehner, and S. Hochreiter. xLSTM scaling laws: Competitive performance with linear time-complexity. In *International Conference on Learning Representations*. OpenReview, 2026. doi: 10.48550/arXiv.2510.02228. URL https://arxiv.org/abs/2510.02228. arXiv:2510.02228 [cs].

A. Behrouz, P. Zhong, and V. Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2501.00663.

A. Behrouz, M. Razaviyayn, P. Zhong, and V. Mirrokni. It's all connected: A journey through test-time memorization, attentional bias, retention, and online optimization. *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2504.13173.

Y. Bisk, R. Zellers, J. Gao, Y. Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

R. Bogacz and M. W. Brown. An anti-Hebbian model of familiarity discrimination in the perirhinal cortex. *Neurocomputing*, 52–54:1–6, 2003. doi: 10.1016/S0925-2312(02)00738-5.

L. Cabannes, M. Beck, M. Lomeli, G. Szilvasy, M. Douze, J. Copet, P.-E. Mazaré, G. Synnaeve, and H. Jégou. Short window attention enables long-term memorization. In *International Conference on Learning Representations*, volume 14, pp. 79759–79774, 2026. URL https://openreview.net/forum?id=btgVfhudI1.

A. Cegielski. *Iterative Methods for Fixed Point Problems in Hilbert Spaces*, volume 2057 of *Lecture Notes in Mathematics*. Springer, 2012.

X. Chen. The Kaczmarz algorithm, row action methods, and statistical learning algorithms. In *Contemporary Mathematics*, volume 706, pp. 115–127, 2018. ISBN 9781470436193. doi: 10.1090/conm/706/14216.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

T. Dao and A. Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Forty-first International Conference on Machine Learning*, volume 235, pp. 10041–10071, Vienna, Austria, 2024. PMLR. URL https://openreview.net/forum?id=ztn8FCR1td.

Y. Fu, R. Panda, X. Niu, X. Yue, H. Hajishirzi, Y. Kim, and H. Peng. Data engineering for scaling language models to 128K context. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14125–14134. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/fu24d.html.

T. Gao, A. Wettig, H. Yen, and D. Chen. How to train long-context language models (effectively). *ArXiv:2410.02660*, 2024. URL https://arxiv.org/abs/2410.02660.

R. Grazzi, J. Siems, A. Zela, J. K. H. Franke, F. Hutter, and M. Pontil. Unlocking State-Tracking in Linear RNNs Through Negative Eigenvalues. In *International Conference on Learning Representations*, volume 13, Singapore, Singapore, 2025. URL https://openreview.net/forum?id=UvTo3tVBk2.

A. Gu and T. Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *Conference on Language Modeling*, volume 1, Philadelphia, PA, USA, 2024. OpenReview. URL https://openreview.net/forum?id=tEYskw1VY2.

H. Guo, S. Yang, T. Goel, E. P. Xing, T. Dao, and Y. Kim. Log-Linear Attention. In *International Conference on Learning Representations*. OpenReview, 2026. URL https://openreview.net/forum?id=mOJgZWkXKW.

A. Hatamizadeh, Y. Choi, and J. Kautz. Gated DeltaNet-2: Decoupling erase and write in linear attention, 2026. URL https://arxiv.org/abs/2605.22791.

S. Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München, 1991.

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1997.9.8.1735. URL https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735.

S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In J. Kolen and S. Kremer (eds.), *A Field Guide to Dynamical Recurrent Networks*. IEEE Press, 2000.

P.J. Hoedt, F. Kratzert, D. Klotz, C. Halmich, M. Holzleitner, G. Nearing, S. Hochreiter, and G. Klambauer. Mc-lstm: Mass-conserving lstm. In *International Conference on Machine Learning*, pp. 4275–4286. PMLR, 2021.

C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *Conference on Language Modeling*, volume 1, Philadelphia, PA, USA, 2024. OpenReview. URL https://openreview.net/pdf?id=kIoBbc76Sy.

S. Kaczmarz. Angenäherte Auflösung von Systemen linearer Gleichungen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres, Classe des Sciences Mathématiques et Naturelles, Série A*, 35:355–357, 1937.

A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In E. H. Daumé III and A. Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165, virtual, 2020. PMLR. URL http://proceedings.mlr.press/v119/katharopoulos20a.html.

Kimi Team, Y. Zhang, Z. Lin, X. Yao, J. Hu, F. Meng, C. Liu, X. Men, S. Yang, Z. Li, W. Li, E. Lu, W. Liu, Y. Chen, W. Xu, L. Yu, Y. Wang, Y. Fan, L. Zhong, E. Yuan, D. Zhang, Y. Zhang, T. Y. Liu, H. Wang, S. Fang, W. He, S. Liu, Y. Li, J. Su, J. Qiu, B. Pang, J. Yan, Z. Jiang, W. Huang, B. Yin, J. You, C. Wei, Z. Wang, C. Hong, Y. Chen, G. Chen, Y. Wang, H. Zheng, F. Wang, Y. Liu, M. Dong, Z. Zhang, S. Pan, W. Wu, Y. Wu, L. Guan, J. Tao, G. Fu, X. Xu, Y. Wang, G. Lai, Y. Wu, X. Zhou, Z. Yang, and Y. Du. Kimi linear: An expressive, efficient attention architecture, 2025. URL https://arxiv.org/abs/2510.26692. arXiv:2510.26692 [cs].

T. Kohonen. Correlation matrix memories. *IEEE Transactions on Computers*, C-21(4):353–359, 1972.

A. Lahoti, K. Y. Li, B. Chen, C. Wang, A. Bick, J. Z. Kolter, T. Dao, and A. Gu. Mamba-3: Improved sequence modeling using state space principles. In *The Fourteenth International Conference on Learning Representations*, volume 14, pp. 86173–86202, Rio de Janeiro, Brazil, 2026. OpenReview. URL https://openreview.net/forum?id=HwCvaJOiCj.

B. Liu, R. Wang, L. Wu, Y. Feng, P. Stone, and Q. Liu. Longhorn: State space models are amortized online learners. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2407.14207.

J. Liu, S. Elflein, O. Litany, Z. Gojcic, and R. Li. Test-time training with KV binding is secretly linear attention. *Forty-third International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.21204.

W. Merrill, Y. Li, T. Romero, A. Svete, C. Costello, P. Dasigi, D. Groeneveld, D. Heineman, B. Kuehl, N. Lambert, C. Li, K. Lo, S. Malik, D. J. Matusz, B. Minixhofer, J. Morrison, L. Soldaini, F. Timbers, P. Walsh, N. A. Smith, H. Hajishirzi, and A. Sabharwal. Olmo Hybrid: From Theory to Practice and Back, 2026. URL http://arxiv.org/abs/2604.03444. arXiv:2604.03444 [cs].

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, 1993. 2nd edition: Cambridge University Press, 2009.

M. Mishra. LM Engine: A Hyper-Optimized Library for Pretraining and Finetuning, 2024. URL https://github.com/open-lm-engine/lm-engine.

NVIDIA. Nemotron 3 Nano: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning, 2025. URL http://arxiv.org/abs/2512.20848. arXiv:2512.20848 [cs].

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card, 2025. URL https://arxiv.org/abs/2508.10925. arXiv:2508.10925 [cs].

D. Paperno, G. Kruszewski, A. Lazaridou, N. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, 2016. doi: 10.18653/v1/P16-1144.

B. Peng, D. Goldstein, Q. Anthony, A. Albalak, E. Alcaide, S. Biderman, E. Cheah, X. Du, T. Ferdinan, H. Hou, P. Kazienko, K. K. GV, J. Kocoń, B. Koptyra, S. Krishna, R. McClelland Jr., J. Lin, N. Muennighoff, F. Obeid, A. Saito, G. Song, H. Tu, C. Wirawan, S. Woźniak, R. Zhang, B. Zhao, Q. Zhao, P. Zhou, J. Zhu, and R.-J. Zhu. Eagle and Finch: RWKV with matrix-valued states and dynamic recurrence, 2024. URL https://arxiv.org/abs/2404.05892.

B. Peng, R. Zhang, D. Goldstein, E. Alcaide, X. Du, H. Hou, J. Lin, J. Liu, J. Lu, W. Merrill, G. Song, K. Tan, S. Utpala, N. Wilce, J. S. Wind, T. Wu, D. Wuttke, and C. Zhou-Zheng. RWKV-7 Goose with expressive dynamic state evolution. *arXiv:2503.14456*, 2025. URL https://arxiv.org/abs/2503.14456.

D. Prados and S. Kak. Neural network capacity using delta rule. *Electronics Letters*, 25(3):197–199, 1989.

K. Sakaguchi, R. Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

I. Schlag, K. Irie, and J. Schmidhuber. Linear transformers are secretly fast weight programmers. In M. Meila and T. Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 9355–9366, virtual, 2021. PMLR. URL https://proceedings.mlr.press/v139/schlag21a.html.

J. Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

N. Schmidinger, L. Schneckenreiter, P. Seidl, J. Schimunek, P.-J. Hoedt, J. Brandstetter, A. Mayr, S. Luukkonen, S. Hochreiter, and G. Klambauer. Bio-xLSTM: Generative modeling, representation and in-context learning of biological and chemical sequences. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=IjbXZdugdj.

T. Schmied, T. Adler, V. P. Patil, M. Beck, K. Pöppel, J. Brandstetter, G. Klambauer, R. Pascanu, and S. Hochreiter. A Large Recurrent Action Model: xLSTM enables Fast Inference for Robotics Tasks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53343–53387, Vancouver, BC, Canada, 2025. PMLR. URL https://proceedings.mlr.press/v267/schmied25a.html.

J. Siems, T. Carstensen, A. Zela, F. Hutter, M. Pontil, and R. Grazzi. DeltaProduct: Improving state-tracking in linear RNNs via Householder products. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, San Diego, CA, USA, 2025. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e1ea2520fbf9cd1600b287dde67e0a3c-Abstract-Conference.html.

T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009. URL https://link.springer.com/article/10.1007/s00041-008-9030-4.

Y. Sun, X. Li, K. Dalal, J. Xu, A. Vikram, G. Zhang, Y. Dubois, X. Chen, X. Wang, S. Koyejo, T. Hashimoto, and C. Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. *arXiv:2407.04620*, 2024. URL https://arxiv.org/abs/2407.04620.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971, 2023a. URL https://arxiv.org/abs/2302.13971.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL https://arxiv.org/abs/2307.09288.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Aidan N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, Long Beach, CA, USA, 2017. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9.

J. von Oswald, M. Schlegel, A. Meulemans, S. Kobayashi, E. Niklasson, N. Zucchet, N. Scherrer, N. Miller, M. Sandler, B. Agüera y Arcas, M. Vladymyrov, R. Pascanu, and J. Sacramento. Uncovering mesa-optimization algorithms in transformers. *arXiv:2309.05858*, 2023. URL https://arxiv.org/abs/2309.05858.

B. Widrow and M. E. Hoff. Adaptive switching circuits. In *1960 IRE WESCON Convention Record, Part 4*, pp. 96–104, 1960.

S. Yang and Y. Zhang. FLA: A Triton-Based Library for Hardware-Efficient Implementations of Linear Attention Mechanism, 2026. URL https://github.com/fla-org/flash-linear-attention.

S. Yang, B. Wang, Y. Zhang, Y. Shen, and Y. Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 115491–115522, Vancouver, BC, Canada, 2024. Curran Associates, Inc. doi: 10.52202/079017-3668. URL https://arxiv.org/abs/2406.06484.

S. Yang, J. Kautz, and A. Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=r8H7xhYPwz.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019. URL https://aclanthology.org/P19-1472/.

CONTENTS

1 Introduction **1**

2 KronLSTM **2**
2.1 Kronecker Delta Rule . . . 3
2.2 The KronLSTM architecture . . . 4

3 Experimental Validation **6**
3.1 Language Modeling . . . 6
3.2 Retrieval Capabilities . . . 6
3.3 Long Context Extension . . . 8
3.4 State Tracking Capabilities . . . 9

4 Conclusion **9**

Appendices **17**

A1 Related Work **17**
A1.1 Long Short-Term Memory . . . 17
A1.2 The delta-rule linear attention family . . . 17
A1.3 The online-learning perspective . . . 17
A1.4 Classical update and learning rules . . . 18
A1.5 Relation to cognitive and neuroscience . . . 18

A2 Kronecker Delta Rule: Setting, Name, Properties **18**
A2.1 Setup and notation. . . . 18
A2.2 Name Kronecker Delta Rule . . . 18
A2.3 Properties . . . 19
A2.4 Minimally Invasive Memory Revision . . . 22
A2.4.1 The admissible set . . . 22
A2.4.2 Minimality in norm . . . 23
A2.4.3 Minimality of support: the correction is surgical . . . 24
A2.4.4 Minimality of interference . . . 24
A2.4.5 Relaxed steps: the correction never moves away from a correct memory . . . 24
A2.4.6 Greedy minimality is global minimality . . . 25
A2.4.7 Less invasive than DeltaNet . . . 25
A2.5 Convergence of the KronLSTM memory . . . 25
A2.5.1 Overview . . . 25
A2.5.2 Setting and notation . . . 26
A2.5.3 Theorems on expected and steady state solution . . . 26
A2.5.4 Proofs of the theorems . . . 27
A2.5.5 Remarks . . . 32
A2.6 Norm Bounds for the KronLSTM with Steady-State Solution . . . 33
A2.6.1 Overview . . . 33
A2.6.2 Setting and nomenclature . . . 33
A2.6.3 An exact energy identity . . . 33
A2.6.4 The Norm Bound . . . 33
A2.6.5 Tightness and special cases . . . 34
A2.6.6 Bound on the memory norm for steady-state and all updates . . . 34

A3 Further Results **37**
A3.1 400M language models . . . 37
A3.2 Ablation on different formulations of KronLSTM . . . 37

A4 Pretraining Details **37**
A4.1 Long-context extension . . . 38
A4.2 State Tracking Experiments . . . 39

Appendices

A1 Related Work

To the best of our knowledge, no existing work uses *exactly* this update that is a scalar bilinear read-out $\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{k}$ regressed onto the constant target 1. The established delta-rule linear attention family uses the vector read-out $\boldsymbol{S}\boldsymbol{k} \to \boldsymbol{v}$. The rule studied here is therefore the *bilinear-readout, scalar-target* member of the delta-rule. The constraints are $\left\langle \boldsymbol{v}\boldsymbol{k}^\top, \boldsymbol{S} \right\rangle_{\mathrm{F}} = 1$ rather than $\boldsymbol{S}\boldsymbol{k} = \boldsymbol{v}$.

A1.1 Long Short-Term Memory

KronLSTM belongs to the xLSTM (Beck et al., 2024) family that revisits LSTM (Hochreiter, 1991; Hochreiter & Schmidhuber, 1997) for the era of large language models (LLMs) with two modifications: exponential gating and a redesigned memory. mLSTM replaces the scalar cell state by a matrix memory and drops memory mixing, making the recurrence fully parallelizable over the sequence. mLSTM enables recurrent LLMs with fast, constant-memory inference at scale (Beck et al., 2025a), and Beck et al. (2026) establish scaling behavior competitive with Transformers at linear time complexity. xLSTM has also been applied to language modeling (Beck et al., 2025b), vision (Alkin et al., 2025), biological and chemical sequences (Schmidinger et al., 2025), and robotics (Schmied et al., 2025).

The matrix memory of mLSTM can be viewed as a classical correlation-based associative memory (Kohonen, 1972; Anderson, 1972) embedded within LSTM-style gating. In addition to providing an efficient mechanism for storing and retrieving associations, the exponential gates allow mLSTM to revise storage decisions. This distinguishes mLSTM from purely accumulative linear attention mechanisms. KronLSTM follows this line of research by retaining the matrix memory and the ability to revise storage decisions, while introducing a different memory-update rule.

A central challenge of recurrent architectures with growing memory is maintaining stable memory dynamics over long sequences. The original LSTM uses a bounded squashing function for the exposed cell state, while xLSTM introduces a normalizer state and a stabilizer state to control the scale of its matrix memory and retrievals. However, normalization based on ratios of large quantities can lead to numerical and optimization difficulties, particularly in large networks (Beck et al., 2025b). MC-LSTM (Hoedt et al., 2021) instead redistributes its cell state with column-stochastic matrices, which conserve mass and never amplify the state. These considerations motivate alternative matrix-memory update rules that retain memory revision while providing stable memory dynamics. The Kronecker Delta Rule addresses this issue and provides convergence to a steady-state distribution with a norm-bounded memory in expectation. Furthermore, KronLSTM yields advantageous gradient flow over time compared with mLSTM (Hochreiter et al., 2000).

A1.2 The delta-rule linear attention family

The direct ancestor is the fast weight programmer perspective of Schlag et al. (2021), who introduced the delta-rule write $\boldsymbol{S}_{t+1} = \boldsymbol{S}_t - \beta\left(\boldsymbol{S}_t\boldsymbol{k}_t - \boldsymbol{v}_t\right)\boldsymbol{k}_t^\top$ into linear attention (Katharopoulos et al., 2020), building on the fast-weight programmers of the early 1990s (Schmidhuber, 1992). Yang et al. (2024) (DeltaNet) made this rule hardware-efficient by exploiting the WY representation of the resulting products of generalized Householder matrices. This is exactly the product structure the bilinear rule inherits, with the Kronecker product $\boldsymbol{k}_t \otimes \boldsymbol{v}_t$ taking the role of $\boldsymbol{k}_t$. Subsequent extensions include Gated DeltaNet (Yang et al., 2025), which adds a scalar forget gate $\alpha_t\boldsymbol{S}_t$; RWKV-7 (Peng et al., 2025), which employs a generalized delta rule with a decoupled in-context learning rate; and DeltaProduct (Siems et al., 2025), which applies multiple Householder factors per token to improve state tracking. This is directly relevant to the negative-eigenvalue regime $\beta\left\|\boldsymbol{a}_t\right\|^2 \in (1, 2)$ of the transition factors $\boldsymbol{I} - \beta\boldsymbol{a}_t\boldsymbol{a}_t^\top$, which allows state-tracking.

A1.3 The online-learning perspective

Longhorn (Liu et al., 2025) derives state-space recurrences explicitly as closed-form solutions of per-token online learning objectives; the bilinear rule fits this template with the objective $\frac{1}{2}\left(\boldsymbol{v}^\top\boldsymbol{S}\boldsymbol{k} - 1\right)^2$. Test-time-training (TTT) layers (Sun et al., 2024) and Titans (Behrouz et al., 2025) treat the memory as a model trained by gradient descent at inference time. Most relevant as a unifying framework is the Miras taxonomy of Behrouz et al. (2026), which classifies memory architectures by their attentional-bias objective, retention mechanism, and optimizer. The new bilinear rule occupies a new place of that taxonomy: it is a *bilinear* attentional bias with constant scalar target. A recent

related observation is that test-time training with key–value binding is secretly linear attention (Liu et al., 2026). The mesa-optimization line (von Oswald et al., 2023) provides the complementary least-squares-solver perspective on such layers.

A1.4 Classical update and learning rules

When vectorized, the rule is the Widrow–Hoff LMS / delta rule (Widrow & Hoff, 1960) applied to the Kronecker product input $\boldsymbol{a}_t = \boldsymbol{k}_t \otimes \boldsymbol{v}_t$ with target 1. The stability range $\beta \|\boldsymbol{a}_t\|^2 \in (0, 2)$ is the classical LMS step-size condition. The projection interpretation is Kaczmarz's method (Kaczmarz, 1937): with the Newton step $\beta^\star = 1/(\|\boldsymbol{k}\|^2 \|\boldsymbol{v}\|^2)$, the update is the orthogonal projection onto the hyperplane $\{\boldsymbol{S} : \langle \boldsymbol{v}\boldsymbol{k}^\top, \boldsymbol{S}\rangle_{\mathrm{F}} = 1\}$, and general β gives relaxed projections. The modern convergence theory, including the linear rate under random sampling of constraints, is due to Strohmer & Vershynin (2009). The connections between Kaczmarz iterations, stochastic gradient descent, and learning are surveyed by Chen (2018). Fejér monotonicity and relaxed projection methods are standard in the projection-methods literature (Cegielski, 2012).

A1.5 Relation to cognitive and neuroscience

Because $\boldsymbol{S}$ scores key–value pairs rather than retrieving values, the closest relative in the cognitive and neuroscience literature is familiarity (recognition) memory: the anti-Hebbian model of familiarity discrimination in perirhinal cortex of Bogacz & Brown (2003) shows that scoring memories attain a far higher capacity than retrieval memories, which is the same d^2-versus-d capacity separation derived for the bilinear rule. The classical retrieval-side background is given by correlation matrix memories (Kohonen, 1972; Anderson, 1972) for the Hebbian write, and by delta-rule / pseudoinverse associative memories for capacity under Widrow–Hoff learning (Prados & Kak, 1989).

A2 Kronecker Delta Rule: Setting, Name, Properties

A2.1 Setup and notation.

We use the notation of the main paper (Section 2.1), specialized to $d_k = d_v =: d$ as in our experiments (where $d_{qk} = d_v$); the general $d_v \times d_k$ case is identical with d^2 replaced by $d_k d_v$. Let $\boldsymbol{S}_t \in \mathbb{R}^{d\times d}$ be the memory matrix, $\boldsymbol{k}_t \in \mathbb{R}^d$ the key and $\boldsymbol{v}_t \in \mathbb{R}^d$ the value at step t, $c_t \in \mathbb{R}$ a scalar target, and β_t a step size. The scalar $\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t$ measures how strongly the memory associates $\boldsymbol{k}_t$ with $\boldsymbol{v}_t$; the Kronecker Delta Rule corrects it towards the target c_t along the outer product $\boldsymbol{v}_t \boldsymbol{k}_t^\top$,

$$\boxed{\boldsymbol{S}_{t+1} \;=\; \boldsymbol{S}_t \;-\; \beta_t \big(\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t \;-\; c_t\big)\, \boldsymbol{v}_t \boldsymbol{k}_t^\top \,,} \tag{A1}$$

which is Eq. (3) of the main paper. For the analysis we abbreviate the scalar read-out and its residual,

$$y_t \;:=\; \boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t \qquad \text{(scalar read-out)}, \tag{A2}$$

$$r_t \;:=\; y_t - c_t \qquad \text{(residual)}, \tag{A3}$$

so that the update reads compactly $\boldsymbol{S}_{t+1} = \boldsymbol{S}_t - \beta_t\, r_t\, \boldsymbol{v}_t \boldsymbol{k}_t^\top$. The target sequence c_t subsumes different variants: $c_t \equiv 1$ is the basic Kronecker Delta Rule with target 1, $c_t \equiv c > 0$ the Constant-Target Kronecker Delta Rule, $c_t = \|\boldsymbol{v}_t\|^2$ the Norm-Target Kronecker Delta Rule, and $c_t = \|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|$ the Double-Norm-Target Kronecker Delta Rule. The target c_t may also be produced by a gate or a network, so that *what* is written becomes itself input-dependent. Throughout we use $\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{k} = \mathrm{Tr}\big(\boldsymbol{S}\, \boldsymbol{k}\boldsymbol{v}^\top\big) = \big\langle \boldsymbol{v}\boldsymbol{k}^\top, \boldsymbol{S}\big\rangle_{\mathrm{F}}$.

A2.2 Name Kronecker Delta Rule

We use the Kronecker product

$$\boldsymbol{a}_t \;=\; \boldsymbol{k}_t \otimes \boldsymbol{v}_t \tag{A4}$$

in a vectorized update rule. $\boldsymbol{a}_t$ is a vector of length d^2, where the vector $\boldsymbol{v}_t$ is multiplied by each component of $\boldsymbol{k}_t$ and then stacked to one vector. However, we have $\boldsymbol{v}_t \boldsymbol{k}_t^\top = \boldsymbol{v}_t \otimes \boldsymbol{k}_t^\top$. Let $\boldsymbol{s} = \mathrm{vec}(\boldsymbol{S}) \in \mathbb{R}^{d^2}$ and $\boldsymbol{a}_t = \boldsymbol{k}_t \otimes \boldsymbol{v}_t \in \mathbb{R}^{d^2}$, so that $\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t = \boldsymbol{a}_t^\top \boldsymbol{s}$. The rule becomes

$$\boldsymbol{s}_{t+1} \;=\; \big(\boldsymbol{I} - \beta\, \boldsymbol{a}_t \boldsymbol{a}_t^\top\big)\, \boldsymbol{s}_t + \beta\, c_t\, \boldsymbol{a}_t, \tag{A5}$$

which is exactly *Delta Rule with Kronecker product* input $\boldsymbol{a}_t$ and target c_t.

A2.3 PROPERTIES

Property A1 (Memory Space is Hilbert Space and Kronecker Update Rule is a Projection onto a Hyperplane)**.** *The memory space $\mathbb{R}^{d\times d}$ is a Hilbert space endowed with the Frobenius inner product*

$$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle_{\mathrm{F}} \;=\; \mathrm{Tr}\left(\boldsymbol{X}^\top \boldsymbol{Y}\right) \,,$$

where $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d\times d}$. Hence the target condition

$$\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t \;=\; c_t$$

is a hyperplane *in this Hilbert space. The update*

$$\boldsymbol{S}_{t+1} \;=\; \boldsymbol{S}_t - \beta \left(\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t - c_t\right) \boldsymbol{v}_t \boldsymbol{k}_t^\top$$

is a orthogonal projection *onto it, if*

$$\beta \;=\; \frac{1}{\|\boldsymbol{k}_t\|_2^2 \, \|\boldsymbol{v}_t\|_2^2} \,.$$

The update

$$\boldsymbol{S}_{t+1} \;=\; \boldsymbol{S}_t - \beta \left(\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t - c_t\right) \boldsymbol{v}_t \boldsymbol{k}_t^\top$$

is a relaxed orthogonal projection *onto it, with relaxation parameter*

$$\lambda_t \;=\; \beta \, \|\boldsymbol{k}_t\|_2^2 \, \|\boldsymbol{v}_t\|_2^2 \,.$$

Property A2 (Exact online gradient descent: a bilinear delta rule with target c_t)**.** *The rule* (A1) *is stochastic gradient descent with step size β on the instantaneous loss*

$$\mathrm{L}_t(\boldsymbol{S}) = \tfrac{1}{2}\left(\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t - c_t\right)^2, \qquad \nabla_S \mathrm{L}_t(\boldsymbol{S}) = \left(\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t - c_t\right) \boldsymbol{v}_t \boldsymbol{k}_t^\top . \tag{A6}$$

It differs from DeltaNet, which performs SGD on $\frac{1}{2}\|\boldsymbol{S}\boldsymbol{k}_t - \boldsymbol{v}_t\|^2$ with update $\boldsymbol{S}_{t+1} = \boldsymbol{S}_t - \beta(\boldsymbol{S}_t\boldsymbol{k}_t - \boldsymbol{v}_t)\boldsymbol{k}_t^\top$: there the target is the vector $\boldsymbol{v}_t$ *(imposing d constraints per pair), whereas here the target is the* scalar c_t *and the pair $(\boldsymbol{k}_t, \boldsymbol{v}_t)$ jointly forms the input, imposing a single bilinear constraint $\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t = c_t$ per pair.*

Property A3 (Superposition: the memory is linear in the target sequence)**.** *For a fixed pair sequence $(\boldsymbol{k}_t, \boldsymbol{v}_t)_{t=1}^T$ and step size β, the map from the target sequence $\boldsymbol{c} = (c_1, \ldots, c_T)$ to the final memory $\boldsymbol{S}_T(\boldsymbol{c})$ with $\boldsymbol{S}_0 = \boldsymbol{0}$ is* linear*:*

$$\boldsymbol{S}_T(\boldsymbol{c} + \boldsymbol{c}') \;=\; \boldsymbol{S}_T(\boldsymbol{c}) \,+\, \boldsymbol{S}_T(\boldsymbol{c}'), \qquad \boldsymbol{S}_T(\alpha\, \boldsymbol{c}) \;=\; \alpha\, \boldsymbol{S}_T(\boldsymbol{c}) \,, \tag{A7}$$

explicitly, with $\boldsymbol{a}_t = \boldsymbol{k}_t \otimes \boldsymbol{v}_t$,

$$\mathrm{vec}\left(\boldsymbol{S}_T(\boldsymbol{c})\right) \;=\; \beta \sum_{t=1}^{T} c_t \left[\prod_{\tau=t+1}^{T} \left(\boldsymbol{I} - \beta \boldsymbol{a}_\tau \boldsymbol{a}_\tau^\top\right)\right] \boldsymbol{a}_t \,. \tag{A8}$$

Memories written with different target sequences over the same pairs superpose exactly. The pure-gain homogeneity of the constant-target rule ($\boldsymbol{S}_t^{(c)} = c\, \boldsymbol{S}_t^{(1)}$) is the special case $\boldsymbol{c} = c\,\boldsymbol{1}$. The pair sequence determines where *the memory can write (the reachable subspace* $\mathrm{span}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T\}$ *and the interference pattern); the target sequence linearly selects* what *is written into it.*

Property A4 (Per-step reparameterization to the unit target)**.** *For every step with $c_t \neq 0$, rule* (A1) *with pair $(\boldsymbol{k}_t, \boldsymbol{v}_t)$ and step β coincides exactly with the unit-target rule applied to the rescaled value $\tilde{\boldsymbol{v}}_t = \boldsymbol{v}_t / c_t$ with the* time-varying *effective step $\tilde{\beta}_t = \beta\, c_t^2$:*

$$\boldsymbol{S} - \beta\left(\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{k} - c_t\right)\boldsymbol{v}\boldsymbol{k}^\top \;=\; \boldsymbol{S} - \tilde{\beta}_t \left(\tilde{\boldsymbol{v}}^\top \boldsymbol{S}\boldsymbol{k} - 1\right) \tilde{\boldsymbol{v}}\boldsymbol{k}^\top . \tag{A9}$$

In contrast to the constant-target case, the effective step size now varies over time, so the time-varying rule is not *globally a reparameterized unit-target rule; steps with $c_t = 0$ have no unit-target counterpart at all (see Property A5 (iii)).*

Property A5 (Distinguished targets)**.** *Three choices of c_t are structurally special.*

1. *(i)* Norm target, $c_t = \|\boldsymbol{v}_t\|^2$. *The residual becomes the value-projected retrieval error $r_t = \boldsymbol{v}_t^\top(\boldsymbol{S}_t\boldsymbol{k}_t - \boldsymbol{v}_t)$; every exact retrieval memory ($\boldsymbol{S}\boldsymbol{k}_i = \boldsymbol{v}_i$) satisfies all constraints, and in the autoassociative case $\boldsymbol{k}_t = \boldsymbol{v}_t$ the identity $\boldsymbol{S} = \boldsymbol{I}$ is a universal fixed point.*

(ii) Correlation target, $c_t = \boldsymbol{v}_t^\top \boldsymbol{k}_t$. *The identity* $\boldsymbol{S} = \boldsymbol{I}$ *is a universal fixed point even* hetero*associatively:* $\boldsymbol{v}^\top \boldsymbol{I}\boldsymbol{k} = \boldsymbol{v}^\top \boldsymbol{k}$ *for all pairs. The rule then learns deviations of the bilinear map from the standard inner product, i.e.* $\boldsymbol{S}$ *is naturally parameterized around the identity, and the read-out* $\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{k} - \boldsymbol{v}^\top \boldsymbol{k}$ *measures the* learned *(in-context) part of the similarity.*

(iii) Erasure target, $c_t = 0$. *The update drives* $\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t \to 0$*: with the Newton step* $\beta^\star = 1/(\|\boldsymbol{k}_t\|^2 \|\boldsymbol{v}_t\|^2)$ *the association is* deleted in one shot*, while every read-out with* $\boldsymbol{k}' \perp \boldsymbol{k}_t$ *or* $\boldsymbol{v}' \perp \boldsymbol{v}_t$ *is exactly unchanged. Hence,* $c_t = 0$ *is a targeted unlearning of a single key–value binding. Alternating write targets (*$c_t \neq 0$*) and erasure targets (*$c_t = 0$*) turns the memory into an editable scalar store (see also Property A10 on overwriting).*

Property A6 (One-step error contraction and stability range for β)**.** *The target never enters the homogeneous part of the dynamics. Re-reading with the same pair and the same target after one update,*

$$\boldsymbol{v}_t^\top \boldsymbol{S}_{t+1}\boldsymbol{k}_t \;=\; y_t - \beta\, r_t\, \|\boldsymbol{k}_t\|^2\, \|\boldsymbol{v}_t\|^2 \tag{A10}$$

$$\iff \boldsymbol{v}_t^\top \boldsymbol{S}_{t+1}\boldsymbol{k}_t - c_t \;=\; y_t - c_t - \beta\, r_t\, \|\boldsymbol{k}_t\|^2\, \|\boldsymbol{v}_t\|^2 \tag{A11}$$

$$\iff r_{t+1} \;=\; r_t - \beta\, r_t\, \|\boldsymbol{k}_t\|^2\, \|\boldsymbol{v}_t\|^2 \tag{A12}$$

$$\iff r_{t+1} \;=\; \left(1 - \beta\, \|\boldsymbol{k}_t\|^2\, \|\boldsymbol{v}_t\|^2\right) r_t\,, \tag{A13}$$

where r_{t+1} *is computed for the same* $(\boldsymbol{k}_t, \boldsymbol{v}_t, c_t)$*. The residual contracts if and only if*

$$0 \;<\; \beta \;<\; \frac{2}{\|\boldsymbol{k}_t\|^2\, \|\boldsymbol{v}_t\|^2}, \tag{A14}$$

and the loss shrinks by the factor $\left(1 - \beta\, \|\boldsymbol{k}_t\|^2\, \|\boldsymbol{v}_t\|^2\right)^2$*. Since* L_t *is quadratic along the ray* $\boldsymbol{S} - \eta\, r_t\, \boldsymbol{v}_t \boldsymbol{k}_t^\top$*,, the choice*

$$\beta^\star \;=\; \frac{1}{\|\boldsymbol{k}_t\|^2\, \|\boldsymbol{v}_t\|^2} \tag{A15}$$

is a Newton step: the constraint $\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t = c_t$ *is interpolated exactly in one update, for every value of* c_t*.*

Property A7 (Delta Rule / Widrow–Hoff LMS in vectorized form)**.** *Let* $\boldsymbol{s} = \mathrm{vec}(\boldsymbol{S}) \in \mathbb{R}^{d^2}$ *and* $\boldsymbol{a}_t = \boldsymbol{k}_t \otimes \boldsymbol{v}_t \in \mathbb{R}^{d^2}$ *the Kronecker product, so that* $\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t = \boldsymbol{a}_t^\top \boldsymbol{s}$ *and* $\|\boldsymbol{a}_t\|^2 = \|\boldsymbol{k}_t\|^2\, \|\boldsymbol{v}_t\|^2$*. The rule becomes*

$$\boldsymbol{s}_{t+1} \;=\; \left(\boldsymbol{I} - \beta\, \boldsymbol{a}_t \boldsymbol{a}_t^\top\right) \boldsymbol{s}_t \;+\; \beta\, c_t\, \boldsymbol{a}_t\,, \tag{A16}$$

which is exactly the Widrow–Hoff LMS/delta rule with input $\boldsymbol{a}_t$ *and target* c_t*. This is the classical setting of LMS, which regresses a* time-varying *desired signal onto the input. The transition matrix* $\boldsymbol{I} - \beta \boldsymbol{a}_t \boldsymbol{a}_t^\top$ *does not depend on* c_t*: it has eigenvalue* $1 - \beta\, \|\boldsymbol{a}_t\|^2$ *on* $\mathrm{span}(\boldsymbol{a}_t)$ *and eigenvalue* 1 *on the orthogonal complement (dimension* $d^2 - 1$*). Unrolling the recurrence,*

$$\boldsymbol{s}_T \;=\; \prod_{t=1}^{T} \left(\boldsymbol{I} - \beta \boldsymbol{a}_t \boldsymbol{a}_t^\top\right) \boldsymbol{s}_0 \;+\; \beta \sum_{t=1}^{T} c_t \left[\prod_{\tau=t+1}^{T} \left(\boldsymbol{I} - \beta \boldsymbol{a}_\tau \boldsymbol{a}_\tau^\top\right)\right] \boldsymbol{a}_t\,, \tag{A17}$$

gives the same product-of-generalized-Householder structure as DeltaNet. Hence, the WY-representation / chunkwise-parallel training techniques apply verbatim with $\boldsymbol{a}_t = \boldsymbol{k}_t \otimes \boldsymbol{v}_t$ *in place of* $\boldsymbol{k}_t$*, and the targets* c_t *enter only the driving (inhomogeneous) term. Expressivity consequence: for* $\beta\, \|\boldsymbol{a}_t\|^2 \in (1, 2)$ *the factor has a negative eigenvalue (reaching* -1 *at* $\beta\, \|\boldsymbol{a}_t\|^2 = 2$*), the same mechanism that gives DeltaNet-style models state-tracking power beyond linear RNNs restricted to nonnegative eigenvalues. This is independent of the target sequence.*

Property A8 (Homogeneous part of the update in non-expansive)**.** *For* $\beta\, \|\boldsymbol{a}_t\|^2 \in [0, 2]$ *every factor* $\boldsymbol{I} - \beta \boldsymbol{a}_t \boldsymbol{a}_t^\top$ *has spectral norm at most* 1*, so the homogeneous part of the recurrence never amplifies and* $\|\boldsymbol{S}_t\|$ *cannot blow up. In contrast the Hebbian write* $\boldsymbol{S}_{t+1} = \boldsymbol{S}_t + \boldsymbol{v}_t \boldsymbol{k}_t^\top$ *of vanilla linear attention leads to states that grow without bound. Moreover, the update is* surgical*: it modifies* $\boldsymbol{S}$ *only along the rank-one direction* $\boldsymbol{v}_t \boldsymbol{k}_t^\top$*. For any test pair with* $\boldsymbol{k}' \perp \boldsymbol{k}_t$ **or** $\boldsymbol{v}' \perp \boldsymbol{v}_t$ *the read-out* $\boldsymbol{v}'^\top \boldsymbol{S}\boldsymbol{k}'$ *is exactly unchanged.*

Property A9 (Doubly gated interference)**.** *For a previously stored pair $(\boldsymbol{k}_i, \boldsymbol{v}_i)$, an update with $(\boldsymbol{k}_t, \boldsymbol{v}_t, c_t)$ perturbs the stored read-out by*

$$\Delta\big(\boldsymbol{v}_i^\top \boldsymbol{S}\boldsymbol{k}_i\big) \;=\; -\,\beta\, r_t\, \big(\boldsymbol{v}_i^\top \boldsymbol{v}_t\big)\big(\boldsymbol{k}_t^\top \boldsymbol{k}_i\big)\,, \tag{A18}$$

i.e. interference is proportional to the product *of the key overlap and the value overlap. Orthogonality in either factor suffices to protect a memory, which is a milder interference profile than the Hebbian rule (crosstalk from key overlap alone) and than DeltaNet (gated by key overlap only). The formula is target-independent, since the target enters only through the residual r_t. Consequently, writes whose read-out is already correct ($r_t \approx 0$) cause no interference at all, regardless of c_t.*

Property A10 (Relaxed Kaczmarz iteration, Fejér monotonicity, convergence, and overwriting)**.** *The constraint $\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t = c_t$ defines the hyperplane $\mathcal{H}_t = \{\boldsymbol{S} : \langle \boldsymbol{v}_t\boldsymbol{k}_t^\top, \boldsymbol{S}\rangle_{\mathrm{F}} = c_t\}$ in $\mathbb{R}^{d\times d}$ with Frobenius geometry. With $\beta = \beta^\star$ the update (A1) is the orthogonal projection onto $\mathcal{H}_t$; general β gives a relaxed projection. In contrast to a constant target, consistency of the system $\{\boldsymbol{v}_i^\top \boldsymbol{S}\boldsymbol{k}_i = c_i\}_i$ is now a genuine condition: it holds if and only if the target vector $(c_i)_i$ lies in the range of the matrix stacking the rows $\boldsymbol{a}_i^\top$. In particular, this always holds, whenever the $\{\boldsymbol{a}_i\}$ are linearly independent (at most d^2 of them). If the system is consistent, then for every solution $\boldsymbol{S}^\star$ and any $\beta_t \in \big[0,\, 2/\,\|\boldsymbol{a}_t\|^2\big]$,*

$$\|\boldsymbol{S}_{t+1} - \boldsymbol{S}^\star\|_{\mathrm{F}} \;\le\; \|\boldsymbol{S}_t - \boldsymbol{S}^\star\|_{\mathrm{F}} \qquad \textit{(Fejér monotonicity)}, \tag{A19}$$

and cycling or randomly sampling over the pairs converges linearly to a solution, at a rate governed by the smallest singular value of the stacked $\boldsymbol{a}_i^\top$. From $\boldsymbol{S}_0 = \boldsymbol{0}$ the limit is the minimum-Frobenius-norm *solution. The canonical* inconsistent *case is a repeated pair with a changed target, $(\boldsymbol{k}, \boldsymbol{v}, c)$ followed later by $(\boldsymbol{k}, \boldsymbol{v}, c')$ with $c' \neq c$: the two hyperplanes are parallel, and the iteration resolves the conflict by* overwriting*. The contraction of Property A6 drives the read-out to the most recent target, in one shot for $\beta = \beta^\star$. In expectation over a stationary distribution of triples $(\boldsymbol{k}, \boldsymbol{v}, c)$, the fixed-point condition is*

$$\mathrm{E}\big[(\boldsymbol{a}^\top \boldsymbol{s} - c)\boldsymbol{a}\big] = \boldsymbol{0}\,, \tag{A20}$$

the normal equation of the least-squares regression of the target c on the feature $\boldsymbol{a} = \boldsymbol{k} \otimes \boldsymbol{v}$: the memory is an online linear regressor *that learns, in context, to predict the scalar signal c_t from the key–value pair.*

Property A11 (Capacity: memory can store d^2 associations)**.** *Each triple $(\boldsymbol{k}_i, \boldsymbol{v}_i, c_i)$ contributes one linear constraint on the d^2 parameters of $\boldsymbol{S}$, so up to d^2* arbitrary *scalars c_i can be stored exactly for pairs with linearly independent $\{\boldsymbol{k}_i \otimes \boldsymbol{v}_i\}$ (d^2 scalar components). Compare to DeltaNet with goal $\boldsymbol{S}\boldsymbol{k}_i = \boldsymbol{v}_i$ that represent d value vectors. In an orthonormal design ($\{\boldsymbol{k}_i\}$ orthonormal and $\{\boldsymbol{v}_i\}$ orthonormal), the minimum-norm solution has the closed form*

$$\boldsymbol{S}^\star \;=\; \sum_i c_i\, \boldsymbol{v}_i \boldsymbol{k}_i^\top\,, \qquad \boldsymbol{v}_j^\top \boldsymbol{S}^\star \boldsymbol{k}_i \;=\; c_i\, \delta_{ij}\,, \tag{A21}$$

a target-weighted Hebbian superposition *realizing the learned Kronecker delta with per-pair gains. The memory is thus best understood as a* pair-addressed scalar table*. It stores a number c_i at the address $(\boldsymbol{k}_i, \boldsymbol{v}_i)$, writes by rank-one updates, reads by the bilinear probe $\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{k}$, rewrites by presenting the same pair with a new target, and deletes via $c = 0$ (Property A5 (iii)).*

Next, we exemplify the fact that KronLSTM can store d^2 numbers. Assume d orthonormal keys $\{\boldsymbol{k}_1, \ldots, \boldsymbol{k}_d\}$ and d orthonormal values $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d\}$. Then we can build d^2 key-value pairs $(\boldsymbol{k}_i, \boldsymbol{v}_j)$ with their associated d^2 numbers c_{ij}. We set $\beta = 1$ and start with $\boldsymbol{S}_0 = \boldsymbol{0}$, then we iteratively update $\boldsymbol{S}_t$ with randomly chosen i and j:

$$\boldsymbol{S}_{t+1} \;=\; \boldsymbol{S}_t - \beta\, \big(\boldsymbol{v}_j^\top\, \boldsymbol{S}_t\, \boldsymbol{k}_i \;-\; c_{ij}\big)\, \boldsymbol{v}_j\, \boldsymbol{k}_i^\top\,. \tag{A22}$$

If all pairs $(\boldsymbol{k}_i, \boldsymbol{v}_j)$ or numbers c_{ij} are seen at least once, the update converged. If pair $(\boldsymbol{k}_i, \boldsymbol{v}_j)$ has already been used in the update rule, then $\boldsymbol{v}_j^\top \boldsymbol{S}_t \boldsymbol{k}_i = c_{ij}$ and $\boldsymbol{S}_{t+1} = \boldsymbol{S}_t$. Therefore, if all pairs are seen, $\boldsymbol{S}_t$ does not change any more and the update converged. If pair $(\boldsymbol{k}_i, \boldsymbol{v}_j)$ has NOT been stored, then $\boldsymbol{v}_j^\top \boldsymbol{S}_t \boldsymbol{k}_i = 0$ and $\boldsymbol{S}_{t+1} = \boldsymbol{S}_t + c_{ij}\boldsymbol{v}_j\boldsymbol{k}_i^\top$. Therefore, if all pairs $(\boldsymbol{k}_i, \boldsymbol{v}_j)$ or numbers c_{ij} are seen at least once, the memory matrix $\boldsymbol{S}_t$ is

$$\boldsymbol{S}_t \;=\; \sum_{i=1,j=1}^{d,d} c_{ij}\, \boldsymbol{v}_j\, \boldsymbol{k}_i^\top \;=\; \boldsymbol{V}\, \boldsymbol{C}^\top\, \boldsymbol{K}^\top\,, \tag{A23}$$

with

$$\boldsymbol{K} \;=\; [\boldsymbol{k}_1,\,\ldots,\,\boldsymbol{k}_d] \in \mathbb{R}^{d\times d}, \quad \boldsymbol{V} \;=\; [\boldsymbol{v}_1,\,\ldots,\,\boldsymbol{v}_d] \in \mathbb{R}^{d\times d}, \quad \boldsymbol{C} \;=\; (c_{ij})_{i,j=1}^{d} \in \mathbb{R}^{d\times d}. \tag{A24}$$

Since $\boldsymbol{v}_j^\top \boldsymbol{S}_t \boldsymbol{k}_i = c_{ij}$, we can now retrieve d^2 stored numbers.
We have

$$\boldsymbol{C}^\top \;=\; \boldsymbol{V}^\top \boldsymbol{S}_t\, \boldsymbol{K}\,, \tag{A25}$$

since $\boldsymbol{K}$ and $\boldsymbol{V}$ are orthogonal. The transpose appears because the first index of $\boldsymbol{C}$ is the key index while the first index of $\boldsymbol{S}_t$ is the value index.

Property A12 (Norm and energy dynamics)**.** *Using* $\langle \boldsymbol{v}_t\boldsymbol{k}_t^\top, \boldsymbol{S}_t\rangle_{\mathrm{F}} = y_t = r_t + c_t$,

$$\|\boldsymbol{S}_{t+1}\|_{\mathrm{F}}^2 - \|\boldsymbol{S}_t\|_{\mathrm{F}}^2 \;=\; -2\beta\, r_t(r_t + c_t) \;+\; \beta^2 r_t^2\, \|\boldsymbol{k}_t\|^2\, \|\boldsymbol{v}_t\|^2\,, \tag{A26}$$

so within the stable range, the state norm stays bounded whenever the constraint system is consistent. This is a consequence of Fejér monotonicity: the distance to any fixed solution is non-increasing.

Summary. The time-varying-target rule (A1) is the most general member of the Kronecker Delta Rule family: SGD on the scalar quadratic loss $\frac{1}{2}(\boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t - c_t)^2$, equivalently Widrow–Hoff LMS on the feature $\boldsymbol{a}_t = \boldsymbol{k}_t \otimes \boldsymbol{v}_t$ with desired signal c_t, equivalently relaxed Kaczmarz on the hyperplanes $\langle \boldsymbol{v}_t\boldsymbol{k}_t^\top, \boldsymbol{S}\rangle_{\mathrm{F}} = c_t$. The entire analytic framework is target-independent. That comprises stability range and Newton step $\beta^\star$, generalized-Householder transition factors with their parallelization and negative-eigenvalue state-tracking regime, doubly gated interference, Fejér-monotone linear convergence, d^2 capacity. The targets enter only the driving term, *linearly*: memories superpose in (c_t), and in an orthonormal design the memory converges to the target-weighted Hebbian superposition $\sum_i c_i \boldsymbol{v}_i\boldsymbol{k}_i^\top$ realizing $\boldsymbol{v}_j^\top \boldsymbol{S}\boldsymbol{k}_i = c_i\,\delta_{ij}$. Distinguished targets recover $c_t \equiv 1$, $c_t \equiv c$, $c_t = \|\boldsymbol{v}_t\|^2$ and add two new members: the correlation target $c_t = \boldsymbol{v}_t^\top \boldsymbol{k}_t$, for which the identity is a universal fixed point and $\boldsymbol{S}$ parameterizes learned deviations from the standard inner product, and the erasure target $c_t = 0$, which deletes a single binding in one Newton step. Functionally, $\boldsymbol{S}$ is a pair-addressed scalar table with write, read, rewrite, and delete operations, therefore it is an editable d^2-slot scalar memory. In the stochastic view, it is an online in-context regressor of the signal c_t on the bilinear feature.

A2.4 Minimally Invasive Memory Revision

The Kronecker Delta Rule is not merely *a* correction of the addressed association: at the Newton step it is *the* smallest possible one, in every sense in which smallness can reasonably be asked for. This section elaborates on this topic. Throughout, $\boldsymbol{S}_t \in \mathbb{R}^{d\times d}$, the read-out of the pair $(\boldsymbol{k}_t, \boldsymbol{v}_t)$ is $\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t$, the target is c_t, and

$$r_t \;:=\; \boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t \;-\; c_t \tag{A27}$$

is the residual that should be pushed to zero.

A2.4.1 The admissible set

A correction is a matrix $\boldsymbol{E} \in \mathbb{R}^{d\times d}$ added to the memory. It *corrects the addressed association* if $\boldsymbol{v}_t^\top(\boldsymbol{S}_t + \boldsymbol{E})\boldsymbol{k}_t = c_t$. Using $\boldsymbol{v}_t^\top \boldsymbol{E}\boldsymbol{k}_t = \langle \boldsymbol{v}_t\boldsymbol{k}_t^\top, \boldsymbol{E}\rangle_{\mathrm{F}}$, the set of such corrections is

$$\mathcal{E}_t \;:=\; \Big\{ \boldsymbol{E} \in \mathbb{R}^{d\times d} \;:\; \langle \boldsymbol{v}_t\boldsymbol{k}_t^\top,\, \boldsymbol{E}\rangle_{\mathrm{F}} \;=\; -\,r_t \Big\}, \tag{A28}$$

an affine hyperplane in the Hilbert space $(\mathbb{R}^{d\times d}, \langle\cdot,\cdot\rangle_{\mathrm{F}})$: it is nonempty, and it is a proper affine subspace because $\boldsymbol{v}_t\boldsymbol{k}_t^\top \neq \boldsymbol{0}$ whenever the pair is nondegenerate. Every statement below compares members of $\mathcal{E}_t$. We hold the *requirement* fixed (the addressed association must come out right) and ask which correction achieves it with the least collateral change. Write

$$\boldsymbol{E}^\star \;:=\; -\,\frac{r_t}{\|\boldsymbol{k}_t\|^2\,\|\boldsymbol{v}_t\|^2}\,\boldsymbol{v}_t\boldsymbol{k}_t^\top \;=\; \frac{c_t - \boldsymbol{v}_t^\top \boldsymbol{S}_t\boldsymbol{k}_t}{\|\boldsymbol{k}_t\|^2\,\|\boldsymbol{v}_t\|^2}\,\boldsymbol{v}_t\boldsymbol{k}_t^\top\,, \tag{A29}$$

which is exactly the Kronecker Delta Rule update at the Newton step $\beta = \beta^\star = 1/(\|\boldsymbol{k}_t\|^2\,\|\boldsymbol{v}_t\|^2)$.

A2.4.2 MINIMALITY IN NORM

Theorem A1 (The Newton step is the smallest correction)**.** *$\boldsymbol{E}^\star \in \mathcal{E}_t$, and for every $\boldsymbol{E} \in \mathcal{E}_t$*

$$\|\boldsymbol{E}\|_{\mathrm{F}} \;\geq\; \frac{|r_t|}{\|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|} \;=\; \|\boldsymbol{E}^\star\|_{\mathrm{F}}\,, \tag{A30}$$

with equality if and only if $\boldsymbol{E} = \boldsymbol{E}^\star$.

Proof. Admissibility is a computation: $\langle \boldsymbol{v}_t\boldsymbol{k}_t^\top, \boldsymbol{E}^\star\rangle_{\mathrm{F}} = -r_t \left\|\boldsymbol{v}_t\boldsymbol{k}_t^\top\right\|_{\mathrm{F}}^2 / (\|\boldsymbol{k}_t\|^2\,\|\boldsymbol{v}_t\|^2) = -r_t$, using $\left\|\boldsymbol{v}_t\boldsymbol{k}_t^\top\right\|_{\mathrm{F}} = \|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|$. For the bound, apply Cauchy–Schwarz to the constraint in (A28):

$$|r_t| \;=\; \left|\langle \boldsymbol{v}_t\boldsymbol{k}_t^\top, \boldsymbol{E}\rangle_{\mathrm{F}}\right| \;\leq\; \left\|\boldsymbol{v}_t\boldsymbol{k}_t^\top\right\|_{\mathrm{F}}\,\|\boldsymbol{E}\|_{\mathrm{F}} \;=\; \|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|\,\|\boldsymbol{E}\|_{\mathrm{F}}\,, \tag{A31}$$

which is (A30). Cauchy–Schwarz holds with equality exactly when $\boldsymbol{E}$ is a scalar multiple of $\boldsymbol{v}_t\boldsymbol{k}_t^\top$, and the constraint then pins that scalar to $-r_t/(\|\boldsymbol{k}_t\|^2\,\|\boldsymbol{v}_t\|^2)$, i.e. $\boldsymbol{E} = \boldsymbol{E}^\star$. □

Equivalently the geometric view: $\boldsymbol{S}_t + \boldsymbol{E}^\star$ is the orthogonal projection of $\boldsymbol{S}_t$ onto the constraint hyperplane $\mathcal{H}_t = \{\boldsymbol{S} : \boldsymbol{v}_t^\top \boldsymbol{S}\boldsymbol{k}_t = c_t\}$, so the Newton step moves the memory along the shortest path from where it is to where the constraint is satisfied.

Theorem A2 (Minimality in *every* unitarily invariant norm)**.** *Let $|||\cdot|||$ be any unitarily invariant norm on $\mathbb{R}^{d\times d}$. Then*

$$\min_{\boldsymbol{E}\in\mathcal{E}_t} |||\boldsymbol{E}||| \;=\; |||\boldsymbol{E}^\star|||\,, \tag{A32}$$

i.e. the same *$\boldsymbol{E}^\star$ is optimal simultaneously for all of them. In particular, for every Schatten-p norm, $1 \leq p \leq \infty$,*

$$\min_{\boldsymbol{E}\in\mathcal{E}_t} \|\boldsymbol{E}\|_{S_p} \;=\; \|\boldsymbol{E}^\star\|_{S_p} \;=\; \frac{|r_t|}{\|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|}\,, \tag{A33}$$

the same number for all p. Therefore, the nuclear, Frobenius and spectral norms of the optimal correction coincide. The minimizer is unique for strictly convex norms (Schatten $1 < p < \infty$). For the spectral and nuclear norms $\boldsymbol{E}^\star$ is a minimizer but need not be the only one.

Proof. Let $|||\cdot|||'$ be the dual norm, so that $|\langle \boldsymbol{A}, \boldsymbol{E}\rangle_{\mathrm{F}}| \leq |||\boldsymbol{A}|||'|||\boldsymbol{E}|||$ for all $\boldsymbol{A}, \boldsymbol{E}$. Applying this to the constraint gives, for every $\boldsymbol{E} \in \mathcal{E}_t$,

$$|||\boldsymbol{E}||| \;\geq\; \frac{|r_t|}{\left|\left|\left|\boldsymbol{v}_t\boldsymbol{k}_t^\top\right|\right|\right|'}\,. \tag{A34}$$

By von Neumann's theorem a unitarily invariant norm is a symmetric gauge function Φ of the singular values, and the dual norm is the gauge Φ' of the singular values. The matrix $\boldsymbol{v}_t\boldsymbol{k}_t^\top$ has the single nonzero singular value $\|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|$, so

$$\left|\left|\left|\boldsymbol{v}_t\boldsymbol{k}_t^\top\right|\right|\right| = \|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|\;\Phi(\boldsymbol{e}_1)\,, \qquad \left|\left|\left|\boldsymbol{v}_t\boldsymbol{k}_t^\top\right|\right|\right|' = \|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|\;\Phi'(\boldsymbol{e}_1)\,. \tag{A35}$$

Now $\Phi(\boldsymbol{e}_1)\Phi'(\boldsymbol{e}_1) = 1$: indeed $\Phi'(\boldsymbol{e}_1) = \sup\{x_1 : \Phi(\boldsymbol{x}) \leq 1\}$, and since a symmetric gauge is nondecreasing in each $|x_i|$, replacing $\boldsymbol{x}$ by $(x_1, 0, \ldots, 0)$ does not increase Φ, so the supremum is attained on the first axis and equals $1/\Phi(\boldsymbol{e}_1)$. Since $\boldsymbol{E}^\star$ is rank one with singular value $|r_t|/(\|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|)$,

$$|||\boldsymbol{E}^\star||| \;=\; \frac{|r_t|}{\|\boldsymbol{k}_t\|\,\|\boldsymbol{v}_t\|}\,\Phi(\boldsymbol{e}_1) \;=\; \frac{|r_t|}{\left|\left|\left|\boldsymbol{v}_t\boldsymbol{k}_t^\top\right|\right|\right|'}\,, \tag{A36}$$

which is exactly the lower bound (A34): $\boldsymbol{E}^\star$ attains it. The Schatten case follows from $\Phi_p(\boldsymbol{e}_1) = 1$ for every p. Uniqueness for a strictly convex norm holds because $\mathcal{E}_t$ is convex and a strictly convex function has at most one minimizer on a convex set; for the spectral norm uniqueness fails already at $d = 2$ with $\boldsymbol{k} = \boldsymbol{v} = \boldsymbol{e}_1$ and $r_t = -1$, where $\boldsymbol{E} = \mathrm{diag}(1, \delta)$ is admissible and has spectral norm 1 for every $|\delta| \leq 1$. □

A2.4.3 MINIMALITY OF SUPPORT: THE CORRECTION IS SURGICAL

Norm-optimality says the correction is small. The next theorem has a stronger statement even: it is the only correction that stays inside the addressed directions at all. No optimization is involved.

Theorem A3 (Unique correction with addressed support)**.** *$\boldsymbol{E}^\star$ is the unique $\boldsymbol{E} \in \mathcal{E}_t$ whose column space is contained in $\mathrm{span}\{\boldsymbol{v}_t\}$ and whose row space is contained in $\mathrm{span}\{\boldsymbol{k}_t\}$.*

Proof. A matrix has column space in $\mathrm{span}\{\boldsymbol{v}_t\}$ and row space in $\mathrm{span}\{\boldsymbol{k}_t\}$ if and only if it is of the form $\boldsymbol{E} = \gamma\, \boldsymbol{v}_t \boldsymbol{k}_t^\top$ for some $\gamma \in \mathbb{R}$. Admissibility forces $\gamma \left\|\boldsymbol{k}_t\right\|^2 \left\|\boldsymbol{v}_t\right\|^2 = -r_t$, hence $\gamma = -r_t/(\left\|\boldsymbol{k}_t\right\|^2 \left\|\boldsymbol{v}_t\right\|^2)$ and $\boldsymbol{E} = \boldsymbol{E}^\star$. □

Every other admissible correction like every DeltaNet-style correction necessarily writes into at least one key direction other than $\boldsymbol{k}_t$ or one value direction other than $\boldsymbol{v}_t$, i.e. into directions the task never asked about.

A2.4.4 MINIMALITY OF INTERFERENCE

Smallness in norm translates into smallness of the damage done to *other* associations.

Theorem A4 (Interference caused by the correction)**.** *Let $(\boldsymbol{k}', \boldsymbol{v}')$ be any probe pair and let $\Delta(\boldsymbol{k}', \boldsymbol{v}') := \boldsymbol{v}'^\top \boldsymbol{E} \boldsymbol{k}'$ be the change its read-out suffers.*

(i) Exact form and protection. *For $\boldsymbol{E} = \boldsymbol{E}^\star$,*

$$\Delta(\boldsymbol{k}', \boldsymbol{v}') \;=\; -\,\frac{r_t}{\left\|\boldsymbol{k}_t\right\|^2 \left\|\boldsymbol{v}_t\right\|^2} \left(\boldsymbol{v}'^\top \boldsymbol{v}_t\right)\left(\boldsymbol{k}_t^\top \boldsymbol{k}'\right), \tag{A37}$$

*which vanishes whenever $\boldsymbol{v}' \perp \boldsymbol{v}_t$ **or** $\boldsymbol{k}' \perp \boldsymbol{k}_t$, and obeys $|\Delta(\boldsymbol{k}', \boldsymbol{v}')| \le |r_t|\, \|\boldsymbol{k}'\|\, \|\boldsymbol{v}'\| / (\|\boldsymbol{k}_t\|\, \|\boldsymbol{v}_t\|)$. The condition is a property of the inputs alone: it can be checked in advance and does not depend on what $\boldsymbol{S}_t$ currently holds.*

(ii) Minimum expected interference. *Suppose probe pairs are drawn with $\boldsymbol{k}'$ and $\boldsymbol{v}'$ independent and isotropic, $\mathrm{E}[\boldsymbol{k}'\boldsymbol{k}'^\top] = \sigma_k^2 \boldsymbol{I}$ and $\mathrm{E}[\boldsymbol{v}'\boldsymbol{v}'^\top] = \sigma_v^2 \boldsymbol{I}$. Then for* every $\boldsymbol{E} \in \mathbb{R}^{d \times d}$

$$\mathrm{E}\Big[\left(\boldsymbol{v}'^\top \boldsymbol{E} \boldsymbol{k}'\right)^2\Big] \;=\; \sigma_k^2\, \sigma_v^2 \;\|\boldsymbol{E}\|_{\mathrm{F}}^2 \,. \tag{A38}$$

Consequently

$$\boldsymbol{E}^\star \;=\; \operatorname*{argmin}_{\boldsymbol{E} \in \mathcal{E}_t} \mathrm{E}\Big[\left(\boldsymbol{v}'^\top \boldsymbol{E} \boldsymbol{k}'\right)^2\Big], \qquad \min_{\boldsymbol{E} \in \mathcal{E}_t} \mathrm{E}\big[\Delta^2\big] \;=\; \frac{\sigma_k^2 \sigma_v^2\, r_t^2}{\left\|\boldsymbol{k}_t\right\|^2 \left\|\boldsymbol{v}_t\right\|^2}, \tag{A39}$$

and the minimizer is unique.

Proof. (i) Substitute (A29) and use $\boldsymbol{v}'^\top(\boldsymbol{v}_t \boldsymbol{k}_t^\top)\boldsymbol{k}' = (\boldsymbol{v}'^\top \boldsymbol{v}_t)(\boldsymbol{k}_t^\top \boldsymbol{k}')$; the inequality is Cauchy–Schwarz on each factor. (ii) Vectorize: with $\boldsymbol{e} = \mathrm{vec}(\boldsymbol{E})$ and $\boldsymbol{a}' = \boldsymbol{k}' \otimes \boldsymbol{v}'$ we have $\boldsymbol{v}'^\top \boldsymbol{E} \boldsymbol{k}' = \boldsymbol{a}'^\top \boldsymbol{e}$, hence

$$\mathrm{E}\big[(\boldsymbol{a}'^\top \boldsymbol{e})^2\big] \;=\; \boldsymbol{e}^\top\, \mathrm{E}\big[\boldsymbol{a}'\boldsymbol{a}'^\top\big]\, \boldsymbol{e} \;=\; \boldsymbol{e}^\top \Big(\mathrm{E}[\boldsymbol{k}'\boldsymbol{k}'^\top] \otimes \mathrm{E}[\boldsymbol{v}'\boldsymbol{v}'^\top]\Big) \boldsymbol{e} \;=\; \sigma_k^2 \sigma_v^2 \left\|\boldsymbol{e}\right\|^2, \tag{A40}$$

using independence and $\|\boldsymbol{e}\| = \|\boldsymbol{E}\|_{\mathrm{F}}$. The objective is therefore a positive multiple of $\|\boldsymbol{E}\|_{\mathrm{F}}^2$, so minimizing it over $\mathcal{E}_t$ is the problem already solved by Theorem A1, with the same unique solution. □

Part (ii) is what justifies the name: for isotropic probe statistics, *minimum norm and minimum expected interference are literally the same objective*, so the Newton step is minimally invasive in the operational sense and not merely in the geometric one. It causes the least expected damage to the rest of the memory.

A2.4.5 RELAXED STEPS: THE CORRECTION NEVER MOVES AWAY FROM A CORRECT MEMORY

In practice β is a learned gate rather than $\beta^\star$. Write $\beta = \lambda \beta^\star$, so that the actual update is $\boldsymbol{S}_{t+1} = \boldsymbol{S}_t + \lambda \boldsymbol{E}^\star$ with relaxation $\lambda = \beta \left\|\boldsymbol{k}_t\right\|^2 \left\|\boldsymbol{v}_t\right\|^2$.

Theorem A5 (Exact Fejér identity)**.** *Let $\boldsymbol{S}^\circ$ be* any *memory that already satisfies the addressed association, $\boldsymbol{v}_t^\top \boldsymbol{S}^\circ \boldsymbol{k}_t = c_t$. Then*

$$\|\boldsymbol{S}_{t+1} - \boldsymbol{S}^\circ\|_{\mathrm{F}}^2 \;=\; \|\boldsymbol{S}_t - \boldsymbol{S}^\circ\|_{\mathrm{F}}^2 \;-\; \lambda\,(2-\lambda)\,\frac{r_t^2}{\|\boldsymbol{k}_t\|^2\,\|\boldsymbol{v}_t\|^2}\,. \tag{A41}$$

Hence for every $\lambda \in (0,2)$ the update strictly decreases *the distance to every correct memory, the decrease is largest at the Newton step $\lambda = 1$, and at $\lambda \in \{0,2\}$ the distance is unchanged.*

Proof. Expand, using $\langle \boldsymbol{E}^\star, \boldsymbol{S}_t - \boldsymbol{S}^\circ\rangle_{\mathrm{F}} = -\frac{r_t}{\|\boldsymbol{k}_t\|^2\|\boldsymbol{v}_t\|^2}\left(\boldsymbol{v}_t^\top \boldsymbol{S}_t \boldsymbol{k}_t - \boldsymbol{v}_t^\top \boldsymbol{S}^\circ \boldsymbol{k}_t\right) = -r_t^2/(\|\boldsymbol{k}_t\|^2\,\|\boldsymbol{v}_t\|^2) = -\|\boldsymbol{E}^\star\|_{\mathrm{F}}^2$:

$$\|\boldsymbol{S}_t - \boldsymbol{S}^\circ + \lambda \boldsymbol{E}^\star\|_{\mathrm{F}}^2 = \|\boldsymbol{S}_t - \boldsymbol{S}^\circ\|_{\mathrm{F}}^2 + 2\lambda\langle \boldsymbol{E}^\star, \boldsymbol{S}_t - \boldsymbol{S}^\circ\rangle_{\mathrm{F}} + \lambda^2\,\|\boldsymbol{E}^\star\|_{\mathrm{F}}^2\,, \tag{A42}$$

which is (A41) after inserting $\|\boldsymbol{E}^\star\|_{\mathrm{F}}^2 = r_t^2/(\|\boldsymbol{k}_t\|^2\,\|\boldsymbol{v}_t\|^2)$. □

This is the strongest sense in which a single write is harmless: it does not merely perturb other associations little, it moves the memory *closer* to every memory that is consistent with the association being written and that simultaneously, for all of them.

A2.4.6 Greedy minimality is global minimality

Per-step optimality does not by itself imply that a *sequence* of such steps is optimal. We show that our does.

Proposition 1 (The limit is the minimum-norm memory)**.** *Let $\{(\boldsymbol{k}_i, \boldsymbol{v}_i, c_i)\}_{i=1}^n$ be a consistent system of associations. Apply the rule from $\boldsymbol{S}_0 = \boldsymbol{0}$ with relaxation $\lambda \in (0,2)$, visiting each pair infinitely often either cyclically or by random sampling. Then $\boldsymbol{S}_t$ converges to*

$$\boldsymbol{S}^\sharp \;=\; \operatorname{argmin}\left\{\,\|\boldsymbol{S}\|_{\mathrm{F}} \;:\; \boldsymbol{v}_i^\top \boldsymbol{S}\boldsymbol{k}_i = c_i \;\; \textit{for all } i\right\}, \tag{A43}$$

the minimum Frobenius-norm memory satisfying all *the associations.*

Proof. This is the classical convergence theorem for relaxed Kaczmarz iterations (Kaczmarz, 1937; Cegielski, 2012), applied to the vectorized system $\boldsymbol{a}_i^\top \boldsymbol{s} = c_i$ with $\boldsymbol{a}_i = \boldsymbol{k}_i \otimes \boldsymbol{v}_i$. Each iterate stays in $\boldsymbol{s}_0 + \operatorname{span}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$, which for $\boldsymbol{s}_0 = \boldsymbol{0}$ is the row space of the system; the iterates converge to a solution, and the unique solution in the row space is the minimum-norm one. □

So the memory built by a stream of minimally invasive revisions is itself the least-norm memory compatible with everything it was asked to store: locally and globally, the rule writes as little as the task permits.

A2.4.7 Less invasive than DeltaNet

Theorems A1–A4 hold *given the requirement* encoded in $\mathcal{E}_t$: make the scalar read-out of the addressed pair equal to c_t, and do nothing else. DeltaNet's step has to solve the d constraints $\boldsymbol{S}\boldsymbol{k}_t = \boldsymbol{v}_t$, therefore its correction is necessarily larger. The claim here is therefore not that DeltaNet is suboptimal for its own task, but that the two tasks differ: a recognition memory needs one scalar corrected, and the Kronecker Delta Rule corrects it with provably minimal disturbance, whereas a retrieval memory must rewrite an entire value vector and pays for it in footprint.

A2.5 Convergence of the KronLSTM memory

We analyze the expected solution and steady state of the randomized Kronecker Delta Rule with more than d^2 associations. Running memories i.e. online updates are covered via the steady state considerations.

A2.5.1 Overview

We analyze the Kronecker Delta Rule $\boldsymbol{S}_{t+1} = \boldsymbol{S}_t - \beta\,(\boldsymbol{v}_{i_t}^\top \boldsymbol{S}_t \boldsymbol{k}_{i_t} - c_{i_t})\,\boldsymbol{v}_{i_t}\boldsymbol{k}_{i_t}^\top$ with randomly drawn indices when the number of associations exceeds the capacity: $n > d^2$. The system is then generically *inconsistent*, and the iterates no longer converge to a point. Instead we have: (i) The **expected iterate** converges linearly to the sampling-weighted **least-squares solution** $\boldsymbol{S}^{\mathrm{LS}} = \operatorname{argmin}_{\boldsymbol{S}} \sum_i p_i (\boldsymbol{v}_i^\top \boldsymbol{S}\boldsymbol{k}_i - c_i)^2$. This is for uniform sampling exactly the minimizer of the loss L. (ii) The iterates converge in distribution to a unique **stationary distribution** whose mean is *exactly* $\boldsymbol{S}^{\mathrm{LS}}$ (no bias at any order in β) and whose covariance is $O(\beta)$, given by a Lyapunov-type equation

driven by the residuals at $\boldsymbol{S}^{\mathrm{LS}}$. (iii) The steady-state loss exceeds the minimum by the classical LMS misadjustment, $\mathrm{E}[\mathrm{L}] - \mathrm{L}_{\min} = \frac{\beta}{2}\,\mathrm{E}\big[\rho_i^2\,\|\boldsymbol{k}_i\|^2\,\|\boldsymbol{v}_i\|^2\big] + O(\beta^2)$, where ρ_i are the optimal residuals. This appendix contains **complete proofs** of both theorems: the mean dynamics are solved exactly; existence, uniqueness, and geometric convergence of the stationary distribution are established by a synchronous-coupling contraction in the Wasserstein-2 metric; the stationary mean and the exact covariance equation are derived with all cross terms accounted for; the misadjustment follows from an exact trace identity with two-sided bounds; and the transient is controlled by an explicit geometric-plus-horizon estimate. Exact convergence can be recovered by diminishing step sizes or by iterate averaging, and it survives $n > d^2$ unharmed whenever the targets remain consistent. Notably, exact convergence is always obtain for the correlation target $c_i = \boldsymbol{v}_i^\top \boldsymbol{k}_i$ and the norm target $c_i = \|\boldsymbol{v}_i\|^2$ whenever an exact retrieval memory exists.

A2.5.2 Setting and notation

Let $n > d^2$ associations $(\boldsymbol{k}_i, \boldsymbol{v}_i, c_i)$, $1 \le i \le n$, be given. Vectorized, with $\boldsymbol{s} = \mathrm{vec}(\boldsymbol{S}) \in \mathbb{R}^{d^2}$ and $\boldsymbol{a}_i = \boldsymbol{k}_i \otimes \boldsymbol{v}_i$, therefore $\|\boldsymbol{a}_i\| = \|\boldsymbol{k}_i\|\,\|\boldsymbol{v}_i\|$. The associations form the linear system $\boldsymbol{A}\boldsymbol{s} = \boldsymbol{c}$ with $\boldsymbol{A} \in \mathbb{R}^{n\times d^2}$ stacking the rows $\boldsymbol{a}_i^\top$. We assume $\boldsymbol{A}$ has full column rank d^2 i.e. the features span $\mathbb{R}^{d^2}$. Since $n > d^2$, the system is generically *inconsistent*: no $\boldsymbol{S}$ satisfies all associations, and

$$\mathrm{L}(\boldsymbol{S}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{v}_i^\top \boldsymbol{S}\boldsymbol{k}_i - c_i\big)^2 \tag{A44}$$

has a positive minimum $\mathrm{L}_{\min} > 0$. At each step an index i_t is drawn with probabilities $p_i > 0$, independently of the past, and the rule is applied with fixed step size β satisfying

$$0 \;<\; \beta\,\|\boldsymbol{a}_i\|^2 \;<\; 2 \qquad \text{for all } i\,. \tag{A45}$$

We define

$$\bar{\boldsymbol{A}} \;:=\; \mathrm{E}\big[\boldsymbol{a}\boldsymbol{a}^\top\big] \;=\; \sum_{i=1}^{n} p_i\,\boldsymbol{a}_i\boldsymbol{a}_i^\top \;\succ\; \boldsymbol{0}\,, \qquad \bar{\boldsymbol{b}} \;:=\; \mathrm{E}\big[c\,\boldsymbol{a}\big] \;=\; \sum_{i=1}^{n} p_i\,c_i\,\boldsymbol{a}_i\,, \tag{A46}$$

$$\boldsymbol{s}^{\mathrm{LS}} \;:=\; \bar{\boldsymbol{A}}^{-1}\bar{\boldsymbol{b}} \;=\; \operatorname*{argmin}_{\boldsymbol{s}} \sum_{i=1}^{n} p_i\big(\boldsymbol{a}_i^\top \boldsymbol{s} - c_i\big)^2\,, \quad \rho_i \;:=\; c_i - \boldsymbol{a}_i^\top \boldsymbol{s}^{\mathrm{LS}} \quad \text{(optimal residuals)}\,, \tag{A47}$$

so that the normal equations read

$$\mathrm{E}\big[\rho\,\boldsymbol{a}\big] \;=\; \sum_{i=1}^{n} p_i\,\rho_i\,\boldsymbol{a}_i \;=\; \boldsymbol{0} \qquad\Longleftrightarrow\qquad \bar{\boldsymbol{b}} \;=\; \bar{\boldsymbol{A}}\,\boldsymbol{s}^{\mathrm{LS}}\,. \tag{A48}$$

For uniform sampling, $\bar{\boldsymbol{A}} = \frac{1}{n}\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{s}^{\mathrm{LS}} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{c}$ is the ordinary least-squares solution, i.e. the unique minimizer of L. Throughout, expectations without subscript are over $i \sim p$, and E_π denotes expectation of a random state distributed according to a law π. Vectorized, the rule and its error form are

$$\boldsymbol{s}_{t+1} \;=\; \boldsymbol{M}_{i_t}\boldsymbol{s}_t + \beta\,c_{i_t}\boldsymbol{a}_{i_t}\,, \qquad \boldsymbol{e}_{t+1} \;=\; \boldsymbol{M}_{i_t}\,\boldsymbol{e}_t \;+\; \beta\,\rho_{i_t}\,\boldsymbol{a}_{i_t}\,, \qquad \boldsymbol{M}_i \;:=\; \boldsymbol{I} - \beta\,\boldsymbol{a}_i\boldsymbol{a}_i^\top\,, \tag{A49}$$

with $\boldsymbol{e}_t := \boldsymbol{s}_t - \boldsymbol{s}^{\mathrm{LS}}$; the error form follows by inserting $c_i = \boldsymbol{a}_i^\top\boldsymbol{s}^{\mathrm{LS}} + \rho_i$. The constants

$$\gamma_\beta \;:=\; \min_i\,\beta\,\big(2 - \beta\,\|\boldsymbol{a}_i\|^2\big) \;>\; 0\,, \qquad \bar{\rho} \;:=\; \gamma_\beta\,\lambda_{\min}(\bar{\boldsymbol{A}}) \;\in\; (0,1]\,, \qquad \sigma^2 \;:=\; \mathrm{E}\big[\rho^2\,\|\boldsymbol{a}\|^2\big] \tag{A50}$$

appear throughout ($0 < \bar{\rho} \le 1$ is shown in Lemma A1).

A2.5.3 Theorems on expected and steady state solution

Theorem A6 (Expected solution: exact, unbiased least squares)**.** *Under* (A45)*, the expected iterate obeys the* exact *deterministic recursion*

$$\mathrm{E}\big[\boldsymbol{s}_{t+1}\big] \;=\; \big(\boldsymbol{I} - \beta\bar{\boldsymbol{A}}\big)\,\mathrm{E}\big[\boldsymbol{s}_t\big] \;+\; \beta\,\bar{\boldsymbol{b}} \qquad\Longleftrightarrow\qquad \mathrm{E}\big[\boldsymbol{e}_{t+1}\big] \;=\; \big(\boldsymbol{I} - \beta\bar{\boldsymbol{A}}\big)\,\mathrm{E}\big[\boldsymbol{e}_t\big]\,, \tag{A51}$$

whose unique fixed point is $\boldsymbol{s}^{\mathrm{LS}}$*, and*

$$\big\|\mathrm{E}[\boldsymbol{S}_t] - \boldsymbol{S}^{\mathrm{LS}}\big\|_{\mathrm{F}} \;\le\; r^t\,\big\|\boldsymbol{S}_0 - \boldsymbol{S}^{\mathrm{LS}}\big\|_{\mathrm{F}}\,, \qquad r \;:=\; \max_k\,\big|1 - \beta\lambda_k(\bar{\boldsymbol{A}})\big| \;<\; 1\,, \tag{A52}$$

so $\mathrm{E}[\boldsymbol{S}_t] \to \boldsymbol{S}^{\mathrm{LS}}$ *linearly. The convergence of the mean is* exactly unbiased*: the fixed point of* (A51) *is* $\boldsymbol{s}^{\mathrm{LS}}$ *itself, with no* $O(\beta)$ *correction.*

Theorem A7 (Steady state: stationary distribution around $\boldsymbol{S}^{\mathrm{LS}}$)**.** *Under (A45), the iterates form a Markov chain on $\mathbb{R}^{d^2}$ that has a unique stationary distribution π_β with finite second moment, and the law of $\boldsymbol{s}_t$ converges to π_β geometrically in the Wasserstein-2 metric:*

$$\mathcal{W}_2\big(\mathit{law}(\boldsymbol{s}_t),\,\pi_\beta\big)\;\le\;\big(1-\bar\rho\big)^{t/2}\,\mathcal{W}_2\big(\mathit{law}(\boldsymbol{s}_0),\,\pi_\beta\big)\,. \tag{A53}$$

Under π_β, with $\boldsymbol{e}:=\boldsymbol{s}-\boldsymbol{s}^{\mathrm{LS}}$:

(i) **Mean:** $\mathrm{E}_{\pi_\beta}[\boldsymbol{S}]=\boldsymbol{S}^{\mathrm{LS}}$ *exactly.*

(ii) **Covariance:** $\boldsymbol{V}_\beta:=\mathrm{E}_{\pi_\beta}[\boldsymbol{e}\boldsymbol{e}^\top]$ *satisfies the* exact *stationarity equation*

$$\bar{\boldsymbol{A}}\boldsymbol{V}_\beta+\boldsymbol{V}_\beta\bar{\boldsymbol{A}}\;-\;\beta\,\mathrm{E}\big[\big(\boldsymbol{a}^\top\boldsymbol{V}_\beta\,\boldsymbol{a}\big)\,\boldsymbol{a}\boldsymbol{a}^\top\big]\;=\;\beta\,\mathrm{E}\big[\rho^2\,\boldsymbol{a}\boldsymbol{a}^\top\big]\,, \tag{A54}$$

and obeys the two-sided trace bounds

$$\frac{\beta\,\sigma^2}{2}\;\le\;\mathrm{tr}\,\big(\bar{\boldsymbol{A}}\boldsymbol{V}_\beta\big)\;\le\;\frac{\beta\,\sigma^2}{2-\beta\max_i\|\boldsymbol{a}_i\|^2}\,, \tag{A55}$$

$$\mathrm{E}_{\pi_\beta}\Big[\big\|\boldsymbol{S}-\boldsymbol{S}^{\mathrm{LS}}\big\|_{\mathrm{F}}^2\Big]\;\le\;\frac{\beta\,\sigma^2}{\big(2-\beta\max_i\|\boldsymbol{a}_i\|^2\big)\,\lambda_{\min}(\bar{\boldsymbol{A}})}\,;$$

in particular $\boldsymbol{V}_\beta=O(\beta)$, and to leading order (A54) is the Lyapunov equation $\bar{\boldsymbol{A}}\boldsymbol{V}+\boldsymbol{V}\bar{\boldsymbol{A}}=\beta\mathrm{E}[\rho^2\boldsymbol{a}\boldsymbol{a}^\top]$.

(iii) **Steady-state loss (misadjustment):** *for uniform sampling,*

$$\mathrm{E}_{\pi_\beta}\big[\mathrm{L}(\boldsymbol{S})\big]\;-\;\mathrm{L}_{\min}\;=\;\mathrm{tr}\,\big(\bar{\boldsymbol{A}}\boldsymbol{V}_\beta\big)\;=\;\frac{\beta}{2}\Big(\sigma^2+\mathrm{E}\big[\big(\boldsymbol{a}^\top\boldsymbol{V}_\beta\boldsymbol{a}\big)\,\|\boldsymbol{a}\|^2\big]\Big)\,, \tag{A56}$$

an exact identity, whence

$$\frac{\beta}{2}\,\sigma^2\;\le\;\mathrm{E}_{\pi_\beta}\big[\mathrm{L}(\boldsymbol{S})\big]-\mathrm{L}_{\min}\;\le\;\frac{\beta}{2}\,\sigma^2\cdot\frac{2}{2-\beta\max_i\|\boldsymbol{a}_i\|^2}\,,\qquad\sigma^2=\mathrm{E}\big[\rho_i^2\,\|\boldsymbol{k}_i\|^2\,\|\boldsymbol{v}_i\|^2\big]\,, \tag{A57}$$

i.e. $\mathrm{E}_{\pi_\beta}[\mathrm{L}]=\mathrm{L}_{\min}+\frac{\beta}{2}\sigma^2+O(\beta^2)$. This is the classical Widrow–Hoff LMS misadjustment for the bilinear feature $\boldsymbol{a}=\boldsymbol{k}\otimes\boldsymbol{v}$.

(iv) **Transient:** *with $\boldsymbol{w}:=\mathrm{E}\big[\rho\,\|\boldsymbol{a}\|^2\,\boldsymbol{a}\big]$,*

$$\mathrm{E}\,\|\boldsymbol{e}_t\|^2\;\le\;\Big(1-\tfrac{\bar\rho}{2}\Big)^t\,\|\boldsymbol{e}_0\|^2\;+\;\frac{2\beta^2\sigma^2}{\bar\rho}\;+\;\frac{4\beta^4\,\|\boldsymbol{w}\|^2}{\bar\rho^2}\,, \tag{A58}$$

a geometric approach to an $O(\beta)$ horizon (note $\bar\rho=\Theta(\beta)$): overload does not slow the transient, it only replaces the limit point by a limit distribution.

A2.5.4 Proofs of the theorems

Throughout, $\mathcal{F}_t:=\sigma(\boldsymbol{s}_0,i_0,\dots,i_{t-1})$ denotes the filtration of the past; i_t is independent of $\mathcal{F}_t$ by assumption. We repeatedly use that (A49) is *linear* in the state with coefficients independent of the state.

Two lemmas crucial for the proofs of the theorems.

Lemma A1 (One-step mean-square contraction of the homogeneous part)**.** *For every fixed $\boldsymbol{x}\in\mathbb{R}^{d^2}$,*

$$\mathrm{E}\big[\,\|\boldsymbol{M}_i\,\boldsymbol{x}\|^2\,\big]\;=\;\|\boldsymbol{x}\|^2\;-\;\sum_{i=1}^{n}p_i\,\beta\big(2-\beta\,\|\boldsymbol{a}_i\|^2\big)\big(\boldsymbol{a}_i^\top\boldsymbol{x}\big)^2\;\le\;\big(1-\bar\rho\big)\,\|\boldsymbol{x}\|^2\,, \tag{A59}$$

and $\bar\rho\in(0,1]$.

Proof. Since $\boldsymbol{M}_i$ is symmetric,

$$\|\boldsymbol{M}_i\boldsymbol{x}\|^2\;=\;\boldsymbol{x}^\top\big(\boldsymbol{I}-\beta\boldsymbol{a}_i\boldsymbol{a}_i^\top\big)^2\boldsymbol{x}\;=\;\|\boldsymbol{x}\|^2-2\beta\big(\boldsymbol{a}_i^\top\boldsymbol{x}\big)^2+\beta^2\,\|\boldsymbol{a}_i\|^2\,\big(\boldsymbol{a}_i^\top\boldsymbol{x}\big)^2\,, \tag{A60}$$

which is the equality in (A59) after averaging over $i\sim p$. By (A45), $\beta(2-\beta\,\|\boldsymbol{a}_i\|^2)\ge\gamma_\beta>0$ for every i, and $\sum_i p_i(\boldsymbol{a}_i^\top\boldsymbol{x})^2=\boldsymbol{x}^\top\bar{\boldsymbol{A}}\boldsymbol{x}\ge\lambda_{\min}(\bar{\boldsymbol{A}})\,\|\boldsymbol{x}\|^2$, hence the inequality with $\bar\rho=\gamma_\beta\lambda_{\min}(\bar{\boldsymbol{A}})$. Finally $\bar\rho\le1$ because the left-hand side of (A59) is nonnegative for every $\boldsymbol{x}$. We assumed that $\bar{\boldsymbol{A}}$ has full rank, therefore $0<\lambda_{\min}(\bar{\boldsymbol{A}})$ and with $0<\gamma_\beta$ we have $0<\bar\rho$. □

Lemma A2 (The chain lives in $\mathcal{P}_2$ and its kernel is a $\mathcal{W}_2$-contraction)**.** *Let P denote the transition kernel of (A49) and let $\mathcal{P}_2$ be the space of probability laws on $\mathbb{R}^{d^2}$ with finite second moment, equipped with the Wasserstein-2 metric $\mathcal{W}_2$. Then P maps $\mathcal{P}_2$ into $\mathcal{P}_2$, and for all $\mu, \nu \in \mathcal{P}_2$,*

$$\mathcal{W}_2\big(\mu P,\, \nu P\big) \;\le\; \sqrt{1-\bar{\rho}}\;\mathcal{W}_2(\mu,\nu)\,. \tag{A61}$$

Consequently P has a unique invariant law $\pi_\beta \in \mathcal{P}_2$ and (A53) holds.

Proof. P preserves $\mathcal{P}_2$: for $\boldsymbol{e} \sim \mu \in \mathcal{P}_2$ and an independent index i, $\mathrm{E}\,\|\boldsymbol{M}_i\boldsymbol{e} + \beta\rho_i\boldsymbol{a}_i\|^2 \le 2\,\mathrm{E}\,\|\boldsymbol{M}_i\boldsymbol{e}\|^2 + 2\beta^2\,\mathrm{E}[\rho^2\,\|\boldsymbol{a}\|^2] \le 2\,\mathrm{E}\,\|\boldsymbol{e}\|^2 + 2\beta^2\sigma^2 < \infty$, using Lemma A1.
Contraction: let $(\boldsymbol{x}, \boldsymbol{y})$ be an optimal $\mathcal{W}_2$-coupling of (μ, ν) and couple the two chains *synchronously*: draw one index i (independent of $(\boldsymbol{x}, \boldsymbol{y})$) and set $\boldsymbol{x}' = \boldsymbol{M}_i\boldsymbol{x} + \beta\rho_i\boldsymbol{a}_i$, $\boldsymbol{y}' = \boldsymbol{M}_i\boldsymbol{y} + \beta\rho_i\boldsymbol{a}_i$. The additive terms cancel in the difference,

$$\boldsymbol{x}' - \boldsymbol{y}' \;=\; \boldsymbol{M}_i\,(\boldsymbol{x} - \boldsymbol{y})\,, \tag{A62}$$

so by Lemma A1, first conditionally on $(\boldsymbol{x}, \boldsymbol{y})$ and then in expectation,

$$\mathrm{E}\,\|\boldsymbol{x}' - \boldsymbol{y}'\|^2 \;\le\; (1-\bar{\rho})\,\mathrm{E}\,\|\boldsymbol{x} - \boldsymbol{y}\|^2 \;=\; (1-\bar{\rho})\,\mathcal{W}_2(\mu,\nu)^2\,. \tag{A63}$$

Since $(\boldsymbol{x}', \boldsymbol{y}')$ is *some* coupling of $(\mu P, \nu P)$, $\mathcal{W}_2(\mu P, \nu P)^2 \le \mathrm{E}\,\|\boldsymbol{x}' - \boldsymbol{y}'\|^2$, which is (A61).
Conclusion: $(\mathcal{P}_2, \mathcal{W}_2)$ is a complete metric space (see, e.g., (Villani, 2009, Theorem 6.18)), so Banach's fixed-point theorem yields a unique $\pi_\beta \in \mathcal{P}_2$ with $\pi_\beta P = \pi_\beta$ and $\mathcal{W}_2(\mu P^t, \pi_\beta) \le (1-\bar{\rho})^{t/2}\mathcal{W}_2(\mu, \pi_\beta)$ for every $\mu \in \mathcal{P}_2$, which is (A53). Convergence in $\mathcal{W}_2$ implies convergence in distribution together with convergence of first and second moments. □

Comments to the proof of Lemma A2 in 9 bullet points.
• *1) Overview.* How a mixture of non-contractive matrices produces a contraction: the kernel P and the coupling argument. The matrices $\boldsymbol{M}_i$ are not themselves contractions - by Lemma A1, $\|\boldsymbol{M}_i\|_2 = 1$ for *every* single i, since $\boldsymbol{M}_i$ has eigenvalue 1 on $\boldsymbol{a}_i^{\perp}$. Contractivity in Lemma A2 is a property of the *mixture* P, not of any individual $\boldsymbol{M}_i$, and P is not a matrix at all: it is a Markov kernel acting on laws in $\mathcal{P}_2$, and the relevant notion of size for it is the Wasserstein metric $\mathcal{W}_2$ on $\mathcal{P}_2$ itself - not any norm on $\mathbb{R}^{d^2}$ from which P's contraction could be read off the $\boldsymbol{M}_i$'s directly. This note spells out how P is built from the $\boldsymbol{M}_i$'s, in what sense P does and does not have an "operator norm," and why the coupling argument nonetheless produces the contraction (A61).
• *2) Constructing P from the $\boldsymbol{M}_i$.* For a fixed state $\boldsymbol{x} \in \mathbb{R}^{d^2}$, define the n deterministic affine maps

$$\Phi_i(\boldsymbol{x}) \;:=\; \boldsymbol{M}_i\boldsymbol{x} + \beta\rho_i\boldsymbol{a}_i\,, \qquad i = 1, \dots, n\,. \tag{A64}$$

One step of the chain from $\boldsymbol{x}$ lands at $\Phi_i(\boldsymbol{x})$ with probability p_i, so the transition kernel is the mixture of point masses

$$P(\boldsymbol{x}, \cdot) \;:=\; \sum_{i=1}^{n} p_i\,\delta_{\Phi_i(\boldsymbol{x})}(\cdot)\,, \tag{A65}$$

and, for a law μ, μP is defined by $(\mu P)(B) = \int P(\boldsymbol{x}, B)\,\mu(d\boldsymbol{x})$ - equivalently, μP is the law of $\Phi_I(\boldsymbol{X})$ for $\boldsymbol{X} \sim \mu$ and $I \sim p$ drawn *independently* of $\boldsymbol{X}$.
B is a Borel (measurable) subset of $\mathbb{R}^{d^2}$ - an arbitrary *test set*, held fixed, to which the kernel assigns a probability. Concretely,

$$P(\boldsymbol{x}, B) \;:=\; \Pr\big(\Phi_I(\boldsymbol{x}) \in B\big) \;=\; \sum_{i=1}^{n} p_i\;\mathbb{1}\big[\Phi_i(\boldsymbol{x}) \in B\big] \;=\; \sum_{i=1}^{n} p_i\;\delta_{\Phi_i(\boldsymbol{x})}(B)\,, \tag{A66}$$

where $\Phi_i(\boldsymbol{x}) = \boldsymbol{M}_i\boldsymbol{x} + \beta\rho_i\boldsymbol{a}_i$ is the (fixed, deterministic) image of $\boldsymbol{x}$ under transition i, and $I \sim p$ is the random index selecting which transition fires. In words: $P(\boldsymbol{x}, B)$ is the probability that one step of the chain, started at $\boldsymbol{x}$, lands in B - obtained by summing the weights p_i of every index i whose image $\Phi_i(\boldsymbol{x})$ happens to fall inside B.
P as a function of each argument separately. The kernel $P : \mathbb{R}^{d^2} \times \mathcal{B}(\mathbb{R}^{d^2}) \to [0, 1]$ plays a different role in each of its two slots:

- Fixed $\boldsymbol{x}$, varying B: $P(\boldsymbol{x}, \cdot)$ is itself a probability measure on $\mathbb{R}^{d^2}$ - concretely, the discrete measure

$$P(\boldsymbol{x}, \cdot) \;=\; \sum_{i=1}^{n} p_i \, \delta_{\Phi_i(\boldsymbol{x})}(\cdot) \tag{A67}$$

  putting mass p_i on the point $\Phi_i(\boldsymbol{x})$ for each i (masses are summed if several $\Phi_i(\boldsymbol{x})$ coincide). This is the law of the chain's next state, given that it is currently at $\boldsymbol{x}$.
- Fixed B, varying $\boldsymbol{x}$: $P(\cdot, B)$ is a bounded measurable function of $\boldsymbol{x}$, since each indicator $\mathbb{1}[\Phi_i(\boldsymbol{x}) \in B]$ is measurable in $\boldsymbol{x}$ and the sum is finite.

This is the standard convention for a Markov transition kernel: measurable in the first argument, a probability measure in the second.

• *3) the pushforward measure μP is well defined.* For a law μ of the current state, the one-step-ahead law μP is defined by integrating the *second* argument B (held fixed) against μ acting on the *first* argument $\boldsymbol{x}$:

$$(\mu P)(B) \;:=\; \int_{\mathbb{R}^{d^2}} P(\boldsymbol{x}, B)\, \mu(d\boldsymbol{x})\,. \tag{A68}$$

Equivalently, μP is the law of $\Phi_I(\boldsymbol{X})$ for $\boldsymbol{X} \sim \mu$ and $I \sim p$ drawn independently of $\boldsymbol{X}$: draw a random current state $\boldsymbol{X}$, draw an independent random index I, and output $\Phi_I(\boldsymbol{X})$. The two roles of B above are exactly what make this integral meaningful: for each fixed $\boldsymbol{x}$ in the integral, $P(\boldsymbol{x}, B)$ is a number in $[0, 1]$ (the probability mass the one-step law from $\boldsymbol{x}$ places on B); integrating that number against $\mu(d\boldsymbol{x})$ averages it over the randomness in the current state, and the result, as a function of B, is again a probability measure. It is the law μP appearing throughout the steady-state analysis of the rule.

• *4) P has no operator norm on $\mathbb{R}^{d^2}$.* Thus P is linear as an operator on (signed) measures ($(\alpha\mu + (1-\alpha)\nu)P = \alpha\mu P + (1-\alpha)\nu P$), but this linearity lives on the infinite-dimensional space of measures, not on $\mathbb{R}^{d^2}$: there is no $d^2 \times d^2$ matrix representing P, so P has *no* operator norm on $(\mathbb{R}^{d^2}, \|\cdot\|)$ the way each $\boldsymbol{M}_i$ does. What P *does* have is an operator norm in the sense appropriate to its actual domain: a Lipschitz modulus on the metric space $(\mathcal{P}_2, \mathcal{W}_2)$, namely the smallest L such that $\mathcal{W}_2(\mu P, \nu P) \le L\, \mathcal{W}_2(\mu, \nu)$ for all $\mu, \nu \in \mathcal{P}_2$. That modulus is exactly what (A61) bounds by $\sqrt{1-\bar{\rho}} < 1$. It cannot be read off a per-map bound such as $\max_i \|\boldsymbol{M}_i\|_2$, which equals 1 by Lemma A1 and would certify no contraction whatsoever; obtaining a modulus strictly below 1 requires the coupling argument below, precisely because it is a statement about P acting on laws, not about any $\boldsymbol{M}_i$ acting on $\mathbb{R}^{d^2}$.

• *5) The Wasserstein metric is essential to prove the contraction.* The definition

$$\mathcal{W}_2(\mu, \nu)^2 \;=\; \inf_{\gamma \in \Gamma(\mu,\nu)} \mathrm{E}_\gamma\big[\, \|\boldsymbol{X} - \boldsymbol{Y}\|^2 \,\big] \tag{A69}$$

is an *infimum over all couplings* γ of (μ, ν) (joint laws with the prescribed marginals). This has an immediate but crucial consequence: to upper-bound $\mathcal{W}_2(\mu P, \nu P)$ it suffices to exhibit *any single* coupling of $(\mu P, \nu P)$ and compute its second moment. Note that we never need the optimal coupling of the outputs, let alone an explicit formula for P as an operator.

• *6) The synchronous coupling.* We are free to choose how the randomness that drives one step of *each* chain is generated, and we exploit that freedom by reusing the same random index for both chains. Concretely: let $(\boldsymbol{x}, \boldsymbol{y})$ be an *optimal* coupling of (μ, ν), i.e. a joint law with $\boldsymbol{x} \sim \mu$, $\boldsymbol{y} \sim \nu$, and $\mathrm{E}\,\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \mathcal{W}_2(\mu, \nu)^2$ (such a coupling exists because $\mathbb{R}^{d^2}$ is Polish; see (Villani, 2009, Theorem 4.1)). Draw a single index $i \sim p$, independent of $(\boldsymbol{x}, \boldsymbol{y})$, and set

$$\boldsymbol{x}' \;:=\; \Phi_i(\boldsymbol{x}) \;=\; \boldsymbol{M}_i \boldsymbol{x} + \beta\rho_i \boldsymbol{a}_i\,, \qquad \boldsymbol{y}' \;:=\; \Phi_i(\boldsymbol{y}) \;=\; \boldsymbol{M}_i \boldsymbol{y} + \beta\rho_i \boldsymbol{a}_i\,. \tag{A70}$$

Marginally, $\boldsymbol{x}' \sim \mu P$ and $\boldsymbol{y}' \sim \nu P$ by the very definition (A65) of μP, applied separately to each starting law; hence $(\boldsymbol{x}', \boldsymbol{y}')$ *is* a valid coupling of $(\mu P, \nu P)$. This coupling is not necessarily the optimal one, but that is all the argument requires.

• *7) The additive target terms cancel.* Because $\boldsymbol{x}'$ and $\boldsymbol{y}'$ are built from the *same* draw i, the inhomogeneous term $\beta\rho_i \boldsymbol{a}_i$is identical in both and cancels exactly in the difference:

$$\boldsymbol{x}' - \boldsymbol{y}' \;=\; \big(\boldsymbol{M}_i \boldsymbol{x} + \beta\rho_i \boldsymbol{a}_i\big) - \big(\boldsymbol{M}_i \boldsymbol{y} + \beta\rho_i \boldsymbol{a}_i\big) \;=\; \boldsymbol{M}_i\,(\boldsymbol{x} - \boldsymbol{y})\,. \tag{A71}$$

The part of the update that cancels out is exactly the part carrying the targets c_i. This is the crux of the construction: it converts a statement about two *inhomogeneous, randomly driven* affine updates

into a statement about a single *homogeneous linear* map, $\boldsymbol{M}_i$, applied to the fixed vector $\boldsymbol{x}-\boldsymbol{y}$. This is exactly the object Lemma A1 controls. Had we instead coupled the two chains using *independent* indices for $\boldsymbol{x}'$ and $\boldsymbol{y}'$, the noise terms would not cancel and the argument would not go through; the whole content of the synchronous-coupling trick is this cancellation.

• *8) Deriving the contraction.* Since i is independent of $(\boldsymbol{x},\boldsymbol{y})$, condition on $(\boldsymbol{x},\boldsymbol{y})$ first and average over i:

$$\mathrm{E}\big[\,\|\boldsymbol{x}'-\boldsymbol{y}'\|^2\,\big|\,\boldsymbol{x},\boldsymbol{y}\big] \;=\; \sum_{i=1}^{n} p_i\,\|\boldsymbol{M}_i(\boldsymbol{x}-\boldsymbol{y})\|^2 \;\le\; \big(1-\bar{\rho}\big)\,\|\boldsymbol{x}-\boldsymbol{y}\|^2 \tag{A72}$$

by Lemma A1, applied pointwise to the vector $\boldsymbol{x}-\boldsymbol{y}$ (fixed once we condition). Taking the outer expectation over $(\boldsymbol{x},\boldsymbol{y})$ and using that this coupling was chosen optimal,

$$\mathrm{E}\,\|\boldsymbol{x}'-\boldsymbol{y}'\|^2 \;\le\; \big(1-\bar{\rho}\big)\,\mathrm{E}\,\|\boldsymbol{x}-\boldsymbol{y}\|^2 \;=\; \big(1-\bar{\rho}\big)\,\mathcal{W}_2(\mu,\nu)^2\,. \tag{A73}$$

Finally, since $(\boldsymbol{x}',\boldsymbol{y}')$ is *a* coupling of $(\mu P,\nu P)$, the infimum defining $\mathcal{W}_2(\mu P,\nu P)^2$ is at most the cost of this particular coupling:

$$\mathcal{W}_2(\mu P,\nu P)^2 \;\le\; \mathrm{E}\,\|\boldsymbol{x}'-\boldsymbol{y}'\|^2 \;\le\; \big(1-\bar{\rho}\big)\,\mathcal{W}_2(\mu,\nu)^2\,, \tag{A74}$$

and taking square roots recovers (A61) exactly.

• *9) Summary of the approach.* Three ingredients combine, none of which is a bound on P as an operator on $\mathbb{R}^{d^2}$: (1) $\mathcal{W}_2$ is an infimum over couplings, so any one coupling of the images gives a valid upper bound; (2) the synchronous coupling is reusing the same random index for both copies and is a legitimate choice precisely because P is defined by mixing over that same randomness, and it makes the target-dependent noise cancel identically; (3) what remains is a bound on $\mathrm{E}_i\,\|\boldsymbol{M}_i\boldsymbol{z}\|^2$ for the *fixed* vector $\boldsymbol{z}=\boldsymbol{x}-\boldsymbol{y}$, which is exactly the pointwise mean-square contraction already established in Lemma A1. This lemma is a statement purely about the $\boldsymbol{M}_i$'s and $\bar{\boldsymbol{A}}=\sum_i p_i\boldsymbol{a}_i\boldsymbol{a}_i^\top$, requiring no coupling of its own. The contraction of P in $\mathcal{W}_2$ is thus assembled entirely from the per-index contraction of Lemma A1, transported to the space of distributions by the coupling device. It is not, and could not be, a statement recoverable from $\max_i\|\boldsymbol{M}_i\|_2$ alone.

Proof of Theorem A6.

Proof. **Exact mean recursion.** The update (A49) is linear in $\boldsymbol{s}_t$, and i_t is independent of $\mathcal{F}_t\supseteq\sigma(\boldsymbol{s}_t)$. Hence, conditioning on $\mathcal{F}_t$,

$$\mathrm{E}\big[\boldsymbol{s}_{t+1}\,\big|\,\mathcal{F}_t\big] \;=\; \mathrm{E}\big[\boldsymbol{M}_{i_t}\big]\,\boldsymbol{s}_t+\beta\,\mathrm{E}\big[c_{i_t}\boldsymbol{a}_{i_t}\big] \;=\; \big(\boldsymbol{I}-\beta\bar{\boldsymbol{A}}\big)\,\boldsymbol{s}_t+\beta\,\bar{\boldsymbol{b}}\,, \tag{A75}$$

and the tower property gives (A51). This equation holds with equality and is not an approximation: no linearization was needed because the update itself is linear. Subtracting the fixed-point relation $\boldsymbol{s}^{\mathrm{LS}}=(\boldsymbol{I}-\beta\bar{\boldsymbol{A}})\boldsymbol{s}^{\mathrm{LS}}+\beta\bar{\boldsymbol{b}}$, which holds by the normal equations (A48) ($\bar{\boldsymbol{b}}=\bar{\boldsymbol{A}}\boldsymbol{s}^{\mathrm{LS}}$), yields the error form and, by induction,

$$\mathrm{E}\big[\boldsymbol{e}_t\big] \;=\; \big(\boldsymbol{I}-\beta\bar{\boldsymbol{A}}\big)^t\,\mathrm{E}\big[\boldsymbol{e}_0\big]\,. \tag{A76}$$

Spectral radius. $\bar{\boldsymbol{A}}$ is symmetric positive definite; let $0<\lambda_{\min}=\lambda_1\le\cdots\le\lambda_{d^2}=\lambda_{\max}$ be its eigenvalues. For any unit vector $\boldsymbol{x}$, $\boldsymbol{x}^\top\bar{\boldsymbol{A}}\boldsymbol{x}=\sum_i p_i(\boldsymbol{a}_i^\top\boldsymbol{x})^2\le\max_i\|\boldsymbol{a}_i\|^2$, so

$$\beta\,\lambda_{\max} \;\le\; \beta\,\max_i\|\boldsymbol{a}_i\|^2 \;<\; 2 \tag{A77}$$

by (A45). Hence, every eigenvalue $1-\beta\lambda_k$ of $\boldsymbol{I}-\beta\bar{\boldsymbol{A}}$ lies in $(-1,1)$: $1-\beta\lambda_k<1$ since $\beta\lambda_k>0$, and $1-\beta\lambda_k>-1$ since $\beta\lambda_k<2$. The matrix is symmetric, so its spectral norm equals $r=\max_k|1-\beta\lambda_k|<1$ and $\|\mathrm{E}[\boldsymbol{e}_t]\|\le r^t\,\|\boldsymbol{e}_0\|$, i.e. linear convergence of $\mathrm{E}[\boldsymbol{S}_t]$ to $\boldsymbol{S}^{\mathrm{LS}}$ in Frobenius norm ($\|\boldsymbol{s}\|_2=\|\boldsymbol{S}\|_{\mathrm{F}}$).

Exact unbiasedness. The fixed point of the affine map $\boldsymbol{z}\mapsto(\boldsymbol{I}-\beta\bar{\boldsymbol{A}})\boldsymbol{z}+\beta\bar{\boldsymbol{b}}$ is unique ($\beta\bar{\boldsymbol{A}}$ invertible) and equals $\bar{\boldsymbol{A}}^{-1}\bar{\boldsymbol{b}}=\boldsymbol{s}^{\mathrm{LS}}$ for *every* β in the admissible range: the limit of the mean does not depend on β and carries no $O(\beta)$ correction. □

Proof of Theorem A7.

Proof. Existence, uniqueness, membership in $\mathcal{P}_2$, and the geometric $\mathcal{W}_2$-convergence (A53) are Lemma A2. We have to prove (i)–(iv). In stationarity we write $\boldsymbol{e}\sim\pi_\beta$ (centered at $\boldsymbol{s}^{\mathrm{LS}}$), draw a fresh index i independent of $\boldsymbol{e}$, and use that

$$\boldsymbol{e}' \;:=\; \boldsymbol{M}_i\,\boldsymbol{e}+\beta\,\rho_i\,\boldsymbol{a}_i \;\overset{d}{=}\; \boldsymbol{e}\,, \tag{A78}$$

with all first and second moments finite ($\pi_\beta \in \mathcal{P}_2$). Here $\overset{d}{=}$ means equality in distribution.

(i) Stationary mean. Taking expectations in (A78) and using the independence of i and $\boldsymbol{e}$ together with the normal equations $\mathrm{E}[\rho \boldsymbol{a}] = \mathbf{0}$:

$$\mathrm{E}[\boldsymbol{e}] \;=\; \mathrm{E}[\boldsymbol{e}'] \;=\; \mathrm{E}\big[\boldsymbol{M}_i\big]\,\mathrm{E}[\boldsymbol{e}] + \beta\,\mathrm{E}\big[\rho_i \boldsymbol{a}_i\big] \;=\; \big(\boldsymbol{I} - \beta \bar{\boldsymbol{A}}\big)\,\mathrm{E}[\boldsymbol{e}]\,, \tag{A79}$$

hence $\beta \bar{\boldsymbol{A}}\,\mathrm{E}[\boldsymbol{e}] = \mathbf{0}$, and since $\bar{\boldsymbol{A}}$ is invertible, $\mathrm{E}[\boldsymbol{e}] = \mathbf{0}$, i.e. $\mathrm{E}_{\pi_\beta}[\boldsymbol{S}] = \boldsymbol{S}^{\mathrm{LS}}$, which holds exactly.

(ii) Exact covariance equation. Expand $\boldsymbol{e}'\boldsymbol{e}'^{\top}$ from (A78):

$$\boldsymbol{e}'\boldsymbol{e}'^{\top} \;=\; \boldsymbol{M}_i \boldsymbol{e}\boldsymbol{e}^{\top} \boldsymbol{M}_i \;+\; \beta\,\rho_i\,\boldsymbol{M}_i \boldsymbol{e}\,\boldsymbol{a}_i^{\top} \;+\; \beta\,\rho_i\,\boldsymbol{a}_i \boldsymbol{e}^{\top} \boldsymbol{M}_i \;+\; \beta^2 \rho_i^2\,\boldsymbol{a}_i \boldsymbol{a}_i^{\top}\,. \tag{A80}$$

Take expectations term by term, using the independence of i and $\boldsymbol{e}$ and writing $\boldsymbol{V}_\beta = \mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\top}]$.

Homogeneous term. Since $\boldsymbol{M}_i \boldsymbol{V} \boldsymbol{M}_i = \boldsymbol{V} - \beta\big(\boldsymbol{a}_i \boldsymbol{a}_i^{\top} \boldsymbol{V} + \boldsymbol{V} \boldsymbol{a}_i \boldsymbol{a}_i^{\top}\big) + \beta^2 \big(\boldsymbol{a}_i^{\top} \boldsymbol{V} \boldsymbol{a}_i\big) \boldsymbol{a}_i \boldsymbol{a}_i^{\top}$ (using $\boldsymbol{a}_i \boldsymbol{a}_i^{\top} \boldsymbol{V} \boldsymbol{a}_i \boldsymbol{a}_i^{\top} = (\boldsymbol{a}_i^{\top} \boldsymbol{V} \boldsymbol{a}_i)\,\boldsymbol{a}_i \boldsymbol{a}_i^{\top}$),

$$\mathrm{E}\big[\boldsymbol{M}_i \boldsymbol{e}\boldsymbol{e}^{\top} \boldsymbol{M}_i\big] \;=\; \boldsymbol{V}_\beta - \beta\big(\bar{\boldsymbol{A}} \boldsymbol{V}_\beta + \boldsymbol{V}_\beta \bar{\boldsymbol{A}}\big) + \beta^2\,\mathrm{E}\big[\big(\boldsymbol{a}^{\top} \boldsymbol{V}_\beta \boldsymbol{a}\big) \boldsymbol{a}\boldsymbol{a}^{\top}\big]\,. \tag{A81}$$

Cross terms. They are *linear* in $\boldsymbol{e}$, so independence factorizes them through $\mathrm{E}[\boldsymbol{e}]$, which vanishes by part (i):

$$\mathrm{E}\big[\rho_i\,\boldsymbol{M}_i \boldsymbol{e}\,\boldsymbol{a}_i^{\top}\big] \;=\; \mathrm{E}_i\big[\rho_i\,\boldsymbol{M}_i\,\mathrm{E}[\boldsymbol{e}]\,\boldsymbol{a}_i^{\top}\big] \;=\; \mathbf{0}\,, \tag{A82}$$

and likewise for its transpose. This is the step where exact unbiasedness pays off: no $O(\beta^2)$ cross contribution survives.

Stationarity. Setting $\mathrm{E}[\boldsymbol{e}'\boldsymbol{e}'^{\top}] = \boldsymbol{V}_\beta$ and collecting,

$$\boldsymbol{V}_\beta \;=\; \boldsymbol{V}_\beta - \beta\big(\bar{\boldsymbol{A}} \boldsymbol{V}_\beta + \boldsymbol{V}_\beta \bar{\boldsymbol{A}}\big) + \beta^2 \mathrm{E}\big[\big(\boldsymbol{a}^{\top} \boldsymbol{V}_\beta \boldsymbol{a}\big) \boldsymbol{a}\boldsymbol{a}^{\top}\big] + \beta^2 \mathrm{E}\big[\rho^2 \boldsymbol{a}\boldsymbol{a}^{\top}\big]\,; \tag{A83}$$

dividing by β gives (A54), which is an exact equation, not an expansion.

Trace bounds. Take the trace of (A54):

$$2\,\mathrm{tr}\,\big(\bar{\boldsymbol{A}} \boldsymbol{V}_\beta\big) \;=\; \beta\,\mathrm{E}\big[\big(\boldsymbol{a}^{\top} \boldsymbol{V}_\beta \boldsymbol{a}\big)\,\|\boldsymbol{a}\|^2\big] \;+\; \beta\,\sigma^2\,. \tag{A84}$$

Since $\boldsymbol{V}_\beta \succeq \mathbf{0}$, we have $\boldsymbol{a}^{\top} \boldsymbol{V}_\beta \boldsymbol{a} \geq 0$, which gives the lower bound in (A55); and $\mathrm{E}[(\boldsymbol{a}^{\top} \boldsymbol{V}_\beta \boldsymbol{a})\,\|\boldsymbol{a}\|^2] \leq \max_i \|\boldsymbol{a}_i\|^2\,\mathrm{E}[\boldsymbol{a}^{\top} \boldsymbol{V}_\beta \boldsymbol{a}] = \max_i \|\boldsymbol{a}_i\|^2\,\mathrm{tr}(\bar{\boldsymbol{A}} \boldsymbol{V}_\beta)$, so

$$\big(2 - \beta \max_i \|\boldsymbol{a}_i\|^2\big)\,\mathrm{tr}\,\big(\bar{\boldsymbol{A}} \boldsymbol{V}_\beta\big) \;\leq\; \beta\,\sigma^2\,, \tag{A85}$$

the upper bound (the prefactor is positive by (A45)). Finally $\mathrm{tr}\,\boldsymbol{V}_\beta \leq \mathrm{tr}(\bar{\boldsymbol{A}} \boldsymbol{V}_\beta)/\lambda_{\min}(\bar{\boldsymbol{A}})$ gives the Frobenius bound in (A55), and $\mathrm{tr}\,\boldsymbol{V}_\beta = O(\beta)$ makes the quartic term in (A54) $O(\beta^2)$, leaving the Lyapunov equation at leading order.

(iii) Misadjustment. For uniform sampling, $\bar{\boldsymbol{A}} = \frac{1}{n} \boldsymbol{A}^{\top} \boldsymbol{A}$ and $\mathrm{L}(\boldsymbol{s}) = \frac{1}{n}\,\|\boldsymbol{A}\boldsymbol{s} - \boldsymbol{c}\|^2$. Decompose with $\boldsymbol{A}\boldsymbol{s} - \boldsymbol{c} = \boldsymbol{A}\boldsymbol{e} - \boldsymbol{\rho}$, where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)^{\top}$: since $\boldsymbol{A}^{\top} \boldsymbol{\rho} = n\,\mathrm{E}[\rho \boldsymbol{a}] = \mathbf{0}$,

$$\mathrm{L}(\boldsymbol{s}) \;=\; \tfrac{1}{n}\,\|\boldsymbol{A}\boldsymbol{e}\|^2 - \tfrac{2}{n}\,\boldsymbol{e}^{\top} \boldsymbol{A}^{\top} \boldsymbol{\rho} + \tfrac{1}{n}\,\|\boldsymbol{\rho}\|^2 \;=\; \tfrac{1}{n}\,\|\boldsymbol{A}\boldsymbol{e}\|^2 + \mathrm{L}_{\min}\,, \tag{A86}$$

where $\mathrm{L}_{\min} = \frac{1}{n}\,\|\boldsymbol{\rho}\|^2$. This also re-proves that $\boldsymbol{s}^{\mathrm{LS}}$ minimizes L. Taking stationary expectations,

$$\mathrm{E}_{\pi_\beta}[\mathrm{L}] - \mathrm{L}_{\min} \;=\; \tfrac{1}{n}\,\mathrm{E}\big[\boldsymbol{e}^{\top} \boldsymbol{A}^{\top} \boldsymbol{A}\,\boldsymbol{e}\big] \;=\; \mathrm{tr}\,\big(\bar{\boldsymbol{A}}\,\boldsymbol{V}_\beta\big)\,, \tag{A87}$$

and inserting the exact trace identity (A84) gives (A56). The two-sided bound (A57) follows from the trace bounds (A55). Since the upper and lower bounds differ by a factor $2/(2 - \beta \max_i \|\boldsymbol{a}_i\|^2) = 1 + O(\beta)$, the excess loss equals $\frac{\beta}{2}\sigma^2(1 + O(\beta))$.

(iv) Transient. Condition on $\mathcal{F}_t$ and expand (A49):

$$\mathrm{E}\big[\,\|\boldsymbol{e}_{t+1}\|^2 \,\big|\, \mathcal{F}_t\big] \;=\; \mathrm{E}\big[\,\|\boldsymbol{M}_i \boldsymbol{e}_t\|^2 \,\big|\, \mathcal{F}_t\big] \;+\; 2\beta\,\mathrm{E}\big[\rho_i\,\boldsymbol{a}_i^{\top} \boldsymbol{M}_i\big]\,\boldsymbol{e}_t \;+\; \beta^2 \sigma^2\,. \tag{A88}$$

For the middle term, $\boldsymbol{a}_i^{\top} \boldsymbol{M}_i = (1 - \beta\,\|\boldsymbol{a}_i\|^2)\,\boldsymbol{a}_i^{\top}$, so by the normal equations

$$\mathrm{E}\big[\rho_i\,\boldsymbol{a}_i^{\top} \boldsymbol{M}_i\big] \;=\; \mathrm{E}\big[\rho\,\boldsymbol{a}\big]^{\top} - \beta\,\mathrm{E}\big[\rho\,\|\boldsymbol{a}\|^2\,\boldsymbol{a}\big]^{\top} \;=\; -\beta\,\boldsymbol{w}^{\top}\,, \tag{A89}$$

with $\boldsymbol{w} = \mathrm{E}[\rho\,\|\boldsymbol{a}\|^2\,\boldsymbol{a}]$ as in the statement. Outside stationarity this cross term does *not* vanish, and it is $O(\beta^2)$ in total. Lemma A1 and Young's inequality $-2\beta^2 \boldsymbol{w}^{\top} \boldsymbol{e} \leq \frac{\bar{\rho}}{2}\,\|\boldsymbol{e}\|^2 + \frac{2\beta^4}{\bar{\rho}}\,\|\boldsymbol{w}\|^2$ give

$$\mathrm{E}\big[\,\|\boldsymbol{e}_{t+1}\|^2 \,\big|\, \mathcal{F}_t\big] \;\leq\; \Big(1 - \tfrac{\bar{\rho}}{2}\Big)\,\|\boldsymbol{e}_t\|^2 \;+\; \beta^2 \sigma^2 \;+\; \tfrac{2\beta^4}{\bar{\rho}}\,\|\boldsymbol{w}\|^2\,. \tag{A90}$$

Taking full expectations and iterating the affine recursion (its geometric series is bounded by $2/\bar{\rho}$) yields (A58). Since $\bar{\rho} = \gamma_\beta \lambda_{\min} = \Theta(\beta)$, the horizon $2\beta^2 \sigma^2/\bar{\rho} + 4\beta^4\,\|\boldsymbol{w}\|^2/\bar{\rho}^2$ is $O(\beta)$, consistent with (A55). □

Corollary A1 (Recovering exact convergence to $\boldsymbol{S}^{\mathrm{LS}}$)**.** *Three standard mechanisms remove the $O(\sqrt{\beta})$ fluctuation:*

(i) Diminishing step sizes *(Robbins–Monro): $\beta_t \downarrow 0$, $\sum_t \beta_t = \infty$ (e.g. $\beta_t \propto 1/t$) gives $\boldsymbol{S}_t \to \boldsymbol{S}^{\mathrm{LS}}$ almost surely.*

(ii) Iterate averaging *(Polyak–Ruppert / ergodic average): with constant β, the time average $\bar{\boldsymbol{S}}_T = \frac{1}{T}\sum_{t\le T} \boldsymbol{S}_t \to \mathrm{E}_{\pi_\beta}[\boldsymbol{S}] = \boldsymbol{S}^{\mathrm{LS}}$ almost surely - and by Theorem A7 (i) the average is exactly unbiased.*

(iii) Small-β limit: *by (A55), π_β concentrates on $\boldsymbol{S}^{\mathrm{LS}}$ as $\beta \to 0$, with the explicit bias–variance trade-off: larger β tracks nonstationary targets faster, smaller β reduces the steady-state loss excess (A57), which is the standard LMS trade-off, inherited verbatim.*

Corollary A2 (Consistent targets)**.** *The inconsistency, not the count $n > d^2$, is what forbids exact convergence. Let us assume that the targets happen to be realizable, that is $c_i = \boldsymbol{v}_i^\top \boldsymbol{S}^\circ \boldsymbol{k}_i$ for some $\boldsymbol{S}^\circ$. In this case $\mathrm{L}_{\min} = 0$, all residuals ρ_i vanish, the driving noise in (A49) disappears, and the Fejér-monotone argument of the consistent case applies unchanged: $\boldsymbol{S}_t \to \boldsymbol{S}^\circ$ (unique by full column rank) almost surely, linearly, for* arbitrarily many *associations. Two distinguished targets realize this structurally:*

(i) the correlation target *$c_i = \boldsymbol{v}_i^\top \boldsymbol{k}_i$, solved by $\boldsymbol{S} = \boldsymbol{I}$ for every n;*

(ii) the norm target *$c_i = \|\boldsymbol{v}_i\|^2$, solved by any exact retrieval memory $\boldsymbol{S}\boldsymbol{k}_i = \boldsymbol{v}_i$ (in particular autoassociatively by $\boldsymbol{S} = \boldsymbol{I}$).*

For these targets the recognition memory has effectively unbounded *association count as long as realizability holds.*

A2.5.5 Remarks

Memory for the overloaded setting. Beyond capacity, $\boldsymbol{S}$ is no longer a table but an *online linear regressor*: the read-out $\boldsymbol{v}^\top \boldsymbol{S}^{\mathrm{LS}} \boldsymbol{k}$ is the best linear prediction (in the p-weighted least-squares sense) of the scalar signal c from the bilinear feature $\boldsymbol{k} \otimes \boldsymbol{v}$. Degradation is graceful since overload produces averaging plus an $O(\sqrt{\beta})$ jitter, never blow-up, since the homogeneous factors keep spectral norm ≤ 1.

Assumptions for the proofs. Independence of i_t from the past makes the mean recursion exact and factorizes the covariance cross terms; the normal equations kill the driving term in the mean and the cross terms in stationarity (via $\mathrm{E}[\boldsymbol{e}] = \boldsymbol{0}$); the step condition (A45) enters three times: (1) $\gamma_\beta > 0$ (contraction), (2) $\beta\lambda_{\max}(\bar{\boldsymbol{A}}) < 2$ (mean stability), and (3) $2 - \beta \max_i \|\boldsymbol{a}_i\|^2 > 0$ (trace upper bound). Full column rank gives $\lambda_{\min}(\bar{\boldsymbol{A}}) > 0$, used for uniqueness of $\boldsymbol{s}^{\mathrm{LS}}$, for $\bar{\rho} > 0$, and to convert $\mathrm{tr}(\bar{\boldsymbol{A}}\boldsymbol{V}_\beta)$ into a bound on $\mathrm{tr}\,\boldsymbol{V}_\beta$. Linearity of the update in the state is what makes every step of the argument exact rather than a Taylor expansion which we obtain from the quadratic loss.

Conflicting targets on the same pair: averaging vs. overwriting. The canonical inconsistency is one pair $(\boldsymbol{k}, \boldsymbol{v})$ presented with different targets. Under *random* sampling with probabilities p_j over targets c_j, the stationary mean read-out is the weighted average, $\boldsymbol{v}^\top \mathrm{E}_{\pi_\beta}[\boldsymbol{S}]\,\boldsymbol{k} = \sum_j p_j c_j / \sum_j p_j$ (the LS solution for identical rows). Under *sequential* presentation the same rule *overwrites*: the contraction drives the read-out to the most recent target. Random replay averages; ordered replay recencies, therefore the sampling protocol, not the rule, decides about overwrite behavior.

Forget gate: recency-weighted least squares. With the scalar gated variant $\boldsymbol{S}_{t+1} = \mathrm{f}_t \boldsymbol{S}_t - \beta(\boldsymbol{v}_t^\top \mathrm{f}_t \boldsymbol{S}_t \boldsymbol{k}_t - c_t)\boldsymbol{v}_t \boldsymbol{k}_t^\top$, $\mathrm{f}_t \in (0, 1)$, the steady state is the corresponding *exponentially weighted* least-squares solution: old associations are geometrically down-weighted, so the effective association count is finite ($\sim 1/(1-\mathrm{f})$) and the memory tracks a drifting target distribution. Therefore, the overloaded regime is exactly where the forget gate is most relevant.

Summary. For $n > d^2$ associations the randomized Kronecker Delta Rule with constant β does not converge to a point but to a stationary distribution: its mean is *exactly* the sampling-weighted least-squares memory $\boldsymbol{S}^{\mathrm{LS}} = \mathrm{argmin} \sum_i p_i (\boldsymbol{v}_i^\top \boldsymbol{S}\boldsymbol{k}_i - c_i)^2$ (the minimizer of L for uniform sampling), its covariance is $O(\beta)$ and solves the exact stationarity equation (A54) driven by the optimal residuals, and its loss sits at $\mathrm{L}_{\min} + \frac{\beta}{2}\mathrm{E}[\rho_i^2 \|\boldsymbol{k}_i\|^2 \|\boldsymbol{v}_i\|^2]$ up to an explicit $1 + O(\beta)$ factor. All of this is proven from first principles: an exact linear mean recursion with spectral radius < 1; a synchronous-coupling $\mathcal{W}_2$-contraction plus Banach fixed point for existence, uniqueness, and geometric convergence of π_β; exact moment identities in stationarity for mean and covariance; an exact

trace identity for the misadjustment; and a Young-inequality recursion for the transient. Averaging or annealing β recovers $\boldsymbol{S}^{\mathrm{LS}}$ exactly; consistent target structures (correlation target, norm target with a retrieval solution) escape the capacity limit altogether; and a forget gate turns the steady state into recency-weighted least squares.

A2.6 Norm Bounds for the KronLSTM with Steady-State Solution

We derive a two-sided norm bound for the steady state of the randomized Kronecker Delta Rule with more than d^2 associations. Running memories or data stream i.e. online updates are covered via the steady state, therefore we obtain an expected bound on the online updated memory.

A2.6.1 Overview

For the overloaded KronLSTM (n associations, features $\boldsymbol{a}_i = \boldsymbol{k}_i \otimes \boldsymbol{v}_i$, sampling weights p_i) the randomized Kronecker Delta Rule converges around the weighted least-squares memory $\boldsymbol{S}^{\mathrm{LS}}$. We derive a *two-sided* bound on its Frobenius norm from an exact energy (Pythagoras) identity:

$$\frac{\mathrm{E}[c^2] - \mathrm{E}[\rho^2]}{\lambda_{\max}(\bar{\boldsymbol{A}})} \;\le\; \left\|\boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}}^2 \;\le\; \frac{\mathrm{E}[c^2] - \mathrm{E}[\rho^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})}\,,$$

where ρ_i are the optimal residuals and $\bar{\boldsymbol{A}} = \mathrm{E}[\boldsymbol{a}\boldsymbol{a}^\top]$ is the feature second moment. The numerator is the *explained target energy*; the bound is attained (both sides coincide) for isotropic designs, and in the orthonormal design it reproduces $\|\boldsymbol{S}^\star\|_{\mathrm{F}}^2 = \sum_i c_i^2$ of the target-weighted Hebbian superposition exactly. Combined with the steady-state analysis, it yields an a-priori bound on the expected memory norm of the running iterates.

A2.6.2 Setting and nomenclature

As before, we assume n associations $(\boldsymbol{k}_i, \boldsymbol{v}_i, c_i)$ with features $\boldsymbol{a}_i = \boldsymbol{k}_i \otimes \boldsymbol{v}_i \in \mathbb{R}^{d^2}$, $\|\boldsymbol{a}_i\| = \|\boldsymbol{k}_i\|\,\|\boldsymbol{v}_i\|$, sampling probabilities $p_i > 0$, and

$$\bar{\boldsymbol{A}} \;=\; \mathrm{E}\big[\boldsymbol{a}\boldsymbol{a}^\top\big] \;=\; \sum_{i=1}^{n} p_i\, \boldsymbol{a}_i \boldsymbol{a}_i^\top\,, \qquad \boldsymbol{s}^{\mathrm{LS}} \;=\; \operatorname*{argmin}_{\boldsymbol{s}}\; \mathrm{E}\big[(\boldsymbol{a}^\top \boldsymbol{s} - c)^2\big] \;=\; \bar{\boldsymbol{A}}^{-1}\,\mathrm{E}\big[c\,\boldsymbol{a}\big]\,, \tag{A91}$$

$$\rho_i \;=\; c_i - \boldsymbol{a}_i^\top \boldsymbol{s}^{\mathrm{LS}}\,, \qquad \mathrm{E}\big[\rho\,\boldsymbol{a}\big] \;=\; \boldsymbol{0} \quad \text{(normal equations)}\,, \tag{A92}$$

assuming $\bar{\boldsymbol{A}} \succ \boldsymbol{0}$ (features span $\mathbb{R}^{d^2}$; the rank-deficient case is treated in the remarks). All expectations are over $i \sim p$. Recall $\|\boldsymbol{s}\|_2 = \|\boldsymbol{S}\|_{\mathrm{F}}$ for $\boldsymbol{s} = \mathrm{vec}(\boldsymbol{S})$, and that for uniform sampling $\mathrm{E}[\rho^2] = \mathrm{L}_{\min}$, the minimum of the loss L.

A2.6.3 An exact energy identity

Lemma A3 (Energy identity)**.** *Let $y_i := \boldsymbol{a}_i^\top \boldsymbol{s}^{\mathrm{LS}} = \boldsymbol{v}_i^\top \boldsymbol{S}^{\mathrm{LS}} \boldsymbol{k}_i$ be the optimal read-outs. Then $\mathrm{E}[\rho\, y] = 0$ and*

$$\mathrm{E}\big[c^2\big] \;=\; \mathrm{E}\big[y^2\big] \;+\; \mathrm{E}\big[\rho^2\big]\,, \qquad \big(\boldsymbol{s}^{\mathrm{LS}}\big)^\top \bar{\boldsymbol{A}}\, \boldsymbol{s}^{\mathrm{LS}} \;=\; \mathrm{E}\big[y^2\big] \;=\; \mathrm{E}\big[c^2\big] - \mathrm{E}\big[\rho^2\big]\,. \tag{A93}$$

Proof. $\mathrm{E}[\rho\, y] = \mathrm{E}[\rho\, \boldsymbol{a}]^\top \boldsymbol{s}^{\mathrm{LS}} = 0$ by the normal equations; with $c = y + \rho$ this gives the first identity. The second follows from $\mathrm{E}[y^2] = (\boldsymbol{s}^{\mathrm{LS}})^\top \mathrm{E}[\boldsymbol{a}\boldsymbol{a}^\top]\, \boldsymbol{s}^{\mathrm{LS}}$. □

The stored *read-out energy* of the least-squares memory equals the target energy minus the unexplained (residual) energy. Therefore, the memory can only be as large as the part of the target signal that the bilinear features can explain.

A2.6.4 The Norm Bound

Theorem A8 (Two-sided Frobenius-norm bound)**.** *With $\lambda_{\min}, \lambda_{\max}$ the extreme eigenvalues of $\bar{\boldsymbol{A}}$,*

$$\boxed{\frac{\mathrm{E}\big[c^2\big] - \mathrm{E}\big[\rho^2\big]}{\lambda_{\max}(\bar{\boldsymbol{A}})} \;\le\; \left\|\boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}}^2 \;\le\; \frac{\mathrm{E}\big[c^2\big] - \mathrm{E}\big[\rho^2\big]}{\lambda_{\min}(\bar{\boldsymbol{A}})}\,.} \tag{A94}$$

In particular, since $\sigma_{\max}(\boldsymbol{S}^{\mathrm{LS}}) \le \left\|\boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}}$,

$$\sigma_{\max}\big(\boldsymbol{S}^{\mathrm{LS}}\big) \;\le\; \sqrt{\frac{\mathrm{E}\big[c^2\big] - \mathrm{E}\big[\rho^2\big]}{\lambda_{\min}(\bar{\boldsymbol{A}})}} \;\le\; \sqrt{\frac{\mathrm{E}\big[c^2\big]}{\lambda_{\min}(\bar{\boldsymbol{A}})}}\,. \tag{A95}$$

Proof. Sandwich the quadratic form: $\lambda_{\min} \left\|\boldsymbol{s}^{\mathrm{LS}}\right\|^2 \le (\boldsymbol{s}^{\mathrm{LS}})^\top \bar{\boldsymbol{A}}\, \boldsymbol{s}^{\mathrm{LS}} \le \lambda_{\max} \left\|\boldsymbol{s}^{\mathrm{LS}}\right\|^2$, and insert (A93). □

Corollary A3 (First-moment variant)**.** *From* $\lambda_{\min} \|\boldsymbol{s}\|^2 \le \boldsymbol{s}^\top \bar{\boldsymbol{A}} \boldsymbol{s} = \boldsymbol{s}^\top \mathrm{E}[c\boldsymbol{a}] \le \|\boldsymbol{s}\|\, \mathrm{E}[|c|\, \|\boldsymbol{a}\|]$,

$$\left\|\boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}} \;\le\; \frac{\mathrm{E}\big[\,|c_i|\, \|\boldsymbol{k}_i\|\, \|\boldsymbol{v}_i\|\,\big]}{\lambda_{\min}(\bar{\boldsymbol{A}})}\,. \tag{A96}$$

The accumulated write mass $\beta\, c_t\, \|\boldsymbol{k}_t\|\, \|\boldsymbol{v}_t\|$ *is exactly the quantity whose first moment appears here. The second moment bound* (A94) *is usually much sharper (it divides by* $\lambda_{\min}$ *only once, under the square root).*

A2.6.5 Tightness and special cases

Isotropic design gives equality. If $\bar{\boldsymbol{A}} = \frac{\mathrm{E}[\|\boldsymbol{a}\|^2]}{d^2}\, \boldsymbol{I}$ (features spread evenly over $\mathbb{R}^{d^2}$), then $\lambda_{\min} = \lambda_{\max}$ and (A94) collapses to the identity

$$\left\|\boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}}^2 \;=\; \frac{d^2\,\big(\mathrm{E}[c^2] - \mathrm{E}[\rho^2]\big)}{\mathrm{E}[\|\boldsymbol{a}\|^2]}\,. \tag{A97}$$

Orthonormal design recovers the Hebbian superposition norm. For $n = d^2$ with orthonormal $\{\boldsymbol{k}_i\}$, orthonormal $\{\boldsymbol{v}_i\}$, and uniform sampling, the features are orthonormal, $\bar{\boldsymbol{A}} = \boldsymbol{I}/d^2$, the system is consistent ($\rho \equiv 0$), and (A94) holds with equality:

$$\|\boldsymbol{S}^\star\|_{\mathrm{F}}^2 \;=\; d^2\, \mathrm{E}[c^2] \;=\; \sum_{i=1}^{d^2} c_i^2\,, \tag{A98}$$

exactly the Frobenius norm of the target-weighted Hebbian superposition $\boldsymbol{S}^\star = \sum_i c_i \boldsymbol{v}_i \boldsymbol{k}_i^\top$. The bound is therefore attained, not merely an estimate.

Unit-norm keys and values give a d^2 conditioning factor. If $\|\boldsymbol{k}_i\| = \|\boldsymbol{v}_i\| = 1$ then $\operatorname{tr} \bar{\boldsymbol{A}} = \mathrm{E}[\|\boldsymbol{a}\|^2] = 1$, so $\lambda_{\min} \le 1/d^2 \le \lambda_{\max}$: the lower bound in (A94) is at most $\mathrm{E}[c^2] - \mathrm{E}[\rho^2]$, the upper bound at least $d^2(\mathrm{E}[c^2] - \mathrm{E}[\rho^2])$, with both collapsing to the latter in the isotropic case. Ill-spread designs pay in $\lambda_{\min}$: the worst case is a target profile aligned with the least-covered feature direction, and the $1/\lambda_{\min}$ dependence is unavoidable in general.

Consistent case where association errors are zero. If the system is consistent ($\rho \equiv 0$; e.g. $n \le d^2$ with independent features, or the correlation/norm targets), then $\boldsymbol{S}^{\mathrm{LS}} = \boldsymbol{S}^\star$ is the exact solution and

$$\|\boldsymbol{S}^\star\|_{\mathrm{F}}^2 \;\le\; \frac{\mathrm{E}[c^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})}\,. \tag{A99}$$

Residual energy can only *shrink* the memory: overload (inconsistency) makes the least-squares memory smaller than a consistent memory with the same target energy, never larger.

Rank-deficient features. If the features do not span $\mathbb{R}^{d^2}$ ($n < d^2$ or dependent $\boldsymbol{a}_i$), replace $\boldsymbol{S}^{\mathrm{LS}}$ by the minimum-Frobenius-norm least-squares solution and $\lambda_{\min}$ by the smallest *nonzero* eigenvalue of $\bar{\boldsymbol{A}}$: since $\boldsymbol{s}^{\mathrm{LS}} \in \operatorname{span}\{\boldsymbol{a}_i\}$ (as do the iterates from $\boldsymbol{S}_0 = \boldsymbol{0}$), the sandwich argument applies on that subspace and (A94) holds verbatim with $\lambda_{\min}^+$.

A2.6.6 Bound on the memory norm for steady-state and all updates

Bound on the memory norm for steady-state distribution. First, we consider a bound on the memory norm under the steady-state distribution.

Corollary A4 (Bound on the steady-state memory norm)**.** *Under the randomized rule with step size* $0 < \beta < 2/\max_i \|\boldsymbol{a}_i\|^2$*, the stationary distribution* π_β *has mean exactly* $\boldsymbol{S}^{\mathrm{LS}}$*, so the second moment decomposes orthogonally,* $\mathrm{E}_{\pi_\beta} \|\boldsymbol{S}\|_{\mathrm{F}}^2 = \left\|\boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}}^2 + \operatorname{tr} \boldsymbol{V}_\beta$*, and with the covariance trace bound of the steady-state note,*

$$\mathrm{E}_{\pi_\beta}\Big[\|\boldsymbol{S}\|_{\mathrm{F}}^2\Big] \;\le\; \frac{\mathrm{E}[c^2] - \mathrm{E}[\rho^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})} \;+\; \frac{\beta\, \mathrm{E}\big[\rho^2\, \|\boldsymbol{k}\|^2\, \|\boldsymbol{v}\|^2\,\big]}{\big(2 - \beta \max_i \|\boldsymbol{a}_i\|^2\big)\, \lambda_{\min}(\bar{\boldsymbol{A}})}\,. \tag{A100}$$

The first term is the size of the memory content, the second the $O(\beta)$ *fluctuation; for* $\beta \to 0$ *or in the consistent case the bound reduces to Theorem A8.*

Theorem A9 (Two-sided memory norm bound for the steady state)**.** *Let π_β be the steady-state law of (A16), $\boldsymbol{V}_\beta = \mathrm{E}_{\pi_\beta}[\boldsymbol{e}\boldsymbol{e}^\top]$ its covariance as in Theorem A7 (ii) (the two coincide because the stationary mean is exactly $\boldsymbol{s}^{\mathrm{LS}}$), $\sigma^2 = \mathrm{E}[\rho^2 \|\boldsymbol{a}\|^2]$, and write*

$$A_{\max}^2 := \max_i \|\boldsymbol{a}_i\|^2 = \max_i \|\boldsymbol{k}_i\|^2 \|\boldsymbol{v}_i\|^2, \qquad \kappa := \beta A_{\max}^2 \in (0,2), \tag{A101}$$

the last inclusion being the step-size condition (A45). Then the energy identity

$$\boldsymbol{s}^{\mathrm{LS}\top} \bar{\boldsymbol{A}}\, \boldsymbol{s}^{\mathrm{LS}} = \mathrm{E}[c^2] - \mathrm{E}[\rho^2] \tag{A102}$$

holds, hence the least-squares memory is pinned on both sides,

$$\frac{\mathrm{E}[c^2] - \mathrm{E}[\rho^2]}{\lambda_{\max}(\bar{\boldsymbol{A}})} \le \left\|\boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}}^2 \le \frac{\mathrm{E}[c^2] - \mathrm{E}[\rho^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})}, \tag{A103}$$

and the steady-state memory itself satisfies

$$\mathrm{E}_{\pi_\beta}\left[\|\boldsymbol{S}\|_{\mathrm{F}}^2\right] = \left\|\boldsymbol{S}^{\mathrm{LS}}\right\|_{\mathrm{F}}^2 + \operatorname{tr} \boldsymbol{V}_\beta, \qquad \operatorname{tr} \boldsymbol{V}_\beta \le \frac{\beta\, \mathrm{E}\left[\rho^2 \|\boldsymbol{a}\|^2\right]}{(2-\kappa)\, \lambda_{\min}(\bar{\boldsymbol{A}})}. \tag{A104}$$

The fluctuation term admits the exact decomposition

$$\operatorname{tr}\left(\bar{\boldsymbol{A}} \boldsymbol{V}_\beta\right) = \frac{\beta}{2}\sigma^2 + R_\beta, \tag{A105}$$

$$R_\beta := \frac{\beta}{2}\, \mathrm{E}\left[\left(\boldsymbol{a}^\top \boldsymbol{V}_\beta \boldsymbol{a}\right) \|\boldsymbol{a}\|^2\right], \qquad 0 \le R_\beta \le \frac{\kappa}{2-\kappa} \cdot \frac{\beta\, \sigma^2}{2}.$$

R_β is bounded by $\beta^2 A_{\max}^2 \sigma^2 / \big(2(2-\kappa)\big)$. Consequently, since $\mathrm{E}[\rho^2] \le \mathrm{E}[c^2]$,

$$\mathrm{E}_{\pi_\beta}\left[\|\boldsymbol{S}\|_{\mathrm{F}}^2\right] \le \frac{1}{\lambda_{\min}(\bar{\boldsymbol{A}})} \left(\mathrm{E}[c^2] - \frac{2\,(1-\kappa)}{2-\kappa}\, \mathrm{E}[\rho^2]\right) \le \frac{\mathrm{E}[c^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})} \max\left\{1, \frac{\kappa}{2-\kappa}\right\}, \tag{A106}$$

and in particular, whenever the step size satisfies $\beta \|\boldsymbol{a}_t\|^2 \le 1$ (i.e. $\kappa \le 1$, the non-expansive regime),

$$\boxed{\mathrm{E}_{\pi_\beta}\left[\|\boldsymbol{S}\|_{\mathrm{F}}^2\right] \le \frac{\mathrm{E}[c^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})}.} \tag{A107}$$

All statements are exact and non-asymptotic: they hold for every β admitted by (A45).

Proof. The normal equations $\bar{\boldsymbol{A}}\boldsymbol{s}^{\mathrm{LS}} = \boldsymbol{b}$ give $\mathrm{E}[\rho^2] = \mathrm{E}[c^2] - 2\boldsymbol{b}^\top \boldsymbol{s}^{\mathrm{LS}} + \boldsymbol{s}^{\mathrm{LS}\top} \bar{\boldsymbol{A}} \boldsymbol{s}^{\mathrm{LS}} = \mathrm{E}[c^2] - \boldsymbol{s}^{\mathrm{LS}\top} \bar{\boldsymbol{A}} \boldsymbol{s}^{\mathrm{LS}}$, which is (A102); bounding the Rayleigh quotient $\boldsymbol{s}^{\mathrm{LS}\top} \bar{\boldsymbol{A}} \boldsymbol{s}^{\mathrm{LS}}$ between $\lambda_{\min}(\bar{\boldsymbol{A}}) \left\|\boldsymbol{s}^{\mathrm{LS}}\right\|^2$ and $\lambda_{\max}(\bar{\boldsymbol{A}}) \left\|\boldsymbol{s}^{\mathrm{LS}}\right\|^2$, and using $\|\boldsymbol{s}\|_2 = \|\boldsymbol{S}\|_{\mathrm{F}}$, gives (A103). The first part of (A104) is the bias–variance split $\mathrm{E}\|\boldsymbol{s}\|^2 = \|\mathrm{E}\boldsymbol{s}\|^2 + \mathrm{TrCov}(\boldsymbol{s})$ together with the fact that the stationary mean is *exactly* $\boldsymbol{s}^{\mathrm{LS}}$ (Theorem A7 (i)).

Exact stationarity equation. Rather than the leading-order Lyapunov equation, we use the *exact* stationarity equation (A54) of Theorem A7 (ii),

$$\bar{\boldsymbol{A}}\boldsymbol{V}_\beta + \boldsymbol{V}_\beta \bar{\boldsymbol{A}} - \beta\, \mathrm{E}\left[\left(\boldsymbol{a}^\top \boldsymbol{V}_\beta\, \boldsymbol{a}\right) \boldsymbol{a}\boldsymbol{a}^\top\right] = \beta\, \mathrm{E}\left[\rho^2\, \boldsymbol{a}\boldsymbol{a}^\top\right], \tag{A108}$$

whose trace is $2\operatorname{tr}(\bar{\boldsymbol{A}}\boldsymbol{V}_\beta) = \beta\sigma^2 + \beta \mathrm{E}[(\boldsymbol{a}^\top \boldsymbol{V}_\beta \boldsymbol{a}) \|\boldsymbol{a}\|^2]$. This is (A105), and it identifies the remainder: $R_\beta \ge 0$ because $\boldsymbol{V}_\beta \succeq \boldsymbol{0}$, while $\mathrm{E}[(\boldsymbol{a}^\top \boldsymbol{V}_\beta \boldsymbol{a}) \|\boldsymbol{a}\|^2] \le A_{\max}^2\, \mathrm{E}[\boldsymbol{a}^\top \boldsymbol{V}_\beta \boldsymbol{a}] = A_{\max}^2 \operatorname{tr}(\bar{\boldsymbol{A}}\boldsymbol{V}_\beta)$ gives $(2-\kappa) \operatorname{tr}(\bar{\boldsymbol{A}}\boldsymbol{V}_\beta) \le \beta\sigma^2$, the prefactor being positive by (A45). Substituting this bound back into $R_\beta = \frac{\beta A_{\max}^2}{2} \operatorname{tr}(\bar{\boldsymbol{A}}\boldsymbol{V}_\beta)$'s upper estimate yields the stated bound on R_β, and $\lambda_{\min}(\bar{\boldsymbol{A}}) \operatorname{tr} \boldsymbol{V}_\beta \le \operatorname{tr}(\bar{\boldsymbol{A}}\boldsymbol{V}_\beta)$ yields the second part of (A104).

Deriving the clean bound. $\boldsymbol{S} = \boldsymbol{0}$ is feasible, so $\mathrm{E}[\rho^2] \le \mathrm{E}[c^2]$; moreover $\sigma^2 = \mathrm{E}[\rho^2 \|\boldsymbol{a}\|^2] \le A_{\max}^2 \mathrm{E}[\rho^2]$, hence $\beta\sigma^2 \le \kappa\, \mathrm{E}[\rho^2]$ and

$$\operatorname{tr} \boldsymbol{V}_\beta \le \frac{\kappa}{2-\kappa} \cdot \frac{\mathrm{E}[\rho^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})}. \tag{A109}$$

Adding this to (A103) gives

$$\mathrm{E}_{\pi_\beta}\big[\,\|\boldsymbol{S}\|_{\mathrm{F}}^2\,\big] \;\le\; \frac{1}{\lambda_{\min}(\bar{\boldsymbol{A}})}\left(\mathrm{E}[c^2]-\mathrm{E}[\rho^2]+\frac{\kappa}{2-\kappa}\mathrm{E}[\rho^2]\right), \tag{A110}$$

and $-1+\kappa/(2-\kappa) = -2(1-\kappa)/(2-\kappa)$ is the coefficient in (A106). For $\kappa \le 1$ that coefficient is ≤ 0, so dropping the $\mathrm{E}[\rho^2]$ term gives (A107); for $\kappa > 1$, using $\mathrm{E}[\rho^2] \le \mathrm{E}[c^2]$ gives the factor $1+2(\kappa-1)/(2-\kappa) = \kappa/(2-\kappa)$. □

Residuals lower the bound. The $\mathrm{E}[\rho^2]$ term in (A106) enters with a *negative* sign throughout the non-expansive regime $\kappa \le 1$: overload shrinks the bound but does not inflate it, because residual energy is energy the bilinear features cannot explain and therefore cannot store. The two effects of ρ are less memory content (A103) and more steady-state jitter (A104). They cancel exactly at $\kappa = 1$ and the content term dominates below it. This is why the constant in (A107) is 1 rather than the $3/2$ or 2 obtained by bounding the two contributions separately.

Bound on the memory norm for all updates. Next, we consider the bound even before the memory $\boldsymbol{S}_t$ reaches the steady state but is in the update phase.

Theorem A10 (The bound applies to the *running* or *online* memory at every step: not only in the steady state)**.** *Let $\mu_t := law(\boldsymbol{s}_t) = \mu_0 P^t \in \mathcal{P}_2$ be the law of the iterate at step t under (A16), so that π_β is the unique fixed point of the same kernel, and set*

$$q_\beta \;:=\; 1 \;-\; \beta\,(2-\kappa)\,\lambda_{\min}(\bar{\boldsymbol{A}}) \;\in\; [0,1)\,. \tag{A111}$$

Started at $\boldsymbol{S}_0 = \boldsymbol{0}$, the running memory obeys, for every t and in the non-expansive regime $\kappa \le 1$,

$$\boxed{\mathrm{E}\big[\,\|\boldsymbol{S}_t\|_{\mathrm{F}}^2\,\big] \;\le\; \left(1+q_\beta^{t/2}\right)^2 \frac{\mathrm{E}[c^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})} \;\le\; \frac{4\,\mathrm{E}[c^2]}{\lambda_{\min}(\bar{\boldsymbol{A}})}\,,} \tag{A112}$$

where the transient factor expands as $\left(1+q_\beta^{t/2}\right)^2 = 1+2q_\beta^{t/2}+q_\beta^t \downarrow 1$, so the bound decays geometrically from 4 to the steady-state constant 1 of (A107). For a general step size $\kappa \in (0,2)$ the same statement holds with $\mathrm{E}[c^2]$ replaced by $\mathrm{E}[c^2]\max\{1,\kappa/(2-\kappa)\}$; for a general deterministic start $\boldsymbol{S}_0$ it reads

$$\mathrm{E}\big[\,\|\boldsymbol{S}_t\|_{\mathrm{F}}^2\,\big] \;\le\; \left(B_\beta \;+\; q_\beta^{t/2}\sqrt{\|\boldsymbol{S}_0-\boldsymbol{S}^{\mathrm{LS}}\|_{\mathrm{F}}^2+\operatorname{tr}\boldsymbol{V}_\beta}\right)^2, \qquad B_\beta := \mathrm{E}_{\pi_\beta}\big[\,\|\boldsymbol{S}\|_{\mathrm{F}}^2\,\big]^{1/2}. \tag{A113}$$

Nothing on the right-hand side depends on the sequence length: the expected squared memory norm never exceeds four times its steady-state bound from the very first step, and relaxes geometrically to that bound.

Proof. Rate. Couple two copies of (A16) synchronously, i.e. drive them with the same $(\boldsymbol{a}_t, c_t)$. The additive term cancels in the difference $\boldsymbol{d}_t$, leaving $\boldsymbol{d}_{t+1} = (\boldsymbol{I}-\beta\boldsymbol{a}_t\boldsymbol{a}_t^\top)\boldsymbol{d}_t$, and

$$\mathrm{E}\big[(\boldsymbol{I}-\beta\boldsymbol{a}\boldsymbol{a}^\top)^2\big] \;=\; \boldsymbol{I}-2\beta\bar{\boldsymbol{A}}+\beta^2\mathrm{E}\big[\,\|\boldsymbol{a}\|^2\,\boldsymbol{a}\boldsymbol{a}^\top\big] \;\preceq\; \boldsymbol{I}-\beta\big(2-\kappa\big)\bar{\boldsymbol{A}}\,, \tag{A114}$$

using $\beta^2\mathrm{E}[\|\boldsymbol{a}\|^2\,\boldsymbol{a}\boldsymbol{a}^\top] \preceq \beta\kappa\bar{\boldsymbol{A}}$. Hence $\mathrm{E}\,\|\boldsymbol{d}_{t+1}\|^2 \le q_\beta\,\mathrm{E}\,\|\boldsymbol{d}_t\|^2$, and taking the infimum over couplings gives $\mathcal{W}_2(\mu_t,\pi_\beta) \le q_\beta^{t/2}\,\mathcal{W}_2(\mu_0,\pi_\beta)$. (Equivalently $q_\beta = 1-\bar\rho$ with the $\bar\rho = \gamma_\beta\lambda_{\min}(\bar{\boldsymbol{A}})$ of Lemma A1, since $\gamma_\beta \ge \beta(2-\kappa)$; q_β is written out here only to keep the statement self-contained.) Since $\boldsymbol{s} \mapsto \|\boldsymbol{s}\|$ is 1-Lipschitz, the reverse triangle inequality in L^2 gives $\mathrm{E}[\|\boldsymbol{S}_t\|_{\mathrm{F}}^2]^{1/2} \le B_\beta + \mathcal{W}_2(\mu_t,\pi_\beta)$.

The initial term is not an estimate. For a deterministic start the only coupling of $\mu_0 = \delta_{\boldsymbol{S}_0}$ with π_β is the product, so the infimum defining $\mathcal{W}_2$ is attained and

$$\mathcal{W}_2(\mu_0,\pi_\beta)^2 \;=\; \mathrm{E}_{\pi_\beta}\big[\,\|\boldsymbol{S}-\boldsymbol{S}_0\|_{\mathrm{F}}^2\,\big] \;=\; \big\|\boldsymbol{S}_0-\boldsymbol{S}^{\mathrm{LS}}\big\|_{\mathrm{F}}^2 \;+\; \operatorname{tr}\boldsymbol{V}_\beta\,, \tag{A115}$$

the second equality being the bias–variance split around the stationary mean $\boldsymbol{S}^{\mathrm{LS}}$ (Theorem A7 (i)). At $\boldsymbol{S}_0 = \boldsymbol{0}$ the right-hand side is exactly $\big\|\boldsymbol{S}^{\mathrm{LS}}\big\|_{\mathrm{F}}^2 + \operatorname{tr}\boldsymbol{V}_\beta = B_\beta^2$ by (A104), i.e. $\mathcal{W}_2(\mu_0,\pi_\beta) = B_\beta$: the distance from the empty memory to the steady state *is* the steady-state size. The two terms therefore collapse into one,

$$\mathrm{E}\big[\,\|\boldsymbol{S}_t\|_{\mathrm{F}}^2\,\big]^{1/2} \;\le\; \left(1+q_\beta^{t/2}\right)B_\beta\,, \tag{A116}$$

and squaring and inserting $B_\beta^2 \le \mathrm{E}[c^2]/\lambda_{\min}(\bar{\boldsymbol{A}})$ from (A107) gives (A112). Squaring preserves the inequality since both sides are nonnegative. Finally $q_\beta^{t/2} \le 1$ bounds the transient factor by 4. The general-κ variant replaces B_β^2 by the bound of (A106), and the general-$\boldsymbol{S}_0$ variant (A113) keeps $\mathcal{W}_2(\mu_0, \pi_\beta)$ in the form (A115) instead of collapsing it. □

Summary. The least-squares memory satisfies the exact energy identity $(\boldsymbol{s}^{\mathrm{LS}})^\top \bar{\boldsymbol{A}}\, \boldsymbol{s}^{\mathrm{LS}} = \mathrm{E}[c^2] - \mathrm{E}[\rho^2]$, which sandwiches its Frobenius norm between $(\mathrm{E}[c^2] - \mathrm{E}[\rho^2])/\lambda_{\max}$ and $(\mathrm{E}[c^2] - \mathrm{E}[\rho^2])/\lambda_{\min}$. This is the explained target energy divided by the extreme feature-covariance eigenvalues. The bound is an equality for isotropic designs, reproduces $\sum_i c_i^2$ in the orthonormal design, extends to the running iterates with an additive $O(\beta)$ fluctuation term, shrinks under overload since the residual energy is subtracted.

A3 Further Results

A3.1 400M language models

Models with 400M/40B (Table A1) are closer to each other in their performance. KronLSTM has the best average and performs best on HellaSwag, ARC-easy, ARC-challenge, and WinoGrande. Mamba-3 is best on PIQA, KDA on LAMBADA, and GDN-2 on WinoGrande where it ties with KronLSTM. Compared with mLSTM, the matrix-memory xLSTM variant that KronLSTM builds on, KronLSTM is better on five of the six tasks, raises the average by 2.1 points, and lowers LAMBADA perplexity from 43.40 to 28.28. mLSTM is only marginally ahead on PIQA.

Table A1: Downstream performance of 400M/40B linear models. Avg-6 is the mean accuracy over the five commonsense reasoning tasks and LAMBADA. Best per column in **bold**, second best underlined.

Model	Hella	PIQA	ARC-e	ARC-c	Wino	LMB acc	Avg-6 ↑	LMB ppl ↓
GDN	41.28	66.05	57.87	31.48	49.64	32.31	46.44	34.36
Mamba-3	41.89	**67.25**	60.82	31.23	52.33	35.11	48.10	29.51
KDA	42.08	66.16	60.48	31.40	51.30	**36.70**	48.02	**27.84**
GDN-2	42.33	66.87	60.23	29.69	**52.41**	33.84	47.56	29.14
mLSTM	40.29	67.19	60.77	29.78	50.75	29.87	46.29	43.40
KronLSTM	**43.09**	67.03	**61.03**	**32.00**	**52.41**	35.05	**48.43**	28.28

A3.2 Ablation on different formulations of KronLSTM

Table A2 ablates the components of the KronLSTM write rule at 1.5B/100B. We report Avg-6, the mean accuracy over the five commonsense reasoning tasks and LAMBADA, together with LAMBADA perplexity, the average over the six-task real-world recall suite at 2k context, and the average over the six RULER NIAH tasks of Table 3 at 4k, 8k and 16k. Three variants are compared with the base model of (10):
(1) the ℓ_2-normalization of $\boldsymbol{v}_t$ is removed;
(2) erase and write are coupled by using the target c_t instead of $\tilde{c}_t = c_t/\beta_t$, so that β_t scales both terms as in Gated DeltaNet;
(3) the write gate is fixed to the unit target $c_t \equiv 1$.
On language modeling, variants (1) and (2) stay within 0.3 accuracy points and 0.25 perplexity points of the base model, with the coupled variant marginally ahead on both. Fixing the target costs 1.6 accuracy points and 0.8 perplexity points, the largest degradation among the ablations. On the recall suite, variants (1) and (2) gain 1.0 and 0.5 points, respectively. On NIAH the ordering reverses: at 4k the base model leads the coupled variant by 12.6 points and the unnormalized variant by 19.9 points, at 8k the unnormalized variant is ahead by 2.2 points, and at 16k the base model leads all three variants. The learned write gate and the decoupled write therefore matter most for retrieval within the pretraining context, while their effect on language modeling is small.

A4 Pretraining Details

All models are pretrained from scratch using `lm-engine` Mishra (2024) with Triton kernels from `flash-linear-attention` Yang & Zhang (2026) whenever available. Every mixer in a com-

Table A2: Write-rule ablation at 1.5B/100B, relative to KRONLSTM. Avg-6 is the mean of the five commonsense reasoning tasks plus LAMBADA accuracy. Recall is the six-task real-world recall suite at 2k context. RULER is averaged at each evaluation length. Best per column in **bold**, second best underlined.

Ablations	Avg-6 ↑	LMB ppl ↓	Recall ↑	NIAH-Avg@4k	@8k	@16k
base	<u>59.72</u>	<u>11.716</u>	30.3	**59.5**	<u>30.5</u>	**14.3**
no ℓ_2-norm on $\boldsymbol{v}_t$	59.41	11.844	**31.3**	39.6	**32.7**	<u>12.6</u>
β on both sides	**60.00**	**11.479**	<u>30.8</u>	<u>46.9</u>	29.4	11.8
non-learnable target	58.17	12.495	30.1	41.9	26.2	10.1

Table A3: RULER NIAH for the 32k-context-extended linear models, evaluated on identical documents. Better of the two per task and length in **bold**.

Model	Len.	S1	S2	S3	MK	MQ	MV
GDN-2	4k	**100.0**	74.0	68.2	28.4	**56.5**	**62.5**
	8k	**100.0**	55.6	17.6	23.2	**38.8**	**41.9**
	16k	**100.0**	10.2	7.4	9.4	7.4	8.9
	32k	**97.8**	2.6	0.8	3.0	1.5	7.2
KronLSTM	4k	**100.0**	**99.6**	**92.6**	**30.8**	37.1	43.9
	8k	99.4	**94.0**	**61.2**	**28.4**	33.5	36.8
	16k	94.6	**27.6**	**41.8**	**18.0**	**21.5**	**30.1**
	32k	81.4	**5.6**	**8.6**	**5.0**	**3.0**	**11.3**

parison shares the corpus, the tokenizer, the optimizer, the schedule, the batch size, and the seed; we only change the sequence mixer blocks. The models were trained on 16 nodes of 4×NVIDIA GH200. Training is pure data parallelism under the FSDP-2 with bf16 mixed-precision.

Data and tokenization We train on the FineWeb-Edu 100BT sample, tokenized with the Llama 2 tokenizer with its standard vocab size of 32,000. Documents are concatenated, separated by the EOS token, and packed into fixed-length sequences of 8192 tokens; This is applied identically to every experiment.

Architecture All models are decoder-only and identical outside the sequence mixer. We implemented the block structures of every baseline as in the original works. Mamba-3 was trained using the standard expansion factor of 2, a dstate of 128, and a head dimension of 64. For Mamba-3, we utilized the SISO formulation. For the other baselines and our method, we utilized $d_{qk} = d_v = 128$. This makes the state size equal across all models. 1.5B models have 24 layers, a hidden dimension of 2048 with 16 heads. The 400M models have 24 layers with a hidden dimension of 1024 and 4 heads (here $d_v = 256$).

Optimization

optimizer	AdamW, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-10}$
peak learning rate	3×10^{-4}
schedule	linear warmup over 10% of steps, then cosine to $0.1\times$ peak (3×10^{-5})
weight decay	0.1, excluding norms, biases, $A_{\log}$, b_Δ and the gate bias
gradient clipping	global norm 1.0
sequence length	8192
global batch	128 sequences = 1,048,576 tokens per step
precision	bf16 mixed precision, fp32 optimiser state

A4.1 LONG-CONTEXT EXTENSION

The context-extended models of Section 3.3 branch from the 1.5B pretraining runs at 75B tokens and are trained for a further 25B tokens at a sequence length of 32,768, so that both extended models finish at exactly 100B tokens and are token-matched to the 8k-pretrained models. The architecture,

tokenizer, and all model hyperparameters are unchanged. The global batch is held at the same 1M tokens per step, now as 32 sequences of 32,768 tokens instead of 128 sequences of 8,192.

Optimization. The optimizer state is re-initialized at the branch point for both models and the schedule restarts: linear warmup over 5% to a peak learning rate of 8.4×10^{-5}, which is the value the parent cosine schedule had reached at the checkpoint, followed by cosine decay over 23,800 steps to $0.1\times$ the peak. This mirrors the warmup-and-decay shape used by Gao et al. (2024). AdamW hyperparameters, weight decay, and gradient clipping are as in Appendix A4.

Data. The corpus takes the four long-document domains of ProLong-64K (Gao et al., 2024), books, code, textbooks, and arXiv, re-tokenized with the Llama 2 tokenizer, and blends them at the ratios of Gao et al. (2024) with FineWeb-Edu in place of their short-context mixture. Documents are concatenated and packed into sequences of 32,768 tokens as in pretraining.

Systems. Training runs on 8 nodes of $4\times$NVIDIA GH200 with ZeRO stage 2 and block-wise gradient checkpointing under bf16 mixed precision.

A4.2 State Tracking Experiments

State-tracking tasks. Each sample is a token sequence of length $\sim \mathcal{U}[2, 128]$ padded to the context length, with a single supervised position at its end. *Parity*: i.i.d. uniform bits, label the parity of the number of ones (chance 50%). S_3: i.i.d. uniform transpositions $(0\,1), (0\,2), (1\,2)$, label the prefix product $x_1 \circ \cdots \circ x_\ell$ as one of the six group elements (chance 16.7%; tracking parity alone gives 33.3%). *Modular arithmetic*: an expression $d_1\, o_1\, d_2 \ldots d_k\, =$ with uniform digits in $\{0, \ldots, 4\}$ and operators in $\{+, -, \times\}$, evaluated with operator precedence modulo 5, the result supervised at the $=$ token (chance 20%). Samples are generated on the fly and deterministically from the run seed (seeds 1-3); evaluation uses $50{,}000$ fresh sequences per length $L \in \{128, 256, 512, 1024\}$.

Models and training. All models share a 128-dimensional embedding, pre-norm residual blocks (RMSNorm, no feed-forward sublayer) and a linear read-out; they differ only in the mixer: Gated DeltaNet with the eigenvalue range extended to $[-1, 1]$ (Grazzi et al., 2025), Mamba-3 (SISO, expansion 2, state size 128, data-dependent RoPE), and KronLSTM. For optimizer we used AdamW with weight decay 0.1, batch size 128, bf16. Parity and S_3: two blocks, constant learning rate $\in \{10^{-3}, 10^{-4}\}$, 50k steps, gradient clipping 100. Modular arithmetic requires depth and a longer budget for all three architectures: three blocks, 150k steps. We report end-of-training accuracy, mean $\pm$ standard deviation over three seeds, at the learning rate with the best mean accuracy over $L \in \{256\}$.