IMPLICIT BIAS OF OPTIMIZERS GUIDES LENGTH GENERALIZATION IN FORGETTING ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Whether a Transformer trained on bounded-length inputs stays accurate on much longer ones depends on its design and training choices. We study this *length generalization* from the optimization perspective, through a task in which a sequence of key–value pairs may reassign a key and a query asks for its key's latest value. The model must reject stale assignments while retaining distant values behind newer records of other keys. A recency bias helps with the first demand but can hinder the second. We analyze this tension in a two-layer attention model with fixed attention masks and value projections, which learns key–value binding, query–key matching, a forget gate, and a linear decoder from binary answer labels. Our main result separates gradient descent (GD) from Adam on this model under suitable per-parameter-group learning rates. With high probability over a shared random orthogonal initialization of the query and key embeddings, a common $1/(t+1)$ schedule yields the following asymptotic separation. GD's worst-case error tends to zero through two records beyond the largest training target lag; at each greater lag, a fixed sequence has correct-answer probability tending to zero. Adam, with its stabilizer ϵ set to zero on the scalar parameters and positive on the embedding matrices, instead guarantees a retrieval distance proportional to its cumulative base learning rate, which diverges. Both success guarantees are uniform over arbitrary earlier prefixes. Adam's guarantee extends to independent Gaussian initialization. SGD inherits GD's cutoff with increasing batch sizes and exact averages over all key pairs. Controlled experiments compare finite retrieval ranges and isolate the role of the scalar stabilizer ϵ.

1 INTRODUCTION

In sequence learning, transformers use attention to retrieve information from earlier tokens in the input (Vaswani et al., 2017). Whether computations learned from short examples remain reliable on longer inputs is the problem of *length generalization* (Anil et al., 2022; Kazemnejad et al., 2023). Models trained on finite data often need to process inputs longer than those seen during training. Despite its importance, there are many documented failures to length generalize both theoretically and empirically (Anil et al., 2022; Weiss et al., 2021; Zhou et al., 2024; Hahn, 2020; Veličković et al., 2025). Nevertheless, architectural designs such as recency biases, positional representations, and scratchpad prompting support substantial finite-length gains (Press et al., 2022; Ruoss et al., 2023; Kazemnejad et al., 2023; Anil et al., 2022), and learned forgetting improves long-context language modeling (Lin et al., 2025).

Existing theory gives conditions under which training on bounded-length sequences is sufficient for generalization to longer ones (Huang et al., 2025; Izzo et al., 2026). However, these results leave open how optimizer choice affects the retrieval range learned during training. In classification, studies of implicit bias show that different optimizers can favor different decision boundaries and learn different features (Zhang et al., 2024; Vasudeva et al., 2025). We ask whether optimizer choice similarly affects length generalization in attention in a specific task: latest-value retrieval.

In latest-value retrieval, an input sequence assigns values to keys, and a query asks for its key's latest value. Retrieving the target requires distinguishing it from both stale assignments of the same key and newer records of other keys (Figure 1). A preference for recent records helps reject stale assignments.

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.

This preference appears in ALiBi[1], which subtracts a linear distance penalty, and in the Forgetting Transformer (FoX), whose learned forget gates multiplicatively decay unnormalized attention weights (Press et al., 2022; Lin et al., 2025). We study data-independent gates, which give a linear penalty with a learned slope. The two have equivalent dynamics only under the same slope parameterization.

Architecture. We study how this retrieval rule is learned using a two-layer attention model (Figure 3). The first layer learns to associate each value with its key. The second matches these keys to the query and combines their values to predict the answer. This design draws on prior studies of how attention learns relationships between tokens (Elhage et al., 2021; Bietti et al., 2023; Nichani et al., 2024). The model learns from correct-answer labels, with attention masks and value projections held fixed. How far it can retrieve depends on the balance it learns between recognizing the correct key and favoring recent records.

Our contribution. For the bounded-length populations in Figure 2, the *lag* of a target is one plus the number of records after it, and $R \geq 2$ is the largest training recall lag.

- **An optimizer-dependent retrieval boundary.** We prove a high-probability separation between full-batch gradient descent (GD) and Adam from a shared random orthogonal initialization (Theorem 2). Both use the same data, objective, and suitably chosen learning-rate multipliers and $1/(t+1)$ base schedule, fixed before training. GD's worst-case retrieval error tends to zero through lag $R+2$. At every fixed lag $r \geq R+3$, however, there is a fixed sequence whose correct-answer probability tends to zero. Adam's worst-case error tends to zero over a growing range of lags proportional to its cumulative base learning rate, which grows as $\log t$. The success bounds are uniform over sequences in the guaranteed ranges, including arbitrarily long prefixes before the target.
- **Learning key matching from answer labels.** We jointly analyze how the model learns query/key representations, soft key–value binding, the forgetting gate, and the decoder from answer labels, with masks and value projections held fixed. Their asymptotic behavior along training determines the retrieval ranges.
- **Initialization and sampling extensions.** Adam's growing-range guarantee also holds with high probability from independent Gaussian query/key initialization under a separately chosen fixed rate schedule (Theorem 4). GD's conclusion extends with high probability to SGD with exact averaging over all key pairs and growing batches (Corollary 3). The Adam guarantees use zero scalar stabilizers. With a fixed positive decoder stabilizer, our guarantee no longer applies; whether Adam's retrieval range can still grow without bound remains open.

We complement the theory with experiments in the structured model and in fully learned Transformers. In the structured model, changing Adam's scalar stabilizer substantially changes retrieval despite nearly identical training losses. The fully learned Transformer experiments show no consistent Adam–SGD separation under the tested protocols when target predecessors are unrestricted. These findings motivate evaluating optimizer choices and hyperparameters on longer inputs, since short-sequence training loss alone need not reveal retrieval quality.

1.1 Related work

Forgetting and learned retrieval. ALiBi uses fixed linear distance penalties (Press et al., 2022), and FoX learns forget gates whose data-independent special case gives the penalty studied here (Lin et al., 2025). When ALiBi's slope is learned with the same parameterization, the two implementations have identical learning dynamics (Section 2.2). Induction-head analyses study the acquisition of associative recall (Bietti et al., 2023), while Nichani et al. (2024) analyze how a two-layer Transformer learns causal structure. Vural et al. (2026) give finite-sample recall guarantees with nonorthogonal embeddings. We track learned binding, matching, forgetting, and decoding together when repeated keys have conflicting values.

Length generalization. Huang et al. (2025) identify conditions under which an idealized regularized learner determines long-sequence behavior from sufficiently long training examples. Izzo et al. (2026) give quantitative bounds on the required training length through simulation. GD can learn

[1]Note that vanilla ALiBi's fixed slopes fall outside our analysis. Here we consider its learnable slope variant.

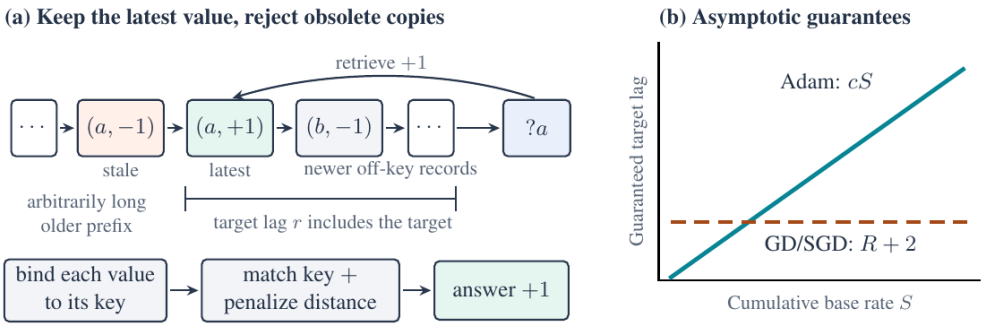

Figure 1: Latest-value retrieval and its optimizer-dependent range. The target competes with stale copies of its key and newer records of other keys. For maximum training recall lag $R \geq 2$, the right panel shows the $R + 2$ cutoff for GD and for SGD with averaging over all key pairs and growing batches, alongside Adam's growing guarantee cS for a protocol-dependent constant $c > 0$. Here $S \geq 0$ denotes the cumulative base learning rate. The curves describe asymptotic guarantees under our specified schedules.

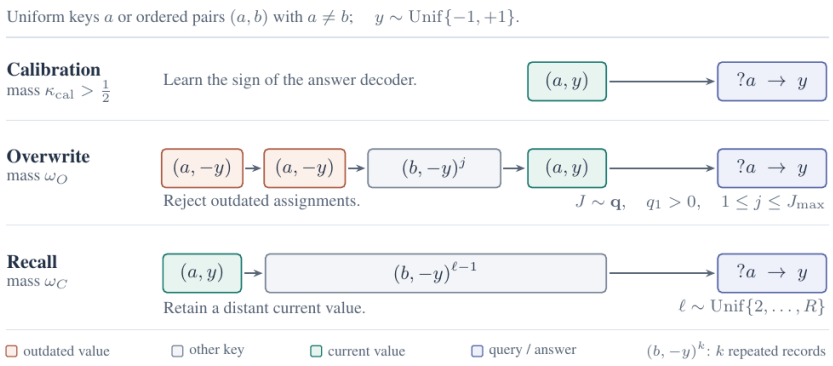

Figure 2: Training dataset components: calibration, overwrite, and recall rows teach three parts of latest-value retrieval. Both answer signs are included. Recall lag ℓ is uniform on $\{2,\ldots,R\}$, and overwrite spacing j follows $\mathbf{q}$.

length-generalizing sparse token selection (Wang et al., 2024). For fixed-offset retrieval, Singh et al. (2026) study positional encodings and implicit bias. Our comparison fixes the architecture and answer objective and derives optimizer-dependent retrieval boundaries under specified schedules. Our guarantees do not require uniformly bounded attention logits: the recency penalty grows with distance, and learned margins grow with training. These features distinguish the result from bounded-logit softmax limitations (Veličković et al., 2025). Forgetting controls arbitrarily long older prefixes, while the learned margins support increasing target lags across trained checkpoints.

Implicit bias. Attention dynamics have been linked to margin maximization and factorized query–key regularization (Tarzanagh et al., 2023; Sheen et al., 2024). Zero-stabilizer Adam has a maximum-ℓ_∞-margin bias for separable linear classification (Zhang et al., 2024); Adam and GD can also differ in feature learning and generalization in two-layer ReLU models (Vasudeva et al., 2025). Analyses of quasi-homogeneous networks and continuous-time Adam on smooth homogeneous networks describe parameterization-dependent margin bias (Kunin et al., 2023; Gronich & Vardi, 2026). The nested attention and answer decoder here fall outside a homogeneous predictor model. The proof must track both recency growth and amplified binding error, then resolve the logarithmic correction that determines the endpoint of the retrieval cutoff.

2 LATEST-VALUE RETRIEVAL WITH LEARNED RECENCY

2.1 TASK, TRAINING POPULATION, AND TEST ERROR

We now detail our setting concerning a particular failure case of length generalization in associative recalls. The task revolves around two conflicting demands on memory: retain a distant assignment and reject an obsolete assignment to the same key. For a vocabulary $\mathcal{K} = [n]$ of $n \geq 2$ keys, an input X has $L \geq 1$ binary-valued records and a query key a that occurs among them. Its latest matching index i_*, answer y, and target lag r are illustrated in Figure 1 and formally

$$X = ((a_1, v_1), \ldots, (a_L, v_L); ?a), \qquad a_i, a \in \mathcal{K}, \quad v_i \in \{-1, +1\}, \tag{1}$$

$$i_*(X) = \max\{i : a_i = a\}, \qquad y(X) = v_{i_*(X)}, \qquad r(X) = L - i_*(X) + 1. \tag{2}$$

For sanity check, the newest record has lag one, and note that older records before the target can make the context arbitrarily long without changing its lag.

To train both of these conflicting demands using answer labels, combine calibration rows S, recall rows C, and overwrite rows O (illustrated in Figure 2). Fix a maximum training recall lag $R \geq 2$ and overwrite count $J_{\max} \geq 1$. For distinct keys a, b, the row templates are in Figure 2 but formally,

$$S^a = ((a, +1); ?a), \tag{3}$$

$$C_\ell^{a,b} = ((a, +1), (b, -1)^{\ell-1}; ?a), \qquad 2 \leq \ell \leq R, \tag{4}$$

$$O_j^{a,b} = ((a, -1), (a, -1), (b, -1)^j, (a, +1); ?a), \qquad 1 \leq j \leq J_{\max}. \tag{5}$$

Exponents denote repetitions. Each displayed sequence and its global value complement (every value sign reversed) are equally likely. The training distribution mixes three families: calibration examples orient the decoder, recall examples encourage retention, and overwrite examples teach forgetting stale records. Their mixture weights are the probabilities of selecting each family: $\kappa_{\text{cal}} = \kappa$ for calibration, ω_C for recall, and ω_O for overwrite. We assume $\kappa > \frac{1}{2}, \omega_C, \omega_O > 0, \kappa + \omega_C + \omega_O = 1$. Calibration dominates the dataset to orient the decoder before accurate retrieval. Within each family, keys or ordered distinct key pairs are uniform. Independently, recall lag ℓ is uniform on $\{2, \ldots, R\}$ and overwrite spacing j follows the fixed law $\mathbf{q} = (q_1, \ldots, q_{J_{\max}})$ with $q_1 > 0$. The shortest spacing, present because $q_1 > 0$, will determine GD's balance.

To compare training and test performance, let $\mathcal{D}_{R,\mathbf{q}}$ denote this distribution, θ the trainable parameters, and P_θ the answer probability defined below. Define cross-entropy training loss $F_{R,\mathbf{q}}$ and worst-case test error $\mathcal{E}_\theta(r)$ for integer target lag bound $r \geq 1$ by

$$F_{R,\mathbf{q}}(\theta) = \mathbb{E}_{X \sim \mathcal{D}_{R,\mathbf{q}}}[-\log P_\theta(y(X) \mid X)], \tag{6}$$

$$\mathcal{E}_\theta(r) = \sup_{X : r(X) \leq r} \big(1 - P_\theta(y(X) \mid X)\big). \tag{7}$$

Remark 1. The test error supremum allows arbitrary keys and values outside the target and arbitrarily many older records. Small training loss alone leaves this quantity undetermined and the optimizer's *implicit bias* has to select how matching and forgetting extrapolate beyond the training lags.

2.2 A LEARNED TWO-LAYER MODEL

We use the model illustrated in Figure 3. On the sequence $a_1, v_1, \ldots, a_L, v_L, ?, a$, the first attention layer binds values to keys; the second retrieves the query key's value for a learned decoder. The $2nd+6$ trainable parameters are two query and key embedding tables $Q, K \in \mathbb{R}^{n \times d}$ of representation dimension $d \geq 1$, and six scalars q, p, u, v, x, w. The rows Q_a and K_a are the query and key representations of key a.

For a fixed scale $c_0 > 0$, the scalars define the matching gain m, binding penalty g, and record-recency penalty h by

$$m = c_0 pq, \qquad g = \text{softplus}(uv), \qquad h = \text{softplus}(x), \qquad \text{softplus}(z) = \log(1 + e^z). \tag{8}$$

For binding, each value attends to the preceding keys one and three tokens away, when present. The older key receives a fixed logit offset $D \in \mathbb{R}$. Together with the learned per-token forget factor e^{-g},

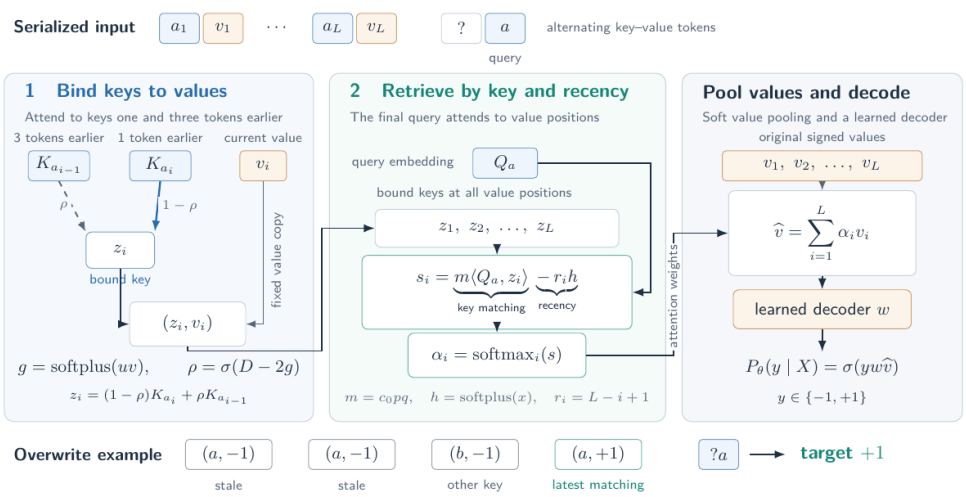

Figure 3: The two-layer architecture used in this paper.

this determines its weight, or *leakage*, ρ. Using the sigmoid $\sigma(z) = (1 + e^{-z})^{-1}$, define the leakage and bound key representation z_i by

$$\rho = \frac{e^{D-3g}}{e^{-g} + e^{D-3g}} = \sigma(D - 2g), \tag{9}$$

$$z_1 = K_{a_1}, \qquad z_i = (1 - \rho)K_{a_i} + \rho K_{a_{i-1}} \quad (i \geq 2). \tag{10}$$

Increasing g reduces leakage. The query uses the bound keys to attend to value positions. Its retrieval scores s_i, attention weights α_i, pooled value $\widehat{v}_\theta(X)$, and answer probability, with learned decoder weight w, are

$$s_i = m\langle Q_a, z_i\rangle - (L - i + 1)h, \qquad \alpha_i = \frac{e^{s_i}}{\sum_{j=1}^{L} e^{s_j}}, \tag{11}$$

$$\widehat{v}_\theta(X) = \sum_i \alpha_i v_i, \qquad P_\theta(y \mid X) = \sigma(yw\widehat{v}_\theta(X)). \tag{12}$$

Thus m amplifies matching while h penalizes distance. Amplification also magnifies binding errors, especially if table norms grow; the proof must therefore track both. The input format, attention masks, and signed value projections are fixed; learning acts through the two tables and six scalars (Appendix B.1).

Equivalent learned linear biases. These penalties arise from data-independent per-token FoX factors e^{-g} and $e^{-h/2}$. With two tokens per record, the second gives $e^{-(L-i+1)h}$ from value position $2i$ to query position $2L+2$. ALiBi-style token-distance slopes $g = \mathrm{softplus}(uv)$ and $h/2 = \mathrm{softplus}(x)/2$ thus give identical losses, gradients, and trajectories under the same masks and parameterizations. The equivalence excludes fixed slopes and direct optimization of slopes or native FoX gate biases. We next compare how GD and Adam balance matching against recency under a common training protocol, for either implementation.

3 Optimizer-dependent retrieval ranges

To compare how optimization balances matching and recency, Algorithms 1–2 in the Appendix specify GD/SGD and Adam updates on the answer loss. The diagonal matrix Λ_k contains the groupwise learning rates, each a fixed positive multiplier times the common base schedule. All initialization and schedule choices precede the random draw. Adam uses group-specific stabilizers and coordinatewise division, with $0/0 = 0$ for zero stabilizers.

3.1 INITIALIZATION, SCHEDULE, AND DATA ASSUMPTIONS

We compare the two optimizers under one unchanged schedule for all $2nd + 6$ parameters. For the fixed positional offset D, assume that the fixed training distribution $\mathcal{D}_{R,\mathbf{q}}$ has sufficient recall weight:

$$\omega_C > \frac{16}{27}(R-1)e^D\omega_O. \tag{13}$$

This makes the mean matching pressure positive at zero matching, before binding is accurate. The scalar factors are initialized $q_0 = p_0 > 0$, $u_0 = v_0 > 0$, with finite x_0 and $w_0 = 0$.

To obtain tractable GD geometry, initialize the stacked table $W_0 = [Q_0; K_0] \in \mathbb{R}^{2n\times d}$ by a uniform random rotation of $2n$ orthogonal rows of squared norm $s_0 > 0$:

$$W_0W_0^\top = s_0I_{2n}, \qquad Q_0K_0^\top = 0, \qquad d \geq 2n. \tag{14}$$

Write $\Delta_{ab,t} = \langle Q_{a,t}, K_{a,t} - K_{b,t}\rangle$ for the matching advantage, or *raw gap*, of key a over key $b \neq a$. By (14), every raw gap is initially zero. In the GD analysis they remain equal, with common value b_t. Let $\eta_t = \eta^*/(t+1)$ be the base learning rate, with fixed amplitude $\eta^* > 0$, and $S_t = \sum_{s<t}\eta_s$ its cumulative sum. Both optimizers use learning rates $\alpha_m\eta_t$ for (p, q), $\alpha_g\eta_t$ for (u, v), $\alpha_x\eta_t$ for x, $\alpha_w\eta_t$ for w, and $\lambda\eta_t$ for (Q, K). The α's and λ are called *multipliers*.

For Adam, we assume zero initial moments, no weight decay, and

$$0 \leq \beta_1 < 1, \quad \beta_1^2 < \beta_2 < 1, \qquad \epsilon_{q,p,u,v,x,w} = 0, \quad \epsilon_{Q,K} = E > 0. \tag{15}$$

We use $0/0 = 0$ for zero scalar gradients and moments. The table stabilizer permits a finite-time comparison with ordinary matrix gradients while the scalar updates retain Adam's normalization.

3.2 MAIN RESULTS

Theorem 2 (A shared initialization and one fixed schedule)**.** *Fix the population above, positive balanced initial scalar factors, $s_0, E > 0$, and $\zeta \in (0, 1)$. There are common fixed positive multipliers $\alpha_m, \alpha_g, \alpha_x, \alpha_w$ and λ, a sufficiently small $\eta^* > 0$, and a sufficiently large finite d such that full-batch GD and Adam from the same draw of* (14) *satisfy the following with probability at least $1 - \zeta$.*

(a) GD learns a common positive raw gap b_t and, for the effective matching advantage $M_t = m_tb_t$,

$$M_t - (R+1)h_t - \tfrac{3}{2}\log M_t \longrightarrow \log\frac{2\sqrt{c_0\alpha_m\lambda/(n-1)}\,\omega_C}{(R+1)(R-1)\alpha_x\omega_Oq_1}. \tag{16}$$

Consequently test error $\mathcal{E}_{\theta_t}(R+2) \to 0$, while every fixed $r \geq R+3$ has a fixed sequence with correct-answer probability tending to zero. The raw table norms diverge.

(b) Adam learns and retains positive raw gaps, its raw tables converge, and $(q_t, p_t, u_t, v_t, h_t, w_t)/S_t$ converges to $(\alpha_m, \alpha_m, \alpha_g, \alpha_g, \alpha_x, \alpha_w)$. For some deterministic $c > 0$, test error $\mathcal{E}_{\theta_t}(\lfloor cS_t\rfloor) \to 0$.

Both conclusions allow arbitrary older prefixes. The objective, multipliers, stabilizers, and schedule remain unchanged from initialization, with no stopping rule or moment reset.

The theorem requires a large binding multiplier α_g, small recency and decoder multipliers α_x and α_w, and a small common step amplitude η^*. It establishes that such a hierarchy exists. Note that we do not get a bound on necessary width or running time. Since $S_t \sim \eta^*\log t$, Adam's guaranteed lag grows only logarithmically in the number of updates. Appendices D and E prove the theorem and check that one choice of multipliers serves both optimizers.

For Adam and GD, the entire population loss is computed exactly. We also obtain a result for SGD when recall lags ℓ and overwrite spacings j (Figure 2) are sampled. Since calibration examples have identical gradients, we include their population contribution directly. The proof is in Appendix F.

Corollary 3 (SGD with averaging over all key pairs)**.** *The GD conclusion also holds with probability at least $1 - \zeta$ when each sampled lag ℓ and spacing j tuple is averaged over all $n(n-1)$ ordered key pairs and the batch schedule satisfies $N_t \geq \lceil N_0(t+1)^2\log(t+2)\rceil$ for sufficiently large N_0. The initialization and fixed multipliers are unchanged.*

Independent Gaussian initialization for Adam. Adam's proof tracks row norms and pairwise inner products over a finite interval. At large width, their initial values concentrate near (14), allowing approximate orthogonality. The proof also bounds the smallest singular values of the effective factors q_tQ_t and p_tK_t. For $A \in \mathbb{R}^{n\times d}$, write $\sigma_{\min}(A) = \sqrt{\lambda_{\min}(AA^\top)}$ for its smallest row singular value, with $\lambda_{\min}$ denoting the least eigenvalue. We also have the following results.

Theorem 4 (Adam from independent Gaussian tables)**.** *Fix the population satisfying* (13)*, $q_0 = p_0 > 0$, $u_0 = v_0 > 0$, $x_0 \in \mathbb{R}$, $w_0 = 0$, $s_0, E > 0$, the moment parameters in* (15)*, a decay exponent $\gamma \in (2/3, 1]$, and $\zeta \in (0, 1)$. There exist fixed positive multipliers, a sufficiently small $\eta^* > 0$, and a finite $d_0 \geq 2n$ such that, for every $d \geq d_0$, full-batch Adam with zero initial moments, stabilizers* (15)*, base rate $\eta_t = \eta^*(t+1)^{-\gamma}$, and independent $N(0, s_0/d)$ entries in Q_0, K_0 has the following properties with probability at least $1 - \zeta$. The raw tables converge, all raw matching gaps eventually exceed a deterministic $\delta_* > 0$, and the convergences in Theorem 2(b) hold. For a deterministic $c > 0$, with $S_t = \sum_{s<t} \eta_s$,*

$$\mathcal{E}_{\theta_t}(\lfloor cS_t \rfloor) \longrightarrow 0, \tag{17}$$

$$\min\{\sigma_{\min}(Q_t), \sigma_{\min}(K_t)\} \geq \tfrac{1}{2}\sqrt{s_0}, \tag{18}$$

$$\min\{\sigma_{\min}(q_tQ_t), \sigma_{\min}(p_tK_t)\} \geq \tfrac{1}{4}\alpha_m\sqrt{s_0}\, S_t \tag{19}$$

The spectral bounds hold for all sufficiently large t. The objective and multipliers remain fixed, and the schedule is chosen before the random draw.

The growing factors are precisely those used in the content scores, $m_tQ_tK_t^\top = c_0(q_tQ_t)(p_tK_t)^\top$. Theorem 20 and Corollary 24 prove these claims with a separately chosen fixed rate hierarchy. The corresponding Gaussian GD/SGD boundary remains open. All results apply unchanged to the matched learned-slope ALiBi implementation in Section 2.2.

Remark 5 (A fixed scalar stabilizer ϵ excludes the full convergence in Theorem 2(b)). If the decoder uses a fixed positive stabilizer and the raw tables stay bounded with matching gaps bounded below by a positive constant, the six linear scalar-speed limits in Theorem 2 cannot all hold (Proposition 12). Those limits would force the decoder gradient and its Adam direction to vanish, contradicting $w_t/S_t \to \alpha_w$. This rules out the stated speed law but slower unbounded growth remains possible.

4 Why the optimizers select different ranges: proof sketches

From matching gaps to uniform retrieval. The optimizer comparison reduces to how quickly matching grows relative to recency, provided binding errors vanish. Let B_t bound the raw query/key row norms and let $\Delta_{\min,t} = \min_{a\neq b}\Delta_{ab,t}$. Summing the competing attention weights as a geometric series

$$\mathcal{B}_t(r) = e^{4B_t^2 m_t\rho_t}\,\frac{e^{-m_t\Delta_{\min,t}+(r-1)h_t} + e^{-h_t}}{1 - e^{-h_t}} \tag{20}$$

gives the sufficient condition

$$\begin{aligned} h_t, w_t \to \infty, \qquad B_t^2 m_t\rho_t \to 0, \qquad m_t\Delta_{\min,t} - (r_t - 1)h_t \to \infty \\ \implies \quad \mathcal{E}_{\theta_t}(r_t) \to 0. \end{aligned} \tag{21}$$

This is because for every stream with target lag at most r, the total non-target attention weight divided by the target weight is at most $\mathcal{B}_t(r)$. Consequently, $P_{\theta_t}(y \mid X) \geq \mathrm{sigmoid}(w_t(1 - \mathcal{B}_t(r))/(1 + \mathcal{B}_t(r)))$. The geometric denominator makes this bound independent of the older prefix length. Appendix B.2 proves this criterion. We now explain why GD and Adam produce different matching margins.

GD couples matching and table growth. Calibration makes the decoder weight w positive, and the recall-weight condition then favors matching even before binding is accurate. Uniform key averaging preserves the symmetry of the jointly orthogonal initialization: with $J = \mathbf{1}\mathbf{1}^\top/n$ and $P = I - J$, the exact GD iterates satisfy

$$Q_tQ_t^\top = K_tK_t^\top = A_tP + s_0J, \qquad Q_tK_t^\top = b_tP. \tag{22}$$

Thus every raw matching gap equals b_t, the effective matching advantage is $M_t = m_tb_t$, and A_t controls the table row norms.

The Gram identities (Lemma 14) constrain scalar and table growth together: as $M_t \to \infty$, $m_t, b_t, A_t \asymp \sqrt{M_t}$. Both scalar-factor and table updates therefore contribute a matching coefficient of order $M^{3/2}$. For the corresponding gradient flow in cumulative time $S_t = \sum_{j<t} \eta_j$, the matching field is

$$\frac{dM}{dS} = -\mu_M F_M, \qquad \mu_M \asymp M^{3/2}, \tag{23}$$

where F_M is the derivative of the reduced population loss with respect to M. The same leading coefficient governs discrete GD, with the discretization errors controlled by the argument below.

The balance gives a finite GD boundary. After removing a common loss-derivative factor, the longest recall row contributes matching pressure proportional to $e^{-M+(R-1)h}$, while the shortest overwrite, with mass $q_1 > 0$, contributes recency pressure proportional to e^{-2h}. Together with the matching coefficient, this suggests the balance

$$M^{3/2} e^{-M+(R-1)h} \asymp e^{-2h}. \tag{24}$$

Taking logarithms motivates the residual $f = M - (R+1)h - \frac{3}{2}\log M$.

The main technical step is to establish this balance along the actual GD trajectory. The fixed multiplier hierarchy (the α's and λ) brings the iterates into a forward-invariant region in finite time. Within this region, binding leakage and discrete errors remain controlled, $M_t, h_t, w_t \to \infty$, and $B_t^2 m_t \rho_t \to 0$. A restoring drift forces $f_t \to C_*$, where C_* is the constant in (16). Appendix D proves entrance and convergence under the same unchanged rates.

Consequently,

$$M_t - (r-1)h_t = (R+2-r)h_t + \tfrac{3}{2}\log M_t + C_* + o(1). \tag{25}$$

At $r = R+2$, the logarithmic correction still diverges, so (21) gives uniform retrieval through $R+2$. For any fixed $r \geq R+3$, the negative linear term dominates because $\log M_t = o(h_t)$. On the stream $((a,+1),(b,-1)^{r-1};?a)$ with $a \neq b$, the newest wrong record then overwhelms the target, and the correct-answer probability tends to zero.

Normalized scalar growth produces a growing range for Adam. Adam changes this balance through normalized scalar updates $d_{i,t} = -\widehat{M}_{i,t}/\sqrt{\widehat{V}_{i,t}}$ and the chosen stabilizer setting (0 for scalar, $E > 0$ for table parameters). At large width and sufficiently small steps, Adam's initial table updates approximate ordinary gradient descent with multiplier λ/E, using the scalar values generated by Adam. Calibration and favorable matching pressure then create a positive match gap in this reference dynamics, which transfer to Adam's.

To preserve this gap, the continuation argument (Appendix E) jointly bounds the query/key row norms by B, controls Adam's first and second moments for the scalar parameters, and bounds the binding error ρ_t. A sufficiently large fixed α_g keeps binding ahead of matching, ensuring $\rho \leq \rho_* e^{-4B^2 m}$. The moment bounds give $q, p, u, v = \Theta(S)$, hence $m = \Theta(S^2)$, while $h, w = O(S)$. The table gradients therefore satisfy

$$\|\nabla_{Q,K} F\|_F \leq C(1+S)^3 e^{-cS^2+CS}. \tag{26}$$

This integrable bound and $E > 0$ control the total remaining table movement, including the effect of earlier gradients stored in Adam's moments. This movement is too small to erase the learned matching gaps, and the tables converge.

Quadratic matching also suppresses the adverse recall contribution to the recency gradient; the recency and decoder gradients eventually become favorable. Relative changes in all scalar gradients vanish. Tracking both moments with their original bias corrections then gives $d_{i,t} \to 1$ for every scalar coordinate i. Thus

$$\begin{gathered}(q_t, p_t, u_t, v_t, h_t, w_t)/S_t \longrightarrow (\alpha_m, \alpha_m, \alpha_g, \alpha_g, \alpha_x, \alpha_w), \\ m_t \sim c_0 \alpha_m^2 S_t^2, \qquad h_t \sim \alpha_x S_t.\end{gathered} \tag{27}$$

Normalization cancels the shrinking scalar gradient magnitudes: both matching factors grow linearly, so their product grows quadratically, while the recency penalty grows linearly. With $\Delta_{\min,t} \geq \delta_* > 0$, choose $r_t = \lfloor c_* S_t \rfloor$ for $0 < c_* < \delta_* c_0 \alpha_m^2/(2\alpha_x)$. Then

$$m_t \Delta_{\min,t} - (r_t - 1)h_t \geq (\delta_* c_0 \alpha_m^2 - c_* \alpha_x + o(1)) S_t^2 \to \infty. \tag{28}$$

The binding region also gives $B^2 m_t \rho_t \to 0$. Together with $h_t, w_t \to \infty$, (21) proves a retrieval range growing proportionally to S_t. Appendix E gives the full argument.

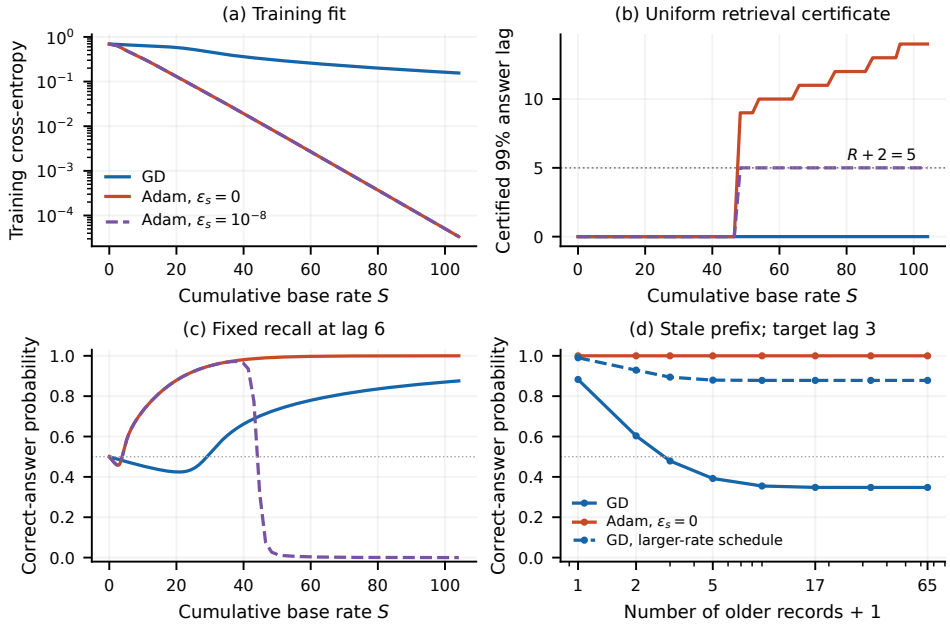

Figure 4: Seed-1 results: (a) training loss, (b) sufficient uniform 99% answer lag, (c) lag-six recall, and (d) stale prefixes at fixed lag three. The Adam control changes only its scalar stabilizer; dashed GD uses tenfold larger rates from initialization.

5 Experiments

For a proof-of-concept experiment, we train all parameters of the structured model for 10^5 updates with recall lags through $R = 3$, comparing GD and Adam from shared initializations and schedules.

Finite-time retrieval. Applying Proposition 6 to the learned parameters, all three Adam seeds certify uniform 99% correct-answer probability through lag 14, including arbitrary older prefixes (Figure 4). GD also learns positive gaps, but its minimum training probability is 0.5852, versus 0.999966 for Adam, so this comparison includes unequal training fit. Prepending 64 stale records to a lag-three training recall gives correct-answer probabilities 0.3476 for GD and 0.999966 for Adam (seed 1).

Stabilizer sensitivity at matched fit. Changing only Adam's scalar stabilizer from zero to 10^{-8} leaves training loss near 3.36×10^{-5}, but reduces its certified 99% lag from 14 to 5 and its lag-six correct-answer probability from 0.999966 to 4.57×10^{-5}. This isolates a finite-time retrieval difference despite nearly identical training fit; it does not establish a permanent retrieval limit.

Scope. Larger rates change both optimizers' finite-budget results (Appendix G.2). A fully learned two-layer Transformer under separate stochastic protocols shows no consistent Adam–SGD separation when evaluation allows arbitrary target predecessors (Appendix G.3). These results support the conventional wisdom that enabling length generalization requires careful learning designs, including that of the optimizers. Appendix G gives the detailed protocols and controls.

6 Discussion

Optimizer choice can change length generalization even when the architecture, training objective, and initialization are held fixed. Our theory establishes this dependence in a structured model under specialized training conditions. Extending the guarantees to standard minibatch training and less constrained Transformers remains open. The fully learned Transformer experiments show no consistent Adam–SGD separation under the tested protocols with unrestricted target predecessors. These findings motivate evaluating length generalization directly when selecting optimizers and hyperparameters.

REPRODUCIBILITY STATEMENT

Section 3.1 states the initialization, optimizer schedules, and assumptions of the main results; Algorithms 1 and 2 specify the updates. Appendix B derives the architecture and retrieval guarantees, and Appendices C–F give the optimization proofs. Appendix G describes the synthetic tasks, experimental configurations, seed allocation, rate selection, numerical implementation, and evaluation metrics. The accompanying anonymous code supplement provides Python implementations of the structured model, GD, Adam, the SGD controls, the learned Transformer, synthetic-data generation, retrieval evaluation, and table aggregation.

AI USE STATEMENT

In this work, we used OpenAI Codex and Claude to help, in part, and under our guidance, develop the theoretical model and refine hypotheses, formulate mathematical claims, develop key proof arguments and draft proofs, design experiments, implement methods and code for generating synthetic data, and interpret results. These contributions concerned the GD–Adam separation, its stochastic and initialization extensions, and the empirical analysis. We did not use generative AI for translation; dataset cleaning and reformatting and qualitative or thematic data analysis were not applicable to this work. Additionally, we used these tools to create scientific figures, draft manuscript text, edit the paper for readability, and identify relevant literature. We have reviewed all AI-assisted work, including the mathematical arguments, code, and empirical claims. The implementation was checked against literal attention computations, finite-difference derivatives, and optimizer recurrences, and citations were checked against primary sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fb7451e43f9c1c35b774bcfad7a5714b-Abstract-Conference.html.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0561738a239a995c8cd2ef0e50cfa4fd-Abstract-Conference.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Omer Elkabetz and Nadav Cohen. Continuous vs. discrete optimization of deep neural networks, 2021. URL https://arxiv.org/abs/2107.06608.

Eitan Gronich and Gal Vardi. The implicit bias of Adam and Muon on smooth homogeneous neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DpIc1cpNKG.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306. URL https://aclanthology.org/2020.tacl-1.11/.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. In *International Conference on Learning Representations*,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/928170bcb050fe64a63fe781b82265aa-Abstract-Conference.html.

Zachary Izzo, Eshaan Nichani, and Jason D. Lee. Quantitative bounds for length generalization in transformers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4a765a1bbf8332154e58bf452fbf806c-Abstract-Conference.html.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24892–24928, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e85362c02172c0c6567ce593122d31c-Abstract-Conference.html.

Daniel Kunin, Atsushi Yamamura, Chao Ma, and Surya Ganguli. The asymmetric maximum margin bias of quasi-homogeneous neural networks. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IM4xp7kGI5V.

Zhixuan Lin, Evgenii Nikishin, Xu Owen He, and Aaron Courville. Forgetting transformer: Softmax attention with a forget gate. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/add3d389197ad2267f660ad060ef61f4-Abstract-Conference.html.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks, 2020. URL https://arxiv.org/abs/1906.05890.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38018–38070. PMLR, 2024. URL https://proceedings.mlr.press/v235/nichani24a.html.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1889–1903. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.161. URL https://aclanthology.org/2023.acl-short.161/.

Heejune Sheen, Siyu Chen, Tianhao Wang, and Harrison H. Zhou. Implicit regularization of gradient flow on one-layer softmax attention. *arXiv preprint arXiv:2403.08699*, 2024. URL https://arxiv.org/abs/2403.08699.

Subham Singh, Ashutosh Mishra, and Subha Raut. Relative positions generalize, absolute positions memorize: An implicit-bias account of length generalization in attention. *arXiv preprint arXiv:2607.18759*, 2026. URL https://arxiv.org/abs/2607.18759.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023. URL https://arxiv.org/abs/2308.16898.

Bhavya Vasudeva, Jung Hoon Lee, Vatsal Sharan, and Mahdi Soltanolkotabi. The rich and the simple: On the implicit bias of Adam and SGD. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ec9b2a6ad5444caeff75efaa6176b3e4-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61190–61211. PMLR, 2025. URL https://proceedings.mlr.press/v267/velickovic25a.html.

Mert Vural, Alberto Bietti, Mahdi Soltanolkotabi, and Denny Wu. Learning to recall with transformers beyond orthogonal embeddings. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79ce24f9e8d3c4ff5919240eac78a782-Abstract-Conference.html.

Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D. Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51854–51912, 2024. URL https://proceedings.mlr.press/v235/wang24ca.html.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11080–11090. PMLR, 2021. URL https://proceedings.mlr.press/v139/weiss21a.html.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of Adam on separable data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2ac79356a03fe5e9250e5e77ebc76e6e-Abstract-Conference.html.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Joshua M. Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/45ed1a72597594c097152ef9cc187762-Abstract-Conference.html.

ORGANIZATION OF THE APPENDICES

Appendix A contains the full optimization algorithms for completeness. Appendix B derives the architecture (B.1) and converts learned score margins into a uniform bound for retrieval errors (B.2). Appendix C provides bounds and asymptotics of the population derivatives and Adam moment estimates. These enter the GD and Adam proofs and the SGD proof with averaging over all key pairs (Appendices D–F). Appendix G specifies the experiments and their additional diagnostics.

A OPTIMIZATION ALGORITHMS

Algorithm 1 GD/SGD on answer loss

Require: T, θ_0, rates $\{\Lambda_k\}$,
loss $F_{R,\mathbf{q}}$, batch sizes $\{N_k\}$

for $k = 0, \dots, T-1$ **do**
  Sample $\widehat{F}_k$ as in Corollary 3.
  $G_k \leftarrow \nabla \widehat{F}_k(\theta_k)$
  $\theta_{k+1} \leftarrow \theta_k - \Lambda_k G_k$
end for
return θ_T
GD: use $G_k = \nabla F_{R,\mathbf{q}}(\theta_k)$.

Algorithm 2 Full-batch Adam on answer loss

Require: T, θ_0, rates $\{\Lambda_k\}$, loss $F_{R,\mathbf{q}}$,
β_1, β_2, stabilizers $\{\epsilon_k\}$

$M_{-1} \leftarrow 0, V_{-1} \leftarrow 0$
for $k = 0, \dots, T-1$ **do**
  $G_k \leftarrow \nabla F_{R,\mathbf{q}}(\theta_k)$
  $M_k \leftarrow \beta_1 M_{k-1} + (1-\beta_1) G_k$
  $V_k \leftarrow \beta_2 V_{k-1} + (1-\beta_2) G_k^2$
  $\theta_{k+1} \leftarrow \theta_k - \Lambda_k \dfrac{M_k/(1-\beta_1^{k+1})}{\sqrt{V_k/(1-\beta_2^{k+1})} + \epsilon_k}$
end for
return θ_T

B POSITIONAL ROUTING AND UNIFORM RETRIEVAL BOUNDS

The optimization proofs use two facts about the stated architecture. Positional routing realizes its prescribed binding masks, and a geometric tail bound transfers learned margins to answer accuracy. The grammar, attention masks, and value projections are fixed as specified in Section 2; learning those components is outside the optimization theorem.

B.1 A POSITIONAL IMPLEMENTATION OF THE PRESCRIBED ARCHITECTURE

The first head must distinguish key and value positions at every length. The following construction implements that distinction from a fixed positional feature, before applying the learned tables.

Number raw token positions starting at one, with alternating keys at odd positions and values at even positions. The single fixed feature $\cos(\pi t)$ gives exact role bits

$$r_K(t) = \frac{1-\cos(\pi t)}{2}, \qquad r_V(t) = \frac{1+\cos(\pi t)}{2}. \tag{29}$$

For a bounded raw token coordinate $z \in [-B, B]$ and $r \in \{0, 1\}$, the pointwise ReLU expression

$$\mathrm{route}(z, r) = \mathrm{ReLU}(z + Br) - \mathrm{ReLU}(z - Br) - Br \tag{30}$$

is zero when $r = 0$ and equals z when $r = 1$. Thus fixed bounded one-hot token coordinates and signed value bits can be routed before the learned linear tables are applied; the routing acts on the bounded token coordinates, so it does not require bounded learned tables. A residual copy of the raw token supplies the final query key even when that query has the other parity.

At a value position the first head has eligible raw-token lags $1, 3$. Let c_1, c_3 be fixed positional logits and set $D = c_3 - c_1$. With relative logits $c_j - jg$, the probability of the lag-three candidate is exactly

$$\rho = \mathrm{sigmoid}(D - 2g). \tag{31}$$

At the first record, lag three is outside the input and is masked, so the first bound key vector equals its preceding key. Strict causality excludes self-attention. The second head's record-position mask excludes keys and the final query from its candidate values; its recency penalty h is measured per record, whereas g above is per raw-token lag. These rules give the boundary convention and the factors of two in the soft-binding formulas. The masks depend only on positions in the specified grammar. Finite positional logits alone cannot produce exact zero attention weights, so the proof retains these masks explicitly. Nor does it assert that a learned finite absolute-position table learns an extrapolating parity rule.

For signed values let $\widehat{v} = \sum_i \alpha_i v_i$. Independent two-class linear logit weights θ_+, θ_- give

$$P(y \mid X) = \mathrm{sigmoid}(yw\widehat{v}), \qquad w = \theta_+ - \theta_-. \tag{32}$$

The loss depends on the two weights only through w. Their gradients and first moments are opposite and their second moments are equal, provided they start with zero buffers and use the same moment coefficients and stabilizer. With rates $\alpha_+\eta_t$ and $\alpha_-\eta_t$, their difference therefore follows exact scalar Adam with learning-rate multiplier $\alpha_w = \alpha_+ + \alpha_-$. No tying of the weights or opposite initialization is required; an arbitrary finite common mode cancels from the softmax probabilities. If a bias is included, zero initial bias and exact value-complement balance make its gradient cancel at every step: complementation sends $(y, \widehat{v})$ to $(-y, -\widehat{v})$ and reverses the bias derivative. For a positive stabilizer this zero update is automatic. With zero stabilizer one can omit the bias, or use the zero-update convention for a zero-over-zero coordinate. Arbitrary nonzero bias initialization is not covered by this symmetry argument.

B.2 Uniform Retrieval Bound

We derive a sufficient condition to obtain uniform bound on the prediction error $\mathcal{E}_{\theta_t}(r)$ after training iteration t. At training iteration t, let $\Delta_{ab,t} = \langle Q_{a,t}, K_{a,t} - K_{b,t} \rangle$ be the matching gap between keys $a \neq b$, $\Delta_{\min,t} = \min_{a \neq b} \Delta_{ab,t}$, and $\mathcal{M}_t = m_t \Delta_{\min,t}$. The main idea of the uniform bound is by bounding the worst-case non-target-to-target attention odds, which can be used to bound $\mathcal{E}_{\theta_t}(r)$.

Proposition 6 (Uniform retrieval under learned matching and binding)**.** *Suppose after iteration t, $m_t, h_t > 0, w_t \geq 0, \Delta_{\min,t} \geq 0$, and all query and key row norms are at most $B_t > 0$. Let*

$$\mathcal{B}_t(r) = e^{4B_t^2 m_t \rho_t} \frac{e^{-\mathcal{M}_t + (r-1)h_t} + e^{-h_t}}{1 - e^{-h_t}},$$

then

$$\mathcal{E}_{\theta_t}(r) \leq \mathrm{sigmoid}\left[-w_t \frac{1 - \mathcal{B}_t(r)}{1 + \mathcal{B}_t(r)}\right] \tag{33}$$

Consequently, for any integer sequence $r_t \geq 1$,

$$h_t, w_t \to \infty, \qquad B_t^2 m_t \rho_t \to 0, \qquad \mathcal{M}_t - (r_t - 1)h_t \to \infty \tag{34}$$

imply $\mathcal{E}_{\theta_t}(r_t) \to 0$.

Proof. For simplicity, suppress t and denote the target index by i^*. For $i \geq 2$, binding (10) gives $z_i - K_{a_i} = \rho(K_{a_{i-1}} - K_{a_i})$, so $\|z_i - K_{a_i}\| \leq \rho(\|K_{a_i}\| + \|K_{a_{i-1}}\|) \leq 2B\rho$. For $i = 1$, $z_1 = K_{a_1}$, so the same bound holds.

Fix an input X whose target index satisfies $L - i^* + 1 \leq r$, and let $y = v_{i^*}$ be its correct answer. Define the exact-binding scores $s_i^0 = m\langle Q_a, K_{a_i} \rangle - (L - i + 1)h$. The actual non-target-to-target attention odds are

$$A = \sum_{i \neq i^*} e^{s_i - s_{i^*}} = \sum_{i \neq i^*} \exp\left(m\langle Q_a, z_i - z_{i^*} \rangle + (i - i^*)h\right).$$

With $\Delta_{aa} = 0$, the exact-binding score difference is

$$s_i^0 - s_{i^*}^0 = -m\langle Q_a, K_a - K_{a_i}\rangle + (i - i^*)h = -m\Delta_{aa_i} + (i - i^*)h.$$

For older positions this is at most $(i - i^*)h$, since all gaps are nonnegative. Every newer position has $a_i \neq a$ because i^* is the latest match, so its score difference is at most $-\mathcal{M}+(i-i^*)h$. Consequently,

$$A^0 = \sum_{i<i^*} e^{s_i^0 - s_{i^*}^0} + \sum_{i>i^*} e^{s_i^0 - s_{i^*}^0} \tag{35}$$

$$\leq \sum_{i<i^*} e^{(i-i^*)h} + e^{-\mathcal{M}} \sum_{j=1}^{r-1} e^{jh} \tag{36}$$

$$\leq \frac{e^{-h}}{1 - e^{-h}} + \frac{e^{-\mathcal{M}+(r-1)h}}{1 - e^{-h}}. \tag{37}$$

The binding-error bound gives $|s_i - s_i^0| = m|\langle Q_a, z_i - K_{a_i}\rangle| \leq 2B^2 m\rho$. Thus $s_i - s_{i^*} \leq s_i^0 - s_{i^*}^0 + 4B^2 m\rho$, and (37) implies $A \leq \mathcal{B}(r)$.

Since $yv_{i^*} = 1$ and $yv_i \geq -1$,

$$\begin{aligned} y\widehat{v} &= \alpha_{i^*} + \sum_{i \neq i^*} \alpha_i y v_i \\ &\geq \alpha_{i^*} - \sum_{i \neq i^*} \alpha_i = 2\alpha_{i^*} - 1. \end{aligned}$$

The identity $\alpha_{i^*} = 1/(1 + A)$ and monotonicity of $(1 - z)/(1 + z)$ give $y\widehat{v} \geq (1 - A)/(1 + A) \geq (1 - \mathcal{B}(r))/(1 + \mathcal{B}(r))$. Hence, since $w \geq 0$,

$$1 - P_\theta(y \mid X) = \sigma(-wy\widehat{v}) \leq \sigma\left(-w\frac{1 - \mathcal{B}(r)}{1 + \mathcal{B}(r)}\right).$$

Taking the supremum over X proves (33). □

C Technical tools: Population Gradient and Adam Moments

In this section, we derive the properties of parameter gradients and Adam moments used in analyzing later training arguments. Throughout this section, we analyze using the loss $F_{R,\mathbf{q}}$ (6) and the fixed training distribution $\mathcal{D}_{R,\mathbf{q}}$ defined in Section 2.1.

C.1 Exact odds and matching acquisition

Lemma 7 (Overwrite and Recall odds)**.** *For a matching gap $z = \langle Q_a, K_a - K_b\rangle$, the overwrite odds are $T_{O,j}(z) = U_j(z) + B_j(z) + E_j(z)$, where*

$$U_j(z) = e^{zm\rho}\big(e^{-(j+1)h} + e^{-(j+2)h}\big),$$

$$B_j(z) = e^{-zm(1-2\rho)-jh}, \qquad E_j(z) = e^{-zm(1-\rho)} \sum_{k=1}^{j-1} e^{-kh}.$$

and the recall odds are

$$T_{C,\ell}(z) = e^{-zm(1-\rho)+h} + e^{-zm} \sum_{k=2}^{\ell-1} e^{kh}.$$

Proof. Consider an overwrite sequence $(a, -y), (a, -y), (b, -y), (b, -y), \ldots, (b, -y), (a, y)$ with j intermediate records of $(b, -y)$. Because the target follows a $b-$record, its bound key is $\widetilde{K} = (1 - \rho)K_a + \rho K_b$. The overwrite odds can then be analyzed by considered three parts of the sequence

- **The first two stale records** $(a, -y)$ **contributes** $U_j(z)$**.** The first record binds to K_a, and the second record binds to $(1-\rho)K_a + \rho K_a = K_a$. Their content-score advantage over the target is therefore $m\langle Q_a, K_a - \widetilde{K}\rangle = m\rho\langle Q_a, K_a - K_b\rangle = mz\rho$. Since they lie $j+2$ and $j+1$ records before the target, we have $U_j(z) = e^{zm\rho}\big(e^{-(j+1)h} + e^{-(j+2)h}\big)$.
- **The first record** $(b, -y)$ **contributes** $B_j(z)$**.** Its bound key is $K' = (1-\rho)K_b + \rho K_a$. Hence, $K' - \widetilde{K} = -(1-2\rho)(K_a - K_b)$, and this contributes odds of $B_j(z) = e^{-zm(1-2\rho)-jh}$.
- **The remaining** $b-$**records contributes** $E_j(z)$**.** Their bound keys is exactly K_b, and thus $K_b - \widetilde{K} = -(1-\rho)(K_a - K_b)$, which implies the odds $E_j(z) = e^{-zm(1-\rho)}\sum_{k=1}^{j-1} e^{-kh}$.

Summing these parts together, we get $T_{O,j}(z) = U_j(z) + B_j(z) + E_j(z)$, as desired.

Consider now a recall sequence $(a, y), (b, -y), \ldots, (b, -y)$. The recall odds can be calculated similarly as above, by noting that the first $b-$record contributes odds of $e^{-zm(1-\rho)+h}$, and the remaining $b-$records contribute odds of $e^{-zm}\sum_{k=2}^{\ell-1} e^{kh}$. □

For odds T, define the signed pooled value, answer loss, and loss derivative

$$\begin{aligned} s(T) &= \frac{1-T}{1+T}, & \ell(T, w) &= \mathrm{softplus}[-ws(T)], \\ K(T, w) = \ell_T(T, w) &= \frac{2w\,\mathrm{sigmoid}[-ws(T)]}{(1+T)^2}, & -\ell_w(T, w) &= s(T)\,\mathrm{sigmoid}[-ws(T)]. \end{aligned} \tag{38}$$

With $\vartheta = (q, p, u, v, x, w)$, $N_{\mathrm{pair}} = n(n-1)$, and $\kappa_{\mathrm{cal}} = \kappa$, the population loss based on data distribution $\mathcal{D}_{R,\mathbf{q}}$ is

$$\begin{aligned} \phi_\vartheta(z) &= \omega_O \sum_{j=1}^{J_{\max}} q_j \ell(T_{O,j}(z), w) + \frac{\omega_C}{R-1}\sum_{\ell=2}^{R} \ell(T_{C,\ell}(z), w), \\ F_{R,q} &= \kappa_{\mathrm{cal}}\,\mathrm{softplus}(-w) + \frac{1}{N_{\mathrm{pair}}}\sum_{a\neq b}\phi_\vartheta(\Delta_{ab}) \end{aligned} \tag{39}$$

The following lemma allows us to argue about loss gradient at zero gap, which happens at initialization.

Lemma 8 (Loss derivative near zero gap)**.** *Suppose $q, p, u, v, w > 0$, x is finite, $\omega_C/(R-1) > (16/27)e^D\omega_O$, and $0 \le \rho \le e^D/(4+e^D)$. Let $d_\vartheta = -\phi'_\vartheta$. The followings simultaneously hold*

(a) $d_\vartheta(0) > 0$.

(b) On any fixed compact positive intervals for m, h and bounded interval $0 \le w \le W$, the ratio $d_\vartheta(z)/w$ extends continuously to $w = 0$, and there exist $z_*, c_-, c_+ > 0$ such that for $|z| \le z_*$,

$$c_- w \le d_\vartheta(z) \le c_+ w \tag{40}$$

(c) For the score matrix $H = QK^\top$ and centered projector $P = I - \mathbf{1}\mathbf{1}^\top/n$, its gradient at zero scores is

$$\nabla_H F_{R,\mathbf{q}}|_{H=0} = -\frac{d_\vartheta(0)}{n-1}P. \tag{41}$$

Proof. Let $t = e^{-h} \in (0, 1)$. By direct differentiation,

$$m^{-1}\frac{d}{dz}\ell(T_{C,2}(z), w)\Big|_{z=0} = -(1-\rho)t^{-1}K(t^{-1}, w) < 0,$$

and

$$m^{-1}\frac{d}{dz}T_{O,j}(z)\Big|_{z=0} = \rho\sum_{k=1}^{j+2} t^k - (1-\rho)t^j - \sum_{k=1}^{j-1} t^k \le \rho\sum_{k=1}^{j+2} t^k.$$

Denote $\mathrm{t}_j = \sum_{k=1}^{j+2} t^k$ and $f(s) = \frac{s}{(1+s)^2}$. By some standard calculus, we can show that $f(\mathrm{t}_j) \leq \frac{32}{27} f(t)$ for all $t \in [0,1]$. Now, by chain rule

$$\begin{aligned}\partial_z \ell(T_{O,j}(z), w)\Big|_{z=0} &= K(\mathrm{t}_j, w) \frac{d}{dz} T_{O,j}(z)\Big|_{z=0} \\ &\leq m\rho \mathrm{t}_j K(\mathrm{t}_j, w) \\ &= 2mw\rho f(\mathrm{t}_j)\sigma(-ws(\mathrm{t}_j)) \\ &\leq 2mw\rho f(\mathrm{t}_j)\end{aligned}$$

On the other hand,

$$-\frac{d}{dz}\ell(T_{C,2}(z), w)\Big|_{z=0} = m(1-\rho)t^{-1}K(t^{-1}, w) \geq mw(1-\rho)f(t).$$

Therefore,

$$\frac{\partial_z \ell(T_{O,j}(z), w)\big|_{z=0}}{-\frac{d}{dz}\ell(T_{C,2}(z), w)\big|_{z=0}} \leq 2\frac{\rho}{1-\rho}\frac{f(\mathrm{t}_j)}{f(t)} \leq \frac{64}{27}\frac{\rho}{1-\rho} \leq \frac{16e^D}{27}.$$

Now, differenting the population loss, we have

$$\begin{aligned}d_\vartheta(0) &= -\phi'_\vartheta(0) \\ &= \frac{\omega_C}{R-1}\sum_{\ell=2}^{R} -\partial_z \ell(T_{C,\ell}(z), w)\big|_{z=0} - \omega_O \sum_j q_j \partial_z \ell(T_{O,j}(z), w)\big|_{z=0} \\ &\geq -\partial_z \ell(T_{C,2}(z), w)\big|_{z=0}\left[\frac{\omega_C}{R-1} - \frac{16}{27}e^D\omega_O\right] > 0.\end{aligned}$$

Now note that by chain rule,

$$\begin{aligned}\frac{d_\vartheta(z)}{w} =& -\omega_O \sum_j q_j \frac{K(T_{O,j}(z), w)}{w} T'_{O,j}(z) \\ &- \frac{\omega_C}{R-1}\sum_{\ell=2}^{R}\frac{K(T_{C,\ell}(z), w)}{w} T'_{C,\ell}(z),\end{aligned}$$

where $\frac{K(T,w)}{w} = \frac{2\sigma(-ws(T))}{(1+T)^2}$ for any $w > 0$. Hence, the limit of $d_\vartheta(z)/w$ is well-defined as $w \to 0$. At $z = 0$, the preceding estimates give

$$\frac{d_\vartheta(0)}{w} \geq \left(\frac{\omega_C}{R-1} - \frac{16}{27}e^D\omega_O\right) m(1-\rho)\frac{e^h}{(1+e^h)^2} \geq c_{\mathrm{loc}} > 0,$$

uniformly for m, h in the fixed positive compact intervals, $0 \leq w \leq W$, and $0 \leq \rho \leq e^D/(4+e^D)$, using the continuous extension at $w = 0$. Uniform continuity therefore gives $z_* > 0$ such that $d_\vartheta(z)/w \geq c_{\mathrm{loc}}/2 =: c_-$ for $|z| \leq z_*$. Continuity on the same compact region also supplies a finite upper bound c_+. This proves (40).

Now, note that

$$F_{R,\mathbf{q}}(H) = \kappa_{\mathrm{cal}}\mathrm{softplus}(-w) + \frac{1}{n(n-1)}\sum_{a\neq b}\phi_\vartheta(H_{aa} - H_{ab}).$$

For $a \neq b$, $\frac{\partial F}{\partial H_{ab}} = -\frac{\phi'_\vartheta(H_{aa}-H_{ab})}{n(n-1)} = \frac{d_\vartheta(\Delta_{ab})}{n(n-1)}$. Meanwhile, $\frac{\partial F}{\partial H_{aa}} = \frac{1}{n(n-1)}\sum_{b\neq a}\phi'_\vartheta(H_{aa} - H_{ab}) = -\frac{1}{n(n-1)}\sum_{b\neq a} d_\vartheta(\Delta_{ab})$. At $H = 0$, $\Delta_{ab} = 0$ for all a, b, implying (41). □

C.2 Relative gradients under controlled leakage

Write $L(z) = e^{-zm(1-\rho)+h}$ for the first recall-odds term in Lemma 7. Differentiating the odds with respect to leakage gives the binding derivatives used below:

$$\partial_\rho T_{O,j} = zm(U_j + 2B_j + E_j), \qquad \partial_\rho T_{C,\ell} = zmL. \tag{42}$$

We provide the bound on gradients with conditions that can be chosen by the training proof.

Lemma 9 (Relative gradient bounds)**.** *Fix $B, \delta, k_g > 0$ and positive lower bounds on q, p, u, v, w. Suppose tables Q, K have row norms at most B, every gap lies in $[\delta, 2B^2]$, and*

$$\rho \le \rho_* e^{-4B^2 m}, \qquad \rho_* = \tfrac{1}{4} e^{-k_g}. \tag{43}$$

Then $a_i = -F_i > 0$ for $i \in \{q, p, u, v\}$, and

$$\|\nabla_\vartheta \log a_i\| \le C(1 + \|\vartheta\|)^2, \qquad \|\nabla_{Q,K} \log a_i\|_F \le C(1 + \|\vartheta\|)^3. \tag{44}$$

Note that no lower bound on x or $h = \text{softplus}(x)$ is required. The constant C depends only on the fixed bounds and finite population.

Proof. The matching derivatives are

$$\begin{aligned} -\partial_m T_{O,j} &= z\big[(1-2\rho)B_j + (1-\rho)E_j - \rho U_j\big], \\ -\partial_m T_{C,\ell} &= z\left[(1-\rho)L + e^{-zm}\sum_{k=2}^{\ell-1} e^{kh}\right]. \end{aligned} \tag{45}$$

Using $z \le 2B^2$, $h > 0$ and (43),

$$\frac{\rho U_j}{(1-2\rho)B_j} = \frac{\rho}{1-2\rho} e^{zm(1-\rho)}(e^{-h} + e^{-2h}) \le 4\rho_* = e^{-k_g} < 1. \tag{46}$$

.

Thus, $-\partial_m T_{O,j} > 0$, and also $-\partial_m T_{C,\ell} > 0$. Note that $K(T, w) > 0$, so by using the chain rule, all the factors are positive, and thus $a_i > 0$ for all $i \in \{q, p, u, v\}$.

Using $zm\rho \le \rho_*/(2e)$ and the finite age bounds, elementary odds terms and chain-rule factors have scalar logarithmic derivatives bounded by $C(1 + \|\vartheta\|)$ and table derivatives bounded by $C(1 + m)$. Positive sums preserve these bounds by weighted averaging; (46) controls the single difference. Finally,

$$|\partial \log K(T, w)| \le C|\partial w| + C(1 + w)|\partial \log T|, \tag{47}$$

since $T/(1+T)^2 \le 1/4$ and w is bounded below. Apply this to each row gradient, average its positive contributions, and use $m = c_0 qp \le C\|\vartheta\|^2$ to obtain (44). No small-odds assumption is needed. □

C.3 Adam Moments through Training iterations

Let $0 \le \beta_1 < 1$, $\beta_1^2 < \beta_2 < 1$, and choose $\gamma > 0$ with $\beta_1 e^\gamma, \beta_2 e^{2\gamma} < 1$. Use a one-based update index $k \ge 1$ and initialize $M_0 = V_0 = 0$. The moments from Algorithm 2, with its indices shifted by one, and their interval bounds are

$$M_k = \beta_1 M_{k-1} + (1-\beta_1)G_k, \qquad V_k = \beta_2 V_{k-1} + (1-\beta_2)G_k^2, \tag{48}$$

$$\begin{aligned} M_- &= \frac{1-\beta_1}{1-\beta_1 e^{-\gamma}}, \quad M_+ = \frac{1-\beta_1}{1-\beta_1 e^{\gamma}}, \\ V_-^2 &= \frac{1-\beta_2}{1-\beta_2 e^{-2\gamma}}, \quad V_+^2 = \frac{1-\beta_2}{1-\beta_2 e^{2\gamma}}. \end{aligned} \tag{49}$$

With this indexing, bias correction gives $\widehat{M}_k = M_k/(1-\beta_1^k)$, $\widehat{V}_k = V_k/(1-\beta_2^k)$. For $k \ge 1$, define $a_k = -G_k$. The following lemma allows us to bound ratios of moments and gradients if the gradients are negative.

Lemma 10. *Let $k \geq 2$, assume that $a_{k-1}, a_k > 0$, and $|\log(a_k/a_{k-1})| \leq \gamma$. If*

$$-\frac{M_{k-1}}{a_{k-1}} \in [M_-, M_+], \frac{V_{k-1}}{a_{k-1}^2} \in [V_-^2, V_+^2],$$

then

$$-\frac{M_k}{a_k} \in [M_-, M_+], \qquad \frac{V_k}{a_k^2} \in [V_-^2, V_+^2]. \tag{50}$$

When $\beta_1 = 0$, the first-moment interval is the singleton $\{1\}$, and its equality holds exactly.

Proof. Omitted. Direct algebra check. □

For $0 \leq \epsilon_k/a_k \leq z_\epsilon$, define $d_k = -\widehat{M}_k/(\sqrt{\widehat{V}_k} + \epsilon_k)$ as the increasing update direction, and the common bias-correction factor by

$$f_k = \frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k}. \tag{51}$$

If M_k, V_k satisfy the conclusion of Lemma 10,

$$f_k \frac{M_-}{V_+ + z_\epsilon} \leq d_k \leq f_k \frac{M_+}{V_-} \tag{52}$$

which will be used to analyze the affine binding boundary (Appendix E).

Lemma 11 (Tracking a slowly varying gradient)**.** *Suppose $G_k = -a_k < 0$ eventually, $a_k > 0$, $|\log(a_k/a_{k-1})| \to 0$, and $\epsilon_k/a_k \to 0$. Then the globally bias-corrected moments satisfy*

$$\frac{-\widehat{M}_k}{a_k} \longrightarrow 1, \qquad \frac{\widehat{V}_k}{a_k^2} \longrightarrow 1, \qquad \frac{-\widehat{M}_k}{\sqrt{\widehat{V}_k} + \epsilon_k} \longrightarrow 1. \tag{53}$$

Consequently, for positive rates $\alpha\eta_k$ with $S_t = \sum_{k \leq t} \eta_k \to \infty$, the corresponding coordinate satisfies $y_t/S_t \to \alpha$.

Proof. Choose N so that $a_k > 0$ for $k \geq N$ and $|\log(a_k/a_{k-1})| \leq \gamma$ for $k > N$. Let $X_{1,k} = -M_k$ and $X_{2,k} = V_k$. For $r \in \{1, 2\}$ and $k > N$,

$$\frac{X_{r,k}}{a_k^r} = \frac{\beta_r^{k-N} X_{r,N}}{a_k^r} + (1-\beta_r) \sum_{j=0}^{k-N-1} \beta_r^j \left(\frac{a_{k-j}}{a_k}\right)^r.$$

For fixed j, $a_{k-j}/a_k \to 1$, and this ratio is at most $e^{\gamma j}$ when $0 \leq j \leq k-N$. Hence

$$\left|\frac{\beta_r^{k-N} X_{r,N}}{a_k^r}\right| \leq \frac{|X_{r,N}|}{a_N^r} \beta_r^{k-N} e^{r\gamma(k-N)} = \frac{|X_{r,N}|}{a_N^r} (\beta_r e^{r\gamma})^{k-N} \to 0.$$

The summands are dominated by the summable geometric sequence $(\beta_r e^{r\gamma})^j$. Dominated convergence therefore gives

$$(1-\beta_r) \sum_{j=0}^{k-N-1} \beta_r^j \left(\frac{a_{k-j}}{a_k}\right)^r \longrightarrow (1-\beta_r) \sum_{j=0}^{\infty} \beta_r^j = 1.$$

Thus $-M_k/a_k \to 1$ and $V_k/a_k^2 \to 1$. Lastly, the bias correction scale $\frac{1}{1-\beta_1^k}, \frac{1}{1-\beta_2^k}$ tends to 1 as $k \to \infty$, we obtain

$$\frac{-\widehat{M}_k}{a_k} \to 1, \quad \frac{\widehat{V}_k}{a_k^2} \to 1, \quad \frac{-\widehat{M}_k}{\sqrt{\widehat{V}_k} + \epsilon_k} = \frac{-\widehat{M}_k/a_k}{\sqrt{\widehat{V}_k/a_k^2} + \epsilon_k/a_k} \to 1.$$

□

C.4 What a fixed scalar stabilizer rules out

The theorem uses a fixed table stabilizer and zero scalar stabilizers.

Proposition 12 (Nonpersistence of the full linear speed law)**.** *Let Adam have fixed positive scalar multipliers and $S_t = \sum_{s<t} \eta_s \to \infty$. Suppose $\epsilon_{w,t} \geq \epsilon_* > 0$ and, eventually, all raw table rows are bounded and all matching gaps have a positive lower bound. Then the following limits cannot all hold:*

$$\frac{(q_t, p_t, u_t, v_t, h_t, w_t)}{S_t} \longrightarrow (\alpha_q, \alpha_p, \alpha_u, \alpha_v, \alpha_x, \alpha_w) > 0 \tag{54}$$

Proof. If these limits held, then $m_t \asymp S_t^2$, $\rho_t \leq Ce^{-cS_t^2}$, $h_t, w_t \asymp S_t$. The gap bounds and odds (Lemma 7) imply vanishing training odds, hence $y\widehat{v}_t \to 1$. The finite decoder-gradient sum (38) and its bias-corrected moments therefore tend to zero. Thus

$$\left| \frac{-\widehat{M}_{w,t}}{\sqrt{\widehat{V}_{w,t}} + \epsilon_{w,t}} \right| \leq \frac{|\widehat{M}_{w,t}|}{\epsilon_*} \longrightarrow 0, \quad \Longrightarrow \quad \frac{w_t}{S_t} \longrightarrow 0$$

by positive weighted averaging, contradicting $\alpha_w > 0$. □

This excludes the displayed speed law under a fixed positive scalar stabilizer. It does not exclude a slower unbounded retrieval range.

D Proof of Theorem 2(a)

The main motivation for gradient descent analysis is by exploiting strong geometry of orthogonal initialization. Gram identities (Lemma 14) obtained from symmetry of initialization control growth rate of matching gaps, table row norms, and derivatives of trainable parameters. Then, we define an invariant region $\mathcal{R}$ (70) that gives us the desired asymptotics on matching, binding, and recency, such that once GD steps into this region, it remains inside there (Lemma 16). We then finish the proof by showing that there exists a sufficient choice of the learning rate multipliers that can guide GD to guarantee hitting this invariant region (Lemma 18). Section D.5 details how to construct the multiplier hierarchy that works as well for Adam analysis.

D.1 Initialization and rates

Let $n \geq 2$, $d \geq 2n$, $\mathsf{J} = \mathbf{1}\mathbf{1}^\top/n$, and $\mathsf{P} = I_n - \mathsf{J}$. Assume

$$Q_0 Q_0^\top = K_0 K_0^\top = s_0 I_n, \qquad Q_0 K_0^\top = 0, \qquad s_0 > 0. \tag{55}$$

Thus all initial matching gaps vanish. GD below applies to every orientation of this frame, including the shared Haar initialization of Theorem 2. Retain the finite population of Section 2: $R \geq 2$, $J_{\max} < \infty$, $q_1 > 0$, uniform ordered distinct key pairs, uniform recall lags, and exact value-complement balance, with

$$\kappa > \tfrac{1}{2}, \qquad \omega_C, \omega_O > 0, \qquad \kappa + \omega_C + \omega_O = 1, \qquad \frac{\omega_C}{R-1} > \frac{16}{27} e^D \omega_O.$$

All updates minimize $F_{R,q}$. Initialize $q_0 = p_0 > 0$, $u_0 = v_0 > 0$, $w_0 = 0$, and $x_0 \in \mathbb{R}$, and use fixed positive multipliers:

$$\begin{aligned} \eta_t &= \frac{\eta^*}{t+1}, & S_t &= \sum_{s<t} \eta_s, \\ \eta_{Q,t} &= \eta_{K,t} = \lambda \eta_t, & \eta_{q,t} &= \eta_{p,t} = \alpha_m \eta_t, \\ \eta_{u,t} &= \eta_{v,t} = \alpha_g \eta_t, & \eta_{x,t} &= \alpha_x \eta_t, \qquad \eta_{w,t} = \alpha_w \eta_t. \end{aligned} \tag{56}$$

Every rate sum diverges, and $q_t = p_t$, $u_t = v_t$ persist.

We prove here a stronger version of Theorem 2(a), by not only controls the asymptotics growth of matching-versus-recency gap, but also the scaled leakage ρ, gate g, and predictor w. Using this and Proposition 6, we conclude the retrieval range for GD.

Theorem 13. *Under the preceding initialization and population assumptions, there exist fixed positive multipliers in* (56) *and* $\eta^*_{\max} > 0$ *such that every* $\eta^* \in (0, \eta^*_{\max}]$ *gives*

$$M_t - (R+1)h_t - \frac{3}{2}\log M_t \longrightarrow \log \frac{\mathfrak{a}_\infty \omega_C}{2(R+1)(R-1)\alpha_x \omega_O q_1},$$
$$\mathfrak{a}_\infty = 4\sqrt{\frac{c_0\alpha_m\lambda}{n-1}}, \qquad \frac{M_t}{h_t} \longrightarrow R+1, \qquad M_t^{3/2}\rho_t \longrightarrow 0. \tag{57}$$

Moreover $M_t, h_t, g_t, w_t \to \infty$. *The uniform retrieval guarantee holds through record lag* $R+2$, *including arbitrary older prefixes. For each fixed lag* $r \geq R+3$, *the explicit sequence constructed below has correct-answer probability tending to zero.*

The table norms grow in this theorem. The proof therefore tracks their geometry instead of requiring finite table displacement. We establish the exact discrete geometry first; the subsequent entrance argument explains why the theorem does not assume an acquired matching representation.

D.2 Gram Dynamics and The Effective Matching Rate

By symmetricity, at every training iteration k (Algorithm 1), every distinct key pair has the same gap $\Delta_{ab,k} = b_k$. Within the symmetric Gram family, denote

$$d_{\mathrm{match},k} := -\phi'_{\vartheta_k}(b_k), \qquad e_k := \frac{\eta_k \lambda d_{\mathrm{match},k}}{n-1}, \qquad \eta_k = \frac{\eta^*}{k+1},$$

where $\vartheta_k = (q_k, p_k, u_k, v_k, x_k, w_k)$ and ϕ_ϑ the per-pair loss in (39). The next lemma reduces the table geometry to its centered scale A_k and common matching gap b_k, then controls their relation to m_k.

Lemma 14 (Discrete Gram identities)**.** *Starting from* (55)*, exact GD preserves, at every iteration* k*,*

$$Q_k Q_k^\top = K_k K_k^\top = A_k \mathsf{P} + s_0 J, \qquad Q_k K_k^\top = b_k \mathsf{P}. \tag{58}$$

Here $A_0 = s_0$, $b_0 = 0$, *and* $A_k = \|\mathsf{P}Q_k\|_F^2/(n-1) = \|\mathsf{P}K_k\|_F^2/(n-1)$. *At steps with* $m_k = c_0 q_k p_k = c_0 q_k^2 > 0$, *the exact recurrences are*

$$\begin{aligned} A_{k+1} &= (1+e_k^2)A_k + 2e_k b_k, \\ b_{k+1} &= (1+e_k^2)b_k + 2e_k A_k, \\ q_{k+1} &= q_k + \eta_k \alpha_m \frac{b_k d_{\mathrm{match},k}}{q_k}, \\ m_{k+1} &= m_k + 2\eta_k c_0 \alpha_m b_k d_{\mathrm{match},k} + \eta_k^2 c_0^2 \alpha_m^2 \frac{b_k^2 d_{\mathrm{match},k}^2}{m_k}. \end{aligned} \tag{59}$$

Define the centered Gram determinant and scalar–table imbalance by

$$\mathcal{D}_k := A_k^2 - b_k^2, \qquad I_k := m_k - \chi A_k, \qquad \chi := \frac{c_0\alpha_m(n-1)}{\lambda}.$$

For $A_k > 0$, *these quantities satisfy*

$$\begin{aligned} \mathcal{D}_{k+1} &= (1-e_k^2)^2 \mathcal{D}_k, \\ I_{k+1} &= (1-r_k)I_k - r_k \chi \frac{\mathcal{D}_k}{A_k}, \qquad r_k := \frac{\chi e_k^2 A_k}{m_k} = \eta_k^2 c_0^2 \alpha_m^2 d_{\mathrm{match},k}^2 \frac{A_k}{\chi m_k}. \end{aligned} \tag{60}$$

If $m_k > 0$, $|e_k| \leq 1/2$, *and* $r_k \leq 1$ *at every step, then*

$$0 < \mathcal{D}_k \leq s_0^2, \qquad \min\{I_0, -\chi s_0\} \leq I_k \leq \max\{I_0, 0\}. \tag{61}$$

Proof. We will proceed to prove this by induction on k. Note that (58) holds at $k = 0$. Since every ordered pair gap $(H_k)_{aa} - (H_k)_{ab} = b_k$ for $a \neq b$, by Lemma 8(c), we have $G_{H,k} := (\nabla_H F_{R,\mathbf{q}})|_k = -\frac{d_{\mathrm{match},k}}{n-1}\mathsf{P}$. The matrix chain rule gives $(\nabla_Q F_{R,q})|_k = G_{H,k}K_k$ and $(\nabla_K F_{R,q})|_k = G_{H,k}^\top Q_k$. Hence

$$Q_{k+1} = Q_k + e_k \mathsf{P} K_k, \qquad K_{k+1} = K_k + e_k \mathsf{P} Q_k.$$

Using $\mathsf{P}^2 = \mathsf{P}, \mathsf{P}\mathsf{J} = 0$, we have

$$\begin{aligned} Q_{k+1}Q_{k+1}^\top &= Q_kQ_k^\top + e_kQ_kK_k^\top\mathsf{P} + e_k\mathsf{P}K_kQ_k^\top + e_k^2\mathsf{P}K_kK_k^\top\mathsf{P} \\ &= \big[(1+e_k^2)A_k + 2e_kb_k\big]\mathsf{P} + s_0\mathsf{J}, \\ Q_{k+1}K_{k+1}^\top &= Q_kK_k^\top + e_kQ_kQ_k^\top\mathsf{P} + e_k\mathsf{P}K_kK_k^\top + e_k^2\mathsf{P}K_kQ_k^\top\mathsf{P} \\ &= \big[(1+e_k^2)b_k + 2e_kA_k\big]\mathsf{P}. \end{aligned}$$

This implies the induction for (58) with $A_{k+1} = (1+e_k^2)A_k + 2e_kb_k$ and $b_{k+1} = (1+e_k^2)b_k + 2e_kA_k$, proving the first two recursions of (59).

Differentiate with respect to the separate parameters q, p before using $q_k = p_k$. The loss depends on (q, p, b) through $M = c_0qpb$, so

$$\partial_q F|_k = \frac{b_k}{q_k}\phi'_{\vartheta_k}(b_k) = -\frac{b_kd_{\mathrm{match},k}}{q_k}. \tag{62}$$

Since the p update is identical $q_{k+1} = p_{k+1}$. Using the learning rates in (56), we have $q_{k+1} = q_k + \eta_k\alpha_m\frac{b_kd_{\mathrm{match},k}}{q_k}$. Therefore,

$$m_{k+1} = c_0q_{k+1}^2 = m_k + 2\eta_kc_0\alpha_mb_kd_{\mathrm{match},k} + \eta_k^2c_0^2\alpha_m^2\frac{b_k^2d_{\mathrm{match},k}^2}{m_k},$$

implying the other two recursions in (59). Using the first two recursions, we have $A_{k+1} \pm b_{k+1} = (1 \pm e_k)^2(A_k \pm b_k)$. This yields $D_{k+1} = (1-e_k^2)^2D_k$. Since $\mathcal{D}_0 = s_0^2$ and $|e_k| \le 1/2$, induction gives $A_k > |b_k|$ and $0 < \mathcal{D}_k \le s_0^2$ at every finite iteration. Using the recursions in (59),

$$\begin{aligned} I_{k+1} - I_k &= \frac{\chi e_k^2}{m_k}(\chi b_k^2 - m_kA_k) \\ &= -\frac{\chi e_k^2}{m_k}(A_kI_k + \chi\mathcal{D}_k) = -r_kI_k - r_k\chi\frac{\mathcal{D}_k}{A_k}. \end{aligned}$$

This implies (60). Lastly, since

$$0 < \frac{\mathcal{D}_k}{A_k} \le \sqrt{\mathcal{D}_k} \le s_0, \qquad 0 \le r_k \le 1,$$

I_{k+1} is a convex combination of I_k and a point in $[-\chi s_0, 0]$. Induction proves the second bound in (61).

For linear interpolation of the original parameters between iterations k and $k+1$, keep the gradients evaluated at k and replace η_k by $\xi\eta_k$, $0 \le \xi \le 1$. Then $e_k \mapsto \xi e_k$ and $r_k \mapsto \xi^2 r_k$, so the same identities and bounds hold along the entire segment. □

The coefficient of $-F_M$ in the first-order matching field is

$$\mathfrak{a}_k = 2c_0\alpha_mm_kb_k^2 + \frac{2\lambda}{n-1}m_k^2A_k. \tag{63}$$

Uniformly under (61), $m_k, b_k > 0$ and $M_k = m_kb_k \to \infty$ imply

$$b_k \sim A_k \sim \sqrt{M_k/\chi}, \qquad m_k \sim \sqrt{\chi M_k}, \qquad \frac{\mathfrak{a}_k}{M_k^{3/2}} \longrightarrow \mathfrak{a}_\infty. \tag{64}$$

Indeed $m_k = \chi A_k + I_k$ and $A_k^2 - b_k^2 = \mathcal{D}_k$ have bounded remainders; no positive limiting lower bound on $\mathcal{D}$ is needed. For continuous flow, the corresponding invariants are exact:

$$\mathcal{D} = s_0^2, \qquad I = m_0 - \chi s_0, \qquad M = b\left[m_0 + \chi\left(\sqrt{s_0^2 + b^2} - s_0\right)\right]. \tag{65}$$

The last expression is strictly increasing for $b \ge 0$; thus (63) defines a smooth flow mobility $\mathfrak{a}(M) > 0$, including $\mathfrak{a}(0) > 0$, with $\mathfrak{a}(M) \sim \mathfrak{a}_\infty M^{3/2}$. We call $\mathfrak{a}_k$ the *matching mobility*, measuring how quickly matching grows relative to recency.

D.3 AN INVARIANT REGION AND THE RETRIEVAL BOUNDARY

We next control the interaction between effective matching, recency, and binding. Recall the loss $B_j, U_j, E_j, T_{O,j}$, and $T_{C,\ell}$ in Lemma 7, which depends on matching only through $M = mb$. Denote $K_{O,j} = K(T_{O,j}, w)$ and $K_{C,\ell} = K(T_{C,\ell}, w)$ be their loss coefficients. Define the sums

$$\begin{aligned} Z &= \omega_O \sum_j q_j K_{O,j}(B_j + E_j) + \frac{\omega_C}{R-1} \sum_{\ell=2}^{R} K_{C,\ell} T_{C,\ell}, \\ Y &= \omega_O \sum_j q_j K_{O,j}(U_j + 2B_j + E_j) + \frac{\omega_C}{R-1} \sum_{\ell=2}^{R} K_{C,\ell} L. \end{aligned} \tag{66}$$

These quantities control the matching, leakage, and recency derivative,

$$\begin{aligned} -F_M &= Z - \rho Y, \qquad F_\rho = MY, \\ F_h &= \omega_O \sum_j q_j K_{O,j} \partial_h T_{O,j} + \frac{\omega_C}{R-1} \sum_{\ell=2}^{R} K_{C,\ell} \partial_h T_{C,\ell}. \end{aligned} \tag{67}$$

We are interested in invariant properties of matching, recency, and binding over training time. To study this, we analyze the trajectory evolution over cumulative training steps $S_t = \sum_{s<t} \eta_s$. To differentiate the GD trajectory, let θ_t be its original parameters, and define the continuous interpolation

$$\theta(S) = \theta_t + \frac{S - S_t}{\eta_t}(\theta_{t+1} - \theta_t), \quad S \in [S_t, S_{t+1}].$$

Thus, $\theta'(S) = (\theta_{t+1} - \theta_t)/\eta_t$ inside each interval. This gives a continuous interpolation of the discrete GD trajectory. The prime notation means derivative in this continuous variable S, and at grid points use right derivatives. We use similar ideas from (Elkabetz & Cohen, 2021), (Lyu & Li, 2020) to analyze the errors bound between gradient flow and discrete GD, as well as asymptotics bounds on derivatives.

For the corresponding gradient-flow field under the multipliers in (56),

$$q' = -\alpha_m F_q, \quad p' = -\alpha_m F_p, \quad u' = -\alpha_g F_u, \quad v' = -\alpha_g F_v, \quad x' = -\alpha_x F_x, \quad w' = -\alpha_w F_w.$$

The following identities hold for this field and for right derivatives at GD grid points; between grid points, the residual bounds below control the difference. These identities will be used in analyzing invariance in Lemma 16.

Lemma 15. *We have*

$$\begin{aligned} M' &= \mathfrak{a}_t(Z - \rho Y), & w' &= -\alpha_w F_w, \\ h' &= -\alpha_x(1 - e^{-h})^2 F_h, & & \\ g' &= 4\alpha_g \,\mathrm{softplus}^{-1}(g)(1 - e^{-g})^2 \rho(1-\rho) MY. & & \end{aligned} \tag{68}$$

Proof. Recall that $M = mb, m = c_0 qp$, and from (62), $F_q = c_0 pbF_M, F_p = c_0 qbF_M$. By (59), $b_{t+1} - b_t = 2e_t A_t + e_t^2 b_t$, where $e_t = \frac{\eta_t \lambda d_{\text{match},t}}{n-1}$. Since $d_{\text{match}} = -mF_M$, the corresponding original-parameter field gives $b' = -2\lambda m A F_M/(n-1)$. By using chain rule, we have $m' = c_0(pq' + qp') = -2c_0\alpha_m mbF_M$, and $M' = bm' + mb' = -\mathfrak{a}_t F_M = \mathfrak{a}_t(Z - \rho Y)$.

For recency h, we have $h' = \sigma(x)x' = -\alpha_x \sigma(x)^2 F_h = -\alpha_x(1 - e^{-h})^2 F_h$.

Finally, for the gate g, we have $g' = \sigma(uv)(vu' + uv') = 2\alpha_g(u^2 + v^2)\sigma(uv)^2 \rho(1-\rho)F_\rho$, and conclusion follows by $u^2 + v^2 = 2\mathrm{softplus}^{-1}(g), \sigma(uv) = 1 - e^{-g}$. This concludes 15. □

From (64), matching mobility grows in an order of $M^{3/2}$. Accordingly define the residual matching advantage f, the amplified leakage Ψ, and the gradient scale

$$f = M - (R+1)h - \tfrac{3}{2}\log M, \qquad \Psi = M^{3/2}\rho, \qquad \mathcal{K} = 2we^{-w-2h}. \tag{69}$$

To control every subsequent update, define the region

$$\mathcal{R} = \{M \geq M_0,\ h \geq H,\ w \geq W,\ f \geq f_-,\ 0 < \Psi \leq \varepsilon,\ we^{-2h} \leq \bar{\eta}\}. \tag{70}$$

We show in the next lemma that there exists a choice of constants $(M_0, H, W, f_-, \varepsilon, \bar{\eta})$ such that $\mathcal{R}$ is *invariant*, in a sense that for every new iteration, the matching, recency, and binding remains inside $\mathcal{R}$ once they enter $\mathcal{R}$.

Lemma 16 (Continuation and its stable perturbations)**.** *Assume* (58)*,* (61)*,* $q = p > 0$*, and* $u = v > 0$*. For fixed positive multipliers, the constants in* (70) *can be chosen so that every balanced flow, or parameter interpolation satisfying* (75)*, entering a fixed compact subset of* $\mathcal{R}^\circ$ *remains in* $\mathcal{R}$ *and satisfies* (57)*. Gradient descent with orthogonal initialization* (55) *satisfies this conclusion under the specified learning rate schedule* (56) *whenever its amplitude* η^* *is sufficiently small.*

Proof. The proof of this is accomplished in three parts. Firstly, for gradient flow, we show that there exists boundary constants $(M_0, H, W, f_-, \varepsilon, \bar{\eta})$ such that derivatives of M, h, w, f, Ψ, and we^{-2h} pointing inwards into the region $\mathcal{R}$. Hence, when the gradient flow steps into this invariant region $\mathcal{R}$, it remains inside and retains the relevant asymptotics properties. Secondly, we control the error terms between the interpolated gradient flow and the discrete GD to make sure the results hold also for discrete GD. Lastly, we derive the limits and retrieval boundary once discrete GD stays in the region.

1. Inward boundary derivatives. Throughout, $c, C > 0$ are uniform in time and the state in $\mathcal{R}$; they may vary between estimates and depend only on fixed problem data, rates, (61), and region thresholds. On $\mathcal{R}$, the identities

$$e^{-M+(R-1)h} = \frac{e^{-f-2h}}{M^{3/2}}, \qquad M\rho = \frac{\Psi}{\sqrt{M}} \leq \frac{\varepsilon}{\sqrt{M_0}}$$

and $\sum_{a=0}^{N} e^{-ah} \leq (1 - e^{-H})^{-1}$ give

$$T_{C,\ell} \leq C\frac{e^{-2h}}{M^{3/2}}, \qquad T_{O,j} \leq Ce^{-2h}.$$

for $C = (2 + M_0^{-3/2})e^{2\varepsilon/\sqrt{M_0}+\max\{0,-f_-\}}/(1 - e^{-H})$. Note that $\frac{K(T,w)}{2we^{-w}} = \frac{\exp\left(\frac{2wT}{1+T}\right)}{(1+T)^2(1+e^{-ws(T)})}$. Hence, for large enough H and small enough $\bar{\eta}$, $T \leq \frac{1}{2}$, $s(T) > 0$, and $wT \leq C\bar{\eta}$. In particular, $\frac{2}{9} \leq \frac{K(T,w)}{2we^{-w}} \leq e^{2C\bar{\eta}}$, implying that $K_{C,\ell}, K_{O,j} \asymp 2we^{-w}$.

The longest recall sequence of lag R contributes a term comparable to $\frac{\omega_C}{R-1}2we^{-w}e^{-M+(R-1)h}$, and this is the largest entry in Z. Hence, $Z \asymp \frac{\omega_C}{R-1}2we^{-w}e^{-M+(R-1)h} \asymp \mathcal{K}e^{-f}M^{-3/2}$. The $j = 1$ overwrite contribution gives $Y \asymp \mathcal{K}$ and $-F_h \asymp \mathcal{K}$ on the region. Their exact asymptotic ratios are established below after proving $M, h, w \to \infty$ and $we^{-2h} \to 0$. Combining this with Lemma 15, we obtain the asymptotics

$$w' \asymp \alpha_w e^{-w}, \qquad h' \asymp \alpha_x \mathcal{K}, \qquad Y \asymp \mathcal{K}, \qquad |M'| \leq C\mathcal{K}. \tag{71}$$

The chain rule gives

$$f' = \left(1 - \frac{3}{2M}\right)M' - (R+1)h', \qquad (\log \Psi)' = \frac{3M'}{2M} - 2(1-\rho)g'. \tag{72}$$

The leakage bound $\Psi \leq \varepsilon$ implies that leakage ρ is small when $M \geq M_0$ is big enough. Let M_0 sufficiently big so that $g \geq 1$, so $\mathrm{softplus}^{-1}(g)(1 - e^{-g})^2 \asymp g$. Hence,

$$\frac{f'}{\mathcal{K}} \geq ce^{-f} - C\Psi - C\alpha_x, \qquad \frac{(\log \Psi)'}{\mathcal{K}} \leq \frac{C}{M} - \frac{c\alpha_g g\Psi}{\sqrt{M}}. \tag{73}$$

We can choose f_- sufficiently small and M_0 sufficiently large, so that $ce^{-f_-} \geq 2C(\varepsilon + \alpha_x)$ and $\sqrt{M_0} \geq \frac{2C}{c\alpha_g \varepsilon}$. Then $f' > 0$ on the face $f = f_-$ and $\Psi' < 0$ on the face $\Psi = \varepsilon$ by (73). On the other hand, (71) implies $h', w' > 0$. On the boundary face $we^{-2h} = \bar{\eta}$,

$$(we^{-2h})' = e^{-2h}(w' - 2wh') \leq e^{-w-2h}(C\alpha_w - c\alpha_x w^2 e^{-2h}) < 0, \tag{74}$$

whenever $W\bar{\eta} > \frac{C}{c}\frac{\alpha_w}{\alpha_x}$, which can be ensured by increasing W or decreasing α_w/α_x while keeping $\bar{\eta}$ small. Finally, we can choose $H > \frac{M_0 - \frac{3}{2}\log M_0 - f_-}{R+1}$, so that every point with $M = M_0$ and $h \geq H$ has $f < f_-$. Hence, $M > M_0$ in this region, and due to continuity, any point inside this regions can only have point heading inwards the region, and never cross M_0, avoids needing to prove $M' > 0$ at that boundary.

2. Discrete errors and continuation

For $S \in (S_i, S_{i+1})$, primes denote actual derivatives of the GD interpolation. Evaluate every unprimed quantity below at $\theta(S)$, define the errors

$$r_h := h' + \alpha_x(1 - e^{-h})^2 F_h, \qquad r_w := w' + \alpha_w F_w,$$
$$r_f := f' - \left(1 - \frac{3}{2M}\right)\mathfrak{a}_t(Z - \rho Y) - (R+1)\alpha_x(1 - e^{-h})^2 F_h,$$
$$r_\Psi := \Psi' - \Psi\left[\frac{3\mathfrak{a}_t}{2M}(Z - \rho Y) - 8\alpha_g \operatorname{softplus}^{-1}(g)(1 - e^{-g})^2 M\rho(1-\rho)^2 Y\right],$$

in a sense that for sufficiently small δ_0, h', f', Ψ', w', g' retains the directions pointing inwards to the region $\mathcal{R}$.

Every residual is the actual derivative minus the gradient-flow field at the same state, using Lemma 15 and (72). They vanish for gradient flow and for the right derivatives at GD grid point. The strict boundary signs persist for sufficiently small $\delta_0 > 0$ when the normalized residuals satisfy

$$\max\left\{\frac{|r_f| + |r_h|}{\mathcal{K}}, \quad e^w|r_w|, \quad \frac{\sqrt{M}|r_\Psi|}{\Psi\mathcal{K}}\right\} \leq \delta_0 e^{-S/2}. \tag{75}$$

Now, (71) and (75) gives $|M'| \leq Ch', h' \leq Cwe^{-w}$. Thus, by taking derivative and integration, we have

$$e^{w(S)} \asymp 1 + S, \qquad M(S) + e^{g(S)/2} \leq C[1 + h(S)] \leq C[1 + \log^2(2+S)]. \tag{76}$$

For the gate bound g, the induced error in g' is $O(\delta_0 e^{-S/2}\mathcal{K}/\sqrt{M})$. Hence, $(e^g)' \leq cM\mathcal{K} + C\delta_0 e^{-S/2}\mathcal{K}e^g$. Since $M \leq C(1+h), \mathcal{K} \leq Ch'$, and $\int_0^S e^{-s/2}\mathcal{K}(s)ds \leq C$, by Grönwall's inequality, we have $e^g \leq C(1+h^2)$. Note that from (61), we can bound $A, b, m \leq C(1+\sqrt{M})$. We can then derive from the bound on M that

$$\max\left\{|q|, |p|, |u|, |v|, |x|, |w|, \max_i \|Q_i\|, \max_i \|K_i\|\right\} \leq C(1 + \log^2(2+S)). \tag{77}$$

Denote t_0 the first time discrete GD enters $\mathcal{R}$ and $j = t - t_0 + 1$. Note that $S_t - S_{t_0} \leq \eta^*(1 + \log j)$, so

$$\max\left\{|q|, |p|, |u|, |v|, |x|, |w|, \max_i \|Q_i\|, \max_i \|K_i\|\right\} \leq C(1 + \log^2(2 + 3 + \log j)).$$

Since the finite row formula, their derivatives, and the repriprocals of the scales in (76) involves finite sums, products, exponentials, and logistic functions of these parameters, they can be bounded by

$$\exp\big(C[1 + \log^2(3 + \log j)]^C\big) = j^{o(1)}. \tag{78}$$

Using chain rule, we can deduce that

$$\max\left\{\frac{|r_f| + |r_h|}{\mathcal{K}}, \; e^w|r_w|, \frac{\sqrt{M}|r_\Psi|}{\Psi\mathcal{K}}\right\} \leq C\eta_t j^{o(1)} \leq C\eta^* j^{-1/2} \leq C\eta^* e^{-S_t/2}.$$

Here the fixed entrance-clock factor is absorbed into C, which does not depend on the training step t. Hence, with sufficiently small η^*, $C\eta^* < \delta_0$, and (75) holds for any later t. Note that (71) and (61) also gives $|d_{\text{match}}| \leq C\mathcal{K}M$. Small η^* also ensures that $|e_t| \leq 1/2$, $r_t < 1$, and $|\eta_t \alpha_m c_0 b F_M| < 1/2$, preserving (61) and the positive multiplicative updates of q, p, and gradient of g preserve $u, v > 0$. The inward margins, growth bounds, and improved errors hence retains through every iteration $t > t_0$, as desired.

3. Limits and retrieval boundary. From (76), we have $w \asymp \log(1+S) \to \infty$. If h is bounded, (71) gives $h' \geq cwe^{-w}$, whose integral diverges. Thus $h \to \infty$. $f \geq f_-$ and $\Psi \leq \varepsilon$ then force $M, g \to \infty$. Equations (73) and (75) give

$$(\log \Psi)' \leq \frac{\mathcal{K}}{\sqrt{M}} \left[\frac{C}{\sqrt{M}} - c\alpha_g g \Psi + o(1) \right], \qquad \int_0^\infty \frac{\mathcal{K}}{\sqrt{M}}\, dS = \infty.. \tag{79}$$

Indeed, $M \leq C(1+h)$ and $\mathcal{K} \geq ch'$ bound this integral below by $c \int h'/\sqrt{1+h}\, dS = \infty$. Above any fixed positive threshold for Ψ, the bracket is eventually strictly negative, so $\Psi \to 0$. Also $dh/dw \geq cwe^{-2h}$ implies $e^{2h} \geq cw^2$, hence $we^{-2h} \to 0$.

Consequently $K_{C,\ell}, K_{O,j} \sim 2we^{-w}$. The longest recall and $j = 1$ overwrite dominate the finite odds sums, giving

$$\begin{aligned} \frac{M'}{\mathcal{K}} &= \frac{\mathfrak{a}_t}{M^{3/2}} \left[\frac{\omega_C}{R-1} e^{-f} - \omega_O q_1 \Psi \right] + o(1), \\ \frac{h'}{\mathcal{K}} &= 2\alpha_x \omega_O q_1 + o(1). \end{aligned} \tag{80}$$

The recall contribution to $h'/\mathcal{K}$ is $O(e^{-f} M^{-3/2}) = o(1)$. Using (64), (72), and $\Psi \to 0$ now yields

$$\frac{f'}{\mathcal{K}} = \frac{\mathfrak{a}_\infty \omega_C}{R-1} e^{-f} - 2(R+1)\alpha_x \omega_O q_1 + o(1), \qquad \int_0^\infty \mathcal{K}\, dS = \infty. \tag{81}$$

The integral diverges because $h' \asymp \mathcal{K}$ and $h \to \infty$. Outside any neighborhood of its unique equilibrium, the right-hand side has the strict sign pointing toward that equilibrium. Hence f converges to the constant in (57).

Moreover $e^w w' \to \alpha_w$, since $\kappa + \omega_C + \omega_O = 1$. Integrating this and (80) gives

$$\frac{e^{w_t}}{S_t} \longrightarrow \alpha_w, \qquad \frac{e^{2h_t}}{(\log S_t)^2} \longrightarrow \frac{4\alpha_x \omega_O q_1}{\alpha_w}, \qquad M_t \sim (R+1) \log\log S_t. \tag{82}$$

At lag $R+2$, the target's score advantage is $\frac{3}{2}\log M + O(1) \to \infty$. The maximum raw row norm B_t satisfies $B_t^2 m_t = O(M_t)$ by (58); hence $M_t \rho_t = \frac{\Psi}{\sqrt{M_t}} \leq \frac{\varepsilon}{\sqrt{M_t}} \to 0$ controls the binding error in Proposition 6. Its geometric prefix bound and $w_t \to \infty$ give uniform success through $R+2$.

For distinct keys a, b and fixed $r \geq R+3$, define

$$\begin{aligned} X_r &= ((a,+1), (b,-1)^{r-1}; ?a), \\ s_{\text{target}} - s_{\text{newest}} &= (R+2-r)h + \tfrac{3}{2}\log M + O(1) \longrightarrow -\infty. \end{aligned} \tag{83}$$

On X_r, every non-target value is -1; the diverging wrong-score advantage gives correct-answer probability tending to zero. □

D.4 Entering the invariant region under sufficient choice of fixed multipliers

The continuation lemma assumes an invariant region of the learned trajectory. We prove that some fixed multipliers choice bring the zero-score initialization into that region. The main idea is that we want to maintain the hierarchy control: we start with obtaining small matching signal M, then use that to control small leakage ρ and increases matching M, keeping everything else small. Once matching signal is sufficient, we grow recency h while keeping decoder w small, and finally grows w to be sufficiently large.

Consider the invariant region $\mathcal{R}(M_0, H, W, f_-, \varepsilon, \bar{\eta})$ defined by Lemma 16. With the matching mobility $\mathfrak{a}(M)$ defined by (63) and (65), define

$$f_a = M - (R+1)h - \log \mathfrak{a}(M), \qquad \Psi_a = \mathfrak{a}(M)\rho. \tag{84}$$

Recall h_0 the initial recency strength $h_0 = \text{softplus}(x_0) > 0$.

Lemma 17 (Invariant small-decoder region). *Fix $W_+ > W, 0 < h_{\min} < h_0$. There exists constants F and ε_a, a choice of multiplier α_x sufficiently small, and α_g sufficiently large, such that the region*

$$\mathcal{R}_a = \{h \geq h_{\min},\ 0 < w \leq W_+,\ f_a \geq F,\ \Psi_a \leq \varepsilon_a\} \tag{85}$$

is invariant for the symmetric gradient flow until w reaches W_+. Specifically, $h' > 0$ throughout the region, $f_a' > 0$ on the face $f_a = F$, and $(\log \Psi_a)' < 0$ on the face $\Psi_a = \varepsilon_a$.

Proof. Choose $M_{\min}$ large and $F > 0$ satisfying

$$F > 2 + \max_{0 \le M \le M_{\min}} \{M - \log \mathfrak{a}(M)\}. \tag{86}$$

Then $f_a \ge F$ and $h \ge h_{\min}$ imply $M > M_{\min}$. We have $M\rho = \frac{M}{\mathfrak{a}(M)}\Psi_a \le \varepsilon_a \sup_{r \ge M_{\min}} \frac{r}{\mathfrak{a}(r)}$. Since $\mathfrak{a}(r) \asymp r^{3/2}$, $\sup_{r \ge M_{\min}} \frac{r}{\mathfrak{a}(r)}$ is bounded, and thus with sufficiently small ε_a, $M\rho \le 1$. The exact recall formula gives

$$\begin{aligned} T_{C,\ell} &= e^{-M(1-\rho)+h} + e^{-M}\sum_{k=2}^{\ell-1} e^{kh} \\ &\le e^{-M+h+1} + e^{-M}\sum_{k=2}^{R-1} e^{kh} \\ &\le \frac{e}{1-e^{-h_{\min}}} e^{-M+(R-1)h}, \end{aligned}$$

and for overwrite formula,

$$\begin{aligned} T_{O,j} &= U_j + B_j + E_j \\ &\le 2e \cdot e^{-2h} + e^2 e^{-M}\sum_{k=1}^{j} e^{-kh} \\ &\le e^{-2h}\left(2e + \frac{e^2}{1-e^{-h_{\min}}}\right). \end{aligned}$$

Hence, $T_{C,\ell} \le Ce^{-M+(R-1)h}, T_{O,j} \le Ce^{-2h}$ for some time-independent constant $C > 0$. By Lemma 8(b), $K(T,w)/w$ extends positively and continuously to $w = 0$, and we have for $z_-, z_+, y_-, y_+, h_-, h_+ > 0$,

$$\begin{aligned} z_- we^{-M+(R-1)h} \le Z \le z_+ we^{-M+(R-1)h}, \\ y_- we^{-2h} \le Y \le y_+ we^{-2h}, \quad \alpha_x h_- we^{-2h} \le h' \le \alpha_x h_+ we^{-2h}. \end{aligned} \tag{87}$$

The last inequality is achieved by $h' = -\alpha_x(1-e^{-h})^2 F_h$, and since the shortest overwrite example contributes at least $2\omega_O q_1 K(T,w)e^{-2h}$ to $-F_h$, while the adverse recall contribution has magnitude at most $O\left(we^{-F}e^{-2h}/\mathfrak{a}(M)\right)$, we can choose F sufficiently large such that $-F_h \ge \Omega(we^{-2h})$.

The constants are uniform over $0 < \alpha_x \le 1$ and all gate rates. Choose

$$\varepsilon_a \le \frac{z_- e^{-F}}{4y_+}, \qquad \alpha_x \le \frac{z_- e^{-F}}{16(R+1)h_+}. \tag{88}$$

From Lemma 15,

$M' = \mathfrak{a}(M)(Z - \rho Y) \ge \mathfrak{a}(M) z_- we^{-M+(R-1)h} - \mathfrak{a}(M)\rho y_+ we^{-2h} = [z_- e^{-f_a} - y_+\Psi_a]we^{-2h}$.

At $f_a = F, \Psi_a \le \varepsilon_a$, we have

$$M' \ge [z_- e^{-F} - y_+\varepsilon_a]we^{-2h} \ge \frac{3}{4} z_- e^{-F} we^{-2h}.$$

Since $\mathfrak{a}'/\mathfrak{a} \le 1/2$ for large $M_{\min}$, this makes f_a' strictly positive. At $\Psi_a = \varepsilon_a$, differentiation gives

$$(\log \Psi_a)' = \mathfrak{a}'(M)(Z - \rho Y) - 8\alpha_g G(g)\frac{M}{\mathfrak{a}(M)}\Psi_a(1-\rho)^2 Y, \quad G(g) = u^2\,\mathrm{sigmoid}(u^2)^2. \tag{89}$$

The gate is increasing whenever $M > 0$, so $G(g)$ retains a positive lower bound. The positive term on is at most $\mathfrak{a}' z_+ e^{-F} we^{-2h}/\mathfrak{a}$, while the magnitude of the negative term is at least $c\alpha_g M\varepsilon_a we^{-2h}/\mathfrak{a}$ for a fixed $c > 0$. Since $\sup_{M \ge M_{\min}} \mathfrak{a}'(M)/M < \infty$, a sufficiently large fixed α_g makes $(\log \Psi_a)' < 0$. By (87), $h' > 0$. Thus this region is invariant until w reaches W_+. □

Lemma 18 (Acquisition without intervention)**.** *The fixed multipliers in* (56) *can be chosen so that the symmetric flow from* (55) *enters a fixed compact subset of $\mathcal{R}$ in finite cumulative time. For every sufficiently small single amplitude η^*, exact GD enters the same region under its unchanged learning rate schedule* (56)*.*

Proof. We construct a region valid while the decoder is still small. Inside it, recency can reach the level needed for continuation before the decoder grows. Fix $W_+ > W, 0 < h_{\min} < h_0$, then by Lemma 17, we can construct the invariant region $\mathcal{R}_a$ as in (85) with certain choices of α_x, α_g. Now we show that from initialization with matching $M = 0$, we can reach the auxiliary region $\mathcal{R}_a$. Introduce the analytical clock $\tau(S) = \int_0^S w(s)ds$. At $w = 0$, every gradient except the decoder gradient vanishes, and GD makes w positive at the first update. Indeed,

$$-F_w \geq \kappa\sigma(-w) - (1-\kappa)\sigma(w) = \kappa - \sigma(w).$$

Choose $0 < w_b < \log\frac{\kappa}{1-\kappa}$ sufficiently small so that $w_b \leq W_+$. Then, while $w \leq w_b$, $-F_w > 0$. In particular, $w'(0) > 0$, so $w(S) > 0$ for $S > 0$. Retain $h \in [h_{\min}, h_0 + 1]$ during the initial finite interval, and choose a finite target M_{tar} with

$$M_{\text{tar}} - (R+1)(h_0+1) - \log \mathfrak{a}(M_{\text{tar}}) > F + 2, \tag{90}$$

and choose a target leakage $\rho_{\text{tar}} > 0$ with $\mathfrak{a}(M_{\text{tar}})\rho_{\text{tar}} < \varepsilon_a/4$, small enough that the compact-region lower bound for $-F_M/w$ displayed below also holds. This choice precedes the gate multiplier choice.

By Lemma 8(b), choose a sufficiently small fixed $M_1 > 0$ so that, uniformly for $0 \leq M \leq M_1$, $h \in [h_{\min}, h_0 + 1]$, $0 \leq w \leq w_b$, and $0 \leq \rho \leq \rho_0$, the ratio d_{match}/w has positive upper and lower bounds, using its continuous extension at $w = 0$. Since $d_{\text{match}} = -mF_M$, we have $dM/d\tau = [\mathfrak{a}(M)/m(M)]d_{\text{match}}/w$. The ratio $\mathfrak{a}(M)/m(M)$ is also bounded above and below by positive constants on $[0, M_1]$. Thus there exist $v_-, v_+ > 0$ such that $v_-\tau \leq M(\tau) \leq v_+\tau$ while this initial interval is retained. Choose $\tau_0 > 0$ with $v_+\tau_0 < M_1$ so matching remains in this interval through clock τ_0.

Note that GD maintains $u = v$ always due to symmetry, we have $\rho = \sigma(D - 2\text{softplus}(u^2))$, so $\rho \leq \rho_{\text{tar}}$ iff $u \geq u_{\text{tar}}$ for some u_{tar}. Now, note that

$$\frac{du}{d\tau} = 2\alpha_g u\sigma(u^2)\rho(1-\rho)M\frac{Y}{w} \geq \alpha_g c_g M \geq \alpha_g c_g v_-\tau, \qquad c_g > 0. \tag{91}$$

The inequality is taken at the time before ρ reaches ρ_{tar}, which gives a deterministic lower bound for $u\sigma(u^2)\rho(1-\rho)\frac{Y}{w}$, noting that $u \in [u_0, u_{\text{tar}}]$ for $u_0 > 0$ and Y/w is positive and continuous on the fixed parameter range, including its limit at $w = 0$. Hence, $u(\tau) - u_0 \geq \frac{\alpha_g c_g v_-}{2}\tau^2$. Choose α_g that is larger than the constant in Lemma 17 and $\frac{\alpha_g c_g v_-}{2}\tau_0^2 > u_{\text{tar}} - u_0$. If the gate u has not reached u_{tar} by τ_0, this inequality would force $u(\tau_0) > u_{\text{tar}}$, a contradiction. Thus leakage reaches ρ_{tar} by τ_0, while $M(\tau) \leq v_+\tau_0$.

Compactness gives $c_M, \rho_{\text{tar}} > 0$ such that

$$-F_M/w \geq c_M \quad \text{on } [0, M_{\text{tar}}+1] \times [h_{\min}, h_0+1] \times [0, w_b] \times [0, \rho_{\text{tar}}], \qquad \mathfrak{a}(M_{\text{tar}})\rho_{\text{tar}} < \varepsilon_a/4.$$

This implies that after leakage is small, $dM(\tau)/d\tau \geq c_M \min_{0 \leq M \leq M_{\text{tar}}+1} \mathfrak{a}(M) > 0$. Matching therefore reaches M_{tar} within a finite amount of time. Note that, we need to retain h inside the interval $[h_{\min}, h_0 + 1]$ throughout the entire time, which we can reduce α_x further to control this. By reducing α_w as well, we can also ensure that $w < w_b$ throughout the acquisition. These sufficient choices of multipliers control the entire acquisition to be in the compact region specified above.

With the desired leakage and matching, we can now conclude that $f_a > F, \Psi_a \leq \frac{\varepsilon_a}{4}$, placing us into the invariant region $\mathcal{R}_a$.

Now, we control the recency h and decoder target w. Since $|F_w| \leq 1$, $d(w^2)/d\tau = -2\alpha_w F_w \leq 2\alpha_w$.Within $\mathcal{R}_a$, $dh/d\tau \geq \alpha_x h_- e^{-2h}$; any finite H_{tar} is therefore reached within additional clock

$$\tau_H \leq \frac{e^{2H_{\text{tar}}} - e^{2h_{\min}}}{2\alpha_x h_-}. \tag{92}$$

Choose fixed $\alpha_w > 0$ small enough to retain $w < w_b/2$ until then. Calibration bounds $-F_w$ positively there, so this finite τ horizon also has finite S. Choose H_{tar} so subsequent training odds are $< 1/4$ and $We^{-2H_{\text{tar}}} < \bar{\eta}$. Every signed pooled value is then $> 3/5$, giving $-F_w \geq c(W_+) > 0$ while $w \leq W_+$. The decoder reaches a value strictly above W in finite time, with $\mathcal{R}_a$ preserved throughout.

To avoid circular rate choices, first choose continuation thresholds uniformly over

$$0 < \alpha_x \le 1, \qquad \alpha_g \ge 1, \qquad 0 < \alpha_w \le c_{\mathrm{dec}}\alpha_x. \tag{93}$$

Use $\alpha_x = 1$ in the lower f' estimate and $\alpha_g = 1$ in the leakage estimate; take $W = 2$, choose c_{dec} small enough for (74), then choose H. Select the small-decoder targets next and impose (93) on the final α_w. Perturbation tolerances depend on these fixed rates; select η^* last. For $M \ge M_{\min}$, fix $a_- > 0$ with $\mathfrak{a}(M)/M^{3/2} \ge a_-$; then

$$f = f_a + \log\frac{\mathfrak{a}(M)}{M^{3/2}} \ge f_a + \log a_-, \qquad \Psi = \frac{\Psi_a}{\mathfrak{a}(M)/M^{3/2}} \le \frac{\Psi_a}{a_-}. \tag{94}$$

Impose $F > f_- - \log a_- + 2$ and $\varepsilon_a < a_-\varepsilon/4$. Increasing the recency and decoder targets as needed gives strict entrance into $\mathcal{R}$. □

D.5 Common multipliers for the optimizer comparison

The preceding construction must be compatible with Adam's finite comparison of pairwise inner products. The following order of constants makes that compatibility explicit.

Lemma 19 (Shared fixed multipliers)**.** *Lemma 16, 18 Theorem 20 admit the same fixed positive multipliers and the same sufficiently small base amplitude η^* for the same learning rate schedule* (56)*. With the common Haar orthogonal initialization satisfying* (55)*, every prescribed confidence is attained by Adam at sufficiently large finite width; GD holds for every frame orientation.*

Proof. Identify $\alpha_q = \alpha_p = \alpha_m$ and $\alpha_u = \alpha_v = \alpha_g$. Choose constants in the following order.

First fix small positive α_m, retaining a compact positive (q, p) range through Adam's first unit clock interval. Restrict $\alpha_x \le 1$ and α_w below a fixed small calibration bound. Adam's local raw-gap upper bound is uniform over these choices and the full leakage interval, so fix $\lambda > 0$ now.

This fixes the GD mobility. Select its uniform region constants, then α_x, α_w by Lemma 18. Its finite τ bound is uniform above the sufficient gate-rate threshold: increasing α_g reduces leakage, the only dependence of the other fields on u, v. Adam's early gap in (117) is now fixed and positive, proportional to $\lambda s_0 \alpha_w$ (with the table stabilizer absorbed in the constant). Lemma 23 therefore fixes its finite score-pressure integral and row ceiling before α_g. Choose α_g above both optimizer thresholds; the preceding gap, ceiling, and GD entrance bounds remain valid.

Next choose a common η^* satisfying GD's interpolation and Adam's relative-moment requirements. Finally choose finite $d \ge 2n$ large enough: a Haar frame has exact Grams (55) and maximum coordinate $O(\sqrt{\log(nd/\zeta)/d})$ with probability at least $1 - \zeta$. This gives Adam's finite-horizon coordinate bound and GD's exact symmetry under the same initialization, loss, and nominal rates. Adam retains its stated group-dependent stabilizers. □

Combining Lemmas 18 and 16 proves Theorem 13. Lemma 19 ensures the common-rate comparison in Theorem 2.

E Proof of Theorem 2(b)

The Adam conclusions of Theorems 2 and 20 require matching and binding to emerge under the same update rule used for subsequent training. We prove this by comparing their pairwise inner products with a reference trajectory that treats all keys equally, for a finite deterministic interval. Width controls this comparison even though the scalar updates need not have a continuous limit near zero gradient. Once the matching factors are large, the answer gradients of the tables have an integrable tail. This tail preserves the learned representation for all later updates without a change of rate.

E.1 Statement and optimizer conventions

The population must provide a matching signal before a representation has been learned. We use the training distribution $\mathcal{D}_{R,\mathbf{q}}$ and require its calibration, recall, and overwrite weights to satisfy

$$\kappa > \tfrac{1}{2}, \qquad \omega_C, \omega_O > 0, \qquad \kappa + \omega_C + \omega_O = 1, \qquad q_1 > 0, \qquad \frac{\omega_C}{R-1} > \frac{16}{27}e^D\omega_O. \tag{95}$$

The last inequality is the strict mean matching-pressure condition of Lemma 8. It permits arbitrary positive initial binding factors; it does not assume that the initial binding error is small. Throughout this appendix, $\vartheta = (q, p, u, v, x, w)$ denotes the six scalar parameters and $Z = (Q, K)$ denotes the two learned tables.

The table stabilizer prevents a small coordinate gradient from producing a large coordinate update during representation learning. Fix $E > 0$, $0 \le \beta_1 < 1$, and $\beta_1^2 < \beta_2 < 1$. Starting with zero moment buffers, process the full gradient $G_{i,t}$ at update t by

$$\mathsf{M}_{i,t+1} = \beta_1 \mathsf{M}_{i,t} + (1-\beta_1) G_{i,t}, \qquad \mathsf{V}_{i,t+1} = \beta_2 \mathsf{V}_{i,t} + (1-\beta_2) G_{i,t}^2, \tag{96}$$

$$\widehat{\mathsf{M}}_{i,t+1} = \frac{\mathsf{M}_{i,t+1}}{1-\beta_1^{t+1}}, \qquad \widehat{\mathsf{V}}_{i,t+1} = \frac{\mathsf{V}_{i,t+1}}{1-\beta_2^{t+1}}. \tag{97}$$

All products and divisions for table coordinates are entrywise. For a common step amplitude $\eta^* > 0$ and fixed positive learning-rate multipliers α_i, λ, the base rate, its cumulative clock, and the parameter rates are

$$\eta_t = \eta^* (t+1)^{-r}, \qquad \frac{2}{3} < r \le 1, \qquad S_t = \sum_{s<t} \eta_s, \qquad r_{i,t} = \alpha_i \eta_t, \qquad \nu_t = \lambda \eta_t. \tag{98}$$

Here $r_{i,t}$ is the rate of scalar i and ν_t is the rate of both tables. Scalar coordinates use zero stabilizer; table coordinates use E. A direction with zero numerator and denominator is defined to be zero. There is no weight decay, and neither the objective nor any optimizer parameter changes during the run.

We track $QQ^\top$, $KK^\top$, and $QK^\top$, whose entries are the query–query, key–key, and query–key inner products, respectively. Two initialization laws control their deviations from fixed initial matrices. For a fixed variance scale $s_0 > 0$, the Gaussian law has independent $N(0, s_0/d)$ entries in both tables. The orthogonal law draws a Haar row-orthogonal matrix $U \in \mathbb{R}^{2n \times d}$, with $d \ge 2n$, and sets the stacked matrix $(Q_0^\top, K_0^\top)^\top = \sqrt{s_0} U$. The latter law has zero initial content logits:

$$Q_0 Q_0^\top = K_0 K_0^\top = s_0 I_n, \qquad Q_0 K_0^\top = 0. \tag{99}$$

Both laws leave all $2nd$ table entries trainable.

Theorem 20. *Fix a finite population satisfying* (95)*, $r \in (2/3, 1]$, $E, s_0 > 0$, the moment parameters above, and initial scalars $q_0 = p_0 > 0$, $u_0 = v_0 > 0$, $x_0 \in \mathbb{R}$, and $w_0 = 0$. For every $\zeta \in (0,1)$ there exist fixed positive multipliers $\alpha_q = \alpha_p$, $\alpha_u = \alpha_v$, α_x, α_w, λ, a sufficiently small $\eta^* > 0$, and a finite width threshold $d_0 \ge 2n$ with the following property. For every $d \ge d_0$, either initialization law followed by* (97)–(98) *on the unchanged loss $F_{R,\mathbf{q}}$ satisfies, with probability at least $1 - \zeta$, the following conclusions.*

There are deterministic $T, B, \delta_* > 0$ such that, at every update with $S_t \ge T$, all raw row norms are at most B and $\Delta_{ab,t} \ge \delta_*$ for every $a \ne b$. The tables converge in Frobenius norm. The scalar speeds satisfy

$$\frac{(q_t, p_t, u_t, v_t, x_t, w_t)}{S_t} \longrightarrow (\alpha_q, \alpha_p, \alpha_u, \alpha_v, \alpha_x, \alpha_w). \tag{100}$$

Consequently, for a deterministic $c_* > 0$,

$$m_t \sim c_0 \alpha_q \alpha_p S_t^2, \qquad h_t \sim \alpha_x S_t, \qquad \mathcal{E}_{\theta_t}(\lfloor c_* S_t \rfloor) \longrightarrow 0. \tag{101}$$

Every nominal rate sum diverges, as does every table coordinate's preconditioned rate sum. All constants, the width, and the rates are selected before the random draw.

The theorem concerns all $2nd + 6$ learned parameters and answer supervision. The proof gives existence of a width threshold and a hierarchy of multipliers; it does not give a useful polynomial width or training-time bound. The argument does not provide a quantitative bound on d_0/n. The orthogonal and Gaussian laws are distinct: only the former satisfies the inner-product identities (99) exactly. The common multiplier choice needed for the GD comparison is addressed separately in Lemma 19.

E.2 Finite histories and comparison of pairwise inner products

We first control the optimizer before any favorable gradient sign has been established. Cauchy–Schwarz in the raw geometric sums gives the uniform coordinate bound

$$\left|\frac{\widehat{\mathsf{M}}_{i,t+1}}{\sqrt{\widehat{\mathsf{V}}_{i,t+1}}+\epsilon_i}\right| \le C_\beta, \qquad C_\beta = \left[(1-\beta_2)(1-\beta_1^2/\beta_2)\right]^{-1/2}, \tag{102}$$

for every $\epsilon_i \ge 0$. Thus each scalar interpolation is Lipschitz in the cumulative clock, regardless of its initial gradient signs.

The next lemma supplies positive scalar speeds after favorable matching has been acquired. Its window begins at a positive clock because all attention gradients vanish at $w_0 = 0$.

Lemma 21 (Tracking a favorable finite window)**.** *Fix $0 < \varepsilon < T$. On a stopped history through clock T, suppose a scalar's favorable gradient magnitude $g = -G$ belongs to $[g_*, H]$ on $[\varepsilon, T]$, where $g_* > 0$, and is Lipschitz in the clock with constant L. Suppose all earlier gradients have magnitude at most H. For every $0 < \ell < 1 < b$, a sufficiently small maximum step η^* makes its normalized increasing direction belong to $[\ell, b]$ at every mesh point in $[2\varepsilon, T]$. The statement uses the full inherited history and the global bias corrections.*

Proof. We separate the recent favorable gradients from the earlier history. Let $N = \lfloor \varepsilon/(2\eta^*) \rfloor - 2$. At a mesh point in $[2\varepsilon, T]$, the last N gradients belong to the favorable interval. Exact corrected geometric averages, and the bound that a lag of j updates spans at most $\eta^* j$ clock units, give

$$|\widehat{\mathsf{M}} + g| \le \frac{L\eta^*\beta_1}{1-\beta_1} + \frac{2H\beta_1^N}{1-\beta_1}, \tag{103}$$

$$|\widehat{\mathsf{V}} - g^2| \le \frac{2HL\eta^*\beta_2}{1-\beta_2} + \frac{2H^2\beta_2^N}{1-\beta_2}. \tag{104}$$

The history term for $\beta_1 = 0$ is absent. For any fixed $\xi > 0$, choose η^* so the right sides are at most ξg_* and ξg_*^2, respectively. Zero scalar stabilizer then gives

$$\frac{1-\xi}{\sqrt{1+\xi}} \le \frac{-\widehat{\mathsf{M}}}{\sqrt{\widehat{\mathsf{V}}}} \le \frac{1+\xi}{\sqrt{1-\xi}}. \tag{105}$$

Taking ξ small proves the claimed interval. No sign restriction was placed on the discarded part of the history. □

The table denominator allows a different comparison: at large width, each coordinate gradient is small, so the fixed E dominates its second moment on any fixed clock interval. Let $H = QK^\top$ denote the score matrix, and let $\mathcal{C}(H, \vartheta) = -\nabla_H F_{R,\mathbf{q}}$ denote its negative gradient. We use the realized scalar path as an input to the following reference; no scalar limiting differential equation is needed.

Lemma 22 (Finite comparison of pairwise inner products)**.** *Fix a finite clock horizon T and stop when scalar coordinates or raw row norms leave prescribed compact sets. Matrix norms in this lemma are operator norms. Define reference matrices $\overline{X}, \overline{Y}, \overline{H} \in \mathbb{R}^{n\times n}$ for $QQ^\top$, $KK^\top$, and $QK^\top$, respectively, by*

$$\overline{X}' = \frac{\lambda}{E}(\mathcal{C}\overline{H}^\top + \overline{H}\mathcal{C}^\top), \qquad \overline{Y}' = \frac{\lambda}{E}(\mathcal{C}^\top\overline{H} + \overline{H}^\top\mathcal{C}), \tag{106}$$

$$\overline{H}' = \frac{\lambda}{E}(\mathcal{C}\overline{Y} + \overline{X}\mathcal{C}), \qquad (\overline{X}_0, \overline{Y}_0, \overline{H}_0) = (s_0 I_n, s_0 I_n, 0), \tag{107}$$

where $\mathcal{C} = \mathcal{C}(\overline{H}, \vartheta(S))$ and $\vartheta(S)$ is the actual scalar interpolation. For either initialization law, with probability at least $1-\zeta$,

$$\sup_{S\le T}\left(\|QQ^\top - \overline{X}\| + \|KK^\top - \overline{Y}\| + \|QK^\top - \overline{H}\|\right) \le C_T\left(\eta^* + \sqrt{\frac{\log(nd/\zeta)}{d}}\right). \tag{108}$$

The constant may depend on the fixed compact sets, n, s_0, E, and the multipliers, but is independent of η^* and d.

Proof. We bound coordinate sizes before replacing the adaptive denominator. Concentration of Gaussian inner products gives the initial error in (108); the orthogonal law has zero initial error. For both laws, a union bound over the Gaussian or spherical coordinate tails gives

$$\max_{a,j}\{|Q_{aj,0}|, |K_{aj,0}|\} \le C\sqrt{\frac{\log(nd/\zeta)}{d}}. \tag{109}$$

Constants absorb the fixed n and s_0. On the stopped compact set, every gradient column is a bounded linear combination of the corresponding columns of Q and K. Since the table denominator is at least E, a discrete Gronwall estimate bounds the maximum coordinate over the whole history by its initial maximum times C_T. Corrected second moments are convex averages of squared gradients, hence

$$\max_{i,S_t\le T}\sqrt{\widehat{\mathsf{V}}_{i,t}} \le C_T\sqrt{\frac{\log(nd/\zeta)}{d}}. \tag{110}$$

We next control the first moments without a factor depending on width. The row bound gives a Frobenius bound for table gradients, and the fixed denominator gives table increments $O(\eta_t)$ in Frobenius norm. Equation (102) gives the same clock order for the six scalars. Smoothness of the finite score formulas therefore bounds successive table-gradient differences by $C_T\eta_t$. The bounded mean lag of the corrected geometric average implies

$$\|\widehat{\mathsf{M}}_{Z,t+1} - G_{Z,t}\|_F \le C_T\eta^*, \qquad \left\|\frac{\widehat{\mathsf{M}}_{Z,t+1}}{\sqrt{\widehat{\mathsf{V}}_{Z,t+1}}+E} - \frac{G_{Z,t}}{E}\right\|_F \le C_T\left(\eta^* + \sqrt{\frac{\log(nd/\zeta)}{d}}\right). \tag{111}$$

The second bound uses the maximum variance in (110), multiplied by a bounded Frobenius norm. Passing from the tables to these three matrix products introduces no factor $\sqrt{d}$. Their interpolated equations differ from (107) by the same error order. The reference and actual matrix products have the same realized scalar input. Gronwall on the stopped compact set proves (108). □

E.3 A REPRESENTATION BOUND BEFORE THE GATE MULTIPLIER IS CHOSEN

Permutation symmetry makes the reference explicit. Let $P = I_n - \mathbf{1}\mathbf{1}^\top/n$ be the centering projection and let $J = I_n - P$. Let $\phi_\vartheta(z)$ be the population loss averaged over one ordered distinct key pair with raw matching gap z, excluding the calibration term, and define the favorable score pressure $c(z,\vartheta) = -\phi'_\vartheta(z)/(n-1)$. Then $\mathcal{C}(bP,\vartheta) = c(b,\vartheta)P$. Uniqueness in (107) gives scalar functions A, b satisfying

$$\overline{X} = \overline{Y} = AP + s_0 J, \qquad \overline{H} = bP, \tag{112}$$

$$A' = 2(\lambda/E)c(b,\vartheta)b, \qquad b' = 2(\lambda/E)c(b,\vartheta)A, \qquad A^2 - b^2 = s_0^2. \tag{113}$$

In particular, the reference starts with zero raw matching gap. Its subsequent positive gap must be produced by the answer gradient.

The next estimate separates the choice of the gate multiplier from representation acquisition. This order is needed because the gate may have to grow much faster than the matching factors.

Lemma 23 (Acquisition and a uniform reference ceiling)**.** *Choose sufficiently small positive $\alpha_q = \alpha_p$, α_x, and α_w. There exist $\lambda, \delta_1 > 0$ and a finite row ceiling B, selected before $\alpha_u = \alpha_v$, with the following properties for scalar interpolations satisfying $|\vartheta_i'(S)| \le C_\beta\alpha_i$ on each mesh interval. The reference acquires $b(1) \ge \delta_1$ whenever the following input-path conditions hold through clock one: q, p stay in fixed compact positive intervals containing their initial values, x stays in a fixed compact interval, and $u, v > 0$. For a fixed $0 < w_b < \log(\kappa/(1-\kappa))$ and $c_w = \kappa - \mathrm{sigmoid}(w_b) > 0$, the decoder path must satisfy*

$$c_w\alpha_w(S-\eta^*)_+ \le w(S) \le C_\beta\alpha_w(S+\eta^*) \le w_b/2, \qquad 0 \le S \le 1, \quad \eta^* \le 1/2, \tag{114}$$

where $(z)_+ = \max\{z, 0\}$. These are conditional bounds on the realized scalar path; the stopped Adam comparison below verifies them. If the four favorable factor directions lie between $1/2$ and C_β from a clock no later than $1/4$ onward, and the gate satisfies

$$\rho \le \rho_* e^{-4B^2 m}, \qquad 0 < \rho_* < 1/4, \tag{115}$$

from clock one onward, then $b \geq \delta_1$ thereafter and every reference row norm squared is below $B^2 - 2$. A sufficiently large fixed $\alpha_u = \alpha_v$ forces (115) *from clock one onward once the favorable-direction comparison begins at a sufficiently small positive clock.*

Proof. We first obtain a gap bound that is uniform over large positive gates. Small matching, recency, and decoder multipliers keep q, p in a fixed positive interval, h in a fixed interval, and $0 \leq w \leq w_b/2$ up to clock one under the universal direction bound. Positivity of u, v bounds the binding error by $\overline{\rho} = e^D/(4 + e^D) < 1$. This interval includes the limit of arbitrarily strong binding. Lemma 8, after dividing its pressure by w and extending continuously to $w = 0$, provides $b_*, c_-, c_+ > 0$ such that

$$c_- w \leq c(b, \vartheta) \leq c_+ w \quad \text{for } 0 \leq b \leq b_*, \quad 0 \leq \rho \leq \overline{\rho}. \tag{116}$$

These constants do not depend on an upper gate bound. Calibration gives $-F_w \geq \kappa - \mathrm{sigmoid}(w) \geq c_w$ while $w \leq w_b$. Since $|F_w| \leq 1$ and the decoder starts with zero moments, its corrected increasing direction is at least c_w on this initial interval. Together with (102), this yields (114). Choose $\lambda > 0$ sufficiently small that (113) stays in $b \leq b_*/2$ through clock one. Its lower comparison then gives, for all sufficiently small η^*,

$$b(S) \geq \frac{\lambda s_0 c_- c_w \alpha_w}{4E} S^2 \quad (2\eta^* \leq S \leq 1), \qquad \delta_1 = \frac{\lambda s_0 c_- c_w \alpha_w}{4E} > 0. \tag{117}$$

Indeed, integration of $b' \geq 2(\lambda/E) s_0 c_- w$ gives the lower bound with $(S - \eta^*)^2 \geq S^2/4$. Both bounds precede the choice of gate multiplier.

We next bound score pressure without differentiating a score with respect to a table row. On raw row norms at most B, all gaps are at most $2B^2$. Write the exact overwrite odds as $U_J + B_J + E_J$ and the recall odds as C_ℓ, with the terms in Lemma 7. The stale term is U_J, the isolated off-key term is B_J, and E_J is the remaining off-key sum. For a raw gap z, their derivatives are

$$\partial_z T_{O_J} = m\{\rho U_J - (1 - 2\rho) B_J - (1 - \rho) E_J\}, \tag{118}$$

$$\partial_z C_\ell = -m \left\{ (1 - \rho) e^{-zm(1-\rho)+h} + e^{-zm} \sum_{k=2}^{\ell-1} e^{kh} \right\}. \tag{119}$$

The first term in the overwrite derivative is the contribution of the learned first-layer binding error. Condition (115) bounds $zm\rho$ by a universal constant, suppresses this adverse term, and makes each matching pressure favorable. Since the ordinary loss coefficient satisfies $K(T, w) \leq 2w$, for $z \geq \delta_1/2$ we obtain

$$0 < c(z, \vartheta) \leq Cwm \exp\{-c_\delta m + (R - 1)h\}. \tag{120}$$

Here $C, c_\delta > 0$ depend on the fixed population and δ_1, but not on B or the gate multiplier: $zm\rho \leq \rho_*/(2e)$, the stale derivative is at most $Ce^{-4B^2 m}$, and all other terms have a decay coefficient bounded below by a multiple of δ_1.

The hypothesized factor directions give $m \geq c_m(1 + S)^2$ for $S \geq 1$. Their upper bounds, and the universal recency and decoder bounds, give $m \leq C_m(1 + S)^2$ and $w, h \leq C_0(1 + S)$. These constants depend on the matching, recency, and decoder multipliers, not on the gate multiplier. Consequently (120) has a deterministic integrable upper bound $H_c(S)$, also independent of that multiplier. Set $D_c = \int_1^\infty H_c(S)\, dS$. The hyperbolic coordinate in (113) therefore satisfies

$$\mathrm{arsinh}(b(S)/s_0) \leq \mathrm{arsinh}(b_*/s_0) + 2(\lambda/E) D_c. \tag{121}$$

Let b^* be s_0 times the hyperbolic sine of this right side, and choose $B^2 > s_0 + b^* + 2$. Because $A \leq s_0 + b$, every reference row norm squared is then below $B^2 - 2$. This choice of B precedes the choice of the gate multiplier.

Finally, choose a positive affine slope A_0 and positive offset b_0 with $A_0^2 \geq 2B^2 c_0$ and $2b_0^2 > D + \log(1/\rho_*)$. The affine binding region is defined by $u \geq A_0 q + b_0$ and $v \geq A_0 p + b_0$. These inequalities imply (115). Fix the finite-window lower direction $\ell = 1/2$, and select permanent raw moment intervals whose normalized lower direction is greater than $3/4$ and upper direction is less than $3/2$. Choose $\alpha_u = \alpha_v$ so large that

$$\frac{\alpha_u}{4} > A_0(q_0 + C_\beta \alpha_q) + b_0 + 4, \qquad \ell \alpha_u > A_0 C_\beta \alpha_q. \tag{122}$$

A sufficiently short initial interval can lose less than $u_0/4$ of u. Once the lower direction holds, it gives $u(S) \geq \alpha_u S/4$ for $S \geq 1$. The first inequality yields $u \geq A_0 q + b_0 + 4$ at clock one and later, while the second preserves the affine face under the permanent moment bounds. Thus the gate has strict startup slack, completing the stated comparisons. □

E.4 Acquisition and binding under the actual updates

Lemma 23 gave comparisons conditional on favorable scalar directions. We now verify those conditions without assuming favorable initial gradients. The distinction between clock one and a much shorter initial interval prevents a circular choice of multipliers.

Choose all constants through B and α_u in the order of Lemma 23. Next choose an initial comparison clock $\varepsilon > 0$ so small that

$$\varepsilon < \frac{1}{8}, \qquad 2C_\beta \alpha_u \varepsilon < \frac{u_0}{4}, \qquad 2C_\beta \alpha_q \varepsilon < \frac{q_0}{4}. \tag{123}$$

Stop also if an input-path condition of Lemma 23 fails before clock one. Small matching and recency multipliers retain their compact intervals, and calibration gives (114). Before clock 2ε, the unconditional direction bound keeps all four factors above half their initial values even if their gradients initially have adverse signs. At clock ε, the lower bound in (117) is a strictly positive deterministic number. For any fixed later horizon T, choose the error tolerance in (108) below a fixed fraction of this number. The actual gaps are then positive from clock ε until a possible first exit. Before clock one they remain in the local interval of (116); after clock one the affine binding region makes their matching pressures favorable.

These positive gaps give positive actual gate gradients, not merely positive reference gradients. If $d_{ab} = -\phi'_\vartheta(\Delta_{ab})$ and the average below is over ordered distinct pairs, the exact identities are

$$-F_q = q^{-1} \operatorname*{avg}_{a \neq b} \Delta_{ab} d_{ab}, \qquad -F_p = p^{-1} \operatorname*{avg}_{a \neq b} \Delta_{ab} d_{ab}, \tag{124}$$

$$-F_u = 2v \operatorname{sigmoid}(uv)\rho(1-\rho)F_\rho, \qquad -F_v = 2u \operatorname{sigmoid}(uv)\rho(1-\rho)F_\rho. \tag{125}$$

Here $F_\rho > 0$ whenever the raw gaps and w are positive. On the finite stopped interval $[\varepsilon, T]$, all scalar coordinates and row norms lie in fixed compact sets, and all gaps have a fixed positive lower bound. Before clock one use the local pressure set; after it use the binding region. Each of the four magnitudes in (125) therefore has a positive compact minimum. That minimum can be extremely small when the gate multiplier or T is large; we require no lower bound uniform in either quantity.

The compact bounds also verify the clock regularity needed by Lemma 21. Scalar increments are $O(\eta_t)$ by (102). A fixed table stabilizer and bounded table gradients give Frobenius increments $O(\eta_t)$, hence increments of the same order in the three matrix products. The four gradients are smooth functions of the finite score matrix and the scalars. Their clock Lipschitz constants are therefore finite and independent of width. Choosing η^* sufficiently small makes their directions arbitrarily close to one throughout $[2\varepsilon, T]$. The possible adverse history before ε is already included in (104); its remaining movement fits (123).

The decoder stays positive throughout the comparison. Its initial calibration bound gives $w(\varepsilon) \geq c_w \alpha_w \varepsilon/2$ for $\eta^* < \varepsilon/2$. On any later downward entry into $w \leq w_b$, all new gradients favor increasing w. The adverse inherited first moment decays at rate β_1, while its inherited second moment supplies a denominator at rate $\sqrt{\beta_2}$. Its entire remaining downward movement, including the crossing step, is bounded by

$$\frac{C_\beta \eta^* \alpha_w}{1 - \beta_1/\sqrt{\beta_2}}. \tag{126}$$

Choose this below $w_b/2$. Every such visit stays above $w_b/2$, so $w \geq \min\{c_w \alpha_w \varepsilon/2, w_b/2\} > 0$ after that initial interval.

We can now close the finite stopped comparison. Up to clock 2ε use the unconditional factor bounds; from then until clock one use the positive local matching pressures and the finite-window moment lemma. Equation (122) places the actual scalars strictly inside the binding region at clock one. For later clocks, the same moment comparison preserves the scalar faces, while Lemma 23 bounds the reference matrices. Making (108) smaller than the strict row and gap slack keeps the actual row norms and raw gaps within their prescribed bounds. This establishes the comparison through any predetermined finite T. Its constants are chosen before the final small step and large width. The algorithm checks none of these proof clocks.

E.5 THE INFINITE TAIL WITH THE SAME TABLE MULTIPLIER

The finite comparison alone does not control an infinite run. The learned gate supplies the additional fact needed for that step: table gradients become integrable in cumulative learning time. On row norms at most B, gaps at least $\delta_1/2$, and the binding region, (119) and $K(T,w) \le 2w$ give

$$\|G_{Z,t}\|_F \le C_B w_t m_t e^{-c_\delta m_t + (R-1)h_t} \le H_Z(S_t), \qquad H_Z(s) = C(1+s)^3 e^{-cs^2+Cs}. \tag{127}$$

The bound retains the stale leakage derivative. The factor comparisons give the quadratic lower bound for m; the universal direction bound gives the linear upper bounds for w, h. Enlarging C covers the initial finite interval, so this is a uniform estimate on every stopped history.

To control the full inherited table history, expand the raw first moment from an arbitrary global age k. Monotonicity of η_t and an exchange of finite sums give

$$\sum_{t\ge k} \eta_t \|\mathsf{M}_{Z,t+1}\|_F \le \frac{\eta_k \beta_1}{1-\beta_1} \|\mathsf{M}_{Z,k}\|_F + \sum_{t\ge k} \eta_t \|G_{Z,t}\|_F, \tag{128}$$

$$\sum_{t\ge k} \|Z_{t+1} - Z_t\|_F \le \frac{\lambda}{E(1-\beta_1)} \left\{ \frac{\eta_k \beta_1}{1-\beta_1} \|\mathsf{M}_{Z,k}\|_F + \sum_{t\ge k} \eta_t H_Z(S_t) \right\}. \tag{129}$$

The inequalities also hold for a finite stopping horizon. The second uses the actual bias correction and only the lower bound E for the denominator; it does not approximate or reset the second moment.

We choose a movement allowance that also controls intermediate table segments. Set $k = \min\{t : S_t \ge T\}$ for a deterministic late clock T. Take $\eta^* \le 1$ and apply the finite comparison through $T+1$, which contains S_k. Tighten this comparison so rows at age k are at most $B_0 = \sqrt{B^2-1}$ and gaps are at least $3\delta_1/4$. Identify $Z = (Q,K)$ with its stacked matrix and choose

$$0 < \varepsilon_Z \le \min\left\{1, \frac{B-B_0}{2}, \frac{\delta_1}{4(4B+2)}\right\}. \tag{130}$$

Every point in the convex ball $\|Z - Z_k\|_F \le \varepsilon_Z$ has rows at most B and gaps at least $\delta_1/2$, since

$$|\Delta_{ab}(Z) - \Delta_{ab}(Z_k)| \le 4B_0\varepsilon_Z + 2\varepsilon_Z^2 \le \delta_1/4. \tag{131}$$

The allowance is independent of T. Choose T so the gradient-tail integral in (129), multiplied by its displayed constant, is smaller than $\varepsilon_Z/4$. Beyond the maximum of H_Z, each discrete left sum is bounded by its integral plus $\eta^* H_Z(T)$. The inherited moments are bounded on the stopped history; decreasing η^* makes the joint-table bound smaller than $\varepsilon_Z/2$. At global age k, the separate first- and second-moment estimates in (104) place the processed raw ratios $-\mathsf{M}_{i,k+1}/g_{i,k}$ and $\mathsf{V}_{i,k+1}/g_{i,k}^2$ inside the permanent intervals, with strict slack on each nondegenerate interval (the first-moment identity is exact when $\beta_1 = 0$): choose the relative moment errors sufficiently small and the global age sufficiently large that conversion from corrected to raw moments preserves the slack. All these requirements are imposed on the one initial value of η^*.

For the four factors, the processed raw ratios are $-\mathsf{M}_{i,t+1}/g_{i,t}$ and $\mathsf{V}_{i,t+1}/g_{i,t}^2$, where $g_{i,t} = -G_{i,t} > 0$. Let $L_\gamma = M_-/V_+ > 3/4$ and $U_\gamma = M_+/V_- < 3/2$ be the direction bounds from (49). Increase the starting global age so the bias-correction factor f_{t+1} in (51) is at least $2/3$. Whenever the raw intervals hold,

$$q_{t+1} - q_t \ge \tfrac{1}{2}\alpha_q \eta_t, \tag{132}$$

$$(u_{t+1} - A_0 q_{t+1} - b_0) - (u_t - A_0 q_t - b_0) \ge \eta_t f_{t+1}(\alpha_u L_\gamma - A_0 \alpha_q U_\gamma) > 0. \tag{133}$$

The strict inequality follows from (122) and $C_\beta \ge 1$. The same comparisons apply to p, v. Thus the raw intervals give positive factor growth and preserve both affine faces with the actual global bias corrections.

It remains to preserve those intervals under the nonsummable table rates. For the four favorable factor magnitudes $g_{i,t} = -G_{i,t}$, Lemma 9, with row bound B and the binding region, gives scalar logarithmic derivatives of order $(1+S_t)^2$ and table derivatives of order $(1+S_t)^3$. Complete the square in (127), then split the first-moment convolution at half the global age. Since $S_{\lfloor t/2 \rfloor} \ge S_t/3$ for all sufficiently large t,

$$\|\mathsf{M}_{Z,t+1}\|_F \le C\{\beta_1^{t/2} + e^{-c_1 S_t^2}\}, \qquad \|Z_{t+1} - Z_t\|_F \le C\eta_t\{\beta_1^{t/2} + e^{-c_1 S_t^2}\}. \tag{134}$$

For $\beta_1 = 0$, omit the geometric term at positive ages. The adjacent factor-gradient changes consequently satisfy

$$|\log g_{i,t+1} - \log g_{i,t}| \le C\eta_t(1+S_t)^2 + C\eta_t(1+S_t)^3\{\beta_1^{t/2} + e^{-c_1 S_t^2}\}. \tag{135}$$

The table term has supremum $O(\eta^*)$ because a Gaussian or geometric decay bounds its polynomial multiplier. The scalar term tends to zero and has arbitrarily small supremum for $r > 2/3$. The exact geometric moment recurrences therefore preserve the raw intervals of (50). Their common global-age factor cancels in the binding-region boundary comparisons, and (122) preserves the affine binding region.

We induct on the global update age, starting at k. Gradients already evaluated on the retained history give the finite-sum version of (129), keeping the next tables in the convex ball. Equation (133) keeps the next factors in the positive affine region, and the decoder barrier keeps w positive. The whole straight parameter segment therefore has the row, gap, positivity, and binding bounds needed for the logarithmic derivative estimate. On this segment, (135) is at most γ after the one initial amplitude is made sufficiently small. The exact moment recurrences then place the next processed ratios in their intervals. This closes the simultaneous induction without using a future gradient before its state has been controlled. Absolute convergence of the tables follows from (129). Their gradients and second moments tend to zero, giving, for each table coordinate j,

$$\frac{\nu_t}{\sqrt{\widehat{\mathsf{V}}_{j,t+1}} + E} \sim \frac{\lambda}{E}\eta_t, \qquad \sum_t \frac{\nu_t}{\sqrt{\widehat{\mathsf{V}}_{j,t+1}} + E} = \infty. \tag{136}$$

It is gradient decay, rather than a summable effective rate, that bounds the table motion.

E.6 Scalar speeds, uniform answers, and spectral growth

The retained raw moment intervals already give linear upper and lower factor growth. To obtain their exact speeds, retain one positive-gap short recall term in Lemma 7. Its positive matching and gate derivatives, the lower decoder bound, and the finite row bounds give

$$g_{i,t} \ge Le^{-C(1+S_t)^2} > 0, \qquad i \in \{q, p, u, v\}. \tag{137}$$

For the gate derivative this estimate includes the factor $\rho \ge ce^{-2uv}$; hence it remains valid when binding becomes strong. Equation (135) tends to zero, and $S_t^2 = o(t)$. The finite-history tracking argument (53) therefore gives normalized factor directions tending to one, including the finite earlier signed history. Weighted averaging against η_t proves the first four limits in (100).

The recency and decoder coordinates acquire their favorable signs later. The positive overwrite contribution P_h and adverse recall contribution N_h to $-F_h$ satisfy

$$P_h \ge ce^{-w-2h}, \qquad N_h \le Cwe^{-c_\delta m+(R-1)h}, \qquad N_h/P_h \longrightarrow 0. \tag{138}$$

The lower bound uses the positive mass q_1; the upper bound uses the learned gap. Thus eventually $-F_x > 0$, with a lower bound $Le^{-C(1+S_t)}$. The finite positive pressure sums have the logarithmic derivative orders used in (135); subtraction is safe once $N_h \le P_h/2$. Equation (53) yields $x_t/S_t, h_t/S_t \to \alpha_x$. The overwrite odds are then $O(e^{-2h} + e^{-c_\delta m})$, while the recall odds are $O(e^{-c_\delta m+(R-1)h})$. They tend uniformly to zero, as do their products with $w = O(S_t)$. Every decoder gradient term is eventually favorable, $-F_w \ge ce^{-w}$, and the same positive-sum logarithmic estimate applies. Tracking gives $w_t/S_t \to \alpha_w$, completing (100).

The scalar speeds imply the uniform retrieval claim through Proposition 6. The bounded row norms and binding region give $B^2 m_t \rho_t \to 0$. For a constant $0 < c_* < \delta_* c_0 \alpha_q \alpha_p/(2\alpha_x)$, the same speeds give

$$\delta_* m_t - (\lfloor c_* S_t \rfloor - 1)h_t \ge \tfrac{1}{2}\delta_* c_0 \alpha_q \alpha_p S_t^2 \quad \text{eventually.} \tag{139}$$

Since $h_t, w_t \to \infty$, the proposition yields $\mathcal{E}_{\theta_t}(\lfloor c_* S_t \rfloor) \to 0$, uniformly over older prefixes. This proves Theorem 20.

The inner-product comparison also answers a spectral question about the learned queries. The scalar amplification must be included: the raw tables converge, whereas the effective query and key matrices grow.

Corollary 24 (Spectral growth of effective queries and keys)**.** *The constants in Theorem 20 can be chosen so that, on the same probability event, the smallest row singular values satisfy, for all sufficiently large t,*

$$\sigma_{\min}(Q_t),\ \sigma_{\min}(K_t) \geq \tfrac{1}{2}\sqrt{s_0}, \tag{140}$$

$$\sigma_{\min}(q_t Q_t) \geq \tfrac{1}{4}\alpha_q \sqrt{s_0}\, S_t, \qquad \sigma_{\min}(p_t K_t) \geq \tfrac{1}{4}\alpha_p \sqrt{s_0}\, S_t. \tag{141}$$

Here $\sigma_{\min}$ is the smallest of the n row singular values. The matrices $q_t Q_t, p_t K_t$ are the effective factors in $m_t Q_t K_t^\top = c_0 (q_t Q_t)(p_t K_t)^\top$.

Proof. The reference matrices $\overline{X}, \overline{Y}$ in (113) have least eigenvalue at least s_0, since $A \geq s_0$. Tighten the comparison error in (108) at the late clock to $s_0/4$, and tighten the tail movement allowance in (129) to $\sqrt{s_0}/8$. The row singular values are then at least $\sqrt{3s_0/4} - \sqrt{s_0}/8 > \sqrt{s_0}/2$ by Weyl's inequality. Multiplication by the scalar speeds in (100) gives the remaining bounds. □

For clarity, the choices in the proof have a fixed order. The population, initial scalars, s_0, E, and moment parameters come first. We then choose the small matching, recency, and decoder multipliers and the table multiplier, obtaining δ_1. The pressure integral fixes B; only then is the gate multiplier chosen. The short initial comparison clock and late tail clock come next. Finally we choose the one small amplitude and large finite width that satisfy all finite-window and tail estimates. No constant depends on a realized training gap, and no proof clock changes the training protocol.

F SGD WITH AVERAGING OVER ALL KEY PAIRS

Appendix D proves a finite retrieval boundary while both factorized tables keep growing. We extend that argument to stochastic gradients by averaging over keys and sampling record lengths. This sampling rule preserves the inner-product identities (58) needed for the scalar reduction; increasing the number of sampled lengths controls the remaining noise. The result uses the same answer loss and one schedule from initialization.

F.1 SAMPLING RULE AND GUARANTEE

To preserve the uniform treatment of keys, each sampled length pair includes all ordered distinct key pairs. Write $N_{\mathrm{pair}} = n(n-1)$ and let $\mathcal{L}_\theta(X) = -\log P_\theta(y(X) \mid X)$ denote the answer loss of a training sequence. Let N_t be the number of sampled length pairs at update t, and let $\mathcal{F}_t$ contain initialization and all preceding samples. Draw N_t independent pairs $(J_{t,r}, L_{t,r})$, with independent spacing $J_{t,r} \sim \mathbf{q}$ and recall lag $L_{t,r}$ uniform on $\{2, \ldots, R\}$. Calibration is evaluated exactly. Global value complementation leaves the answer loss unchanged, so each summand represents both answer signs. The sampled objective is

$$\widehat{F}_t(\theta) = \kappa_{\mathrm{cal}}\, \mathrm{softplus}(-w) + \frac{1}{N_t N_{\mathrm{pair}}} \sum_{r=1}^{N_t} \sum_{a \neq b} \left[\omega_C \mathcal{L}_\theta(C^{a,b}_{L_{t,r}}) + \omega_O \mathcal{L}_\theta(O^{a,b}_{J_{t,r}}) \right],$$

$$\mathbb{E}[\nabla \widehat{F}_t(\theta_t) \mid \mathcal{F}_t] = \nabla F_{R,\mathbf{q}}(\theta_t). \tag{142}$$

Each update evaluates $N_t N_{\mathrm{pair}}$ paired summands, each containing one recall and one overwrite loss. Thus N_t counts length samples, rather than individual training examples.

The average over key pairs preserves the inner-product identities (58) from initialization (55). Thus the sampled updates can use the same rates and entrance argument as GD.

Corollary 25 (SGD with averaging over all key pairs)**.** *Use the training distribution, initialization, and fixed positive learning-rate multipliers constructed in Theorem 13. For each confidence level $\zeta \in (0,1)$, there are a sufficiently small $\eta^* > 0$ and sufficiently large $N_0 > 0$ such that SGD on* (142)*, with rates and a predetermined integer batch schedule*

$$\eta_t = \frac{\eta^*}{t+1}, \qquad \eta_{Q,t} = \eta_{K,t} = \lambda \eta_t, \qquad \eta_{i,t} = \alpha_i \eta_t, \qquad N_t \geq \left\lceil N_0 (t+1)^2 \log(t+2) \right\rceil, \tag{143}$$

satisfies all conclusions of Theorem 13 with probability at least $1-\zeta$. Here i ranges over the six scalar parameters, with $\alpha_q = \alpha_p = \alpha_m$ and $\alpha_u = \alpha_v = \alpha_g$. All constants are chosen before sampling.

In particular, the same sharp balance (57) gives uniform retrieval through lag $R+2$ and a fixed failure witness at every larger lag. The multiplier hierarchy is unchanged: binding learns quickly, recency more slowly, and the decoder more slowly still. Only the length average is replaced by sampling.

F.2 Exact geometry and control of sampling errors

The key average preserves the deterministic reduction sample by sample. Let $P = I - \mathbf{1}\mathbf{1}^\top/n$ and $J = I - P$ be the centered and constant projectors. At a state satisfying (58), write the negative sampled score gradient as $\widehat{c}_t P$, where $\widehat{c}_t$ is its scalar coefficient. Therefore the exact updates preserve

$$Q_t Q_t^\top = K_t K_t^\top = A_t P + s_0 J, \qquad Q_t K_t^\top = b_t P. \tag{144}$$

Put $z_t = \lambda \eta_t \widehat{c}_t$. Lemma 14 applies to each realized batch with $e = z_t$ and $d_{\mathrm{match}} = (n-1)\widehat{c}_t$. In particular, the imbalance $\mathcal{I}_t = m_t - \chi A_t$, with $\chi = c_0 \alpha_m (n-1)/\lambda$, and determinant $\mathcal{D}_t = A_t^2 - b_t^2$ satisfy

$$\begin{aligned} \mathcal{D}_{t+1} &= (1 - z_t^2)^2 \mathcal{D}_t, \\ \mathcal{I}_{t+1} - \mathcal{I}_t &= z_t^2 \left(\frac{\chi^2 b_t^2}{m_t} - \chi A_t \right). \end{aligned} \tag{145}$$

Sampling therefore introduces no first-order error in these two relations.

Only the length frequencies remain random. Let p_j and $\widehat{p}_{t,j}$ denote the population and empirical probabilities of the finitely many length pairs, and let N_{len} be their number. For a prescribed tolerance $\varepsilon_0 > 0$, define the event

$$\mathcal{G} = \left\{ \max_j |\widehat{p}_{t,j} - p_j| \le \frac{\varepsilon_0}{t+1} \text{ for every } t \ge 0 \right\}. \tag{146}$$

Hoeffding's inequality and a union bound imply

$$\mathbb{P}(\mathcal{G}^c) \le 2 N_{\mathrm{len}} \sum_{t \ge 0} (t+2)^{-2 N_0 \varepsilon_0^2}. \tag{147}$$

Once $2N_0\varepsilon_0^2 \ge 2$, the series is dominated by $\sum_{t\ge 0}(t+2)^{-2}$. Dominated convergence as $N_0 \to \infty$ therefore makes this probability at most ζ. The event controls the sampled coefficients at every parameter value, so it remains valid along the adaptive trajectory.

We next verify that this event supplies the relative errors required by the GD continuation argument. Work on a trajectory stopped at its first exit from the basin in (70) and from slightly enlarged compact intervals for $\mathcal{I}_t, \mathcal{D}_t$, with $\mathcal{D}_t > 0$. Also stop if the sum of normalized errors on the left of (151) exceeds $2\varepsilon_{\mathrm{num}} e^{-\widetilde{S}/2}$, including inside a parameter interpolation segment. Here $\widetilde{S}$ is cumulative time since entrance, and $\varepsilon_{\mathrm{num}} > 0$ is small enough that the inward field comparisons tolerate this relaxed bound. Finite-entrance compactness and small η^*, ε_0 establish it initially. Recall the effective mobility $\mathfrak{a}_t$ and set

$$\begin{aligned} U &= e^{-2h}, \qquad k_0 = 2we^{-w}, \qquad k = k_0 U, \\ f &= M - (R+1)h - \tfrac{3}{2}\log M, \qquad \Psi = M^{3/2}\rho. \end{aligned} \tag{148}$$

On these intervals, $m, A, b \asymp \sqrt{M}$ and $\mathfrak{a}_t \sim \mathfrak{a}_\infty M^{3/2}$ uniformly. For each length atom the exact odds satisfy $T_C \le CU/M^{3/2}$ and $T_O \le CU$. If $\widehat{Y} = M^{-1}\partial_\rho \widehat{F}$ denotes its gate pressure, these odds bounds give

$$\begin{aligned} \left|\partial_M \widehat{F}\right| &\le \frac{Ck}{M^{3/2}}, \qquad \left|\partial_h \widehat{F}\right| \le Ck, \\ \left|\partial_w \widehat{F}\right| &\le Ce^{-w}, \qquad 0 \le \widehat{Y} \le Ck, \quad ck \le Y \le Ck. \end{aligned} \tag{149}$$

Here Y is the population gate pressure. Its lower bound uses $q_1 > 0$; individual longer-spacing atoms need not satisfy that lower bound. The matching bound includes the derivative of first-layer leakage, since $\rho\widehat{Y} = \Psi\widehat{Y}/M^{3/2}$.

These bounds make the invariant drift summable. Since $\widehat{c} = m(-\partial_M \widehat{F})/(n-1)$, the exact identities imply

$$|z_t| \le C\eta_t k_t/M_t, \qquad |\mathcal{I}_{t+1} - \mathcal{I}_t| \le C\eta_t^2 k_t^2/M_t^{3/2}, \\ \left|\log \frac{\mathcal{D}_{t+1}}{\mathcal{D}_t}\right| \le C\eta_t^2 k_t^2/M_t^2 \quad \text{when } |z_t| < \tfrac{1}{2}. \tag{150}$$

Thus small η^* keeps both invariants in their designated intervals, $\mathcal{I}_t$ converges, and $\mathcal{D}_t$ has a positive limit.

The scalar errors are controlled on their natural scales. Let r_h, r_f, r_w, r_Ψ denote errors in the derivatives of h, f, w, Ψ, respectively, along the interpolation of the parameter iterates. These errors include sampling and discretization, including the second-order terms in (145). On $\mathcal{G}$, Taylor expansion and (149) give

$$\frac{|r_h| + |r_f|}{k} + \frac{|r_w|}{e^{-w}} + \frac{\sqrt{M}\,|r_\Psi|}{\Psi Y} \le C\left(\eta_t + \frac{\varepsilon_0}{t+1}\right)\mathcal{P}(\widetilde{S}). \tag{151}$$

Here $\widetilde{S}$ is cumulative time since entry into the basin, and $\log \mathcal{P}(s) = o(s)$. This envelope follows from the stopped bound $C[1 + \log^2(2+s)]$ on every scalar and table-row norm established after (76). Derivatives of the finite odds formulas, including the reciprocals of the positive error scales, have at most exponential dependence on fixed powers of this bound. There is no extra inverse-leakage factor: each gate update contains ρ, which cancels when the error is divided by Ψ. Equivalently, derivatives of $\log \rho$ contribute only polynomial factors in $u = v$, and every gate increment retains its factor ρ.

The offset $1/t$ steps decay exponentially in $\widetilde{S}$. Consequently the right side of (151) is at most $\varepsilon_{\mathrm{num}} e^{-\widetilde{S}/2}$ after choosing η^* and ε_0 sufficiently small. This strictly improves the provisional error bound on every interpolation segment. The inward field comparisons exclude exit through the scalar faces, and (150) preserves the invariant intervals. Thus the joint stop never occurs. Lemma 16 therefore applies, with the actual global update index and unchanged rates.

F.3 Proof of the corollary and its scope

Proof of Corollary 25. We first use the deterministic entrance to learn matching and binding, then apply the preceding estimates to all subsequent updates. Choose the fixed multipliers and the finite flow entrance time from Lemma 18. Its endpoint lies strictly inside the joint continuation basin. Choose a compact neighborhood of the finite reference trajectory and continuation constants uniform on a neighborhood of that endpoint. On $\mathcal{G}$, finite-horizon Euler and gradient-error bounds put the sampled trajectory inside this neighborhood when η^* and ε_0 are sufficiently small. Orthogonal initialization and the key average make (144) exact throughout this argument.

The preceding stopped estimates apply from the first mesh point past the reference entrance time. They preserve the basin and make the relative errors vanish in the divergent clock $v = \int k\, d\widetilde{S}$. The imbalance and determinant remain bounded, so the matching mobility has the GD coefficient $\mathfrak{a}_\infty$ in (57). Let $\Gamma = \omega_C/(R-1)$ and $\Omega = \omega_O q_1$ be the longest-recall and shortest-overwrite masses. The residual equation is

$$\frac{df}{dv} = \mathfrak{a}_\infty \Gamma e^{-f}(1 + o(1)) - 2(R+1)\alpha_x \Omega + o(1). \tag{152}$$

The inward-band argument of Lemma 16 proves (57). At lag $R+2$ the target advantage is $\frac{3}{2}\log M + O(1)$; at fixed lag $r \ge R+3$ the single-competitor witness has advantage $(R+2-r)h + \frac{3}{2}\log M + O(1) \to -\infty$. Together with $M\rho \to 0$ and $w \to \infty$, the retrieval bounds yield the claimed answer probabilities.

Finally choose one N_0 making (147) at most ζ. The same event covers the finite entrance and infinite continuation. All constants are fixed before sampling; the proof's entrance time causes no operation in the algorithm. This proves the result from initialization. □

The growing-batch condition is sufficient, not a lower bound on the batch size needed for the conclusion. It permits the more conservative quartic schedule used in Appendix G. The average over key pairs preserves (58) at every update. With iid key minibatches, independent Gaussian tables, or

Table 1: Training-only selection of shared multipliers. The selection loss is the larger maximum family loss across GD and Adam after 100,000 updates. The six paired candidates use the same population, initialization seed, and base schedule.

α_m	α_g	α_x	α_w	λ	Selection loss
0.2	2	0.1	0.05	1	0.6653
0.2	5	0.1	0.05	1	0.6168
0.2	5	0.05	0.025	1	0.6866
0.2	5	0.1	0.05	4	0.5859
0.5	5	0.1	0.05	4	0.5833
0.2	10	0.2	0.1	4	0.4432

fixed length batches, a finite-time approximation does not establish the same infinite-time balance. Those extensions require a separate stability argument.

G Experimental setups and supporting results

We specify the protocols and retrieval metrics for Section 5, including the controls and Transformer comparison that delimit its conclusions.

G.1 Uninterrupted training of the learned model

Architecture, population, and initialization. The model has $n = 4$ keys and table width $d = 128$, so its two learned tables and six scalars contain $2nd + 6 = 1{,}030$ parameters. The constants are $c_0 = 1$ and $D = 0$. The calibration, recall, and overwrite families receive masses $(\kappa, \omega_C, \omega_O) = (0.6, 0.3, 0.1)$. Recall lag is uniform on $2, \dots, R$, with $R = 3$ except in the $R = 5$ control. Overwrite spacing has law $\mathbf{q} = (0.5, 0.3, 0.2)$ on $1, 2, 3$. Each family averages uniformly over its admissible keys and includes both value complements. These masses satisfy (13) for both values of R. GD and Adam evaluate the exact finite population loss, including each nonlinear per-length loss before averaging.

For each paired comparison, both optimizers receive identical initial arrays. A random orthogonal frame gives the stacked table $W_0 = [Q_0; K_0]$ eight orthonormal rows, with $s_0 = 1$. The scalar initialization is $q_0 = p_0 = 1$, $u_0 = v_0 = 0.25$, $x_0 = -1$, and $w_0 = 0$. Orthogonality holds to floating-point precision; small residuals in $Q_0 K_0^\top$ can produce nonzero initial scalar gradients under zero-epsilon Adam. We retain these updates without thresholding. The separate Gaussian control uses independent $N(0, 1/128)$ table entries and therefore does not start from exactly zero matching.

Updates and rate selection. Every parameter group uses its fixed learning-rate multiplier times $\eta_t = 40/(t + 8000)$, with $t = 0, \dots, 99{,}999$, giving cumulative rate $S = 104.11$. The selected multipliers are $\alpha_m = 0.2$, $\alpha_g = 10$, $\alpha_x = 0.2$, $\alpha_w = 0.1$, and $\lambda = 4$. Adam uses $(\beta_1, \beta_2) = (0.9, 0.99)$, zero scalar stabilizers, table stabilizer $E = 1$, and zero initial moments. Neither optimizer uses clipping, weight decay, phase changes, frozen parameters, or moment resets.

To choose shared multipliers without selecting for a retrieval gap, six candidates are evaluated on initialization seed 0 (Table 1). Eligibility requires both optimizers to finish with finite parameters, positive minimum raw matching gap, and every training-family mean loss below $\log 2$. Among eligible candidates, the rule minimizes the larger of the two optimizers' maximum family losses. Loss ties within 0.001 are broken by the smaller maximum of $B_t^2 m_t \rho_t$ across optimizers, with B_t the largest table-row norm, then by table order. All six candidates are eligible; the last row is selected. Retrieval certificates and test probabilities do not enter the rule. The primary evaluation uses fresh frame seeds 1, 2, and 3.

Numerical implementation. Parameters are stored in float64. Scalar gradients and Adam moments use signed logarithmic arithmetic to represent values below the ordinary floating-point range. This implements the usual bias-corrected Adam recurrence without rescaling the loss or normalizing gradients. Zero scalar gradients with zero moments produce a zero update. All endpoint quantities are

recomputed from saved tables and scalars. Independent checks compare literal attention, gradients, optimizer recurrences, inner-product identities, and certificate boundaries against their analytical expressions. These checks concern finite computations; the theorem's sufficient numerical constants remain unverified. The primary Adam runs need not enter the proof's conservative binding region for the directly evaluated finite certificates to apply. The accompanying anonymous code supplement provides Python modules for structured-model and Transformer training, synthetic-data generation, retrieval evaluation, and table aggregation. The structured implementation includes SGD controls with averaging over all key pairs. The experimental configurations and evaluation protocols are specified in this appendix.

G.2 Retrieval metrics and controls

Uniform confidence certificate. To distinguish a uniform guarantee from sampled accuracy, we evaluate (33) using the actual largest query/key row norm B_t and smallest raw gap $\Delta_{\min,t}$. For a desired correct-answer probability $\gamma \in (1/2, 1)$, let $r_\gamma(t)$ denote the largest certified integer lag:

$$r_\gamma(t) = \max\left(\{0\} \cup \left\{r \geq 1 : \operatorname{sigmoid}\left[-w_t \frac{1 - \mathcal{B}_t(r)}{1 + \mathcal{B}_t(r)}\right] \leq 1 - \gamma\right\}\right). \tag{153}$$

The maximum is over integers, and the bound is used only when Proposition 6's sign conditions hold; otherwise we report zero. Monotonicity in r makes the certificate cover every smaller lag. The calculation includes first-layer binding error and arbitrarily long older prefixes. A zero value means this sufficient bound certifies no positive lag, rather than that retrieval is impossible. The integer boundary is checked without imposing a finite search cap.

Fixed-sequence tests. The lag-r recall witness is $((a, +1), (b, -1)^{r-1}; ?a)$ for the first two vocabulary keys a, b. To isolate extrapolation in the older prefix while preserving a training suffix, we also prepend L copies of $(a, -1)$ to the lag-three witness. The additional records leave the target and its lag unchanged. Because each added record and the target bind to K_a, their exact additional non-target-to-target attention odds are $\sum_{k=1}^{L} e^{-kh_t}$. This expression includes the first-layer forget gate. It agrees with literal attention evaluation on 192 finite sequences to absolute error below 3.6×10^{-15}. This post-training diagnostic uses the fixed grid $L \in \{0, 1, 2, 4, 8, 16, 32, 64\}$ and $r \in \{1, 3, 5, 6, 8, 10, 16, 32\}$ across all evaluation runs; it does not influence rate selection.

Primary endpoints. All three fresh Adam seeds have $r_{0.99} = 14$, $r_{0.9} = 15$, and minimum raw gaps between 0.83085 and 0.83115. GD's scalar endpoints agree across these frames, as the identities (58) predict; Adam's small seed variation is empirical. Both optimizers learn positive matching gaps. GD classifies every training example correctly, but its minimum training probability is 0.5852, compared with 0.999966 for Adam. The primary comparison therefore partly reflects unequal training fit.

GD's zero certificate alone does not establish retrieval failure, so we also evaluate the stale-prefix diagnostic directly. Prepending 64 records $(a, -1)$ to the training recall $((a, +1), (b, -1), (b, -1); ?a)$ increases its length from 3 to 67 records while keeping the target at lag three. For seed 1, GD's correct-answer probability falls to 0.3476, compared with 0.999966 for Adam.

Stabilizer, initialization, and gate controls. To compare retrieval at matched training fit, we change only Adam's scalar stabilizer from zero to 10^{-8}. Training loss changes from 3.3624×10^{-5} to 3.3634×10^{-5}, while $r_{0.99}$ falls from 14 to 5 and correct-answer probability on the fixed lag-six recall falls from 0.999966 to 4.57×10^{-5}. Both settings handle the lag-three stale-prefix test accurately; the positive-stabilizer failure concerns the larger target lag. This control separates finite retrieval behavior at nearly identical training fit, but does not establish that a positive stabilizer bounds the retrieval range forever.

The separate Gaussian Adam control also has $r_{0.99} = 14$. Reducing the gate multiplier to $\alpha_g = \alpha_m = 0.2$ retains $r_{0.99} = 14$, so these finite experiments do not show that the proof's larger gate multiplier is necessary. For $R = 5$, Adam certifies $r_{0.99} = 16$, whereas GD's certificate remains zero at minimum training probability 0.5729. Each control uses seed 1 and 100,000 updates.

Larger-rate and small-rate checks. A prespecified sensitivity uses $\eta_t = 400/(t + 8000)$ from the same initialization for both optimizers, triggered when the selected GD candidate's largest family loss exceeds 0.1. It changes the step size as well as the cumulative rate, giving $S = 1041.10$ at 100,000 updates. GD reaches loss 0.01031 and minimum training probability 0.9772; its correct-answer probability on the same test with 64 stale records is 0.8781. Its balance residual in (16) is 1.4410, compared with the predicted limit 1.3540. Its lag-six correct probability remains 0.9868. This sequence has the witness form for eventual failure above $R+2$, but that failure has not appeared at this checkpoint. The theorem does not require failure on every sequence. Adam reaches loss 1.30×10^{-39} but certifies only $r_{0.99} = 7$. Its matching-to-forgetting ratio $m_t\Delta_{\min,t}/h_t$ is 6.088, below GD's 7.908; finite-time ordering of this ratio is therefore not uniform across the tested schedules.

A separate paired seed-0 reference uses the first row of Table 1 and $\eta_t = 0.01/(t + 1)$. After 100,000 updates its cumulative rate is only 0.12090, and both optimizers remain in early acquisition.

SGD with averaging over all key pairs. Two 4,096-update stochastic runs use the selected rates and $N_t = 16(t + 1)^4$ sampled lag/spacing tuples, with exact calibration and an average over all ordered distinct key pairs. Each update samples a multinomial histogram of the joint lag/spacing distribution; gradients use these random counts, rather than their expectations. This represents large logical batches without enumerating examples. The largest number of sampled tuples is 2^{52}, within exact integer representation; each tuple represents $2n(n - 1) = 24$ example losses. Both runs closely track the matched short GD trajectory at $S = 16.5382$, with loss 0.6059 and minimum training probability 0.4719. The matched Adam run has loss 0.1799. All four short runs have $r_{0.99} = 0$. These checks concern early dynamics of the specified estimator that averages over all key pairs. They do not test fixed small batches or establish that batch growth is necessary. The quartic schedule grows faster than the sufficient quadratic-logarithmic bound in Corollary 25. All 31 runs are retained: fourteen for rate selection and references, thirteen longer evaluation runs, and these four short checks.

G.3 A FULLY LEARNED TWO-LAYER TRANSFORMER

The structured model supplies positional routing and value-copy projections. To test the effect of learning those operations, we train a Transformer with eight keys, width 64, two attention layers, two heads, pre-layer root-mean-square (RMS) normalization, and no feedforward network. Each head has learned gains on RMS-normalized queries and keys. The first layer excludes self-attention whenever an earlier token exists and uses forgetting $\mathrm{softplus}(uv)$. Second-layer forgetting has one softplus parameter. A learned RMS normalization and full-vocabulary linear decoder produce the answer. Every parameter is trained from random initialization on consecutively serialized key–value tokens. Projection and embedding weights have Gaussian standard deviation 0.02; the initial forgetting penalties are 1 and 0.1 in the two layers.

Each online batch has 128 histories. Target lag is uniform on $1, \ldots, 4$; prefix and filler lengths are independently uniform on $0, \ldots, 4$. Prefix records independently use the query key with probability $1/2$, except that the target's immediate predecessor has another key. Newer records and fillers use other keys; values are random. Paired optimizers share initial weights and data sequences within each seed. Adam uses rate 10^{-3}, moments $(0.9, 0.95)$, epsilon 10^{-300}, and 200 warmup updates. SGD uses rate 0.3, gradient clipping at norm one, and 100 warmup updates. Both receive 20,000 float64 updates without weight decay. Their rates and warmup schedules differ, so this comparison does not isolate only the update rule.

Evaluation measures sampled classification accuracy on three fixed panels of 128 histories per lag: eight older off-key records, or 16 or 64 older records with stale-key probability $1/2$. Non-target values oppose the target. The panels are shared across optimizers and seeds. The empirical 90% range is the largest tested lag for which that lag and every smaller tested lag achieve at least 90% accuracy on all panels. The grid is $1, \ldots, 10, 12, 14, 16, 20, 24, 32$; zero denotes a failed lag-one panel. Table 2 compares the training predecessor restriction with evaluation that removes it. The latter gives no consistent Adam–SGD separation.

Table 2: Empirical 90% classification ranges for Transformer seeds 0, 1, and 2 after 20,000 updates. This finite-panel accuracy metric differs from the uniform correct-answer probability certificate in Figure 4.

Evaluation predecessors	Adam	SGD
Same restriction as training	$10, 10, 6$	$7, 6, 7$
Unrestricted	$5, 0, 5$	$5, 3, 5$