RECURRENT MEMORY REPAIR: WHEN LOCAL SUFFICIENCY FAILS AT FINITE BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how shared retention gates can improve prediction while limiting changes in the influence of past writes. For normalized recurrent memory with fixed candidate values and write strengths, we constrain changes in their allocation weights by Kullback–Leibler (KL) divergence or the operator norm of the sensitivity change. We construct a sigmoid-gate family with a shared bias whose best first-order loss decrease at the baseline matches that of freely adjustable allocations at equal Fisher cost. Yet we prove that its best loss decrease vanishes as sequences grow at a fixed sufficiently small positive budget. Adding the squared difference between consecutive candidates to the sigmoid input with one new shared coefficient yields a loss decrease bounded away from zero with length, at the same budget and from the same pre-update gates and allocation. For this construction, the fixed-task improvement survives all separately bounded candidate-value perturbations within a positive radius independent of length and budget, with pre- and post-update gate values held fixed. General results characterize how gate changes accumulate to consume the budget and which gate features realize prescribed allocation changes. In a 20-seed synthetic study with a frozen trained gate-generating network, quadratic augmentation attains larger calibration loss decreases in every seed than equally sized random augmentation at matched local task efficiency and mean KL cap.

1 INTRODUCTION

Retaining information from past inputs and learning to change its contribution to prediction pose different questions for recurrent memory models (Gu et al., 2020b; Zucchet & Orvieto, 2024). Our main analysis concerns normalized recurrent memory, whose output is the weighted sum $m_T(\theta) = \sum_i \pi_i(\theta) z_i$ of fixed bounded candidate values $z_{1:T}$. Here, we call the normalized weight vector $\pi(\theta)$ the allocation; its components satisfy

$$\pi_i(\theta) \propto \iota_i \prod_{k=i+1}^{T} f_k(\theta), \qquad \sum_i \pi_i(\theta) = 1.$$

Retention gates $f_t(\theta)$ are generated using the same parameter vector θ at every time step. The write strengths $\iota_i > 0$ are the coefficients multiplying candidate values when they are added to memory, and are held fixed here.

With gates fixed, perturbing z_i by δz_i changes the output by $\pi_i \delta z_i$. A gate-parameter update changes the allocation and hence these sensitivities. To study prediction improvement while controlling this change, we impose an allocation-KL or exact operator-norm budget. Local sufficiency means that a shared family's best first-order loss decrease matches that of freely adjustable allocations at equal Fisher cost derived from allocation KL (Section 2). Does this first-order match ensure a nonvanishing loss decrease as sequences grow under a fixed sensitivity budget?

To answer this question, we construct two alternating-sign scalar candidate sequences with different amplitudes, indexed by $a = 1, 2$. Let $v_{a,j} = z_{a,j+1} - z_{a,j}$ denote the difference between consecutive candidates. As our central construction separating local sufficiency from useful improvement under a finite budget, we use retention gates

$$f_{a,j+1}(\theta) = \mathrm{sigm}\left(u_*^{\mathrm{ret}} + \frac{\vartheta v_{a,j}}{1 - f_*} + \theta_{\mathrm{bias}} \right).$$

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.

Here $f_* = \mathrm{sigm}(u_*^{\mathrm{ret}})$ and the write strengths $\iota_i = f_*^{i-1}$ are fixed, and $\theta = (\vartheta, \theta_{\mathrm{bias}})$ is shared across time and the two sequences. For the constructed task, this difference term gives local sufficiency at $\theta = 0$. Nevertheless, we prove that, for this family, the best loss decrease tends to zero as sequence length grows at a fixed sufficiently small positive budget. By contrast, we use the squared difference between consecutive candidates, $v_{a,j}^2 = (z_{a,j+1} - z_{a,j})^2$, as an additional feature and add the quadratic term $\theta_{\mathrm{sq}} v_{a,j}^2$ to the sigmoid input, where the new coefficient θ_{sq} is shared across time and the two sequences. This yields a loss decrease bounded away from zero as sequence length grows, at the same budget and while preserving identical pre-update gates and allocation (Section 3). With each time step's pre- and post-update gate values, write strengths, and prediction target fixed, this improvement survives all per-write bounded candidate-value perturbations within a positive radius independent of length and budget (Corollary 3.2).

Sections 4–5 explain how changes in retention-gate values induced by a shared-parameter update compound along the sequence and consume the sensitivity budget, and characterize the features needed in the gate-generating rule to realize prescribed allocation changes. In synthetic experiments with a frozen trained network that generates retention gates, quadratic augmentation attains greater calibration loss decreases than equal-width random augmentation in all 20 seeds at matched local task efficiency and mean allocation-KL cap (Section 6).

2 Memory and sensitivity-constrained updates

Normalized recurrent memory. Fix $T \geq 2$, $d \geq 1$, $R > 0$, candidate values $z_t \in \mathbb{R}^d$ with $\|z_t\| \leq R$, and positive write strengths ι_t. With retention gates f_t, cumulative-normalized memory (CNM) is

$$c_t = f_t c_{t-1} + \iota_t z_t, \quad n_t = f_t n_{t-1} + \iota_t, \quad m_t = c_t / n_t, \quad c_0 = n_0 = 0, \quad f_t > 0. \tag{1}$$

Here c_t accumulates candidate values and n_t their total weight; their normalized ratio is the bounded output. The accumulators themselves need not be bounded. Equation (1) yields $m_1 = z_1$ and, for $t \geq 2$,

$$m_t = (1 - \gamma_t) m_{t-1} + \gamma_t z_t, \qquad \gamma_t = \iota_t / n_t \in (0, 1). \tag{2}$$

Conversely, any prescribed $(\gamma_t)_{t=2}^T \in (0,1)^{T-1}$ is realized by suitable positive gates and write strengths, for every candidate sequence and every time. This does not identify prescribed shared families (Appendix A.1). For the prefix $z_{1:t} = (z_1, \ldots, z_t)$, unrolling gives the allocation over indices $1, \ldots, t$:

$$w_i^{(t)} \coloneqq \iota_i \prod_{k=i+1}^{t} f_k, \qquad p_i^{(t)} \coloneqq \frac{w_i^{(t)}}{\sum_{j=1}^t w_j^{(t)}}, \qquad m_t = Z_t p^{(t)}, \quad Z_t \coloneqq [z_1, \ldots, z_t]. \tag{3}$$

The empty product is one. Thus the allocation records the cumulative action of the retention gates on individual writes. Set $Z \coloneqq Z_T$, $\pi \coloneqq p^{(T)}$, $L_T(p) \coloneqq L_T^{\mathrm{cons}}(p) = \|Zp - y_T\|^2/2$ for fixed output target y_T; $\|x\| = \|x\|_2$ for vectors. Baseline gates f^0, allocation $p^0 = \pi^0$ and output m^0 are pre-update quantities, not necessarily from an untrained model; $+$ denotes post-update quantities. Appendix A.1 covers weighted observations. The normalized scalar accumulation in sLSTM and Forgetting Attention at a fixed terminal query and fixed keys provide conditional correspondences (Beck et al., 2024; Lin et al., 2025); Appendix N.2 specifies the computation paths and fixed quantities.

Conditional sensitivity and budgets. For $\mathbf{z} = (z_i)_{i=1}^T$, set $A_p \mathbf{z} \coloneqq \sum_i p_i z_i$. During a gate-parameter update, candidates and write strengths stay fixed while the induced allocation changes. For subsequent candidate-value perturbations, keep each time step's pre- and post-update gate values f^0, f^+, the write strengths and prediction target y_T fixed, comparing

$$m_T^{0,+}(\mathbf{z} + \delta\mathbf{z}) = \sum_i p_i^{0,+}(z_i + \delta z_i), \qquad \delta\mathbf{z} \in \mathcal{E}_{\varepsilon_z} \coloneqq \{e \in (\mathbb{R}^d)^T : \max_i \|e_i\|_2 \leq \varepsilon_z\}.$$

Both outputs use the same perturbed candidates, without recomputing input-dependent gate values. The update effect $E_{\mathrm{upd}}(\mathbf{z}) \coloneqq (A_{p^+} - A_{p^0})\mathbf{z}$ changes by $(A_{p^+} - A_{p^0})\delta\mathbf{z}$. With $\|\delta\mathbf{z}\|_{\infty,2} \coloneqq \max_i \|\delta z_i\|_2$, norm duality gives the exact identity (Appendix A.2), while Pinsker's inequality (Sason & Verdú, 2016) gives the KL bound:

$$\sup_{\|\delta\mathbf{z}\|_{\infty,2} \leq \varepsilon_z} \|(A_p - A_q)\delta\mathbf{z}\|_2 = \varepsilon_z \|p - q\|_1 \leq \varepsilon_z \sqrt{2 D_{\mathrm{KL}}(p\|q)}. \tag{4}$$

We impose $D_{\mathrm{KL}}(p^+\|p^0) \le \tau$ or $\|p^+ - p^0\|_1 \le \delta$, where $D_{\mathrm{KL}}(p\|q) := \sum_i p_i \log(p_i/q_i)$ on a common positive support. These allocation-KL and operator budgets control conditional sensitivity changes. Costs compare the two allocations, not parameter-curve length or individual operator norms. The choices in $\mathcal{E}_{\varepsilon_z}$ are separate by index, not probabilistically independent; perturbed values may leave the original ball. Loss bounds also account for the perturbed baseline residual (Appendix A.7).

Log-retentions and local response. Set $b_j := \log f_{j+1}$, $a_i^w := \log \iota_i$, and $H_{ij} := \mathbf{1}\{i \le j\}$ for $H \in \mathbb{R}^{T\times(T-1)}$. Then $\pi = \mathrm{softmax}(a^w + Hb)$ and

$$C_\pi := \mathrm{Diag}(\pi) - \pi\pi^\top, \qquad J_b = ZC_\pi H, \qquad F_b := H^\top C_\pi H \succ 0. \tag{5}$$

Here $J_b := D_b m_T$ and F_b is the Fisher information of an indexed write drawn from $\pi(b)$, not predictive-label Fisher information or an empirical matrix of observed loss-gradient outer products (Martens, 2020, Sections 5 and 11). For differentiable sharing, put $U_0 := D_\theta b(\theta^0)$ and

$$F_\theta := U_0^\top F_b U_0, \qquad K_U := J_b U_0 F_\theta^\dagger U_0^\top J_b^\top. \tag{6}$$

Here $\dagger$ is the Moore–Penrose inverse and $\ker F_\theta = \ker U_0$. Let $g \in \mathbb{R}^d$ denote a task direction; for the squared loss, $g = Z\pi - y_T$ at the baseline. At local Fisher radius r_{loc}, the best first-order decrease is $r_{\mathrm{loc}}\sqrt{g^\top K_U g}$. Freely adjustable allocations attain $r_{\mathrm{loc}}\sqrt{g^\top \Sigma_\pi g}$ at the same radius, where $\Sigma_\pi := ZC_\pi Z^\top$. The ratio $g^\top K_U g/(g^\top \Sigma_\pi g)$, for positive denominator, is the local task efficiency; its square root is the response ratio. Efficiency one is local sufficiency for that task; $K_U = \Sigma_\pi$ is the stronger equality for all task directions.

Finite realization under affine log-retention sharing. For $h = Z^\top g$, define $(\Delta h)_j := h_j - h_{j+1}$. The identities $\Delta H = I$ and $H\Delta h = h - h_T\mathbf{1}$ give

$$q_{\lambda,i} := \frac{\pi_i e^{-\lambda h_i}}{\sum_j \pi_j e^{-\lambda h_j}}, \qquad b^+ = b - \lambda\Delta h. \tag{7}$$

Appendix A.3 also proves $J_b F_b^{-1} J_b^\top = \Sigma_\pi$.

Theorem 2.1 (Affine shared closure)**.** *For $b = b^0 + U\theta$ with unrestricted θ, the following are equivalent: $K_U = \Sigma_\pi$ at one positive allocation; $\mathrm{range}(\Delta Z^\top) \subseteq \mathrm{range}\, U$; and realization in the same family of every finite entropic tilt. The minimum number of arbitrary added columns is $\mathrm{rank}[U, \Delta Z^\top] - \mathrm{rank}\, U$.*

Natural-gradient projection and exponential tilting combine with the recurrent inverse (van Oostrum et al., 2023; Kivinen & Warmuth, 1997; Wainwright & Jordan, 2008). For fixed positive-variance g, taskwise equality holds iff $-\Delta Z^\top g \in \mathrm{range}\, U$, realizing its full tilt curve (Appendix A.4). Gates $f < 1$ additionally require $b^+ < 0$. Appendix I treats joint updates of retention coefficients and a shared query-generating matrix at multiple times, with keys fixed.

3 A FINITE-RESPONSE BARRIER AND A SAME-BASELINE REPAIR

3.1 TWO AMPLITUDES WITH AN EXISTING SHARED BIAS

Section 2 links taskwise local sufficiency and finite entropic realization for affine log-retentions. We now address the Introduction's question with a separation construction: sharing is affine in gate logits but nonlinear in log-retentions. Let $\mathrm{sigm}(u) := (1+e^{-u})^{-1}$, $\ell_{\mathrm{sig}} := \log\mathrm{sigm}$, and $\mathrm{logit}(x) := \log\{x/(1-x)\}$. Fix $T = 2n \ge 4$ and candidate amplitudes $0 < c_1 < c_2 \le 1$. For sequences $a \in \{1,2\}$, set $z_{a,i} = c_a(-1)^i$, $v_{a,j} = z_{a,j+1} - z_{a,j}$, and write strengths $\iota_i = f_*^{i-1}$ with $f_* \in (0,1)$. The baseline gate logit $u_*^{\mathrm{ret}} = \mathrm{logit}\, f_*$ gives uniform allocations $\mathcal{U}_T$ and zero means in both sequences, where $\mathcal{U}_N$ denotes the uniform law on N indices. Here $i = 1,\dots,T$ indexes writes and $j = 1,\dots,T-1$ the gate at time $j+1$, with $u_{a,j}^{\mathrm{ret}} = \mathrm{logit}\, f_{a,j+1}$. Real difference, bias, and square coefficients are shared across all gates in both sequences; u_*^{ret} is fixed. Compare

$$\begin{aligned} \mathcal{F}_{\mathrm{bias}}: &\quad u_{a,j}^{\mathrm{ret}} = u_*^{\mathrm{ret}} + \frac{\vartheta v_{a,j}}{1-f_*} + \theta_{\mathrm{bias}}, \\ \mathcal{F}_{\mathrm{sq}}: &\quad u_{a,j}^{\mathrm{ret}} = u_*^{\mathrm{ret}} + \frac{\vartheta v_{a,j}}{1-f_*} + \theta_{\mathrm{bias}} + \theta_{\mathrm{sq}} v_{a,j}^2. \end{aligned} \tag{8}$$

The real coefficient vectors are $\theta = (\vartheta, \theta_{\text{bias}}) \in \Theta_{\text{bias}} \coloneqq \mathbb{R}^2$ and $\theta = (\vartheta, \theta_{\text{bias}}, \theta_{\text{sq}}) \in \Theta_{\text{sq}} \coloneqq \mathbb{R}^3$. Gates $f_{a,j+1}(\theta) = \text{sigm}(u^{\text{ret}}_{a,j}(\theta))$ induce $p_a(\theta)$ by (3) and $m_{a,T}(\theta) = \sum_i p_{a,i}(\theta) z_{a,i}$, with $p_a(0) = \mathcal{U}_T$, $m_{a,T}(0) = 0$. Same-baseline repair starts both families at these identical gates, allocation and output; their updated allocations may differ. The scaling gives $\partial_\vartheta \log f_{a,j+1}|_0 = v_{a,j}$, the adjacent-difference direction. Define the linear-output gain $G(\theta) \coloneqq -\frac{1}{2}\sum_a m_{a,T}(\theta)$ and mean costs over the two sequences

$$\mathcal{K}_T(\theta) \coloneqq \frac{1}{2}\sum_a D_{\text{KL}}(p_a(\theta)\|\mathcal{U}_T), \qquad \mathcal{V}_T(\theta) \coloneqq \frac{1}{2}\sum_a \|p_a(\theta) - \mathcal{U}_T\|_1.$$

For each family, let $B^{\text{KL}}_{\mathcal{F}}(T,\tau) \coloneqq \sup_{\mathcal{K}_T \le \tau} G$ and $B^{\text{op}}_{\mathcal{F}}(T,\delta) \coloneqq \sup_{\mathcal{V}_T \le \delta} G$ over its entire coefficient space. These are best linear gains; Section 3.2 connects them to the fixed squared-loss task used in the Introduction.

Theorem 3.1 (Minimal repair with a shared bias)**.** *The common-gradient task in $\mathcal{F}_{\text{bias}}$ has unit local task efficiency and a rank-two two-output Jacobian. Uniformly over even $T \ge 4$, sufficiently small positive budgets, and compact structural sets separated from gate endpoints, zero amplitude, and equal amplitudes,*

$$B^{\text{KL}}_{\text{bias}} \asymp \min\{\sqrt{\tau}, \tau^{1/4}/\sqrt{T}\}, \quad B^{\text{KL}}_{\text{sq}} \asymp \sqrt{\tau},$$
$$B^{\text{op}}_{\text{bias}} \asymp \min\{\delta, \sqrt{\delta/T}\}, \qquad B^{\text{op}}_{\text{sq}} \asymp \delta.$$

Here $\asymp$ denotes two-sided bounds up to positive constant factors independent of T and the budget. For every fixed finite $\tau > 0$, the original family's KL response is $\Theta_{\tau,f_*,c_1,c_2}(T^{-1/2})$ as $T \to \infty$. The added squared-difference feature attains length-independent response with bounded coefficients. One addition is minimal for nonvanishing response in this construction.

Why the bias is insufficient. Bias supplies $(1,1)$, while quadratic drift follows (c_1^2, c_2^2). Two sigmoid-product identities localize all low-drift controls, including unbounded sequences, before the quadratic expansion. Square repairs both products exactly (Appendices C.3.1–C.3.2). Local equality is taskwise; minimum added dimension does not make the square formula unique (Appendix L.2).

3.2 A FIXED TASK UNDER CANDIDATE-VALUE PERTURBATIONS

To express the separation as prediction improvement, fix target -1 in both sequences. The nominal squared-loss decrease is $G - \frac{1}{4}\sum_a m^2_{a,T} \le G$, and is at least $G/2$ on the repair curve. Thus, at fixed small positive budgets, the original family's best loss decrease vanishes with length, whereas the repaired curve has a positive length-independent lower bound (Appendix C.3.1). Choose the gate update using the original candidate sequences. During candidate-value perturbations, keep every time step's pre- and post-update retention-gate values, the write strengths and prediction target fixed; do not recompute the difference features in (8). Put $m^\theta_{a,T}(\delta\mathbf{z}) \coloneqq \sum_i p_{a,i}(\theta)(z_{a,i} + \delta z_{a,i})$ and

$$\mathcal{D}^{\text{pert}}_T(\theta, \delta\mathbf{z}) \coloneqq \frac{1}{4}\sum_a \left[(1 + m^0_{a,T}(\delta\mathbf{z}))^2 - (1 + m^\theta_{a,T}(\delta\mathbf{z}))^2\right].$$

Using the original mean costs $\mathcal{K}_T, \mathcal{V}_T$, define

$$B^{\text{pert,KL}}_{\mathcal{F}}(T,\tau;\varepsilon_z) \coloneqq \sup_{\theta\in\Theta_{\mathcal{F}}:\mathcal{K}_T(\theta)\le\tau} \inf_{\max_{a,i}|\delta z_{a,i}|\le\varepsilon_z} \mathcal{D}^{\text{pert}}_T(\theta, \delta\mathbf{z}),$$

and define $B^{\text{pert,op}}_{\mathcal{F}}$ using $\mathcal{V}_T \le \delta$. Here $\Theta_{\mathcal{F}}$ denotes the corresponding coefficient space: $\Theta_{\text{bias}} = \mathbb{R}^2$ for $\mathcal{F}_{\text{bias}}$ and $\Theta_{\text{sq}} = \mathbb{R}^3$ for $\mathcal{F}_{\text{sq}}$ (Section 3.1).

Corollary 3.2 (A nonshrinking candidate-value perturbation radius)**.** *In the compact structural range of Theorem 3.1, let both amplitudes be at least $c_{\min} > 0$. Uniformly in even $T \ge 4$, sufficiently small positive budgets, and $0 \le \varepsilon_z \le c_{\min}/16$, positive constants C_-, C_+ give*

$$0 \le B^{\text{pert,KL}}_{\text{bias}} \le C_+ \min\{\sqrt{\tau}, \tau^{1/4}/\sqrt{T}\},$$
$$C_-\sqrt{\tau} \le B^{\text{pert,KL}}_{\text{sq}} \le C_+\sqrt{\tau}.$$

For the operator budget, $B^{\text{pert,op}}_{\text{bias}} \le C_+ \min\{\delta, \sqrt{\delta/T}\}$ and $C_-\delta \le B^{\text{pert,op}}_{\text{sq}} \le C_+\delta$. The perturbation radius is independent of length and budget. No matching perturbed-loss lower rate is asserted for the original family.

The explicit repair satisfies $\|p_a - \mathcal{U}_T\|_1 = -m_{a,T}/c_a$, linking its change in write influence to its linear-output gain. Figure 1 evaluates specific updates; Corollary 3.2 bounds the best value over each budget-feasible family. Appendices L.1 and L.4 give evaluation details, Appendix C.4 proves the fixed repair's $G/4$ bound, and Table 4 reports additional comparisons.

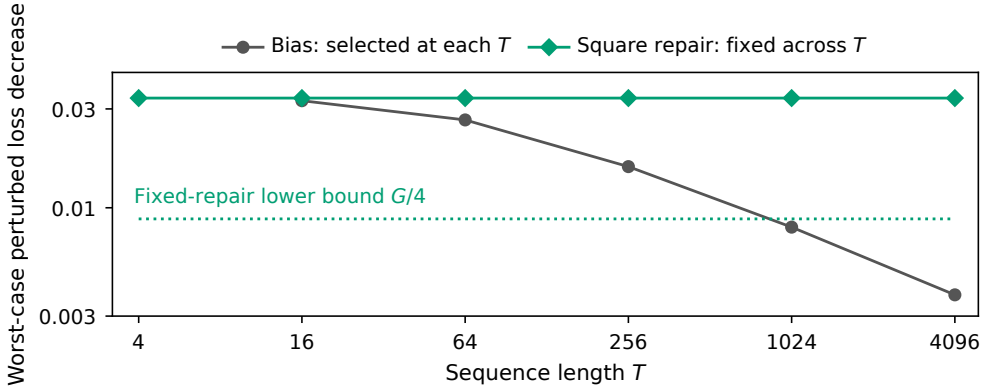


Figure 1: **One fixed repair across lengths.** Bias ($\mathcal{F}_{\text{bias}}$) selects coefficients by nominal (unperturbed) loss at each T; Square ($\mathcal{F}_{\text{sq}}$) reuses its entire $T = 4$ repair coefficient vector at every shown length. Markers show $\inf_{\max_{a,i}|\delta z_{a,i}|\le\varepsilon_z} \mathcal{D}_T^{\text{pert}}(\theta, \delta\mathbf{z})$ at these fixed coefficients: the squared-loss decrease averaged over both sequences and minimized over perturbations. Pre- and post-update gates, write strengths and target -1 stay fixed during perturbations. Candidate amplitudes $(c_1, c_2) = (0.5, 1)$, baseline gate $f_* = 0.5$, and radius $\varepsilon_z = 0.015625$. All updates satisfy $\mathcal{K}_T(\theta) \le \tau = 10^{-3}$, a mean allocation-KL cap over the two sequences. The dotted line is the fixed Square repair's lower bound $G/4$ (Appendix C.4).

4 Cumulative Drift: Mechanism, Global Bounds, and Extensions

4.1 Repeated Blocks and the Finite-Response Laws

To isolate the recurrent accumulation mechanism of Section 3, take finitely many sequences a with fixed positive weights w_a^{seq} summing to one. Each repeats L_{blk} candidates $z_{a,\ell}$ for n blocks, $T = nL_{\text{blk}}$. Let $j = 0, \dots, n-1$ index blocks and $\ell = 1, \dots, L_{\text{blk}}$ their positions (phases). The baseline is $\mathcal{U}_n \otimes \nu_a^0$ with positive phase law ν_a^0. At $i = jL_{\text{blk}} + \ell < T$, the repeated increment $d_{a,\ell}(\theta) = \log[f_{a,i+1}(\theta)/f_{a,i+1}^0]$ is independent of j and zero at $\theta = 0$. Candidate values and write strengths stay fixed; no gate follows the terminal write. Define

$$\beta_a := -\sum_{\ell=1}^{L_{\text{blk}}} d_{a,\ell}, \quad s_{a,\ell}^{\text{phase}} := \sum_{k=\ell}^{L_{\text{blk}}-1} d_{a,k}, \quad \nu_{a,\theta} := \text{softmax}(\log \nu_a^0 + s_a^{\text{phase}}). \tag{9}$$

Write $\beta(\theta) = (\beta_a(\theta))_a$. The last phase score is zero. Unrolling gives, for chronological block position $j = 0, \dots, n-1$,

$$p_{a,n,\theta}(j, \ell) = q_{n,\beta_a}^{\text{time}}(j)\nu_{a,\theta}(\ell), \qquad q_{n,\beta}^{\text{time}}(j) = \frac{e^{\beta j}}{\sum_{k=0}^{n-1} e^{\beta k}}. \tag{10}$$

Here $q_{n,\beta_a}^{\text{time}}$ is the time marginal over blocks j, and $\nu_{a,\theta}$ the phase marginal over positions ℓ. The drift β_a is the deterministic log-weight slope across blocks, not stochastic state drift. For this repeated-block model, write $Z_a = [z_{a,1}, \dots, z_{a,L_{\text{blk}}}]$ and set $G(\theta) = -\sum_a w_a^{\text{seq}} g_a^\top Z_a(\nu_{a,\theta} - \nu_a^0)$ and $\mathcal{K}_{\text{phase}} = \sum_a w_a^{\text{seq}} D_{\text{KL}}(\nu_{a,\theta} \| \nu_a^0)$ for fixed task directions g_a. Let $\mathcal{K}_n$ and $\mathcal{V}_n$ be the weighted KL and ℓ_1 distances from $\mathcal{U}_n \otimes \nu_a^0$, and write $B_n^{\text{KL}}(\tau) = \sup_{\mathcal{K}_n \le \tau} G$ and $B_n^{\text{op}}(\delta) = \sup_{\mathcal{V}_n \le \delta} G$. The operator cost permits separately chosen perturbations for each sequence. Exactly,

$$\mathcal{K}_n = \mathcal{K}_{\text{phase}} + \sum_a w_a^{\text{seq}} D_n^{\text{KL}}(\beta_a), \qquad D_n^{\text{KL}}(\beta) := D_{\text{KL}}(q_{n,\beta}^{\text{time}} \| \mathcal{U}_n). \tag{11}$$

The output $Z_a \nu_{a,\theta}$ depends only on the phase marginal; the time marginal contributes $D_n^{\mathrm{KL}}(\beta_a)$ to that sequence's allocation KL.

Theorem 4.1 (Zero-drift characterization). *Suppose the control domain Θ is compact, contains the baseline $\theta = 0$, and all increments are continuous. For every fixed finite $\tau \geq 0$,*

$$\lim_{n\to\infty} B_n^{\mathrm{KL}}(\tau) = \max\{G(\theta) : \theta \in \Theta,\ \beta(\theta) = 0,\ \mathcal{K}_{\mathrm{phase}}(\theta) \leq \tau\}. \tag{12}$$

At fixed τ, $\mathcal{K}_n \leq \tau$ implies $D_n^{\mathrm{KL}}(\beta_a) \leq \tau / w_a^{\mathrm{seq}}$ by (11), hence $\|\beta\|_\infty = O(n^{-1})$ as $n \to \infty$. Compactness gives a convergent subsequence of maximizing controls (Appendices B.2–B.3). This compact-domain result is not used to impose compactness on the unrestricted sigmoid parameters.

Theorem 4.2 (Drift-order transfer under two budgets). *Suppose some $\epsilon_0, c > 0$ and integer $r \geq 1$ satisfy the following: every control with $G \geq 0$, $\max_a \|\nu_{a,\theta} - \nu_a^0\|_1 \leq \epsilon_0$, and $\|\beta(\theta)\|_\infty \leq \epsilon_0$ obeys*

$$\|\beta(\theta)\|_\infty \geq c G(\theta)^r. \tag{13}$$

Suppose $c_0, C_0, \vartheta_0 > 0$ and a curve $\theta(\vartheta)$, $0 \leq \vartheta \leq \vartheta_0$, satisfy $G \geq c_0 \vartheta$, $\mathcal{K}_{\mathrm{phase}} \leq C_0 \vartheta^2$, and $\|\beta\|_\infty \leq C_0 \vartheta^r$. Uniformly over $n \geq 2$ and sufficiently small positive budgets,

$$B_n^{\mathrm{KL}}(\tau) \asymp \min\{\sqrt{\tau}, \tau^{1/(2r)} n^{-1/r}\}, \qquad B_n^{\mathrm{op}}(\delta) \asymp \min\{\delta, (\delta/n)^{1/r}\}. \tag{14}$$

Constants are independent of length and budget. Actual feasibility gives $n\|\beta\|_\infty \lesssim \sqrt{\tau}$ or $\lesssim \delta$, where $\lesssim$ denotes an upper bound up to a positive constant factor. Equation (13) yields the upper bounds and the curve attains them (Appendix B.4). The drift inequality (13) must hold for every θ satisfying its preceding constraints on G, $\nu_{a,\theta}$ and β, however far from zero: the local quadratic drift $\beta(\vartheta) = \vartheta^2(1 - \vartheta/\varepsilon_{\mathrm{esc}})$ has a useful nonzero zero-drift endpoint. Appendix B.6 gives the recurrent counterexample; Appendix B.5 states when localization makes a reduced local jet sufficient.

4.2 A NONPERIODIC SIGN-BALANCED FAMILY

Theorem 4.3 (Nonperiodic sign-balanced pairs). *Let $T = 2n \geq 4$, choose arbitrary $c_k \in [c_{\min}, c_{\max}] \subset (0, 1]$ for $k = 1, \ldots, n$, and set $z_{2k-1} = -c_k$, $z_{2k} = c_k$, $\iota_i = f_*^{i-1}$ for $i = 1, \ldots, T$, and $u_*^{\mathrm{ret}} = \operatorname{logit} f_*$. For gate indices $j = 1, \ldots, T-1$, use*

$$v_j = z_{j+1} - z_j, \qquad u_j^{\mathrm{ret}} = u_*^{\mathrm{ret}} + \vartheta v_j / (1 - f_*), \qquad \vartheta \in \mathbb{R}. \tag{15}$$

The baseline allocation is uniform with zero mean. Its scalar allocation-Fisher response coefficient is $K = \mathrm{Var}_{\mathcal{U}_T}(z)$. Define $B^{\mathrm{KL}}(T, \tau) = \sup_{\vartheta : D_{\mathrm{KL}}(\pi(\vartheta)\|\mathcal{U}_T) \leq \tau} [-m_T(\vartheta)]$, and define B^{op} with the ℓ_1 budget. Then

$$B^{\mathrm{KL}}(T, \tau) \asymp \min\{\sqrt{\tau}, \tau^{1/4}/\sqrt{T}\}, \qquad B^{\mathrm{op}}(T, \delta) \asymp \min\{\delta, \sqrt{\delta/T}\}. \tag{16}$$

The comparisons are uniform over the amplitude arrays, even $T \geq 4$, sufficiently small positive budgets, and f_* in a fixed compact subset of $(0, 1)$. Allowing the nonlinear inverse rule $u_j^{\mathrm{ret}} = u_*^{\mathrm{ret}} + \psi(\lambda v_j)$, where $\psi(a) = \operatorname{logit}(f_* e^a) - u_*^{\mathrm{ret}}$, on the admissible neighborhood $|\lambda| \|v\|_\infty < -\log f_*$, instead attains rates $\sqrt{\tau}$ and δ uniformly in T.

Conditional sign probabilities localize all positive-gain small-budget controls; concavity then yields nonnegative gate errors accumulating to $T\vartheta^2 \lesssim \sqrt{\tau}$ or $T\vartheta^2 \lesssim \delta$ (Appendix C.1). Sign pairing and the amplitude lower bound remain essential. The nonlinear inverse changes the family; this is not one-square-feature repair for arbitrary amplitudes.

4.3 CROSSOVER AND REFERENCE ALLOCATIONS

For $c_k \equiv 1$, $f_* = 1/2$, compare (15), $u_j^{\mathrm{ret}} = 2\vartheta v_j$, with its bias extension $u_j^{\mathrm{ret}} = 2\vartheta v_j + \theta_{\mathrm{bias}}$. For either budget, the ratio of the original to bias-extended optimal gains tends to one as their common cap vanishes at fixed T, but to zero as $T \to \infty$ at fixed sufficiently small positive cap. Appendix C.2 gives the construction, exact optima, and Figure 4. Unlike the two-amplitude setting of Section 3, a shared bias suffices in this specialization. Appendix H treats nonuniform reference allocations under additional spread and tail conditions; changing that reference is distinct from repair at the same baseline.

5 FINITE REALIZATION AND BUDGET-FEASIBLE IMPROVEMENT

5.1 WHAT FIXED FEATURES MUST REALIZE

We now ask which fixed features in the gate-generating rule can produce a prescribed task-directed allocation curve. For fixed candidate values, positive write strengths, and a fixed task direction g, the target allocation $q_{\lambda,i} \propto \pi_i^0 e^{-\lambda z_i^\top g}$ uniquely requires $b^0 + \lambda v$, where $v = -\Delta Z^\top g$. Realizing this entire curve is stronger than repairing a useful response. Take pre-update retention gates $f_{j+1}^0 = f_* \in (0,1)$ constant over time (not necessarily a uniform allocation), and declare

$$u^{\mathrm{ret}} = u_*^{\mathrm{ret}} \mathbf{1} + X_{\mathrm{ret}}\theta, \qquad S_{\mathrm{ret}} \coloneqq \operatorname{range} X_{\mathrm{ret}}. \tag{17}$$

Rows of fixed $X_{\mathrm{ret}} \in \mathbb{R}^{(T-1)\times d_\theta}$ index gates, columns features; $\theta \in \mathbb{R}^{d_\theta}$ is shared across rows. Although $U_0 = (1-f_*)X_{\mathrm{ret}}$ at the baseline, $b = \ell_{\mathrm{sig}}(u^{\mathrm{ret}})$ is nonlinear. The unique required logit increment is componentwise $\psi(\lambda v)$, analytic for $|\lambda|\|v\|_\infty < -\log f_*$, where

$$\begin{aligned} \psi(a) &\coloneqq a - \log(1 - f_* e^a) + \log(1 - f_*) = \sum_{k\geq 1} \psi_k a^k, \\ \psi_1 &= \frac{1}{1-f_*}, \qquad \psi_2 = \frac{f_*}{2(1-f_*)^2}. \end{aligned} \tag{18}$$

All $\psi_k > 0$; powers are componentwise. For integer $q \geq 1$, set $X_{\mathrm{pow},q} \coloneqq [v, v^{\odot 2}, \ldots, v^{\odot q}]$. Order-$q$ matching means $X_{\mathrm{ret}}\theta(\lambda) - \psi(\lambda v) = O(|\lambda|^{q+1})$ as $\lambda \to 0$ for a smooth coefficient curve in (17).

Theorem 5.1 (Exact and finite-order feature closure)**.** *Exact realization for every sufficiently small signed λ holds iff*

$$\operatorname{span}\{v^{\odot k} : k \geq 1\} \subseteq S_{\mathrm{ret}}. \tag{19}$$

The power span $\operatorname{span}\{v^{\odot k} : k \geq 1\}$ has dimension equal to the number of distinct nonzero values in v. Order-q matching holds iff $\operatorname{range} X_{\mathrm{pow,q}} \subseteq S_{\mathrm{ret}}$. With unrestricted added feature columns, their minimum number is

$$k_q^{\mathrm{ret}} = \operatorname{rank}[X_{\mathrm{ret}}, X_{\mathrm{pow,q}}] - \operatorname{rank} X_{\mathrm{ret}}. \tag{20}$$

These counts measure the additional feature columns needed on the fixed evaluation data. Analytic inversion and Vandermonde rank give the formulas (Appendix D.2; related rank analysis: Madden & Thrampoulidis, 2024). Lemma D.1 gives nonuniform-gate templates used in Section 6.

Stacking Section 3's two sequences yields four distinct nonzero difference values. Their three-dimensional $\operatorname{span}\{\mathbf{1}, v, v^{\odot 2}\}$ repairs response but not the whole entropic curve. Allowing nonlinear $\psi((z_{j+1} - z_j)^\top \zeta)$, $\zeta \in \mathbb{R}^d$, bypasses that linear-feature dimension (Appendix D.4).

5.2 ACTUAL-BUDGET LOSS REDUCTION

We now certify actual budget and squared-loss decrease at a realizable endpoint. At a positive baseline allocation π^0, set $g = Z\pi^0 - y_T$, $h = Z^\top g$, $v = -\Delta h$, $V_h = \operatorname{Var}_{\pi^0}(h) > 0$, and $\Omega_h = \operatorname{osc}(h)$, where $\operatorname{osc}(x) \coloneqq \max_i x_i - \min_i x_i$. For a log-retention vector b_p, let $p = \operatorname{softmax}(a^w + Hb_p)$ be its induced allocation. Define

$$\varepsilon_{\mathrm{sc}}(p,\lambda) \coloneqq \operatorname{osc}(\log p - \log q_\lambda) = \operatorname{osc}\{H(b_p - b^0 - \lambda v)\}. \tag{21}$$

Theorem 5.2 (Actual-budget finite descent)**.** *If $0 < \lambda \leq R^{-2}$, $\lambda\Omega_h \leq 1$, and $\varepsilon_{\mathrm{sc}}(p,\lambda)^2/8 \leq V_h\lambda^2/(4\mathrm{e})$, then*

$$L_T(\pi^0) - L_T(p) \geq V_h\lambda/(4\mathrm{e}) > 0. \tag{22}$$

The actual finite costs satisfy

$$D_{\mathrm{KL}}(p\|\pi^0) \leq \frac{\{\lambda\Omega_h + \varepsilon_{\mathrm{sc}}(p,\lambda)\}^2}{8}, \qquad \|p - \pi^0\|_1 \leq \frac{\lambda\Omega_h + \varepsilon_{\mathrm{sc}}(p,\lambda)}{2}. \tag{23}$$

Together with the descent hypotheses, $\lambda\Omega_h + \varepsilon_{\mathrm{sc}} \leq \sqrt{8\tau}$ or $\leq 2\delta$ certifies the respective cost. Appendix E.2 uses the directed entropic identity and relative smoothness (Bauschke et al., 2017; Lu et al., 2018). This tilt is a linearized-task step, not generally the loss optimum at that cap. Length uniformity needs control of accumulated interval errors, gate margins, value bounds and task variance, not Taylor degree alone (Appendix E.5). Appendix D.5 interprets the local target-KL coefficient.

6 EMPIRICAL EVIDENCE: ATTAINED UPDATES AND CANDIDATE-VALUE PERTURBATIONS

6.1 CONTROLLED SETUP AND UPDATE FAMILIES

We compare finite loss decreases from equal-width Random and Quadratic updates with unit local task efficiency. A nine-input, eight-hidden-unit $\tanh$ multilayer perceptron (MLP) generates scalar retention logits (89 parameters). For $t \geq 2$, its context is

$$\bar{z}_t = \frac{1}{t}\sum_{i=1}^{t} z_i, \qquad x_t^{\mathrm{ctx}} = \left(z_t^\top, z_{t-1}^\top, \bar{z}_t^\top, x_{t,1}^{\mathrm{meta}}, x_{t,2}^{\mathrm{meta}}, (t-1)/(T-1)\right)^\top \in \mathbb{R}^9.$$

Here $z_i \in \mathbb{R}^2$, and the two auxiliary scalars are independent $U[-1,1]$ draws. The controller receives $\bar{z}_t$, not m_t, and the known length T. After 400 Adam updates (Kingma & Ba, 2015) matching teacher memory outputs at $t = 2, \ldots, 64$, we freeze all 89 parameters. Appendix K.1 specifies the teacher, initialization and fixed logit offset.

For each of 20 prespecified seeds, set $y_{a,T} = m_{a,T}^0 - g_s$ with one $\|g_s\| = 0.25$ across sequences, giving pre-update loss 0.03125 at the chosen checkpoint. Family A uses calibration-RMS-scaled fixed features $X_{A,a}$ and coefficients θ_A shared across gates and sequences: $u_{A,a}^{\mathrm{ret},+} = u_a^{\mathrm{ret},0} + X_{A,a}\theta_A$. Sigmoid gates and (3) induce $p_{A,a}(\theta_A)$. For $N_{\mathrm{cal}} = 16$, fit

$$\min_{\theta_A} \frac{1}{2N_{\mathrm{cal}}} \sum_a \|Z_a p_{A,a}(\theta_A) - y_{a,T}\|^2, \qquad \frac{1}{N_{\mathrm{cal}}} \sum_a D_{\mathrm{KL}}(p_{A,a}(\theta_A) \| p_a^0) \leq 10^{-3}.$$

Calibration means coefficient fitting, not probability calibration; candidates, write strengths and controller parameters stay fixed. Gate head uses eight frozen hidden activations and a constant (9 coefficients). First-order adds two inverse-gate linear columns (11). Random and Quadratic each retain these eleven columns and add three sine or inverse-gate quadratic columns, respectively (14 each). Appendix K.3 gives the feature and RMS formulas; Lemma D.1 derives the inverse-gate columns. Features exclude targets and g_s; First-order already contains the task tangent, giving both augmentations unit local task efficiency. Coefficients and scales transfer unchanged to 64 held-out sequences per seed.

Table 1: Checkpoint 400, $T = 64$, mean calibration-KL cap 10^{-3}. ΔL averages within seeds, then over 20 seeds; brackets are 95% seed-bootstrap confidence intervals (CIs). Counts give coefficients / active Fisher rank; local task efficiency is the squared response ratio. Ref. share averages seedwise ratios of attained to independent-gate *feasible* loss decrease. The reference removes coefficient sharing while retaining $0 < f_t < 1$, candidates, write strengths, targets, baseline allocations and the mean KL cap; it is not a shared optimum (Appendix K.5).

Family	Coeff. / rank	Local eff.	$10^3\Delta L$ [95% CI]	Ref. share
Gate head	9 / 9	0.536	4.199 [3.866, 4.519]	68.34%
First-order	11 / 11	$\simeq 1$	5.408 [5.318, 5.491]	87.96%
Random	14 / 14	$\simeq 1$	5.455 [5.367, 5.537]	88.74%
Quadratic	14 / 14	$\simeq 1$	5.977 [5.900, 6.045]	97.22%

6.2 MATCHED FEATURE COUNT AND LOCAL TASK EFFICIENCY

With $\Delta L_{A,s}$ the seed-mean calibration decrease, all 20 differences $D_s = \Delta L_{\mathrm{quad},s} - \Delta L_{\mathrm{rand},s}$ are positive, averaging 5.2139×10^{-4} with 95% seed-bootstrap interval $[4.8143, 5.6385] \times 10^{-4}$ (10,000 resamples). Its lower end exceeds the prespecified, study-specific margin 3.125×10^{-4} (Figure 2(a)). All 80 selected updates satisfy the mean KL cap. The comparison uses the sequential least squares programming (SLSQP) selection procedure in Appendices K.5–K.6; Appendix M.6 reports the endpoint diagnostics.

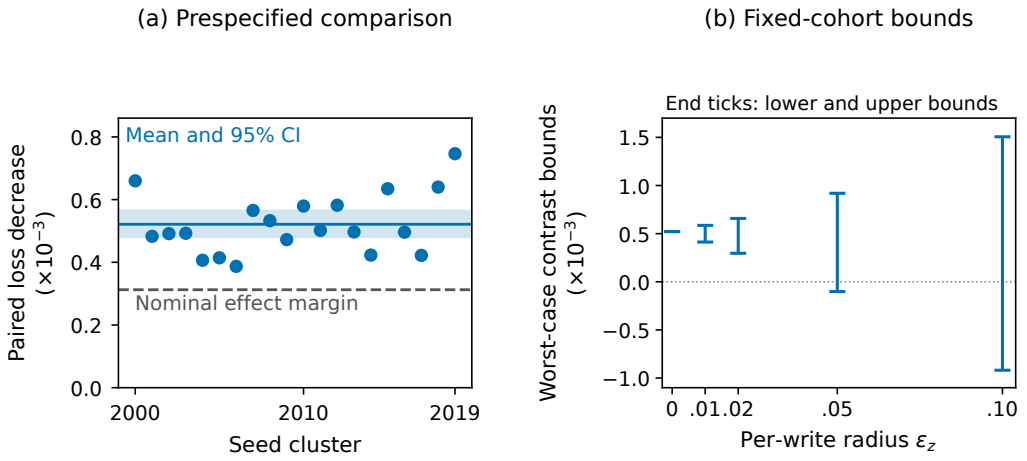


Figure 2: **Attained improvement and fixed-update perturbation bounds.** The same 20 calibration seeds, checkpoint 400, $T = 64$, mean allocation-KL cap 10^{-3}. (a) D_s from Section 6.2: solid mean, 95% seed-bootstrap band, dashed nominal margin 3.125×10^{-4}. (b) Bars bound $\mathcal{C}_{\mathrm{pert}}(\varepsilon_z) = 20^{-1} \sum_s (\mathcal{D}^{\mathrm{worst}}_{\mathrm{quad},s} - \mathcal{D}^{\mathrm{worst}}_{\mathrm{rand},s})$ via (122). Each $\mathcal{D}^{\mathrm{worst}}_{A,s}$ averages sequencewise loss-decrease infima over $\max_i \|\delta z_{a,i}\|_2 \le \varepsilon_z$, separately for each arm and sequence. These finite-precision, fixed-update bounds are not CIs (Appendices L.3 and M.1).

6.3 Perturbation, realization, and transfer analyses

Fixed-update perturbations. To assess the selected improvements under candidate changes, freeze both gate sequences, write strengths and targets. Figure 2(b) bounds $\mathcal{C}_{\mathrm{pert}}(\varepsilon_z)$ using analytic lower bounds and upper values from admissible perturbations. At $\varepsilon_z = 0.02$, the enclosure is $[2.9609, 6.5779] \times 10^{-4}$; at 0.05 and 0.1 it includes zero, leaving the ordering unresolved (Appendices L.3–L.4).

Prescribed-allocation realization. To test Section 5's finite-order predictions, both arms use degree-two paths $u^{\mathrm{ret}}_A(\lambda) = u^0 + X_A(\lambda\theta_{A,1} + \lambda^2\theta_{A,2})$ toward $q_{a,\lambda} \propto p^0_a e^{-\lambda Z_a^\top g_s}$, with Fisher-score projections (Appendix K.7). The calibration-mean error $N^{-1}_{\mathrm{cal}} \sum_a D_{\mathrm{KL}}(p_{A,a}(\lambda) \| q_{a,\lambda})$, not the baseline-KL cost, has median empirical orders 4.00 (Random) and 6.00 (Quadratic) as λ decreases (Figure 6). These paths are separate from the loss-optimized updates.

Transfer. Held-out coefficients and scales stay fixed: Random and Quadratic exceed the calibration mean KL cap in 18/20 and 15/20 seeds (Appendix K.8). Length comparisons refit updates; checkpoints redefine targets, with positive contrasts already at initialization (Appendices K.9–K.10). Figure 1 instead reuses one explicit repair across lengths.

7 Discussion and conclusion

For the constructed normalized recurrent memory, we prove that local sufficiency can coexist with a vanishing best loss decrease at a fixed small positive sensitivity budget. In the two-amplitude sigmoid family, a squared-difference feature cancels cumulative drift at the same baseline and preserves fixed-task improvement under the fixed-gate candidate perturbations of Corollary 3.2. The 20-seed study complements this construction: quadratic augmentation attains larger calibration improvements at matched width and local task efficiency; endpoint bounds and held-out costs quantify the perturbation and transfer outcomes.

The extensions identify how observations, reference spread, and shared query/retention updates change finite realization (Appendices F–J). Together, these results separate local response, target realization and finite sensitivity-constrained improvement.

REPRODUCIBILITY STATEMENT

Appendices A–J provide the mathematical definitions, assumptions, proofs and counterexamples; Appendix A.8 fixes notation and conventions. Appendix K specifies the synthetic experiment, update-selection procedure and seedwise analysis. Appendix L reports constructed controls and fixed-gate candidate-value perturbation evaluations; its learned-update analysis reuses the original 20-seed cohort. Appendices M.2–M.5 document deterministic numerical checks, Appendix M.6 describes independent checks from saved update parameters, and Appendix M.7 reports software and resources. The anonymous supplement separates saved-state verification, statistical reaggregation and figure-coordinate rendering. It supplies scientific code, configurations, all controller checkpoints, primary and constructed-update states, and source tables. Learned perturbation rows are regenerated from fixed primary states; secondary condition-level tables are supplied without all nonprimary endpoint caches. Appendix M.7 specifies the package paths, commands and scope.

ARTIFICIAL INTELLIGENCE (AI) USE STATEMENT

AI tools assisted theoretical exploration, literature discovery, mathematical drafting, programming, and finite-precision checking. Mathematical arguments, executed numerical checks, the completed synthetic empirical study, and unperformed extensions are identified separately. The authors take full responsibility for the correctness, novelty, and reporting of the work.

REFERENCES

Shun-ichi Amari. Differential Geometry of Curved Exponential Families—Curvatures and Information Loss. *The Annals of Statistics*, 10(2):357–385, 1982. doi: 10.1214/aos/1176345779. URL https://doi.org/10.1214/aos/1176345779.

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2): 251–276, 1998. doi: 10.1162/089976698300017746. URL https://doi.org/10.1162/089976698300017746.

Douglas M. Bates and Donald G. Watts. Relative Curvature Measures of Nonlinearity. *Journal of the Royal Statistical Society: Series B*, 42(1):1–16, 1980. doi: 10.1111/j.2517-6161.1980.tb01094.x. URL https://doi.org/10.1111/j.2517-6161.1980.tb01094.x.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. doi: 10.1287/moor.2016.0817. URL https://doi.org/10.1287/moor.2016.0817.

Amir Beck and Marc Teboulle. Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization. *Operations Research Letters*, 31(3):167–175, 2003. doi: 10.1016/S0167-6377(02)00231-6. URL https://doi.org/10.1016/S0167-6377(02)00231-6.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended Long Short-Term Memory. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c2ce2f2701c10a2b2f2ea0bfa43cfaa3-Abstract-Conference.html.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning Long-Term Dependencies with Gradient Descent Is Difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994. doi: 10.1109/72.279181. URL https://doi.org/10.1109/72.279181.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On Identifiability in Transformers. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1908.04211.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069. URL https://doi.org/10.1137/0916069.

Peter G. Casazza and Xuemei Chen. Frame Scalings: A Condition Number Approach. *Linear Algebra and its Applications*, 523:152–168, 2017. doi: 10.1016/j.laa.2017.02.020. URL https://arxiv.org/abs/1510.01653.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1812.07956.

Imre Csiszár and František Matúš. Information Projections Revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003. doi: 10.1109/TIT.2003.810633. URL https://doi.org/10.1109/TIT.2003.810633.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev Training for Neural Networks. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/758a06618c69880a6cee5314ee42d52f-Abstract.html.

Tri Dao and Albert Gu. Transformers Are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models, 2024. URL https://arxiv.org/abs/2402.19427. Preprint.

Bradley Efron. Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency). *The Annals of Statistics*, 3(6):1189–1242, 1975. doi: 10.1214/aos/1176343282. URL https://doi.org/10.1214/aos/1176343282.

Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory Traces in Dynamical Systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975, 2008. doi: 10.1073/pnas.0804451105. URL https://doi.org/10.1073/pnas.0804451105.

Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000. doi: 10.1162/089976600300015015. URL https://doi.org/10.1162/089976600300015015.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent Memory with Optimal Polynomial Projections. In *Advances in Neural Information Processing Systems*, 2020a. URL https://arxiv.org/abs/2008.07669.

Albert Gu, Caglar Gulcehre, Thomas Paine, Matt Hoffman, and Razvan Pascanu. Improving the Gating Mechanism of Recurrent Neural Networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3800–3809, 2020b. URL https://proceedings.mlr.press/v119/gu20a.html.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.00396.

Dhruman Gupta, Aritra Das, and Debayan Gupta. The Query Knows What to Forget: A Second Erase Direction for Linear Attention, 2026. URL https://arxiv.org/abs/2608.13668v1. Preprint.

Ali Hatamizadeh, Yejin Choi, and Jan Kautz. Gated DeltaNet-2: Decoupling Erase and Write in Linear Attention, 2026. URL https://arxiv.org/abs/2605.22791v1. Preprint.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Sarthak Jain and Byron C. Wallace. Attention Is Not Explanation. In *Proceedings of NAACL-HLT*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Robert E. Kass. Canonical Parameterizations and Zero Parameter-Effects Curvature. *Journal of the Royal Statistical Society: Series B*, 46(1):86–92, 1984. doi: 10.1111/j.2517-6161.1984.tb01280.x. URL https://doi.org/10.1111/j.2517-6161.1984.tb01280.x.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165, 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz Constant of Self-Attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5562–5571, 2021. URL https://proceedings.mlr.press/v139/kim21i.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated Gradient versus Gradient Descent for Linear Predictors. *Information and Computation*, 132(1):1–63, 1997. doi: 10.1006/inco.1996.2612. URL https://doi.org/10.1006/inco.1996.2612.

Dieter Kraft. A software package for sequential quadratic programming. Forschungsbericht DFVLR-FB 88-28, Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt (DFVLR), Institut für Dynamik der Flugsysteme, Oberpfaffenhofen, 1988.

Henry Lam. Robust Sensitivity Analysis for Stochastic Systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016. doi: 10.1287/moor.2015.0776. URL https://arxiv.org/abs/1303.0326v2. Author version: arXiv:1303.0326v2.

Eitan Levin, Joe Kileel, and Nicolas Boumal. The Effect of Smooth Parametrizations on Nonconvex Optimization Landscapes. *Mathematical Programming*, 209:63–111, 2025. doi: 10.1007/s10107-024-02058-3. URL https://arxiv.org/abs/2207.03512.

Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. Approximation and Optimization Theory for Linear Continuous-Time Recurrent Neural Networks. *Journal of Machine Learning Research*, 23(42):1–85, 2022. URL https://jmlr.org/papers/v23/21-0368.html.

Zhixuan Lin, Evgenii Nikishin, Xu Owen He, and Aaron Courville. Forgetting Transformer: Softmax Attention with a Forget Gate. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.02130.

Scott W. Linderman, Matthew J. Johnson, and Ryan P. Adams. Dependent multinomial models made easy: Stick-breaking with the Pólya-Gamma augmentation. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/07a4e20a7bbeeb7a736682b26b16ebe8-Abstract.html.

Lorenzo Livi. Learnability Window in Gated Recurrent Neural Networks. *Physical Review E*, 114: 035301, 2026. doi: 10.1103/843n-yshj. URL https://journals.aps.org/pre/abstract/10.1103/843n-yshj.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. doi: 10.1137/16M1099546. URL https://doi.org/10.1137/16M1099546.

Zhi-Quan Luo and Jong-Shi Pang. Error Bounds for Analytic Systems and Their Applications. *Mathematical Programming*, 67:1–28, 1994. doi: 10.1007/BF01582210. URL https://doi.org/10.1007/BF01582210.

Liam Madden and Christos Thrampoulidis. Memory Capacity of Two Layer Neural Networks with Smooth Activations. *SIAM Journal on Mathematics of Data Science*, 6(3):679–702, 2024. doi: 10.1137/23M1599355. URL https://arxiv.org/abs/2308.02001.

James Martens. New Insights and Perspectives on the Natural Gradient Method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://jmlr.org/papers/v21/17-678.html.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting Recurrent Neural Networks for Long Sequences. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26670–26698, 2023. URL https://proceedings.mlr.press/v202/orvieto23a.html.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the Difficulty of Training Recurrent Neural Networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318, 2013. URL https://proceedings.mlr.press/v28/pascanu13.html.

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative Entropy Policy Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1607–1612, 2010. doi: 10.1609/aaai.v24i1.7727. URL https://ojs.aaai.org/index.php/AAAI/article/view/7727.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://papers.nips.cc/paper/2021/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11504. URL https://ojs.aaai.org/index.php/AAAI/article/view/11504.

Igal Sason and Sergio Verdú. f-Divergence Inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. doi: 10.1109/TIT.2016.2603151. URL https://arxiv.org/abs/1508.00335.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear Transformers Are Secretly Fast Weight Programmers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366, 2021. URL https://proceedings.mlr.press/v139/schlag21a.html.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust Region Policy Optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Tomonari Sei. Efron's Curvature of the Structural Gradient Model. *Journal of the Japan Statistical Society*, 41(1):51–66, 2011. doi: 10.14490/jjss.41.051. URL https://doi.org/10.14490/jjss.41.051.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The Pitfalls of Simplicity Bias in Neural Networks. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6cfe0e6127fa25df2a0ef2ae1067d915-Abstract.html.

Yang Song, Jiaming Song, and Stefano Ermon. Accelerating Natural Gradient with Higher-Order Invariance. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4713–4722, 2018. URL https://proceedings.mlr.press/v80/song18a.html.

Rémi Tachet des Combes, Mohammad Pezeshki, Samira Shabanian, Aaron Courville, and Yoshua Bengio. On the Learning Dynamics of Deep Neural Networks, 2018. URL https://arxiv.org/abs/1809.06848. Preprint.

Corentin Tallec and Yann Ollivier. Can Recurrent Neural Networks Warp Time? In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.11188.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems*, volume 19, 2006. URL https://proceedings.neurips.cc/paper/2006/hash/d806ca13ca3449af72a1ea5aedbed26a-Abstract.html.

Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Geometry of Nonlinear Least Squares with Applications to Sloppy Models and Optimization. *Physical Review E*, 83:036701, 2011. doi: 10.1103/PhysRevE.83.036701. URL https://arxiv.org/abs/1010.1449.

Jesse van Oostrum, Johannes Müller, and Nihat Ay. Invariance Properties of the Natural Gradient in Overparametrised Systems. *Information Geometry*, 6:51–67, 2023. doi: 10.1007/s41884-022-00067-9. URL https://doi.org/10.1007/s41884-022-00067-9.

Manon Verbockhaven, Théo Rudkiewicz, Sylvain Chevallier, and Guillaume Charpiat. Growing Tiny Networks: Spotting Expressivity Bottlenecks and Fixing Them Optimally. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=hbtG6s6e7r.

Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001. URL https://doi.org/10.1561/2200000001.

Shida Wang and Qianxiao Li. StableSSM: Alleviating the Curse of Memory in State-Space Models through Stable Reparameterization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50766–50793, 2024. URL https://proceedings.mlr.press/v235/wang24ag.html.

Shida Wang, Zhong Li, and Qianxiao Li. Inverse Approximation Theory for Nonlinear Recurrent Neural Networks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.19190.

Sarah Wiegreffe and Yuval Pinter. Attention Is Not Not Explanation. In *Proceedings of EMNLP-IJCNLP*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Lemeng Wu, Dilin Wang, and Qiang Liu. Splitting Steepest Descent for Growing Neural Architectures. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/3a01fc0853ebeba94fde4d1cc6fb842a-Abstract.html.

Kaori Yamaguchi and Hiraku Nozawa. On statistics which are almost sufficient from the viewpoint of the Fisher metrics. *Information Geometry*, 7:543–553, 2024. doi: 10.1007/s41884-024-00160-1. URL https://doi.org/10.1007/s41884-024-00160-1.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated Linear Attention Transformers with Hardware-Efficient Training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56501–56523, 2024a. URL https://proceedings.mlr.press/v235/yang24ab.html.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing Linear Transformers with the Delta Rule over Sequence Length. In *Advances in Neural Information Processing Systems*, 2024b. URL https://arxiv.org/abs/2406.06484.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.06464.

Shiyu Zhao and Daniel Zelazo. Bearing Rigidity Theory and Its Applications for Control and Estimation of Network Systems: Life Beyond Distance Rigidity. *IEEE Control Systems Magazine*, 39(2):66–83, 2019. doi: 10.1109/MCS.2018.2888681. URL https://arxiv.org/abs/1803.05505.

Nicolas Zucchet and Antonio Orvieto. Recurrent Neural Networks: Vanishing and Exploding Gradients Are Not the End of the Story. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.21064.

APPENDIX GUIDE AND PROOF MAP

The core reading route follows the central separation before its extensions: Appendix C.3.1 and Appendix C.3.2 prove the two-amplitude barrier and exact repair, and Appendix C.4 proves the candidate-value perturbation guarantee. Appendix A develops allocation geometry, B gives the global drift conditions, and C also covers nonperiodic pairs and the canonical crossover, with its exact comparison and figure in Appendix C.2. Figure 3 accompanies the nonperiodic construction. Appendices D–E give finite realization and actual-budget descent. F–G study reset origins and candidate observations. H develops changes of reference allocation, illustrated in Figure 5; I–J treat query, observation and memory routes. K specifies the learned-feature experiment, L its fixed-update perturbation analysis and constructed controls, M the numerical checks and reproduction details, and N the literature comparisons.

Result	Complete proof
Theorem 2.1	Appendix A.4
Theorem 3.1	Appendices C.3.1–C.3.2
Corollary 3.2	Appendix C.4
Theorem 4.1	Appendix B.3
Theorem 4.2	Appendix B.4
Theorem 4.3	Appendix C.1
Theorem C.1	Appendix C.2
Theorem 5.1	Appendix D.2
Theorem 5.2	Appendix E.2, using E.1
Theorem F.4	Appendix F.5
Theorem G.1	Appendix G.1

A FOUNDATIONS: ALLOCATION, SENSITIVITY AND LOCAL RESPONSE

A.1 EQUIVALENT MEMORY FUNCTIONS AND UNAVOIDABLE DILUTION

For $t \geq 2$, division of (1) by n_t yields (2); at $t = 1$, $m_1 = z_1$. Conversely, for any $0 < \gamma_t < 1$, choose $n_1 = \iota_1 = 1$, $f_t = 1$, $n_t = n_{t-1}/(1-\gamma_t)$, and $\iota_t = \gamma_t n_t$. These candidate-independent choices give the same output at every time for every candidate sequence. One can choose another positive f_t and set $n_t = f_t n_{t-1}/(1-\gamma_t)$ if a retention domain is prescribed.

Holding the allocation fixed gives $D_{z_i} m_T = \pi_i I_d$, and therefore $\sum_i \|D_{z_i} m_T\|_{\mathrm{Frob}}^2 = d \sum_i \pi_i^2$. Under uniform allocation each write has derivative I_d/T. Nevertheless, the parameterized change $z_i(\vartheta_{\mathrm{shift}}) = z_i + \vartheta_{\mathrm{shift}} d_{\mathrm{shift}}$ has $D_{\vartheta_{\mathrm{shift}}} m_T = d_{\mathrm{shift}}$. Thus individual-write influence and the response to a shared candidate shift can have different length dependence. When candidates also determine allocation controls, the total derivative includes the derivatives of those controls.

A nonnegative scalar update $m' = \kappa m + \omega z$ preserves the unit ball for every m, z in it exactly when $\kappa + \omega \leq 1$: sufficiency is the triangle inequality, and necessity uses $m = z$ of norm one. The missing weight can be assigned to a zero anchor. Consequently boundedness imposes a keep/write budget, with any missing mass represented by the zero-anchor branch.

For the multi-observation extensions, let $T \in \mathcal{O} \subseteq \{2, \ldots, T\}$ and $w_t^{\mathrm{obs}} > 0$ sum to one. With fixed targets y_t, conservative prefix laws $p_{\mathrm{cons}}^{(t)}$, and $m_t = Z_t p_{\mathrm{cons}}^{(t)}$, use

$$L_{\mathcal{O}} = \tfrac{1}{2} \sum_{t \in \mathcal{O}} w_t^{\mathrm{obs}} \|m_t - y_t\|^2, \qquad D_w(p\|q) = \sum_{t \in \mathcal{O}} w_t^{\mathrm{obs}} D_{\mathrm{KL}}(p_{\mathrm{cons}}^{(t)} \| q_{\mathrm{cons}}^{(t)}). \tag{24}$$

Here D_w first samples an observation with weight w_t^{obs}, then a write; prefixes are not assumed independent. Batch weights are additional explicit sequence weights. The terminal loss is the case $\mathcal{O} = \{T\}$. The inverse identities used in Section 2 are

$$F_b^{-1} = \Delta \, \mathrm{Diag}(1/\pi) \Delta^\top, \qquad J_b F_b^{-1} J_b^\top = \Sigma_\pi \coloneqq Z C_\pi Z^\top. \tag{25}$$

Appendix A.3 proves these inverse identities exactly on the positive simplex.

A.2 Operational sensitivity and a metric control

For any perturbation table with $\max_i \|\delta z_i\| \le \varepsilon_z$, the triangle inequality yields

$$\|\sum_i (p_i - q_i)\delta z_i\| \le \varepsilon_z \sum_i |p_i - q_i|.$$

Take a unit vector e and $\delta z_i = \varepsilon_z \operatorname{sign}(p_i - q_i)e$ to attain equality (any value is allowed at zero differences). Pinsker's inequality proves (4). With the same perturbation table across observations,

$$\sum_t w_t^{\mathrm{obs}} \|(A_{p_{\mathrm{cons}}^{(t)}} - A_{q_{\mathrm{cons}}^{(t)}})\delta z_{:t}\|^2 \le 2\varepsilon_z^2 D_w(p\|q).$$

Equation (4) applies operator-norm duality to the fixed-allocation update effect and motivates the chosen KL budget. The multi-observation expression above provides an upper bound when all observations share the same perturbation table. Recomputing input-dependent gates adds their derivative terms to this conditional response.

For a replication control, repeat an L_{blk}-phase template n times. Let j denote the copy index, ℓ its phase, and use a tied log-retention β so that $p_\beta(j,\ell) \propto \nu_\ell e^{\beta(L_{\mathrm{blk}} j + \ell)}$. Candidates depend only on ℓ. The law factors into a copy marginal and a phase marginal. Its mean, as a function of β, is independent of n, and so are every derivative of a mean-only loss and every deterministic gradient-descent path on that same loss. But

$$F_n(\beta) = \mathrm{Var}(\ell) + L_{\mathrm{blk}}^2 \mathrm{Var}(j), \qquad D_{\mathrm{KL}}(p_\beta\|p_0) = D_{\mathrm{KL}}(p_\beta^{\mathrm{phase}}\|\nu) + D_{\mathrm{KL}}(p_\beta^{\mathrm{copy}}\|\mathcal{U}_n).$$

At zero the extra Fisher is $L_{\mathrm{blk}}^2(n^2-1)/12$. It measures redistribution between individually distinguishable copies. Merging copies changes that metric. In this example, replication increases the allocation metric while leaving the mean-only loss and its deterministic gradient-descent paths unchanged.

A.3 The allocation inverse

Unrolling the recurrence gives unnormalized weights $\iota_i \prod_{k=i+1}^{T} f_k$, hence (5). The inverse positive-simplex coordinates are

$$b = \Delta(\log \pi - a^{\mathrm{w}}).$$

The columns of H together with $\mathbf{1}$ span $\mathbb{R}^T$. Since C_π is positive on the quotient by constants, $F_b \succ 0$. Differentiating the inverse, or using $\Delta H = I$ and $\Delta \mathbf{1} = 0$, gives

$$F_b^{-1} = \Delta \operatorname{Diag}(1/\pi)\Delta^\top, \qquad F_b^{-1} H^\top C_\pi = \Delta.$$

For example, $H\Delta = I - \mathbf{1}e_T^\top$, so

$$\Delta \operatorname{Diag}(1/\pi)\Delta^\top H^\top C_\pi H = \Delta \operatorname{Diag}(1/\pi) C_\pi H = \Delta H = I.$$

It follows that $C_\pi H F_b^{-1} H^\top C_\pi = C_\pi$, which proves the covariance identity. The inverse is tridiagonal. Under sharing, the parameter-space metric is instead formed by the pullback $U_0^\top F_b U_0$.

A.4 Affine sharing and the regression identity

The shared response also has the exact weighted regression representation

$$g^\top K_U g = \mathrm{Var}_\pi(Z^\top g) - \min_{a_{\mathrm{reg}}, \theta_{\mathrm{reg}}} \sum_i \pi_i \{z_i^\top g - a_{\mathrm{reg}} - (HU\theta_{\mathrm{reg}})_i\}^2. \tag{26}$$

The affine minimum completion count in Theorem 2.1 is

$$k_{\min}^{\mathrm{aff}} = \operatorname{rank}[U, \Delta Z^\top] - \operatorname{rank} U. \tag{27}$$

Let $\widetilde{J}_b = J_b F_b^{-1/2}$ and $E = \operatorname{range}(F_b^{1/2} U)$. The Gram-projector identity proves $K_U = \widetilde{J}_b P_E \widetilde{J}_b^\top \preceq \Sigma_\pi$. The projection step is the natural-gradient pushforward identity under a possibly redundant parameterization (van Oostrum et al., 2023, Theorem 1 and Lemma 1). The candidate-difference

criterion below combines this identity with the recurrent inverse $\Delta H = I$. Weighted Cauchy–Schwarz on the parameter quotient gives the radius-r_{loc} coefficient in the main text.

Set $B = \mathrm{Diag}(\sqrt{\pi})(Z^\top - \mathbf{1}m^\top)$ and $E_s = \mathrm{Diag}(\sqrt{\pi})(HU - \mathbf{1}\pi^\top HU)$. Equality of the shared and free operators is equivalent to $(I - P_{\mathrm{range}\, E_s})B = 0$. Because all masses are positive, this is equivalent to $\mathrm{range}\, Z^\top \subseteq \mathrm{range}(HU) + \mathrm{span}\, \mathbf{1}$. Applying Δ yields the inclusion in Theorem 2.1. Conversely, if $UC = \Delta Z^\top$, $\theta^+ = \theta - \lambda C g$ produces (7) exactly, for every magnitude in the unrestricted affine family. Closure itself implies the same inclusion by differentiating at zero, or by taking adjacent differences of any nonzero finite target. Adding a basis of the missing quotient attains the rank-increment minimum; each new column can enlarge the span by at most one.

Projection of the centered vector h onto centered cumulative features gives (26). For even T, $z_i = (-1)^i$, and $U = \mathbf{1}$, HU has entries $T - i$. Its variance is $(T^2 - 1)/12$ and covariance with z is $-1/2$. The shared scalar coefficient is $3/(T^2 - 1)$, although the output derivative is nonzero and output rank is one. The equal-Fisher-radius response ratio is the square root of the squared-response coefficient ratio.

A.5 A GLOBAL OSCILLATION INEQUALITY

For a positive reference q and score increment d, let $p = \mathrm{softmax}(\log q + d)$ and $A(s) = \log \sum_i q_i e^{sd_i}$ and $p_s = \mathrm{softmax}(\log q + sd)$. Since $A''(s) = \mathrm{Var}_{p_s}(d) \le \mathrm{osc}(d)^2/4$, the Bregman integral formulas give

$$D_{\mathrm{KL}}(p\|q) = \int_0^1 sA''(s)\, ds \le \mathrm{osc}(d)^2/8, \qquad D_{\mathrm{KL}}(q\|p) \le \mathrm{osc}(d)^2/8. \tag{28}$$

The bound depends only on score oscillation and applies uniformly over positive reference probabilities. Pinsker gives $\|p - q\|_1 \le \mathrm{osc}(d)/2$. We use these categorical inequalities to control accumulated recurrent error.

For (21), the score change from q_λ to p has oscillation $\varepsilon_{\mathrm{sc}}(\lambda)$. The change from π^0 to p is $-\lambda h$ plus that error, modulo a constant, so it has oscillation at most $\lambda\Omega_h + \varepsilon_{\mathrm{sc}}(\lambda)$. Equation (28) proves both the target mismatch bound and (23).

A.6 COARSENED SENSITIVITY AND A NEGATIVE CONTROL

Let $\mathcal{P}$ be a fixed partition of $\{1, \dots, T\}$ into nonempty disjoint groups covering every write index. Define $p_G^{\mathcal{P}} = \sum_{i \in G} p_i$ and $q_G^{\mathcal{P}} = \sum_{i \in G} q_i$. Restrict perturbations to $\delta z_i = \delta z_G$ for $i \in G \in \mathcal{P}$, with $\max_G \|\delta z_G\| \le \varepsilon_z$. Then

$$(A_p - A_q)\delta z = \sum_{G \in \mathcal{P}} (p_G^{\mathcal{P}} - q_G^{\mathcal{P}})\delta z_G, \qquad p_G^{\mathcal{P}} = \sum_{i \in G} p_i.$$

Norm duality, attained by choosing all δz_G along one unit vector with the sign of $p_G^{\mathcal{P}} - q_G^{\mathcal{P}}$, proves that the supremum is exactly $\varepsilon_z \|p^{\mathcal{P}} - q^{\mathcal{P}}\|_1$. Thus the chosen partition specifies which write perturbations the budget controls.

In the canonical one-feature construction, group the write indices by negative or positive candidate value. The grouped law has binary parameter $s^{\mathrm{sgn}}(\vartheta) = -\vartheta + \log\cosh(2\vartheta)/2$. Its gain is $-\tanh s^{\mathrm{sgn}}(\vartheta)$, its grouped operator distance is the same number for $\vartheta \ge 0$, and its grouped KL is $k(s^{\mathrm{sgn}}(\vartheta))$. Every sufficiently small positive grouped budget is attained by a parameter independent of T. Therefore the coarsened-operator optimum is δ, and the coarsened-KL optimum is $\tanh \lambda_\tau$, both independent of T, even without a new feature. The original barrier lives in the copy redistribution discarded by that grouping. The replication control in Appendix A.2 isolates the same distinction between copy redistribution and the mean-only task.

A.7 WHAT THE OPERATIONAL BUDGET PRESERVES

Let $\eta^0, \eta^+ \in \mathbb{R}^T_{\ge 0}$ be the fixed candidate-coefficient vectors of the nonnegative memories considered here, and set $m^0(z) = \sum_i \eta_i^0 z_i$ and $m^+(z) = \sum_i \eta_i^+ z_i$. Their update effect is $E_{\mathrm{upd}}(z) = m^+(z) -$

$m^0(z)$. Separately selectable changes δz_i to the candidate values give the exact identity

$$\sup_{\max_i \|\delta z_i\|_2 \le \varepsilon_z} \|E_{\mathrm{upd}}(z+\delta z) - E_{\mathrm{upd}}(z)\|_2 = \varepsilon_z \|\eta^+ - \eta^0\|_1. \tag{29}$$

Indeed, $E_{\mathrm{upd}}(z+\delta z) - E_{\mathrm{upd}}(z) = \sum_i (\eta_i^+ - \eta_i^0)\delta z_i$, and aligned signed perturbations attain the triangle bound. Thus a candidate-operator budget limits how much the *effect of the update* can change when candidate values are perturbed separately. The identity holds for every candidate sequence with the two coefficient vectors fixed. The perturbation coordinates are the candidate writes; the zero anchors remain fixed.

When the loss decrease survives a candidate-value perturbation. A loss-decrease bound must also account for the perturbation of the baseline residual. For a fixed target y, let $\mathcal{D}_{\mathrm{loss}}(z) = \frac{1}{2}\|m^0(z) - y\|^2 - \frac{1}{2}\|m^+(z) - y\|^2$, $\Delta m = m^+(z) - m^0(z)$, $\delta_\eta = \|\eta^+ - \eta^0\|_1$, and $M_0 = \sum_i \eta_i^0$. For every $\max_i \|\delta z_i\| \le \varepsilon_z$,

$$\mathcal{D}_{\mathrm{loss}}(z+\delta z) \ge \mathcal{D}_{\mathrm{loss}}(z) - \varepsilon_z M_0 \|\Delta m\| - \varepsilon_z \delta_\eta \|m^+(z) - y\| - \varepsilon_z^2 \left(M_0 \delta_\eta + \tfrac{1}{2}\delta_\eta^2\right). \tag{30}$$

To prove it, set $\delta m^0 = \sum_i \eta_i^0 \delta z_i$ and $\delta E_{\mathrm{upd}} = \sum_i (\eta_i^+ - \eta_i^0)\delta z_i$, so $\|\delta m^0\| \le \varepsilon_z M_0$, $\|\delta E_{\mathrm{upd}}\| \le \varepsilon_z \delta_\eta$. Exact expansion gives

$$\mathcal{D}_{\mathrm{loss}}(z+\delta z) - \mathcal{D}_{\mathrm{loss}}(z) = -\left(\delta m^0\right)^\top \Delta m - (m^+(z) - y)^\top \delta E_{\mathrm{upd}} - \left(\delta m^0\right)^\top \delta E_{\mathrm{upd}} - \tfrac{1}{2}\|\delta E_{\mathrm{upd}}\|^2.$$

Cauchy–Schwarz proves (30). A certified baseline loss margin exceeding its three error terms therefore certifies improvement after every such perturbation. If $\|z_i\| \le R$, also $\|\Delta m\| \le R\delta_\eta$. The target and both gate sequences remain fixed. The supremum is over ambient perturbations of radius ε_z; requiring perturbed values to remain in the original ball restricts that set and preserves this lower bound, but need not preserve equality in (29).

A.8 Notation and model-dependent conventions

A.8.1 Reserved roles and coordinate maps

Typeface and dimension conventions. Scalar, vector and matrix variables use the same ordinary mathematical alphabet; their declared spaces and indices, rather than bold weight, determine their types. Thus $z_i, m_t, y_t \in \mathbb{R}^d$ are vectors, while $Z_t \in \mathbb{R}^{d\times t}$ and $H \in \mathbb{R}^{T\times(T-1)}$ are matrices. In scalar constructions ($d = 1$), the same candidate and output symbols denote scalars. An entry such as H_{ij} or $(z_i)_k$ is scalar. Bold $\mathbf{z}$ denotes the complete candidate tuple, and $\mathbf{1}$ the all-ones vector (or an indicator when indexed by an event or followed by a braced condition). Sets and scalar-valued functionals retain their named calligraphic symbols; $\mathsf{T}^{\mathrm{pos}}$ is a scalar random position. Named operators and descriptive subscripts are upright. The signs $:=$ and $=:$ mark definitions in the indicated direction; ordinary equality is also used in defining recurrences and after explicit "let", "set", or "define" statements.

Bounded affine memory (BAM) and innovation-coordinate memory (ICM) are the recurrence and chart names used below; Appendix F.1 gives their formal definitions.

Notation	Object and interpretation
$z_i,\ Z_t$	Candidate vector $z_i \in \mathbb{R}^d$ and candidate matrix $Z_t = [z_1, \dots, z_t] \in \mathbb{R}^{d\times t}$.
$a_t,\ m_t,\ y_t$	Fixed anchor, memory output, and fixed target; $\widehat{y}_t$ is a predicted readout.
$b_j = \log f_{j+1}$	Log-retention coordinate.
$u_j^{\mathrm{ret}},\ u_t^{\mathrm{inn}}$	Respectively $\operatorname{logit} f_{j+1}$ (when $f_{j+1} < 1$) and $\operatorname{logit} \gamma_t$.
$u_t^{I,\mathrm{k}}, u_t^{I,\mathrm{w}}$	ICM logits $\log(\kappa_t/\sigma_t)$ and $\log(\omega_t/(\kappa_t+\sigma_t))$. The keep logit conditions on the nonwrite branch.
ξ	Stack $(u_2^{I,\mathrm{k}}, u_2^{I,\mathrm{w}}, \dots, u_T^{I,\mathrm{k}}, u_T^{I,\mathrm{w}})^\top$.
$J_{Y\leftarrow x}$	Coordinate matrix of the derivative D_xY; rows are output coordinates and columns are input coordinates.
$F,\ K$	Fisher metric and local output-response Gram, with support/coordinate subscripts. Calligraphic $\mathcal{K}, \mathcal{V}$ denote scalar KL and operator costs.
H, Δ	Cumulative retention-to-score map and adjacent-difference inverse; $\Delta H = I$.
$X_{\mathrm{ret}}, \Phi,\ U$	Fixed sharing in retention logits, ICM logits, and log-retention coordinates, respectively. These are different shared families.
$P_E,\ P_E^{[p^0]}$	Orthogonal projection in Euclidean space, or in $L^2(p^0)$ when the bracket qualifier is present.
$Q,\ E_{\mathrm{key}}$	Shared query-matrix increment in Appendix I.1 and fixed key-embedding matrix; baseline query scores are included in $s^{0,t}$.
$W_n,\ \mathsf{T}^{\mathrm{pos}}$	Reference temporal spread and random block position, respectively.

For a general sharing map, $U_0 = J_{b\leftarrow\theta}(\theta^0)$ is a derivative at the fixed baseline. For the affine family $b = b^0 + U\theta$, $U_0 = U$ everywhere. At that baseline,

$$J^0_{m_T\leftarrow\theta} = J_b U_0, \quad F_\theta = U_0^\top F_b U_0, \quad K_U = J_b U_0 F_\theta^\dagger U_0^\top J_b^\top.$$

Here $J_b \in \mathbb{R}^{d\times(T-1)}$, $U_0 \in \mathbb{R}^{(T-1)\times d_\theta}$, and $K_U \in \mathbb{R}^{d\times d}$. The whitened Jacobian $\widetilde{J}_b = J_b F_b^{-1/2}$ uses transformed input coordinates; $\widetilde{J}_b\widetilde{J}_b^\top$ is its output Gram.

A.8.2 Supports, observations, and Fisher pullbacks

We use p, q for generic probability laws; sequence-indexed p_a and $p_{a,n,\theta}$ also denote terminal allocations. In conservative memory $\pi = p^{(T)}$ and $p^{0,+} = \pi^{0,+}$; with reset origins $\pi^{(t)}$ denotes the conditional candidate allocation below. These supports remain distinct:

Object	Size	Support or meaning
$p^{(t)}_{\mathrm{cons}}$	t	A probability vector over conservative candidate writes.
$p^{(t)}_{\mathrm{orig}}$	$2t-1$	All origin-resolved writes and reset origins.
$\eta^{(t)}$	t	Candidate coefficients; $M_t = \mathbf{1}^\top\eta^{(t)} \le 1$.
$\bar{p}^{(t)}$	$t+1$	$(\eta^{(t)}, 1-M_t)$, merging only the reset origins.
$\pi^{(t)}$	t	Conditional candidate allocation $\eta^{(t)}/M_t$ when $M_t > 0$.

The origin-resolved order is $(\mathsf{w}_1, \mathsf{w}_2, \mathsf{r}_2, \dots, \mathsf{w}_t, \mathsf{r}_t)$, with $x_{\mathsf{w}_i} = z_i$ and $x_{\mathsf{r}_i} = a_i$. Thus terminal origin-resolved masses are $(\eta_1, \eta_2, r_2^{\mathrm{reset}}, \dots, \eta_T, r_T^{\mathrm{reset}})$. The observation map sums the reset masses while retaining every candidate coordinate, including separately perturbable zero-valued candidates. Grouped write perturbations instead use a separately declared partition $\mathcal{P}$ and masses $p^{\mathcal{P}}_G = \sum_{i\in G} p_i$.

For fixed-zero anchors, let $J_{\mathrm{cand}} = D\mathcal{C}(\xi^0) \in \mathbb{R}^{T\times 2(T-1)}$. Then

$$F_\eta^{\mathrm{agg},0} = \operatorname{Diag}(1/\eta^0) + \frac{\mathbf{1}\mathbf{1}^\top}{1-M_0},$$
$$F_\xi^{\mathrm{agg},0} = J_{\mathrm{cand}}^\top F_\eta^{\mathrm{agg},0} J_{\mathrm{cand}},$$
$$\ker(\Phi^\top F_\xi^{\mathrm{agg},0}\Phi) = \ker(J_{\mathrm{cand}}\Phi).$$

The origin-resolved Fisher $F_\xi^{\mathrm{orig},0}$ uses the same logit coordinates but a different probability law. It is positive definite in the independent positive chart. The aggregated Fisher is only semidefinite;

its quadratic form becomes an inner product only after taking the corresponding quotient. The candidate-operator cost $\mathcal{V}_{\mathrm{cand}}(\eta) = \|\eta - \eta^0\|_1$ is not this Fisher quadratic form or the full distance $\|\bar{p} - \bar{p}^0\|_1$.

The losses specify their input: $L_T^{\mathrm{cons}}(p) = \frac{1}{2}\|Zp - y_T\|^2$ and $L_T^{\mathrm{cand}}(\eta) = \frac{1}{2}\|Z\eta - y_T\|^2$; $L_T^{\mathrm{agg}}(\bar{p}(\eta))$ equals the latter. Origin-resolved losses use the matrix of attached write/anchor values. $L_{\mathcal{O}}$ sums the indicated prefix losses with weights w_t^{obs}. The calligraphic map $\mathcal{L}(\eta, r^{\mathrm{reset}})$ denotes the inverse gate map.

A.8.3 Expansion order, probability laws, and time scaling

The necessary origin-resolved-logit increment is

$$\Xi(\lambda) = \sum_{j \geq 1} \lambda^j \Xi_j, \qquad \Xi_j = \Xi^{(j)}(0)/j!.$$

The coefficients $\xi_{[j]}$ describe a chosen logit curve, whereas $\eta^\star_{[j]}$ describe the requested candidate curve. Proposition G.2 ranges over all admissible first-order lifts $\xi_{[1]}$. Both Ξ_j and the inverse-logit coefficients ψ_j are Taylor coefficients, including the factor $1/j!$. Cumulants χ_j and the log-probability derivatives used for statistical curvature are derivatives at zero; the corresponding degree-j Taylor coefficient is obtained by dividing by $j!$.

Prefix statistics use the value pushforward

$$\mu_{\mathrm{val}}^{0,(t)} = \sum_{i \in \mathcal{I}_t} p_{\mathrm{orig},i}^{0,(t)} \delta_{x_i}.$$

The variables $\mathsf{V}_{t-1}^{\mathrm{pref}}$ and $\mathsf{V}_t^{\mathrm{nw}}$ have laws $\mu_{\mathrm{val}}^{0,(t-1)}$ and $f_t^{I,0}\mu_{\mathrm{val}}^{0,(t-1)} + (1 - f_t^{I,0})\delta_{a_t}$, respectively. This pushforward supplies the prefix statistics, while the KL budget remains defined on the origin labels.

A time law q^{time} has position $\mathsf{T}^{\mathrm{pos}}$, spread $W > 0$, and standardized position $\mathsf{T}^{\mathrm{std}} = (\mathsf{T}^{\mathrm{pos}} - \mathbb{E}\mathsf{T}^{\mathrm{pos}})/W$. The two tilt parameters obey

$$q_\beta^{\mathrm{time}}(j) \propto q^{\mathrm{time}}(j) e^{\beta j}, \qquad \widehat{q}_s^{\mathrm{time}}(j) \propto q^{\mathrm{time}}(j) e^{s\mathsf{T}^{\mathrm{std}}(j)}, \qquad q_\beta^{\mathrm{time}} = \widehat{q}_{\beta W}^{\mathrm{time}}.$$

Replacing position by age reverses the sign of the tilt but not the variance. The geometric rate is $\rho_{\mathrm{geo}} \in (0, 1)$, whereas $r \geq 1$ is the unavoidable drift order. The assumptions of Theorem H.2, including tail regularity, comparison of widths, and global drift control, remain unchanged.

Time is denoted by t, a scalar control by ϑ, and a prescribed tilt by λ. Superscripts 0, $+$ and (t) indicate a baseline, an updated endpoint and a prefix, respectively; $\top$ denotes transpose. The input radius ε_z, local Fisher radius r_{loc}, actual budgets τ, δ, and cumulative score error $\varepsilon_{\mathrm{sc}}$ retain their respective definitions. Local abbreviations, such as $A(s)$ for a log-partition function, are defined at their point of use.

A.8.4 Fixed-gate perturbations and numerical enclosures

A candidate-value perturbation is δz, with radius ε_z. A sensitivity budget means either an allocation-KL or an operator budget, as specified. For the two-amplitude theoretical construction, $\mathcal{K}_T$ and $\mathcal{V}_T$ are mean allocation-KL and mean individual-write operator costs. The perturbed loss decrease $\mathcal{D}_T^{\mathrm{pert}}(\theta, \delta z)$ compares the same perturbed candidate values under two frozen gate sequences. The best outer-update/inner-perturbation value is $B_{\mathcal{F}}^{\mathrm{pert,KL}}$ or $B_{\mathcal{F}}^{\mathrm{pert,op}}$; it is distinct from a fixed learned endpoint.

For a selected learned update, $\mathcal{D}_{A,s}^{\mathrm{worst}}$ averages the per-sequence perturbation infima within seed s. Here $A \in \{\mathrm{head}, \mathrm{first}, \mathrm{rand}, \mathrm{quad}\}$ labels Gate head, First-order, Random and Quadratic. The quantities $\underline{\mathcal{D}}_{A,s}$ and $\overline{\mathcal{D}}_{A,s}$ average the analytic lower bounds and feasible-attack upper values. Their gap is bound slack; seed-bootstrap intervals describe sampling variation across the 20 seed clusters. Equation (122) uses the Quadratic/Random pair, while Figure 12 reports the four armwise bounds.

A.8.5 Terminology tied to a map and a budget

CNM, BAM and ICM refer to the recurrences and coordinate charts defined in this paper. Conservative memory has total candidate mass one; boundedness refers to the memory output. Causal feature

computation uses the available prefix, whereas an offline parameter update may use a terminal residual.

Local task efficiency is the squared shared/free response ratio in the declared allocation-Fisher metric, when the free response is nonzero. Local sufficiency means this scalar is one for the requested task; $K_U = \Sigma_\pi$ is the stronger all-direction property. Zero-deficit statements do not require dividing by the free response. Softmax scores are log weights modulo a common constant, and statistical scores are mean-zero derivatives of log probabilities. Drift is the deterministic block log-weight slope. Statistical curvature is defined from the Fisher-normal component of log-probability acceleration in exponential-family geometry (Appendix D.5.2).

Closure under tilts in Theorem 2.1 means realization of every specified entropic tilt within the family. The topological closure in Lemma G.3 consists of limits of budget-feasible shared allocations. A fiber is an inverse image of the observation map; the derivative kernel describes locally invisible variations. The adversarial perturbation set allows a separate bounded choice at each write index. Appendix H.3 defines reference allocations, mean lag and temporal spread in their respective units.

The prefix $z_{1:t}$ is a sequence and Z_t its column matrix. The arithmetic mean $\bar{z}_t$, allocation-weighted output m_t, and origin-weighted prefix moments use the weights specified in their definitions. Mixing coefficients $(\gamma_t)_{t=2}^T$, gate logits and parameter curves describe different stages of the recurrence-to-parameter map.

A baseline is the pre-update gate configuration and its induced allocation and output. Same-baseline families agree at that configuration. A homogeneous retention baseline has constant gates; the induced allocation also depends on the write strengths. A selected endpoint is a fitted coefficient vector with its induced allocation, and an explicit witness is a constructed feasible endpoint. The output target y_t, target allocation q_λ, linear gain G, local response and squared-loss decrease are the quantities defined in their respective sections.

B Cumulative drift: global characterizations and rates

This appendix proves the general drift statements of Section 4.1. The candidate arrays, phase count, reference phase laws, task directions, and positive sequence weights are fixed. Constants are allowed to depend on them. They are independent of the number n of repeated blocks and of budgets in the stated small-budget ranges. The control domain need not be compact for Theorem 4.2; its drift inequality must instead hold over the entire region stated there.

B.1 Exact recurrent factorization

Define $\ell_{\mathrm{ph}}(k) = 1 + ((k-1) \bmod L_{\mathrm{blk}})$. Index a write by $i = jL_{\mathrm{blk}} + \ell$, with $0 \le j < n$ and $1 \le \ell \le L_{\mathrm{blk}}$. The gate after write i has increment $d_{a,\ell}$; there is no gate after the terminal write. Relative to the baseline, the terminal log mass of write (j, ℓ) changes by

$$\sum_{k=i}^{nL_{\mathrm{blk}}-1} d_{a,\ell_{\mathrm{ph}}(k)} = (n-1-j) \sum_{\ell'=1}^{L_{\mathrm{blk}}} d_{a,\ell'} + \sum_{k=\ell}^{L_{\mathrm{blk}}-1} d_{a,k}.$$

The first term is $j\beta_a$ modulo a constant independent of j, ℓ. Since the baseline is $\mathcal{U}_n \otimes \nu_a^0$, normalization separately over j and ℓ proves (10). The last phase increment is a real gate in every block except the last; it is not an additional terminal control. Candidates depend only on phase, so the mean is $Z_a \nu_{a,\theta}$. The KL chain rule for the product gives (11). One CNM realization uses baseline gates f_* and writes $\iota_{jL_{\mathrm{blk}}+\ell} = f_*^{jL_{\mathrm{blk}}+\ell-1} \nu_a^0(\ell)$; all terminal baseline masses have the stated factorization. More general positive baselines are possible.

B.2 Tilted-uniform bounds for both budgets

Recall $q_{n,\beta}^{\mathrm{time}}(j) \propto e^{\beta j}$ on $j = 0, \dots, n-1$, $n \ge 2$. Reflection permits $\beta \ge 0$. With $A_n(\beta) = \log(n^{-1} \sum_j e^{\beta j})$,

$$D_n^{\mathrm{KL}}(\beta) = \beta A_n'(\beta) - A_n(\beta) = \int_0^\beta s \operatorname{Var}_{q_{n,s}^{\mathrm{time}}}(j)\, ds.$$

For $\beta(n-1) \le 1$, every intermediate density relative to $\mathcal{U}_n$ is between e^{-1} and e. Using $\mathrm{Var}_p(j) = \min_c \mathbb{E}_p(j-c)^2$ gives

$$\frac{e^{-1}(n^2-1)\beta^2}{24} \le D_n^{\mathrm{KL}}(\beta) \le \frac{e(n^2-1)\beta^2}{24}. \tag{31}$$

Strict increase of D_n^{KL} on $\beta > 0$ implies that $D_n^{\mathrm{KL}}(\beta) < 1/(24e)$ forces $\beta(n-1) < 1$ and hence $n\beta \lesssim \sqrt{D_n^{\mathrm{KL}}(\beta)}$. The global upper bound $D_n^{\mathrm{KL}}(\beta) \le \beta^2(n-1)^2/8$ follows from (28) in Appendix A.5, since the score increment βj has oscillation $|\beta|(n-1)$.

For the operator budget let $D_n^{\mathrm{op}}(\beta) = \|q_{n,\beta}^{\mathrm{time}} - \mathcal{U}_n\|_1$. The variable $\mathsf{T}^{\mathrm{unit}} = j/(n-1)$ lies in $[0,1]$, so

$$D_n^{\mathrm{op}}(\beta) \ge 2|\mathbb{E}_{q_{n,\beta}^{\mathrm{time}}}\mathsf{T}^{\mathrm{unit}} - \mathbb{E}_{\mathcal{U}_n}\mathsf{T}^{\mathrm{unit}}|.$$

For $\beta(n-1) \le 1$, differentiate its expectation and use the variance lower bound:

$$D_n^{\mathrm{op}}(\beta) \ge \frac{e^{-1}(n+1)\beta}{6}, \qquad D_n^{\mathrm{op}}(\beta) \le \frac{(n-1)\beta}{2}. \tag{32}$$

The upper bound is Pinsker combined with the global KL bound. The expectation of $\mathsf{T}^{\mathrm{unit}}$ is increasing with β at all positive β. Consequently $D_n^{\mathrm{op}}(\beta) < 1/(6e)$ forces $\beta(n-1) < 1$, even without assuming monotonicity of D_n^{op} itself. In particular $n\beta \lesssim D_n^{\mathrm{op}}(\beta)$ in this small-budget region. Reflection gives the bounds for either sign.

For fixed, not necessarily small, finite KL budgets, the reversed copy law is a truncated geometric law. Its entropy is bounded by $-\log(1-e^{-\beta}) + \beta/(e^{\beta}-1) \le -\log(1-e^{-\beta}) + 1$. Thus

$$D_n^{\mathrm{KL}}(\beta) \ge \log\{n(1-e^{-|\beta|})\} - 1. \tag{33}$$

If $D_n^{\mathrm{KL}}(\beta) \le \tau_{\mathrm{cap}}$ and $n \to \infty$, then $1 - e^{-|\beta|} \le e^{\tau_{\mathrm{cap}}+1}/n$, giving $|\beta| = O_{\tau_{\mathrm{cap}}}(n^{-1})$. This implication does not presuppose small controls.

B.3 Proof of the zero-drift characterization

At fixed n, the feasible subset of the compact Θ is nonempty and closed, so it has a maximizer. Every point satisfying $\beta = 0$ and $\mathcal{K}_{\mathrm{phase}} \le \tau$ has $\mathcal{K}_n = \mathcal{K}_{\mathrm{phase}}$ and the same gain at every n. This proves the lower bound in (12). For an arbitrary subsequence approaching the limit superior, choose maximizers θ_n and extract a convergent subsequence $\theta_n \to \bar{\theta}$. Each $D_n^{\mathrm{KL}}(\beta_a(\theta_n)) \le \tau/w_a^{\mathrm{seq}}$; (33) and continuity imply $\beta_a(\bar{\theta}) = 0$. Also $\mathcal{K}_{\mathrm{phase}}(\theta_n) \le \tau$, hence $\mathcal{K}_{\mathrm{phase}}(\bar{\theta}) \le \tau$. The gain converges to a value at most the right side of (12). Together with the lower bound, this proves (12). At $\tau = 0$, positivity of the reference gives the same conclusion directly.

B.4 Proof of the drift-order transfer theorem

Let $w_{\min}^{\mathrm{seq}} = \min_a w_a^{\mathrm{seq}} > 0$. If $\mathcal{K}_n \le \tau$, each phase KL and each copy KL is at most $\tau/w_{\min}^{\mathrm{seq}}$. Choose τ_0 so small that Pinsker places every phase marginal in the region required by (13), and (31) gives $\|\beta\|_\infty \le C\sqrt{\tau}/n \le \epsilon_0$. For $h_{\max} = \max_{a,\ell}|g_a^\top z_{a,\ell}|$, bounded task scores and Pinsker give $G \le h_{\max}\sum_a w_a^{\mathrm{seq}}\|\nu_{a,\theta} - \nu_a^0\|_1 \le h_{\max}\sqrt{2\tau}$. For positive gain, the global drift inequality also gives $G \le C(\sqrt{\tau}/n)^{1/r}$. Zero or negative gain does not enlarge the supremum. These prove the KL upper bound.

Along the specified feasible curve, (11) and the global score upper bound yield $\mathcal{K}_n(\theta(\vartheta)) \le C(\vartheta^2 + n^2\vartheta^{2r})$. Set $\vartheta = c\min\{\sqrt{\tau}, \tau^{1/(2r)}n^{-1/r}\}$. A fixed sufficiently small c makes $\vartheta \le \vartheta_0$ and the KL at most τ for every $n \ge 2$, giving the matching lower bound from $G \ge c_0\vartheta$.

For $\mathcal{V}_n \le \delta$, marginalization contracts the ℓ_1 norm, so both phase and copy distances are at most $\delta/w_{\min}^{\mathrm{seq}}$. Choose δ_0 to apply (32) and the global drift condition. Then $G \le h_{\max}\delta$ and $\|\beta\|_\infty \le C\delta/n$, giving the operator upper bound. For the lower bound, product measures obey

$$\|q^{\mathrm{time}} \otimes \nu - \mathcal{U}_n \otimes \nu^0\|_1 \le \|q^{\mathrm{time}} - \mathcal{U}_n\|_1 + \|\nu - \nu^0\|_1.$$

The phase distance along the curve is at most $\sqrt{2\mathcal{K}_{\mathrm{phase}}/w_a^{\mathrm{seq}}}$, and (32) bounds the copy distance by $n|\beta|/2$. Thus $\mathcal{V}_n(\theta(\vartheta)) \le C(\vartheta + n\vartheta^r)$. The choice $\vartheta = c\min\{\delta, (\delta/n)^{1/r}\}$ completes the proof.

B.5 From reduced jets to a globally applicable drift inequality

For a nuisance Jacobian $J_{\mathrm{aux}} = D_{\theta_{\mathrm{aux}}}\beta(0)$ of full column rank, write $\theta^{\circ}_{\mathrm{aux}}(\vartheta)$ for the local solution of $J^{\dagger}_{\mathrm{aux}}\beta(\vartheta, \theta^{\circ}_{\mathrm{aux}}(\vartheta)) = 0$. The reduced drift condition used below is

$$(I - J_{\mathrm{aux}}J^{\dagger}_{\mathrm{aux}})\beta(\vartheta, \theta^{\circ}_{\mathrm{aux}}(\vartheta)) = a_r\vartheta^r + O(\vartheta^{r+1}), \qquad a_r \neq 0. \tag{34}$$

Proposition B.1 (Reduced drift, localization, and rank-based repair)**.** *Suppose the maps β, ν, G are C^{r+1} near $(\vartheta, \theta_{\mathrm{aux}}) = 0$, with $\beta(0) = 0$, $\nu(0) = \nu^0$, $G(0) = 0$, $\partial_\vartheta\beta(0) = 0$, and $\partial_\vartheta G(0) > 0$. Assume $J_{\mathrm{aux}} = D_{\theta_{\mathrm{aux}}}\beta(0)$ has full column rank (zero nuisance columns are allowed). Let $\theta^{\circ}_{\mathrm{aux}}(\vartheta)$ solve $J^{\dagger}_{\mathrm{aux}}\beta(\vartheta, \theta^{\circ}_{\mathrm{aux}}(\vartheta)) = 0$. If* (34) *holds, there is a neighborhood on which* (13) *holds for nonnegative gain, and the curve $(\vartheta, \theta^{\circ}_{\mathrm{aux}}(\vartheta))$ satisfies the lower-bound hypotheses of Theorem 4.2.*

To conclude its global two-budget rates, assume additionally that for some $\epsilon_0 > 0$, every *parameter with $G \geq 0$, $\max_a \|\nu_a - \nu^0_a\|_1 \leq \epsilon_0$, and $\|\beta\|_\infty \leq \epsilon_0$ lies in that neighborhood. On a compact control domain, uniqueness of the baseline preimage of $(\nu^0, 0)$ is one sufficient condition.*

Alternatively, if $D_{\theta_{\mathrm{aux}}}\beta(0)$ has full row rank and $\partial_\vartheta\beta(0) = 0$, there is a local curve with $\beta = 0$, gain of order $\vartheta > 0$, phase KL $O(\vartheta^2)$, and phase ℓ_1 change $O(\vartheta)$. It yields length-independent small-budget response.

Proof. Write $P_{\mathrm{range}\, J_{\mathrm{aux}}} \coloneqq J_{\mathrm{aux}}J^{\dagger}_{\mathrm{aux}}$. Since $J^{\dagger}_{\mathrm{aux}}J_{\mathrm{aux}} = I$, the implicit function theorem applied to $J^{\dagger}_{\mathrm{aux}}\beta$ gives $\theta^{\circ}_{\mathrm{aux}}(0) = 0$ and $\theta^{\circ\prime}_{\mathrm{aux}}(0) = 0$. More generally $(\vartheta, \theta_{\mathrm{aux}}) \mapsto (\vartheta, \zeta_{\mathrm{aux}})$, where $\zeta_{\mathrm{aux}} = J^{\dagger}_{\mathrm{aux}}\beta(\vartheta, \theta_{\mathrm{aux}})$, is a local diffeomorphism with inverse $\theta_{\mathrm{aux}}(\vartheta, \zeta_{\mathrm{aux}}) = \theta^{\circ}_{\mathrm{aux}}(\vartheta) + O(\|\zeta_{\mathrm{aux}}\|)$ uniformly near zero. The normal component along $\zeta_{\mathrm{aux}} = 0$ is $\beta_\perp(\vartheta) = (I - P_{\mathrm{range}\, J_{\mathrm{aux}}})\beta(\vartheta, \theta^{\circ}_{\mathrm{aux}}(\vartheta))$. Its norm is bounded below by $c|\vartheta|^r$. Smoothness gives

$$\|\beta_\perp(\vartheta)\| \leq \|(I - P_{\mathrm{range}\, J_{\mathrm{aux}}})\beta(\vartheta, \theta_{\mathrm{aux}})\| + C\|\zeta_{\mathrm{aux}}\| \leq C'\|\beta(\vartheta, \theta_{\mathrm{aux}})\|.$$

Hence $|\vartheta|^r \leq C\|\beta\|$. Since $\|\theta_{\mathrm{aux}}\| \leq C(\vartheta^2 + \|\beta\|)$ and $G(0) = 0$, we also have $|G| \leq C(|\vartheta| + \|\beta\|) \leq C'\|\beta\|^{1/r}$ locally. Norm equivalence in the fixed number of sequences proves (13). Along $\theta^{\circ}_{\mathrm{aux}}$, the tangential component of β is zero and the normal component is $O(\vartheta^r)$. Also $G = \partial_\vartheta G(0)\vartheta + O(\vartheta^2)$, while positive fixed phase probabilities imply $\mathcal{K}_{\mathrm{phase}} = O(\vartheta^2)$. This gives attainment.

The additional localization assumption extends the inequality to exactly the region quantified in Theorem 4.2. For the compact sufficient condition, any sequence outside the chosen open neighborhood with $(\nu, \beta) \to (\nu^0, 0)$ would have a convergent subsequence to another baseline preimage, a contradiction. For the full-row-rank alternative, select an invertible minor of $D_{\theta_{\mathrm{aux}}}\beta$, hold the other nuisance coordinates at zero, and solve $\beta = 0$ by the implicit function theorem. The condition $\partial_\vartheta\beta(0) = 0$ makes $\theta'_{\mathrm{aux}}(0) = 0$, preserving the positive gain derivative. □

For $r = 2$, the nonzero projected quadratic drift and the relevant rank conditions persist under sufficiently small smooth perturbations preserving the baseline and tangent identities. The global rates persist when the localization condition also remains valid.

B.6 Why local jets do not decide a global finite optimum

Fix any small $\varepsilon_{\mathrm{esc}} > 0$. Consider a two-phase recurrent family with phase candidates $(-1, 1)$ and

$$d_1(\vartheta) = 2\vartheta, \qquad d_2(\vartheta) = -2\vartheta - \beta(\vartheta), \qquad \beta(\vartheta) = \vartheta^2(1 - \vartheta/\varepsilon_{\mathrm{esc}}), \quad 0 \leq \vartheta \leq \varepsilon_{\mathrm{esc}}.$$

A fixed baseline log-retention such as -1 keeps all gates in $(0, 1)$ for sufficiently small $\varepsilon_{\mathrm{esc}}$; write weights are chosen as in the factorization proof. The phase law is proportional to $(e^{2\vartheta}, 1)$, so $G(\vartheta) = \tanh\vartheta$, while the copy drift is exactly $\beta(\vartheta)$. The tangent allocation-Fisher efficiency is one and the leading local drift is quadratic. Yet $\vartheta = \varepsilon_{\mathrm{esc}}$ has zero drift, gain $\tanh\varepsilon_{\mathrm{esc}}$, and KL $k(\varepsilon_{\mathrm{esc}})$ for *every* n. Choosing $k(\varepsilon_{\mathrm{esc}}) < \tau$ defeats a putative fixed-budget $n^{-1/2}$ bound inferred solely from that local jet. The escape point violates the globally quantified drift inequality of Theorem 4.2. At still smaller budgets excluding this point, a local law may hold.

For comparison, $\beta(\vartheta) = \vartheta^r$ on a fixed small interval $0 \leq \vartheta \leq \vartheta_0$, with $d_1 = 2\vartheta$, $d_2 = -2\vartheta - \vartheta^r$, satisfies the global drift hypothesis for every integer $r \geq 2$. Its exact cost is $k(\vartheta) + D^{\mathrm{KL}}_n(\vartheta^r)$ and gain $\tanh\vartheta$. This realizes all the drift-order exponents in (14) by genuine recurrent controls. For $r > 2$, the second-order statistical curvature at the base vanishes, but a higher-order finite obstruction remains.

C SIGMOID SHARING: GLOBAL BARRIERS AND SAME-BASELINE REPAIRS

C.1 NONPERIODIC SIGN-BALANCED PAIRS

We prove Theorem 4.3. Fix $0 < c_{\min} \le c_{\max} \le 1$ and $0 < f_- \le f_* \le f_+ < 1$. Constants below depend only on these intervals, not on T, the amplitude array, or sufficiently small budgets. The supremum is over all real ϑ in the one-feature family.

C.1.1 CONCAVITY, EXACT PAIRING, AND GLOBAL LOCALIZATION

Put $a_* = 1 - f_*$, $u_*^{\mathrm{ret}} = \operatorname{logit} f_*$, and

$$e_f(a) = a - \{\ell_{\mathrm{sig}}(u_*^{\mathrm{ret}} + a/a_*) - \ell_{\mathrm{sig}}(u_*^{\mathrm{ret}})\}.$$

Since $\ell'_{\mathrm{sig}}(u_*^{\mathrm{ret}}) = a_*$ and ℓ_{sig} is strictly concave, $e_f(a) \ge 0$, with equality only at zero. Its second derivative at zero is $f_*/a_* > 0$. Compactness of the gate interval and Taylor's theorem give constants $c_{\mathrm{rem},-}, c_{\mathrm{rem},+}, a_0 > 0$ with

$$c_{\mathrm{rem},-} a^2 \le e_f(a) \le c_{\mathrm{rem},+} a^2 \qquad (|a| \le a_0). \tag{35}$$

Let $v_i = z_{i+1} - z_i$ for $1 \le i < T$. The sign-paired candidates give

$$v_{2j-1} = 2c_j \quad (1 \le j \le n), \qquad v_{2j} = -(c_j + c_{j+1}) \quad (1 \le j < n).$$

Thus $2c_{\min} \le |v_i| \le 2c_{\max}$ for every gate index $1 \le i < T$. Write

$$d_i(\vartheta) = \ell_{\mathrm{sig}}(u_*^{\mathrm{ret}} + \vartheta v_i/a_*) - \log f_* = \vartheta v_i - e_f(\vartheta v_i), \qquad E_i^{\mathrm{cum}}(\vartheta) = \sum_{k=i}^{T-1} e_f(\vartheta v_k), \quad E_T^{\mathrm{cum}} = 0.$$

Telescoping $Hv = -z + z_T \mathbf{1}$ gives terminal scores, modulo a constant,

$$s_i(\vartheta) = -\vartheta z_i - E_i^{\mathrm{cum}}(\vartheta). \tag{36}$$

For pair j, put $e_j = e_f(2c_j\vartheta)$. The two scores are $\vartheta c_j - E_{2j}^{\mathrm{cum}} - e_j$ and $-\vartheta c_j - E_{2j}^{\mathrm{cum}}$. Thus the pair marginal p_j^{pair} and its conditional negative mean are

$$\begin{aligned}
p_j^{\mathrm{pair}} &= \operatorname{softmax}(s^{\mathrm{pair}})_j, & s_j^{\mathrm{pair}} &= -E_{2j}^{\mathrm{cum}} + s_j^{\mathrm{pair,loc}}, \\
s_j^{\mathrm{pair,loc}} &= -e_j/2 + \log\cosh(c_j\vartheta - e_j/2), & & \\
-\mathbb{E}[z \mid j] &= c_j \tanh(d_{2j-1}(\vartheta)/2), & & \\
d_{2j-1}(\vartheta) &= \ell_{\mathrm{sig}}(u_*^{\mathrm{ret}} + 2c_j\vartheta/a_*) - \log f_*. & &
\end{aligned} \tag{37}$$

If $\vartheta < 0$, every conditional mean is positive, so the negative gain is negative. Such parameters cannot improve the supremum over the feasible baseline. For $\vartheta \ge 0$, all pair gains are nonnegative and increasing lower bounds follow from monotonicity of ℓ_{sig}.

Fix $\vartheta_0 > 0$ so small that $2c_{\max}\vartheta_0 \le a_0$. If $\vartheta \ge \vartheta_0$, then uniformly over f_*, c_j, $d_{2j-1}(\vartheta) \ge d_0 > 0$. The KL chain rule with respect to the uniform pair/sign baseline gives

$$D_{\mathrm{KL}}(p(\vartheta)\|\mathcal{U}_T) = D_{\mathrm{KL}}(p^{\mathrm{pair}}\|\mathcal{U}_n) + \sum_j p_j^{\mathrm{pair}} k(d_{2j-1}(\vartheta)/2) \ge k(d_0/2) > 0.$$

Also $G(\vartheta) \ge c_{\min}\tanh(d_0/2)$, whereas $G(\vartheta) \le c_{\max}\|p(\vartheta) - \mathcal{U}_T\|_1$. Therefore sufficiently small KL or operator budgets force *every* positive-gain feasible parameter into $0 \le \vartheta < \vartheta_0$, independently of length. This is the global localization step. It excludes saturation and remote negative or positive parameters before any small-ϑ expansion is used.

Shrinking ϑ_0 if necessary, on this interval there are constants $g_-, g_+ > 0$ such that

$$g_-\vartheta \le G(\vartheta) = \sum_j p_j^{\mathrm{pair}} c_j \tanh(d_{2j-1}(\vartheta)/2) \le g_+\vartheta. \tag{38}$$

These bounds hold for *any* pair marginal p^{pair}. Conditional sign KL also yields $D_{\mathrm{KL}}(p(\vartheta)\|\mathcal{U}_T) \ge c_{\mathrm{cmp}}\vartheta^2$, so $\vartheta \lesssim \sqrt{\tau}$ on the KL-feasible set. The operator inequality and (38) give $\vartheta \lesssim \delta$ under the operator budget.

C.1.2 CUMULATIVE DRIFT GIVES THE SECOND UPPER BOUND

For $0 \le \vartheta \le \vartheta_0$, (35) implies $c_{\mathrm{acc},-}\vartheta^2 \le e_f(\vartheta v_i) \le c_{\mathrm{acc},+}\vartheta^2$ uniformly, and (37) implies $|s_j^{\mathrm{pair,loc}}| \le C_{\mathrm{pair}}\vartheta^2$. For any $k > j$,

$$s_k^{\mathrm{pair}} - s_j^{\mathrm{pair}} = \sum_{i=2j}^{2k-1} e_f(\vartheta v_i) + s_k^{\mathrm{pair,loc}} - s_j^{\mathrm{pair,loc}} \ge \{2c_{\mathrm{acc},-}(k-j) - 2C_{\mathrm{pair}}\}\vartheta^2. \tag{39}$$

Choose a fixed N large enough that, for $n \ge N$, the first and last $q = \lfloor n/4 \rfloor$ pair indices, denoted A, B, satisfy $s_k^{\mathrm{pair}} - s_j^{\mathrm{pair}} \ge c_{\mathrm{cmp}} n\vartheta^2$ for every $j \in A, k \in B$. Their cardinalities are equal and their baseline masses are $q/n \ge 1/8$. Therefore

$$\frac{p^{\mathrm{pair}}(B)}{p^{\mathrm{pair}}(A)} \ge e^{c_{\mathrm{cmp}} n\vartheta^2}. \tag{40}$$

Under $D_{\mathrm{KL}}(p\|\mathcal{U}_T) \le \tau$, marginalization and Pinsker give $|p^{\mathrm{pair}}(A) - q/n|, |p^{\mathrm{pair}}(B) - q/n| \le \sqrt{\tau/2}$. For a sufficiently small universal-in-n τ_0, both masses remain bounded away from zero and $\log(p^{\mathrm{pair}}(B)/p^{\mathrm{pair}}(A)) \le C\sqrt{\tau}$. Hence $n\vartheta^2 \lesssim \sqrt{\tau}$. Under $\|p - \mathcal{U}_T\|_1 \le \delta$, the event mass errors are at most $\delta/2$; the same argument gives $n\vartheta^2 \lesssim \delta$.

For $2 \le n < N$, the already proved bounds $\vartheta \lesssim \sqrt{\tau}$ and $\vartheta \lesssim \delta$ imply $n\vartheta^2 \lesssim \sqrt{\tau}$ and $n\vartheta^2 \lesssim \delta$ after adjusting constants on this finite set, taking budgets at most one. Combining with (38) proves both upper rates in (16). The argument uses the deterministic event-mass comparison (40).

C.1.3 MATCHING LOWER BOUNDS AND EXACT NONLINEAR REPAIR

For small positive ϑ, the score oscillation in (36) is at most $C(\vartheta + T\vartheta^2)$, since $|z_i| \le c_{\max}$ and E_i^{cum} is a sum of nonnegative remainders. Consequently the global score bound and Pinsker yield

$$D_{\mathrm{KL}}(p(\vartheta)\|\mathcal{U}_T) \le C(\vartheta^2 + T^2\vartheta^4), \qquad \|p(\vartheta) - \mathcal{U}_T\|_1 \le C(\vartheta + T\vartheta^2).$$

Select $\vartheta = c_{\mathrm{cmp}} \min\{\sqrt{\tau}, \tau^{1/4}/\sqrt{T}\}$ or $\vartheta = c_{\mathrm{cmp}} \min\{\delta, \sqrt{\delta/T}\}$, respectively, with a uniform small positive c_{cmp}. These controls are feasible, and (38) yields the lower rates. This completes the original-family proof, including the global supremum.

At the baseline, $D_\vartheta b = v$, so the centered score is $-z$ and $F_\vartheta = \mathbb{E}_{\mathcal{U}_T} z^2 = n^{-1}\sum_j c_j^2$. The output derivative is $-F_\vartheta$ and hence its allocation-Fisher squared-response coefficient is $F_\vartheta^2/F_\vartheta = F_\vartheta$, exactly the free covariance. This proves unit local task efficiency without degenerating the candidate variance.

For the nonlinear repair use $u_i^{\mathrm{ret}}(\lambda) = u_*^{\mathrm{ret}} + \psi(\lambda v_i)$ with ψ in (18). Whenever $\lambda\|v\|_\infty < -\log f_*$, its retention increments equal λv, so the allocation is exactly $q_\lambda \propto e^{-\lambda z}$. The candidate array is sign balanced, so $m^0 = 0$, $\mathrm{Var}_{\mathcal{U}_T}(z) \ge c_{\min}^2$, and for small uniform $\lambda > 0$ the negative mean is $\asymp \lambda$. This can also be seen by integrating $-m'(\lambda) = \mathrm{Var}_{q_\lambda}(z)$ and comparing its density to $\mathcal{U}_T$. Score oscillation bounds the KL by $C\lambda^2$ and the operator cost by $C\lambda$. The choices $\lambda \asymp \sqrt{\tau}$ and $\lambda \asymp \delta$ attain length-independent rates; bounded candidates and Pinsker give the matching unrestricted-distribution upper bounds.

These choices have a uniform gate-domain margin. The repair uses the nonlinear inverse map ψ. For a fixed linear-logit feature space, Theorem 5.1 instead characterizes exact closure through the power-evaluation span of the adjacent differences, whose dimension depends on the amplitude array.

C.1.4 UNIFORMITY OVER SIGN-BALANCED AMPLITUDE ARRAYS

The bounds hold uniformly over all amplitude arrays in the stated interval, including deterministic, adversarial and nonstationary choices. Paired signs give a zero baseline mean within each pair, and the positive amplitude lower bound gives an unavoidable second-order retention error at every gate. These properties let the quarter-event comparison control useful output through ϑ. Without sign balancing, redistribution between temporal regions can directly change the task mean.

Finite-length illustrations. Figure 3 evaluates feasible scalar controls from (15) at $T = 16, 64, 256, 1024, 4096$, with $n = T/2$, $f_* = 0.5$ and KL cap 10^{-4}. The candidates remain $z_{2k-1} = -c_k$, $z_{2k} = c_k$ for $k = 1, \dots, n$. The three amplitude rules are

$$c_k^{\text{random}} = 0.35 + 0.65U_{k-1},$$
$$c_k^{\text{ramp}} = 0.35 + 0.65\frac{k-1}{n-1},$$
$$c_k^{\text{rotation}} = 0.35 + 0.65\{(k-1)\sqrt{2}\},$$

where the $U_j \in [0, 1]$ are the pseudorandom values used in the experiment and braces denote fractional part. The amplitudes lie in $[0.35, 1]$; the displayed gains are $-m_T$ at the feasible scalar updates. Appendix M.2 gives the seed, generation order and numerical checks. Theorem 4.3 provides the uniform family rates.

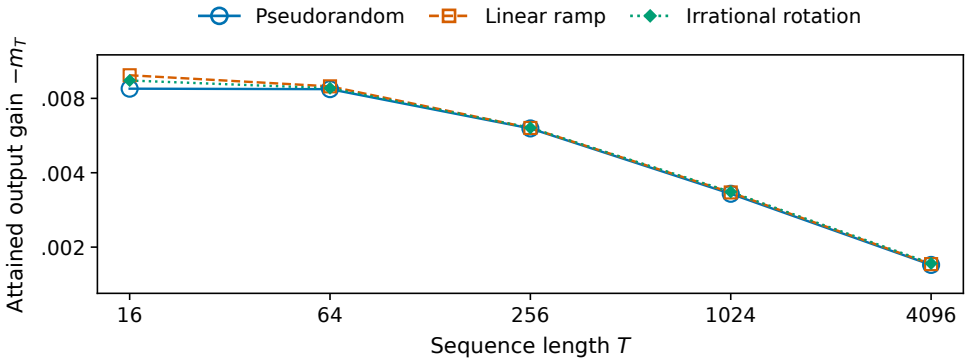


Figure 3: Feasible responses for three nonperiodic amplitude-generation rules. The candidate signs remain paired, $z_{2k-1} = -c_k$ and $z_{2k} = c_k$, with $c_k \in [0.35, 1]$ as defined above. At $f_* = 0.5$ and actual KL cap 10^{-4}, points give the attained gain $-m_T$ at feasible scalar controls for $T = 16, 64, 256, 1024, 4096$. Theorem 4.3 gives the uniform family rates; Appendix M.2 records the generation and numerical checks. The three arrays have their own candidate variances.

C.2 Canonical boundary optima and uniform crossovers

Setting and exact response comparison. The unit-amplitude alternating example, called canonical below, sets $c_k = 1$ and $f_* = 1/2$ in Theorem 4.3. The families are $\mathcal{F}_1 : u_j^{\text{ret}} = 2\vartheta v_j$ and its bias extension $\mathcal{F}_2 : u_j^{\text{ret}} = 2\vartheta v_j + \theta_{\text{bias}}$. Their domains are $\Theta_1 = \mathbb{R}$ for $\theta = \vartheta$ and $\Theta_2 = \mathbb{R}^2$ for $\theta = (\vartheta, \theta_{\text{bias}})$. Sigmoid gates and (3) induce $\pi_k(\theta), m_{k,T}(\theta)$. Define

$$B_k(T, \tau) = \sup_{\theta \in \Theta_k : D_{\text{KL}}(\pi_k(\theta) \| \mathcal{U}_T) \le \tau} [-m_{k,T}(\theta)]. \tag{41}$$

Define $B_k^{\text{op}}(T, \delta)$ using $\|\pi_k(\theta) - \mathcal{U}_T\|_1 \le \delta$. Let $k(x) = x \tanh x - \log\cosh x$ and $k(\lambda_\tau) = \tau$, $\lambda_\tau > 0$.

Theorem C.1 (Canonical sharp crossover)**.** *Both families have local scalar response coefficient $K = 1$. Uniformly over even $T \ge 4$ and sufficiently small positive budgets, the one-feature responses satisfy* (16)*, with the sharp uniform refinement in Appendix C.2. The repaired responses are $B_2 = \tanh \lambda_\tau$ and $B_2^{\text{op}} = \delta$.*

The crossover is between the two scaling regimes of (16): $T\sqrt{\tau} \asymp 1$ for KL, $T\delta \asymp 1$ for operator cost; it is not a curve intersection. Without a sensitivity cap, $\mathcal{F}_1$ has mean $m_T(\vartheta) = \tanh\{-\vartheta + \frac{1}{2}\log\cosh(2\vartheta)\}$, independent of T with derivative -1 at zero. At fixed T, $B_1(T, \tau)/B_2(T, \tau) \to 1$ as $\tau \to 0$; at fixed small positive budget it vanishes with length. Figure 4 shows this nonuniform limit.

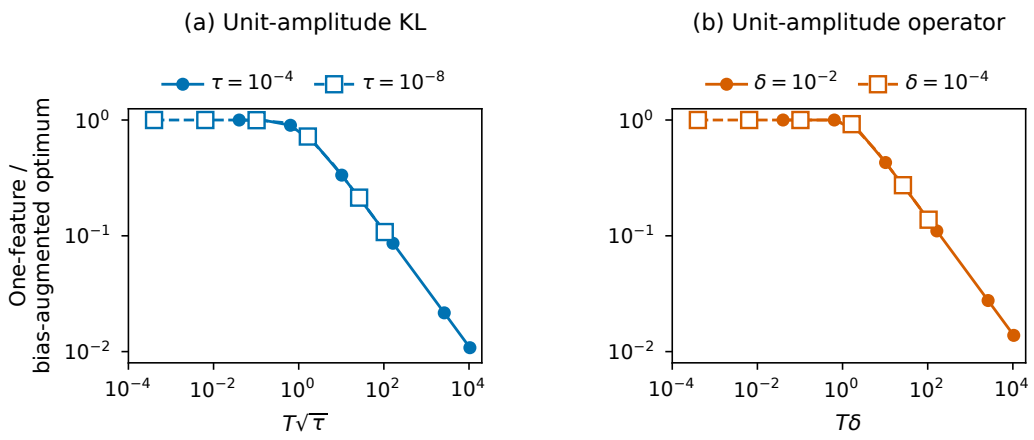


Figure 4: **Unit-amplitude canonical optima.** One-feature response divided by shared-bias-augmented response at matched actual (a) KL or (b) operator cost. Both have unit local task efficiency. Each curve contains all ten $T = 4^k$, $k = 1, \ldots, 10$, points, with six markers. Filled/solid and open/dashed identify larger and smaller budgets. Appendix C.2 characterizes the optima.

The sharp refinement of Theorem C.1, uniform over even $T \geq 4$ and sufficiently small $\tau > 0$, is

$$B_1 = \left(\frac{4\tau}{1 + \sqrt{1 + \frac{8}{3}(T^2 - 4)\tau}} \right)^{1/2} \{1 + O(\sqrt{\tau})\}, \qquad B_2 = \tanh \lambda_\tau. \tag{42}$$

For the canonical model of Theorem C.1, the exact log-sigmoid identity gives

$$b - b^0 = \vartheta v - a(\vartheta)\mathbf{1}, \qquad a(\vartheta) = \log\cosh(2\vartheta) = 2\vartheta^2 + O(\vartheta^4). \tag{43}$$

Consequently

$$\pi_i(\vartheta) \propto e^{-\vartheta z_i - a(\vartheta)(T-i)}, \qquad m_T(\vartheta) = \tanh s^{\mathrm{sgn}}(\vartheta), \quad s^{\mathrm{sgn}}(\vartheta) = -\vartheta + a(\vartheta)/2. \tag{44}$$

The pair-index and sign law factor exactly, and for $n = T/2$,

$$D_{\mathrm{KL}}(\pi(\vartheta) \| \mathcal{U}_T) = k(s^{\mathrm{sgn}}(\vartheta)) + D_n^{\mathrm{KL}}(2a(\vartheta)), \qquad k(x) = x\tanh x - \log\cosh x. \tag{45}$$

These formulas yield the globally certified scalar optimum and the sharp uniform KL expansion below. The canonical operator optimum is proved in Appendix C.2.4; the uniform two-amplitude rates are proved in Appendix C.3.2. Write strengths are represented in log space for numerical evaluation.

C.2.1 A TILTED-UNIFORM CALCULATION

Let $\mathcal{U}_n$ be uniform on $\{0, \ldots, n-1\}$, $n \geq 2$, and $q_{n,\beta}^{\mathrm{time}}(j) \propto e^{\beta j}$. Set $D_n^{\mathrm{KL}}(\beta) = D_{\mathrm{KL}}(q_{n,\beta}^{\mathrm{time}} \| \mathcal{U}_n)$. Reflection of the index shows D_n^{KL} is even. For $\beta > 0$, reversing the index gives a truncated geometric distribution with ratio $e^{-\beta}$ and

$$\mathcal{Z}_n(\beta) = \frac{1 - e^{-n\beta}}{1 - e^{-\beta}},$$
$$D_n^{\mathrm{KL}}(\beta) = \log n - \log \mathcal{Z}_n(\beta) - \beta \left\{ \frac{1}{e^\beta - 1} - \frac{n}{e^{n\beta} - 1} \right\}. \tag{46}$$

The continuous value at $\beta = 0$ is zero. This formula is evaluated with $\mathrm{expm1}$ and increased precision, rather than subtracting almost equal probability sums.

Writing $A_n(\beta) = \log(n^{-1} \sum_j e^{\beta j})$, one has

$$D_n^{\mathrm{KL}}(\beta) = \beta A_n'(\beta) - A_n(\beta) = \int_0^\beta s \, \mathrm{Var}_{q_{n,s}^{\mathrm{time}}}(j)\, ds \quad (\beta \geq 0). \tag{47}$$

Thus D_n^{KL} is strictly increasing for positive β. The two-sided coarse estimate (31) is proved in Appendix B.2; it implies $|\beta|(n-1) = O(\sqrt{\tau})$ whenever $D_n^{\mathrm{KL}}(\beta) \le \tau < 1/(24\mathrm{e})$. Replacing the fixed density-ratio bounds there by $e^{\pm|\beta|(n-1)}$ sharpens the comparison to

$$D_n^{\mathrm{KL}}(\beta) = \frac{(n^2-1)\beta^2}{24}\{1 + O(\sqrt{\tau})\}, \qquad D_n^{\mathrm{KL}}(\beta) \le \tau, \tag{48}$$

with a universal remainder constant for sufficiently small τ.

For arbitrary fixed finite budgets, the entropy bound (33) from Appendix B.2 gives $|\beta| = O_C(n^{-1})$ when $D_n^{\mathrm{KL}}(\beta) \le C$ and $n \to \infty$. This restriction does not assume a local parameter expansion.

C.2.2 Exact alternating factorization

At the baseline $f = 1/2$, every unnormalized terminal write mass is $2^{-(T-1)}$, so $\pi^0 = \mathcal{U}_T$. The same holds within every prefix. For $v_j = -(z_j - z_{j+1})$, the log-sigmoid identity gives (43). Since $Hv = -z + z_T\mathbf{1}$ and $(H\mathbf{1})_i = T - i$, it gives (44).

Within pair $(2j+1, 2j+2)$ the sign scores differ by $2s^{\mathrm{sgn}}(\vartheta)$, while increasing the pair index changes the score by $2a(\vartheta)$. Thus the joint law is a product of a binary sign law and $q^{\mathrm{time}}_{n,2a(\vartheta)}$. The binary mean is $\tanh s^{\mathrm{sgn}}(\vartheta)$, and its KL from uniform is $k(s^{\mathrm{sgn}}(\vartheta))$, proving (45).

For $\vartheta < 0$, $s^{\mathrm{sgn}}(\vartheta) > 0$, so such parameters do not maximize negative output. For $\vartheta \ge 0$, $(s^{\mathrm{sgn}})'(\vartheta) = -1 + \tanh(2\vartheta) < 0$, s^{sgn} decreases from zero to $-\log 2/2$, and $a(\vartheta)$ strictly increases. Both KL terms strictly increase: $k'(x) = x\,\mathrm{sech}^2\,x$ has the sign of x, and (47) handles the other term. Consequently a sufficiently small budget has a unique boundary parameter $\vartheta_{T,\tau} > 0$ and the *global* one-feature optimum is $B_1 = -\tanh s^{\mathrm{sgn}}(\vartheta_{T,\tau})$. This establishes the global supremum in (41).

At zero, the ϑ-score is $Hv = -z + \mathbf{1}$, modulo its mean. Its variance is one and the output derivative is -1. Therefore the one-feature local coefficient is one. The two-feature family contains that score direction and cannot exceed the free-distribution covariance one, so its coefficient is also one. The local comparison uses each family's induced allocation-Fisher metric.

C.2.3 Proof of the uniform length–budget law

Choose a universal $\tau_0 > 0$ smaller than $1/(24e)$ and $k(s^{\mathrm{sgn}}(\vartheta_0))$ for a fixed sufficiently small $\vartheta_0 > 0$. Feasibility implies $0 \le \vartheta \le \vartheta_0$ by the monotonicity of the sign KL, and $|s^{\mathrm{sgn}}(\vartheta)| \asymp \vartheta$. Therefore $\vartheta = O(\sqrt{\tau})$ uniformly. With $\beta = 2a(\vartheta)$, the copy KL is at most τ, so the bounds in Appendix B.2 additionally imply $\beta(n-1) = O(\sqrt{\tau})$. Uniformly on these feasible parameters,

$$k(s^{\mathrm{sgn}}(\vartheta)) = \tfrac{1}{2}\vartheta^2\{1 + O(\sqrt{\tau})\}, \quad \beta = 4\vartheta^2\{1 + O(\vartheta^2)\}, \quad D_n^{\mathrm{KL}}(\beta) = \frac{T^2-4}{6}\vartheta^4\{1 + O(\sqrt{\tau})\}.$$

Their sum gives

$$D_{\mathrm{KL}}(\pi(\vartheta)\|\pi^0) = \left\{\tfrac{1}{2}\vartheta^2 + \frac{T^2-4}{6}\vartheta^4\right\}\{1 + O(\sqrt{\tau})\}, \tag{49}$$

uniformly in T on the small-budget region. Both terms are nonnegative, so separate relative estimates give a relative estimate for their sum. At the boundary the left side equals τ.

Let $\vartheta_*^2 = 4\tau/[1 + \sqrt{1 + (8/3)(T^2-4)\tau}]$, the positive solution of $\vartheta_*^2/2 + (T^2-4)\vartheta_*^4/6 = \tau$. Rescaling ϑ by $1+\delta$ rescales this positive polynomial between factors $(1+\delta)^2$ and $(1+\delta)^4$ for $\delta \ge 0$, with the reversed bounds for $-1 < \delta < 0$. Thus (49) implies $\vartheta_{T,\tau}/\vartheta_* = 1 + O(\sqrt{\tau})$ with universal constants. Finally $-\tanh s^{\mathrm{sgn}}(\vartheta) = \vartheta\{1 + O(\vartheta)\}$, proving the uniform refinement

$$B_1(T,\tau) = \left\{\frac{4\tau}{1 + \sqrt{1 + (8/3)(T^2-4)\tau}}\right\}^{1/2}\{1 + O(\sqrt{\tau})\}. \tag{50}$$

Because $T^2 - 4 \asymp T^2$ for $T \ge 4$, the two-regime comparison in Theorem C.1 follows, with constants independent of both T and τ.

For the second family take $0 < \lambda < \log 2/2$. The target $q_\lambda \propto e^{-\lambda z}$ requires $b_j^* = -\log 2 + \lambda v_j < 0$. Its required logit increments have two values, hence belong to $\mathrm{span}\{v, \mathbf{1}\}$. They are realized by

finite $(\vartheta, \theta_{\text{bias}})$ independent of T. Their output and KL are $-\tanh\lambda$ and $k(\lambda)$. Conversely, data processing from write index to sign gives $D_{\text{KL}}(p\|\mathcal{U}_T) \geq k(\operatorname{arctanh} m)$ for every distribution with scalar mean m. Since k increases with the magnitude of its argument, no p of KL at most τ can have $-m > \tanh\lambda_\tau$. Choose τ_0 smaller if necessary to ensure $\lambda_{\tau_0} < \log 2/2$. This proves the exact formula for B_2 and completes Theorem C.1.

For fixed T, (50) and $k(\lambda) = \lambda^2/2 + O(\lambda^4)$ give $B_1/B_2 \to 1$. For fixed small τ, (31) or (33) gives the $T^{-1/2}$ upper bound; a feasible $\vartheta = c/\sqrt{T}$ gives the matching lower bound. Hence the reversed iterated limit of the ratio is zero. The joint approximation, rather than a fixed-T fourth-order expansion alone, justifies this statement.

C.2.4 Global operator optimum in the canonical plot

The operator optima in Figure 4 follow from monotonicity of the exact finite-threshold formula below. For $\vartheta \geq 0$, both the heavier sign probability and the copy slope $2a(\vartheta)$ increase strictly.

Reverse the copy index so $q^{\text{time}}(k) = e^{-\beta k}/\mathcal{Z}_n$, let p_+, p_- be the sign probabilities, and define

$$N_s^{\text{tail}} = \min\left\{n, \max\left\{0, \left\lceil \frac{\log(2np_s/\mathcal{Z}_n)}{\beta} \right\rceil\right\}\right\}, \qquad \mathcal{Z}_n = \frac{1-e^{-n\beta}}{1-e^{-\beta}}.$$

These count the atoms strictly above $1/(2n)$; including an equality does not change the formula. Then

$$\|q^{\text{time}} \otimes p - \mathcal{U}_{2n}\|_1 = 2 \sum_{s\in\{+,-\}} \left[p_s \frac{1-e^{-\beta N_s^{\text{tail}}}}{1-e^{-n\beta}} - \frac{N_s^{\text{tail}}}{2n} \right].$$

Between threshold changes, any partial copy tail increases strictly with β; the contribution from increasing sign bias is nonnegative because the larger sign probability has at least as large a threshold count. If all thresholds are endpoints, the nonzero distance is the strictly increasing sign bias. Continuity at thresholds completes the monotonicity argument. At $\beta = 0$, the expression is interpreted as the binary distance. Thus a small positive operator budget has the boundary optimum shown in the plot, and negative ϑ has the wrong gain sign.

For the canonical one-feature model, $\beta = 2\log\cosh(2\vartheta) \asymp \vartheta^2$ and $G = -\tanh(-\vartheta + \log\cosh(2\vartheta)/2) \asymp \vartheta$ near zero. The sign marginal globally localizes every small positive-gain feasible point. The operator law follows from Theorem 4.2, or as a special case of Theorem 4.3. In the repaired family, $q_\lambda \propto e^{-\lambda z}$ has $\|q_\lambda - \mathcal{U}_T\|_1 = \tanh\lambda$ and gain $\tanh\lambda$. For small δ take $\lambda = \operatorname{arctanh}\delta < \log 2/2$, giving $B_2^{\text{op}} = \delta$; no distribution can exceed this because $|z_i| = 1$.

C.3 Two amplitudes and a shared bias

C.3.1 A common bias and non-half baseline gates

We prove Theorem 3.1 for arbitrary fixed $0 < f_* < 1$ and $0 < c_1 < c_2 \leq 1$. Write $a_* = 1 - f_*$, $o_{\text{bias},0} = a_*/f_*$, and $u_c^{\text{amp}} = 2c\vartheta/a_*$. In (8), the two logit values in a pair are $u_*^{\text{ret}} + \theta_{\text{bias}} \pm u_c^{\text{amp}}$. Define their log-retention increments

$$\begin{aligned} b_{c,+} &= \ell_{\text{sig}}(u_*^{\text{ret}} + \theta_{\text{bias}} + u_c^{\text{amp}}) - \log f_*, \\ b_{c,-} &= \ell_{\text{sig}}(u_*^{\text{ret}} + \theta_{\text{bias}} - u_c^{\text{amp}}) - \log f_*, \\ \bar{d}_c^{\text{ret}} &= (b_{c,+} + b_{c,-})/2. \end{aligned}$$

The first gate of every pair has increment $b_{c,+}$, giving sign parameter $s_c^{\text{sgn}} = -b_{c,+}/2$. Pair-index slope is $-2\bar{d}_c^{\text{ret}}$; reflection makes its KL insensitive to the sign. Direct pairing yields

$$m_c = c\tanh s_c^{\text{sgn}}, \qquad D_{\text{KL}}(\pi_c\|\mathcal{U}_T) = k(s_c^{\text{sgn}}) + D_n^{\text{KL}}(2\bar{d}_c^{\text{ret}}). \tag{51}$$

If the average KL is at most fixed τ, each term is at most 2τ. By (33), $|\bar{d}_{c_i}^{\text{ret}}| = O_\tau(T^{-1})$ for every feasible parameter pair, including pairs depending on T. This global restriction precedes any expansion at the baseline.

Put $o_{\text{bias}} = e^{-u_*^{\text{ret}} - \theta_{\text{bias}}} = o_{\text{bias},0}e^{-\theta_{\text{bias}}}$. The product of the two sigmoid probabilities gives the exact equations

$$1 + 2o_{\text{bias}}\cosh u_{c_i}^{\text{amp}} + o_{\text{bias}}^2 = f_*^{-2}e^{-2\bar{d}_{c_i}^{\text{ret}}}, \qquad i = 1, 2. \tag{52}$$

Since $\cosh \geq 1$, $o_{\mathrm{bias}} \leq o_{\mathrm{bias},0} + O(T^{-1})$. From the equation for c_2, $o_{\mathrm{bias}} \cosh u_{c_2}^{\mathrm{amp}} \geq o_{\mathrm{bias},0} + O(T^{-1})$, which is bounded below by a positive constant for large T. Subtracting the two equations gives

$$o_{\mathrm{bias}}\{\cosh u_{c_2}^{\mathrm{amp}} - \cosh u_{c_1}^{\mathrm{amp}}\} = O(T^{-1}).$$

Thus $\cosh u_{c_1}^{\mathrm{amp}} / \cosh u_{c_2}^{\mathrm{amp}} = 1 - O(T^{-1})$. For $c_1 < c_2$ this ratio is strictly below one at every nonzero ϑ, bounded away from one on any compact region excluding zero, and tends to zero as $|\vartheta| \to \infty$. It follows that $\vartheta \to 0$. The product equations then force $o_{\mathrm{bias}} \to o_{\mathrm{bias},0}$, hence $\theta_{\mathrm{bias}} \to 0$. With o_{bias} now bounded below, the elementary series inequality

$$\cosh(2c_2\vartheta/a_*) - \cosh(2c_1\vartheta/a_*) \geq \frac{2(c_2^2 - c_1^2)}{a_*^2}\vartheta^2$$

proves $\vartheta^2 = O(T^{-1})$. Comparing (52) with its baseline quadratic $(1 + o_{\mathrm{bias},0})^2 = f_*^{-2}$ gives $o_{\mathrm{bias}} - o_{\mathrm{bias},0} = O(\vartheta^2 + T^{-1})$ and therefore $\theta_{\mathrm{bias}} = O(T^{-1})$. Equation (51) now implies $|m_c| = O(T^{-1/2})$. This bounds the supremum over the entire feasible parameter set.

For the lower bound, set $\theta_{\mathrm{bias}} = 0$ and $\vartheta = c_{\mathrm{step}}/\sqrt{T}$ with $c_{\mathrm{step}} > 0$ small. Taylor expansion gives $\bar{d}_c^{\mathrm{ret}} = -2f_* c^2 \vartheta^2 / a_* + O(\vartheta^4)$ and $s_c^{\mathrm{sgn}} = -c\vartheta + O(\vartheta^2)$. The sign KL is $O(c_{\mathrm{step}}^2/T)$, while the copy-score oscillation is $O(c_{\mathrm{step}}^2)$ and its KL is $O(c_{\mathrm{step}}^4)$. Choosing c_{step} small enough places the average below τ for all sufficiently large T. The negative mean equals $\frac{1}{2}(c_1^2 + c_2^2)c_{\mathrm{step}}/\sqrt{T} + O(T^{-1})$, giving the matching lower bound.

Constructive squared-feature repair. Fix a sufficiently small $\vartheta > 0$, independent of T, and put $c_*^{\mathrm{prod}} = f_*^{-2} - 1$. The positive solution of the zero-drift equation in o_{bias} is

$$o_{\mathrm{bias},c} = \frac{c_*^{\mathrm{prod}}}{\cosh u_c^{\mathrm{amp}} + \sqrt{\cosh^2 u_c^{\mathrm{amp}} + c_*^{\mathrm{prod}}}}, \qquad \theta_{\mathrm{bias},c}(\vartheta) = -\log(o_{\mathrm{bias},c}/o_{\mathrm{bias},0}). \tag{53}$$

These formulas enforce $\mathrm{sigm}(u_*^{\mathrm{ret}} + \theta_{\mathrm{bias},c} + u_c^{\mathrm{amp}})\,\mathrm{sigm}(u_*^{\mathrm{ret}} + \theta_{\mathrm{bias},c} - u_c^{\mathrm{amp}}) = f_*^2$, equivalently $\bar{d}_c^{\mathrm{ret}} = 0$. Solve the two linear equations

$$\theta_{\mathrm{bias}} + 4\theta_{\mathrm{sq}} c_i^2 = \theta_{\mathrm{bias},c_i}(\vartheta), \quad i = 1, 2,$$

by $\theta_{\mathrm{sq}} = (\theta_{\mathrm{bias},c_2} - \theta_{\mathrm{bias},c_1})/[4(c_2^2 - c_1^2)]$ and $\theta_{\mathrm{bias}} = \theta_{\mathrm{bias},c_1} - 4\theta_{\mathrm{sq}} c_1^2$. This is precisely one added $(v^{(c)})^2$ feature. All copy KL vanishes. Because the higher-logit increment is strictly positive for $\vartheta > 0$, both means are strictly negative. Their KLs and means depend only on ϑ, f_*, c_1, c_2, not on T. They tend to zero continuously as $\vartheta \downarrow 0$, so a fixed positive ϑ meets any fixed positive τ. The parameter values remain bounded and all gates remain in the interior. The construction attains nonvanishing response with one additional feature, and the original-family upper bound shows that at least one addition is necessary.

At the base, differentiation of the mean gives

$$(\partial_\vartheta m_c, \partial_{\theta_{\mathrm{bias}}} m_c) = (-c^2, -a_* c/2).$$

The two rows are linearly independent. The common task's desired independent score is $-z^{(c)}$ in each sequence, which is the ϑ-score modulo constants. Therefore its local shared coefficient equals the independent one, $\frac{1}{2}(c_1^2 + c_2^2)$. Thus the common-gradient task has unit local task efficiency alongside a rank-two output Jacobian.

All constants just used can be chosen uniformly on compact subsets with f_* away from zero and one, c_1 bounded below, and $c_2 - c_1$ bounded below. The nonzero bounds on $o_{\mathrm{bias},0}$, the hyperbolic-cosine separation, and the Taylor derivative bounds are then uniform. A common sufficiently small positive ϑ makes the repaired KL at most τ on the compact set, and continuity plus the positive lower amplitude bound gives a uniformly positive mean change. The resulting bounds are uniform over the specified compact structural range.

A fixed squared-loss consequence. The common-gradient task is realized by target $y_c = -1$ in each sequence, because the base means are zero. Its average squared-loss decrease is $-\frac{1}{2}\sum_c m_c - \frac{1}{4}\sum_c m_c^2$, with the sums over the two amplitudes. It is at most the negative mean gain, so the

original-family optimum is $O(T^{-1/2})$; the feasible small-ϑ construction gives the matching lower bound. In the repaired construction $-1 \le m_c < 0$, so the loss decrease is at least half its positive negative-mean gain. It is therefore positive independently of T. This transfers the finite-budget separation to the fixed nonzero prediction task $y_c = -1$.

C.3.2 Applications to sigmoid families

The two-amplitude response laws of Theorem 3.1, with its stated uniformity, are

$$B_{\mathrm{bias}}^{\mathrm{KL}} \asymp \min\{\sqrt{\tau}, \tau^{1/4}/\sqrt{T}\}, \qquad B_{\mathrm{bias}}^{\mathrm{op}} \asymp \min\{\delta, \sqrt{\delta/T}\} \tag{54}$$

For the two-amplitude biased family, put $a_* = 1 - f_*$. The block drifts in Appendix C.3.1 are $\beta_c = -2\bar{d}_c^{\mathrm{ret}}$, and near the base

$$\beta_c(\vartheta, \theta_{\mathrm{bias}}) = -2a_*\theta_{\mathrm{bias}} + \frac{4f_*c^2}{a_*}\vartheta^2 + O(\theta_{\mathrm{bias}}^2 + |\theta_{\mathrm{bias}}|\,\vartheta^2 + \vartheta^4). \tag{55}$$

The nuisance column is $-2a_*(1,1)^\top$; the quadratic vector $(c_1^2, c_2^2)^\top$ has a nonzero projection onto its orthogonal complement. The common gain derivative is $(c_1^2 + c_2^2)/2 > 0$. Thus Proposition B.1 applies with $r = 2$.

For the small-budget transfer, use the same product identity (52), now with $-2\bar{d}_c^{\mathrm{ret}} = \beta_c$. The preceding argument is stronger than localization along one path: replacing $O_\tau(T^{-1})$ by $o(1)$ shows that *any* sequence of unrestricted controls with $\|\beta\|_\infty \to 0$ must satisfy $\vartheta \to 0$ and $\theta_{\mathrm{bias}} \to 0$. Otherwise the cosh ratio is bounded away from one, or tends to zero for unbounded $|\vartheta|$, a contradiction. Compact structural separation makes this implication uniform. Hence the local projected quadratic coefficient in (55) applies to the entire low-cost feasible set, as required by Proposition B.1; Theorem 4.2 yields both small-budget laws.

Appending $\theta_{\mathrm{sq}}(v^{(c)})^2$ adds the drift column $-8a_*(c_1^2, c_2^2)^\top$. Together with the bias column it has rank two. The full-row-rank alternative supplies a zero-drift curve. More strongly, the formulas in Appendix C.3.1 construct one explicitly: for small positive ϑ, $\theta_{\mathrm{bias}}, \theta_{\mathrm{sq}} = O(\vartheta^2)$, the gain is $\asymp \vartheta$, and the phase KL and operator costs are $O(\vartheta^2)$ and $O(\vartheta)$. Choose $\vartheta \asymp \sqrt{\tau}$ or $\vartheta \asymp \delta$. Bounded task scores give the matching general upper bounds. Compact separation of f_* from gate endpoints and of c_1, c_2 from zero and each other makes every constant uniform.

C.4 Proof of the uniform candidate-value perturbation guarantee

Use the families, budgets, fixed target and perturbation order of Corollary 3.2. Let $\underline{c}$ be its common amplitude lower bound, so that $0 < \underline{c} \le c_a \le 1$ for $a = 1, 2$. All comparison constants below are uniform over the compact structural range specified in that corollary.

Proof. Write $G_a = -m_{a,T}(\theta)$ and $G = (G_1 + G_2)/2$. On every endpoint the nominal squared-loss decrease is

$$\mathcal{D}_T^{\mathrm{pert}}(\theta, 0) = G - \frac{1}{4}\sum_a m_{a,T}(\theta)^2 \le G.$$

The perturbation set contains zero. Thus each guaranteed improvement is no larger than the corresponding nominal gain, and Theorem 3.1 gives the two original-family upper bounds. The baseline update gives zero for every perturbation and proves nonnegativity. The unrestricted gain upper bounds also bound the repaired guaranteed improvement by $C\sqrt{\tau}$ or $C\delta$.

For the lower bounds, use the zero-drift squared-feature curve in Appendix C.3.1. Its copy marginal remains uniform. Set $G_a = -m_{a,T}(\theta)$ and $v_a = G_a/c_a \in (0, 1)$. The negative and positive sign probabilities are $(1 + v_a)/2$ and $(1 - v_a)/2$, respectively, and hence

$$\delta_{p,a} \coloneqq \|p_a - \mathcal{U}_T\|_1 = v_a = G_a/c_a.$$

Define the two averages

$$\overline{\delta}_p = \frac{1}{2}\sum_{a=1}^{2}\delta_{p,a} \le \frac{G}{\underline{c}}, \qquad \overline{\delta_p^2} = \frac{1}{2}\sum_{a=1}^{2}\delta_{p,a}^2 \le \overline{\delta}_p.$$

Here $G = (G_1 + G_2)/2$, $0 < G_a \le 1$, $|1 + m_{a,T}| \le 1$, and $\mathcal{D}_T^{\mathrm{pert}}(\theta, 0) \ge G/2$. Apply (30) with $M_0 = 1$ to each sequence and average:

$$\inf_{\max_{a,i} |\delta z_{a,i}| \le \varepsilon_z} \mathcal{D}_T^{\mathrm{pert}}(\theta, \delta z) \ge \frac{G}{2} - \varepsilon_z G - \varepsilon_z \overline{\delta}_p - \varepsilon_z^2 \left(\overline{\delta}_p + \frac{1}{2} \overline{\delta_p^2} \right)$$
$$\ge G \left[\frac{1}{2} - \varepsilon_z - \frac{\varepsilon_z + \frac{3}{2}\varepsilon_z^2}{\underline{c}} \right].$$

For $0 \le \varepsilon_z \le \underline{c}/16$ and $\underline{c} \le 1$, the bracket is at least $1/2 - 1/16 - 1/16 - 3/512 = 189/512 > 1/4$. Appendix C.3.2 supplies feasible choices with $G \ge C_{\mathrm{KL}}\sqrt{\tau}$ under the KL budget and $G \ge C_{\mathrm{op}}\delta$ under the operator budget, for positive constants independent of length and sufficiently small budgets in the stated structural range. The chosen coefficients and costs are independent of T, yielding the two positive lower bounds. □

Uniform perturbed improvement. At a fixed sufficiently small positive budget, the original family's guaranteed improvement tends to zero with length, whereas the repaired family guarantees positive improvement throughout the perturbation set in Corollary 3.2. The local task-efficiency and rank comparisons are evaluated at the original baseline. During the perturbation evaluation, both gate sequences, the write strengths and the target remain fixed. The comparison uses the original candidate identities and nominal individual-write budget. Restricting perturbations to keep candidates in the original unit ball preserves the stated lower and upper bounds: the uncertainty set shrinks and still contains zero.

Role of the amplitude lower bound. Complete erasure of both sequences requires $\varepsilon_z \ge \max(c_1, c_2)$. The choice $\delta z_{a,i} = -z_{a,i}$ then makes both outputs zero and the loss improvement zero. For example, $(c_1, c_2) = (a, 2a)$ is fully erased when $2a \le \varepsilon_z$. Thus a fixed positive radius cannot guarantee positive improvement over amplitude scales tending to zero. Corollary 3.2 uses a compact structural range with both amplitudes bounded below; complete erasure depends on the larger amplitude as well.

D Fixed features, finite inverse curves and local curvature

D.1 The inverse gate map and its analytic coefficients

At homogeneous baseline gates, $\pi_i^0 \propto \iota_i f_*^{T-i}$; it need not be uniform when writes are arbitrary. For $a < -\log f_*$, $0 < f_* e^a < 1$ and

$$\psi(a) = a - \log(1 - f_* e^a) + \log(1 - f_*).$$

On any symmetric neighborhood $|a| < -\log f_*$ it is analytic. Expanding the logarithm in its absolutely convergent series gives

$$\psi(a) = a + \sum_{\ell \ge 1} \frac{f_*^\ell}{\ell} (e^{\ell a} - 1).$$

Termwise differentiation is justified uniformly on compact subintervals. Thus

$$\psi_1 = \frac{1}{1 - f_*}, \qquad \psi_k = \frac{1}{k!} \sum_{\ell \ge 1} f_*^\ell \ell^{k-1} > 0 \quad (k \ge 2).$$

In particular $\psi_2 = f_*/[2(1 - f_*)^2]$ and $\psi_3 = f_*(1 + f_*)/[6(1 - f_*)^3]$. The fixed-write inverse $b = \Delta(\log \pi - a^{\mathrm{w}})$ uniquely determines the required log-retentions. The sigmoid bijection then uniquely determines the required logits.

D.2 Proof of exact closure and the feature count

If $\psi(\lambda v) \in S_{\mathrm{ret}}$ for every sufficiently small λ, all derivatives at zero belong to the closed finite-dimensional space S_{ret}. Since every $\psi_k \ne 0$, this requires $v^{\odot k} \in S_{\mathrm{ret}}$ for every k. Conversely their

inclusion and convergence of the series imply $\psi(\lambda v) \in S_{\mathrm{ret}}$. Applying a fixed right inverse of X_{ret} on S_{ret} constructs an analytic parameter curve. This proves the exact part of Theorem 5.1.

Let the distinct nonzero entries of v be $a_1, \ldots, a_{N_{\mathrm{distinct}}}$. Every power vector is constant on repeated-value groups and is zero on the zero group, giving rank at most N_{distinct}. The first N_{distinct} power columns evaluated at $a_1, \ldots, a_{N_{\mathrm{distinct}}}$ have determinant

$$\det[a_i^k]_{i,k=1}^{N_{\mathrm{distinct}}} = \left(\prod_i a_i\right) \prod_{i<j} (a_j - a_i) \neq 0.$$

Therefore the power span has exactly dimension N_{distinct}. This Vandermonde calculation gives the dimension of the fixed-data power-evaluation span.

For approximation through degree q, let P be an orthogonal projector onto S_{ret}. If a curve in S_{ret} approximates $\psi(\lambda v)$ to $O(|\lambda|^{q+1})$, then $(I - P)\psi(\lambda v) = O(|\lambda|^{q+1})$. Its Taylor coefficients through q must vanish. Conversely, the truncated Taylor curve lies in S_{ret} whenever these power columns do. Necessity follows from distance to the fixed subspace, independently of the choice of parameter lift. Each new feature adds at most one dimension to the missing quotient; adjoining a basis attains (20).

D.3 Nonuniform fixed baseline gates

Lemma D.1 (Required features at a nonuniform fixed baseline)**.** *Fix finite T, positive write strengths, fixed candidate values and g, baseline gates $f_{j+1}^0 \in (0,1)$, $u_j^0 = \operatorname{logit} f_{j+1}^0$, $v = -\Delta Z^\top g$, and a fixed feature space $S = \operatorname{range} X$. Define*

$$\psi_f(a) = a - \log(1 - fe^a) + \log(1 - f), \qquad a < -\log f.$$

For sufficiently small signed λ, the unique logit increment realizing $q_{\lambda,i} \propto \pi_i^0 e^{-\lambda z_i^\top g}$ is

$$\Psi_j(\lambda) = \psi_{f_{j+1}^0}(\lambda v_j) = \sum_{k \geq 1} \lambda^k w_{k,j}, \qquad w_{k,j} = \frac{\psi_{f_{j+1}^0}^{(k)}(0)}{k!} v_j^k.$$

In particular,

$$w_{1,j} = \frac{v_j}{1 - f_{j+1}^0}, \qquad w_{2,j} = \frac{f_{j+1}^0 v_j^2}{2(1 - f_{j+1}^0)^2}.$$

Exact realization for every sufficiently small signed λ in $u^0 + S$ holds iff every $w_k \in S$. Approximation of the required logit path to $O(|\lambda|^{q+1})$ holds iff $w_1, \ldots, w_q \in S$. Writing $X_{\mathrm{inv},q} = [w_1, \ldots, w_q]$, the minimum arbitrary added-feature count is $\operatorname{rank}[X, X_{\mathrm{inv},q}] - \operatorname{rank} X$.

Proof. The fixed-write inverse $b = \Delta(\log p - a^w)$ gives the unique required retention $b^0 + \lambda v$. The sigmoid inverse gives

$$u_j^{\mathrm{req}}(\lambda) = \operatorname{logit}(f_{j+1}^0 e^{\lambda v_j}), \qquad u_j^{\mathrm{req}}(\lambda) - u_j^0 = \psi_{f_{j+1}^0}(\lambda v_j).$$

Finite T and strict baseline gates provide a common positive neighborhood where all arguments lie in $(0,1)$. The coordinate maps are analytic, giving a convergent vector series and the displayed derivatives.

Let P_S be the Euclidean orthogonal projector onto S. If $\Psi(\lambda)$ lies in S on an interval, differentiating $(I - P_S)\Psi(\lambda) = 0$ at zero puts all coefficients in S. Conversely, if every coefficient lies in S, convergence and closedness of the finite-dimensional space give $\Psi(\lambda) \in S$. A fixed right inverse on S, for example $X^\dagger$, supplies an analytic lift $\theta(\lambda) = X^\dagger \Psi(\lambda)$.

For finite order, any curve in S with error $O(|\lambda|^{q+1})$ implies $(I - P_S)\Psi(\lambda) = O(|\lambda|^{q+1})$, whose Taylor coefficients through order q must vanish. Conversely, the truncated series $\sum_{k=1}^q \lambda^k w_k$ lies in S and is lifted by the same fixed right inverse. Each new column contributes at most one missing quotient direction; a basis of $(S + \operatorname{range} X_{\mathrm{inv},q})/S$ attains the rank difference.

□

Baseline derivative and exact inverse. For nonuniform f^0 and $u^{\mathrm{ret}} = u^0 + X_{\mathrm{ret}}\theta$, the baseline retention Jacobian is $U_0 = \mathrm{Diag}(1 - f^0)X_{\mathrm{ret}}$. Lemma D.1 characterizes finite realization using the exact coordinatewise inverse.

All directions and the empirical features. For $d_j^{\mathrm{adj}} = z_{j+1} - z_j$ and $v_j = (d_j^{\mathrm{adj}})^\top g$, polarization yields the coefficient columns

$$j \longmapsto \frac{\psi_{f_{j+1}^0}^{(k)}(0)}{k!}\binom{k}{\alpha}(d_j^{\mathrm{adj}})^\alpha, \qquad |\alpha| = k, \quad 1 \le k \le q.$$

Stack their evaluations across sequences when the coefficients are shared. For $d = 2, q = 2$, these are the two linear and three quadratic columns in Appendix K.3. The generator uses the frozen gate and adjacent inputs; task coefficients supply the dependence on g.

Feature dimension at a nonuniform baseline. Take $v = (1, 1)$ and $(f_2^0, f_3^0) = (1/5, 4/5)$. Then $w_1 = (5/4, 5)^\top$, $w_2 = (5/32, 10)^\top$ and $\det[w_1, w_2] = 375/32 \neq 0$. A single distinct nonzero entry of v does not bound this span's dimension by one. All claims above are finite-T statements; length-uniform domains, conditioning and accumulated-error constants need additional assumptions.

D.4 All directions and causal features

Return to the common baseline f_* of Theorem 5.1; nonuniform baselines use the weighted columns of Lemma D.1. Write $d_j^{\mathrm{adj}} = z_{j+1} - z_j \in \mathbb{R}^d$, so $v_j = (d_j^{\mathrm{adj}})^\top g$. To realize every sufficiently small g-direction through order q, the necessary and sufficient span is the evaluation space of all nonconstant monomials $(d_j^{\mathrm{adj}})^\alpha$ with $1 \le |\alpha| \le q$. Expanding $((d_j^{\mathrm{adj}})^\top g)^k$ proves sufficiency. Differentiating in the components of g, or using polarization of homogeneous polynomials, proves necessity. There are at most $\binom{d+q}{q} - 1$ such monomials, and the minimum additions are the rank increase of their evaluation matrix against X_{ret}.

For exact closure over all directions, the evaluation space of all nonconstant polynomials has dimension equal to the number of distinct nonzero rows d_j^{adj}. To see that they can be separated, choose a vector g outside the finitely many hyperplanes $(d_j^{\mathrm{adj}})^\top g = 0$ and $(d_i^{\mathrm{adj}} - d_j^{\mathrm{adj}})^\top g = 0$ for distinct nonzero rows. Their projections are distinct and nonzero, reducing to the previous Vandermonde argument.

The feature d_j^{adj} is available at gate f_{j+1} from the current and preceding candidates. Its monomials are causal input-only features. Stacking their evaluations across sequences uses the same polynomial coefficients, while the minimum rank is determined by the stated evaluation data. Applying a reduced basis to another input collection requires the corresponding span condition on that collection. The update analysis holds candidate values fixed.

Three realization problems. Realizing a neighborhood of entropic targets requires the full power-evaluation span. Realizing one specified endpoint requires its required-logit vector to be available. A nonlinear feature map can instead evaluate $\psi((d_j^{\mathrm{adj}})^\top \zeta)$ directly. The fixed-linear-feature results above characterize exact neighborhood realization and its finite-order approximation.

Stacking the two amplitudes gives four distinct nonzero adjacent differences $\{-2c_1, 2c_1, -2c_2, 2c_2\}$. Their full power-evaluation span at a common baseline has dimension four, whereas bias, first difference and its square supply $\mathrm{span}\{\mathbf{1}, v, v^{\odot 2}\}$, a three-dimensional space. Thus the square addition can restore a length-independent useful response without realizing the entire common exponential-tilt curve. The minimality in Section 3 concerns nonvanishing response; the feature count here concerns exact curve coverage.

D.5 Leading target error and statistical curvature

D.5.1 The optimal coefficient of the first missing order

Relation to statistical curvature. Write $\pi_{X_{\mathrm{ret}}}(\lambda)$ for the allocation along a shared logit curve and $\mathrm{dist}_{F_b}(x, S_{\mathrm{ret}})^2 = \min_{s \in S_{\mathrm{ret}}}(x - s)^\top F_b(x - s)$ at the base. If $v \in S_{\mathrm{ret}}$, among smooth curves with

the correct first derivative,

$$\inf \lim_{\lambda \to 0} \frac{D_{\mathrm{KL}}(\pi_{X_{\mathrm{ret}}}(\lambda) \| q_\lambda)}{\lambda^4} = \frac{f_*^2}{8(1-f_*)^2} \operatorname{dist}_{F_b}(v^{\odot 2}, S_{\mathrm{ret}})^2. \tag{56}$$

This is a Fisher-normal second-order residual; Appendix D.5.2 gives its relation to classical statistical curvature. The missing logit order need not equal the drift order. The higher-order coefficients are derived below, whereas the finite-response bounds additionally require recurrent accumulation and global localization.

Assume $v, \dots, v^{\odot q} \in S_{\mathrm{ret}}$. Consider local C^{q+1} logit curves in $u_*^{\mathrm{ret}}\mathbf{1} + S_{\mathrm{ret}}$ matching the target through order q. Their first q coefficients are uniquely $\psi_k v^{\odot k}$, while the next coefficient $w \in S_{\mathrm{ret}}$ is free. The required exact curve has next coefficient $\psi_{q+1} v^{\odot(q+1)}$. Since $\ell'_{\mathrm{sig}}(u_*^{\mathrm{ret}}) = 1 - f_*$,

$$b_{X_{\mathrm{ret}}}(\lambda) - b^*(\lambda) = (1-f_*)\lambda^{q+1}\{w - \psi_{q+1} v^{\odot(q+1)}\} + o(\lambda^{q+1}).$$

At a positive base the log-partition Hessian is C_{π^0}, and its pullback through H is $F_b \succ 0$. Hence

$$\lim_{\lambda \to 0} \frac{D_{\mathrm{KL}}(\pi_{X_{\mathrm{ret}}}(\lambda) \| q_\lambda)}{\lambda^{2q+2}} = \frac{(1-f_*)^2}{2} \| w - \psi_{q+1} v^{\odot(q+1)} \|_{F_b}^2.$$

Here $\|x\|_{F_b}^2 := x^\top F_b x$. Minimizing over $w \in S_{\mathrm{ret}}$ is a finite-dimensional weighted projection. Its minimizer is attainable by a polynomial parameter curve after applying a right inverse of X_{ret}. This proves the general coefficient and (56). The infimum is over asymptotic mismatch coefficients of smooth local curves matching the preceding orders.

When the distance is positive the leading order cannot be removed within the current linear feature space. Adding its missing quotient component removes that coefficient. The sharp coefficient measures target-allocation mismatch. Since Z may have a kernel, mean matching can require fewer directions. The finite-budget results separately bound the best output response.

D.5.2 Identifying the classical statistical curvature

For a one-dimensional positive statistical family $p_i(\vartheta)$ with common support, let $s_i^{\mathrm{stat}} = \partial_\vartheta \log p_i(\vartheta)\big|_0$, $a_i^{\mathrm{stat}} = \partial_\vartheta^2 \log p_i(\vartheta)\big|_0$, and $F_\vartheta = \mathbb{E}_{p^0}\left(s^{\mathrm{stat}}\right)^2 > 0$. Efron's statistical curvature is

$$\gamma_{\mathrm{E}}^2 = \frac{\|(\mathrm{Id} - P^{[p^0]}_{\mathrm{span}\{\mathbf{1}, s^{\mathrm{stat}}\}}) a^{\mathrm{stat}}\|_{L^2(p^0)}^2}{F_\vartheta^2},$$

where Id is the identity operator and the projection is in $L^2(p^0)$. This is the one-parameter statistical curvature of Efron (1975); see Sei (2011, Section 2, Eq. (2.1)) for the score-projection formulation. Here a^{stat} is the second log-density derivative, not the quadratic Taylor coefficient. Equivalently the normal residual is $a^{\mathrm{stat}} + F_\vartheta \mathbf{1} - F_\vartheta^{-1} \mathbb{E}_{p^0}[a^{\mathrm{stat}} s^{\mathrm{stat}}] s^{\mathrm{stat}}$, since $\mathbb{E} a^{\mathrm{stat}} = -F_\vartheta$ and $\mathbb{E} s^{\mathrm{stat}} = 0$. The curvature is defined relative to the exponential-family structure.

For the canonical sigmoid family at the uniform baseline, $b'(0) = v$ and $b''(0) = -v^{\odot 2}$. Thus $s^{\mathrm{stat}} = -z$, $F_\vartheta = 1$, and $a^{\mathrm{stat}} = -Hv^{\odot 2}$ modulo a constant. Projecting away the score and the constant yields

$$\gamma_{\mathrm{E}}^2 = \operatorname{dist}_{F_b}(v^{\odot 2}, \operatorname{span} v)^2.$$

Here $v^{\odot 2} = 4\mathbf{1}$, $H\mathbf{1} = T - i$, $\mathrm{Var}(T-i) = (T^2-1)/12$, and $|\mathrm{Cov}(T-i, z)| = 1/2$. Therefore

$$\gamma_{\mathrm{E}}^2 = 16\{(T^2-1)/12 - 1/4\} = \tfrac{4}{3}(T^2-4), \qquad \text{the coefficient in (56)} = \gamma_{\mathrm{E}}^2/8. \tag{57}$$

At $f_* = 1/2$ the multiplier in (56) is $1/8$, which proves the second equality in (57). The local curvature identification uses the classical normal-projection interpretation. The length–budget transition, global feasible-set bounds and nonperiodic extension are established by the recurrent-accumulation and localization arguments in Appendices B–C. The escape example in Appendix B.6 illustrates the role of global localization.

D.5.3 A SHARP TARGET-MISMATCH COEFFICIENT

To match the exponentiated-gradient (EG) mean $-\tanh\lambda$ exactly with the first family, solve $s^{\mathrm{sgn}}(\vartheta) = -\lambda$. The solution is

$$\vartheta(\lambda) = -\tfrac{1}{4}\log(2e^{-2\lambda} - 1), \qquad 0 < \lambda < \log 2/2.$$

The sign marginal now equals the target, and the remaining KL to the target is $D_n^{\mathrm{KL}}(2a(\vartheta(\lambda)))$. At fixed T it is

$$D_{\mathrm{KL}}(\pi(\vartheta(\lambda))\|q_\lambda) = \frac{T^2-4}{6}\lambda^4 + o(\lambda^4).$$

The same coefficient follows from the optimal missing-feature formula. Under $\pi^0 = \mathcal{U}_T$, $F_b = H^\top C_{\pi^0} H$, $v^{\odot 2} = 4\mathbf{1}$, and $Hv = -z + \mathbf{1}$. Direct regression gives

$$\mathrm{dist}_{F_b}(v^{\odot 2}, \operatorname{span} v)^2 = 16\left\{\frac{T^2-1}{12} - \frac{1}{4}\right\} = \frac{4}{3}(T^2-4).$$

The coefficient multiplier at $f_* = 1/2$ is $1/8$, yielding $(T^2-4)/6$. Matching the mean can therefore leave the best possible leading allocation error. This is precisely the distinction between output equality and provenance equality.

D.6 A SEPARATE EUCLIDEAN COEFFICIENT-NORM CONSTRAINT

An explicit coefficient lift is

$$\theta_q = \sum_{k=1}^{q} \psi_k \lambda^k X_{\mathrm{ret}}^\dagger v^{\odot k}, \qquad \|\theta_q\| \le \sum_{k=1}^{q} \psi_k |\lambda|^k \|X_{\mathrm{ret}}^\dagger v^{\odot k}\|.$$

This realizes the prescribed logits because the relevant powers lie in $\operatorname{range} X_{\mathrm{ret}}$. It also exposes feature conditioning. If a Euclidean weight budget $R_\theta > 0$ is additionally required, the right-hand side must be at most R_θ. A sufficient explicit supplement to (63) is

$$\lambda \le 1, \qquad \lambda \le R_\theta \Big/ \sum_{k=1}^{q} \psi_k \|X_{\mathrm{ret}}^\dagger v^{\odot k}\|.$$

This additional constraint controls the coefficient norm through the conditioning of the feature matrix. In the alternating constructions the repaired feature coefficients are bounded independently of length, as their explicit formulas show.

E FINITE DESCENT ON FIXED OBSERVATION SUPPORTS

E.1 NEGATIVE ENTROPY, RELATIVE SMOOTHNESS AND THE DIRECTED KL IDENTITY

On the positive probability simplex, let $\varphi_{\mathrm{ent}}(p) = \sum_i p_i \log p_i$. Its Bregman divergence is $D_{\varphi_{\mathrm{ent}}}(p,q) = \varphi_{\mathrm{ent}}(p) - \varphi_{\mathrm{ent}}(q) - \langle \nabla\varphi_{\mathrm{ent}}(q), p-q\rangle = D_{\mathrm{KL}}(p\|q)$. For $L = L_T^{\mathrm{cons}}$ and $y = y_T$, fix positive p^0, set $g = Zp^0 - y$, $h = Z^\top g$, and take $\lambda > 0$. The tilt $q_{\lambda,i} \propto p_i^0 e^{-\lambda h_i}$ is the unique minimizer of

$$\min_{p \in \Delta_{\mathrm{prob}}^{T-1}} \left\{\langle h, p\rangle + \lambda^{-1} D_{\mathrm{KL}}(p\|p^0)\right\},$$

because this objective is $\lambda^{-1} D_{\mathrm{KL}}(p\|q_\lambda)$ plus a constant. Thus q_λ is the entropic step for the linearized task at p^0. For any p on the same support, substitution gives

$$D_{\mathrm{KL}}(p\|q_\lambda) - D_{\mathrm{KL}}(p^0\|q_\lambda) = D_{\mathrm{KL}}(p\|p^0) + \lambda\langle h, p - p^0\rangle. \tag{58}$$

For linear gain $G = -g^\top(Zp - Zp^0)$, the loss decrease is $G - \|Z(p-p^0)\|^2/2$. Bounded candidates yield

$$L(p) - L(p^0) - \langle h, p - p^0\rangle = \tfrac{1}{2}\|Z(p-p^0)\|^2 \le R^2 D_{\mathrm{KL}}(p\|p^0),$$

using $\|Z(p-p^0)\| \le R\|p-p^0\|_1$ and Pinsker. This establishes R^2-relative smoothness with respect to negative entropy. Relative-smoothness descent is developed by Bauschke et al. (2017); we use the

formulation in Lu et al. (2018, Definition 1.1 and Proposition 1.1). Combining the displayed bound with the directed tilt identity yields

$$L(p^0) - L(p) \geq \frac{D_{\mathrm{KL}}(p^0 \| q_\lambda) - D_{\mathrm{KL}}(p \| q_\lambda)}{\lambda} + (\lambda^{-1} - R^2) D_{\mathrm{KL}}(p \| p^0). \tag{59}$$

Applying (59) at each observation and sequence gives the D_w-weighted inequality. For a realizable shared endpoint, the descent bound uses the actual-to-target mismatch $D_{\mathrm{KL}}(p \| q_\lambda)$. The realization criteria and the actual change from the baseline are treated separately in Appendix D and (23), respectively.

E.2 Proof of actual-budget descent

Here $L = L_T^{\mathrm{cons}}$, $V_h = \mathrm{Var}_{\pi^0}(h)$, and $\Omega_h = \mathrm{osc}(h)$ as in Theorem 5.2. Appendix E.1 proves the descent inequality by the exact directed KL identity and the squared-loss expansion. For $\lambda \Omega_h \leq 1$, the target density ratios along $q_s \propto \pi^0 e^{-sh}$ are at least e^{-1}. Therefore $\mathrm{Var}_{q_s}(h) \geq V_h / e$ and $D_{\mathrm{KL}}(\pi^0 \| q_\lambda) \geq V_h \lambda^2 / (2\mathrm{e})$. The hypothesis on cumulative error gives $D_{\mathrm{KL}}(p \| q_\lambda) \leq V_h \lambda^2 / (4\mathrm{e})$. The remaining KL term in (59) is nonnegative for $\lambda \leq R^{-2}$. Subtracting and dividing by λ proves (22). Equation (23) supplies the corresponding actual-budget bound.

E.3 A finite KL bound without a minimum-mass constant

The global log-weight oscillation inequality (28), proved in Appendix A.5, bounds either directed KL by one eighth of the squared oscillation. Here we apply that bound to the finite Taylor remainder of the recurrent inverse.

For $|\lambda| \|v\|_\infty \leq a_0$, Taylor's theorem gives

$$\|u_q^{\mathrm{ret}} - u_*^{\mathrm{ret}} \mathbf{1} - \psi(\lambda v)\|_\infty \leq c_{q+1}^{\psi} |\lambda|^{q+1} \|v\|_\infty^{q+1}.$$

Because $0 < \ell'_{\mathrm{sig}}(u^{\mathrm{ret}}) < 1$, the same bound holds for $\delta b = b_q - b^*$. Two entries of $H \delta b$ differ by a sum over an interval of at most $T - 1$ entries, so $\mathrm{osc}(H \delta b) \leq (T-1) \|\delta b\|_\infty$. Applying the score bound proves (62), for both KL directions (although descent below needs a particular direction).

For $q \geq 1$ the constant is explicitly evaluable as

$$c_{q+1}^{\psi} = \frac{1}{(q+1)!} \sum_{\ell \geq 1} (f_* e^{a_0})^\ell \ell^q.$$

All terms of $\psi^{(q+1)}$ are positive and increasing on the interval, so its supremum occurs at a_0. For instance, $c_2^{\psi} = f_* e^{a_0} / [2(1 - f_* e^{a_0})^2]$. The constant grows near the gate-domain boundary. Length-uniform steps use the cumulative-error conditions in Appendix E.5; coefficient-norm control is treated in Appendix D.6.

E.4 An explicit finite polynomial step

For completeness, define the realizable degree-q logits and finite Taylor constants

$$u_q^{\mathrm{ret}} = u_*^{\mathrm{ret}} \mathbf{1} + \sum_{k=1}^{q} \psi_k \lambda^k v^{\odot k}, \qquad p_q = \mathrm{softmax}(a^{\mathrm{w}} + H \ell_{\mathrm{sig}}(u_q^{\mathrm{ret}})), \tag{60}$$

$$c_{q+1}^{\psi} = \frac{\sup_{|a| \leq a_0} |\psi^{(q+1)}(a)|}{(q+1)!}, \qquad c_q^{\mathrm{sc}} = (T-1) c_{q+1}^{\psi} \|v\|_\infty^{q+1}, \quad c_q^{\mathrm{KL}} = (c_q^{\mathrm{sc}})^2 / 8. \tag{61}$$

Here $q \geq 1$, $0 < a_0 < -\log f_*$, and all columns in $X_{\mathrm{pow},q}$ are available in the fixed logit span. Taylor's theorem and $0 < \ell'_{\mathrm{sig}} < 1$ give $\varepsilon_{\mathrm{sc}}(\lambda) \leq c_q^{\mathrm{sc}} \lambda^{q+1}$ whenever $\lambda \|v\|_\infty \leq a_0$. In particular

$$D_{\mathrm{KL}}(p_q \| q_\lambda) \leq c_q^{\mathrm{KL}} \lambda^{2q+2}. \tag{62}$$

Let

$$\lambda_0 = \min \left\{ R^{-2}, \frac{a_0}{\|v\|_\infty}, \frac{1}{\Omega_h}, \left(\frac{V_h}{4\mathrm{e} c_q^{\mathrm{KL}}} \right)^{1/(2q)} \right\}. \tag{63}$$

Corollary E.1 (Closed-form step with an actual prescribed budget)**.** *For KL budget $\tau > 0$, every positive λ at most*

$$\min\left\{\lambda_0, \frac{\sqrt{2\tau}}{\Omega_h}, \left(\frac{\sqrt{2\tau}}{c_q^{\mathrm{sc}}}\right)^{1/(q+1)}\right\} \tag{64}$$

satisfies $D_{\mathrm{KL}}(p_q\|\pi^0) \le \tau$ and $L(\pi^0) - L(p_q) \ge V_h\lambda/(4\mathrm{e})$. For operator budget $\delta > 0$, replace the last two terms by δ/Ω_h and $(\delta/c_q^{\mathrm{sc}})^{1/(q+1)}$. The same descent conclusion holds and $\|p_q - \pi^0\|_1 \le \delta$.

Proof. The original radius λ_0 supplies all descent and gate-inversion domain conditions. The two additional KL restrictions each bound one term of $\lambda\Omega_h + c_q^{\mathrm{sc}}\lambda^{q+1}$ by $\sqrt{2\tau}$; their sum is $\sqrt{8\tau}$. For the operator restrictions the two terms are each at most δ. Apply Theorem 5.2. If an exact realization gives $c_q^{\mathrm{sc}} = 0$, omit the corresponding constraints rather than divide by zero. □

The supremum defining c_{q+1}^{ψ} is explicitly evaluable; Appendix D gives its convergent positive series and the effect of a raw weight budget. Exact evaluation of $\varepsilon_{\mathrm{sc}}(\lambda)$ may give a less conservative admissible step than the Taylor majorant. The construction supplies an explicit realizable update with the stated loss-decrease bound. For either step choice, the actual cost is evaluated against the requested budget.

E.5 Length independence is a cumulative-error condition

For any retention error δb,

$$\mathrm{osc}(H\delta b) = \max_{1\le i<j\le T}\left|\sum_{k=i}^{j-1}\delta b_k\right|. \tag{65}$$

This follows because two suffix sums differ by exactly that interval sum. Thus the coarse bound $(T-1)\|\delta b\|_\infty$ can be replaced by an actual bound on interval sums. If the right side is at most $\overline{E}\lambda^{q+1}$ with $\overline{E}$ independent of T, Theorem 5.2 and Corollary E.1 are length uniform whenever $R, \Omega_h, \|v\|_\infty$, the gate margin, and a positive lower bound for V_h are also uniform.

For example, suppose the errors repeat in blocks of L_{blk} and their sum over each full block is zero. Prefix sums return to zero after each block. The difference between any two prefix sums is bounded by $L_{\mathrm{blk}}\max_k|\delta b_k|$, including an incomplete final block: either interval direction is represented by a consecutive segment on the zero-sum cycle. Therefore fixed L_{blk} and $\|\delta b\|_\infty \le C\lambda^{q+1}$ give a length-independent certificate. More generally, a deterministic interval-sum bound uniform in T supplies the same sufficient condition, including for nonperiodic errors. The finite-length Taylor order specifies the power of λ, while the interval-sum bound controls accumulation across length.

E.6 Explicit steps and several observation targets

Appendices E.2–E.4 give a sufficient terminal step with actual-budget descent. The same directed identity extends to several observations as follows.

For several targets on their declared observation supports, write their positive reference laws as $p^{0,(s)}$, task scores as h_s, and normalized weights as w_s^{obs}. Replace V_h by $\sum_s w_s^{\mathrm{obs}}\,\mathrm{Var}_{p^{0,(s)}}(h_s)$ and use a common bound on score oscillations and value norms. The weighted identity certifies an already realizable shared endpoint if its total directed mismatch meets the corresponding bound. For the jointly shared retention/query model, prefix realization obeys the compatibility conditions in Appendix I. Appendix F.7 gives a free-gate construction on origin-resolved prefix supports.

F Reset memory and origin-resolved realization

F.1 Keep/write/reset coefficients and ICM coordinates

Appendix A.1 proves that nonnegative $m^+ = \kappa m + \omega z$ preserves every input pair in a ball if and only if $\kappa + \omega \le 1$; scaling the unit ball gives any radius $R > 0$. Thus the two-term update is the

zero-anchor case of

$$m_1 = z_1, \qquad m_t = \kappa_t m_{t-1} + \omega_t z_t + \sigma_t a_t, \qquad q_t^{\mathrm{gate}} = (\kappa_t, \omega_t, \sigma_t) \in \Delta^2_{\mathrm{prob}}, \quad t \geq 2. \tag{66}$$

Here $\Delta^2_{\mathrm{prob}} = \{q \in \mathbb{R}^3 : q_i \geq 0, \sum_i q_i = 1\}$, and all values have norm at most R. We use BAM for the three-branch recurrence (66) and ICM for its zero-anchor form. The ICM coordinates below also parameterize the branch probabilities when the fixed anchors are nonzero. Initializing at $m_1 = z_1$ agrees with the conservative model. For the following interior-coordinate results all three branches are positive.

Proposition F.1 (ICM is an interior coordinate chart)**.** *The map*

$$(f_t^{\mathrm{I}}, \alpha_t) \mapsto \big(f_t^{\mathrm{I}}(1-\alpha_t), \alpha_t, (1-f_t^{\mathrm{I}})(1-\alpha_t)\big)$$

is a smooth bijection from $(0,1)^2$ onto the interior of Δ^2_{prob}, with inverse $\alpha_t = \omega_t$, $f_t^{\mathrm{I}} = \kappa_t/(\kappa_t+\sigma_t)$. For $a_t = 0$ this is ICM. The conservative face $\sigma_t = 0$, with $\omega_t < 1$, requires $f_t^{\mathrm{I}} = 1$ and is not an interior finite-logit ICM point.

Proof. The displayed coefficients are positive and sum to one. Substituting the inverse recovers each coefficient and each input gate; all denominators are positive in the interior. The assertion about the face follows from $\sigma_t = (1 - f_t^{\mathrm{I}})(1 - \alpha_t)$. □

This is a classical conditional-Bernoulli, or stick-breaking, chart (Linderman et al., 2015). A specified shared family is expressed in its chosen gate coordinates; the affine state recurrence (66) and that parameterization define separate maps.

For a separate scaled-average comparison, fix $0 < M < 1$ and use the zero-anchor recurrence $\widetilde{m}_t = \kappa_t \widetilde{m}_{t-1} + \omega_t z_t$, initialized at $\widetilde{m}_0 = 0$. At $t = 1$, choose $\omega_1 = M$ and $\kappa_1 = \sigma_1 = (1-M)/2$, so that $\widetilde{m}_1 = M z_1$ (equivalently, $\alpha_1 = M$ and $f_1^{\mathrm{I}} = 1/2$). For $t \geq 2$, requiring $\widetilde{m}_t = (M/t)\sum_{i=1}^t z_i$ for every candidate sequence gives $\kappa_t = (t-1)/t$, $\omega_t = M/t$, and $\sigma_t = (1-M)/t$. The ICM coordinates are

$$\alpha_t = M/t, \qquad f_t^{\mathrm{I}} = (t-1)/(t-M), \qquad t \geq 2. \tag{67}$$

All three branches are positive at every finite time, while $\alpha_t \to 0$ and $f_t^{\mathrm{I}} \to 1$ as $t \to \infty$. The example separates representation of a prescribed memory function from uniform interiority of its gate coordinates, complementing the gate-learnability perspective of Gu et al. (2020b).

F.2 Candidate mass, relative allocation, and the chosen budget

Take fixed zero anchors. Expansion of (66) gives candidate coefficients

$$\eta_1 = \prod_{t=2}^{T} \kappa_t, \qquad \eta_i = \omega_i \prod_{t=i+1}^{T} \kappa_t \ (i \geq 2), \qquad M = \sum_i \eta_i, \quad \pi_i = \eta_i / M.$$

Thus $m_T = MZ\pi$. Interior gates and $T \geq 2$ give $0 < M < 1$. The candidate influences form a subprobability vector $\eta = M\pi$, with total mass M. We write $\|A\|_{\infty,2\to 2} \coloneqq \sup_{\|\mathbf{e}\|_{\infty,2}\leq 1} \|A\mathbf{e}\|_2$ for the induced norm from the block input norm of Section 2 to the Euclidean output norm.

Proposition F.2 (Mass–allocation budget decomposition)**.** *Let $\eta = M\pi$, $\eta^0 = M_0\pi^0$, with $0 < M, M_0 < 1$ and positive probability vectors π, π^0. If only candidates may be independently perturbed, then*

$$\|A_\eta - A_{\eta^0}\|_{\infty,2\to 2} = \|M\pi - M_0\pi^0\|_1 =: d_z. \tag{68}$$

In particular $|M - M_0| \leq d_z$ and $M\|\pi - \pi^0\|_1 \leq 2d_z$. If all zero-anchor origins are merged into one atom, $\bar{p} = (M\pi, 1-M)$ satisfies

$$D_{\mathrm{KL}}(\bar{p}\|\bar{p}^0) = d_{\mathrm{Bern}}(M\|M_0) + M D_{\mathrm{KL}}(\pi\|\pi^0). \tag{69}$$

For tangent vectors with $\sum_i \dot{\pi}_i = 0$, its Fisher quadratic form is

$$\sum_i \frac{(\dot{\bar{p}}_i)^2}{\bar{p}_i} = \frac{\dot{M}^2}{M(1-M)} + M\sum_i \frac{\dot{\pi}_i^2}{\pi_i}. \tag{70}$$

Proof. The triangle inequality proves the upper operator bound; choosing every candidate perturbation parallel to a common unit vector, with the sign of $\eta_i - \eta_i^0$, attains it. Summing coefficient differences proves the mass bound. The identity $M(\pi - \pi^0) = (\eta - \eta^0) + (M_0 - M)\pi^0$ proves the second bound. In the KL, expand $\log(M\pi_i/(M_0\pi_i^0))$ and sum. For the Fisher form use $\dot{\eta}_i = \dot{M}\pi_i + M\dot{\pi}_i$; the cross term vanishes and the anchor contributes $\dot{M}^2/(1-M)$. □

Resolving the reset origins. Let $\zeta_t = r_t^{\mathrm{reset}}/(1-M)$ be their conditional law and let the baseline use ζ^0. Permuting the origin coordinates to list writes first gives $p_{\mathrm{orig}} = (M\pi, (1-M)\zeta)$ and

$$D_{\mathrm{KL}}(p_{\mathrm{orig}} \| p_{\mathrm{orig}}^0) = d_{\mathrm{Bern}}(M\|M_0) + MD_{\mathrm{KL}}(\pi\|\pi^0) + (1-M)D_{\mathrm{KL}}(\zeta\|\zeta^0). \tag{71}$$

Indeed, substitute $p_i = M\pi_i$ on writes and $p_t = (1-M)\zeta_t$ on resets in the defining KL sum. Within each group the conditional probabilities sum to one, giving the Bernoulli term and the two conditional KL terms. This identity holds for any positive probability law split into these two groups, independently of the attached values. With zero anchors, the output is $MZ\pi$. Aggregating the reset origins removes the final conditional-KL term; when their attached values differ, the origin-resolved representation also retains the information needed to compute their output contribution.

If $\mu = Z\pi$ and $\Sigma_\pi = ZC_\pi Z^\top$, the free output covariance of $\bar{p}$ is $M\Sigma_\pi + M(1-M)\mu\mu^\top$. At fixed mass its allocation-Fisher response is $M\Sigma_\pi$ under the *full-distribution* KL, but $M^2\Sigma_\pi$ under a separately chosen *conditional* KL on π.

Which anchors are distinguishable? If individual reset origins also remain origin-resolved, their conditional distribution ζ^{reset} adds $(1-M)D_{\mathrm{KL}}(\zeta^{\mathrm{reset}}\|\zeta^{\mathrm{reset},0})$ to (69). If a merged anchor may also be perturbed, its operator contribution is $|M - M_0|$ in addition to d_z. If every anchor is fixed at zero and only candidates vary, neither extra term belongs to the exact candidate operator. For example, the paths

$$q_2^{\mathrm{gate},A} = (.4,.4,.2),\ q_3^{\mathrm{gate},A} = (.5,.2,.3), \qquad q_2^{\mathrm{gate},B} = (1/3,1/3,1/3),\ q_3^{\mathrm{gate},B} = (.6,.2,.2)$$

have the same candidate coefficients $(.2,.2,.2)$ and hence agree for *all* candidate values. Their two reset coefficients are $(.1,.3)$ and $(.2,.2)$, so the fully origin-resolved KL is $.1\log(.1/.2) + .3\log(.3/.2) > 0$. The origin-resolved budget also charges redistribution among the reset labels, beyond the candidate-only sensitivity change.

Discarding memory is a different intervention. Writing $m^0 = M_0\mu^0$ and $m = M\mu$ yields

$$-g^\top(m - m^0) = M[-g^\top(\mu - \mu^0)] - (M - M_0)g^\top\mu^0. \tag{72}$$

The last term can improve the loss without repairing the relative allocation. At a free terminal zero-anchor gate, take $0 < \epsilon < 1$ and replace (κ, ω, σ) by $((1-\epsilon)\kappa, (1-\epsilon)\omega, \sigma + \epsilon(\kappa+\omega))$. This scales the output by $1-\epsilon$. For target zero the loss decrease is

$$\|m^0\|^2(\epsilon - \epsilon^2/2) \geq 0,$$

with strict inequality when $m^0 \neq 0$. If the baseline has no reset support, introducing positive reset mass has infinite actual-to-baseline KL, although the candidate operator change is finite. This free-gate example isolates the mass-change term in (72); finite response under parameter sharing additionally depends on the shared gate family.

F.3 Exact finite tilt and its unique positive branch realization

Use the origin labels $\mathcal{I}_t = \{\mathsf{w}_1, \ldots, \mathsf{w}_t, \mathsf{r}_2, \ldots, \mathsf{r}_t\}$, with $x_{\mathsf{w}_i} = z_i$ and $x_{\mathsf{r}_i} = a_i$. Equal attached values do not identify labels. Probability vectors are ordered $(\mathsf{w}_1, \mathsf{w}_2, \mathsf{r}_2, \ldots, \mathsf{w}_t, \mathsf{r}_t)$. Let $p_{\mathrm{orig}}^{0,(t)}$ be the full origin-resolved law under positive baseline gates $q_t^{\mathrm{gate},0}$. Fix one direction $g \in \mathbb{R}^d$ and put $h_i^{\mathrm{orig}} = g^\top x_i$, $h_{\mathsf{w}_t}^{\mathrm{orig}} = g^\top z_t$, $h_{\mathsf{r}_t}^{\mathrm{orig}} = g^\top a_t$. For $\lambda \in \mathbb{R}$ define $q_{\lambda,\mathrm{orig}}^{(t)}(i) \propto p_{\mathrm{orig}}^{0,(t)}(i)e^{-\lambda h_i^{\mathrm{orig}}}$ and $\mathcal{Z}_t^{\mathrm{orig}}(\lambda) = \sum_{i\in\mathcal{I}_t} p_{\mathrm{orig}}^{0,(t)}(i)e^{-\lambda h_i^{\mathrm{orig}}}$.

Theorem F.3 (Exact BAM tilt and ICM inverse)**.** *The positive gate sequence realizing $q^{(T)}_{\lambda,\mathrm{orig}}$ is unique, also realizes every $q^{(t)}_{\lambda,\mathrm{orig}}$, and is*

$$\mathcal{Z}^{\mathrm{orig}}_1 = e^{-\lambda h^{\mathrm{orig}}_{\mathsf{w}_1}}, \qquad \mathcal{Z}^{\mathrm{orig}}_t = \kappa^0_t \mathcal{Z}^{\mathrm{orig}}_{t-1} + \omega^0_t e^{-\lambda h^{\mathrm{orig}}_{\mathsf{w}_t}} + \sigma^0_t e^{-\lambda h^{\mathrm{orig}}_{\mathsf{r}_t}}, \tag{73}$$

$$(\kappa^\lambda_t, \omega^\lambda_t, \sigma^\lambda_t) = \left(\mathcal{Z}^{\mathrm{orig}}_t\right)^{-1} (\kappa^0_t \mathcal{Z}^{\mathrm{orig}}_{t-1}, \omega^0_t e^{-\lambda h^{\mathrm{orig}}_{\mathsf{w}_t}}, \sigma^0_t e^{-\lambda h^{\mathrm{orig}}_{\mathsf{r}_t}}). \tag{74}$$

In the ICM chart $u^{I,\mathrm{k}}_t = \operatorname{logit} f^{\mathrm{I}}_t$, $u^{I,\mathrm{w}}_t = \operatorname{logit} \alpha_t$, the necessary increments are

$$\begin{aligned} u^{I,\mathrm{k},\lambda}_t - u^{I,\mathrm{k},0}_t &= \log \mathcal{Z}^{\mathrm{orig}}_{t-1} + \lambda h^{\mathrm{orig}}_{\mathsf{r}_t}, \\ u^{I,\mathrm{w},\lambda}_t - u^{I,\mathrm{w},0}_t &= -\lambda h^{\mathrm{orig}}_{\mathsf{w}_t} - \log\{f^{\mathrm{I},0}_t \mathcal{Z}^{\mathrm{orig}}_{t-1} + (1 - f^{\mathrm{I},0}_t) e^{-\lambda h^{\mathrm{orig}}_{\mathsf{r}_t}}\}. \end{aligned} \tag{75}$$

All $\lambda \in \mathbb{R}$ give positive gates. If $\|x_i\| \le R$, then $e^{-|\lambda| R \|g\|} \le \mathcal{Z}^{\mathrm{orig}}_t \le e^{|\lambda| R \|g\|}$ and each logit increment has magnitude at most $2|\lambda| R \|g\|$, independently of T.

Proof. Induction on t proves the formula for $\mathcal{Z}^{\mathrm{orig}}_t$. Multiplying the previously tilted old probabilities by $\kappa^0_t \mathcal{Z}^{\mathrm{orig}}_{t-1} / \mathcal{Z}^{\mathrm{orig}}_t$ cancels their $\mathcal{Z}^{\mathrm{orig}}_{t-1}$; the new write and anchor probabilities have the same required numerator. Thus the tilted gates realize every prefix law. Conversely, from a positive terminal law, the last write and anchor probabilities determine ω_T, σ_T, their complement determines κ_T, and normalization of the older probabilities gives the preceding law. Induction backwards determines all gates uniquely.

Use $u^{I,\mathrm{k}}_t = \log(\kappa_t/\sigma_t)$ and $u^{I,\mathrm{w}}_t = \log(\omega_t/(\kappa_t + \sigma_t))$ in (74); subtraction of the baseline logits gives (75). Each $\mathcal{Z}^{\mathrm{orig}}_t$ is an expectation of numbers in $[e^{-|\lambda| R \|g\|}, e^{|\lambda| R \|g\|}]$. The mixture inside the second logarithm obeys the same bounds, which prove both logit estimates. □

This finite exponential change of branch probabilities is related to the established exponential transformation in KL control (Todorov, 2006); the claim here is its exact recurrent inverse and the restrictions induced by shared ICM features. The construction uses one common direction g across all prefixes. With g and λ fixed and the baseline path available, the forward recurrence requires $O(Td)$ arithmetic and constant auxiliary state besides its output. When g is selected from a terminal residual, this computation implements an offline gate update.

F.4 Necessary prefix statistics and a sharp local mismatch

Stack the increments in (75), ordered $(u^{I,\mathrm{k}}_2, u^{I,\mathrm{w}}_2, \dots, u^{I,\mathrm{k}}_T, u^{I,\mathrm{w}}_T)$, as $\Xi(\lambda) = \sum_{j \ge 1} \lambda^j \Xi_j$ on a neighborhood of zero. This real-analytic expansion exists because its partition functions are positive near zero. Define the value-law pushforward

$$\mu^{0,(t-1)}_{\mathrm{val}} = \sum_{i \in \mathcal{I}_{t-1}} p^{0,(t-1)}_{\mathrm{orig},i} \delta_{x_i},$$

and random values

$$\mathsf{V}^{\mathrm{pref}}_{t-1} \sim \mu^{0,(t-1)}_{\mathrm{val}}, \qquad \mathsf{V}^{\mathrm{nw}}_t \sim f^{I,0}_t \mu^{0,(t-1)}_{\mathrm{val}} + (1 - f^{I,0}_t)\delta_{a_t}.$$

This pushforward is for computing statistics; the KL budget still distinguishes every origin label. Define $\chi_j(\mathsf{V})$ as the jth derivative at zero of $\log \mathbb{E} e^{s\mathsf{V}}$. The first two coefficient blocks are

$$\begin{aligned} \Xi_{1,t} &= \begin{pmatrix} h^{\mathrm{orig}}_{\mathsf{r}_t} - \mathbb{E}[g^\top \mathsf{V}^{\mathrm{pref}}_{t-1}] \\ \mathbb{E}[g^\top \mathsf{V}^{\mathrm{nw}}_t] - h^{\mathrm{orig}}_{\mathsf{w}_t} \end{pmatrix}, \qquad \Xi_{2,t} = \tfrac{1}{2} \begin{pmatrix} \operatorname{Var}(g^\top \mathsf{V}^{\mathrm{pref}}_{t-1}) \\ -\operatorname{Var}(g^\top \mathsf{V}^{\mathrm{nw}}_t) \end{pmatrix}, \\ \Xi_{j,t} &= \frac{(-1)^j}{j!} \begin{pmatrix} \chi_j(g^\top \mathsf{V}^{\mathrm{pref}}_{t-1}) \\ -\chi_j(g^\top \mathsf{V}^{\mathrm{nw}}_t) \end{pmatrix} \quad (j \ge 2). \end{aligned} \tag{76}$$

Thus the second-order features involve prefix variances, not just adjacent-candidate squares.

F.5 SHARED REALIZATION, INFORMATION METRIC AND ALL-TASK TEMPLATES

Theorem F.4 (Finite-order shared ICM completion)**.** *At fixed positive baseline branch probabilities and fixed origin-resolved data, declare the linear-logit family $\xi = \xi^0 + \Phi\theta$, $S_{\mathrm{ICM}} = \operatorname{range}\Phi$. Its local task efficiency in the origin-resolved allocation-Fisher metric is one if and only if $\Xi_1 \in S_{\mathrm{ICM}}$ (assuming positive task variance). Exact realization of all sufficiently small signed tilts is equivalent to $\Xi_j \in S_{\mathrm{ICM}}$ for every $j \geq 1$. Realizing the required logits to $O(|\lambda|^{q+1})$ is equivalent to*

$$\operatorname{range}[\Xi_1, \ldots, \Xi_q] \subseteq S_{\mathrm{ICM}}, \qquad k_q^{\mathrm{ICM,orig}} = \operatorname{rank}[\Phi, \Xi_1, \ldots, \Xi_q] - \operatorname{rank}\Phi \tag{77}$$

being the minimum arbitrary added-feature count. Let $s_{t\to T}^{\mathrm{keep},0} = \prod_{l=t+1}^{T} \kappa_l^0$. The full independent-logit Fisher is positive diagonal, with entries

$$\left(F_\xi^{\mathrm{orig},0}\right)_{u_t^{I,\mathrm{k}} u_t^{I,\mathrm{k}}} = s_{t\to T}^{\mathrm{keep},0}(1-\alpha_t^0) f_t^{\mathrm{I},0}(1-f_t^{\mathrm{I},0}), \qquad \left(F_\xi^{\mathrm{orig},0}\right)_{u_t^{I,\mathrm{w}} u_t^{I,\mathrm{w}}} = s_{t\to T}^{\mathrm{keep},0}\alpha_t^0(1-\alpha_t^0). \tag{78}$$

When $\Xi_1, \ldots, \Xi_q \in S_{\mathrm{ICM}}$, the optimal leading mismatch among smooth local curves that match those orders is

$$\inf \lim_{\lambda\to 0} \frac{D_{\mathrm{KL}}(p_{\mathrm{orig}}^{(T)}(\xi(\lambda)) \| q_{\lambda,\mathrm{orig}}^{(T)})}{\lambda^{2q+2}} = \tfrac{1}{2} \operatorname{dist}_{F_\xi^{\mathrm{orig},0}}(\Xi_{q+1}, S_{\mathrm{ICM}})^2. \tag{79}$$

Proof. The terminal positive simplex and the independent ICM logits are smooth bijective charts by the preceding inverse argument. The score for the target tilt is $-h^{\mathrm{orig}} + \mathbb{E}h^{\mathrm{orig}}$, and its unique independent-logit lift is Ξ_1. The natural-gradient projection therefore loses no response exactly when this lift belongs to S_{ICM}, by the same weighted Cauchy–Schwarz equality criterion used in Theorem 2.1.

At branch t, condition first on whether it is reached from the terminal root. The write-versus-nonwrite Bernoulli score has variance $\alpha_t^0(1-\alpha_t^0)$. The conditional keep-versus-reset score is zero on the write branch and has conditional variance $f_t^{\mathrm{I},0}(1-f_t^{\mathrm{I},0})$ on nonwrite, giving the additional factor $1-\alpha_t^0$. Their covariance is zero. Scores at distinct branches are orthogonal: condition on the later branch, and the earlier conditional score has mean zero within its keep subtree. Multiplying by the reach probability proves (78), including positivity.

For exact closure, projection of $\Xi(\lambda)$ onto $S_{\mathrm{ICM}}^{\perp}$ vanishes on an interval if and only if all analytic coefficients vanish there. For finite order, the same projection must be $O(|\lambda|^{q+1})$, so precisely its first q coefficients vanish. A fixed right inverse of Φ realizes the truncated series, proving sufficiency. Each new column contributes at most one missing dimension, and a basis of the missing quotient attains the stated count.

After matching through degree q, the first logit difference is $\lambda^{q+1}(s - \Xi_{q+1}) + o(\lambda^{q+1})$ for some $s \in S_{\mathrm{ICM}}$. A Taylor expansion of the directed KL of a smooth positive statistical family gives half its squared Fisher norm to leading order. The base probabilities are fixed in this expansion. Minimization is a finite-dimensional $F_\xi^{\mathrm{orig},0}$-orthogonal projection, attained by a polynomial parameter curve; this proves (79). □

Equation (77) counts added columns in the fixed origin-resolved evaluation space. For all task directions, polarization of the cumulant coefficients gives tensor-evaluation features computed from the baseline prefix moments. Under candidate-only observation, the reset redistributions characterized in Appendix G.1 provide additional realizations. Appendices G.3–G.5 relate these realizations to finite-order matching and task response, and Appendix I.6 gives the corresponding all-lifts criterion for jointly shared queries.

Task-independent feature templates at a frozen base. The task-specific rank statement has a constructive all-direction version. Replace λg by $\zeta \in \mathbb{R}^d$ and expand the required vector as $\Xi(\zeta) = \sum_{|\alpha|\geq 1} \Xi_\alpha \zeta^\alpha$, where α is a multi-index. The coefficients are contractions of prefix cumulant tensors; at degree one they are the prefix mean contrasts, and at degree two the covariance entries displayed in (76). Realizing every sufficiently small task tilt through degree q is equivalent to

$$\operatorname{span}\{\Xi_\alpha : 1 \leq |\alpha| \leq q\} \subseteq \operatorname{range}\Phi.$$

There are at most $\binom{d+q}{q} - 1$ such vector-valued evaluation columns before dependencies are removed. To prove necessity, project onto $(\operatorname{range}\Phi)^{\perp}$ and differentiate in the components of ζ at zero; sufficiency is the multivariate Taylor polynomial and a right inverse. Each coefficient column specifies both ICM logit components and all time indices simultaneously. The same polynomial coefficients work across sequences when their prefix-statistic evaluations are stacked.

These templates are causal functions of the baseline prefix law and the current candidate and anchor. Moments through degree q propagate by the same convex-mixture recurrence, and cumulant tensors are obtained algebraically. The feature generator therefore computes and stores these baseline-dependent tensors and is recomputed when the baseline gates or candidate encoder change. A reduced basis is defined on its evaluation data.

F.6 From any realizable BAM endpoint to actual budget and descent

Use $L = L_T^{\mathrm{orig}}$, where $L_T^{\mathrm{orig}}(p) = \|Z_{\mathrm{orig}}p - y_T\|^2/2$, and $y = y_T$. Let $p^0 = p_{\mathrm{orig}}^{0,(T)}$, let Z_{orig} contain all candidate and anchor values, and take $g = Z_{\mathrm{orig}}p^0 - y$, $h^{\mathrm{orig}} = Z_{\mathrm{orig}}^{\top}g$. Put $V_h = \mathrm{Var}_{p^0}(h^{\mathrm{orig}}) > 0$, $\Omega_h = \mathrm{osc}(h^{\mathrm{orig}})$, $q_\lambda \propto p^0 e^{-\lambda h^{\mathrm{orig}}}$. For any positive *realizable* endpoint p, define

$$\varepsilon_{\mathrm{sc}} = \mathrm{osc}(\log p - \log q_\lambda).$$

If $0 < \lambda \le R^{-2}$, $\lambda\Omega_h \le 1$, and $\varepsilon_{\mathrm{sc}}^2/8 \le V_h\lambda^2/(4\mathrm{e})$, then

$$L(p^0) - L(p) \ge \frac{V_h\lambda}{4\mathrm{e}}, \qquad D_{\mathrm{KL}}(p\|p^0) \le \frac{(\lambda\Omega_h + \varepsilon_{\mathrm{sc}})^2}{8}, \qquad \|p - p^0\|_1 \le \frac{\lambda\Omega_h + \varepsilon_{\mathrm{sc}}}{2}. \quad (80)$$

Indeed $p = \mathrm{softmax}(\log q_\lambda + \log p - \log q_\lambda)$, so the global score bound and the same directed entropic identity as Theorem 5.2 apply. The candidate-only operator change is no larger than the last bound. For (74), $\varepsilon_{\mathrm{sc}} = 0$, so an explicitly small λ guarantees an actual budget and positive loss decrease. These guarantees are length uniform when R, Ω_h and a positive lower bound on V_h are uniform. For a finite-order shared approximation, (80) uses its actual cumulative score error $\varepsilon_{\mathrm{sc}}$. Length-uniform application additionally requires uniform control of the displayed constants and that error.

F.7 A two-sweep free-gate update

To include extra forgetting, let $m_1 = z_1$ and for $k \ge 2$ use

$$m_k = \kappa_k m_{k-1} + \omega_k z_k + \sigma_k a_k, \qquad q_k^{\mathrm{gate}} = (\kappa_k, \omega_k, \sigma_k), \quad \kappa_k + \omega_k + \sigma_k = 1,$$

with fixed bounded candidates and anchors. Active branches are strictly positive; CNM uses only keep/write, and ICM has $a_k = 0$, $\omega_k = \alpha_k$, $\kappa_k = f_k^I(1-\alpha_k)$, and $\sigma_k = (1-f_k^I)(1-\alpha_k)$. Unrolling distinguishes each write and anchor as a provenance atom, even if values coincide.

For this sweep, $L = L_{\mathcal{O}}^{\mathrm{orig}}$ is the weighted squared loss of the origin-resolved prefix distributions, with the fixed candidate/anchor value matrix at each observation. Extend D_w to those same prefix supports. At a baseline $q^{\mathrm{gate},0}$, let $g_t = m_t^0 - y_t$ and let $q_{\lambda,\mathrm{orig}}^{(t)}$ be the prescribed exponential tilt of its provenance distribution at observation t. Choose $0 < \lambda \le R^{-2}$. The nonzero observation weights sum to one. First compute the baseline recurrence; at unobserved times the product $w_t^{\mathrm{obs}} g_t$ is defined as zero without evaluating an absent target. Initialize the backward aggregation by

$$w_T^{\mathrm{sweep}} = w_T^{\mathrm{obs}}, \qquad g_T^{\mathrm{sweep,sum}} = w_T^{\mathrm{obs}} g_T.$$

For $k = T-1, \dots, 2$, compute

$$w_k^{\mathrm{sweep}} = w_k^{\mathrm{obs}} + \kappa_{k+1}^0 w_{k+1}^{\mathrm{sweep}},$$
$$g_k^{\mathrm{sweep,sum}} = w_k^{\mathrm{obs}} g_k + \kappa_{k+1}^0 g_{k+1}^{\mathrm{sweep,sum}}.$$

Set $\bar{g}_k = g_k^{\mathrm{sweep,sum}}/w_k^{\mathrm{sweep}}$ when $w_k^{\mathrm{sweep}} > 0$, and set $\bar{g}_k = 0$ when $w_k^{\mathrm{sweep}} = 0$. At unobserved times $w_k^{\mathrm{obs}} = 0$ with the zero-product convention above. Gates with $w_k^{\mathrm{sweep}} = 0$ stay fixed. For numerical evaluation, propagate positive weights by log-addition and normalized averages by their

convex-mixture weights; a tiny positive weight is kept active. A forward sweep, in order $k = 2, \dots, T$, starts with $m_1^+ = z_1$, $e_1^{\mathrm{KL}} = 0$, and normalizes active entries of

$$\begin{aligned} q_k^{\mathrm{gate},+} &\propto \left(\kappa_k^0 e^{-e_{k-1}^{\mathrm{KL}} - \lambda (m_{k-1}^+)^\top \bar{g}_k},\ \omega_k^0 e^{-\lambda z_k^\top \bar{g}_k},\ \sigma_k^0 e^{-\lambda a_k^\top \bar{g}_k}\right), \\ m_k^+ &= \kappa_k^+ m_{k-1}^+ + \omega_k^+ z_k + \sigma_k^+ a_k, \\ e_k^{\mathrm{KL}} &= D_{\mathrm{KL}}(q_k^{\mathrm{gate},+} \| q_k^{\mathrm{gate},0}) + \kappa_k^+ e_{k-1}^{\mathrm{KL}}. \end{aligned} \tag{81}$$

Normalize the active entries in log space by log-sum-exp, using log weights $\log \kappa_k^0 - e_{k-1}^{\mathrm{KL}} - \lambda (m_{k-1}^+)^\top \bar{g}_k$, $\log \omega_k^0 - \lambda z_k^\top \bar{g}_k$, and $\log \sigma_k^0 - \lambda a_k^\top \bar{g}_k$. An inactive branch stays zero; when $w_k^{\mathrm{sweep}} = 0$, keep the baseline triple and still propagate m_k^+ and e_k^{KL}. The outputs are the gate path, prefix means and prefix KL values in (81).

Proposition F.5 (Finite free-path correction)**.** *The two sweeps take $O(Td)$ work and storage, preserve active support, and exactly minimize $D_w(p \| q_{\lambda,\mathrm{orig}})$ one gate at a time. They satisfy*

$$L(p^0) - L(p^+) \geq \lambda^{-1} \sum_{k=2}^{T} w_k^{\mathrm{sweep}} D_{\mathrm{KL}}(q_k^{\mathrm{gate},0} \| q_k^{\mathrm{gate},+}) + (\lambda^{-1} - R^2) \sum_t w_t^{\mathrm{obs}} e_t^{\mathrm{KL}} \geq 0$$

for $0 < \lambda \leq R^{-2}$. If the target provenance family is compatible with one path, the sweep realizes it exactly.

Proof. At coordinate k, earlier gates have been changed and later gates remain at baseline. An observation $t \geq k$ reaches the atoms present at k with baseline probability $\prod_{l=k+1}^{t} \kappa_l^0$. Conditional on reaching them, its branches are the old distribution at $k-1$, the new write, and the new anchor. The conditional old-distribution KL to the tilted target equals

$$e_{k-1}^{\mathrm{KL}} + \lambda (m_{k-1}^+)^\top g_t + \log \mathcal{Z}_{k-1,t}^{\mathrm{orig}},$$

where $\mathcal{Z}_{k-1,t}^{\mathrm{orig}}$ is that branch's tilt normalizer. The target keep weight is proportional to $\kappa_k^0 \mathcal{Z}_{k-1,t}^{\mathrm{orig}}$, so this normalizer cancels from its log score. The write/anchor branches have only their exponential-tilt scores. Aggregating with the baseline reaches gives exactly (81). The remaining coordinate objective is $w_k^{\mathrm{sweep}} D_{\mathrm{KL}}(q_k^{\mathrm{gate}} \| q_k^{\mathrm{gate},+})$ plus a constant. Since the as-yet-unmodified coordinate is $q_k^{\mathrm{gate},0}$, the objective improvement is $w_k^{\mathrm{sweep}} D_{\mathrm{KL}}(q_k^{\mathrm{gate},0} \| q_k^{\mathrm{gate},+})$.

The KL chain rule on the three branches proves $e_k^{\mathrm{KL}} = D_{\mathrm{KL}}(p_{\mathrm{orig}}^{+,(k)} \| p_{\mathrm{orig}}^{0,(k)})$. Summing coordinate improvements gives

$$D_w(p^0 \| q_{\lambda,\mathrm{orig}}) - D_w(p^+ \| q_{\lambda,\mathrm{orig}}) = \sum_k w_k^{\mathrm{sweep}} D_{\mathrm{KL}}(q_k^{\mathrm{gate},0} \| q_k^{\mathrm{gate},+}).$$

Insert this identity into the weighted version of (59). Compatible target conditionals are obtained successively by induction, proving exact realization in that case. Each sweep stores a scalar reach, a vector aggregate, a mean, and a scalar KL at each time, giving the stated complexity. □

The sweep performs an offline sequence of exact single-gate minimizations. It realizes compatible origin-resolved exponential targets, including the common-direction targets in Appendix F.3, and gives the stated coordinatewise improvement for incompatible targets. Active branches retain their support. A previously absent reset branch has zero baseline mass, so introducing positive mass there has infinite actual-to-baseline KL.

G CANDIDATE OBSERVATIONS: ALL REALIZATIONS AND TASK RESPONSE

This appendix characterizes all positive gate paths producing prescribed candidate coefficients and relates their budget-feasible shared image to task response. Candidate observation coarsens the origin-label space, while parameter sharing restricts the gate paths. The KL identities are derived directly in the finite categorical model. Information-projection theory provides broader context through generalized-projection existence results for convex or log-convex sets with finite divergence infimum (Csiszár & Matúš, 2003, Section I-A, Theorem 1), and Yamaguchi & Nozawa (2024) study quantitative near-sufficiency of sample-space statistics.

G.1 THE COMPLETE FINITE FIBER, INCLUDING MULTIPLE OBSERVATIONS

Fix $T \geq 2$ and zero anchors for the candidate-only observation problem. Define the open domain of terminal candidate coefficients by

$$\mathcal{E}_T = \{\eta \in (0,\infty)^T : \mathbf{1}^\top \eta < 1\}.$$

For $\eta \in \mathcal{E}_T$, write $M = \sum_i \eta_i$. Choose positive terminal reset-origin masses $r_2^{\mathrm{reset}}, \ldots, r_T^{\mathrm{reset}}$ with $\sum_{t=2}^T r_t^{\mathrm{reset}} = 1 - M$. Define

$$S_1^{\mathrm{end}} = \eta_1, \qquad S_t^{\mathrm{end}} = \sum_{i=1}^t \eta_i + \sum_{k=2}^t r_k^{\mathrm{reset}}, \qquad S_T^{\mathrm{end}} = 1. \tag{82}$$

The corresponding ICM logits, in the order $(u_2^{I,\mathrm{k}}, u_2^{I,\mathrm{w}}, \ldots, u_T^{I,\mathrm{k}}, u_T^{I,\mathrm{w}})$, are

$$\mathcal{L}(\eta, r^{\mathrm{reset}})_{u_t^{I,\mathrm{k}}} = \log \frac{S_{t-1}^{\mathrm{end}}}{r_t^{\mathrm{reset}}}, \qquad \mathcal{L}(\eta, r^{\mathrm{reset}})_{u_t^{I,\mathrm{w}}} = \log \frac{\eta_t}{S_{t-1}^{\mathrm{end}} + r_t^{\mathrm{reset}}}. \tag{83}$$

The inverse uses terminal candidate and reset masses. Let $\mathcal{O} = \{t_1, \ldots, t_{N_{\mathrm{obs}}}\}$, where $2 \leq t_1 < \cdots < t_{N_{\mathrm{obs}}} = T$, and put $t_0 = 1$. An observed candidate sequence $\eta_{\mathrm{req}}^{(t_j)} \in \mathbb{R}^{t_j}$ specifies the coefficient of every individual write in m_{t_j}, and thereby specifies the conditional input operator.

Theorem G.1 (All candidate-equivalent positive gate paths)**.** *For a terminal table η, every positive gate path realizing it, and no other path, is obtained by choosing the positive reset masses above and using* (83)*. Its gates are*

$$(\kappa_t, \omega_t, \sigma_t) = (S_{t-1}^{\mathrm{end}}, \eta_t, r_t^{\mathrm{reset}})/S_t^{\mathrm{end}}. \tag{84}$$

For several positive observed tables set $\eta = \eta_{\mathrm{req}}^{(T)} \in \mathcal{E}_T$ and $S_0^{\mathrm{obs}} = \eta_1$. The tables are jointly realizable by positive gates iff, for each j, there is $S_j^{\mathrm{obs}} > 0$ such that

$$\eta_{\mathrm{req},i}^{(t_j)} = \eta_i / S_j^{\mathrm{obs}} \ (i \leq t_j), \quad S_{N_{\mathrm{obs}}}^{\mathrm{obs}} = 1, \qquad r_j^{\mathrm{int}} := S_j^{\mathrm{obs}} - S_{j-1}^{\mathrm{obs}} - \sum_{i=t_{j-1}+1}^{t_j} \eta_i > 0. \tag{85}$$

When realizable, all paths are obtained by distributing total mass r_j^{int} positively among the reset indices $t_{j-1}+1, \ldots, t_j$, independently for each interval, and then applying (84)*. The fiber is a product of open simplexes of dimension*

$$\sum_{j=1}^{N_{\mathrm{obs}}} (t_j - t_{j-1} - 1) = T - 1 - N_{\mathrm{obs}}. \tag{86}$$

The map from the $2(T-1)$ independent ICM logits to the observed candidate tables has constant rank $T - 1 + N_{\mathrm{obs}}$.

Proof. Given $(\eta, r^{\mathrm{reset}})$, positivity gives $0 < S_{t-1}^{\mathrm{end}} < S_t^{\mathrm{end}}$. The gates in (84) are positive and sum to one. Products telescope: $\prod_{k=t+1}^T \kappa_k = S_t^{\mathrm{end}}$, so the terminal write and reset masses are exactly $\eta_t, r_t^{\mathrm{reset}}$, with initial mass $S_1^{\mathrm{end}} = \eta_1$. Conversely, expanding any positive gate path gives these terminal masses, and backward recovery in Theorem F.3 yields the same inverse.

At time t_j, a write present there survives to T with the common factor $S_j^{\mathrm{obs}} = \prod_{k=t_j+1}^T \kappa_k = S_{t_j}^{\mathrm{end}}$. Thus all ratios $\eta_i / \eta_{\mathrm{req},i}^{(t_j)}$ are equal to S_j^{obs}, and S_j^{obs} is uniquely determined, for example by $i = 1$. Taking differences of (82) shows that the total reset mass in interval j equals r_j^{int}. Necessity follows. Conversely, choose positive reset masses with these interval totals. Then $S_{t_j}^{\mathrm{end}} = S_j^{\mathrm{obs}}$ and the telescoping products give every specified observed table. An interval with ℓ reset positions has an open simplex of dimension $\ell - 1$, proving (86).

Finally, the full origin-resolved positive simplex is a smooth chart of the logits. In that simplex, $(\eta, S_1^{\mathrm{obs}}, \ldots, S_{N_{\mathrm{obs}}-1}^{\mathrm{obs}})$ are $T + N_{\mathrm{obs}} - 1$ independent coordinates on an open set defined by the displayed strict inequalities; the remaining coordinates are the within-interval reset splits. The observed tables determine these coordinates smoothly, and are recovered smoothly from them by (85). This proves the constant rank. □

For $N_{\mathrm{obs}} = 1$ the reset fiber has dimension $T-2$; for all prefixes $N_{\mathrm{obs}} = T-1$ it is a singleton. These dimensions count free-gate variations invisible to the specified candidate observations. The case $T = 2$ has one reset origin and a zero-dimensional fiber even at the terminal observation.

Exact shared feasibility over all endpoints. For the declared shared family $\xi = \xi^0 + \Phi\theta$ with unrestricted θ, put $S_{\mathrm{ICM}} = \operatorname{range}\Phi$. A terminal candidate sequence η is attainable iff

$$\exists r_t^{\mathrm{reset}} > 0: \quad \sum_{t=2}^{T} r_t^{\mathrm{reset}} = 1 - \sum_i \eta_i, \qquad \mathcal{L}(\eta, r^{\mathrm{reset}}) - \xi^0 \in S_{\mathrm{ICM}}. \tag{87}$$

For multiple observations add the interval-total constraints of Theorem G.1. The finite feasibility criterion quantifies over all positive reset allocations, including those far from the baseline. If every prefix candidate sequence is prescribed by the common-direction tilt, the unique path is the origin-resolved tilt in Theorem F.3. At all prefixes the inverse is explicit: $\kappa_t = \eta^{(t)}_{\mathrm{req},1}/\eta^{(t-1)}_{\mathrm{req},1}$, $\omega_t = \eta^{(t)}_{\mathrm{req},t}$, and $\sigma_t = 1 - \kappa_t - \omega_t$, with $\eta^{(1)}_{\mathrm{req},1} = 1$. Here $\eta^{(t)}_{\mathrm{req}}$ is indexed by recurrent time. This positive analytic inverse also preserves finite Taylor order. The origin-resolved finite-order feature criterion therefore remains necessary for those *candidate-only tables*, even though a KL on origin-resolved resets and a KL on coarsened observations are still different costs.

G.2 Coarsened Fisher geometry and all first-order lifts

Write $\mathcal{C}(\xi) = \eta(\xi)$ for the terminal candidate coefficient map and $\bar{p}(\eta) = (\eta, 1 - \mathbf{1}^\top\eta)$. A superscript 0 denotes the fixed baseline, whose branch probabilities and origin masses are positive. Set

$$J_{\mathrm{cand}} = D\mathcal{C}(\xi^0), \qquad F_\eta^{\mathrm{agg},0} = \operatorname{Diag}(1/\eta^0) + \frac{\mathbf{1}\mathbf{1}^\top}{1 - M_0}, \qquad F_\xi^{\mathrm{agg},0} = J_{\mathrm{cand}}^\top F_\eta^{\mathrm{agg},0} J_{\mathrm{cand}}. \tag{88}$$

Theorem G.1 gives $\operatorname{rank} J_{\mathrm{cand}} = T$ and $\dim\ker J_{\mathrm{cand}} = T-2$. The matrix $F_\eta^{\mathrm{agg},0}$ is positive definite; $F_\xi^{\mathrm{agg},0}$ vanishes precisely on infinitesimal redistributions of the fixed-zero reset origins. For shared coordinates the zero-cost directions are $\ker(\Phi^\top F_\xi^{\mathrm{agg},0}\Phi) = \ker(J_{\mathrm{cand}}\Phi)$, not necessarily $\ker\Phi$. This is the local quotient for the chosen candidate observation; the origin-resolved metric uses a different kernel. Writing $r^{\mathrm{reset}} = (1-M)\zeta^{\mathrm{reset}}$ with $\sum_t \zeta_t^{\mathrm{reset}} = 1$, the full origin-resolved Fisher separates as

$$\dot{\xi}^\top F_\xi^{\mathrm{orig},0}\dot{\xi} = \dot{\eta}^\top F_\eta^{\mathrm{agg},0}\dot{\eta} + (1 - M_0)\sum_{t=2}^{T} \frac{(\dot{\zeta}_t^{\mathrm{reset}})^2}{\zeta_t^{\mathrm{reset},0}}. \tag{89}$$

To verify it, expand the candidate and reset probability scores and use $\sum_t \dot{\zeta}_t^{\mathrm{reset}} = 0$ to cancel the cross term. Under this fixed coarsening, the observed statistical score is the conditional mean of the origin score. The Fisher information loss is therefore the expected conditional score covariance, in the classical information-loss setting of Amari (1982). Equation (89) specializes the identity to the allocation among reset origins.

G.3 Finite-order matching over all first-order lifts

For candidate values Z and output direction g, write $h = Z^\top g$ and consider the *coarsened* entropic target

$$\bar{q}_\lambda \propto \bar{p}^0 \odot \exp\{-\lambda(h_1, \dots, h_T, 0)\}, \qquad \eta^*(\lambda) = (\bar{q}_{\lambda,1}, \dots, \bar{q}_{\lambda,T}) = \eta^0 + \lambda\eta^\star_{[1]} + \lambda^2\eta^\star_{[2]} + O(\lambda^3). \tag{90}$$

Here $\eta^\star_{[1]} = -\big(F_\eta^{\mathrm{agg},0}\big)^{-1}h$ and $\big(\big(F_\eta^{\mathrm{agg},0}\big)^{-1}\big) = \operatorname{Diag}(\eta^0) - \eta^0(\eta^0)^\top$, as direct multiplication verifies. Assume $h^\top\big(F_\eta^{\mathrm{agg},0}\big)^{-1}h > 0$. For a subspace E, define $\operatorname{dist}_{F_\eta^{\mathrm{agg},0}}(v, E)^2 = \inf_{e\in E}(v-e)^\top F_\eta^{\mathrm{agg},0}(v-e)$.

Proposition G.2 (Operational local response and finite-order repair)**.** *For $\xi \in \xi^0 + S_{\mathrm{ICM}}$, local sufficiency for the requested task in the coarsened allocation-Fisher metric is equivalent to $\eta^\star_{[1]} \in J_{\mathrm{cand}}S_{\mathrm{ICM}}$. A second-order match of* (90) *exists iff some $\xi_{[1]} \in S_{\mathrm{ICM}}$ with $J_{\mathrm{cand}}\xi_{[1]} = \eta^\star_{[1]}$ satisfies*

$$\eta^\star_{[2]} - \tfrac{1}{2}D^2\mathcal{C}(\xi^0)[\xi_{[1]}, \xi_{[1]}] \in J_{\mathrm{cand}}S_{\mathrm{ICM}}. \tag{91}$$

When $\eta^\star_{[1]} \in J_{\mathrm{cand}} S_{\mathrm{ICM}}$, the infimum of fourth-order coarsened KL coefficients among C^2 local curves with the correct first derivative is

$$\frac{1}{2} \inf_{\xi_{[1]} \in S_{\mathrm{ICM}}:\, J_{\mathrm{cand}}\xi_{[1]} = \eta^\star_{[1]}} \mathrm{dist}_{F_\eta^{\mathrm{agg},0}}\Big(\eta^\star_{[2]} - \tfrac{1}{2} D^2\mathcal{C}(\xi^0)[\xi_{[1]}, \xi_{[1]}],\, J_{\mathrm{cand}} S_{\mathrm{ICM}}\Big)^2 . \tag{92}$$

If $J_{\mathrm{cand}} S_{\mathrm{ICM}} = \mathbb{R}^T$, every sufficiently small coarsened target has a smooth exact realization in $\xi^0 + S_{\mathrm{ICM}}$. No length-uniform neighborhood is asserted without uniform conditioning and derivative bounds.

Proof. The local objective is $-h^\top \dot\eta$ and its squared budget is $\dot\eta^\top F_\eta^{\mathrm{agg},0} \dot\eta$. The free maximizer is proportional to $-\big(F_\eta^{\mathrm{agg},0}\big)^{-1} h = \eta^\star_{[1]}$. Weighted Cauchy–Schwarz gives equality for the shared image exactly when $\eta^\star_{[1]} \in J_{\mathrm{cand}} S_{\mathrm{ICM}}$. Equivalently, the squared response is $h^\top J_{\mathrm{cand}} \Phi (\Phi^\top J_{\mathrm{cand}}^\top F_\eta^{\mathrm{agg},0} J_{\mathrm{cand}} \Phi)^\dagger \Phi^\top J_{\mathrm{cand}}^\top h$; its free value is $h^\top \big(F_\eta^{\mathrm{agg},0}\big)^{-1} h$.

Expand $\xi(\lambda) = \xi^0 + \lambda\xi_{[1]} + \lambda^2 \xi_{[2]} + o(\lambda^2)$ with $\xi_{[1]}, \xi_{[2]} \in S_{\mathrm{ICM}}$. Then $\mathcal{C}(\xi(\lambda)) = \eta^0 + \lambda J_{\mathrm{cand}}\xi_{[1]} + \lambda^2\{J_{\mathrm{cand}}\xi_{[2]} + D^2\mathcal{C}[\xi_{[1]}, \xi_{[1]}]/2\} + o(\lambda^2)$. The first two matching equations prove the existence criterion; a polynomial logit curve attains an $O(\lambda^3)$ error when both equations are satisfied. For a fixed first lift $\xi_{[1]}$, minimize the squared $F_\eta^{\mathrm{agg},0}$-norm of the second residual over $\xi_{[2]} \in S_{\mathrm{ICM}}$. A Taylor expansion of categorical KL gives half this norm at order λ^4. Taking the infimum over *all* first lifts gives (92). An infimum of zero is not being identified with an attained zero when the first-lift set is unbounded; the existence statement is the existential condition (91). Finally, if $J_{\mathrm{cand}}\Phi$ has row rank T, select T parameter directions with invertible derivative and apply the inverse function theorem to $\mathcal{C}$ restricted to that slice. □

At any finite order q, write the desired coefficients as $\eta^\star_{[j]}$. For a polynomial logit path $\xi^0 + \sum_{j=1}^q \lambda^j \xi_{[j]}$, its order-$j$ coefficient under $\mathcal{C}$ is $J_{\mathrm{cand}}\xi_{[j]} + E_j(\xi_{[1]}, \dots, \xi_{[j-1]})$, where E_j is the coefficient obtained with $\xi_{[j]} = 0$. Thus matching through degree q is equivalent to a simultaneous solution of

$$\xi_{[j]} \in S_{\mathrm{ICM}}, \qquad J_{\mathrm{cand}}\xi_{[j]} = \eta^\star_{[j]} - E_j(\xi_{[1]}, \dots, \xi_{[j-1]}), \quad 1 \le j \le q.$$

Analyticity of $\mathcal{C}$ gives the sufficiency by Taylor's theorem and necessity by differentiation of a matching path through degree q. These conditions characterize matching through a fixed finite order q. Earlier admissible lifts remain joint unknowns in the system. The resulting all-lifts formulation is analogous to the joint-query criterion in Appendix I.6.

G.4 An exact task-gap identity at an actual KL budget

For any declared shared subspace S_{ICM}, define $\mathcal{P}_{S_{\mathrm{ICM}}} = \{\bar p(\mathcal{C}(\xi^0 + s)) : s \in S_{\mathrm{ICM}}\}$, $h = Z^\top g$, and $\bar h = (h, 0)$. On this coarsened support set $\mathcal{K}_{\mathrm{agg}}(p) = D_{\mathrm{KL}}(p\|\bar p^0)$ and $G_h(p) = -\bar h^\top (p - \bar p^0)$. The candidate-operator cost is $\mathcal{V}_{\mathrm{cand}}(\eta) = \|\eta - \eta^0\|_1$, not the full ℓ_1 distance of $\bar p$. Let

$$B^{\mathrm{KL}}_{S_{\mathrm{ICM}}}(\tau) = \sup_{p \in \mathcal{P}_{S_{\mathrm{ICM}}} : \mathcal{K}_{\mathrm{agg}}(p) \le \tau} G_h(p), \qquad B^{\mathrm{op}}_{S_{\mathrm{ICM}}}(\delta) = \sup_{p = \bar p(\eta) \in \mathcal{P}_{S_{\mathrm{ICM}}} : \mathcal{V}_{\mathrm{cand}}(\eta) \le \delta} G_h(p). \tag{93}$$

The free values allow every $\eta \in \mathcal{E}_T$. These are linear task gains; squared loss is treated separately below. For nonconstant $\bar h$, choose $\lambda > 0$ and put

$$\bar q_{\lambda,i} = \frac{\bar p^0_i e^{-\lambda \bar h_i}}{\mathcal{Z}_\lambda}, \qquad \mathcal{Z}_\lambda = \sum_i \bar p^0_i e^{-\lambda \bar h_i}, \qquad \tau = D_{\mathrm{KL}}(\bar q_\lambda \| \bar p^0).$$

This is a nonsaturated free KL budget, specified by a finite positive tilt.

Lemma G.3 (Exact task gap)**.** *The free optimum is $B^{\mathrm{KL}}_{\mathrm{free}}(\tau) = \mathbb{E}_{\bar p^0} \bar h - \mathbb{E}_{\bar q_\lambda} \bar h$. Every $p \in \mathcal{P}_{S_{\mathrm{ICM}}}$ with $\mathcal{K}_{\mathrm{agg}}(p) \le \tau$ satisfies*

$$B^{\mathrm{KL}}_{\mathrm{free}}(\tau) - G_h(p) = \frac{D_{\mathrm{KL}}(p\|\bar q_\lambda) + \tau - \mathcal{K}_{\mathrm{agg}}(p)}{\lambda}. \tag{94}$$

Consequently $B^{\mathrm{KL}}_{S_{\mathrm{ICM}}}(\tau) = B^{\mathrm{KL}}_{\mathrm{free}}(\tau)$ *if and only if*

$$\bar{q}_\lambda \in \overline{\{p \in \mathcal{P}_{S_{\mathrm{ICM}}} : \mathcal{K}_{\mathrm{agg}}(p) \le \tau\}}, \tag{95}$$

where the bar denotes topological closure in the finite-dimensional probability simplex.

Proof. The identity $\log(\bar{q}_\lambda/\bar{p}^0) = -\lambda\bar{h} - \log \mathcal{Z}_\lambda$ gives

$$D_{\mathrm{KL}}(p\|\bar{q}_\lambda) = \mathcal{K}_{\mathrm{agg}}(p) + \lambda\mathbb{E}_p\bar{h} + \log \mathcal{Z}_\lambda, \qquad \tau = -\lambda\mathbb{E}_{\bar{q}_\lambda}\bar{h} - \log \mathcal{Z}_\lambda.$$

Adding proves (94). Both numerator terms are nonnegative for a feasible distribution and vanish at $\bar{q}_\lambda$, proving free optimality. If the shared supremum equals that optimum, a shared feasible sequence with approaching gains has $D_{\mathrm{KL}}(p_j\|\bar{q}_\lambda) \to 0$ by (94), hence $p_j \to \bar{q}_\lambda$ by Pinsker. Conversely a sequence in the feasible shared image converging to $\bar{q}_\lambda$ has convergent gain. □

At matched actual cost $\mathcal{K}_{\mathrm{agg}}(p) = \tau$, the task gap is exactly $D_{\mathrm{KL}}(p\|\bar{q}_\lambda)/\lambda$; unused budget otherwise contributes separately. Realizing this single endpoint suffices for equality. More generally, (95) requires a sequence of shared allocations that remains budget-feasible while approaching $\bar{q}_\lambda$. This is topological closure of the feasible image; closure under tilts in Theorem 2.1 concerns realization of every specified target. Exponential-family duality provides the variational background (Wainwright & Jordan, 2008, Section 3.6); related KL-constrained response characterizations appear in Lam (2016). Here the direct finite-sum identity connects the budget-feasible shared image to the task response.

G.5 Two distinct consequences for realization and response

G.5.1 The same candidate target can have different origin-resolved realizations

Take $T = 3$, scalar candidates $(z_1, z_2, z_3) = (1, -1, 0)$, zero anchors, and baseline $q_2^{\mathrm{gate},0} = q_3^{\mathrm{gate},0} = (1/3, 1/3, 1/3)$. Then $\xi^0 = (0, -\log 2, 0, -\log 2)^\top$. Declare the fixed three-column shared family

$$\xi = \xi^0 + \Phi\theta, \qquad \Phi = \begin{pmatrix} -1 & 0 & 0 \\ 1/2 & -1 & 0 \\ -1/3 & -1/3 & 0 \\ 1/6 & 1/6 & -1 \end{pmatrix}. \tag{96}$$

These are the baseline first-order coefficient columns for the three independent candidate scores (all reset scores are zero). Hence they are task-independent at this frozen baseline. For $g = 1$,

$$\Xi_1 = (-1, 3/2, 0, 0)^\top \in S_{\mathrm{ICM}}, \qquad \Xi_2 = (0, -1/8, 1/3, -1/6)^\top \notin S_{\mathrm{ICM}}.$$

The full origin-resolved Fisher is $F_\xi^{\mathrm{orig},0} = \mathrm{Diag}(1/18, 2/27, 1/6, 2/9)$. Weighted projection gives

$$\tfrac{1}{2}\,\mathrm{dist}_{F_\xi^{\mathrm{orig},0}}(\Xi_2, S_{\mathrm{ICM}})^2 = 3/512 > 0. \tag{97}$$

For example its residual is $(-9/32, -9/64, 3/16, 0)^\top$, which is $F_\xi^{\mathrm{orig},0}$-orthogonal to S_{ICM} and has the stated norm. Thus one additional direction is necessary to match the origin-resolved target through second order.

Nevertheless no additional feature is needed to realize the terminal *candidate* target exactly on a neighborhood. To construct this realization, write $\mathcal{Z}_{\mathrm{raw}} = 2\cosh\lambda + 7$. The required candidate masses are

$$\eta(\lambda) = \mathcal{Z}_{\mathrm{raw}}^{-1}(e^{-\lambda}, e^{\lambda}, 3), \qquad (r_2^{\mathrm{reset}}, r_3^{\mathrm{reset}}) = \mathcal{Z}_{\mathrm{raw}}^{-1}(4\zeta_2^{\mathrm{reset}}, 4(1-\zeta_2^{\mathrm{reset}})), \quad 0 < \zeta_2^{\mathrm{reset}} < 1.$$

The baseline has $\zeta_2^{\mathrm{reset},0} = 1/4$. Since S_{ICM} is the kernel of the row $(-1/2, -1/3, 1, 0)$, the inverse in (83) belongs to $\xi^0 + S_{\mathrm{ICM}}$ iff

$$\begin{aligned} 0 = \varphi(\lambda, \zeta_2^{\mathrm{reset}}) = \log \frac{2\cosh\lambda + 4\zeta_2^{\mathrm{reset}}}{4(1-\zeta_2^{\mathrm{reset}})} + \frac{\lambda}{2} + \frac{1}{2}\log(4\zeta_2^{\mathrm{reset}}) \\ - \frac{1}{3}\left\{\lambda - \log(e^{-\lambda} + 4\zeta_2^{\mathrm{reset}}) + \log 2\right\}. \end{aligned} \tag{98}$$

At $(0, 1/4)$, $\varphi = 0$ and $\partial_{\zeta_2^{\mathrm{reset}}}\varphi = 16/3$. The implicit function theorem supplies a unique analytic $\zeta_2^{\mathrm{reset}}(\lambda)$ locally, with

$$\zeta_2^{\mathrm{reset}}(\lambda) = \tfrac{1}{4} - \tfrac{9}{128}\lambda^2 + O(\lambda^3).$$

This realizes the exact terminal candidate vector in the original shared family, while reallocating only the hidden reset origins relative to the origin-resolved target. In particular, the realization preserves the target's total candidate mass. The origin-resolved mismatch equals $\frac{4}{\mathcal{Z}_{\mathrm{raw}}} d_{\mathrm{Bern}}(\zeta_2^{\mathrm{reset}}(\lambda)\|1/4) = \frac{3}{512}\lambda^4 + O(\lambda^5)$, whereas the coarsened mismatch is identically zero. The output is $-2\sinh\lambda/(2\cosh\lambda + 7)$, which improves the fixed target $y = -1$ for small positive λ and obeys the coarsened budget certificate below. If the time-two candidate sequence is also required to match the tilted base, Theorem G.1 removes the reset freedom; the origin-resolved second-order obstruction returns for that multi-observation realization problem. This multi-observation problem uses the specified observed-candidate budget.

Coarsening does not make every shared family sufficient. Restricting (96) to $S_{\mathrm{ICM},1} = \mathrm{span}\{\Xi_1\}$ leaves the last write gate fixed, so $\eta_3 = 1/3$ at every parameter. The target has $\eta_3^*(\lambda) = 3/(2\cosh\lambda + 7) < 1/3$ for every $\lambda \neq 0$, proving exact candidate-target infeasibility globally in this smaller family. Its first lift is uniquely Ξ_1. At second order, $\eta^\star_{[2]} = (7/162, 7/162, -1/27)^\top$ and $D^2\mathcal{C}(\xi^0)[\Xi_1, \Xi_1]/2 = (1/72, 1/36, 0)^\top$. Projection onto $J_{\mathrm{cand}}\Xi_1$ chooses $\xi_{[2]} = -\Xi_1/16$, leaving candidate residual $(29, 29, -48)^\top/1296$ in the metric $F_\eta^{\mathrm{agg},0} = \mathrm{Diag}(9, 9, 3) + (9/4)\mathbf{1}\mathbf{1}^\top$. Formula (92) therefore gives the strictly positive coefficient $275/41472$. This coefficient quantifies candidate-target mismatch in the one-column family. Appendix G.5.2 determines the corresponding optimal task responses under the two budgets.

G.5.2 One feature: no operator deficit, but a higher-order KL deficit

Use the one-column family $S_{\mathrm{ICM},1} = \mathrm{span}\{\Xi_1\}$ above, the same candidates and baseline, and

$$\xi(\vartheta) = \xi^0 + \vartheta\Xi_1 = \xi^0 + \vartheta(-1, 3/2, 0, 0)^\top, \qquad \vartheta \in \mathbb{R}.$$

Here ϑ is a scalar control, not recurrent time. Put $x_{\mathrm{exp}} = e^{\vartheta/2}$ and $d_{\mathrm{den}}(x_{\mathrm{exp}}) = (2+x_{\mathrm{exp}}^3)(1+x_{\mathrm{exp}}^2)$. For this one-parameter example write $p(\vartheta)$ as a local alias for the aggregate law $\bar{p}(\eta(\vartheta))$. Direct expansion of the two updates gives

$$p(\vartheta) = \left(\frac{2}{3d_{\mathrm{den}}(x_{\mathrm{exp}})}, \frac{x_{\mathrm{exp}}^3}{3(2+x_{\mathrm{exp}}^3)}, \frac{1}{3}, \frac{2 + 4x_{\mathrm{exp}}^2 + x_{\mathrm{exp}}^3 + x_{\mathrm{exp}}^5}{3d_{\mathrm{den}}(x_{\mathrm{exp}})}\right),$$

$$p^0 = (1/9, 1/9, 1/3, 4/9). \tag{99}$$

The third, zero-valued candidate remains distinct from the merged reset atom. For $h = (1, -1, 0)$ set $G(\vartheta) = p_2(\vartheta) - p_1(\vartheta)$, $\mathcal{K}_1(\vartheta) \coloneqq \mathcal{K}_{\mathrm{agg}}(p(\vartheta))$, and $\mathcal{V}_1(\vartheta) \coloneqq \mathcal{V}_{\mathrm{cand}}((p_1(\vartheta), p_2(\vartheta), p_3(\vartheta)))$.

Proposition G.4 (Exact restricted response under two budgets)**.** *Although this family cannot realize the nonzero coarsened entropic target,*

$$B^{\mathrm{op}}_{S_{\mathrm{ICM},1}}(\delta) = B^{\mathrm{op}}_{\mathrm{free}}(\delta) = \delta, \qquad 0 \le \delta < 1/3. \tag{100}$$

For $0 < \tau < \tau_{\mathrm{sat}} \coloneqq \frac{1}{3}\log(9/4)$ the unique positive solution ϑ_τ of $\mathcal{K}_1(\vartheta_\tau) = \tau$ is the globally optimal shared parameter. Thus

$$B^{\mathrm{KL}}_{S_{\mathrm{ICM},1}}(\tau) = G(\vartheta_\tau), \qquad B^{\mathrm{KL}}_{\mathrm{free}}(\tau) - B^{\mathrm{KL}}_{S_{\mathrm{ICM},1}}(\tau) = \frac{275}{1536}\tau^{3/2} + O(\tau^2) \quad (\tau \downarrow 0). \tag{101}$$

At the respective saturation budgets $\delta = 1/3$ and $\tau = \tau_{\mathrm{sat}}$, the shared best gain is the supremum $1/3$, not attained at any finite parameter.

Proof. Candidate-operator budget. For $\vartheta \ge 0$, p_1 strictly decreases, p_2 strictly increases, and p_3 is constant. Therefore

$$\mathcal{V}_1(\vartheta) = (1/9 - p_1(\vartheta)) + (p_2(\vartheta) - 1/9) = G(\vartheta),$$

which strictly increases from zero to $1/3$. Negative ϑ has negative gain and cannot improve on the feasible baseline. Every $0 \le \delta < 1/3$ is attained by a unique nonnegative parameter. Any free endpoint obeys $G_h(\bar{p}(\eta)) \le \|h\|_\infty \|\eta - \eta^0\|_1 = \mathcal{V}_{\mathrm{cand}}(\eta)$, proving (100).

Actual KL and global optimality. Write $p_1 = p_1(\vartheta)$, $p_2 = p_2(\vartheta)$, $p_{\mathrm{reset}} = p_4(\vartheta)$ and $p^0_{\mathrm{reset}} = p_4(0)$. Since $p_3' = 0$ and $p_1' + p_2' + p'_{\mathrm{reset}} = 0$,

$$\mathcal{K}_1'(\vartheta) = p_1'(\vartheta) \log \frac{p_1/p_{\mathrm{reset}}}{p_1^0/p^0_{\mathrm{reset}}} + p_2'(\vartheta) \log \frac{p_2/p_{\mathrm{reset}}}{p_2^0/p^0_{\mathrm{reset}}}. \tag{102}$$

The ratios are

$$\frac{p_1}{p_{\mathrm{reset}}} = \frac{2}{2 + 4x_{\mathrm{exp}}^2 + x_{\mathrm{exp}}^3 + x_{\mathrm{exp}}^5}, \qquad \frac{p_2}{p_{\mathrm{reset}}} = \frac{x_{\mathrm{exp}}^3 + x_{\mathrm{exp}}^5}{2 + 4x_{\mathrm{exp}}^2 + x_{\mathrm{exp}}^3 + x_{\mathrm{exp}}^5}.$$

The first strictly decreases for $x_{\mathrm{exp}} > 0$. The derivative of $(x_{\mathrm{exp}}^3 + x_{\mathrm{exp}}^5)/(2 + 4x_{\mathrm{exp}}^2)$ has positive numerator $6x_{\mathrm{exp}}^2 + 14x_{\mathrm{exp}}^4 + 12x_{\mathrm{exp}}^6$, so the second strictly increases. On $\vartheta > 0$ both terms of (102) are positive. Hence $\mathcal{K}_1$ increases from zero to τ_{sat}, obtained from $p(\vartheta) \to (0, 1/3, 1/3, 1/3)$. Since G increases too and negative controls have negative gain, the KL boundary is the global shared optimum. This also proves the saturation assertions.

The precise small-budget deficit. Expanding the explicit probabilities with natural logarithms gives

$$G(\vartheta) = \frac{2}{9}\vartheta + \frac{1}{72}\vartheta^2 - \frac{7}{216}\vartheta^3 + O(\vartheta^4), \qquad \mathcal{K}_1(\vartheta) = \frac{1}{9}\vartheta^2 + \frac{1}{72}\vartheta^3 - \frac{131}{4608}\vartheta^4 + O(\vartheta^5). \tag{103}$$

Since $G'(0) > 0$, gain itself is a local analytic coordinate. Denote it by G_{val} in the rest of the proof. Inverting the gain series and substituting gives

$$\mathcal{K}^{\mathrm{gain}}_{S_{\mathrm{ICM},1}}(G_{\mathrm{val}}) \coloneqq \mathcal{K}_1(G^{-1}(G_{\mathrm{val}})) = \frac{9}{4}G_{\mathrm{val}}^2 + \frac{11907}{8192}G_{\mathrm{val}}^4 + O(G_{\mathrm{val}}^5). \tag{104}$$

For the free reference, $\bar{h} = (1, -1, 0, 0)$ and the log partition is

$$A(\lambda) = \log \frac{2\cosh\lambda + 7}{9} = \frac{\lambda^2}{9} + \frac{\lambda^4}{324} + O(\lambda^6).$$

Its gain is $G_{\mathrm{val}} = A'(\lambda)$ and its cost is $\lambda A'(\lambda) - A(\lambda)$. The exponential-tilt identity makes this the minimum cost at gain G_{val}: the excess cost of any distribution with that mean score equals its KL to the tilt. Inverting $G_{\mathrm{val}} = A'(\lambda)$ yields

$$\begin{aligned} \mathcal{K}^{\mathrm{gain}}_{\mathrm{free}}(G_{\mathrm{val}}) &= \frac{9}{4}G_{\mathrm{val}}^2 - \frac{81}{64}G_{\mathrm{val}}^4 + O(G_{\mathrm{val}}^6), \\ \mathcal{K}^{\mathrm{gain}}_{S_{\mathrm{ICM},1}}(G_{\mathrm{val}}) - \mathcal{K}^{\mathrm{gain}}_{\mathrm{free}}(G_{\mathrm{val}}) &= \frac{22275}{8192}G_{\mathrm{val}}^4 + O(G_{\mathrm{val}}^5). \end{aligned} \tag{105}$$

A positive inverse of $\mathcal{K}(G_{\mathrm{val}}) = c_2^{\mathrm{cost}}G_{\mathrm{val}}^2 + c_4^{\mathrm{cost}}G_{\mathrm{val}}^4 + O(G_{\mathrm{val}}^5)$ satisfies

$$G_{\mathrm{val}}(\tau) = (c_2^{\mathrm{cost}})^{-1/2}\tau^{1/2} - \frac{c_4^{\mathrm{cost}}}{2(c_2^{\mathrm{cost}})^{5/2}}\tau^{3/2} + O(\tau^2).$$

With $c_2^{\mathrm{cost}} = 9/4$, the coefficient difference is $(22275/8192)/[2(9/4)^{5/2}] = 275/1536$, proving (101). □

Strictness throughout the unsaturated KL interval. For $\tau \in (0, \tau_{\mathrm{sat}})$, let $\lambda > 0$ be the free tilt with actual cost τ. Every shared endpoint has $p_3 = 1/3$, whereas $\bar{q}_{\lambda,3} = 3/(2\cosh\lambda + 7) < 1/3$. Applying KL data processing to the event consisting of this third candidate in (94) gives

$$B^{\mathrm{KL}}_{\mathrm{free}}(\tau) - B^{\mathrm{KL}}_{S_{\mathrm{ICM},1}}(\tau) \geq \frac{d_{\mathrm{Bern}}\big(1/3 \,\|\, 3/(2\cosh\lambda + 7)\big)}{\lambda} > 0. \tag{106}$$

Equation (106) gives a directly evaluable nonasymptotic lower bound, complementing the sharp leading coefficient in (101). The separately perturbable zero-valued candidate is not merged with the fixed reset atom.

Actual loss and the scope of the contrast. For target $y = -1$, the squared-loss decrease is $G - G^2/2$, strictly increasing on $0 \le G \le 1/3$; negative controls have negative gain. Thus the same shared endpoints optimize actual loss in this example, by the scalar monotonicity of $G - G^2/2$ on this interval. At small KL budgets both gains start at $(2/3)\sqrt{\tau}$, and their ratio is

$$\frac{B^{\mathrm{KL}}_{S_{\mathrm{ICM},1}}(\tau)}{B^{\mathrm{KL}}_{\mathrm{free}}(\tau)} = 1 - \frac{275}{1024}\tau + O(\tau^{3/2}).$$

The family misses a prescribed coarsened entropic curve yet loses no optimal response under the candidate-operator budget. KL instead charges a higher-order deficit; Lemma G.3 connects it to target mismatch at matched cost. For this fixed-$T = 3$ family, the candidate-operator and KL budgets yield distinct task-response conclusions. Together with Appendix G.5.1, this separates feature requirements for exact target realization from those for optimizing the specified task.

G.6 Finite costs and certificates on the correct support

For a fixed terminal candidate vector η, let $\mathcal{R}_{S_{\mathrm{ICM}}}(\eta)$ be the positive reset-conditionals ζ^{reset} for which (87) holds. The exact conditional KL decomposition gives

$$\begin{aligned}&\inf_{\xi \in \xi^0 + S_{\mathrm{ICM}}:\, \mathcal{C}(\xi) = \eta} D_{\mathrm{KL}}(p^{(T)}_{\mathrm{orig}}(\xi) \| p^{0,(T)}_{\mathrm{orig}}) \\ &\qquad = D_{\mathrm{KL}}(\bar{p}(\eta) \| \bar{p}^0) + (1 - M) \inf_{\zeta^{\mathrm{reset}} \in \mathcal{R}_{S_{\mathrm{ICM}}}(\eta)} D_{\mathrm{KL}}(\zeta^{\mathrm{reset}} \| \zeta^{\mathrm{reset},0}),\end{aligned} \tag{107}$$

with infimum $+\infty$ for an empty feasible set. Without sharing the last infimum is zero, uniquely at $\zeta^{\mathrm{reset},0}$. Under sharing that conditional reset distribution need not be feasible. Thus an origin-resolved-budget obstruction cannot be transferred to the candidate-only problem merely by data processing, which gives the wrong direction for such an upper bound. For example, if every coarsened-feasible candidate endpoint has a shared lift whose extra term is at most c times its coarsened KL, then its gain is bounded by the origin-resolved optimum at budget $(1 + c)\tau$. This sufficient transfer comparison is uniform when a common constant c applies across the settings being compared.

Use $L = L^{\mathrm{agg}}_T$, with $L^{\mathrm{agg}}_T(\bar{p}(\eta)) = \|Z\eta - y_T\|^2/2$. A realizable coarsened endpoint $\bar{p} = \bar{p}(\mathcal{C}(\xi))$ has the same mean as its origin-resolved lift. Apply Theorem 5.2 on support $(z_1, \dots, z_T, 0)$ with $g = m^0_T - y_T$, $\bar{h} = (Z^\top g, 0)$, $V_h = \mathrm{Var}_{\bar{p}^0}(\bar{h}) > 0$, $\Omega_h = \mathrm{osc}(\bar{h})$, and $\bar{q}_\lambda \propto \bar{p}^0 e^{-\lambda \bar{h}}$. With $\varepsilon_{\mathrm{sc}} = \mathrm{osc}(\log \bar{p} - \log \bar{q}_\lambda)$, $0 < \lambda \le R^{-2}$, $\lambda\Omega_h \le 1$, and $\varepsilon^2_{\mathrm{sc}}/8 \le V_h\lambda^2/(4\mathrm{e})$, one has

$$L(\bar{p}^0) - L(\bar{p}) \ge \frac{V_h\lambda}{4\mathrm{e}}, \quad D_{\mathrm{KL}}(\bar{p} \| \bar{p}^0) \le \frac{(\lambda\Omega_h + \varepsilon_{\mathrm{sc}})^2}{8}, \quad \|\eta - \eta^0\|_1 \le \frac{\lambda\Omega_h + \varepsilon_{\mathrm{sc}}}{2}. \tag{108}$$

The proof applies the categorical identity on the coarsened support, whose KL cost tracks candidate masses and the merged reset mass. In the three-column example of Appendix G.5.1, the score error is zero; the one-column task optima in Appendix G.5.2 instead use their actual costs and (94). Applying (108) across lengths requires uniform control of its displayed constants and score error.

H Regular nonuniform reference allocations and temporal spread

Relation to the uniform-reference laws. For nonuniform block weights, let $\mathsf{T}^{\mathrm{pos}} \sim q^{\mathrm{time}}_{a,n}$ on $\{0, \dots, n-1\}$ and set $W^2_{a,n} = \mathrm{Var}_{q^{\mathrm{time}}_{a,n}}(\mathsf{T}^{\mathrm{pos}})$. Under comparable widths $W_{a,n} \asymp W_n \ge w_* > 0$, uniform standardized exponential tails, and the original global drift and attainment conditions, Theorem H.2 replaces n by W_n in (14). Uniform allocation has $W_n \asymp n$; a fixed-rate geometric allocation has bounded spread. Variance alone is insufficient, and changing the reference allocation is not same-baseline repair (Appendices H.3 and H.4).

H.1 Uniform local bounds with global small-budget localization

Let q^{time} be a probability law on a finite subset of $\mathbb{R}$, write $\mathsf{T}^{\mathrm{pos}}$ for its coordinate, and assume $W^2 = \mathrm{Var}_{q^{\mathrm{time}}}(\mathsf{T}^{\mathrm{pos}}) > 0$. Define $\mathsf{T}^{\mathrm{std}} = (\mathsf{T}^{\mathrm{pos}} - \mathbb{E}_{q^{\mathrm{time}}}\mathsf{T}^{\mathrm{pos}})/W$. Exponential tilting by $\beta\mathsf{T}^{\mathrm{pos}}$ is equivalent to tilting $\mathsf{T}^{\mathrm{std}}$ by $s = \beta W$. For constants $c_{\mathrm{tail}} > 0$, $C_{\mathrm{tail}} < \infty$, impose

$$\mathbb{E}_{q^{\mathrm{time}}} \exp(c_{\mathrm{tail}}|\mathsf{T}^{\mathrm{std}}|) \le C_{\mathrm{tail}}. \tag{109}$$

The same constants will apply to all reference laws being compared. This controls standardized tails and is satisfied by bounded standardized support as well as by uniformly exponentially decaying standardized tails.

Lemma H.1 (Regular exponential tilts)**.** *Under* (109)*, there exist positive constants* $s_0, c, C, \tau_0, \delta_0$*, depending only on* $c_{\mathrm{tail}}, C_{\mathrm{tail}}$*, such that for* $q_\beta^{\mathrm{time}}(j) \propto q^{\mathrm{time}}(j)e^{\beta j}$ *and* $|\beta|W \le s_0$*,*

$$c\beta^2 W^2 \le D_{\mathrm{KL}}(q_\beta^{\mathrm{time}} \| q^{\mathrm{time}}) \le C\beta^2 W^2, \qquad c|\beta|W \le \|q_\beta^{\mathrm{time}} - q^{\mathrm{time}}\|_1 \le C|\beta|W. \tag{110}$$

Moreover, either $D_{\mathrm{KL}}(q_\beta^{\mathrm{time}} \| q^{\mathrm{time}}) \le \tau_0$ *or* $\|q_\beta^{\mathrm{time}} - q^{\mathrm{time}}\|_1 \le \delta_0$ *forces* $|\beta|W \le s_0$*. This implication holds over all real* β*, not just an assumed local interval.*

Proof. Define $\widehat{q}_s^{\mathrm{time}}(j) = q^{\mathrm{time}}(j)\exp\{s\mathsf{T}^{\mathrm{std}}(j) - A(s)\}$, so $q_\beta^{\mathrm{time}} = \widehat{q}_{\beta W}^{\mathrm{time}}$, where $A(s) = \log \mathbb{E}e^{s\mathsf{T}^{\mathrm{std}}}$. Since $\mathbb{E}\mathsf{T}^{\mathrm{std}} = 0$ and $\mathbb{E}\big(\mathsf{T}^{\mathrm{std}}\big)^2 = 1$, $A(0) = A'(0) = 0$ and $A''(0) = 1$. The exponential moment controls $\mathbb{E}|\mathsf{T}^{\mathrm{std}}|^k e^{c_{\mathrm{tail}}|\mathsf{T}^{\mathrm{std}}|/2}$ for every fixed k uniformly. Jensen gives $\mathbb{E}e^{s\mathsf{T}^{\mathrm{std}}} \ge 1$, so the third derivative of A, a third centered tilted moment, is uniformly bounded on $|s| \le c_{\mathrm{tail}}/2$. Choose a uniform $s_0 > 0$ such that $1/2 \le A''(s) \le 2$ there. The identity

$$D_{\mathrm{KL}}(q_\beta^{\mathrm{time}} \| q^{\mathrm{time}}) = sA'(s) - A(s) = \int_0^s s' A''(s')\, ds'$$

gives the quadratic bounds for both signs, and Pinsker gives the upper ℓ_1 bound.

For the lower ℓ_1 bound, $\mathbb{E}(\mathsf{T}^{\mathrm{std}})^4$ is bounded by (109). Hölder's inequality yields $\mathbb{E}|\mathsf{T}^{\mathrm{std}}| \ge (\mathbb{E}(\mathsf{T}^{\mathrm{std}})^4)^{-1/2} > 0$ uniformly. The inequality $|e^y - 1 - y| \le y^2 e^{|y|}/2$, with $y = s\mathsf{T}^{\mathrm{std}} - A(s)$ and $A(s) = O(s^2)$, gives

$$\mathbb{E}\left|e^{s\mathsf{T}^{\mathrm{std}} - A(s)} - 1 - s\mathsf{T}^{\mathrm{std}}\right| \le C_1 s^2.$$

After decreasing s_0, the triangle inequality proves $\mathbb{E}|e^{s\mathsf{T}^{\mathrm{std}} - A(s)} - 1| \ge c_1|s|$.

It remains to rule out remote feasible tilts. The directed KL has derivative $sA''(s)$, so it increases away from zero on each side. For $s > 0$, the likelihood ratio is increasing in $\mathsf{T}^{\mathrm{std}}$; hence half the ℓ_1 distance is the supremum over upper-tail events $\{\mathsf{T}^{\mathrm{std}} > c\}$ of $\widehat{q}_s^{\mathrm{time}}(\mathsf{T}^{\mathrm{std}} > c) - q^{\mathrm{time}}(\mathsf{T}^{\mathrm{std}} > c)$. Each difference is nondecreasing in s because its derivative is $\mathrm{Cov}_{\widehat{q}_s^{\mathrm{time}}}(\mathbf{1}_{\mathsf{T}^{\mathrm{std}} > c}, \mathsf{T}^{\mathrm{std}}) \ge 0$. The lower tails give the corresponding statement for $s < 0$ as $|s|$ increases. Thus both distances outside $[-s_0, s_0]$ are bounded below by their uniformly positive boundary values. Choosing budgets smaller than these values proves global localization. □

These exponential-family bounds make the recurrence-induced transfer uniform over reference laws satisfying (109). Related KL-neighborhood sensitivity expansions appear in Lam (2016).

H.2 Response laws for nonuniform reference allocations

Use the fixed phase laws, candidates, controls, gain G, phase cost $\mathcal{K}_{\mathrm{phase}}$, and drift β_a of Section 4.1, but take baseline $q_{a,n}^{\mathrm{time}} \otimes \nu_a^0$ in place of $\mathcal{U}_n \otimes \nu_a^0$. The law $q_{a,n}^{\mathrm{time}}$ is positive on $\{0, \dots, n-1\}$. Unrolling gives

$$p_{a,n,\theta} = q_{a,n,\beta_a(\theta)}^{\mathrm{time}} \otimes \nu_{a,\theta}, \qquad q_{a,n,\beta}^{\mathrm{time}}(j) \propto q_{a,n}^{\mathrm{time}}(j)e^{\beta j}.$$

Here $j \in \{0, \dots, n-1\}$ is chronological block position, and $k = n - 1 - j$ is its block lag at the terminal observation. Reflection leaves the variance unchanged and reverses the tilt sign. Thus $W_{a,n}$ measures spread in *block* units. Token lag $T - i = (n - 1 - j)L_{\mathrm{blk}} + (L_{\mathrm{blk}} - \ell)$ is measured in token units and also includes the within-block phase.

One fixed-horizon CNM realization uses writes proportional to $q_{a,n}^{\mathrm{time}}(j)\nu_a^0(\ell)f_*^{jL_{\mathrm{blk}}+\ell-1}$; the controls change only the periodic retention increments. References across different horizons need not represent the same running-average function.

Theorem H.2 (Temporal-spread transfer)**.** *Suppose all standardized* $q_{a,n}^{\mathrm{time}}$ *satisfy* (109) *with the same* $c_{\mathrm{tail}}, C_{\mathrm{tail}}$*. Assume their standard deviations* $W_{a,n}$ *satisfy*

$$c_W W_n \le W_{a,n} \le C_W W_n, \qquad W_n \ge w_* > 0$$

with constants independent of a, n. If the globally quantified drift inequality and the attaining-curve assumptions of Theorem 4.2 hold, then the nonuniform-reference optima satisfy, uniformly for sufficiently small positive budgets,

$$B_n^{\mathrm{KL}}(\tau) \asymp \min\{\sqrt{\tau}, \tau^{1/(2r)} W_n^{-1/r}\}, \qquad B_n^{\mathrm{op}}(\delta) \asymp \min\{\delta, (\delta/W_n)^{1/r}\}. \tag{111}$$

Proof. The KL chain rule still splits the phase and copy costs. Marginalization bounds each copy KL by τ/w_a^{seq} and each copy ℓ_1 distance by $\delta/w_a^{\mathrm{seq}}$. Lemma H.1 forces $W_{a,n}|\beta_a| \lesssim \sqrt{\tau}$ or $\lesssim \delta$. The positive lower bound on W_n, and sufficiently small budgets, place all drifts in the physical neighborhood required by the globally quantified inequality. Phase marginalization similarly localizes the phase laws and bounds G by $C\sqrt{\tau}$ or $C\delta$. Combining with $\|\beta\|_\infty \geq cG^r$ proves the two upper bounds.

On the attaining curve choose $\vartheta = c\min\{\sqrt{\tau}, \tau^{1/(2r)} W_n^{-1/r}\}$ for KL. Then $W_{a,n}|\beta_a| \leq CW_n\vartheta^r \leq Cc^r\sqrt{\tau}$, so the local upper cost bound of the lemma applies. The full cost is at most $C(\vartheta^2 + W_n^2\vartheta^{2r}) \leq \tau$ for uniform small c. Its gain is at least $c_0\vartheta$. For the operator budget choose $\vartheta = c\min\{\delta, (\delta/W_n)^{1/r}\}$; the product triangle inequality bounds its cost by $C(\vartheta + W_n\vartheta^r) \leq \delta$. These establish attainment with constants uniform in n and the budgets. □

The theorem changes the reference time law and its metric scale while retaining the global drift and attaining-curve conditions on the nonlinear control family.

H.3 Explicit references and what forgetting changes

For the uniform reference $W_n^2 = (n^2-1)/12$, and standardized support is uniformly bounded. Thus Theorem H.2 recovers Theorem 4.2. It also applies whenever $q_{a,n}^{\mathrm{time}}(j)$ is bounded above and below by fixed positive multiples of $1/n$: the variance is comparable to n^2 and the standardized support is bounded. This allows nonuniform long-span references without changing the length exponent.

For a geometrically decaying *block-lag* law $q_n^{\mathrm{time}}(k) \propto \rho_{\mathrm{geo}}^k$, $0 \leq k < n$, with $\rho_{\mathrm{geo}} \in [\rho_{\mathrm{geo},-}, \rho_{\mathrm{geo},+}] \subset (0,1)$, the variances are bounded above and away from zero uniformly for $n \geq 2$. Indeed $q_n^{\mathrm{time}}(0) \geq 1 - \rho_{\mathrm{geo},+}$ and $q_n^{\mathrm{time}}(1) \geq \rho_{\mathrm{geo},-}(1 - \rho_{\mathrm{geo},+})$ give a positive variance lower bound via the independent-copy formula; geometric moments give an upper bound. The uniformly geometric tail then gives (109) for a sufficiently small common c_{tail}. As $n \to \infty$,

$$W_n^2 \longrightarrow \frac{\rho_{\mathrm{geo}}}{(1-\rho_{\mathrm{geo}})^2}.$$

The chronological index is $j = n - 1 - k$, whose variance is the same. Consequently a fixed-rate geometrically decaying reference has bounded temporal spread and length-independent response rates under these hypotheses. This reference changes the distribution of retained writes. In contrast, $\rho_{\mathrm{geo},n} = e^{-c/n}$ with fixed bounded $c \geq 0$ has reference probabilities comparable to $1/n$ and retains $W_n \asymp n$.

In particular the canonical one-feature sigmoid family has $\beta = 2\log\cosh(2\vartheta) \asymp \vartheta^2$ and phase gain $\asymp \vartheta$. Its sign marginal supplies localization for every copy reference. Theorem H.2 therefore applies with $r = 2$ to these nonuniform baselines; adding a bias still cancels the copy drift and realizes the binary entropic optimum. Bounded temporal spread removes this length growth by changing the reference allocation, whereas the added feature repairs the response at the *same* baseline. Each response comparison is therefore defined relative to its own baseline allocation.

The normalized allocation is determined jointly by the gates and the write strengths. With $f_t = f_*$ and $\iota_i = f_*^{i-1}$, the unnormalized terminal mass is

$$\iota_i \prod_{k=i+1}^{T} f_k = f_*^{T-1},$$

so the allocation is uniform. Fixed gates together with constant write strengths provide one realization of geometrically decaying token-lag allocation. For the untruncated geometric block-lag law $(1-\rho_{\mathrm{geo}})\rho_{\mathrm{geo}}^k$, both the mean $\rho_{\mathrm{geo}}/(1-\rho_{\mathrm{geo}})$ and the standard deviation $\sqrt{\rho_{\mathrm{geo}}}/(1-\rho_{\mathrm{geo}})$ are bounded at a fixed rate; they are different quantities.

Corollary H.3 (A sharp nonuniform canonical crossover)**.** *Use the canonical sigmoid features and phase candidates $(-1, 1)$, but take baseline $q_n^{\mathrm{time}} \otimes \mathcal{U}_2$, where q_n^{time} satisfies the hypotheses of Theorem H.2. Let $W_n^2 = \mathrm{Var}_{q_n^{\mathrm{time}}}(\mathsf{T}^{\mathrm{pos}})$. Then the global one-feature KL optimum has the uniform refinement*

$$B_1(q_n^{\mathrm{time}}, \tau) = \left\{ \frac{4\tau}{1 + \sqrt{1 + 128 W_n^2 \tau}} \right\}^{1/2} \{1 + O(\sqrt{\tau})\}, \qquad B_2(q_n^{\mathrm{time}}, \tau) = \tanh \lambda_\tau. \quad (112)$$

Here B_2 includes the bias feature and $k(\lambda_\tau) = \tau$, as in Theorem C.1. The constants in the remainder depend only on the reference regularity constants and the lower spread bound, not on n or τ.

Proof. The exact cost is $k(s^{\mathrm{sgn}}(\vartheta)) + D_{\mathrm{KL}}(q^{\mathrm{time}}_{n,\beta(\vartheta)} \| q_n^{\mathrm{time}})$, with $s^{\mathrm{sgn}}(\vartheta) = -\vartheta + \log\cosh(2\vartheta)/2$ and $\beta(\vartheta) = 2\log\cosh(2\vartheta)$. For chronological indexing this tilt is positive; using block-lag indexing reverses its sign and leaves the argument unchanged. Negative ϑ has the wrong gain sign. On $\vartheta \geq 0$, both the binary cost and the copy cost increase strictly, while the negative mean increases. Hence the feasible boundary is the global optimum.

The sign cost and Lemma H.1 give $\vartheta = O(\sqrt{\tau})$ and $|\beta| W_n = O(\sqrt{\tau})$ throughout the small-budget feasible set. The uniform third-cumulant bound in the lemma sharpens the copy KL to $W_n^2 \beta^2 \{1 + O(|\beta| W_n)\}/2$. Since $\beta = 4\vartheta^2\{1 + O(\vartheta^2)\}$ and $k(s^{\mathrm{sgn}}(\vartheta)) = \vartheta^2\{1 + O(\vartheta)\}/2$, the full cost is

$$\left(\vartheta^2/2 + 8 W_n^2 \vartheta^4\right)\{1 + O(\sqrt{\tau})\}.$$

The positive quartic has root squared $4\tau/(1 + \sqrt{1 + 128 W_n^2 \tau})$. Multiplicative inversion, as in Appendix C.2, and $-\tanh s^{\mathrm{sgn}}(\vartheta) = \vartheta\{1 + O(\vartheta)\}$ prove the first formula. The bias realizes the sign tilt without changing q_n^{time}; data processing to the sign marginal gives the matching unrestricted-distribution upper bound for the second. □

Finite-budget reference comparison. For block lag $k = 0, \ldots, n-1$ and $T = 2n$, define $q_n(k)$ by normalizing each of

$$1, \qquad 0.65^k, \qquad 0.90^k, \qquad e^{-3k/n}, \qquad 1.2 + 0.45\cos(\sqrt{2}\, k).$$

Here $n \in \{2, 8, 32, 128, 512, 2048\}$ and $\tau \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Let $W_n = \sqrt{\mathrm{Var}_{q_n}(k)}$, measured in block units. Figure 5(a) compares the five reference laws against sequence length at $\tau = 10^{-3}$. Larger open circles and a dashed line keep the Uniform reference identifiable near the Bounded irregular reference. Panels (b)–(f) show all three budgets separately for each reference against $W_n\sqrt{\tau}$, using markers without connecting segments. The 30 conditions in (a) are included among these 90 conditions. Color identifies the reference and marker shape identifies the KL cap. Each numerator–denominator comparison uses the same reference allocation; the reference changes between the five laws. Corollary H.3 explains the scaled comparison with its finite-budget remainder, under Theorem H.2's regularity and control conditions. Appendix M.4 reports the numerical checks.

H.4 Variance alone does not justify the law

Let a reference on lags $\{0, n\}$ assign mass n^{-2} to lag n and $1 - n^{-2}$ to lag zero. Its variance is $1 - n^{-2} \to 1$. Tilt by $\beta = -1$. The rare-lag probability becomes

$$q_{-1}^{\mathrm{time}}(n) = \frac{n^{-2} e^{-n}}{1 - n^{-2} + n^{-2} e^{-n}}.$$

The operator distance is $2(n^{-2} - q_{-1}^{\mathrm{time}}(n)) \to 0$, and

$$D_{\mathrm{KL}}(q_{-1}^{\mathrm{time}} \| q^{\mathrm{time}}) = -n q_{-1}^{\mathrm{time}}(n) - \log(1 - n^{-2} + n^{-2} e^{-n}) \longrightarrow 0.$$

But $|\beta| W \to 1$. Thus no uniform small-budget localization, nor a uniform quadratic comparison in terms of variance alone, holds over all positive finite-support lag laws. The standardized exponential moment diverges in this example, identifying the role of the tail-regularity assumption (109) in Theorem H.2.

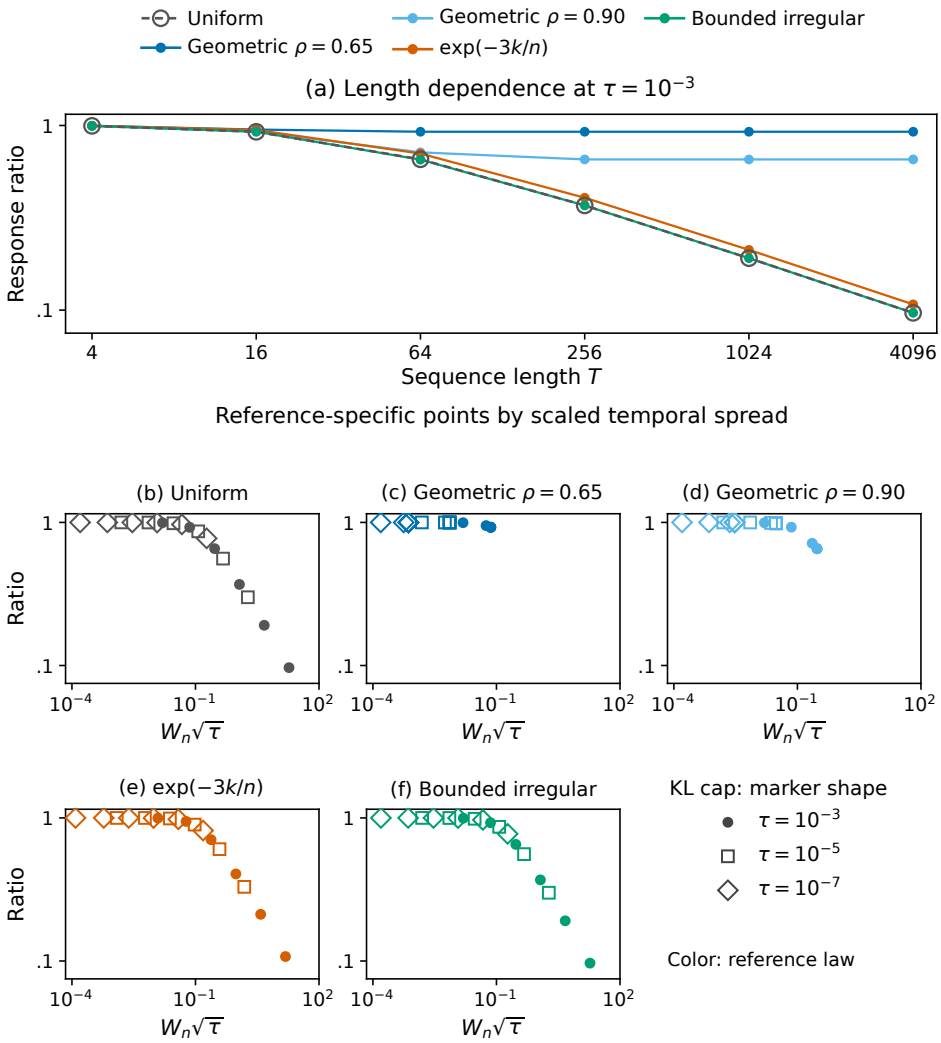


Figure 5: **Reference allocations and scaled temporal spread.** Each ratio compares the canonical one-feature and bias-augmented optima under the same reference allocation. (a) All five references are shown against $T = 2n$ at $\tau = 10^{-3}$. Uniform is distinguished by larger open circles and a dashed line; all other references use smaller filled circles and solid lines. (b)–(f) Each reference has a marker-only panel with three KL caps and six lengths. Color denotes reference allocation, and marker shape denotes KL cap: circles for 10^{-3}, squares for 10^{-5}, and diamonds for 10^{-7}. The 30 conditions in (a) are a subset of the 90 conditions in (b)–(f). Here W_n is the reference block-lag standard deviation. Corollary H.3 gives the finite-budget approximation under Theorem H.2's regularity and control assumptions; Appendix M.4 describes the numerical evaluation.

I JOINT SHARING AND OBSERVATION COMPATIBILITY

I.1 SEVERAL READOUTS WITH JOINTLY SHARED QUERIES

Let $\mathcal{O} \subseteq \{2, \dots, T\}$ contain T, with positive observation weights. Let $E_{\mathrm{key}} \in \mathbb{R}^{T \times d_{\mathrm{key}}}$ be fixed keys, $E_{\mathrm{key},t} = E_{\mathrm{key}}[1:t,:]$, and $\phi_t^Q \in \mathbb{R}^{d_{\phi,Q}}$ fixed query features. Use coordinates (θ, Q) centered at the joint baseline $(0,0)$, where $Q \in \mathbb{R}^{d_{\mathrm{key}} \times d_{\phi,Q}}$ denotes an increment of the shared query-generating matrix. The fixed score vector $s^{0,t}$ includes the baseline retention and query contributions, so $p_{\mathrm{cons}}^{0,(t)} = \mathrm{softmax}(s^{0,t})$. Varying only the retention and query increments gives

$$p_{\mathrm{cons}}^{(t)}(\theta, Q) = \mathrm{softmax}\{s^{0,t} + H_{:t,:}U\theta + E_{\mathrm{key},t}Q\phi_t^Q\}. \tag{113}$$

The target is $q_{\lambda,\mathrm{cons}}^{(t)} = \mathrm{softmax}(s^{0,t} - \lambda Z_t^\top g_t)$, for prescribed g_t. For proper observed prefixes, define and stack

$$\begin{aligned} A_t &= (\phi_t^Q - \phi_T^Q)^\top \otimes (\Delta_t E_{\mathrm{key},t}), & d_t &= \Delta_t Z_t^\top (g_t - g_T), \\ B_T &= \left(\phi_T^Q\right)^\top \otimes \Delta_T E_{\mathrm{key}}, & d_T^{\mathrm{ret}} &= \Delta_T Z^\top g_T. \end{aligned} \tag{114}$$

Here Δ_t takes adjacent differences within a prefix, A, d stack A_t, d_t, and N spans $\ker A$. With $\vartheta_Q = \mathrm{vec}\, Q$ in column-major order, exact score matching modulo one constant per observation is equivalent to the full linear system

$$A\vartheta_Q = -\lambda d, \qquad U\theta + B_T\vartheta_Q = -\lambda d_T^{\mathrm{ret}}. \tag{115}$$

Theorem I.1 (Joint repair and its minimum completion). *For $\lambda \neq 0$, exact realization of all target allocations $q_{\lambda,\mathrm{cons}}^{(t)}$, $t \in \mathcal{O}$, by the family in* (113)*, and zero local allocation-Fisher deficit for the requested score direction, are equivalent to*

$$d \in \mathrm{range}\, A, \qquad -d_T^{\mathrm{ret}} + B_T A^\dagger d \in E_{\mathrm{term}}, \qquad E_{\mathrm{term}} = \mathrm{range}[U, B_T N]. \tag{116}$$

For a linear family of demands, let $D_{\mathcal{G}}^{\mathrm{pre}}, D_{\mathcal{G}}^{\mathrm{term}}$ collect the corresponding d, d_T^{ret}. If $\mathrm{range}\, D_{\mathcal{G}}^{\mathrm{pre}} \not\subseteq \mathrm{range}\, A$, no retention augmentation suffices. Otherwise put $D_{\mathcal{G}}^{\mathrm{rem}} = -D_{\mathcal{G}}^{\mathrm{term}} + B_T A^\dagger D_{\mathcal{G}}^{\mathrm{pre}}$. The minimum number of arbitrary retention columns to add is

$$k_{\min}^{\mathrm{joint}} = \dim(E_{\mathrm{term}} + \mathrm{range}\, D_{\mathcal{G}}^{\mathrm{rem}}) - \dim E_{\mathrm{term}}. \tag{117}$$

Equality of positive softmax outputs is equality of scores modulo constants. Adjacent differences remove those constants; subtracting the terminal equation removes retention from earlier discrepancies. Query solutions differ by $N\theta_{\ker}$, explaining the terminal freedom $B_T N$. This is the part that a completion against U alone would miss.

The criterion is affine feasibility in a recurrent score quotient. Requiring $f_j < 1$ adds $b^0 + U\theta < 0$; the domain-constrained problem ranges over all solutions of (115). The following sections give metric-correct completion and the nonlinear all-lifts condition.

I.2 PROOF OF THE JOINT CRITERION

All readouts here have positive weight, and T is the last readout. The prefix convention $H_{:t,:}$ includes later retention columns as constants; they do not affect the softmax. It satisfies $\Delta_t H_{:t,:} = [I_{t-1}, 0]$.

Write $\vartheta_Q = \mathrm{vec}\, Q$ with column-major vectorization. Equality of positive allocations at time t is equivalent to a scalar s_t^{off} such that

$$H_{:t,:}U\theta + E_{\mathrm{key},t}Q\phi_t^Q = -\lambda Z_t^\top g_t + s_t^{\mathrm{off}}\mathbf{1}.$$

Taking adjacent differences at T gives $U\theta + B_T\vartheta_Q = -\lambda d_T^{\mathrm{ret}}$. Subtracting its first $t-1$ rows from the corresponding earlier equation gives $A_t\vartheta_Q = -\lambda d_t$. Thus the complete equivalent linear system is precisely (115). The first equation has a solution exactly when $d \in \mathrm{range}\, A$. In that case every solution is $\vartheta_Q = -\lambda A^\dagger d + N\theta_{\ker}$. Substituting in the second equation proves (116). A different generalized right inverse changes $B_T A^\dagger d$ only by an element of $\mathrm{range}\, B_T N$, so the quotient condition does not depend on that choice.

For a linear family of demands, the first equation fails for some demand unless $\operatorname{range} D^{\mathrm{pre}}_{\mathcal{G}} \subseteq \operatorname{range} A$, regardless of retention width. Otherwise adjoining retention columns X^b_{add} changes the terminal space to $E_{\mathrm{term}} + \operatorname{range} X^b_{\mathrm{add}}$. All targets are feasible exactly when this contains $\operatorname{range} D^{\mathrm{rem}}_{\mathcal{G}}$. Dimension counting gives (117), attained by an orthonormal basis of $(I - P_{E_{\mathrm{term}}})D^{\mathrm{rem}}_{\mathcal{G}}$. The count is for arbitrary fixed feature columns evaluated on the stated input collection.

I.3 The local metric is the same score quotient

Let J_{score} stack the original score maps $(\theta, \vartheta_Q) \mapsto H_{:t,:}U\theta + E_{\mathrm{key},t}Q\phi^Q_t$ and let s_{req} stack the requested increments $-Z_t^\top g_t$. Define

$$F^0_{\mathrm{score}} = \operatorname{blockdiag}(w^{\mathrm{obs}}_t C_{p^{0,(t)}_{\mathrm{cons}}}).$$

Its kernel is exactly one constant per observation. The free categorical squared-response coefficient is $s_{\mathrm{req}}^\top F^0_{\mathrm{score}} s_{\mathrm{req}}$; the shared coefficient is $s_{\mathrm{req}}^\top \big(F^0_{\mathrm{score}}\big)^{1/2} P_{\operatorname{range}(\left(F^0_{\mathrm{score}}\right)^{1/2} J_{\mathrm{score}})} \big(F^0_{\mathrm{score}}\big)^{1/2} s_{\mathrm{req}}$. The nonnegative deficit is zero exactly when the requested score belongs to the shared score span modulo constants. This is the same full score system as above at unit tilt. It proves the local equivalence in Theorem I.1, including rank-deficient features.

The eliminated equations test feasibility. The response deficit is computed in the original stacked score space with metric F^0_{score}. Numerical evaluation removes the exact constant subspace before taking the metric square root, so numerical errors in the constant directions do not create additional score directions.

I.4 A metric-correct error–dimension tradeoff

Let $H_{\mathcal{O}} = \operatorname{col}_{t\in\mathcal{O}} H_{:t,:}$ map an arbitrary retention increment to the full stacked scores. Put

$$\begin{aligned}
\mathcal{E}_{\mathrm{score}} &= \operatorname{range}\big((F^0_{\mathrm{score}})^{1/2} J_{\mathrm{score}}\big),\\
\mathcal{E}_{\infty} &= \operatorname{range}\big[(F^0_{\mathrm{score}})^{1/2} J_{\mathrm{score}}, (F^0_{\mathrm{score}})^{1/2} H_{\mathcal{O}}\big],\\
\mathcal{E}_{\mathrm{add}} &= \mathcal{E}_{\infty} \cap \mathcal{E}^{\perp}_{\mathrm{score}}.
\end{aligned}$$

These $\mathcal{E}$ spaces live in the weighted full-score space, distinct from the terminal space E_{term} in (116). Suppose $S^{\mathrm{req}}_{\mathcal{G}}$ maps task coefficients $\mathsf{c}_{\mathrm{task}}$ to requested full scores and specify $\mathbb{E}\mathsf{c}_{\mathrm{task}} = 0$, $\mathbb{E}\mathsf{c}_{\mathrm{task}}\mathsf{c}_{\mathrm{task}}^\top = I$. With $C = \big(F^0_{\mathrm{score}}\big)^{1/2} S^{\mathrm{req}}_{\mathcal{G}}$, define

$$B_{\mathrm{out}} = (I - P_{\mathcal{E}_\infty})C, \qquad B_{\mathrm{add}} = (P_{\mathcal{E}_\infty} - P_{\mathcal{E}_{\mathrm{score}}})C.$$

Proposition I.2 (Average deficit after k retention additions)**.** *Over arbitrary additional retention columns, retaining the query family,*

$$\inf_{k \text{ additions}} \mathbb{E}[\textit{local squared-response deficit}] = \|B_{\mathrm{out}}\|^2_{\mathrm{Frob}} + \sum_{j>k} \mathrm{sv}_j(B_{\mathrm{add}})^2.$$

When $B_{\mathrm{out}} = 0$, zero deficit for all demands requires exactly $\operatorname{rank} B_{\mathrm{add}}$ *additional columns.*

Proof. Any enlarged weighted score space is $\mathcal{E}_{\mathrm{score}}$ plus at most k directions in $\mathcal{E}_{\mathrm{add}}$. Its residual decomposes orthogonally into B_{out} and the unrepaired part of B_{add}. Every direction in $\mathcal{E}_{\mathrm{add}}$ is attainable by a retention addition because

$$\operatorname{range}\{(I - P_{\mathcal{E}_{\mathrm{score}}})\big(F^0_{\mathrm{score}}\big)^{1/2} H_{\mathcal{O}}\} = \mathcal{E}_{\mathrm{add}}.$$

Indeed, projecting a sum of an old-score vector and a retention-score vector onto $\mathcal{E}^{\perp}_{\mathrm{score}}$ leaves only the latter. An inverse on this range lifts any chosen direction. The leading left singular vectors of B_{add} minimize the squared Frobenius residual; the singular-value tail gives both lower bound and attainment. The covariance convention on $\mathsf{c}_{\mathrm{task}}$ converts expected squared norm into Frobenius norm. If $B_{\mathrm{out}} = 0$, the zero-error count is the dimension of the missing quotient, equivalent to the eliminated criterion. □

The singular-value tail gives the minimum average squared-response deficit for the stated task distribution. This standard singular value decomposition (SVD) approximation is applied in the Fisher-weighted recurrent score space. Attainment uses unrestricted additional retention columns; constraints on causal feature generation or coefficient norms define a smaller admissible class.

I.5 Finite gate domains and observation conflicts

If the affine retention endpoint must obey $b \le -\mu\mathbf{1}$ for $\mu > 0$, append $U\theta \le -\mu\mathbf{1} - b^0$ to (115). This is an all-solutions linear feasibility test. For example, with $T = 3$, $E_{\mathrm{key}} = (0, 1, 0)^\top$, $Z^\top = (0, 1/2, 1)^\top$, $U = \mathbf{1}$, common $\phi_t^Q = 1$, $g_t = 1$, and $b^0 = -.11\mathbf{1}$, $\mu = .01$, the terminal equations are $\theta - Q = \lambda/2$ and $\theta + Q = \lambda/2$. Thus $Q = 0$, $\theta = \lambda/2$, and the admissible boundary is $\lambda = .20$. Summing the two gate inequalities separates every larger target. Replacing E_{key} by $(0, 1, 2)^\top$ and $Z^\top$ by $E_{\mathrm{key}}/2$ permits a query-only solution $Q = -\lambda/2$ without moving retention. A failed no-query lift would miss that feasible solution.

Without queries, a conservative scalar path satisfies conditional-prefix consistency. Independent entropic targets at $s < t$ agree under restriction precisely when

$$(z_i - z_j)^\top (g_t - g_s) = 0 \quad (i, j \le s).$$

Joint realization requires this cross-prefix compatibility condition. Conversely, consistent positive targets come from one path by setting $\gamma_k = p^{(T)}_{\mathrm{cons},k} / \sum_{i \le k} p^{(T)}_{\mathrm{cons},i}$. Mean feasibility is treated separately in Appendix I.8, which accounts for the kernel of the mean map.

I.6 Nonlinear joint repair over all first-order lifts

For a fixed baseline logit vector $u^{\mathrm{ret},0}$, let $f^0 = \mathrm{sigm}(u^{\mathrm{ret},0})$ and $A_0 = \mathrm{Diag}(1 - f^0)$. Stack observations in increasing time order and define

$$A_{\mathrm{query}} = \mathrm{col}_{t \in \mathcal{O}}\big((\phi_t^Q)^\top \otimes E_{\mathrm{key},t}\big), \qquad \vartheta_Q = \mathrm{vec}\, Q,$$

using column-major vectorization. Thus $A_{\mathrm{query}}\vartheta_Q = \mathrm{col}_{t \in \mathcal{O}}(E_{\mathrm{key},t} Q \phi_t^Q)$, in the same order as $H_{\mathcal{O}}$ and F^0_{score}. Consider stacked scores

$$s(\theta, \vartheta_Q) = s^0 + H_{\mathcal{O}}\{\ell_{\mathrm{sig}}(u^{\mathrm{ret},0} + X_{\mathrm{ret}}\theta) - \ell_{\mathrm{sig}}(u^{\mathrm{ret},0})\} + A_{\mathrm{query}}\vartheta_Q.$$

Let s_{req} be a prescribed first-order full-score request and F^0_{score} the base score Fisher. Define

$$\widetilde{J}_{\mathrm{score}} = \big(F^0_{\mathrm{score}}\big)^{1/2}[H_{\mathcal{O}} A_0 X_{\mathrm{ret}}, A_{\mathrm{query}}], \qquad E = \mathrm{range}\, \widetilde{J}_{\mathrm{score}}.$$

A first-order lift $(\theta_{[1]}, \vartheta_{Q,[1]})$ satisfies $\widetilde{J}_{\mathrm{score}}(\theta_{[1]}, \vartheta_{Q,[1]}) = \big(F^0_{\mathrm{score}}\big)^{1/2} s_{\mathrm{req}}$. Its second-order contribution before correction is

$$c(\theta_{[1]}) = -\tfrac{1}{2}\big(F^0_{\mathrm{score}}\big)^{1/2} H_{\mathcal{O}}\, \mathrm{Diag}(f^0 \odot (1 - f^0))(X_{\mathrm{ret}}\theta_{[1]})^{\odot 2}.$$

Proposition I.3 (Second-order joint realization)**.** *There is a smooth local parameter curve realizing $s^0 + \lambda s_{\mathrm{req}}$ modulo score constants through second order if and only if some first-order lift $(\theta_{[1]}, \vartheta_{Q,[1]})$ satisfies $(I - P_E)c(\theta_{[1]}) = 0$. Polynomial parameter curves then attain a score residual $O(\lambda^3)$.*

Proof. Write $\theta = \lambda\theta_{[1]} + \lambda^2\theta_{[2]} + O(\lambda^3)$ and $\vartheta_Q = \lambda\vartheta_{Q,[1]} + \lambda^2\vartheta_{Q,[2]} + O(\lambda^3)$. Taylor expansion of ℓ_{sig} gives the first-order equation above and second-order equation $\widetilde{J}_{\mathrm{score}}(\theta_{[2]}, \vartheta_{Q,[2]}) + c(\theta_{[1]}) = 0$. The latter is solvable exactly when $c(\theta_{[1]}) \in E$. A quadratic curve with such coefficients proves sufficiency. Conversely any twice differentiable curve matching through second order gives these equations by differentiation. Vanishing Fisher-weighted residual is exactly vanishing modulo constants because all baseline masses are positive. □

The feasibility condition ranges over all first-order lifts satisfying the tangent equation. A small exact control makes the distinction concrete: take $T = 3$, $X_{\mathrm{ret}} = x = (1, 2)^\top$, $f^0 = 1/2$, and a query score $Hx\,\vartheta_Q$. For target Hx, the tangent equation is $\theta_{[1]}/2 + \vartheta_{Q,[1]} = 1$. Its minimum Euclidean norm solution $(\theta_{[1]}, \vartheta_{Q,[1]}) = (.4, .8)$ has a nonzero residual from $x^{\odot 2}$ outside $\mathrm{span}\, x$. Yet $(\theta_{[1]}, \vartheta_{Q,[1]}) = (0, 1)$ is query-only and realizes the target exactly at every magnitude. The query-only lift therefore realizes the target within the same shared family.

I.7 Observation-dependent output rank

Use the conservative recurrence (2) with $m_1 = z_1$ and independent mixing coefficients $\gamma_k \in (0,1)$ for $k = 2, \dots, T$. Let $\mathcal{O} = \{t_1, \dots, t_{N_{\mathrm{obs}}}\}$, with $1 = t_0 < t_1 < \cdots < t_{N_{\mathrm{obs}}} = T$ and positive weights w_t^{obs}. Let $d_k^{\mathrm{inn}} = z_k - m_{k-1}$, $u_k^{\mathrm{inn}} = \operatorname{logit} \gamma_k$, and $J_{\mathcal{O}} = D_{u^{\mathrm{inn}}}(\sqrt{w_t^{\mathrm{obs}}} m_t)_{t \in \mathcal{O}}$. Then

$$\operatorname{rank} J_{\mathcal{O}} = \sum_{j=1}^{N_{\mathrm{obs}}} \dim \operatorname{span}\{d_k^{\mathrm{inn}} : t_{j-1} < k \le t_j\}. \tag{118}$$

To prove it, put $s_{1\to t}^{\mathrm{keep}} = \prod_{l=2}^{t}(1-\gamma_l) > 0$. For $t \ge k$,

$$\partial_{u_k^{\mathrm{inn}}} m_t = s_{1\to t}^{\mathrm{keep}} \frac{\gamma_k(1-\gamma_k)}{s_{1\to k}^{\mathrm{keep}}} d_k^{\mathrm{inn}},$$

and the derivative is zero for $t < k$. A gate in an observation interval contributes d_k^{inn} tensored with a horizon vector that is zero before that interval and equals $\sqrt{w_t^{\mathrm{obs}}} s_{1\to t}^{\mathrm{keep}}$ afterwards. These horizon vectors are linearly independent by triangularity. An invertible transform on the horizon factor makes the column space a direct sum of the interval-wise spans, proving (118). Terminal-only observation therefore has output-Jacobian rank at most d. Observing every prefix has rank $T-1$ when $d_k^{\mathrm{inn}} \neq 0$ at every $k = 2, \dots, T$. Equation (118) counts attainable local variations of the observed means under the free mixing coordinates. Assessing a specified shared family additionally requires its parameter map and the chosen response metric.

I.8 Mean feasibility is weaker than allocation feasibility

Keep the candidate and anchor values fixed and allow each active per-time branch coefficient to vary freely. The following criterion concerns mean realizability before imposing a shared-feature constraint or a sensitivity budget. Let desired means be $\mu_{\mathrm{req}}^{(t_j)}$ at the observation times above, with $\mu_{\mathrm{req}}^{(t_0)} = z_1$. For the conservative mixing recurrence define

$$\mathcal{H}_j = \operatorname{conv}\{\mu_{\mathrm{req}}^{(t_{j-1})}, z_{t_{j-1}+1}, \dots, z_{t_j}\}.$$

For zero-anchor ICM include zero in this convex hull. All desired means are realizable by strictly positive active gates exactly when $\mu_{\mathrm{req}}^{(t_j)} \in \operatorname{ri} \mathcal{H}_j$ for every interval. For closed nonnegative gates, ordinary hull membership is necessary and sufficient.

Necessity follows by unrolling each interval: its final state is a convex combination of the preceding state and all new atoms, with positive weights for active branches. Strictly positive combinations of all generators fill exactly the relative interior of their finite convex hull. For sufficiency, take positive desired endpoint weights $\eta_{\mathrm{old}}, \eta_1^{\mathrm{new}}, \dots, \eta_l^{\mathrm{new}}$. Define partial totals $S_i^{\mathrm{int}} = \eta_{\mathrm{old}} + \sum_{h \le i} \eta_h^{\mathrm{new}}$ and choose keep $S_{i-1}^{\mathrm{int}}/S_i^{\mathrm{int}}$ and write $\eta_i^{\mathrm{new}}/S_i^{\mathrm{int}}$. The products telescope to those endpoint weights. For an active zero anchor, split its positive weight into positive $r_i^{\mathrm{int,reset}}$ and use $(S_{i-1}^{\mathrm{int}}, \eta_i^{\mathrm{new}}, r_i^{\mathrm{int,reset}})/S_i^{\mathrm{int}}$, now with $S_i^{\mathrm{int}} = \eta_{\mathrm{old}} + \sum_{h \le i}(\eta_h^{\mathrm{new}} + r_h^{\mathrm{int,reset}})$. Intervals compose because the preceding desired mean has been realized. Closed feasibility follows by approximation and compactness of the closed gate domain.

For $L = L_{\mathcal{O}}$ with targets $\mu_{\mathrm{req}}^{(t_j)}$, if $e_j = \|m_{t_j} - \mu_{\mathrm{req}}^{(t_j)}\|$, unrolling an actual interval and replacing its starting state by $\mu_{\mathrm{req}}^{(t_{j-1})}$ changes its endpoint by at most e_{j-1}. Consequently $e_j + e_{j-1} \ge \operatorname{dist}(\mu_{\mathrm{req}}^{(t_j)}, \mathcal{H}_j)$ and

$$L \ge \min_{e \ge 0,\ e_j + e_{j-1} \ge \delta_j} \tfrac{1}{2} \sum_j w_{t_j}^{\mathrm{obs}} e_j^2, \qquad 0 \le \delta_j \le \operatorname{dist}(\mu_{\mathrm{req}}^{(t_j)}, \mathcal{H}_j),\ e_0 = 0.$$

An allocation-consistency gap alone does not imply a mean error floor. For $z = (-1, 1, 0)$ with uniform prefix bases, different tilts at times two and three can give inconsistent target allocations but means $\mu_{\mathrm{req}}^{(2)} = -\tanh(.2)$ and $\mu_{\mathrm{req}}^{(3)} = -2\sinh(.1)/(1 + 2\cosh(.1))$. The gates $\gamma_2 = (\mu_{\mathrm{req}}^{(2)} + 1)/2$ and $\gamma_3 = 1 - \mu_{\mathrm{req}}^{(3)}/\mu_{\mathrm{req}}^{(2)}$ are interior and realize both means. This exact example is why the finite-budget theorem bounds the best *output* at a declared allocation cost, rather than equating a distribution mismatch with an output impossibility.

J COORDINATE WITNESSES AND SELECTED MATRIX-MEMORY ROUTES

J.1 THE BOUNDED TWO-DIMENSIONAL COORDINATE WITNESS

Let cond_2 denote the spectral condition number. The following statement studies Euclidean coordinate scaling for a fixed memory function. For every $T \geq 3$, bounded candidates in $\mathbb{R}^2$ with covariance $I_2/4$ can satisfy

$$\inf_{D_{\mathrm{diag}} \succ 0 \text{ diagonal}} \mathrm{cond}_2(J_b D_{\mathrm{diag}} J_b^\top) = \cot^2 \frac{\pi}{2T}, \tag{119}$$

while the same allocation in independent innovation logits has output condition number below two. We prove the statement and then examine its dependence on the candidate array.

For $T \geq 3$, put $\alpha = \pi/(2T)$, take uniform allocation, and define

$$z_i = 2^{-1/2}\big(\cos(\pi(i-\tfrac{1}{2})/T), \cos(\pi(T-1)(i-\tfrac{1}{2})/T)\big)^\top.$$

Discrete cosine orthogonality gives mean zero and covariance $I_2/4$. In fact the second coordinate is $(-1)^{i+1}\sin(\pi(i-\frac{1}{2})/T)/\sqrt{2}$, so every candidate has norm $1/\sqrt{2}$. Centered prefix sums yield

$$(J_b)_{1j} = \frac{\sin(\pi j/T)}{2\sqrt{2}T\sin\alpha}, \qquad (J_b)_{2j} = \frac{(-1)^{j+1}\sin(\pi j/T)}{2\sqrt{2}T\cos\alpha}.$$

For every positive diagonal D_{diag}, $(J_b D_{\mathrm{diag}} J_b^\top)_{11} = \cot^2\alpha\,(J_b D_{\mathrm{diag}} J_b^\top)_{22}$. Extreme Rayleigh quotients imply the lower bound in (119). With $D_{\mathrm{diag}} = I$, sine orthogonality makes the off-diagonal entry zero and attains the bound. Moreover

$$J_b J_b^\top = \mathrm{Diag}\left(\frac{1}{16T\sin^2\alpha}, \frac{1}{16T\cos^2\alpha}\right).$$

Thus the same fixed direction e_2 has the squared-response bound

$$\frac{e_2^\top J_b D_{\mathrm{diag}} J_b^\top e_2}{\lambda_{\max}(J_b D_{\mathrm{diag}} J_b^\top)} \leq \tan^2\alpha.$$

The ratio of response magnitudes is at most $\tan\alpha$. The direction need not be an eigenvector for nonconstant D_{diag}.

For $2 \leq k \leq T$, let $u_k^{\mathrm{inn}} = \mathrm{logit}\,\gamma_k$ be independent innovation logits, and set $J_{u^{\mathrm{inn}}} := D_{u^{\mathrm{inn}}} m_T$. Writing $S_k = \sum_{i \leq k}\pi_i$, their score at index i is $-\gamma_k$ for $i < k$, $1-\gamma_k$ for $i = k$, and zero for $i > k$. These score columns are mutually orthogonal under π, with Fisher diagonal $d_k = \pi_k S_{k-1}/S_k$. At the uniform allocation, $1/(2T) \leq d_k \leq (T-1)/T^2$. Since innovation coordinates also parametrize the full positive simplex, $J_{u^{\mathrm{inn}}}\,\mathrm{Diag}(d_k^{-1})(J_{u^{\mathrm{inn}}})^\top = \Sigma_\pi$. Consequently

$$I_2/(8T) \preceq J_{u^{\mathrm{inn}}}(J_{u^{\mathrm{inn}}})^\top \preceq (T-1)I_2/(4T^2),$$

proving its condition number is below two. The smallest raw eigenvalues in the two charts are both of order T^{-1}; this coordinate comparison uses the raw output Grams. Column scaling itself is an established problem (Casazza & Chen, 2017).

The sharp witness is not uniformly robust to arbitrary vanishing perturbations. For $T \geq 16$ divisible by four, perturb z_1 by $T^{-1/2}e_2$ and z_T by its negative. Norms remain at most one and covariance tends to $I_2/4$. Denote the perturbed Jacobian by J_b^{pert}. Every prefix Jacobian column receives $T^{-3/2}e_2$. The normalized columns 1 and $T/2$ now form an angle of order $T^{-1/2}$. Their two outer products have smallest eigenvalue of order T^{-1}. In D_{diag}, weight each of these columns by its inverse squared norm and all others by T^{-4} times inverse squared norm. The resulting perturbed output Gram $J_b^{\mathrm{pert}} D_{\mathrm{diag}}(J_b^{\mathrm{pert}})^\top$ has condition number $O(T)$: the two chosen columns bound the minimum from below, and the other $T-3$ normalized contributions have total norm $O(T^{-3})$. This gives an achievable condition-number upper bound for the perturbed candidate array.

J.2 Shared matrix-write controls and response metrics

For comparison with a non-scalar memory, consider a fixed-transition channel-write recurrence

$$S_t = A_t S_{t-1} + k_t (w_t^{\mathrm{write}} \odot v_t)^\top, \qquad w_t^{\mathrm{write}} = \mathrm{sigm}(A_{\mathrm{write}} \phi_t^{\mathrm{in}}),$$

with key-by-value state, fixed inputs, keys, values, transitions, and initial state relative to the chosen parameter group A_{write}. With the other routes fixed as above, this recurrence includes a selected channel-write parameter group of Gated DeltaNet-2 (Hatamizadeh et al., 2026, Section 3.1, Eqs. (10)–(11)). A readout is $o_s = S_s^\top q_s$, followed by the differentiable map $\widehat{y}_s = \mathcal{R}_s^{\mathrm{out}}(o_s)$; y_s remains the fixed target. Independent write-logit perturbations at (t, j) have observed source

$$D\mathcal{R}_s^{\mathrm{out}}(o_s) e_j (q_s^\top A_s \cdots A_{t+1} k_t) v_{tj} w_{tj}^{\mathrm{write}} (1 - w_{tj}^{\mathrm{write}}), \qquad t \le s.$$

Let $\varpi_{a,s} \ge 0$ be the product of the sequence and observation weights for observation s of sequence a. Use the weighted output stack $\widehat{y} = \mathrm{col}_{a,s}(\sqrt{\varpi_{a,s}}\, \widehat{y}_{a,s})$. Multiply each displayed source block by $\sqrt{\varpi_{a,s}}$ before stacking it into J_j^{write}. Its columns index all sequence–write-time pairs, with zero blocks outside the corresponding sequence; X_{in} uses the same column-index order for its rows. If X_{in} stacks the rows $(\phi_t^{\mathrm{in}})^\top$, then $\delta u_{:,j}^{\mathrm{write}} = X_{\mathrm{in}} \delta(A_{\mathrm{write}})_{j,:}^\top$. Use $\theta_{\mathrm{write}} = \mathrm{vec}(A_{\mathrm{write}}^\top)$, with column-major vectorization, so the selected-group Jacobian is

$$J_{\widehat{y} \leftarrow \theta_{\mathrm{write}}} = [J_1^{\mathrm{write}} X_{\mathrm{in}} \ \cdots \ J_{d_v}^{\mathrm{write}} X_{\mathrm{in}}].$$

The raw-weight output Gram and the induced-logit response are, respectively,

$$K_{\mathrm{write}}^{\mathrm{raw}} = \sum_j J_j^{\mathrm{write}} X_{\mathrm{in}} X_{\mathrm{in}}^\top (J_j^{\mathrm{write}})^\top, \qquad K_{\mathrm{write}}^{\mathrm{logit}} = \sum_j J_j^{\mathrm{write}} P_{\mathrm{range}\, X_{\mathrm{in}}} (J_j^{\mathrm{write}})^\top.$$

The latter uses the induced metric $X_{\mathrm{in}}^\top X_{\mathrm{in}}$ on each channel and is bounded above by $\sum_j J_j^{\mathrm{write}} (J_j^{\mathrm{write}})^\top$. For every nonzero row space choose B_j^{orth} with

$$(B_j^{\mathrm{orth}})^\top B_j^{\mathrm{orth}} = I, \qquad \mathrm{range}\, B_j^{\mathrm{orth}} = \mathrm{range}\big((J_j^{\mathrm{write}})^\top\big).$$

A sufficient relative lower bound is

$$\min_{j:\, \mathrm{rank}\, J_j^{\mathrm{write}} > 0} \lambda_{\min}\big((B_j^{\mathrm{orth}})^\top P_{\mathrm{range}\, X_{\mathrm{in}}} B_j^{\mathrm{orth}}\big).$$

Zero row spaces contribute zero and are omitted. The bound depends on the alignment of the feature span with the observed write-Jacobian row spaces.

These identities follow from the selected-group derivative and the projector formula. When the same parameters also affect values, queries, transitions or later features, the total Jacobian includes those chain-rule contributions, and its Gram includes their cross terms. The readout derivative $D\mathcal{R}_s^{\mathrm{out}}$ includes any residual or feedforward wrapper.

J.3 Normalization is an additional observation constraint

Exact root-mean-square (RMS) normalization $N_0(o) = \sqrt{d_v}\, o / \|o\|$ has derivative

$$DN_0(o) = \frac{\sqrt{d_v}}{\|o\|} \left(I - \frac{o o^\top}{\|o\|^2} \right), \qquad o \ne 0.$$

It removes radial state changes. At a fitted rank-one association, scalar writing may change only output scale, while independent channelwise writing can produce a $(d_v - 1)$-dimensional nonradial response if all value entries are nonzero. Thus normalization maps the available state edits to their observed nonradial components.

For a matrix with nonzero rows, coordinate queries plus pair-sum queries on a connected graph of nonparallel row pairs have only a global-scale first-order kernel. Coordinate queries force each invisible row perturbation to be parallel to its row; a pair of independent rows forces the two scale factors to coincide. Connectivity then equates all factors. The conclusion is conditional on availability of the stated coordinate and pair-sum queries. Direction-only observations and scale ambiguities have established precedents (Zhao & Zelazo, 2019).

For $\varepsilon_{\mathrm{RMS}} > 0$, define $N_{\varepsilon_{\mathrm{RMS}}}(o) = o/s_o$ with $s_o^2 = \|o\|^2/d_v + \varepsilon_{\mathrm{RMS}}$. Its derivative is defined for every $o \in \mathbb{R}^{d_v}$:

$$DN_{\varepsilon_{\mathrm{RMS}}}(o) = s_o^{-1} I - \frac{oo^\top}{d_v s_o^3}, \qquad DN_{\varepsilon_{\mathrm{RMS}}}(0) = \varepsilon_{\mathrm{RMS}}^{-1/2} I.$$

For $o \neq 0$, putting $P_o^{\mathrm{rad}} = oo^\top/\|o\|^2$ gives the equivalent decomposition

$$DN_{\varepsilon_{\mathrm{RMS}}}(o) = s_o^{-1}(I - P_o^{\mathrm{rad}}) + \varepsilon_{\mathrm{RMS}} s_o^{-3} P_o^{\mathrm{rad}}.$$

At any fixed nonzero o, the radial singular value tends to zero as $\varepsilon_{\mathrm{RMS}} \downarrow 0$, even though the derivative has full rank for each positive stabilizer. The displayed radial singular value therefore quantifies how a full-rank normalization derivative can still attenuate an observed shared update.

J.4 Conditional perturbation sensitivity of linear memories

For linear memory with input-independent transitions $s_t = A_t(\theta)s_{t-1} + B_t(\theta)z_t$ and fixed readout $o_T = Cs_T$, the conditional candidate coefficient is $J_{o_T \leftarrow z_i}(\theta) = CA_T \cdots A_{i+1}B_i$. Between two controls set $D_i = J_{o_T \leftarrow z_i}(\theta) - J_{o_T \leftarrow z_i}(\theta^0)$. Euclidean duality gives

$$\left\| \delta z \mapsto \sum_i D_i \delta z_i \right\|_{\infty,2\to 2} = \max_{\|u\|_2=1} \sum_i \|D_i^\top u\|_2. \tag{120}$$

To prove it, maximize $\sum_i u^\top D_i \delta z_i$ jointly over the unit output dual vector and independently bounded input vectors. Each input maximum is $\|D_i^\top u\|_2$. Scalar $D_i = (\eta_i - \eta_i^0)I$ recovers (68). Equation (120) defines the conditional sensitivity-change norm for the linear recurrence. For example, a two-coordinate shift register can retain z_t, z_{t-1} in separate coordinates while remaining bounded; the scalar keep/write budget does not describe that operation. Likewise state-dependent gates add derivative paths. In $\mathcal{M}(m,z) = (1-\gamma(m,z))m + \gamma(m,z)z$, $D_m\mathcal{M} = (1-\gamma)I + (z-m)(\nabla_m\gamma)^\top$. Appendices J.2–J.3 apply the corresponding selected-group derivatives to matrix writes and normalized observations; (120) supplies the conditional operator for the general linear recurrence above.

K Learned-feature experiment: methods, primary result and transfer

K.1 Data, teacher and retention-controller fitting

Data and causal gate context. For each sequence draw independent $e_t \sim N(0, I_2)$, define $\widetilde{z}_1 = e_1$ and $\widetilde{z}_t = 0.65\widetilde{z}_{t-1} + 0.65e_t$ for $t \geq 2$, and set $z_t = \widetilde{z}_t / \max\{1, \|\widetilde{z}_t\|\}$. The recurrence is applied to $\widetilde{z}$, before the norm restriction. Independently draw $x_{t,1}^{\mathrm{meta}}, x_{t,2}^{\mathrm{meta}} \sim U[-1,1]$. At gate time $t \geq 2$, use

$$x_t^{\mathrm{ctx}} = \left(z_t^\top, z_{t-1}^\top, \left(t^{-1} \sum_{i=1}^t z_i \right)^\top, x_{t,1}^{\mathrm{meta}}, x_{t,2}^{\mathrm{meta}}, \frac{t-1}{T-1} \right)^\top \in \mathbb{R}^9.$$

The gate receives the known sequence length and the externally computed arithmetic prefix mean $\bar{z}_t = t^{-1}\sum_{i=1}^t z_i$ as context inputs, alongside the other displayed coordinates. The recurrent memory output is m_t. The fixed log-write strengths are $a_t^w = (t-1)\log(0.9)$.

Teacher and student memory functions. Both models use the CNM recurrence with the same candidates and write strengths. Teacher logits are

$$u_t^{\mathrm{teacher}} = \mathrm{logit}(0.9) + 0.12 \tanh\{1.2x_{t,1}^{\mathrm{meta}} - 0.7(z_t)_1 + 0.3(z_{t-1})_2\}.$$

The student computes

$$u_t(\theta_{\mathrm{MLP}}) = \mathrm{logit}(0.9) + A_2 \tanh(A_1 x_t^{\mathrm{ctx}} + b_1) + b_2, \qquad f_t(\theta_{\mathrm{MLP}}) = \mathrm{sigm}\, u_t(\theta_{\mathrm{MLP}}),$$

where $A_1 \in \mathbb{R}^{8\times 9}$, $b_1 \in \mathbb{R}^8$, $A_2 \in \mathbb{R}^{1\times 8}$, $b_2 \in \mathbb{R}$. Here u_t indexes the gate time: $u_t(\theta_{\mathrm{MLP}}) = u_{t-1}^{\mathrm{ret}}$ in the gate-index notation used elsewhere. The fixed output offset is not trainable; the total is $72 + 8 + 8 + 1 = 89$ trainable parameters. Initial weights are independent centered Gaussians with standard deviations $0.2/\sqrt{9}$ and $0.02/\sqrt{8}$ for the two layers; both biases are zero. All reported computations use float64 except the separately identified higher-precision checks.

Teacher-fitting objective and optimization. Generate 256 training sequences of length 64 and an independent 128-sequence validation set. The controller is fitted to the teacher's memory outputs at every time $t = 2, \dots, 64$. For a minibatch $\mathcal{B}$ of 32 sampled training indices, the loss is

$$\widehat{L}_{\mathrm{fit}}(\theta_{\mathrm{MLP}}) = \frac{1}{2|\mathcal{B}|(T-1)} \sum_{a \in \mathcal{B}} \sum_{t=2}^{T} \|m_{a,t}(\theta_{\mathrm{MLP}}) - m_{a,t}^{\mathrm{teacher}}\|^2.$$

Each of 400 updates samples indices with replacement. Adam uses learning rate 0.01, momentum parameters $(0.9, 0.999)$, stabilizer 10^{-8}, zero weight decay and `foreach=False`; gradient norm is clipped to 1. Checkpoints $0, 100, 400$ are fixed in advance rather than selected by validation loss. Table 2 reports the full-set teacher-fitting objective at the three fixed checkpoints. The later common-shift task uses separately fitted retention-logit updates.

Table 2: Teacher-output fitting at the three fixed checkpoints. Entries are means with pointwise 95% seed-bootstrap confidence intervals over the same 20 seeds, in units of 10^{-6}. The full-set objective is one-half of the squared Euclidean teacher–student memory-output error, averaged over $t = 2, \dots, 64$ and the sequences in the indicated split (Appendix K.1). The later common-shift task fits separate retention-logit update coefficients.

Controller updates	Training mean [95% CI]	Validation mean [95% CI]
0	13.5821 [13.1059, 14.0547]	13.3303 [12.8271, 13.8324]
100	0.2541 [0.2361, 0.2746]	0.2749 [0.2503, 0.3034]
400	0.1178 [0.0898, 0.1519]	0.1226 [0.0966, 0.1505]

Randomization units. Use seed clusters $s = 2000, \dots, 2019$. The generator stream offsets are $100000 + s$ for training, $200000 + s$ for validation, $300000 + s$ for initialization, $400000 + s$ for minibatches, $500000 + s$ for calibration, $600000 + s$ for held-out data, $700000 + s$ for random projections and $800000 + s$ for the task direction. The settings are in `configs/r4.json`; the generator and ordered random calls are implemented in `code/brmexp/data.py` and `code/brmexp/experiments.py`. Measurements at multiple checkpoints and lengths within one seed remain one statistical cluster.

K.2 Terminal task and frozen quantities

After freezing a checkpoint, draw a Gaussian task direction for seed s and normalize it to $\|g_s\| = 0.25$. Define $y_{a,T} = m_{a,T}^0 - g_s$ for each calibration and held-out sequence, with one common g_s within the seed. The initial terminal squared loss is therefore $\|g_s\|^2/2 = 0.03125$ on every sequence. These are baseline-dependent common-shift targets for the finite-update task. Checkpoint comparisons redefine the targets from that checkpoint's baseline output.

At each evaluation length $T \in \{64, 256, 1024\}$, fit update coefficients on 16 calibration sequences and evaluate unchanged coefficients and scales on 64 held-out sequences. Calibration here refers to fitting the shared update coefficients on the calibration sequences.

There are three distinct operations. Controller training changes all 89 MLP parameters to reduce error against the teacher outputs at times $2, \dots, 64$. The finite intervention freezes the learned features, candidates and write strengths, and fits 9, 11 or 14 shared coefficients that change the retention logits. Candidate-value perturbation evaluation subsequently fixes both complete gate sequences and the target, and changes only the candidate values.

K.3 Update features and calibration scaling

Let a index a sequence and $j = 1, \dots, T-1$ index the gate at recurrent time $j+1$. Write $d_{a,j}^{\mathrm{adj}} = z_{a,j+1} - z_{a,j}$ and use the baseline retention $f_{a,j+1}^0$. Gate head consists of the eight frozen hidden activations and a constant. First-order adds the two columns

$$\phi_{a,j,k}^{(1)} = \frac{d_{a,j,k}^{\mathrm{adj}}}{1 - f_{a,j+1}^0}, \qquad k = 1, 2.$$

Quadratic augmentation appends the three columns

$$\phi^{(2)}_{a,j,k\ell} = \frac{(2 - \mathbf{1}\{k = \ell\}) f^0_{a,j+1}}{2(1 - f^0_{a,j+1})^2} d^{\mathrm{adj}}_{a,j,k} d^{\mathrm{adj}}_{a,j,\ell}, \qquad 1 \le k \le \ell \le 2.$$

Random appends $\phi^{\mathrm{rand}}_{a,j,r} = \sin((\Omega_s)^\top_{:,r} x^{\mathrm{ctx}}_{a,j+1})$, $r = 1, 2, 3$. The seed-fixed $\Omega_s \in \mathbb{R}^{9\times 3}$ has independent standard Gaussian entries from stream $700000 + s$ (Appendix K.1) and is shared across sequences, splits, lengths and checkpoints. There is no phase offset or pre-sine scaling; the RMS scaling below applies to all columns. Both augmentations retain all First-order columns. Thus the four update widths are $9, 11, 14, 14$. The feature generator uses only the fixed context and baseline gates; task dependence enters through the fitted coefficients.

Stack all calibration gate rows into a raw matrix $\widetilde{X}_A$. For column k compute

$$r_{A,k} = \sqrt{\frac{1}{N_{\mathrm{cal}}(T-1)} \sum_{a,j} \widetilde{X}^2_{A,aj,k}}, \qquad s_{A,k} = \begin{cases} r_{A,k}, & r_{A,k} > 10^{-12}, \\ 1, & r_{A,k} \le 10^{-12}. \end{cases}$$

Use $X_A = \widetilde{X}_A \,\mathrm{Diag}(s_A^{-1})$. Apply these same scales to held-out rows. One coefficient vector is shared across all sequences: $u^{\mathrm{ret},+}_{A,a} = u^{\mathrm{ret},0}_a + X_{A,a}\theta_A$.

The inverse-gate coefficients in Appendix D.3 explain these position-dependent linear and quadratic columns. The first-order columns already supply the common task's required tangent direction; therefore all augmented families have unit local task efficiency by construction.

K.4 Finite update objectives and actual budgets

For each declared update family A, use $p_{A,a}(\theta_A) = \mathrm{softmax}\{a^w_a + H\ell_{\mathrm{sig}}(u^{\mathrm{ret},0}_a + X_{A,a}\theta_A)\}$. The loss-fitting objective is

$$\widehat{L}_A(\theta_A) = \frac{1}{2N_{\mathrm{cal}}} \sum_a \|Z_a p_{A,a}(\theta_A) - y_{a,T}\|^2.$$

The two actual calibration costs are

$$C^{\mathrm{KL}}_A = \frac{1}{N_{\mathrm{cal}}} \sum_a D_{\mathrm{KL}}(p_{A,a} \| p^0_a), \qquad C^{\mathrm{op}}_A = \frac{1}{N_{\mathrm{cal}}} \sum_a \|p_{A,a} - p^0_a\|_1.$$

Constrain either $C^{\mathrm{KL}}_A \le \tau$ or $C^{\mathrm{op}}_A \le \delta$; the radii are not numerically equated. The primary condition is checkpoint 400, $T = 64$, $\tau = 10^{-3}$ and squared-loss fitting. Secondary KL caps are $10^{-4}, 10^{-3}, 10^{-2}$; operator caps are $0.005, 0.02, 0.08$. Feasibility is imposed on the within-seed calibration mean across sequences.

Actual-to-target KL fitting is a separate retained objective, $N^{-1}_{\mathrm{cal}} \sum_a D_{\mathrm{KL}}(p_{A,a}(\theta_A) \| q_{a,\lambda})$. In the source tables and endpoint filenames, `fit` denotes this target-KL objective and `loss` denotes squared-loss fitting; neither denotes a data split. The primary comparison uses `loss`. Every reported loss decrease is the baseline loss minus the evaluated endpoint loss.

For the operator-constrained solver, the existing smooth upper approximation is

$$\widetilde{C}^{\mathrm{op}}_A = \frac{1}{N_{\mathrm{cal}}} \sum_{a,i} \sqrt{(p_{A,a,i} - p^0_{a,i})^2 + \varepsilon^2_{\mathrm{sm}}}, \qquad \varepsilon_{\mathrm{sm}} = \delta\rho_{\mathrm{sm}}/T,$$

with successive $\rho_{\mathrm{sm}} = 0.01, 0.0001$. Its excess over the exact cost is at most $\delta\rho_{\mathrm{sm}}$. All selected endpoints are reevaluated using the exact unsmoothed cost. Ray-based searches reevaluate this actual cost at each trial.

K.5 Constrained solver and independent-control references

Metric coordinates and the starting pool. The declared policy uses sequential least squares programming (SLSQP) (Kraft, 1988) as implemented in SciPy 1.17.0.[1] The update-selection procedure combines SLSQP with Fisher-based coordinates, seeded and inherited starts, and exact-cost

[1] Implementation reference: `https://docs.scipy.org/doc/scipy-1.17.0/reference/optimize.minimize-slsqp.html`.

feasibility correction, as specified below. For a frozen family, let $J_a^{\log p}(\theta)$ have row $\partial_\theta \log p_{a,i}(\theta)$. Stack $A_a = \mathrm{Diag}(\sqrt{p_a^0/N_{\mathrm{cal}}})J_a^{\log p}(0)$ into A. The baseline logit derivative enters $J_a^{\log p}(0)$ through $(1 - f_a^0)X_a$ followed by cumulative summation and centering under p_a^0. If $A = U_A \,\mathrm{Diag}(s_A)V_A^\top$, retain $s_{A,j} > 10\max(\mathrm{shape}\, A)\epsilon_{\mathrm{mach}} s_{A,1}$. Active rank is the number of singular values of the baseline Fisher-score embedding A that pass this cutoff. Write

$$B_A = \begin{cases} \sqrt{2\tau}\, V_{A,+}\,\mathrm{Diag}(s_{A,+}^{-1}), & \mathrm{KL}, \\ \delta\, V_{A,+}\,\mathrm{Diag}(s_{A,+}^{-1}), & \mathrm{operator}, \end{cases} \qquad \theta_A = B_A x.$$

A zero-rank family returns the baseline. The covariance and local response are computed in the same metric. Reported local task efficiency is the common task's *squared* response ratio; its square root is the response ratio at equal Fisher radius.

The feasible pool initially contains zero. Families are solved in the order Head, First-order, Random augmentation, Quadratic augmentation. A larger family receives the smaller family's best endpoint; the two augmentations each receive First-order, not one another. An already solved lower budget supplies a same-family candidate. An inherited logit increment is mapped by least squares with relative cutoff 10^{-13} and accepted only if its relative reconstruction discrepancy is at most 2×10^{-10}. Each inherited candidate is reevaluated using its actual cost in the receiving problem.

For each objective let $d_0 = -B_A^\top \nabla_\theta L_A(0)$, normalized when nonzero. With NumPy's `default_rng`, the four ray directions are d_0 and the normalized $d_0 + 0.2\epsilon_k$, $\epsilon_k \sim N(0, I)$, for $k = 1, 2, 3$. A zero direction is replaced by a fresh Gaussian draw. The stream seed is $900000 + s + 100 j_b$, where j_b is the zero-based budget index within the metric. Starting at unit length, a ray is doubled at most 35 times until its actual cost exceeds half the budget, then halved at most 60 times until it is at most 0.7 times the budget. These exact-cost probes generate starting points for the subsequent constrained solve. Supplied candidates add one start at the best incumbent in metric coordinates using a pseudoinverse cutoff 10^{-14}.

A further start projects the required inverse-logit increment onto the raw calibration feature table. Its tilt magnitude is $\min(\lambda_{\mathrm{ref}}, 0.9\lambda_{\mathrm{dom}})$ when a positive finite domain bound exists. For $v_{a,j} = (z_{a,j+1} - z_{a,j})^\top g_a$, use $b^\star = b^0 + \lambda v$ and $u^\star = b^\star - \log(-\operatorname{expm1} b^\star)$ only when every entry of $b^\star$ is negative. Least squares uses cutoff 10^{-12}, followed by the feasible-endpoint correction below. The domain-based tilt restriction applies only to this starting-point construction; the declared target, requested budget and dataset are unchanged.

Objective, trials and endpoint selection. The loss objective is divided by $\max(G_{\mathrm{ref}}, 10^{-10})$; the target-KL objective is divided by $\max(\mathrm{budget}, 10^{-10})$. Costs are divided by the budget. SLSQP uses analytic gradients, 180 iterations per start and tolerance 10^{-11}. Operator runs use the two smooth upper costs from Appendix K.4, continuing from the first solution to the second. During selection, every start, intermediate feasible incumbent and final candidate is considered regardless of the optimizer's success flag. `code/brmexp/optimize.py`, `code/brmexp/experiments.py` and `configs/r4.json` specify the procedure and ordered random calls. The release retains selected states and compact trial diagnostics, not complete optimizer trajectories. Exact-cost correction and objective comparison determine which candidate is selected.

For each candidate θ, set $\alpha = 1$ and target $c_* = 0.9999998\,\mathrm{budget}$. Reevaluate the exact cost. If it exceeds c_*, update at most 80 times

$$u = \max\{0.1, (c_*/\max(C(\alpha\theta), \mathrm{tiny}))^\gamma\},$$
$$\alpha \leftarrow \alpha \min\{1 - 10^{-10}, u(1 - 10^{-10})\}, \quad \gamma = \tfrac{1}{2}\ (\mathrm{KL}), \quad \gamma = 1\ (\mathrm{operator}).$$

Reevaluate at every step; if no feasible candidate is found, use zero. The returned cost must be finite and no greater than the nominal budget. Select the strictly smallest objective among this corrected pool, retaining the earlier incumbent on a tie. A separate diagnostic pool allows tolerance $\max(5 \times 10^{-14}, 2 \times 10^{-7}\mathrm{budget})$, but never determines a claimed strictly feasible result. No such correction is performed after transfer to held-out inputs.

With scaled objective gradient a and constraint gradient c, the recorded multiplier is $\widehat{\mu} = \max(0, -a^\top c/\max(\|c\|^2, 10^{-300}))$. The saved stationarity and complementary-slackness values are $\|a + \widehat{\mu}c\|_\infty$ and $|\widehat{\mu}(C/\mathrm{budget} - 1)|$ at the unrescaled trial endpoint. These are numerical first-order

diagnostics for the single-constraint problem. The combined selected-status flag also checks whether feasibility shrinkage was required. Appendix M.6 reports optimizer termination and final feasibility correction separately.

The independent-control comparison has a different feasible set. For positive baseline p_a^0, define $\mathcal{D}_a$ to be either the full probability simplex or the same-retention-domain polytope

$$\mathcal{D}_a = \left\{p \geq 0 : \mathbf{1}^\top p = 1,\ \iota_{a,j+1} p_j \leq \iota_{a,j} p_{j+1} \quad (1 \leq j < T)\right\}.$$

The latter is the closure, in allocation space, of the distributions induced by independent retention gates $0 < f_t < 1$, $t = 2, \ldots, T$, at the fixed write strengths. Mixing with the strictly interior baseline approaches boundary endpoints using finite logits. This independent-gate feasible set contains the shared-logit family. The reported independent-control reference values and numerical gaps are for calibration KL at $T = 64$.

For squared loss, the primal value is

$$P_* = \min_{p_a \in \mathcal{D}_a,\, N^{-1}\sum_a D_{\mathrm{KL}}(p_a \| p_a^0) \leq \tau} \frac{1}{2N} \sum_a \|Z_a p_a - y_a\|^2.$$

For $\nu \geq 0$ and residual vectors r_a, a lower dual objective is

$$D(r, \nu) = \frac{1}{N} \sum_a \left[-r_a^\top y_a - \frac{1}{2}\|r_a\|^2 + \min_{p \in \mathcal{D}_a} \{r_a^\top Z_a p + \nu D_{\mathrm{KL}}(p \| p_a^0)\} \right] - \nu\tau.$$

Fenchel's identity $\|m - y\|^2/2 = \sup_r \{r^\top (m - y) - \|r\|^2/2\}$ proves the bound, with equality under the usual finite-dimensional interior feasibility conditions here. For $\nu > 0$, the unconstrained inner solution is proportional to $p_a^0 \exp(-Z_a^\top r_a/\nu)$. For the retention-domain polytope, set q to this unnormalized vector and pool adjacent violating values of $\log(q_i/\iota_i)$ with weights ι_i. The nondecreasing pooled levels v_i give $p_i \propto \iota_i e^{v_i}$. This follows because, on a pooled block, the derivative of $\sum_i \iota_i t \log(\iota_i t/q_i)$ vanishes at a common logarithmic level after the normalization multiplier; it is weighted isotonic pooling of *log ratios*, not probability values.

The implemented loss dual uses the bound-constrained limited-memory Broyden–Fletcher–Goldfarb–Shanno method (L-BFGS-B) (Byrd et al., 1995) with 250 iterations, $\mathrm{ftol} = 10^{-14}$, $\mathrm{gtol} = 2 \times 10^{-10}$ and 40 line-search steps. The positive multiplier is scaled by $\max\{\sqrt{\mathrm{Var}(h)/(2\tau)}, 10^{-5}\}$ and bounded below by 10^{-10} in scaled units. Loss gradients are $(m_a - y_a - r_a)/N$ and mean KL minus τ. The fixed-direction gain reference instead fixes $r_a = g_a$ and optimizes only ν.

A returned primal is mixed with p^0 if needed; the gated mixture initially stays 10^{-8} inside the closed endpoint and a scalar root targets 0.9999998τ with absolute tolerance 10^{-14}. For any saved primal cost P and dual value D, $L_0 - P \leq L_0 - P_* \leq L_0 - D$. Report both endpoints and their numerical gap. The brackets are evaluated in float64 with the finite-precision convention in Appendix M.1. For the primary setting, the mean free and same-domain feasible loss decreases are 0.0061482233 and 0.0061468403.

Reference fractions. The seedwise attained shared loss decrease divided by the same-domain independent feasible loss decrease, averaged over the 20 seeds, is 68.34%, 87.96%, 88.74% and 97.22% for Head, First-order, Random augmentation and Quadratic augmentation. The denominator in each seed is the same-retention-domain independent-gate feasible loss decrease. This reference column is separated from the main local task-efficiency comparison.

K.6 Statistical unit and prospective decision

The primary endpoint is checkpoint 400, $T = 64$, common shift, actual mean KL cap 10^{-3}, loss fitting, calibration. For each seed s, set $D_s = \Delta L_{s,\mathrm{quad}} - \Delta L_{s,\mathrm{rand}}$ after averaging over that seed's 16 calibration sequences. The estimator is $20^{-1} \sum_s D_s$. The 95% percentile bootstrap resamples 20 seed indices with replacement 10,000 times using stream seed 33211. The seed is the resampling unit. Checkpoints, lengths, budgets and held-out measurements within a seed belong to the same cluster; secondary intervals are descriptive and pointwise.

The rule fixed before the confirmation run uses effect margin 3.125×10^{-4} and target interval half-width 1.5625×10^{-4}. These are study-specific design values, respectively 1% and 0.5% of baseline loss 0.03125. The observed paired difference is $5.2138925424 \times 10^{-4}$, with interval $[4.8143406136, 5.6384601749] \times 10^{-4}$ and half-width 4.12060×10^{-5}. The lower endpoint exceeds the margin and precision is met; all 20 paired differences are positive. The analysis uses the fixed 20-seed cohort in full.

The two augmentations have the same width and retained Fisher rank, and both have unit local task efficiency by construction. The estimand is the mean paired loss-decrease difference attained by the two feature families with the fitting and selection procedure in Appendix K.5. `data/r4/all_rows.csv` contains all 17,280 seed/condition-level measurements, including held-out cost overruns and unfavorable contrasts; each row already averages over its sequences. Table 1 reports armwise intervals; the primary interval uses paired seed differences. `scripts/reproduce_tables.py` reconstructs both summaries and the primary seedwise values from these rows.

K.7 Prescribed-path diagnostic

At checkpoint 400 and $T = 64$, the prescribed logit curves are constructed using the following Fisher-score projection procedure. For each seed and arm use $u^{\mathrm{ret}}(\lambda) = u^0 + X(\lambda\theta_1 + \lambda^2\theta_2)$ for degree two and omit the quadratic term for degree one. Here X is the arm's fixed calibration-RMS-scaled feature matrix, shared across its 16 calibration sequences and all gates. Let $a_{a,j} = 1 - f^0_{a,j+1}$ and define required logit coefficients $r_{1,a,j} = (g_s^\top(z_{a,j+1} - z_{a,j}))/a_{a,j}$ and $r_{2,a,j} = f^0_{a,j+1}(g_s^\top(z_{a,j+1} - z_{a,j}))^2/(2a^2_{a,j})$. The fixed target is $q_{a,\lambda} \propto p^0_a \exp(-\lambda Z_a^\top g_s)$. Define $\mathcal{E}(v)_{a,i} = \sqrt{p^0_{a,i}/16}\left([H(a_a \odot v_a)]_i - \sum_k p^0_{a,k}[H(a_a \odot v_a)]_k\right)$ and stack all sequences and writes. The columns of A_{score} are the images of the columns of X under $\mathcal{E}$. The executed calls solve $\theta_k = \arg\min_\theta \|A_{\mathrm{score}}\theta - \mathcal{E}(r_k)\|_2^2$, $k = 1, 2$, using `numpy.linalg.lstsq` with `rcond=1e-12`. This is a Fisher-weighted allocation-score projection. Its relative singular-value cutoff is distinct from the optimizer's active-rank cutoff. The minimum-Euclidean-norm coefficient is returned when needed; null directions have zero embedded score. The first lift is treated as exact when its residual norm is at most $10^{-9}\max(\|\mathcal{E}(r_1)\|_2, 10^{-12})$. The curves are evaluated against their corresponding exponential-tilt targets at

$$\lambda \in \{0.005, 0.010, 0.020\}.$$

The plotted error is $\mathcal{D}_{\mathrm{KL}}(p(\lambda)\|q_\lambda)$, averaged over the calibration sequences within each seed. Figure 6 shows the median and the empirical 10th–90th percentiles of the resulting 20 seed values. For each seed, form $\log_2(e_s(0.01)/e_s(0.005))$ from its sequence-averaged error $e_s(\lambda)$. Report the median of these 20 seedwise orders. The range below is the minimum and maximum of the quadratic seedwise orders. The reported smaller-scale orders are 3.99985 for the first-degree curve, 3.99920 for the random degree-2 curve and 5.99769 for the quadratic degree-2 curve; the quadratic seed range is 5.98045–6.01517. A separate long-double Bregman-integrand calculation checked all 480 primary path evaluations, including values near 10^{-15}, with maximum relative difference 6.78×10^{-10}. All values retain their original numerical precision, including the smallest positive errors.

The prescribed curves assess allocation-realization accuracy; Table 1 assesses finite loss improvement at separately optimized shared updates.

K.8 Held-out coefficients and their actual costs

Held-out evaluation applies the primary coefficients and column scales unchanged to 64 new sequences per seed. The mean paired loss-decrease advantage is 5.2764×10^{-4}, while the Random and Quadratic mean held-out KL costs are approximately 1.2721×10^{-3} and 1.0598×10^{-3}. Their seed-mean costs exceed the nominal calibration cap in $18/20$ and $15/20$ seeds, and only one seed is within the cap for both. Table 3 and Figure 7 show improvement and actual cost together.

Calibration feasibility is also a mean constraint: 154 of the 320 quadratic calibration sequence costs exceed 10^{-3} although each seed mean is feasible. These results distinguish average feasibility from per-sequence feasibility and fixed-coefficient transfer from budget-rematched transfer.

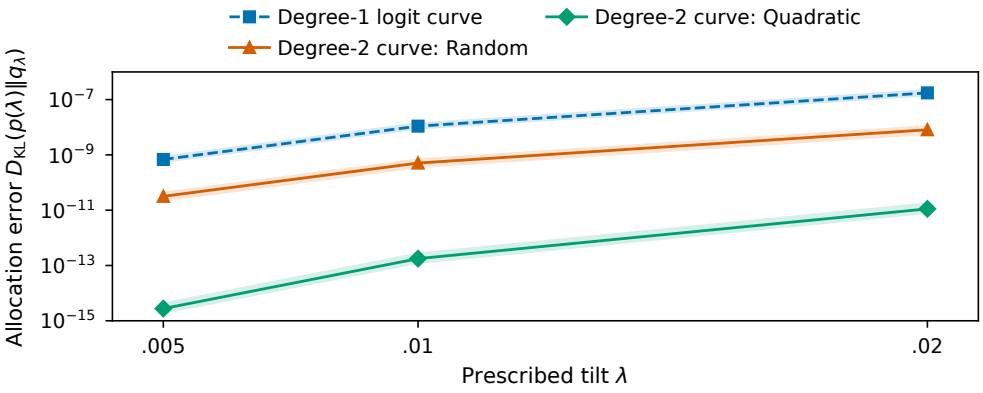


Figure 6: **Prescribed logit curves on calibration sequences.** Checkpoint 400, $T = 64$. Degree-one/two curves are defined by the Fisher-score projections in Appendix K.7, on RMS-scaled fixed features and the same q_λ. Each positive error is averaged over 16 sequences per seed, then shown as the median and 10–90% range across 20 seeds. All three prescribed tilt magnitudes are retained. Table 1 instead reports loss-optimized updates.

Table 3: Fixed-coefficient transfer: improvement and actual allocation KL. The primary calibration coefficients and feature scales are applied unchanged to 64 held-out sequences per seed. A seed is counted over budget when its mean held-out KL exceeds the calibration cap 10^{-3}.

Update family	$10^3 \Delta L_{\text{held-out}}$	$10^3 \overline{D}_{\text{KL,held-out}}$	Over-budget seeds
Gate head	4.279	1.405	19/20
First-order	5.414	1.230	17/20
Random	5.441	1.272	18/20
Quadratic	5.968	1.060	15/20

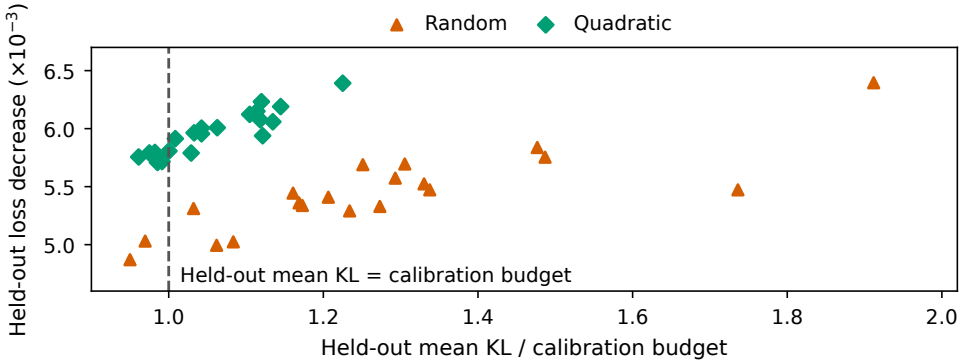


Figure 7: Held-out improvement and actual allocation cost for unchanged fitted updates. Each point is a seed mean over 64 held-out sequences, using the coefficients and feature scales fitted at checkpoint 400, $T = 64$ and mean calibration-KL cap 10^{-3}. The vertical line marks equality of mean held-out KL and that cap. Random and Quadratic exceed it in 18/20 and 15/20 seeds, respectively; one seed is within the cap for both arms. All 20 seeds are shown for each arm, reporting performance and realized cost together.

K.9 Length-specific recalibration and budget sensitivity

At each evaluation length, a new calibration set is used and update coefficients are refitted. Figure 8 groups the results by update family, with three length rows per family and the controller fixed at checkpoint 400.

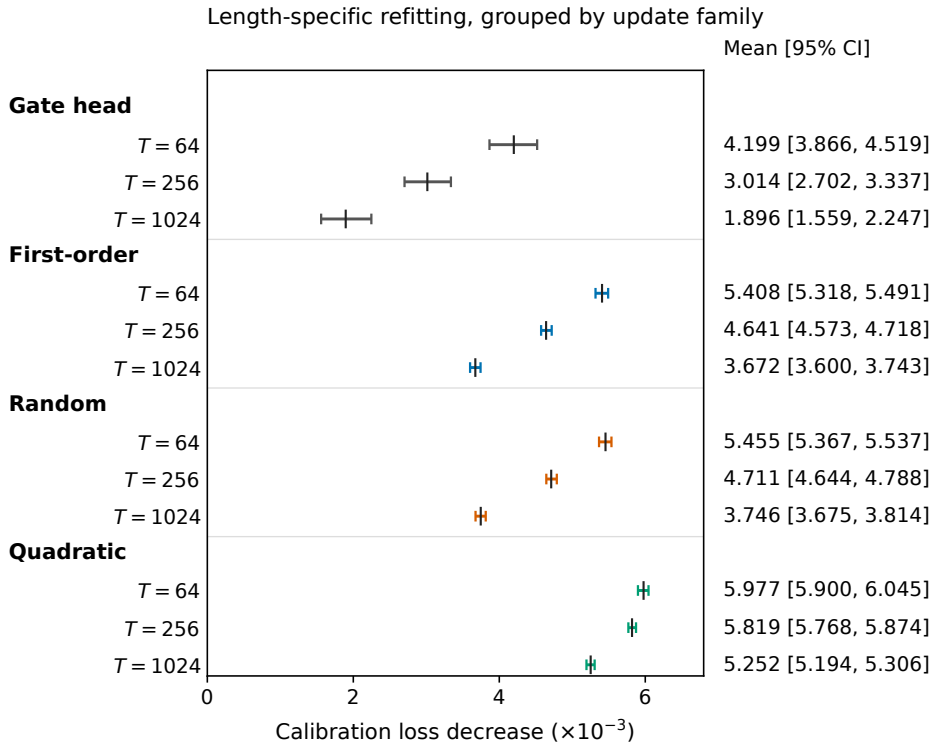


Figure 8: **Finite improvement after length-specific refitting.** Each family block contains $T = 64, 256, 1024$. Black ticks show mean calibration loss decreases over the same 20 seeds; capped horizontal segments give pointwise 95% seed-bootstrap confidence intervals. Printed values use 10^{-3} units, and the plot uses the full precision of the saved values. The controller was trained at length 64 and is fixed at checkpoint 400. Coefficients are fitted to 16 calibration sequences per seed at each evaluation length under a mean allocation-KL cap of 10^{-3}. Appendix K.3 defines the feature families; Figure 2(a) reports the paired primary inference.

Figure 9(a) uses mean allocation-KL caps and panel (b) uses mean candidate-operator caps in separate fitting runs at checkpoint 400 and $T = 64$. Rows distinguish the calibration budget and the Fit (calibration) or Held-out split; here the capitalized plot label Fit does not denote the `fit` objective. Only the Fit row at $\tau = 10^{-3}$ in panel (a) is the prespecified primary comparison. The other intervals are pointwise descriptive seed-bootstrap intervals. Held-out evaluation uses the fitted coefficients and scales without cost rematching. Its mean at the smallest operator cap is negative and its interval includes zero. Appendix K.10 treats the checkpoint comparison separately.

K.10 Checkpoint comparisons

At checkpoints 0, 100 and 400, fixed before teacher fitting, the targets are formed from that checkpoint's baseline outputs and update coefficients are refitted on calibration sequences at each length under mean KL cap 10^{-3}. Figure 10 reports paired loss-decrease contrasts on the calibration sequences. A positive contrast is already present at initialization. Each checkpoint therefore defines its own finite-update task, while Table 2 summarizes fitting to the fixed teacher outputs.

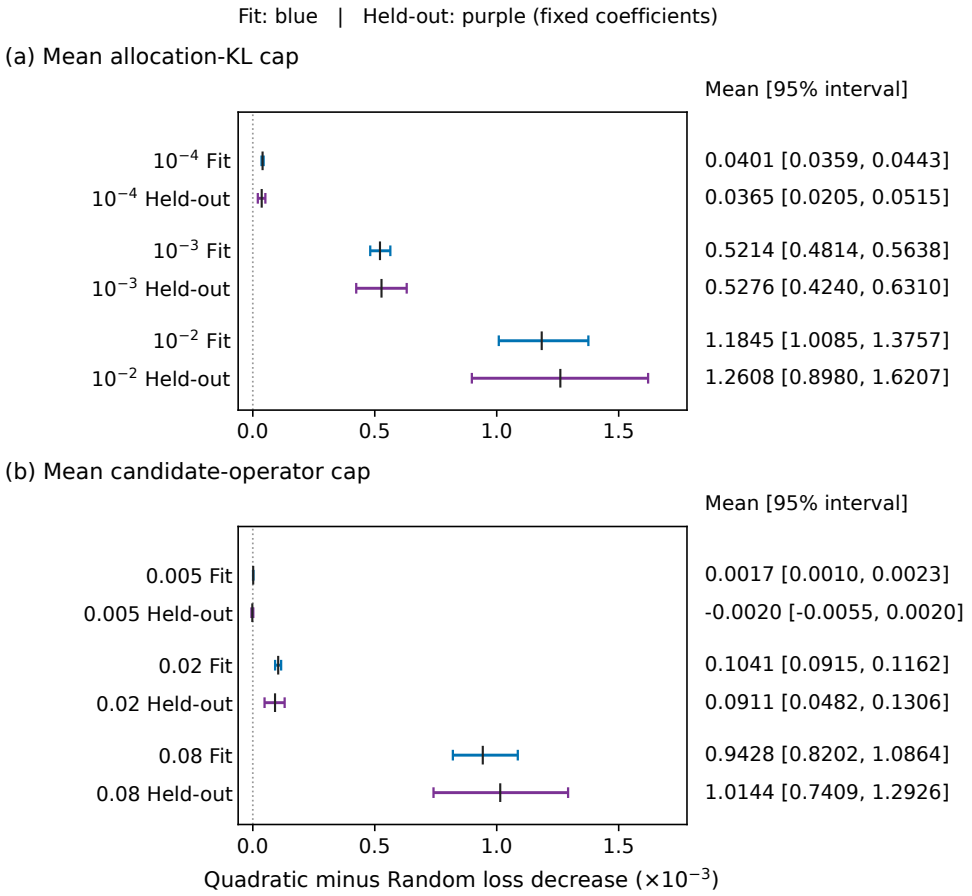


Figure 9: **Budget sensitivity of attained paired improvements.** Checkpoint 400 and $T = 64$. Panel (a) varies the mean allocation-KL cap; panel (b) varies the mean candidate-operator cap in separate fitting runs. Each cap has separate Fit (calibration) and Held-out rows. Black ticks show mean Quadratic-minus-Random loss-decrease differences over the same 20 seeds; capped horizontal segments give pointwise 95% seed-bootstrap confidence intervals. Numerical values use the same 10^{-3} units as the axis. Held-out evaluation reuses the fitted coefficients and scales without cost rematching. The displayed caps constrain the calibration means during fitting. Only the Fit row at $\tau = 10^{-3}$ in (a) is the prespecified primary comparison. At operator cap 0.005, the held-out mean is negative and its interval includes zero.

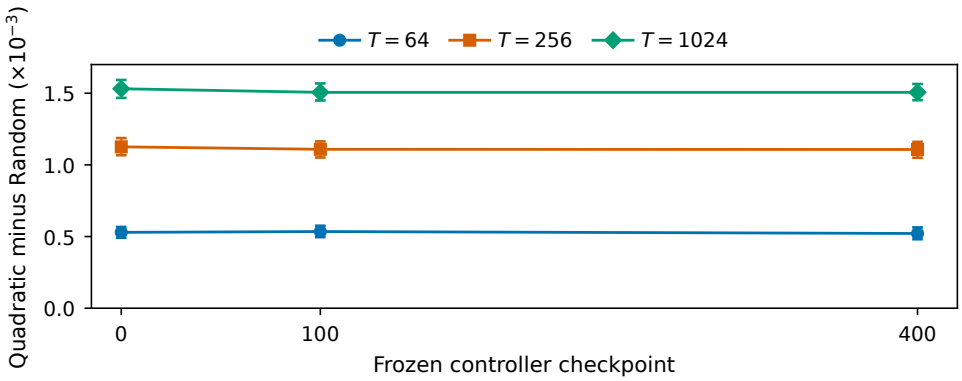


Figure 10: Finite-update contrasts for checkpoint-specific tasks, evaluated on calibration sequences. At checkpoints 0, 100 and 400, common-shift targets are formed from that checkpoint's baseline outputs and update coefficients are refitted for each displayed length under mean calibration-KL cap 10^{-3}. Points show paired Quadratic-minus-Random means over the same 20 seeds; bars are pointwise descriptive 95% seed-bootstrap intervals. A positive contrast is already present at initialization. Appendix K.10 specifies the fitting and evaluation splits.

L Fixed-gate candidate-value perturbations and constructed controls

L.1 A fixed repair over lengths and perturbation radii

For each $f_* \in \{0.2, 0.5, 0.8\}$, amplitude pair in $\{(0.25, 0.75), (0.5, 1), (0.35, 0.9)\}$, and budget, we selected the explicit zero-drift coefficients once at $T = 4$ and reused them at $T = 4, 16, 64, 256, 1024, 4096$. The KL caps are $10^{-6}, 10^{-4}, 10^{-3}$ and the operator caps are $0.001, 0.005, 0.02$. The per-write perturbation radii are $0, \varepsilon_*/2, \varepsilon_*, 2\varepsilon_*$ with $\varepsilon_* = 0.25/16$.

We report the three radii up to the grid-wide bound ε_* separately from $2\varepsilon_*$. For amplitudes $(0.5, 1)$, the latter equals the instance-specific radius bound $\min(c_1, c_2)/16$. This radius comparison concerns the perturbation-radius condition of Corollary 3.2; its other structural and budget conditions remain as stated there. The ratio $\varepsilon_z/(\min(c_1, c_2)/16)$ expresses the radius relative to each instance-specific bound.

The fixed witness's scalar perturbation minimum is evaluated by (121), and the minimizing perturbation is checked with a separate forward recurrence. Appendix C.4 proves the length-uniform lower guarantee.

Exact evaluation on the repaired witness. The scalar reduction below evaluates the explicit zero-drift witness.

For one sequence of amplitude $0 < c \le 1$, let $\upsilon = -m/c \in [0, 1]$ along the zero-drift witness and let $0 \le \varepsilon_z < c$. The perturbation averages in the negative and positive groups, denoted $\bar{e}_-$ and $\bar{e}_+$, each range over $[-\varepsilon_z, \varepsilon_z]$. Write $x = (\bar{e}_- + \bar{e}_+)/2$ and $d_e = (\bar{e}_- - \bar{e}_+)/2$. Then $|x| + |d_e| \le \varepsilon_z$, and the perturbed loss decrease is

$$\upsilon(c - d_e)(1 + x) - \frac{1}{2}\upsilon^2(c - d_e)^2.$$

Since $c - d_e > 0$, its minimum over x is at $x = -\varepsilon_z + |d_e|$. On $d_e \le 0$ the resulting function is decreasing; on $d_e \ge 0$ it is concave. Its minimum therefore occurs at $d_e = 0$ or $d_e = \varepsilon_z$. Thus the exact worst decrease is

$$\min\left\{c\upsilon(1 - \varepsilon_z) - \frac{1}{2}c^2\upsilon^2, (c - \varepsilon_z)\upsilon - \frac{1}{2}(c - \varepsilon_z)^2\upsilon^2\right\}. \tag{121}$$

Both choices are attained by valid individual-write perturbation tables. Because the witness preserves the copy marginal, the full per-write perturbation box reduces to the two sign-group averages. For two sequences, separately selectable perturbation tables give the average of the two exact minima.

Witness evaluation. The evaluation was run on Aurora. For $f_* = 0.5$, amplitudes $(0.5, 1)$ and $\varepsilon_z = 0.015625$, the following values are evaluated at $T = 4$. The coefficients are reused at all six lengths through 4096.

Budget	Nominal decrease	Worst perturbed	$G/4$
KL 10^{-3}	0.03449899	0.03383883	0.00883681
Operator 0.02	0.01647512	0.01616330	0.00416597

All 972 cells at radii up to the grid-wide bound satisfy the evaluated witness inequality, with minimum margin 4.5281×10^{-4}. Across structural configurations, the largest within-configuration change of worst perturbed loss decrease across lengths is 6.01×10^{-14}. The 324 cells at $2\varepsilon_*$ are reported separately.

Figure 1 shows the mean-KL-cap 10^{-3} case at $(c_1, c_2) = (0.5, 1)$, $f_* = 0.5$ and $\varepsilon_z = 0.015625$. It combines the fixed witness evaluated here with the Bias updates selected separately by nominal loss at each length. Table 4 reports the additional Time and length-specific Square comparisons.

L.2 Constructed control families and selected endpoints

Feature controls at fixed two-amplitude baselines. For $c_1 = 0.5, c_2 = 1$, write $v_{a,j} = z_{a,j+1} - z_{a,j} \in \{\pm 1, \pm 2\}$. Each family contains $v/(1 - f_*)$; Bias additionally contains $\mathbf{1}$. Square, Fourth and Cube append v^2, v^4, v^3, respectively. Time nuisance appends $\sin(\sqrt{2}\, j_0)$, $j_0 = 0, \ldots, T - 2$, broadcast identically across the two amplitudes. Duplicate appends another copy of $v/(1 - f_*)$. The constructed features use their original scales. Duplicate has the same span and rank as Bias, while the Time comparison measures the attained improvement from an additional time-dependent direction.

Here $v^4 = 5v^2 - 4$, so Square and Fourth have identical logit spans after Bias is included. For general two amplitudes, $v^4 = 4(c_1^2 + c_2^2)v^2 - 16c_1^2c_2^2$. Feasible coefficients are transferred between the two bases in both directions, and each larger family receives the Bias incumbent. An exact-product construction supplies the Square witness. Cube also cancels drift and supplies a distinct useful repair. The global upper bound for the unaugmented biased family is analytic; Figure 11(a) compares the additional directions supplied by Time, Square and Cube against Bias. Panel (b) reports the signed saved-value differences Duplicate minus Bias and Fourth minus Square at all five lengths. These comparisons use the same Bias and Square values as panel (a), in the original nominal-loss units. The span identities and feasible-coefficient transfers described above define the two representation checks.

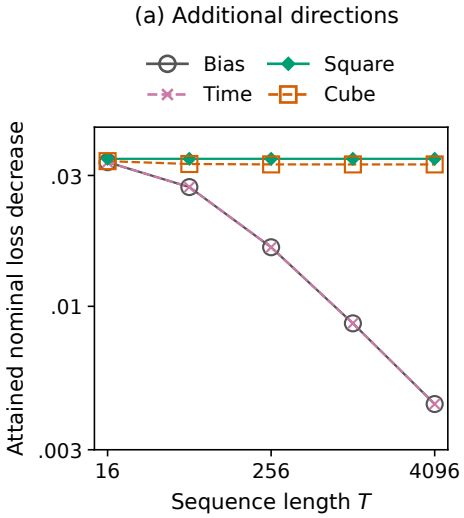
For the Square witness, put $x_a = 2c_a\vartheta/(1 - f_*)$, $C_* = f_*^{-2} - 1$, $Q_0 = (1 - f_*)/f_*$. The block-product equation gives

$$Q_a = \frac{C_*}{\cosh x_a + \sqrt{\cosh^2 x_a + C_*}}, \qquad s_a = -\log(Q_a/Q_0),$$
$$\theta_{\mathrm{sq}} = \frac{s_2 - s_1}{4(c_2^2 - c_1^2)}, \qquad \theta_{\mathrm{bias}} = s_1 - 4c_1^2\theta_{\mathrm{sq}}.$$

A scalar root chooses the control using the *actual* cost. Cube instead solves the two-amplitude linear system for the contrast supplied by the columns $2c_a/(1 - f_*)$ and $8c_a^3$, followed by the common product-preserving bias. The retained constructed finite-budget comparison uses $f_* \in \{0.2, 0.5, 0.8\}$, $T \in \{16, 64, 256, 1024, 4096\}$ and either a KL budget of 10^{-3} or an operator budget of 0.02. The analytic construction targets a cost of 0.9999998 times the corresponding budget; this choice is distinct from the candidate-value perturbation witness below. All 180 selected endpoints satisfy the nominal cap. After exact basis transfer, the 30 Square/Fourth loss differences have maximum approximately 2.5×10^{-12}.

Fixed-task perturbation controls. The perturbation experiment uses the same Time and Duplicate functions. In its controls, SLSQP has 180 iterations, four ray starts and stream seed 20260917;

each larger family inherits Bias, and Square receives the analytic witness. Coefficients are selected by nominal loss before candidate values are perturbed. Their scalar worst-perturbation values are evaluated after selection. The explicit witness used for length transfer is selected once at $T = 4$ with target $(1 - 10^{-6})$ times the budget; a bracket starts at 0.001 and is expanded to at most 1.6 with up to 30 attempts, then Brent's method uses absolute tolerance 10^{-15} and relative tolerance 10^{-13}. The same coefficients are used for every subsequent length. Appendix M.6 reports the run coverage and



(b) Representation check

T	Duplicate minus Bias	Fourth minus Square
16	8.65×10^{-14}	0
64	-1.11×10^{-16}	0
256	0	0
1024	2.66×10^{-15}	0
4096	1.44×10^{-15}	0

Figure 11: **Constructed directions and representation checks.** Amplitudes $(0.5, 1)$, $f_* = 0.5$, and mean allocation-KL cap 10^{-3}. (a) Bias, Time, Square, and Cube show attained nominal squared-loss decreases at selected feasible updates for the five lengths. Cube is a distinct useful repair. (b) The table gives signed differences of saved nominal decreases, Duplicate minus Bias and Fourth minus Square, in the original loss units. Bias and Square use the same values as (a). Duplicate repeats the difference column, and $v^4 = 5v^2 - 4$ on this support; Appendix L.2 describes the feasible coefficient transfers. Printed zero refers to the stored floating-point evaluations. Table 4 reports the separate perturbation-minimum evaluation.

result-integrity checks for the fixed-gate evaluations. The results below concern the fixed witness and the separately selected constructed controls.

The same repair across lengths. The zero-drift construction in Appendix C.3.1 is evaluated in Appendix L.1 using coefficients selected once at $T = 4$ and reused across lengths. The grid covers both budgets. The largest probability discrepancy against the independently implemented normalized recurrence is 7.72×10^{-14}, and the largest scalar perturbed-loss discrepancy is 6.53×10^{-13}. Corollary 3.2 and its proof in Appendix C.4 provide the length-uniform guarantee.

Separately selected nominal endpoints. All 60 selected family/configuration endpoints satisfy the actual nominal budget. Of the 510 solver trials, 493 terminate successfully and 17 have other statuses, including iteration limits. Endpoint selection uses the feasible candidate pool, which includes evaluated iterates and inherited witnesses. The aggregate status also records final feasibility rescaling. Square receives the explicit zero-drift witness as a starting candidate. Appendix M.6 links these constructed trials to their selection records.

At mean KL cap 10^{-3}, the selected Square updates in Table 4 have worst perturbed decreases of 0.033839 at every displayed length. The separately constructed fixed witness in Appendix L.1 has value 0.03383883, agreeing at the displayed precision. The Bias values decrease from 0.032871 at $T = 16$ to 0.003799 at $T = 4096$, and the Time values are close to Bias. The two Square selection policies remain distinct: at operator cap 0.02, the fixed witness has value 0.01616330, while the selected Square update at $T = 4096$ has value 0.016111. Selection uses nominal loss; the table subsequently evaluates each selected update under perturbations.

Table 4: Constructed controls at $\varepsilon_z = 0.015625$, $f_* = 0.5$, amplitudes $(0.5, 1)$. Entries are the exact scalar perturbation minimum evaluated at the selected nominal-loss endpoint. The inner perturbation minimum uses the analytic reduction in Appendix L.4; each outer update is selected from the feasible pool by nominal loss. The other two budget values and all radii remain in the source tables.

Budget	T	Bias	Time feature	Square
KL 10^{-3}	16	0.032871	0.032871	0.033839
KL 10^{-3}	64	0.026545	0.026545	0.033839
KL 10^{-3}	256	0.015798	0.015798	0.033839
KL 10^{-3}	1024	0.008063	0.008063	0.033839
KL 10^{-3}	4096	0.003799	0.003799	0.033839
Op. 0.02	16	0.016139	0.016139	0.016378
Op. 0.02	64	0.016064	0.016064	0.016172
Op. 0.02	256	0.011289	0.011289	0.016125
Op. 0.02	1024	0.005898	0.005898	0.016114
Op. 0.02	4096	0.002850	0.002850	0.016111

L.3 Estimands, per-endpoint bounds and paired contrasts

Fix a split, a seed s and its recorded sequence set $\mathcal{B}_s$. For $A \in \{\mathrm{head}, \mathrm{first}, \mathrm{rand}, \mathrm{quad}\}$, labeling Gate head, First-order, Random and Quadratic, define

$$\mathcal{D}^{\mathrm{worst}}_{A,s}(\varepsilon_z) = \frac{1}{|\mathcal{B}_s|} \sum_{a \in \mathcal{B}_s} \inf_{\max_i \|\delta z_{a,i}\| \le \varepsilon_z} \mathcal{D}^{\mathrm{pert}}_{A,s,a}(\delta z_a).$$

Each arm and sequence has a separately chosen perturbation table. For each arm and sequence, the same perturbed candidate values are evaluated under the pre-update and post-update gates, with both gate sequences, the write strengths and the target held fixed. Let $\underline{\mathcal{D}}_{A,s}$ average the analytic lower bounds and $\overline{\mathcal{D}}_{A,s}$ average the values at the saved feasible attacks. Since $\underline{\mathcal{D}}_{A,s} \le \mathcal{D}^{\mathrm{worst}}_{A,s} \le \overline{\mathcal{D}}_{A,s}$,

$$\underline{\mathcal{D}}_{\mathrm{quad},s} - \overline{\mathcal{D}}_{\mathrm{rand},s} \le \mathcal{D}^{\mathrm{worst}}_{\mathrm{quad},s} - \mathcal{D}^{\mathrm{worst}}_{\mathrm{rand},s} \le \overline{\mathcal{D}}_{\mathrm{quad},s} - \underline{\mathcal{D}}_{\mathrm{rand},s}. \tag{122}$$

For the calibration cohort, define

$$\mathcal{C}_{\mathrm{pert}}(\varepsilon_z) = \frac{1}{20} \sum_s \left[\mathcal{D}^{\mathrm{worst}}_{\mathrm{quad},s}(\varepsilon_z) - \mathcal{D}^{\mathrm{worst}}_{\mathrm{rand},s}(\varepsilon_z) \right].$$

Averaging (122) gives its lower endpoint $20^{-1} \sum_s (\underline{\mathcal{D}}_{\mathrm{quad},s} - \overline{\mathcal{D}}_{\mathrm{rand},s})$ and upper endpoint $20^{-1} \sum_s (\overline{\mathcal{D}}_{\mathrm{quad},s} - \underline{\mathcal{D}}_{\mathrm{rand},s})$. Figure 2(b) displays these endpoints at all five radii. They evaluate analytic inequalities and feasible attacks at the saved inputs and updates; Appendix M.1 specifies their arithmetic precision and distinguishes them from seed-bootstrap intervals.

At radius zero the conditional enclosure reduces to the cohort's empirical paired mean. Its seed-bootstrap interval in Figure 2(a) measures a different quantity: variation under seed resampling. At $\varepsilon_z = 0.02$, the mean lower endpoint is $2.9609 \times 10^{-4} > 0$. The primary nominal effect margin is 3.125×10^{-4}, so this enclosure establishes positive worst-value ordering while leaving achievement of that margin under perturbations unresolved. The smallest seedwise positive lower endpoint at this radius is 1.7258×10^{-4}, compared with diagnostic tolerance 10^{-10}. All arms, sequences and radii are retained in the averages.

L.4 Random perturbations and the finite attack procedure

Random intervention. For each fixed sequence draw 128 perturbation tables. Each per-write direction is a normalized standard Gaussian in $\mathbb{R}^d$, with radius $\varepsilon_z U^{1/d}$ for $U \sim U[0,1]$. Draws are independent across writes within this stochastic stress test, and the entire table is reused across update families. Calibration and held-out streams have offsets $920000 + s$ and $1020000 + s$. This stress test samples the product of uniform-ball distributions. The adversarial evaluation instead takes an infimum over the full product of per-write balls.

Finite projected attack. Use six starts and 60 steps. Start zero is exactly the zero perturbation; the other starts use uniform-ball tables with stream offsets $930000 + s$ (calibration) and $1030000 + s$ (held-out). Each arm is optimized separately from common initial tables. At iteration $k = 0, \ldots, 60$, evaluate the loss decrease and retain the lowest value so far, separately for each sequence and over all starts. For fixed before/after probabilities, the gradient with respect to write i is

$$p_i^0(m^0(\delta z) - y) - p_i(m^+(\delta z) - y).$$

When $k < 60$, normalize each write's gradient by $\max(\|\nabla_i \mathcal{D}\|, 10^{-30})$, take a descent step of size $0.2\varepsilon_z/(1 + k/20)$, and project each perturbation onto its radius-ε_z Euclidean ball. The perturbed candidates are not clipped back into the original unit ball. At radius zero the perturbation is identically zero. The complete gates, target and update coefficients are fixed throughout; neither gate generation nor endpoint fitting is rerun.

The best feasible attack gives an upper value for the true minimum loss decrease, and (30) gives a lower bound. Including the zero start keeps the attack value at most the nominal improvement. Equation (122) combines these values using a separate infimum for each arm and sequence. All finite outcomes are included in the reported results.

For the scalar constructed models, the full perturbation box maps to the zonotope of $((p^0)^\top \delta z, p^\top \delta z)$. The objective is a quadratic on this two-dimensional set. Its interior stationary point, when present, cannot supply a strict minimum because the Hessian has one nonpositive direction; minimizing the restricted quadratic on every boundary segment therefore suffices, including degenerate one-dimensional cases. Vertices, stationary points lying on each segment, and endpoints are evaluated. The scalar minimum is obtained by numerically evaluating this exact analytic reduction. For learned updates, the projected attacks provide feasible upper values and (30) provides lower bounds, with the fixed-gate evaluation defined in Appendix L.3.

L.5 Results for selected updates across radii and splits

Endpoint-level results. Table 5 reports Random and Quadratic at each perturbation radius and split. Values are averaged within seeds and then over the same 20 seed clusters. Lower and Upper use the analytic inequality (30) and the feasible-attack procedure in Appendices L.3–L.4. The evaluation uses ambient per-write Euclidean balls with both gate sequences fixed; Appendix M.1 states the finite-precision convention.

At $\varepsilon_z = 0.02$, the mean calibration attack contrast is 4.7918×10^{-4} with descriptive paired seed-bootstrap interval $[4.4330, 5.1894] \times 10^{-4}$; every seed contrast is positive. The corresponding held-out contrast is 4.9505×10^{-4} with interval $[3.9844, 5.8934] \times 10^{-4}$. These intervals summarize the paired contrasts between attained attack values.

Mean response to random perturbations. The mean improvement under the sampled perturbations remains close to the nominal improvement. For independent uniform-ball perturbations in d dimensions, mean zero and covariance $\varepsilon_z^2 I/(d + 2)$ give exactly $\mathbb{E}\mathcal{D}^{\mathrm{pert}}(\delta z) = \mathcal{D}^{\mathrm{pert}}(0) + d\varepsilon_z^2(\|p^0\|_2^2 - \|p\|_2^2)/(2(d + 2))$ at fixed gates. The supplied seedwise Monte Carlo summaries in `data/learned/seed_means.csv` are distinct from this elementary expectation calculation. Sequence-level outcomes are regenerated by `scripts/replay_learned.py` from the fixed primary inputs and updates.

For calibration at radii 0.01 and 0.02, every seed has $\underline{\mathcal{D}}_{\mathrm{quad},s} - \overline{\mathcal{D}}_{\mathrm{rand},s} > 0$, so the bound distinguishes the empirical worst-case improvements of these recorded endpoints. At radii 0.05 and 0.1, the cohort-mean enclosure includes zero, so the worst-value ordering remains unresolved by these bounds.

Sequence-level outcomes at radius 0.1. At radius 0.1, Head has negative saved attack improvements on $2/320$ calibration sequences and $42/1280$ held-out sequences. The corresponding nonpositive certificate counts are $22/320$ and $207/1280$. On held-out, First has $13/1280$ nonpositive certificates and Random $16/1280$; Random has one negative saved attack. Quadratic has positive lower bounds on all 320 calibration and 1280 held-out sequences at that radius. These held-out results use unchanged coefficients and the actual costs reported in Appendix K.8: the mean KL exceeds the calibration cap in 18/20 Random seeds and 15/20 Quadratic seeds. One held-out nominal paired seed contrast is negative.

Table 5: Fixed learned updates under candidate-value perturbations. Loss-decrease columns use 10^{-3} units, averaged within each seed and then over the same 20 seeds. The Monte Carlo (MC) mean averages the saved random-perturbation evaluations. Lower and Upper are the evaluated analytic and feasible-attack bounds on the mean per-sequence worst-case improvement, respectively, not confidence intervals. N_+/N counts sequences with positive analytic lower bounds, among $N = 320$ calibration or $N = 1280$ held-out sequences. Held-out coefficients are unchanged, without budget rematching.

			Loss decrease (10^{-3} units)			
Split	ε_z	Update family	MC mean	Lower	Upper	N_+/N
Cal.	0	Random	5.4553	5.4553	5.4553	320/320
Cal.	0	Quadratic	5.9767	5.9767	5.9767	320/320
Cal.	0.01	Random	5.4552	5.1344	5.2198	320/320
Cal.	0.01	Quadratic	5.9765	5.6325	5.7202	320/320
Cal.	0.02	Random	5.4551	4.8059	4.9845	320/320
Cal.	0.02	Quadratic	5.9764	5.2806	5.4637	320/320
Cal.	0.05	Random	5.4547	3.7754	4.2798	320/320
Cal.	0.05	Quadratic	5.9759	4.1787	4.6947	320/320
Cal.	0.1	Random	5.4540	1.9075	3.1068	320/320
Cal.	0.1	Quadratic	5.9751	2.1882	3.4134	320/320
Held-out	0	Random	5.4408	5.4408	5.4408	1280/1280
Held-out	0	Quadratic	5.9684	5.9684	5.9684	1280/1280
Held-out	0.01	Random	5.4408	5.1082	5.1998	1280/1280
Held-out	0.01	Quadratic	5.9684	5.6222	5.7114	1280/1280
Held-out	0.02	Random	5.4408	4.7674	4.9594	1280/1280
Held-out	0.02	Quadratic	5.9685	5.2682	5.4545	1280/1280
Held-out	0.05	Random	5.4408	3.6950	4.2403	1280/1280
Held-out	0.05	Quadratic	5.9686	4.1590	4.6842	1280/1280
Held-out	0.1	Random	5.4408	1.7416	3.0450	1264/1280
Held-out	0.1	Quadratic	5.9687	2.1530	3.4012	1280/1280

Table 6: Bounds on the mean Quadratic-minus-Random worst-case contrast in Eq. (122), in 10^{-3} units. Lower and Upper average the seedwise contrast bounds over 20 seeds. $n_+/20$ counts seeds whose own contrast lower bound is positive. These finite-precision conditional bounds concern the saved inputs and updates and are not bootstrap confidence intervals. All five radii are retained for each split.

Split	ε_z	Lower	Upper	$n_+/20$
Calibration	0	0.5214	0.5214	20/20
Calibration	0.01	0.4127	0.5858	20/20
Calibration	0.02	0.2961	0.6578	20/20
Calibration	0.05	-0.1011	0.9193	3/20
Calibration	0.1	-0.9186	1.5059	0/20
Held-out	0	0.5276	0.5276	19/20
Held-out	0.01	0.4224	0.6031	19/20
Held-out	0.02	0.3088	0.6871	19/20
Held-out	0.05	-0.0813	0.9892	6/20
Held-out	0.1	-0.8920	1.6596	0/20

Figure 2(b) shows the calibration Quadratic-minus-Random enclosures from (122); Figure 12 shows the four armwise bounds. Table 6 also reports the held-out enclosures for the unchanged fitted updates, with actual held-out costs documented in Appendix K.8.

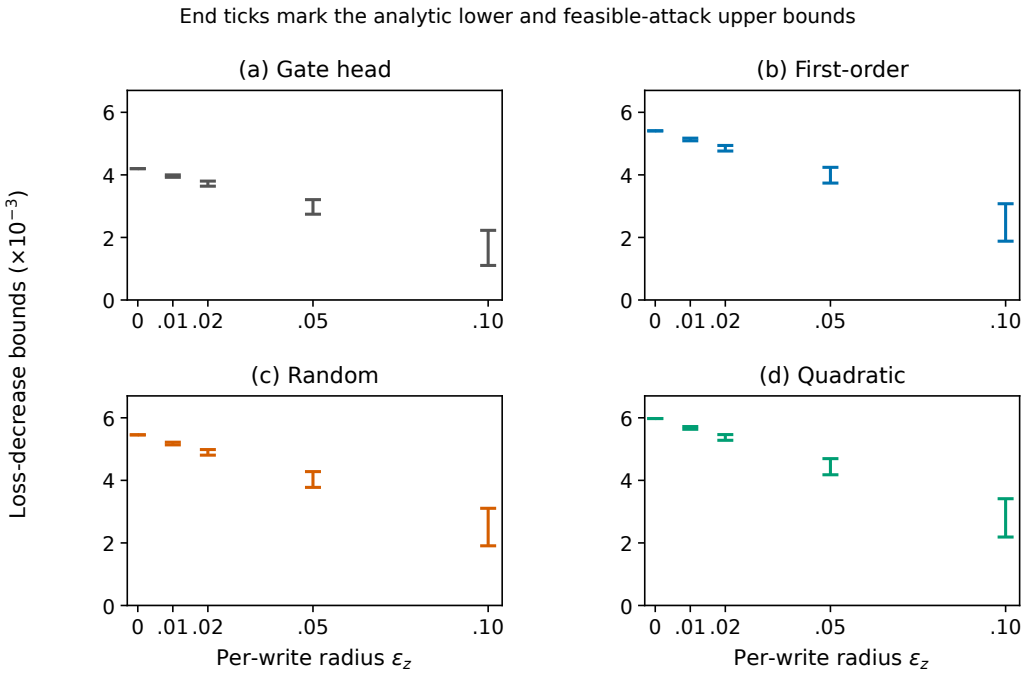


Figure 12: **Bounds for all four fixed learned updates.** Calibration cohort, checkpoint 400, $T = 64$, and original mean allocation-KL cap 10^{-3}. The four panels share axes and contain all five radii. Each capped vertical interval runs from the evaluated analytic lower bound to the feasible-attack upper value. Its endpoints are averaged within 16 sequences per seed and then across 20 seeds, with both gate sequences, write strengths, and target fixed. At zero radius the endpoints agree. These finite-precision conditional bounds are not confidence intervals. Figure 2(b) combines two arms' bounds. Appendix L.5 reports sequence-level exceptions and aggregate results; the release supplies seedwise summaries and code to regenerate individual outcomes. Appendix M.1 specifies the numerical evaluation.

L.6 Data-field conventions

The retained identifiers `record_correction`, `corr`, and `corrected` refer to the candidate-value perturbations denoted by superscript pert. In `data/controls/rows.csv` and `data/witness/rows.csv`, `worst_corrected_loss` stores a loss decrease, not the remaining loss. In `data/learned/seed_means.csv`, `nominal_loss`, `random_mean_loss` and `attack_min_loss` likewise store decreases; the last is a feasible-attack upper value, not a certified perturbation minimum. All are in original loss units before displayed rescaling. `data/learned/paired_contrasts.csv` stores descriptive paired summaries. `DATA_DICTIONARY.md` maps `first_random` and `first_second` to Random and Quadratic and defines the fields; `split` distinguishes `calibration` from `heldout`.

M Numerical consistency checks and computational resources

M.1 Three evidence roles

The paper combines analytic proofs, numerical consistency checks, and a 20-seed synthetic comparison. The proofs establish the stated family-wide and asymptotic claims. The numerical checks

compare independently implemented recurrence formulas, derivatives, inverses and finite-budget calculations at specified configurations and precisions. Check counts describe computational coverage. The empirical comparison uses 20 independently trained seed clusters; settings within a seed share that statistical unit. Candidate-value perturbations are post hoc evaluations of the same selected updates. Bootstrap intervals quantify variation across seed clusters, while perturbation enclosures are deterministic bounds conditional on the saved inputs and updates. Their numerical values use finite-precision arithmetic without outward rounding. The tables report the precision and error of each calculation. The public `VALIDATION.json` records completed package checks; original execution histories are retained privately. Package paths in this appendix and Appendices K–L are relative to the supplement root.

M.2 Mechanism and inverse-map checks

This section cross-checks the recurrence identities, response laws and inverse maps used in the theory. `audit/` supplies the independent implementations, seeds and tolerances; `data/theory/` retains compact inputs and results. `ENVIRONMENT.json` and `SOURCE_HASHES.json` identify software versions and source snapshots. The discrepancies below summarize the supplied numerical-check snapshot; reruns are assessed against the stated tolerances. The learned-controller experiment is described in Section 6 and Appendix K.

`audit/audit_extensions.py` contains 1,142 passing assertions. Its tests cover exact repeated-block factorization for block lengths two, three, and five; drift orders two, three, and four under both budgets; a direct finite-sum cross-check of a geometric-tail ℓ_1 formula; arbitrary-amplitude sign-balanced pairs; the nonlinear inverse repair; classical statistical curvature; optimal local missing-jet coefficients; actual-budget descent; and the analytic escape example.

Table 7: Numerical consistency checks. Configs. counts configurations, while the implementation-check counts for the eight groups in row order are 108, 184, 450, 45, 10, 6, 324 and 15. Observed maxima are rounded upward; zeros use the indicated precision. The paired entries in the drift/operator, curvature and escape rows keep the order stated in their error definitions.

Check / comparison	Configs.	Arithmetic / error	Tolerance	Observed
Block identity: unroll vs. product	36	float64; max atom abs. error	10^{-13}	3.93×10^{-15}
Drift / operator: boundary roots; finite sums	60 / 64	80 digits: max abs. budget residual; float64: max abs. L1 discrepancy	10^{-40} / 2×10^{-14}	3.60×10^{-52} / 2.23×10^{-16}
Nonperiodic: unroll vs. pair law	90	float64; max atom abs. error	4×10^{-12}	5.63×10^{-15}
Inverse repair: logits vs. prescribed tilt	45	float64; max atom abs. error	2×10^{-13}	5.56×10^{-17}
Curvature: score projection vs. formula	5	float64; abs. curvature / KL coefficient	10^{-9} / 2×10^{-10}	0 / 0 (reported)
First missing jet: projected coefficient vs. finite KL	3 orders; 3 steps	80 digits; relative coefficient error at smallest step	2×10^{-3}	4.78×10^{-5}
Finite descent: actual cost and loss vs. bounds	54	float64; max violated slack (positive part)	2×10^{-14}	4.33×10^{-16}
Remote escape: zero drift / constant KL / gain	5	80 digits; abs. drift and KL differences	0 / 10^{-50}	0 / 0 (reported)

Results in `data/theory/results/extensions.json`, in row order: `block/recurrence`; `orders` and `L1`; `aperiodic`; `inverse`; `curvature`; `jets`; `certificate`; `escape`. The drift/operator errors are $\max|C(t)-b|$ over 60 computed roots t (cost C, budget b), and the maximum absolute discrepancy between the threshold formula and direct finite sum over 64 ℓ_1 checks.

`audit/audit_gate_geometry.py` passes 90 high-precision checks of the reduced two-amplitude drift coefficient, the zero-drift square repair, and the coarsened-sensitivity negative control. Its reduced-jet relative discrepancy is below 1.46×10^{-5} at the tested finite steps. `audit/make_figures.py` separately checks 40 canonical boundary roots at 80-digit precision, with direct operator-formula cross-checks already included in the extension script. These are reported separately from the 1,142 extension assertions.

The nonperiodic tests use the amplitude rules defined beside Figure 3 in Appendix C.1.4. NumPy's generator is seeded once with 884119 before the ascending-T loop; each generated array is reused across gate bases and budgets at that length. The ramp includes both endpoints. Gate bases are $0.2, 0.5, 0.8$, lengths are $16, 64, 256, 1024, 4096$, and budgets are KL 10^{-4} or operator 0.005. The scalar boundary search supplies feasible responses; Appendix C.1 proves the global upper bounds.

The new block/recurrence probability discrepancy is at most 3.93×10^{-15}, and the nonperiodic direct-recurrence discrepancy is at most 5.63×10^{-15}. The geometric-tail operator formula agrees with direct sums to 2.23×10^{-16}. High-precision optimal-jet comparisons use step sizes $10^{-4}, 5 \times 10^{-5}, 2.5 \times 10^{-5}$; the final relative discrepancy from the limiting coefficient is below 4.78×10^{-5}. That discrepancy includes both the finite-step asymptotic remainder and floating-point error. Actual-budget certificate inequalities have maximum negative numerical slack below 4.33×10^{-16}, within the declared 2×10^{-14} tolerance.

M.3 Supporting numerical checks

Four supporting scripts were reexecuted without altering their mathematical configurations. The core geometry and keep/write/reset script passes 590 checks, with maximum discrepancy 3.33×10^{-13} and minimum observed sweep-descent slack 0.0017242. The shared-query/retention script passes 4,162 checks, including 162 shared bases and 2,592 finite target operations. The earlier finite-reconstruction script passes 209 checks of the exact crossover, non-half-baseline square repair, finite radii, and an all-lifts control. A further supporting replay reproduces polynomial KL halving slopes 3.999904, 6.000095, and 8.000092 and spectral-completion error below 4.45×10^{-16}. The final replay reports slopes and reconstruction error separately from the assertion counts above.

High-precision calculations use 80 or 90 decimal digits; direct linear-algebra and probability comparisons use NumPy/SciPy floating-point arithmetic. The four scripts are in `audit/retained/code/`, with compact results under `data/theory/retained/results/`. `scripts/run_audits.py` executes each suite separately and creates per-suite logs in a fresh output directory.

M.4 Finite-realization and temporal-spread checks

The script `audit/audit_revision.py` checks BAM realization and temporal-spread formulas using seed 20260916, NumPy/SciPy calculations of probabilities and quadrature, and 90-digit mpmath calculations of small-KL asymptotics. Its 5,246 implementation checks cover the configurations below; inputs and evaluated updates are supplied under `data/theory/results/revision/`; reruns use a separate output directory.

BAM configurations. Ten seeds, lengths $T \in \{3, 8, 24\}$ and dimensions $d \in \{1, 3\}$ give 60 bases. Four signed steps per base yield 240 tilt evaluations. The same-count comparison has 60 rows, comparing one unrelated column with the predicted second-order column while holding the first-order column fixed. Exact, first-order and second-order realizations at each base give 180 budget-and-descent evaluations. Orders $q = 1, 2, 3$, with four steps each, give 12 sharp-ICM-jet evaluations in 90-digit arithmetic.

Reference configurations. Five reference families and six lengths give 30 settings. Four signed standardized tilts give 120 regular time-position-tilt evaluations, and three actual KL budgets give 90 crossover evaluations. A separate rare-tail control uses five lengths violating standardized-tail regularity.

The stored base paths are generated from positive Dirichlet branch probabilities. Candidate and anchor vectors are bounded in the unit ball; half the seed groups have zero anchors. For each case the terminal inverse is checked by backwards probability reconstruction, and the analytic ICM Fisher is checked against a separately formed score Gram matrix. Multiplication of the first required logit coefficient by that score matrix recovers the centered negative task score. First- and second-degree all-task templates are also evaluated independently from prefix mean and covariance tensors.

The maximum signed-prefix allocation discrepancy is 2.23×10^{-16}, and the maximum backwards inverse discrepancy is below 3.34×10^{-16}. These discrepancies are measured at the 60 specified bases. The same-count feature comparison evaluates the predicted local normal residual.

For the sharp ICM coefficients, use baseline branch rows $(.55, .25, .20)$, $(.45, .35, .20)$, $(.60, .15, .25)$ and origin-resolved task values $(-.8, .7, 0, -.5, 0, .9, 0)$. The available space contains the first q required coefficients, and the next coefficient is fitted by its Fisher projection. At steps $.01, .005, .0025, .00125$, the final halving exponents are

$$4.00002075, \qquad 6.00028551, \qquad 7.99912755.$$

The corresponding predicted orders are four, six, and eight. The maximum relative discrepancy from the limiting coefficient at the smallest step is 6.05×10^{-4}; it includes the finite-step remainder and arithmetic error.

Actual-budget descent evaluations. Each of the 180 BAM endpoints uses the actual target $g = m_T^0 - y$ and requested budgets $\tau = 10^{-3}$ and $\delta = .04$. An explicit initial radius is reduced until the score-error certificate holds for the chosen exact or truncated realization. All recorded endpoints satisfy both actual budgets and the positive loss lower bound. Backtracking supplies a feasible endpoint satisfying these checks. Input configurations, step sizes, reduction counts, target variance, cumulative error, both distances, and loss decrease are all recorded. The origin-resolved operator bound also bounds the candidate-only operator change.

Nonuniform references. The five reference distributions and the 6×3 length–budget grid are defined in Appendix H.3 and displayed in Figure 5. Probabilities are stored and evaluated in log space to preserve the positive tail support. Direct KL agrees with integration of the tilted variance to 8.06×10^{-16}. Canonical updates solve the monotone scalar boundary equation characterized by Corollary H.3; their largest absolute budget residual is 3.14×10^{-15}. The sharp horizon approximation has maximum relative discrepancy 0.02176 on this grid.

Reference and reset controls. The rare-tail example has $W_n \to 1$ while a fixed tilt has vanishing KL and operator change, demonstrating failure of variance-only localization. The fixed-zero-anchor label control and mass/Fisher decompositions are evaluated in `audit/audit_bam_retained.py`. Its outputs identify that suite separately from the revision suite and the operational suite in Appendix M.5. The supporting scripts in Appendix M.3 use their original mathematical configurations.

Connection to the theory. The BAM checks evaluate finite realization for the specified features and observation maps. The temporal-spread checks evaluate the response formulas under the regularity and control conditions of Theorem H.2. Section 6 and Appendix K separately assess finite updates in a learned retention controller with fixed candidate values and write strengths.

M.5 Reset realizations, observation ranks, and perturbation checks

Operational suite. The script `audit/audit_operational.py` imports no retained audit module. It compares direct probability recurrence, the explicit reset-fiber inverse, complex-step probability derivatives, analytic second-order recurrence coefficients, and a high-precision implicit reset solve. Its 60 bases are 12 fixed seeds at $T \in \{2, 3, 5, 9, 16\}$, with positive gates drawn from a declared Dirichlet generator. Each base uses terminal-only, one-intermediate, and all-prefix schedules; coincident schedules at the smallest horizons are marked as duplicates.

Realizations and local geometry. The complete reset-fiber check uses 180 observation configurations and has maximum observed-candidate discrepancy 2.23×10^{-16}. Every observation rank agrees with $T - 1 + N_{\mathrm{obs}}$ at the recorded singular-value threshold. Comparing the complex-step Jacobian Gram with the coarse/reset Fisher split gives maximum discrepancy 1.67×10^{-16}. The all-first-order-lifts check includes a noncanonical nullspace lift for $T > 2$ and 60 second-order coefficient comparisons, with discrepancy below 2.09×10^{-17}.

Three-write separation and mismatch. Among nine signed endpoints, six are positive descent steps; candidate targets and the fixed shared-feature constraint agree to below 2.25×10^{-16}. The sharp origin-resolved mismatch is evaluated at four positive step sizes in 90-digit arithmetic. Its final halving order is 3.99995, approaching four, with coefficient approaching $3/512$.

A one-column negative control has globally fixed last-write mass and cannot attain the nonzero target; its coarsened KL coefficient $275/41472$ is checked separately, with final halving order

4.00001. The 2,134 implementation checks evaluate the reset realizations, observation ranks and local expansion coefficients above. The scalar root calculation evaluates the analytic curve established by the implicit-function argument in Appendix G.5.1, and the rank calculations cross-check the formula in Theorem G.1 at the sampled bases. Stored results are in `data/theory/results/operational/`; the public runner records each new suite execution separately.

Separate task-response and candidate-value perturbation checks. The package includes symbolic and 90-digit numerical implementations under `audit/task_response/`. They derive the probabilities independently from the two gates, recover the two cost expansions and $275/1536$ exactly, check 10 signed recursion points, 6 candidate-operator endpoints, 10 actual-KL endpoints, and 40 task-gap cases including unused budget. The strict event-based lower bound (106) is also checked at each KL endpoint. The stored symbolic and numerical summaries are in `data/theory/task_response/`.

A separate implementation, `audit/audit_r3.py`, checks the operator norm and its grouped analogue, singular local Fisher response, coarsened Fisher inversion, and (30). It uses 24 fixed positive baselines and 192 independently bounded candidate-value perturbation cases, recording 840 implementation assertions. Maximum discrepancy in the exact squared-loss expansion is 1.34×10^{-15}; every recorded uniform lower bound and positive perturbed-loss certificate holds within the stated tolerance. The script implements the formulas separately; compact inputs and outputs are in `data/theory/results/r3/`. Its source specifies arithmetic precision and tolerances. `scripts/run_audits.py` runs this and the earlier suites in a disposable source copy without changing the supplied data.

M.6 Saved update parameters and result validation

Original run validation checked configuration and file hashes, expected measurement-key sets and nonfinite values. Learned sequence-level checks used $(\text{seed}, \text{split}, \text{sequence}, \text{arm}, \varepsilon_z)$; the R4 aggregate instead indexes seed, checkpoint, length, task, arm, metric, budget, objective and split. The public `reproduce.sh` command `verify` checks `MANIFEST.json`, the retained condition keys and saved parameter states. Finite outcomes of either sign remain valid measurements.

The original confirmation has 20 training and 180 probe tasks, with 17,280 seed/condition-level measurement rows and 20 independent seed clusters. Independent recomputation covers 160 primary endpoints evaluated on both splits (320 comparisons), 480 prescribed-path values, and 80 independent-control reference problems. The additional fixed-gate evaluation has 94 tasks: 54 witness, 20 constructed-control and 20 learned-update tasks, producing 1,296, 240 and 32,000 rows, respectively. The learned evaluation uses the original trained cohort and its selected updates. Measurement-key checks cover the full configured Cartesian sets within each task. The original workflow tests injected deleted or duplicated rows, an unexpected sequence and a not-a-number (NaN) value to verify defect detection; they are distinct from the scientific unit tests in the public release.

Selected endpoint diagnostics. All 80 primary loss endpoints came from normal final optimizer trials; all 480 underlying trials returned normal termination. Their final feasibility factors lie between 0.9999998543 and 0.9999998999. The maximum loss penalties are $6.29, 6.87, 6.96, 7.84 \times 10^{-10}$ for Head, First-order, Random and Quadratic. The corresponding maximum saved scaled stationarity diagnostics are $1.58, 1.68, 1.59, 1.64 \times 10^{-7}$. These first-order diagnostics are evaluated at the unrescaled trial endpoint. Full-precision measurements are supplied as comma-separated-values (CSV) tables; per-seed `selection.json` files within `data/r4/primary/` and `data/controls/diagnostics/` retain compact JavaScript Object Notation (JSON) trial and selection diagnostics. Complete optimizer trajectories are retained privately. The 17 non-successful constructed-control trials belong to the separate 510-trial collection in Appendix L.2; its selected updates can come from evaluated iterates, inherited endpoints or analytic candidates.

The release retains all 60 controller states under `data/r4/checkpoints/`, primary states and frozen inputs under `data/r4/primary/`, 60 constructed states under `data/controls/states/`, and 180 representation-control states under `data/e2/states/`. `scripts/verify_primary.py`, `scripts/verify_constructed.py` and `scripts/verify_diagnostics.py` recompute the retained states and diagnostics. The full secondary condition table is supplied without all nonprimary endpoint caches; reoptimizing those conditions is

a new computation, not replay of an omitted cache. `scripts/replay_learned.py` regenerates the 32,000 learned perturbation rows from the fixed primary inputs and updates, while seedwise summaries are supplied. Author identifiers and machine-specific paths are removed; the authors retain the original run files and the mapping to this release.

M.7 Recorded computation and public reproduction

Software and execution. The main run used Python 3.12.12, NumPy 2.2.6, SciPy 1.17.0, PyTorch 2.10.0 development build (commit prefix 449b176), mpmath 1.3.0, SymPy 1.14.0 and pandas 3.0.0 in the site framework environment. Primary training used logical devices of PyTorch's Intel graphics processing unit (GPU) backend, named XPU; constrained coefficient optimization and reference calculations used central processing unit (CPU) float64. High-precision checks specify their arithmetic separately. Main training used 12 logical tile ranks per node, and probes used 11 CPU ranks. `ENVIRONMENT.json` records these original versions separately from the tested CPU replay environment (Python 3.13.5); `requirements-tested.txt` pins the replay dependencies. These replay versions are not presented as the original training environment.

Allocated resources. Main training, probing and aggregation used approximately 0.3714 node-hours in the scheduler history, excluding queue waiting and development runs. The separate candidate-value perturbation evaluation used CPU for scalar constructed models and XPU for the original cohort's fixed learned updates. Its five jobs were device validation, constructed evaluation, learned evaluation, an aggregation that reached its walltime limit, and a successful aggregation retry. Including the failed allocation, this evaluation used approximately 0.344 node-hours. Both totals report allocated node-time from elapsed scheduler allocations. For the first aggregation, the scheduler records a walltime failure despite a conflicting shell receipt; the results use the successful retry. The authors retain the underlying scheduler records privately; those operational records are not required by the public numerical replay commands.

Public reproduction. From the supplement root, `bash reproduce.sh verify` checks retained states and diagnostics, `bash reproduce.sh tables` regenerates the six numbered empirical tables and seed-bootstrap summaries, and `bash reproduce.sh figures` renders the numerical figure assets from `data/figure_exports/`. `FIGURE_MAP.md` maps the internal export IDs to paper figure numbers and documents coordinate provenance; those IDs are not the paper numbering. The CSV fields `source_path` and `source_sha256` describe earlier editorial provenance, not files opened at runtime. Plotting uses the supplied coordinates and does not promise pixel-identical layouts or equality to every earlier audit table.

The `reproduce.sh` modes `learned`, `witness`, `tests` and `audits` respectively regenerate the fixed-learned perturbation rows, evaluate the explicit witness across lengths, run the scientific unit tests and rerun the independent numerical suites. Mode `all` executes all seven modes in fresh output directories without modifying retained source or data. `README.md` specifies installation, output locations and optional new training or optimization through `scripts/optional_new_run.py`. `DATA_DICTIONARY.md` defines the data and `CODE_CHANGES.md` documents the release adaptations. Original archives, scheduler logs and full optimizer histories are omitted; adverse scientific outcomes, budget overruns and compact trial-status diagnostics remain available.

Hardware inventory. The records identify CPU versus logical XPU use, logical device assignments, software builds and scheduler node counts. CPU model, physical GPU-package model and count, and memory capacity were not recorded. Per-worker Basic Linear Algebra Subprograms (BLAS) and OpenMP thread limits are one where set by the launch environment. The recorded environment identifies the historical run; the separately tested CPU environment supports public fixed-state and summary replay. Bitwise full-training or optimization equivalence across environments is not asserted.

N Related work: response geometry and recurrent memory

N.1 Detailed comparisons underlying the main interpretation

Recurrent learnability and starvation. Classical analyses study temporal gradients (Bengio et al., 1994; Pascanu et al., 2013), gated long-term retention (Hochreiter & Schmidhuber, 1997), and learnable forgetting (Gers et al., 2000). Further perspectives include time-scale-aware gates (Tallec & Ollivier, 2018), linear and nonlinear approximation theory (Li et al., 2022; Wang et al., 2024), stable reparameterization (Wang & Li, 2024), and noisy temporal learnability (Livi, 2026). Gu et al. (2020b) study how saturation hinders learning of long-retention gates; Zucchet & Orvieto (2024) analyze state sensitivity to parameter changes in long-memory networks even without exploding temporal gradients. Gradient-starvation work studies useful features failing to acquire an effective learning response (Tachet des Combes et al., 2018; Pezeshki et al., 2021), with related questions in simplicity bias (Shah et al., 2020). Tangent-kernel and lazy-training analyses characterize their specified training regimes (Jacot et al., 2018; Chizat et al., 2019).

Our question concerns how much a specified shared recurrent family can improve prediction under a fixed sensitivity budget, even when its local task efficiency is one. Theorem 3.1 gives the finite-response separation, and Section 3.2 its fixed-task loss consequence. In the canonical construction, the baseline output derivative and loss gradient are nonzero independently of length, while the finite-budget response decays. Appendix H.3 separately defines temporal spread from the reference allocation and explains how that allocation depends on the gate and write sequences.

Sensitivity and conditional responses. Sobolev training fits available derivative targets (Czarnecki et al., 2017), whereas input-gradient regularization penalizes sensitivity to improve robustness and interpretability (Ross & Doshi-Velez, 2018). Attention allocations can be nonidentifiable from their outputs (Brunner et al., 2020); their explanatory role has been debated (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Separately, Kim et al. (2021) show that standard dot-product self-attention is not Lipschitz on an unbounded input domain and construct a Lipschitz L^2 alternative. Our budgets constrain the change between two conditional write-response operators, with the candidate perturbations evaluated under fixed pre-update and post-update gate sequences.

Local geometry and finite KL budgets. Natural gradient uses a statistical metric (Amari, 1998; Martens, 2020); its pushforward through a redundant parameterization gives a tangent-space projection (van Oostrum et al., 2023). Exponential-family duality and information projection provide related tools (Wainwright & Jordan, 2008; Csiszár & Matúš, 2003), as do relative-entropy policy updates (Peters et al., 2010; Schulman et al., 2015). KL-neighborhood sensitivity theory characterizes small-budget optimal responses and bounds KL-normalized parametric first-order sensitivities by unrestricted responses (Lam, 2016, arXiv v2, Sections 3.1 and 7, Proposition 7.1). A statistic on the sample space instead induces conditional-score information loss (Amari, 1982); quantitative near-sufficiency is studied by Yamaguchi & Nozawa (2024). Our candidate observation merges reset origins, and the recurrent inverse identifies the hidden directions compatible with that observation. The task-gap identity is derived directly for the finite categorical model and its budget-feasible shared image. Lam's multitime construction constrains a shared input marginal for independent and identically distributed (i.i.d.) observations. Here the feasible set is the allocation image of recurrently shared controls, and the length-uniform bounds use the global drift conditions of Section 4.

Curvature and finite response. Statistical curvature (Efron, 1975; Amari, 1982; Sei, 2011), parameter-effects curvature (Bates & Watts, 1980; Kass, 1984), and model-manifold geometry (Transtrum et al., 2011) distinguish tangent information from nonlinear behavior. Appendix D.5 uses statistical curvature in the exponential-family geometry. Higher-order natural-gradient methods improve finite-step reparameterization behavior (Song et al., 2018), and smooth-parametrization results study optimization landscapes under lifts (Levin et al., 2025). Our bounds additionally control recurrence-amplified discrepancy throughout the actual budget-feasible shared family.

Realization and descent. Exponentiated gradient and mirror descent provide the entropic update (Kivinen & Warmuth, 1997; Beck & Teboulle, 2003); relative smoothness provides a reference-function descent framework (Bauschke et al., 2017; Lu et al., 2018). We derive the negative-entropy bound and its constant for the present loss, then use the recurrent inverse and cumulative

endpoint error to certify a realizable update. Analytic-system error bounds (Luo & Pang, 1994) relate residuals to solution sets. Our length-uniform conclusions combine local expansions with global drift and reference-tail conditions. The escape construction in Appendix B.6 and rare-tail reference in Appendix H.4 identify the roles of these two global conditions.

What an added feature must accomplish. Network growth adds missing functional directions (Verbockhaven et al., 2024); neuron splitting uses second-order functional descent and characterizes local improvements (Wu et al., 2019). These methods motivate a comparison of what an additional direction must achieve. In our sigmoid construction, local task efficiency is already one, while accumulated finite drift still limits the response. The added direction cancels that drift without changing the baseline. CNM and origin-resolved ICM targets require different statistics; candidate observations admit other realizations, and a task optimum need not realize the selected target. Appendix D.2 gives the fixed-data rank calculation; related Hadamard-power and Khatri–Rao rank analyses appear in Madden & Thrampoulidis (2024). The next subsection gives concrete recurrent computations and their conditional update maps.

N.2 Recurrent architectures and conditional update maps

The correspondences below concern specified computation paths with the other paths fixed. Theorem 3.1 applies to its declared two-amplitude sigmoid family; applying its finite-response laws to another architecture requires a corresponding shared-family and budget analysis.

Memory capacity, projection, and state-space design. Ganguli et al. (2008) quantify retained input information in dynamical systems. Our allocation Fisher measures local changes in the distribution of individual-write influences. High-order polynomial projection operators (HiPPO) specify recurrent memory through polynomial projection (Gu et al., 2020a), and the structured state-space sequence model (S4) develops efficient computation for long sequences (Gu et al., 2022). The linear recurrent unit (LRU) emphasizes linear recurrence, initialization and normalization (Orvieto et al., 2023), while Mamba uses input-dependent selective state-space parameters (Gu & Dao, 2024). These systems motivate comparisons of memory functions and their parameterizations beyond nonnegative categorical averages. For the linear recurrence specified in Appendix J.4, (120) gives the conditional candidate-sensitivity operator. Griffin combines gated linear recurrence with local attention (De et al., 2024). For this comparison, the response of a stacked architecture is computed from the derivative of its composite map.

Normalization, attention, and shared forgetting. The recurrence interpretation of linear attention (Katharopoulos et al., 2020), data-dependent forgetting in gated linear attention (Yang et al., 2024a), and structured state-space duality (Dao & Gu, 2024) connect scan implementations to sequence-wide coefficient matrices. The response budget is specified separately for the coefficient map being compared. Extended long short-term memory (xLSTM) includes scalar and matrix memories with stabilization and additional state-dependent routes (Beck et al., 2024). The scalar sLSTM accumulation before its output gate has the form

$$c_t = f_t c_{t-1} + i_t z_t, \qquad n_t = f_t n_{t-1} + i_t, \qquad \widetilde{h}_t = c_t / n_t,$$

which is the CNM update with write strength i_t (Beck et al., 2024). The conditional intervention varies the specified retention controls while keeping the precomputed candidate and write-strength sequences fixed, including when their generators depend on state. The full-network response also includes the derivatives of any recomputed generators and downstream readout maps.

Forgetting Transformer introduces forgetting gates into softmax attention (Lin et al., 2025). At a fixed terminal query, let $\widetilde{q}_T = q_T / \sqrt{d_{\mathrm{key}}}$. Forgetting Attention assigns positive weights proportional to

$$\exp(\widetilde{q}_T^{\top} k_i) \prod_{s=i+1}^{T} f_s.$$

With the query and keys fixed, the first factor acts as a positive write strength (Lin et al., 2025). Across prefixes, varying queries generally give prefix-dependent write scores. Appendix I treats

jointly varying query and retention controls across observed prefixes and characterizes the resulting compatibility conditions, including cases where retention additions alone cannot repair a prefix discrepancy. When candidate changes also affect queries or keys, the full derivative includes those additional paths.

Delta memories and editable directions. The fast-weight view of linear transformers (Schlag et al., 2021), parallel delta-rule computation (Yang et al., 2024b), and gated delta networks (Yang et al., 2025) study structured updates of an online matrix state. Decoupled channel-wise erase and write (Hatamizadeh et al., 2026) modify that update rule. Query-derived Erase Direction (QED) adds a key-orthogonal, query-derived component to the erase vector while retaining a rank-one state edit (Gupta et al., 2026, arXiv v1, Eqs. (11)–(12)); this changes the residual written along the current key. Our intervention varies shared controls that determine a recurrent allocation, with candidates and other specified routes fixed. Appendices J.2–J.4 formulate shared matrix-write derivatives, normalized observations and the linear-memory conditional operator as the corresponding objects for these comparisons.

Coordinate scaling and observed directions. Appendix J.1 gives a two-dimensional recurrent witness for diagonal coordinate scaling, within the broader setting of frame scaling (Casazza & Chen, 2017). Appendix J.3 studies the directions retained by normalized observations and their conditioning; direction-only observations are related to scale ambiguities in bearing rigidity (Zhao & Zelazo, 2019). These extensions specify how the parameter group, response metric and observation map determine the local update response.