When Does Momentum Help Online Associative Memory? A Random-Query Risk Analysis

Anonymous authors
Paper under double-blind review

Abstract

We study online associative memories through their retrieval error on queries that arrive at random times while key–value pairs stream in. The delta rule corrects along the arriving key, while heavy-ball momentum also carries decayed corrections from earlier keys. We ask when this accumulated correction lowers the expected retrieval error, and find that the answer depends first on when the memory updates. When momentum evolves continuously between key changes, we derive leading-order risks in closed form for two nearly collinear keys under Markov access and an exponential query time. From the weak direction, a fixed momentum rule beats every continuous-time delta rule with gain below one. An over-relaxed reflect–project schedule beats that momentum rule for every nonzero initial error, although momentum with large β overtakes it from the weak direction. With each family tuned at a fixed key separation, continuous-time momentum beats the best constant gain found only for nearly identical keys (cosine above 0.98 at every tested ρ) under reconstruction. For $K > 2$ keys, a compatibility matrix that pairs the access modes with the keys' weak directions decides the comparison at small separation. Exact risks on frozen CIFAR-100 and Omniglot features reproduce these effects, and there the tuned delta rule selects a gain above one under reconstruction and below one under the associative query. With one update per token, the comparison reverses: tuned momentum attains lower mean risk than the tuned constant-gain delta rule in every tested synthetic and frozen-feature condition, by up to 0.034 in normalized reconstruction risk.

1 Introduction

Fast-weight memory updates the weights of a neural network incrementally during forward propagation, serving as an online associative memory (Kohonen, 1972; Schmidhuber, 1992; Ba et al., 2016). Schlag et al. (2021a) show that linear attention (Katharopoulos et al., 2020) is equivalent to this formulation, since its running sum of key–value outer products forms the fast weights. They further adopt the delta rule (Widrow & Hoff, 1960; Sayed, 2008), which is more effective than Hebbian learning (Hebb, 1949) for memory updating. More recently, Titans (Behrouz et al., 2025b) and MDN (Huang et al., 2026) incorporate *heavy-ball momentum* (*momentum* for short) into the delta rule, so that each update also draws on past key–value pairs. Specifically, let $W \in \mathbb{R}^{m\times d}$ denote the memory, which returns $\hat{v} = Wk$ for a key $k \in \mathbb{R}^d$. Given a target value $v \in \mathbb{R}^m$, the delta rule updates the memory with the residual $r = v - \hat{v}$ along k:

$$W^+ = W + \theta\, r k^\top,$$

where θ is the learning rate, usually taken in $(0, 1)$ as in DeltaNet and Gated DeltaNet (Yang et al., 2024; 2025). Titans maintains a momentum state $S \in \mathbb{R}^{m\times d}$ with coefficient $\eta \in [0, 1)$, accumulating past residuals:

$$S^+ = \eta S + \theta\, r k^\top, \qquad W^+ = W + S^+. \tag{1}$$

Momentum retains an exponential moving average of past gradients, permitting larger step sizes without divergence and thereby accelerating convergence on deterministic strongly convex objectives (Polyak, 1964; Qian, 1999). In stochastic optimization this advantage is less clear: under small

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

constant step sizes, *stochastic gradient descent with momentum* (SGDM) is equivalent to SGD with a rescaled step size (Yuan et al., 2016; Paquette & Paquette, 2021), and on quadratic objectives, SGD attains the same sample complexity lower bound as SGDM once both bias and variance are accounted for (Kidambi et al., 2018; Ganesh et al., 2023). A suitably tuned delta rule might match this benefit without momentum. With a fixed target, online associative memory is a stochastic least-squares problem with loss $\frac{1}{2}\|Wk_t - v_t\|^2$ at step t, but it is judged by its retrieval error when queried, at a random time and under a query that need not match the updates. This naturally raises the following question on the role of momentum in online memory updates:

When does momentum achieve lower expected retrieval error than the delta rule?

Our main contribution is an exact risk analysis in a tractable model, together with a criterion for nearly collinear keys and numerical evaluations on frozen image features. We first establish it in a special two-key setting, then extend it to the nearly collinear multi-key case under reversible access. We begin by unrolling equation 1, which separates what momentum adds to the delta rule:

$$S_t = \theta\, r_t k_t^\top + \sum\nolimits_{j<t} \eta^{t-j}\, \theta\, r_j k_j^\top .$$

Beyond the delta rule's current update, momentum carries all preceding residuals, decayed by η. How strongly each past residual acts at k_t depends on the angle between k_j and k_t. To analyze this cleanly, we take two nearly collinear unit keys $k_\pm \propto (1, \pm\varepsilon)^\top$ with $0 < \varepsilon \ll 1$. These two keys share a strong direction e_{s} and split along a weak direction e_{w}. Both keys strongly excite e_{s}, so every rule removes most of the error there, while along e_{w} the rules differ only at order ε^2. We therefore compare the rules first from e_{w}, where the smaller retrieval risk R decides which rule is superior, and treat e_{s} separately.

To make this comparison concrete, we assume that the key presented to the memory switches between k_+ and k_- according to a two-state *Continuous-Time Markov Chain* (CTMC) with transition rate ρ, and that the memory is queried at an independent unit-rate exponential time. Let $\mathcal{D}$ denote the family of continuous-time delta flows, each member $d \in \mathcal{D}$ with a rate $\gamma_t \geq 0$ and an effective gain in $[0, 1)$. Its expected query risk from e_{w} has an exact positive infimum, $1 - (1 + 4\rho)\varepsilon^2 + O(\varepsilon^4)$ (Theorem 3.1). Momentum carries its correction across key changes, and for suitable (α, β) this lets a fixed momentum rule beat every delta rule with gain below one from e_{w} (Section 3).

The delta rule's learning rate θ is conventionally confined to $(0, 1)$. Recent studies of the delta rule have begun to loosen it: Grazzi et al. (2025) extend the parameterization to admit negative eigenvalues, and Siems et al. (2025) enlarge state tracking through products of Householder reflections. They lead us to suspect that the range $\theta \in (0, 1)$ is the constraint, and to ask: does a wider range of learning rates admit a lower risk, and how would it compare with momentum? We widen the range to $[0, 2]$ and prove that reflect–project (RP), a momentum-free rule built to isolate the mechanism, beats the momentum rule $H(1, 30)$ for every nonzero initial error on an explicit range of ρ, although momentum with large β overtakes it from e_{w} (Section 3). Here the comparison is at a random query time, which can fall between a reflection and its complementary projection, whereas reflection–projection schemes and heavy-ball Kaczmarz are usually analyzed for convergence (Bauschke & Kruk, 2004; Bauschke et al., 2016; Steinerberger, 2021b; Loizou & Richtárik, 2020; Alderman et al., 2024). On frozen CIFAR-100 (hereafter CIFAR) and Omniglot features, the delta rule with a single learning rate θ tuned over $[0.05, 2]$ attains lower mean risk than the momentum rule $H(\alpha, \beta)$ tuned over the tested grids of (α, β) in all four source–query settings, with the selected θ above one under the reconstruction query and below one under the associative query, and the result persists on class-disjoint splits and when both rules are tuned on a separate selection split (Appendix A.3). With one update per token, this reverses on synthetic and frozen natural keys (Section 5). The answer therefore depends first on when the memory updates: our theory treats momentum as a flow between key changes, where tuned constant gains win except for nearly identical keys (Appendix D.4.2), while the per-token regime, where momentum wins, is compared numerically. Extending to K keys ($K > 2$) with multiple weak directions, accessed through a K-state CTMC, we give a criterion, for nearly collinear keys, under which a momentum-free rule attains lower retrieval risk than momentum for every nonzero initial error, in terms of the keys' weak directions and the access modes.

We test four factors that move the comparison, namely the retrieval query, the key geometry, the access order, and the elapsed time, on P, RP, and $H(1, 30)$, using frozen CIFAR and Omniglot features and specified geometries (Section 5). For nearly collinear keys under the reconstruction

query, the criterion predicts which of the two rule families wins. Appendix F gives the full account of related work and how this paper differs from it.

2 PRELIMINARIES

2.1 MEMORY UPDATES

A memory $W \in \mathbb{R}^{m\times d}$ returns $\hat{v} = Wk$ on a nonzero key $k \in \mathbb{R}^d$, with residual $r = v - Wk$ and the orthogonal projection matrix $P_k = kk^\top/(k^\top k)$ onto $\mathrm{span}\{k\}$. The update rule with learning rate, or gain, θ gives $W^+ = W + \theta\,(v - Wk)k^\top/(k^\top k)$, which reduces to $W^+ = W + \theta r k^\top$ for a unit key. Under a fixed target memory $W_\star$ with $v = W_\star k$, the parameter error $E = W - W_\star$ becomes

$$E^+ = E\,A_k(\theta), \qquad A_k(\theta) = I - \theta\,P_k, \tag{2}$$

with $A_k(0) = I$, $A_k(1) = I - P_k$, and $A_k(2) = I - 2P_k$ representing wait, projection, and reflection: waiting leaves the error unchanged, projection removes the error along k, and reflection flips its sign along k. Projection is the Kaczmarz step (Strohmer & Vershynin, 2009; Gower & Richtárik, 2015), and $\theta \in (0, 2)$ is the relaxation range of projection methods (Bauschke & Borwein, 1996).

Definition 2.1 (Residual correction)**.** For the nth key–value pair (k_n, v_n), the normalized rank-one residual correction at a memory W is $U_n(W) = (v_n - Wk_n)k_n^\top/(k_n^\top k_n) = -EP_{k_n}$.

The delta rule and momentum evolve according to

$$\begin{aligned} &\text{delta:} && W_n = W_{n-1} + \theta_n U_n(W_{n-1}), \\ &\text{momentum:} && S_n = \eta S_{n-1} + \theta_n U_n(W_{n-1}), \qquad W_n = W_{n-1} + S_n, \end{aligned}$$

with $\eta \in [0, 1)$ and $S_0 = 0$. This is momentum LMS in adaptive filtering (Roy & Shynk, 1990).

2.2 ACCESS MODEL AND RETRIEVAL RISK

Let $(X_t)_{t\ge 0}$ be a stationary continuous-time Markov chain on a finite state space with transition rate ρ and stationary law π, and let $(\tau_j)_{j\ge 1}$ be its increasing key-change times. The active key at time t is k_{X_t}, with projection $P_t = P_{k_{X_t}}$, and the parameter error is time-indexed, $E_t = W_t - W_\star$. Stochastic gradient methods under such Markovian sampling are analyzed by Nagaraj et al. (2020) and Even (2023). We compare three rules driven by this process.

Definition 2.2 (Sub-unit-gain delta flows)**.** The family $\mathcal{D}$ of sub-unit-gain delta flows consists of the first-order dynamics

$$\dot{E}_t = -\gamma_t\,E_t P_t, \qquad \gamma_t \ge 0, \tag{3}$$

whose rate γ_t is a locally integrable function of the keys and values up to time t.

On any interval where the active key is fixed, integrating equation 3 gives $E_t = E_{t_0}A_k(1 - e^{-\bar{\gamma}})$ with $\bar{\gamma} = \int_{t_0}^{t}\gamma_s\,ds$ and effective gain $1 - e^{-\bar{\gamma}} \in [0, 1)$, the continuous-time counterpart of $\theta \in (0, 1)$. Constant gains $\theta \in (1, 2]$, including reflection, lie outside $\mathcal{D}$.

Definition 2.3 (Momentum rule)**.** The momentum rule $H(\alpha, \beta)$, with $\alpha, \beta > 0$, is the single fixed second-order system

$$\ddot{E}_t + \alpha\,\dot{E}_t + \beta\,E_t P_t = 0, \qquad \dot{E}_0 = 0. \tag{4}$$

Here α sets the decay of the momentum $\dot{E}_t$ and β weights the gradient $E_t P_t$, following continuous-time analyses of heavy-ball momentum (Gadat et al., 2018; Orvieto et al., 2020), and Appendix C.2 relates (α, β) to the discrete (η, θ).

Construction 2.1 (Reflect–project rule)**.** Reflect–project (RP) projects at the first update, reflects at the odd-numbered key changes $\tau_1, \tau_3, \ldots$, projects at the even-numbered key changes $\tau_2, \tau_4, \ldots$, and holds the memory in between. Each update multiplies the error by $A_k(1)$ or $A_k(2)$, k the active key.

The memory is queried at an independent time $Q \sim \mathrm{Exp}(\lambda)$ along the query directions in the columns of C, and the initial error E_0 is deterministic. For any rule $a \in \mathcal{D} \cup \{H, \mathrm{RP}\}$, the retrieval risk is

$$R(a; \rho) = \mathbb{E}_{X,Q}\left\|E_a(Q)\,C\right\|_F^2, \qquad X_0 \sim \pi, \tag{5}$$

the expected squared retrieval error over the Markov path X and the independent query time Q. It is a quadratic cost of a Markov jump linear system (Costa et al., 2011; 2013), which yields closed forms.

3 RETRIEVAL RISK ORDERING IN THE TWO-KEY SETTING

We specialize the model of Section 2.2 to the two nearly collinear keys of the introduction and reduce the comparison to a single scalar risk. Throughout this section the query factor is $C = I_2$ and the query rate is $\lambda = 1$. A matrix memory splits into independent rows under equation 2, so we follow one error row $f \in \mathbb{R}^2$ and take the matrix risk as the sum of the row risks.

The two-key setting. Let $\varphi = \arctan \varepsilon$ be the half-angle between the two keys, so that

$$k_+ = (\cos\varphi, \sin\varphi)^\top, \qquad k_- = (\cos\varphi, -\sin\varphi)^\top, \qquad \cos\varphi = \frac{1}{\sqrt{1+\varepsilon^2}}.$$

The active key switches between k_+ and k_- at rate ρ, following the stationary symmetric continuous-time Markov chain of Section 2.2. Write $N(Q)$ for the number of switches before the query time $Q \sim \mathrm{Exp}(1)$. Let $P_{k_+} = k_+ k_+^\top$ and $P_{k_-} = k_- k_-^\top$ be the projectors onto the two keys. Their mean $\bar{P} = \frac{1}{2}(P_{k_+} + P_{k_-}) = \mathrm{diag}(\cos^2\varphi, \sin^2\varphi)$ holds the strong direction $e_{\mathrm{s}} = (1,0)^\top$ with weight $\cos^2\varphi$ and the weak direction $e_{\mathrm{w}} = (0,1)^\top$ with weight $\sin^2\varphi = \varepsilon^2/(1+\varepsilon^2)$. Along e_{s} every rule we compare corrects the error at order one, although momentum retains a fraction $q_{\mathrm{s}} > 0$ of it (Equation (7)). Along e_{w} the rules differ at order ε^2. We therefore start from $f_0 = e_{\mathrm{w}}$ and define, for the error row f of a given rule, the weak-direction risk

$$R^{\mathrm{w}}(\rho, \varepsilon) = \mathbb{E}\,\big\|f(Q)\big\|_2^2,$$

the expected squared error at the query of the error row started along the weak axis, written R^{w}_d, R^{w}_H, and $R^{\mathrm{w}}_{\mathrm{RP}}$ for delta, momentum, and reflect–project, respectively.

A lower bound for the delta family. A projection along a key removes the error along that key and keeps the direction orthogonal to it. Across a switch, the surviving weak signal is scaled by the overlap $c = \cos 2\varphi = (1-\varepsilon^2)/(1+\varepsilon^2)$ of the two keys' untouched directions. This overlap caps how far any flow $d \in \mathcal{D}$ can correct along e_{w}.

Theorem 3.1 (Lower bound L_ε for sub-unit-gain delta flows)**.** *For $\varepsilon \in (0,1)$, every $d \in \mathcal{D}$ satisfies $R^{\mathrm{w}}_d(\rho,\varepsilon) > L_\varepsilon$, and $\inf_{d\in\mathcal{D}} R^{\mathrm{w}}_d(\rho,\varepsilon) = L_\varepsilon$, where*

$$L_\varepsilon = \cos^2\varphi\, \mathbb{E}\big[c^{2N(Q)}\big] = \frac{1+\varepsilon^2}{(1+\varepsilon^2)^2 + 4\rho\varepsilon^2} = 1 - (1+4\rho)\,\varepsilon^2 + O(\varepsilon^4), \tag{6}$$

with $\mathbb{E}\big[c^{2N(Q)}\big] = 1/(1+\rho-\rho c^2)$. The infimum is not attained: constant gains $\gamma \to \infty$ approach it, and it equals the risk of P *(Proposition 3.4).*

Mechanism. On a fixed key, every flow in $\mathcal{D}$ contracts toward the key's projection and never reflects, so along a path the weak response is a nonnegative mixture over subsequences of switches, whose most contracted term $\cos\varphi\, c^{N(Q)}$ gives the bound (Appendix D.2.1). The error of randomized Kaczmarz likewise concentrates along small singular directions (Steinerberger, 2021a).

Momentum below the delta lower bound. For $H(\alpha,\beta)$ of Definition 2.3, averaging over the switching and the query gives the leading-order risk in closed form.

Proposition 3.2 (Momentum risk R^{w}_H)**.** *With $\dot{E}_0 = 0$, sign symmetry decouples the two axes, so the risk started from each is $R_H(e_{\mathrm{s}}) = q_{\mathrm{s}} + O(\varepsilon^2)$ and $R^{\mathrm{w}}_H(\rho,\varepsilon) = 1 - \bar{c}_{\mathrm{H}}(\alpha,\beta,\rho)\,\varepsilon^2 + O(\varepsilon^4)$, where*

$$q_{\mathrm{s}} = \frac{(\alpha+1)(2\alpha+1)+2\beta}{(\alpha+1)(2\alpha+4\beta+1)}, \tag{7}$$

$$\bar{c}_{\mathrm{H}} = \frac{2\beta}{\alpha+1} - \frac{4\beta^2}{(\alpha+1)(2\alpha+4\beta+1)} \cdot \frac{z+2\alpha+2\beta+1}{z^2+\alpha z+\beta}, \qquad z = 1+2\rho. \tag{8}$$

Corollary 3.3 ($R^{\mathrm{w}}_H < L_\varepsilon$)**.** *If $\bar{c}_{\mathrm{H}}(\alpha,\beta,\rho) > 1+4\rho$, then $R^{\mathrm{w}}_H(\rho,\varepsilon) < L_\varepsilon = \inf_{d\in\mathcal{D}} R^{\mathrm{w}}_d(\rho,\varepsilon)$ for all sufficiently small $\varepsilon > 0$, since $L_\varepsilon - R^{\mathrm{w}}_H = (\bar{c}_{\mathrm{H}} - 1 - 4\rho)\,\varepsilon^2 + O(\varepsilon^4)$.*

Mechanism. Momentum carries its accumulated correction $\dot{E}_t$ across a switch, which lets $\bar{c}_{\mathrm{H}}$ exceed the lower-bound coefficient $1+4\rho$.

Reflect–project below the momentum risk. The reflect–project rule of Construction 2.1 carries no momentum, and its reflection $A_k(2) = I - 2P_k$ is one step past projection. We compare it to projecting at every switch and to momentum.

Proposition 3.4 (Reflect–project risk)**.** *Let* P *denote projection initially and at every key change. The weak-direction risks of* P *and* RP *are*

$$R^{\mathrm{w}}_{\mathrm{P}} = \frac{t_1}{1+\varepsilon^2}, \quad t_1 = \frac{1}{1+\rho\sin^2 2\varphi}, \qquad R^{\mathrm{w}}_{\mathrm{RP}} = \frac{t_2}{1+\varepsilon^2}, \quad t_2 = \frac{1+2\rho}{1+2\rho+\rho^2\sin^2 4\varphi},$$

ordered by $R^{\mathrm{w}}_{\mathrm{RP}} < R^{\mathrm{w}}_{\mathrm{P}}$ *if and only if* $2\rho\cos 4\varphi > 1$*, and* $R^{\mathrm{w}}_{\mathrm{RP}}$ *expands as*

$$R^{\mathrm{w}}_{\mathrm{RP}}(\rho,\varepsilon) = 1 - \bar{c}_{\mathrm{RP}}(\rho)\,\varepsilon^2 + O(\varepsilon^4), \qquad \bar{c}_{\mathrm{RP}}(\rho) = 1 + \frac{16\rho^2}{1+2\rho}. \tag{9}$$

Hence $R^{\mathrm{w}}_{\mathrm{P}} = L_\varepsilon$, and the comparison with momentum turns on $\bar{c}_{\mathrm{RP}}$ against $\bar{c}_{\mathrm{H}}$.

Corollary 3.5 ($R^{\mathrm{w}}_{\mathrm{RP}} < R^{\mathrm{w}}_{H}$)**.** *If* $\bar{c}_{\mathrm{RP}}(\rho) > \bar{c}_{\mathrm{H}}(\alpha,\beta,\rho)$*, then* $R^{\mathrm{w}}_{\mathrm{RP}}(\rho,\varepsilon) < R^{\mathrm{w}}_{H}(\rho,\varepsilon)$ *for all sufficiently small* $\varepsilon > 0$*, since* $R^{\mathrm{w}}_{H} - R^{\mathrm{w}}_{\mathrm{RP}} = (\bar{c}_{\mathrm{RP}} - \bar{c}_{\mathrm{H}})\,\varepsilon^2 + O(\varepsilon^4)$.

Mechanism. A completed reflect–project block scales the accumulated weak error by $\cos 4\varphi$, against $\cos^2 2\varphi$ for two projections (Lemma D.4). Reflection contracts more per block, but a query landing mid-block pays for the unfinished correction, and faster contraction pays off where $2\rho\cos 4\varphi > 1$.

The strict ordering. The two comparisons hold together on an explicit range of the switching rate, giving the strict ordering announced in the introduction.

Theorem 3.6 (Strict ordering)**.** *Fix* $\alpha, \beta > 0$ *and suppose*

$$1 + 4\rho < \bar{c}_{\mathrm{H}}(\alpha,\beta,\rho) < \bar{c}_{\mathrm{RP}}(\rho). \tag{10}$$

Then for all sufficiently small $\varepsilon > 0$*,*

$$R^{\mathrm{w}}_{\mathrm{RP}}(\rho,\varepsilon) < R^{\mathrm{w}}_{H}(\rho,\varepsilon) < \inf_{d\in\mathcal{D}} R^{\mathrm{w}}_{d}(\rho,\varepsilon) = L_\varepsilon, \tag{11}$$

and moreover $R_{\mathrm{RP}}(f_0) < R_H(f_0)$ *for every initial error* $f_0 \neq 0$*. For* $(\alpha,\beta) = (1, 30)$*, equation 10 holds on* $\rho \in [2, 5]$*, and exactly on* $(\rho_4, \rho_3) \approx (1.7933,\ 5.9115)$*, bounded by the largest real roots of two polynomials (Appendix D.2.4).*

Appendix D.2.4 proves the theorem, including the comparison for every initial error. For $H(1, 30)$ at $\rho = 3$, the leading terms are within 5% of the exact risks up to $\varepsilon = 0.03$ (Appendix D.3.1).

Scope of the fixed momentum rule. Over the momentum family, $\bar{c}_{\mathrm{H}}$ has supremum $1 + 4\rho + \frac{1}{2}[16\rho^2 + 4\rho - 1]_+ > \bar{c}_{\mathrm{RP}}(\rho)$ for every $\rho > 0$ (Proposition D.5, Figure 13), a supremum approached as $\beta \to \infty$ with a shrinking ε-range (Appendix D.3.1), so stiff momentum beats RP from e_{w}. On a finite grid this holds up to $\varepsilon = 1$, the largest value searched, at every sampled ρ (Appendix D.4.1). From e_{s} it does not: every fixed H has $q_{\mathrm{s}} > 0$, while RP removes the leading strong error. This strong margin comes from the initial projection that P and RP receive and H does not. Giving H the same initial projection (Appendix D.3.2) removes it but leaves the ordering of Theorem 3.6 intact for $H(1, 30)$, numerically: the matched weak coefficient still lies strictly between $1 + 4\rho$ and $\bar{c}_{\mathrm{RP}}$ at $\rho \in \{2, 3, 4, 5\}$, and RP still wins for every nonzero initial error at the checked $\rho \in \{2, 4, 5\}$. For stiff momentum, the same matching turns the split outcome into a win for H for every nonzero initial error. Neither side dominates the momentum family. A constant over-relaxed gain $\theta \in (1, \theta_\star)$ also lies beyond the momentum frontier (Appendix D.4.1, which gives $\theta_\star$), although no constant gain dominates RP for every nonzero initial error. This frontier is for unmatched H. Theorem 3.6 is thus a worked instance for one momentum setting. With each family tuned at fixed ε from $E_0 = -W_\star$, flow momentum beats the best constant gain under reconstruction only for $\varepsilon < \varepsilon^*(\rho) \in [0.02, 0.1]$, which decreases in ρ, and never under the associative query. The best constant gain beats RP in every tested cell, and momentum with one discrete update per key change beats or ties all three (Appendix D.4.2).

4 A RETRIEVAL RISK CRITERION IN THE K-KEY SETTING

Extending the two-key analysis to the continuous-time K-key setting ($K > 2$), we present a criterion under which the reflect–project rule achieves lower retrieval risk than the momentum rule $H(\alpha, \beta)$.

The K-key setting. We retain the reconstruction query $C = I_d$ and the query rate $\lambda = 1$. Let $u_1, \dots, u_K \in \mathbb{R}^{d-1}$ specify the weak components of the keys, and define

$$k_i(\varepsilon) = \frac{1}{\sqrt{1+\varepsilon^2\|u_i\|^2}} \begin{pmatrix} 1 \\ \varepsilon u_i \end{pmatrix}, \qquad i = 1, \dots, K. \tag{12}$$

As $\varepsilon \to 0$, all keys approach the strong direction $e_{\mathrm{s}} = (1, 0, \dots, 0)^\top \in \mathbb{R}^d$. Their components in the weak subspace $e_{\mathrm{s}}^\perp$ are of order ε. Following Section 2.2, X_t indexes the active key $k_{X_t}(\varepsilon)$ in equation 12. The stationary CTMC $(X_t)_{t\geq 0}$ has generator $\rho(T - I_K)$ on $\{1, \dots, K\}$, where T is the transition matrix with stationary law π.

Assumption 4.1 (Transition matrix T)**.** The transition matrix T is irreducible, reversible with respect to its stationary distribution π, so that $\pi_i T_{ij} = \pi_j T_{ji}$, and has zero diagonal.

Reversibility makes T self-adjoint with respect to π, so its eigenvalues are real and its eigenvectors can be chosen orthonormal under π (Levin & Peres, 2017). The access modes are the nonconstant eigenvectors $\phi_\ell \in \mathbb{R}^K$, $\ell = 1, \dots, K-1$, with $T\phi_\ell = t_\ell \phi_\ell$, $t_\ell \in [-1, 1)$, and $\sum_i \pi_i \phi_\ell(i)\phi_{\ell'}(i)$ equal to one for $\ell = \ell'$ and zero otherwise. For the weak components u_i, define their mean under π, their coefficients over the access modes, and the correlation decay rate of each mode,

$$\mu = \sum_{i=1}^K \pi_i u_i, \qquad \mu_\ell = \sum_{i=1}^K \pi_i \phi_\ell(i)\, u_i, \qquad \omega_\ell = \rho(1 - t_\ell).$$

The retrieval risk criterion. We compare $H(\alpha, \beta)$ with the projection family P and the reflect–project family RP. Both start with learning rate $1 + y$ at the first update, where $y \in [-1, 1]$ is fixed independently of the initial error f_0. At subsequent key changes, P projects, while RP alternates reflection and projection, starting with reflection. For fixed y, write $R_j(f_0)$ for the retrieval risk of $j \in \{\mathrm{P}, \mathrm{RP}\}$. Setting $y = 0$ recovers the rules of Section 3. There, from the weak initial error e_{w}, P attains the infimum of the risk over $\mathcal{D}$, which no rule in $\mathcal{D}$ attains (Theorem 3.1), so a lower risk than P in that setting is a lower risk than every $d \in \mathcal{D}$.

Lemma 4.1 (Retrieval risk coefficients)**.** *The strong-direction risk satisfies $R_j(e_{\mathrm{s}}) = s_j + O(\varepsilon^2)$ with $s_{\mathrm{P}} = y^2/(1+\rho)$ and $s_{\mathrm{RP}} = (1+2\rho)y^2/(1+\rho)^2$. For access mode ℓ, let $c_{\mathrm{H}}(\omega_\ell)$ and $c_j(t_\ell, y)$ be the coefficients of its contribution to the order-ε^2 risk reduction in the weak subspace. These coefficients are explicit rational functions of α, β, ρ, t_ℓ, and y, given in equation 25, equation 26, and equation 27 of Appendix E.1.3.*

Collecting the strong and modal coefficients, the retrieval risk of a rule a is a quadratic form $R_a(f_0) = f_0^\top \mathcal{R}_a(\varepsilon) f_0$. Its risk matrix has strong entry $s_a + O(\varepsilon^2)$, cross block $\varepsilon(s_a - 1)\mu + O(\varepsilon^3)$, and weak block $I - \varepsilon^2[(1 - s_a)\mu\mu^\top + \sum_{\ell\geq 1} c_a\, \mu_\ell \mu_\ell^\top] + O(\varepsilon^4)$ (Lemma E.2), where $(s_a, c_a) = (q_{\mathrm{s}}, c_{\mathrm{H}}(\omega_\ell))$ for $H(\alpha, \beta)$, with q_{s} from equation 7, and $(s_a, c_a) = (s_j, c_j(t_\ell, y))$ for family j.

In the two-key setting of Section 3, the single access mode has $t_1 = -1$ and $\omega_1 = 2\rho$, and with $y = 0$ the coefficients reduce to $c_{\mathrm{H}}(2\rho) = \bar{c}_{\mathrm{H}}$, $c_{\mathrm{P}}(-1, 0) = 1 + 4\rho$, and $c_{\mathrm{RP}}(-1, 0) = \bar{c}_{\mathrm{RP}}$.

Fix $j \in \{\mathrm{P}, \mathrm{RP}\}$ and $y \in [-1, 1]$, and let $D(\varepsilon) = \mathcal{R}_H(\varepsilon) - \mathcal{R}_j(\varepsilon)$ be the difference of the two risk matrices, so that $R_H(f_0) - R_j(f_0) = f_0^\top D(\varepsilon) f_0$ and family j attains lower retrieval risk than momentum for every $f_0 \neq 0$ if and only if $D(\varepsilon)$ is positive definite. Subtracting the blocks of $\mathcal{R}_H$ and $\mathcal{R}_j$ term by term introduces the strong margin g_0, the modal margins g_ℓ, and the compatibility matrix M,

$$g_0 = q_{\mathrm{s}} - s_j, \qquad g_\ell = c_j(t_\ell, y) - c_{\mathrm{H}}(\omega_\ell), \qquad M = \sum_{\ell=1}^{K-1} g_\ell \mu_\ell \mu_\ell^\top.$$

The scalar g_0 is the leading difference in strong-direction risk, and g_ℓ is the additional risk reduction of rule j over momentum in access mode ℓ. The strong margin is the same formula at the constant mode $\phi_0 \equiv 1$, with $t_0 = 1$, $\omega_0 = 0$, and $\mu_0 = \mu$, since $c_j(1, y) = 1 - s_j$ and $c_{\mathrm{H}}(0) = 1 - q_{\mathrm{s}}$. The weak block of equation 13 is thus $\varepsilon^2 \sum_{\ell=0}^{K-1} g_\ell \mu_\ell \mu_\ell^\top$ to leading order. The criterion assumes $g_0 > 0$, which holds at $y = 0$, where $s_j = 0$ and $g_0 = q_{\mathrm{s}} > 0$.

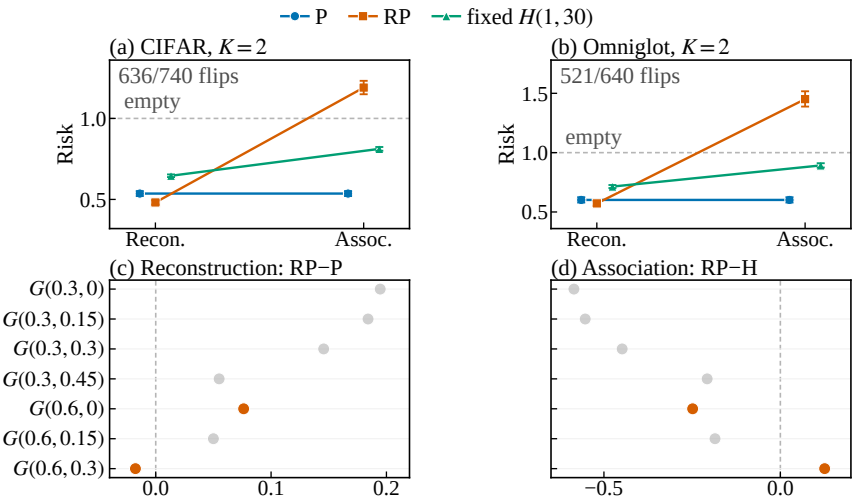

Figure 1: The query and the key geometry each reorder the rules. (a,b) Mean risk at $(\rho, \lambda) = (4, 1)$ on 740 CIFAR and 640 Omniglot two-key dictionaries, with 95% bootstrap intervals. (c,d) RP − P under reconstruction and RP − H under the associative query for the seven $K = 3$ geometries of Table 5. Orange marks $G(0.6, 0)$ and $G(0.6, 0.3)$, which share mean pairwise cosine 0.6 (Appendix A.2.1). $H(1, 30)$ is a fixed setting, and tuned comparators are in Appendix A.3.

Proposition 4.2 (Comparison criterion)**.** *For fixed $j \in \{\mathrm{P}, \mathrm{RP}\}$ and $y \in [-1, 1]$, the risk difference matrix has the block form*

$$D(\varepsilon) = \begin{pmatrix} g_0 + O(\varepsilon^2) & \varepsilon g_0 \mu^\top + O(\varepsilon^3) \\ \varepsilon g_0 \mu + O(\varepsilon^3) & \varepsilon^2 (M + g_0 \mu\mu^\top) + O(\varepsilon^4) \end{pmatrix}, \tag{13}$$

and if $g_0 > 0$ the Schur complement of its strong block is $\varepsilon^2 M + O(\varepsilon^4)$. Consequently:

(i) If $g_0 > 0$ and $\lambda_{\min}(M) > 0$, then $D(\varepsilon)$ is positive definite for all sufficiently small $\varepsilon > 0$, so family j attains strictly lower retrieval risk than $H(\alpha, \beta)$ for every nonzero initial error.

(ii) If $g_0 > 0$ and $w^\top M w < 0$ for some $w \neq 0$, then with strong component $s = -\varepsilon\, \mu^\top w + O(\varepsilon^2)$,

$$(s, w)^\top D(\varepsilon)(s, w) = \varepsilon^2\, w^\top M w + O(\varepsilon^4) < 0,$$

so momentum attains strictly lower risk along that initial error.

Since M is a sum of signed rank-one terms $g_\ell \mu_\ell \mu_\ell^\top$, a positive trace or a positive diagonal of M is compatible with a losing direction, so the comparison depends on the definiteness of M.

Mechanism. The strong margin compares what each rule concedes on the shared strong coordinate. The weak comparison runs mode by mode, and the margins g_ℓ are reassembled into M through the vectors μ_ℓ. The Schur step absorbs the coupling $g_0 \mu$, so the mean direction μ drops out of the weak criterion.

Factors governing retrieval risk. Under the reconstruction query, the criterion depends on the dictionary and the access process only through g_0 and M. The retrieval query enters through $CC^\top$, which Proposition 4.2 does not cover for $C \neq I$. Appendix B.1.1 instead shows in closed form that an associative query at the stored keys reverses P and RP for two keys. The key geometry enters through the vectors μ_ℓ. Their sum $\Sigma = \sum_i \pi_i (u_i - \mu)(u_i - \mu)^\top = \sum_\ell \mu_\ell \mu_\ell^\top$ is the covariance of the weak components, which drops the pairing of each μ_ℓ with its margin g_ℓ, so keys with the same Σ and average similarity can give a positive definite or an indefinite M (Appendix B.2). The access order enters through the eigenvalues t_ℓ, both directly and through ω_ℓ, and through π and ϕ_ℓ, which set μ_ℓ, so a kernel that revisits keys and one that moves on can give different signs of g_ℓ on the same keys. The elapsed time enters because the scalar families depend on the process through the expected update count $1 + \rho/\lambda$, while momentum is a flow and maps to $H(\alpha/\lambda, \beta/\lambda^2)$ at rate ρ/λ (Proposition B.2). At $g_0 = 0$, or when $\lambda_{\min}(M) = 0$, the leading terms leave the comparison open. Section 5 varies these factors on frozen features and specified key geometries.

Table 1: Tuned comparison across update protocols. Each family is tuned separately. Entries give the winner (δ: constant-gain delta rule, M: momentum) and the mean risk gap, or cell counts (n.s.: interval contains zero). The 16 per-step cells hold 14 distinct settings.

Protocol	Keys	Reconstruction	Associative	Appendix
Flow H, δ at key changes	CIFAR, $K = 2$	δ, 0.069	δ, 0.082	A.3
	Omniglot, $K = 3$	δ, 0.0014	δ, 0.078	A.3
	selection split, disjoint	δ in 4/4	δ in 4/4	A.3
One update per step	CIFAR, $K = 2$	M in 16/16	δ 10, M 2, n.s. 4	A.4.1
One update per token	synthetic, $K = 4, 8$	M, 0.0002–0.034	M, 0.0005–0.007	A.4.2
	frozen, $K = 4, 8$	M, 0.003–0.021	M, 0.0002–0.0024	A.4.2
Two keys: flow vs δ	78 (ρ, ε) cells	flow if $\varepsilon < \varepsilon^*(\rho)$	δ in 78/78	D.4.2
Two keys: per change vs all	78 (ρ, ε) cells	M 72, tie 6	M 72, tie 6	D.4.2

5 Factors and Validation

The criterion of Proposition 4.2 reduces the comparison between a momentum-free family and $H(\alpha, \beta)$ to the strong margin g_0 and the compatibility matrix M. On nearly collinear paths ($\varepsilon \leq 1/10$) from 768 frozen natural-feature dictionaries, where the modal margins alone set the sign (Lemma E.3), the finite-ε comparison agrees with sign $\lambda_{\min}(M)$ in every resolved case. A $K > d$ extension with mixed-sign margins (79,850 cases, Appendix A.5.2) tests the geometry: agreement is 99.99% at $\varepsilon = 0.005$, and 99.9% at $\varepsilon = 0.005$ in the 8.8% of cases where M opposes the majority of the g_ℓ, but 30% at $\varepsilon = 1$, so the criterion is a small-ε statement. Each factor of Section 4 gives a testable prediction, checked in Figures 1 and 2.

Retrieval query. Changing the query can reverse the ranking of the rules. The criterion takes $C = I$, whereas an associative query at the stored keys, as in associative-recall benchmarks (Arora et al., 2024), weights the weak directions differently. At $(\rho, \lambda) = (4, 1)$ on 740 CIFAR and 640 Omniglot two-key dictionaries (Figure 1a,b), the mean risk orders $\mathrm{RP} < \mathrm{P} < H$ under reconstruction and $\mathrm{P} < H < \mathrm{RP}$ under the associative query in both sources, so RP moves from best to worst, above the empty-memory risk of one. For two keys the risks depend only on $|k_1^\top k_2|$, and the order changes with the query exactly when $|k_1^\top k_2| > 3/4$ (Proposition B.1, Figure 3).

Key geometry. The ordering depends on the full Gram matrix G with entries $k_i^\top k_j$, not only on mean similarity. For nearly collinear keys, M pairs each μ_ℓ with its margin g_ℓ, so keys with equal average similarity and covariance can fall on opposite sides of the criterion (Appendix B.2). Beyond the nearly collinear regime, among seven prespecified $K = 3$ geometries, $G(0.6, 0.3)$ reverses the signs of $\mathrm{RP} - \mathrm{P}$ under reconstruction and $\mathrm{RP} - H$ under the associative query relative to $G(0.6, 0)$, which has the same mean similarity. The other six keep the order of $G(0.6, 0)$ (Figure 1c,d, Table 5).

Access order. The access kernel matters beyond the switching rate, since the eigenvalues t_ℓ of T enter every modal margin g_ℓ, and its modes set μ_ℓ. On nearly collinear key paths built from 512 Omniglot three-key dictionaries (Figure 2c), the uniform kernel gives $\lambda_{\min}(M) > 0$ for RP against $H(1, 30)$ at $\rho = 4$, and the weighted kernels T_a and T_b give $\lambda_{\min}(M) < 0$ on the same dictionaries. The finite-ε comparison follows this sign in every dictionary.

Elapsed time. Equal expected update counts need not give equal risk, since the scalar families depend on the process only through $1 + \rho/\lambda$ and momentum also depends on λ. On 740 CIFAR and 640 Omniglot two-key dictionaries (Figure 2a,b), the cells $(\rho, \lambda) = (1, 1)$ and $(4, 4)$ share $1 + \rho/\lambda = 2$. There the risks of P and of RP are each unchanged between the two cells, while the reconstruction risk of H rises by 0.066 on CIFAR and 0.034 on Omniglot as the query time $1/\lambda$ drops from 1 to 0.25. At $\lambda = 1$, raising ρ from 1 to 4 changes the reconstruction order from $\mathrm{P} < \mathrm{RP} < H$ to $\mathrm{RP} < \mathrm{P} < H$ in both sources. By Proposition B.2, this concerns fixed (α, β), and rescaling them with λ removes it.

Update granularity. Table 1 tunes every family under three update protocols: the delta rule wins with updates at key changes, and momentum wins under reconstruction with one update per step and

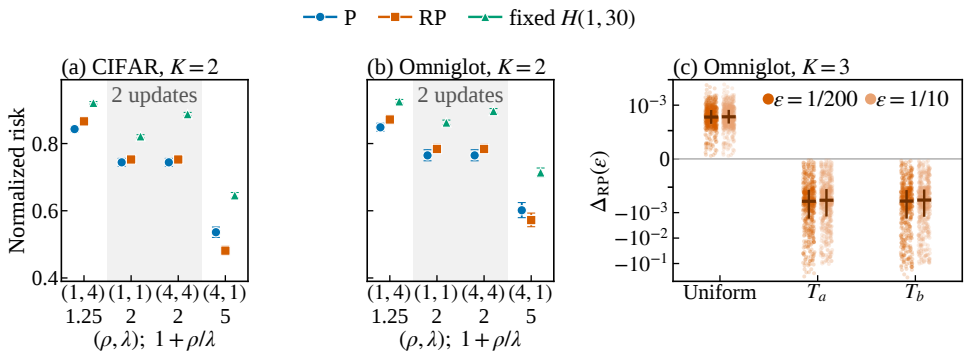

Figure 2: Elapsed time and access order move the comparison. (a,b) Normalized reconstruction risk in four rate cells on the dictionaries of Figure 1. Shaded cells share two expected updates. (c) $\Delta_{\mathrm{RP}}(\varepsilon) = \lambda_{\min}(\mathcal{R}_H - \mathcal{R}_{\mathrm{RP}})/\varepsilon^2$, positive when RP has lower risk for every nonzero initial error, at $\rho = 4$ on 512 Omniglot three-key paths (Appendix A.5.2), on a symmetric logarithmic axis. Each point is one path, and the crosses mark the median and interquartile range. It is positive on all 512 paths for the uniform kernel and negative on all 512 for T_a and T_b. Error bars in (a,b) are 95% paired bootstrap intervals.

under both queries with one update per token, though associative gaps there are small (0.3% to 4.5% on frozen features). Changing only the granularity of momentum in the two-key model gives the same reversal, whereas for $K > 2$ the keys, access, and initial error also change.

6 Limitations

The theory assumes a realizable, noiseless target (Needell, 2010), a reversible access kernel, small ε in continuous time, and one fixed $H(\alpha, \beta)$ per theorem, and Appendix C discusses the first two assumptions. The theorems concern fixed flows, not tuned discrete rules. Some tuned momentum selections lie on the searched boundary (Appendices A.3 and D.4.2), and tuned stochastic heavy-ball can accelerate least squares under anisotropic noise (Pan et al., 2024). The criterion needs the spectrum of the access kernel and the weak directions of the keys, so it cannot be evaluated online during streaming. Appendix B.3 gives only a two-dimensional example. We compare three update granularities, a flow between key changes, one update per step, and one per token, but do not sweep the number of updates per key visit, and the per-token model has no theory. Since the comparison flips between these ends, granularity should be fixed first in an application. The model omits the forgetting gates of Gated DeltaNet and Titans, and no experiment trains a sequence model. The token-level comparison of Appendix A.4.2 uses synthetic keys and frozen natural features under a fixed access process, against a constant gain rather than a per-token gain schedule. The held-out images were used in preliminary experiments, so no split is a blind holdout.

7 Conclusion

We asked when momentum helps online associative memory, and the answer depends first on when the memory updates. With one update per token, tuned momentum attains lower mean risk in every tested synthetic and frozen-feature condition. When the delta rule updates at key changes and momentum evolves continuously between them, the tuned constant-gain delta rule attains lower mean risk than tuned momentum in all four settings, also when both are tuned on a separate selection split, and in the two-key model tuned continuous-time momentum wins only for nearly identical keys. With per-step Markov updates, momentum consistently improves reconstruction, but its effect on associative recall is small and mixed. Within the continuous-time model and for nearly collinear keys, $g_0 > 0$ and a positive definite M suffice for a momentum-free rule to beat a fixed momentum rule for every nonzero initial error, whereas if M has a negative direction, momentum attains lower risk for some initial error. Since key geometry, access order, and the query move the comparison, for small K and a specified access process, exact finite-ε risks compare the rules directly for a given dictionary, and M gives the leading-order comparison as the keys approach collinearity.

AI USE STATEMENT

Claude (Anthropic) helped us brainstorm, screened our conjectures analytically, standardized our proof derivations, supplemented references on the main line of related work, and assisted with formatting. It also verified the closed forms with computer algebra and compared numerical results with the small-ε expansions.

Codex (OpenAI) parallelized our experiment scripts to shorten run time, screened conjectures numerically, suggested figure designs and reference examples, and cross-checked numerical results between two independent solvers.

The main ideas, hypotheses, experimental design, and interpretation of results were not produced by AI, and we have reviewed all AI-assisted work and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

(i) Code. Anonymized source code for the main experiments is available at `https://anonymous.4open.science/r/when-momentum-helps`. It covers Figures 1 and 2 and their factor experiments, Figure 5, the streaming tuning of Table 8 and Figure 6, the per-step and per-token comparisons of Appendices A.4.1 and A.4.2, the checks of Tables 17 and 18 and Figure 10, the matched initialization of Table 20, the fixed-ε comparison of Figure 15, and the summary in Table 1.
(ii) Model and assumptions. Section 2 defines the memory model, the update rules, the access process, and the retrieval risk. Section 6 and Appendix C discuss the assumptions.
(iii) Proofs. Appendices D and E prove the results of Sections 3 and 4, Appendix B the factor results, and Appendix C.2 the discrete correspondence.
(iv) Numerical checks. Appendix A.5.1 checks the expansions, Appendix D.3.1 the momentum frontier, and Appendix A.5.2 the criterion numerically.
(v) Experiments. Each row of Table 1 cites the appendix with its protocol, and Appendices A.2.1 and A.2.2 give the protocols for Figures 1 and 2. Appendix A.1.2 describes the data, and Table 2 records the numerical diagnostics and compute resources.

REFERENCES

Ameya Agaskar, Chuang Wang, and Yue M. Lu. Randomized Kaczmarz algorithms: Exact MSE analysis and optimal sampling probabilities. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 389–393, 2014. doi: 10.1109/globalsip.2014.7032145.

Seth J. Alderman, Roan W. Luikart, and Nicholas F. Marshall. Randomized Kaczmarz with geometrically smoothed momentum. *SIAM Journal on Matrix Analysis and Applications*, 45(4):2287–2313, 2024. doi: 10.1137/24m1633820.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Y. Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, 2024.

Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Kirby Banman, Liam Peet-Pare, Nidhi Hegde, Alona Fyshe, and Martha White. Resonance in weight space: Covariate shift can drive divergence of SGD with momentum. In *International Conference on Learning Representations*, 2022.

Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996. doi: 10.1137/s0036144593251710.

Heinz H. Bauschke and Serge G. Kruk. Reflection-projection method for convex feasibility problems with an obtuse cone. *Journal of Optimization Theory and Applications*, 120(3):503–531, 2004. doi: 10.1023/b:jota.0000025708.31430.22.

Heinz H. Bauschke, José Yunier Bello Cruz, Tran T. A. Nghia, Hung M. Phan, and Xianfu Wang. Optimal rates of linear convergence of relaxed alternating projections and generalized Douglas–Rachford methods for two subspaces. *Numerical Algorithms*, 73:33–76, 2016. doi: 10.1007/s11075-015-0085-4.

Ali Behrouz, Zeman Li, Praneeth Kacham, Majid Daliri, Yuan Deng, Peilin Zhong, Meisam Razaviyayn, and Vahab Mirrokni. ATLAS: Learning to optimally memorize the context at test time. arXiv preprint arXiv:2505.23735, 2025a.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113506–113543. Curran Associates, Inc., 2025b. doi: 10.52202/085713-3786.

Enrico Bernardi, Alberto Lanconelli, and Christopher S. A. Lauria. On the performance of stochastic gradient methods with momentum in time-varying regimes. *Applied Mathematics & Optimization*, 94(3):72, 2026. doi: 10.1007/s00245-026-10517-w.

Raghu Bollapragada, Tyler Chen, and Rachel Ward. On the fast convergence of minibatch heavy ball momentum. *IMA Journal of Numerical Analysis*, 45(3):1397–1424, 2025. doi: 10.1093/imanum/drae033.

Eduardo F. Costa, Alessandro N. Vargas, and João B. R. do Val. Quadratic costs and second moments of jump linear systems with general Markov chain. *Mathematics of Control, Signals, and Systems*, 23(1-3):141–157, 2011. doi: 10.1007/s00498-011-0064-9.

Oswaldo L. V. Costa, Marcelo D. Fragoso, and Marcos G. Todorov. *Continuous-Time Markov Jump Linear Systems*. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-34100-7.

Oswaldo Luiz Valle Costa, Ricardo Paulino Marques, and Marcelo Dutra Fragoso. *Discrete-Time Markov Jump Linear Systems*. Springer London, 2005. doi: 10.1007/b138575.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(101):1–51, 2017.

Ethan N. Epperly, Gil Goldshlager, and Robert J. Webber. Randomized Kaczmarz with tail averaging. *Applied and Computational Harmonic Analysis*, 80:101812, 2026. doi: 10.1016/j.acha.2025.101812.

Mathieu Even. Stochastic gradient descent under Markovian sampling schemes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9412–9439, 2023.

Yuguang Fang and Kenneth A. Loparo. Stabilization of continuous-time jump linear systems. *IEEE Transactions on Automatic Control*, 47(10):1590–1603, 2002. doi: 10.1109/TAC.2002.803528.

Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 658–695. PMLR, 2015.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018. doi: 10.1214/18-ejs1395.

Swetha Ganesh, Rohan Deb, Gugan Thoppe, and Amarjit Budhiraja. Does momentum help in stochastic optimization? A sample complexity analysis. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 602–612. PMLR, 2023.

Benjamin Gess and Sebastian Kassing. Exponential convergence rates for momentum stochastic gradient descent in the overparametrized setting. *Mathematical Programming*, 2026. doi: 10.1007/s10107-025-02308-y.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pp. 310–315, 2015. doi: 10.1109/ecc.2015.7330562.

Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Robert M. Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015. doi: 10.1137/15M1025487.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *International Conference on Learning Representations*, 2025.

Dhruman Gupta, Aritra Das, and Debayan Gupta. The query knows what to forget: A second erase direction for linear attention. arXiv preprint arXiv:2608.13668, 2026.

Mert Gürbüzbalaban, Asuman Ozdaglar, and Pablo A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1-2):49–84, 2021. doi: 10.1007/s10107-019-01440-w.

Jeff Z. HaoChen and Suvrit Sra. Random shuffling beats SGD after finite epochs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2624–2633. PMLR, 2019.

Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, 1949.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991.

Yulong Huang, Xiang Liu, Hongxiang Huang, Xiaopeng Lin, Zunchang Liu, Xiaowen Chu, Zeke Xie, and Bojun Cheng. MDN: Parallelizing stepwise momentum for delta linear attention. arXiv preprint arXiv:2605.05838, 2026.

Kazuki Irie, Imanol Schlag, Róbert Csordás, and Jürgen Schmidhuber. Going beyond linear Transformers with recurrent fast weight programmers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7703–7717. Curran Associates, Inc., 2021.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 545–604. PMLR, 2018.

Björn Johansson, Maben Rabi, and Mikael Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, 2009. doi: 10.1137/08073038x.

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. PMLR, 2020.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, 2018. doi: 10.1109/ita.2018.8503173.

Teuvo Kohonen. Correlation matrix memories. *IEEE Transactions on Computers*, C-21(4):353–359, 1972. doi: 10.1109/tc.1972.5008975.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016. doi: 10.1137/15m1009597.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, second edition, 2017. doi: 10.1090/mbk/107.

Xiao Li, Chengruidong Zhang, Hao Luo, Xi Lin, Zekun Wang, Zihan Qiu, Yunfei Mao, Langshi Chen, Man Yuan, Minmin Sun, Huiqiang Jiang, Siqi Zhang, Rui Men, Wei Hu, Gong Cheng, Bo Zheng, Dayiheng Liu, and Jingren Zhou. Erase-then-delta attention: Decoupling erase and write addresses in delta-rule linear attention. arXiv preprint arXiv:2606.26560, 2026.

Ji Liu and Stephen J. Wright. An accelerated randomized Kaczmarz algorithm. *Mathematics of Computation*, 85(297):153–178, 2016. doi: 10.1090/mcom/2971.

Junchen Liu, Sven Elflein, Or Litany, Zan Gojcic, and Ruilong Li. Test-time training with KV binding is secretly linear attention. arXiv preprint arXiv:2602.21204, 2026.

Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020. doi: 10.1007/s10589-020-00220-z.

Nicholas F. Marshall and Oscar Mickelin. An optimal scheduled learning rate for a randomized Kaczmarz algorithm. *SIAM Journal on Matrix Analysis and Applications*, 44(1):312–330, 2023. doi: 10.1137/22M148803X.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17309–17320. Curran Associates, Inc., 2020.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with Markovian data: Fundamental limits and algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16666–16676. Curran Associates, Inc., 2020.

Ion Necoara. Faster randomized block Kaczmarz algorithms. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1425–1452, 2019. doi: 10.1137/19m1251643.

Deanna Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010. doi: 10.1007/s10543-010-0265-5.

Deanna Needell and Rachel Ward. Two-subspace projection method for coherent overdetermined systems. *Journal of Fourier Analysis and Applications*, 19(2):256–269, 2013. doi: 10.1007/s00041-012-9248-z.

Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 155(1-2):549–573, 2016. doi: 10.1007/s10107-015-0864-7.

Cristiane Nespoli, João B. R. do Val, and Yusef Cáceres. The LQ control problem for Markovian jumps linear systems with horizon defined by stopping times. In *Proceedings of the American Control Conference*, volume 1, pp. 703–707, 2004. doi: 10.23919/ACC.2004.1383686.

James R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 1997.

Antonio Orvieto, Jonas Kohler, and Aurelien Lucchi. The role of memory in stochastic optimization. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 115 of *Proceedings of Machine Learning Research*, pp. 356–366. PMLR, 2020.

Rui Pan, Yuxing Liu, Xiaoyu Wang, and Tong Zhang. Accelerated convergence of stochastic heavy ball method under anisotropic gradient noise. In *International Conference on Learning Representations*, 2024.

Courtney Paquette and Elliot Paquette. Dynamics of stochastic momentum methods on large-scale, quadratic models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9229–9240. Curran Associates, Inc., 2021.

Vivak Patel, Mohammad Jahangoshahi, and Daniel A. Maldonado. An implicit representation and iterative solution of randomly sketched linear systems. *SIAM Journal on Matrix Analysis and Applications*, 42(2):800–831, 2021. doi: 10.1137/19M1259481.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1): 145–151, 1999. doi: 10.1016/s0893-6080(98)00116-6.

Jack Rae, Chris Dyer, Peter Dayan, and Timothy Lillicrap. Fast parametric learning with activation memorization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4228–4237. PMLR, 2018.

Sumit Roy and John J. Shynk. Analysis of the momentum LMS algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(12):2088–2098, 1990. doi: 10.1109/29.61535.

Itay Safran and Ohad Shamir. How good is SGD with random shuffling? In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3250–3284. PMLR, 2020.

Itay Safran and Ohad Shamir. Random shuffling beats SGD only after many epochs on ill-conditioned problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15151–15161. Curran Associates, Inc., 2021.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 2016.

Ali H. Sayed. *Adaptive Filters*. Wiley, 2008. doi: 10.1002/9780470374122.

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. PMLR, 2021a.

Imanol Schlag, Tsendsuren Munkhdalai, and Jürgen Schmidhuber. Learning associative inference using fast weight memory. In *International Conference on Learning Representations*, 2021b.

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

Othmane Sebbouh, Robert M. Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3935–3971. PMLR, 2021.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. DeltaProduct: Improving state-tracking in linear RNNs via Householder products. In *Advances in Neural Information Processing Systems*, volume 38, pp. 153738–153782. Curran Associates, Inc., 2025. doi: 10.52202/085713-5141.

Stefan Steinerberger. Randomized Kaczmarz converges along small singular vectors. *SIAM Journal on Matrix Analysis and Applications*, 42(2):608–615, 2021a. doi: 10.1137/20m1350947.

Stefan Steinerberger. Surrounding the solution of a linear system of equations from all sides. *Quarterly of Applied Mathematics*, 79(3):419–429, 2021b. doi: 10.1090/qam/1587.

Stefan Steinerberger. A weighted randomized Kaczmarz method for solving linear systems. *Mathematics of Computation*, 90(332):2815–2826, 2021c. doi: 10.1090/mcom/3644.

Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009. doi: 10.1007/s00041-008-9030-4.

Tao Sun, Yuejiao Sun, and Wotao Yin. On Markov chain gradient descent. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at test time): RNNs with expressive hidden states. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57503–57522. PMLR, 2025.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1139–1147. PMLR, 2013.

Neehal Tumma, Noel Loo, and Daniela Rus. Preconditioned DeltaNet: Curvature-aware sequence modeling for linear recurrences. arXiv preprint arXiv:2604.21100, 2026.

Ke Alexander Wang, Jiaxin Shi, and Emily B. Fox. Test-time regression: a unifying framework for designing sequence models with associative memory. *Journal of Machine Learning Research*, 27 (169):1–41, 2026.

Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate SGD. In *International Conference on Learning Representations*, 2024.

Bernard Widrow and Marcian E. Hoff, Jr. Adaptive switching circuits. In *IRE WESCON Convention Record*, volume 4, pp. 96–104, 1960.

Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

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*, volume 37, pp. 115491–115522. Curran Associates, Inc., 2024. doi: 10.52202/079017-3668.

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

Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016.

Yun Zeng, Deren Han, Yansheng Su, and Jiaxin Xie. On adaptive stochastic heavy ball momentum for solving linear systems. *SIAM Journal on Matrix Analysis and Applications*, 45(3):1259–1286, 2024. doi: 10.1137/23M1575883.

Yifan Zhang, Steve Ta, Jasper Zhang, Jichen Feng, Shuzhen Li, Yongxin Zhang, Yifeng Liu, Huizhuo Yuan, Mengdi Wang, Quanquan Gu, and Andrew Chi-Chih Yao. Fast weight attention for continual learning. arXiv preprint arXiv:2608.27763, 2026.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P. Foster, and Sham M. Kakade. The benefits of implicit regularization from SGD in least squares problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5456–5468. Curran Associates, Inc., 2021.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Benign overfitting of constant-stepsize SGD for linear regression. *Journal of Machine Learning Research*, 24(326):1–58, 2023.

Anastasios Zouzias and Nikolaos M. Freris. Randomized extended Kaczmarz for solving least squares. *SIAM Journal on Matrix Analysis and Applications*, 34(2):773–793, 2013. doi: 10.1137/120889897.

CONTENTS OF THE APPENDIX

A Experimental details **17**

A.1 Common setting, data, and encoders 17

A.2 Protocols for Figures 1 and 2 20

A.3 Tuned comparators 24

A.4 Update granularity: per-step and per-token rules 27

A.5 Numerical checks of the expansions and the criterion 34

B Results for the four factors **40**

B.1 Retrieval query and elapsed time 40

B.2 Key geometry and access order 41

B.3 The online rule in the indefinite regime 42

C Model assumptions and auxiliary results **44**

C.1 Model assumptions and auxiliary lemmas 44

C.2 Discrete momentum and the momentum rule 45

D Proofs for the two-key setting **47**

D.1 Two-key geometry 47

D.2 Proofs of the results of Section 3 47

D.3 The momentum family and matched initialization 53

D.4 Constant gains and tuned families at fixed ε 56

E Proofs for the K-key setting **60**

E.1 Access modes and the risk matrix 60

E.2 Proofs of Lemma 4.1 and Proposition 4.2 62

E.3 Positive and indefinite compatibility matrices 64

F Extended related work **66**

A EXPERIMENTAL DETAILS

A.1 COMMON SETTING, DATA, AND ENCODERS

A.1.1 COMMON SETTING

Keys, values, and initialization. Each dictionary contains K unit keys $k_i \in \mathbb{R}^d$ and values $v_i = e_i \in \mathbb{R}^K$, where e_i is the ith coordinate vector. The keys are linearly independent, so their similarity matrix G, with entries $G_{ij} = k_i^\top k_j$, is invertible. We use the teacher

$$W_\star = \sum_{i,j=1}^{K} (G^{-1})_{ij}\, e_i k_j^\top, \qquad W_\star k_i = v_i, \qquad W_0 = 0, \qquad E_0 = W_0 - W_\star.$$

Natural-image keys are normalized features from a frozen encoder. Synthetic dictionaries specify G directly. Average similarity denotes the mean pairwise cosine $2\sum_{i<j} k_i^\top k_j/[K(K-1)]$. The active index X_t starts uniformly over the keys. Its holding times are exponential with rate ρ, and its next index follows the access kernel T, whose diagonal is zero. The kernels of Figures 1, 2(a,b), and 4 have uniform stationary distribution $\pi_i = 1/K$, and Appendix A.5.2 describes the kernels of Figure 2(c). The query time $Q \sim \mathrm{Exp}(\lambda)$ is independent of the access process.

Update rules. For a unit key k_i, a residual update with gain θ acts as

$$W^+ = W + \theta(v_i - Wk_i)k_i^\top, \qquad E^+ = E(I - \theta P_{k_i}), \qquad P_{k_i} = k_i k_i^\top.$$

The projection rule P uses $\theta = 1$ at the first update and at every key change. Reflect–project, denoted RP, uses $\theta = 1$ at the first update, $\theta = 2$ at odd-numbered key changes, and $\theta = 1$ at even-numbered key changes. Both rules keep W constant between key changes. These choices set $y = 0$ in the families of Section 4. The momentum comparator is $H(1, 30)$, implemented as

$$\dot{W}_t = S_t, \qquad \dot{S}_t = -\alpha S_t - \beta(W_t k_{X_t} - v_{X_t})k_{X_t}^\top, \qquad (\alpha, \beta) = (1, 30), \qquad S_0 = 0.$$

The state $S_t \in \mathbb{R}^{K\times d}$ persists across key changes. Since $v_i = W_\star k_i$, these equations give equation 4 with $\dot{E}_0 = 0$. The expected update count for P and RP is $1 + \rho/\lambda$, including the first update. The expected duration of the momentum flow is $1/\lambda$.

Queries and normalization. For a rule a, we evaluate the squared retrieval risk

$$R(a; \rho) = \mathbb{E}\|E_a(Q)C\|_F^2,$$

where the columns of C specify the query directions. The reconstruction query uses $C = I_d$. Its dictionary-specific initial risk is $R_{\mathrm{init}} = \|E_0\|_F^2 = \|W_\star\|_F^2$, and its normalized risk is $R(a; \rho)/R_{\mathrm{init}}$. We compute this ratio within each dictionary and then average the ratios with equal dictionary weights. An associative query samples a stored key uniformly, so

$$CC^\top = \frac{1}{K}\sum_{i=1}^{K} k_i k_i^\top, \qquad R(a; \rho) = \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}\|W_a(Q)k_i - v_i\|_2^2.$$

We report this associative risk on its absolute scale. Its initial value is one because $W_0 = 0$ and $\|v_i\|_2 = 1$. Thus both displayed risks start at one, with reconstruction normalized separately for each dictionary. These metrics quantify squared retrieval error in the controlled memory model.

Expected-risk computation. We solve linear equations for the full expectation over the access process and query time. For P and RP, let $N(Q)$ be the number of key changes before the query. The independent exponential clocks give

$$\Pr\{N(Q) = n\} = \frac{\lambda}{\lambda + \rho}\left(\frac{\rho}{\lambda + \rho}\right)^n, \qquad n = 0, 1, 2, \ldots.$$

Conditioning on the next query or key change yields linear equations for the expected quadratic cost at each active key. For RP, the state also records whether the next update is a reflection or a projection. Solving these equations sums over every possible number of key changes. For H, we

Table 2: Recorded numerical diagnostics and resources. Memory is peak resident memory in decimal megabytes. Residuals are relative, and error estimates are posterior floating-point estimates. A dash marks a quantity that was not logged.

Experiment	Worst residual	Worst error estimate	Wall (s)	Peak (MB)
CIFAR held-out pairs (740), continuous time	—	—	439.67[a]	85.77[a]
CIFAR held-out pairs (740), discrete steps	6.369×10^{-17}	—	439.67[a]	85.77[a]
Omniglot pairs (640), main solver	9.708×10^{-17}	—	128.73	116.30
Omniglot pairs (640), second path	2.028×10^{-16}	—	16.18	182.99
CIFAR training-split pairs (640)	1.024×10^{-16}	—	130.21	115.01
Four-key access, CIFAR (256)	8.638×10^{-17}	—	165.43	188.08
Four-key access, Omniglot (256)	7.547×10^{-17}	—	160.41	210.00
Sign check (768 dictionaries)	9.288×10^{-17}	—	696.54	111.02
Tuned comparators, 25-point grid	9.791×10^{-17}	2.585×10^{-6}	96.36	95.65
Tuned comparators, 49-point grid	9.791×10^{-17}	3.253×10^{-5}	165.10	191.50
Tuned comparators, final validation[c]	—	—	453.37	433.23
Tuned comparators, final evaluation[c]	—	—	28.01	104.09
Per-step tuning, common protocol	—	—	3344.19	229.32
Class-disjoint splits[c]	—	—	515.73	771.72
Synthetic geometry $G(0.6, 0)$	—	1.284×10^{-13}	0.90[b]	—
Synthetic geometry $G(0.6, 0.3)$	—	5.620×10^{-13}	0.90[b]	—

[a]The continuous and discrete results of the 740 CIFAR pairs come from one run, whose time and memory appear in both rows. [b]Time of the joint run over all seven valid geometries of Appendix A.2.1. [c]Final successful stage of each.

augment the error with its derivative and solve the second-moment equations of the Markov jump linear system. These are coupled Lyapunov equations with the exponential query discount λ. The initial derivative is zero when we evaluate the resulting quadratic form.

All states remain in the span of the keys, so we evaluate these equations in an orthonormal basis of that span. The reconstruction cost becomes the identity on this span and preserves the full Frobenius risk. With K linearly independent keys, the systems have K^3, $2K^3$, and $4K^3$ unknowns for P, RP, and H, respectively. Reported risks are exact expectations of the model evaluated by floating-point linear solves. The computation uses neither trajectory sampling nor time truncation.

Numerical precision and resources. Table 2 records the numerical diagnostics and resources of the experiments, taken from the run logs. Every linear solve behind a reported risk passes the relative residual threshold 10^{-10}, and no computation uses a GPU.

A.1.2 Data, encoders, and sample reuse

Throughout the paper, "CIFAR" refers to CIFAR-100.

Encoders. Both datasets use the same three-layer convolutional encoder, with 3 input channels for CIFAR-100 and 1 for Omniglot. The three 3×3 convolutions (padding 1) have 16, 32, and 32 output channels, each followed by a ReLU. The first two are followed by 2×2 max pooling, and the network ends with global average pooling and flattening. The 32-dimensional key is the encoder output itself, not the output of a classification or projection head. During training, a linear classification head is attached and the encoder and head are trained jointly with the categorical cross-entropy over the fit classes, using AdamW with learning rate 3×10^{-4}, weight decay 0.01, batch size 64, 5 epochs, a fixed seed, and no data augmentation. For feature extraction the head is discarded and every nonzero feature is normalized to a unit key. CIFAR-100 images enter as RGB values divided by 255. Omniglot images are converted to grayscale, resized bilinearly to 28×28, and divided by 255.

Data splits. Both splits are by class. For CIFAR-100, the 100 fine classes are ordered by the SHA-256 hash of their names: the first 60 classes, taken from the official training set, fit the encoder. The next 20, also from the official training set, form a selection split, and the last 20, taken from the official test set, form the held-out split. These splits contain 30,000, 10,000, and 2,000 images. Removing every member of each group of exact pixel duplicates leaves 29,986,

9,995, and 1,993 images, so the held-out split has 1,993 images. For Omniglot, we keep the official `images_background`/`images_evaluation` boundary. The 30 background alphabets are ordered by the SHA-256 hash of their names. The first 24 fit the encoder and the remaining 6 form the selection split. All 20 evaluation alphabets form the held-out split. These splits contain 14,960, 4,320, and 13,180 images. The held-out classes of both datasets are therefore disjoint from the classes that fit the encoder. The selection splits are used only for the selection-split check of Appendix A.3. We had used both held-out splits in preliminary experiments before drawing the samples below, so neither split is untouched by method development. The rules that select the CIFAR pairs and regroup the image banks below use image indices or identifiers only and apply no label, geometry, feature, or performance filter.

CIFAR held-out pairs. Figures 1(a) and 2(a) use the same 740 disjoint pairs of CIFAR-100 images from the held-out split, which contains 1993 images. A separate sample of 512 held-out images, grouped into 256 pairs, serves as the evaluation set of Appendices A.3 and A.5.2. Excluding these 512 images leaves 1481 eligible images. We order them by a fixed pseudorandom permutation, obtained by hashing each image index with SHA-256 under a prefix fixed before feature extraction, pair adjacent images among the first 1480, and discard the last image. The code release contains this prefix and the selected image indices of every sample below.

The frozen CIFAR encoder supplies all features for these pairs. Inference uses RGB pixel values divided by 255. Each feature has 32 coordinates, and we normalize it to unit length in float64. The selected images are disjoint from the 256-pair sample and from the training split. The unit of comparison is one image pair, with the same pair used for every rule, query, and time cell. The query and time panels, the tuned comparators of Appendix A.3, and the discrete rules of Appendix A.4.1 therefore contribute repeated evaluations of these 740 pairs.

CIFAR training-split images. A second CIFAR bank consists of 1536 distinct images from the training split of the same encoder. None of them coincides with an image of the 740 held-out pairs, by index or by pixel content. Because the encoder was trained on these images, this bank describes features seen during training. It supplies the four-key dictionaries of Figure 4 and the two-key replication of Appendix A.2.3.

Omniglot held-out images. The Omniglot bank consists of 1536 distinct images from the Omniglot held-out split, grouped into 512 triples. A frozen Omniglot encoder, trained on the Omniglot training split, supplies normalized 32-dimensional keys. Both the dataset and the encoder differ from CIFAR. None of the selected images coincides with a training image, by index or by pixel content. The 512 triples serve as three-key dictionaries for the tuned comparators of Appendix A.3 and the sign check of Appendix A.5.2.

Regrouping into dictionaries of size K. For $K \in \{2, 4\}$, we keep the fixed image order of each 1536-image bank and split it into adjacent groups of K images, which gives 768 pairs and 384 quadruples. Within each bank and K, we order the dictionaries by a hash of their ordered image identifiers and reserve the first 128 for validation, which these experiments leave unused. The remaining 640 pairs and 256 quadruples form the evaluation sets of Figures 1(b), 2(b), and 4, and of Appendix A.2.3. The same images appear in the triples, pairs, and quadruples of a bank, so results across K are dependent.

Synthetic geometries. The synthetic experiment specifies eight three-key geometries directly through their similarity matrices. Seven matrices are positive definite and define valid dictionaries. The remaining matrix is not a valid similarity matrix and is not evaluated. The sample unit is one specified geometry. This experiment uses no image dataset or encoder.

Dictionaries for the tuned comparators. The tuned comparisons of Appendix A.3 use the 740 CIFAR held-out pairs above for validation and the 256-pair sample for evaluation. The two sets share no images, and neither shares images with the training split. For Omniglot, the 512 triples above, all drawn from the held-out alphabets, are split into 128 validation and 384 evaluation dictionaries. The two parts share no images but share 233 character classes, so this is not a class-disjoint validation, and through these shared classes 511 of the 512 triples form one connected group. The bootstrap intervals therefore describe variation within these dependent banks.

Table 3: Order of the three rules by key-cosine interval at $(\rho, \lambda) = (4, 1)$, with the number n of dictionaries in each interval. The boundaries are the three crossing cosines 0.750, 0.799, and 0.965.

Source	$\lvert k_1^\top k_2 \rvert$ interval	n	Reconstruction order	Associative order	Changes
CIFAR	$[0, 0.750)$	104	$\mathrm{P} < \mathrm{RP} < H$	$\mathrm{P} < \mathrm{RP} < H$	0
	$[0.750, 0.799)$	58	$\mathrm{RP} < \mathrm{P} < H$	$\mathrm{P} < \mathrm{RP} < H$	**58**
	$[0.799, 0.965)$	446	$\mathrm{RP} < \mathrm{P} < H$	$\mathrm{P} < H < \mathrm{RP}$	**446**
	$[0.965, 0.999)$	132	$\mathrm{RP} < H < \mathrm{P}$	$\mathrm{P} < H < \mathrm{RP}$	**132**
	$[0.999, 1)$	0	$\mathrm{RP} < H < \mathrm{P}$	$\mathrm{P} < H < \mathrm{RP}$	0
Omniglot	$[0, 0.750)$	119	$\mathrm{P} < \mathrm{RP} < H$	$\mathrm{P} < \mathrm{RP} < H$	0
	$[0.750, 0.799)$	41	$\mathrm{RP} < \mathrm{P} < H$	$\mathrm{P} < \mathrm{RP} < H$	**41**
	$[0.799, 0.965)$	240	$\mathrm{RP} < \mathrm{P} < H$	$\mathrm{P} < H < \mathrm{RP}$	**240**
	$[0.965, 0.999)$	222	$\mathrm{RP} < H < \mathrm{P}$	$\mathrm{P} < H < \mathrm{RP}$	**222**
	$[0.999, 1)$	18	$\mathrm{RP} < H < \mathrm{P}$	$\mathrm{P} < H < \mathrm{RP}$	**18**

Table 4: Mean risk at $(\rho, \lambda) = (4, 1)$ in Figure 1(a,b), with 95% paired bootstrap intervals. Reconstruction risk is normalized by the initial risk of each dictionary, and associative risk is absolute. Both risks equal one for the empty memory $W_0 = 0$, so values above one are worse than the empty memory. For P with two keys, values e_i, and $E_0 = -W_\star$, both risks equal $\frac{1}{2}(1 + g^2)\,\mathbb{E}[g^{2N(Q)}]$ with $g = \lvert k_1^\top k_2 \rvert$ and $N(Q)$ the number of key changes before the query, so the two entries coincide exactly. Bold marks the lowest mean in each row.

Source	Query	P	RP	$H(1, 30)$
CIFAR	reconstruction	0.5361 [0.5207, 0.5516]	**0.4813** [0.4682, 0.4945]	0.6460 [0.6379, 0.6543]
CIFAR	associative	**0.5361** [0.5207, 0.5516]	1.1903 [1.1492, 1.2320]	0.8119 [0.8004, 0.8236]
Omniglot	reconstruction	0.6015 [0.5794, 0.6244]	**0.5723** [0.5525, 0.5931]	0.7140 [0.7016, 0.7272]
Omniglot	associative	**0.6015** [0.5794, 0.6244]	1.4509 [1.3883, 1.5172]	0.8922 [0.8746, 0.9108]

A.2 Protocols for Figures 1 and 2

A.2.1 Protocol for Figure 1

Panels (a,b): retrieval query. Panel (a) uses the 740 CIFAR held-out pairs and panel (b) the 640 Omniglot pairs of Appendix A.1.2, at $\rho = 4$ and $\lambda = 1$. The active key switches to the other key at every transition. Each dictionary uses identical keys, values, initialization, and update rules for both queries. Table 4 lists the mean risks. The intervals come from 10,000 bootstrap resamples of whole dictionaries, with one set of resampled indices shared by all rules and queries of a source. A dictionary changes order when the ranking of P, RP, and $H(1, 30)$ under the reconstruction query differs from the ranking under the associative query, with differences within 10^{-10} counted as ties. This happens in 636 of the 740 CIFAR dictionaries and 521 of the 640 Omniglot dictionaries.

For two keys with $E_0 = -W_\star$, every risk depends on the dictionary only through $\lvert k_1^\top k_2 \rvert$: evaluating the closed forms of Proposition B.1 for P and RP and the exact solver for $H(1, 30)$ on a grid of cosines reproduces every dictionary's risk to within 3×10^{-11} (Figure 3). The orders therefore change only at three cosines (Table 3): $\mathrm{P} = \mathrm{RP}$ under reconstruction at $3/4$, as Proposition B.1(i) gives for $(\rho, \lambda) = (4, 1)$, $\mathrm{RP} = H$ under the associative query at 0.799, and $\mathrm{P} = H$ under reconstruction at 0.965. A dictionary changes order exactly when $\lvert k_1^\top k_2 \rvert > 3/4$, so the counts above count cosines above that threshold.

For Omniglot dictionaries 211 and 531, the dominant left eigenvector of the RP second-moment operator misses the residual threshold 10^{-10} in 16 conditions. For these conditions we evaluate the risk by a direct solve of the forward equations in 120-digit arithmetic, whose residual is below 10^{-120}. The spectral abscissa is zero in both computations, at distance at least one from λ. Excluding the two dictionaries gives 521 of 638 and the same mean orders.

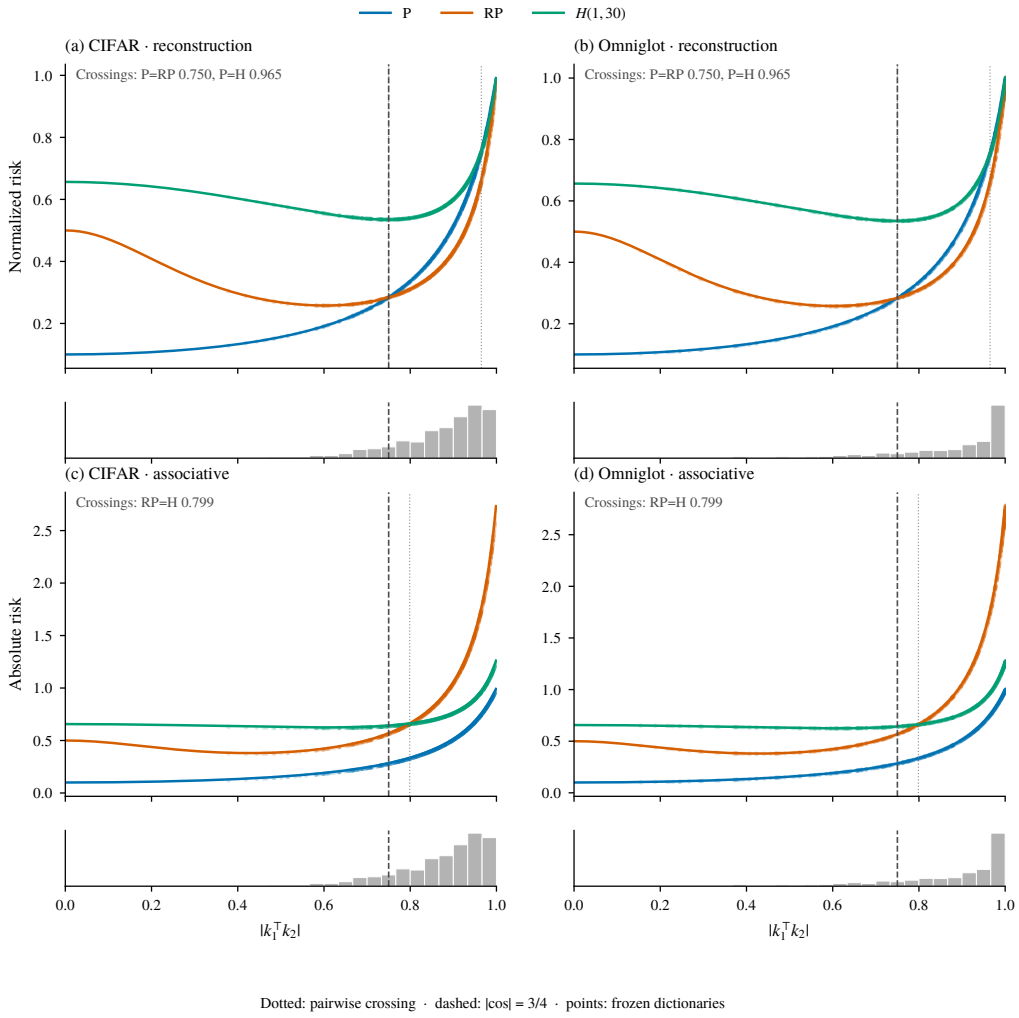

Figure 3: Two-key risk against the key cosine at $(\rho, \lambda) = (4, 1)$. Curves are the closed forms of Proposition B.1 for P and RP and exact solves for $H(1, 30)$. The risks of the 740 CIFAR and 640 Omniglot dictionaries lie on these curves to within 3×10^{-11} and are not drawn separately. The histograms show their cosines.

The dictionary-level result counts the conjunction of two comparisons: $\mathrm{RP} < H(1, 30)$ under reconstruction and $\mathrm{P} < \mathrm{RP}$ under the associative query. Both inequalities hold together in all 740 dictionaries. The sign convention treats a risk difference below -10^{-10} as negative. Proposition B.1(ii) implies that, in this setting, $\mathrm{P} < \mathrm{RP}$ under the associative query holds for every two-key dictionary. The reconstruction comparison between RP and $H(1, 30)$ is evaluated numerically on these dictionaries.

Panels (c,d): key geometry. We construct three-key dictionaries from the similarity matrix

$$G(c, \delta) = \begin{pmatrix} 1 & c + \delta & c - \delta \\ c + \delta & 1 & c \\ c - \delta & c & 1 \end{pmatrix},$$

where c is the mean pairwise cosine and δ redistributes similarity among the three pairs. The prespecified grid uses $c \in \{0.3, 0.6\}$ and $\delta \in \{0, 0.15, 0.30, 0.45\}$. Seven of the eight matrices are positive definite and were evaluated. The matrix $G(0.6, 0.45)$ has off-diagonal entry $1.05 > 1$ and is indefinite, so it is excluded from the risk evaluation.

Table 5: All eight three-key geometries $G(c, \delta)$ of Appendix A.2.1, at $(\rho, \lambda) = (4, 1)$. Columns give $G(c, \delta)$, and rows c_{ij} give the pairwise cosines. Subscripts R and A denote the normalized reconstruction risk and the absolute associative risk. Both equal one for the empty memory. $H = H(1, 30)$. The last geometry is indefinite and is not evaluated. Asterisks mark the two geometries of Figure 1(c,d). Bold marks the two sign reversals.

	$c = 0.3$				$c = 0.6$			
δ	0	0.15	0.3	0.45	0^*	0.15	0.3^*	0.45
c_{12}	0.30	0.45	0.60	0.75	0.60	0.75	0.90	1.05
c_{13}	0.30	0.15	0.00	−0.15	0.60	0.45	0.30	0.15
c_{23}	0.30	0.30	0.30	0.30	0.60	0.60	0.60	0.60
P_R	0.2581	0.2778	0.3534	0.5737	0.3669	0.4432	0.8167	–
RP_R	0.4527	0.4618	0.4990	0.6287	0.4431	0.4932	0.7991	–
H_R	1.0339	0.9997	0.9074	0.8163	0.7849	0.7690	0.8591	–
P_A	0.2639	0.2788	0.3330	0.4749	0.4089	0.4548	0.6641	–
RP_A	0.5253	0.5563	0.6755	1.0458	0.7432	0.8528	1.4897	–
H_A	1.1103	1.1094	1.1239	1.2533	0.9918	1.0379	1.3641	–
$\mathrm{RP}_R - \mathrm{P}_R$	0.1946	0.1841	0.1456	0.0550	0.0762	0.0500	**−0.0176**	–
$\mathrm{RP}_A - H_A$	−0.5850	−0.5531	−0.4484	−0.2075	−0.2486	−0.1852	**0.1255**	–

Table 6: The four time-experiment cells. The update count applies to P and RP and includes the first update. Figure 2(a,b) displays all four cells.

Cell (ρ, λ)	$(1, 1)$	$(1, 4)$	$(4, 1)$	$(4, 4)$
Expected update count $1 + \rho/\lambda$	2	1.25	5	2
Mean query time $1/\lambda$	1	0.25	1	0.25

Table 5 lists all eight geometries. The two marked with an asterisk are displayed in the figure. Both have mean pairwise cosine 0.6, with different off-diagonal entries in G. We use $\rho = 4$, $\lambda = 1$, and uniform transitions to either of the other two keys. The rules, values, and initialization follow Appendix A.1.1. Panel (c) plots the normalized reconstruction-risk difference $\mathrm{RP} - \mathrm{P}$, which changes from 0.076222 to -0.017594 between the two geometries in table order. Panel (d) plots the absolute associative-risk difference $\mathrm{RP} - H(1, 30)$, which changes from -0.248618 to 0.125545. Each point is the expected-risk difference for its specified geometry. Among the seven evaluated geometries, only $G(0.6, 0.3)$ gives $\mathrm{RP} < \mathrm{P}$ under reconstruction and $\mathrm{RP} > H(1, 30)$ under the associative query. The other six keep $\mathrm{P} < \mathrm{RP}$ and $\mathrm{RP} < H(1, 30)$. The sign change therefore rests on one of the seven geometries. $H(1, 30)$ exceeds the empty-memory risk under the associative query in six of them and under reconstruction in one.

A.2.2 Protocol for Figure 2

Panels (a,b): elapsed time. We evaluate the 740 CIFAR pairs and the 640 Omniglot pairs in the four rate cells listed in Table 6, with the same keys, values, and initial states in every cell and both queries. The cell $(4, 1)$ reproduces the setting of Figure 1(a,b).

The cells $(1, 1)$ and $(4, 4)$ share $\rho/\lambda = 1$. By Proposition B.2(i), P and RP have exactly the same risk in these two cells for each dictionary and each query. Their computed maximum relative deviation is zero. For $H(1, 30)$, Proposition B.2(ii) gives the effective parameters α/λ and β/λ^2 after rescaling the query rate to one. Thus the risk depends on the elapsed-time scale as well as ρ/λ. From $(1, 1)$ to $(4, 4)$, the mean normalized reconstruction risk of $H(1, 30)$ changes from 0.8215 to 0.8875 on CIFAR and from 0.8626 to 0.8969 on Omniglot, and its mean associative risk changes from 1.0931 to 0.8616 on CIFAR and from 1.1480 to 0.8683 on Omniglot.

Panel (c): access kernel on nearly collinear paths. Panel (c) displays $\Delta_{\mathrm{RP}}(\varepsilon)$ from the sign check of Appendix A.5.2 for the 512 Omniglot triples at $\rho = 4$, under the uniform kernel and the weighted kernels T_a and T_b, at $\varepsilon = 1/200$ and $1/10$. The same dictionaries enter all three kernels. Table 18 gives the sign of $\lambda_{\min}(M)$ in each cell.

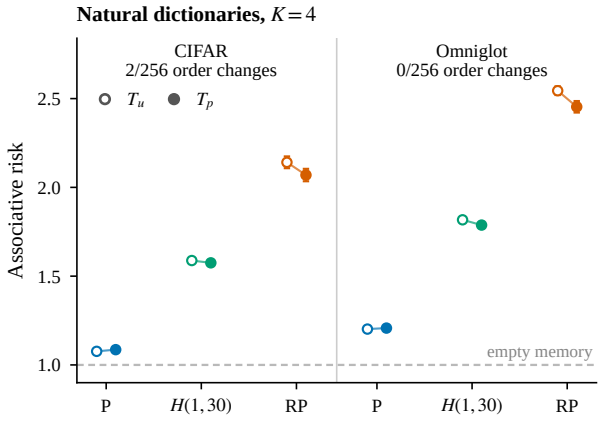

Figure 4: Associative risk at $\rho = 4$ on 256 CIFAR training-split (features of images the encoder was trained on) and 256 held-out Omniglot four-key dictionaries under the uniform kernel T_{u} and the pair kernel T_{p}. The dashed line is the empty-memory risk. Error bars are 95% paired bootstrap intervals.

Table 7: Mean associative risk on 256 four-key dictionaries per source in Figure 4, with 95% paired bootstrap intervals. CIFAR dictionaries come from the encoder's training split. The empty memory $W_0 = 0$ has associative risk one, so every entry is worse than the empty memory. Bold marks the lowest mean in each row.

Source	ρ	Kernel	P	$H(1,30)$	RP
CIFAR	1	T_{u}	**1.2225** [1.2101, 1.2343]	1.6484 [1.6403, 1.6567]	2.1073 [2.0785, 2.1353]
CIFAR	1	T_{p}	**1.2229** [1.2108, 1.2347]	1.6270 [1.6186, 1.6355]	2.0803 [2.0510, 2.1093]
CIFAR	4	T_{u}	**1.0765** [1.0617, 1.0912]	1.5879 [1.5775, 1.5987]	2.1411 [2.1080, 2.1744]
CIFAR	4	T_{p}	**1.0860** [1.0716, 1.1003]	1.5750 [1.5649, 1.5854]	2.0695 [2.0343, 2.1044]
Omniglot	1	T_{u}	**1.2816** [1.2688, 1.2944]	1.7694 [1.7637, 1.7753]	2.2776 [2.2483, 2.3069]
Omniglot	1	T_{p}	**1.2809** [1.2680, 1.2938]	1.7461 [1.7402, 1.7522]	2.2460 [2.2146, 2.2778]
Omniglot	4	T_{u}	**1.2019** [1.1888, 1.2152]	1.8173 [1.8044, 1.8303]	2.5444 [2.5200, 2.5690]
Omniglot	4	T_{p}	**1.2075** [1.1943, 1.2207]	1.7876 [1.7763, 1.7990]	2.4538 [2.4213, 2.4862]

Access kernel on natural dictionaries. We evaluate the 256 CIFAR training-split quadruples and the 256 Omniglot quadruples of Appendix A.1.2 under the uniform kernel T_{u} and the pair kernel T_{p} of Appendix B.2, at $\rho \in \{1, 4\}$ and $\lambda = 1$, with the associative query. The four images of each quadruple occupy the states 1 to 4 in stored order, so T_{p} mostly moves within the first two and within the last two images. Both kernels are symmetric and satisfy Assumption 4.1 with uniform π. We use four keys because for $K = 3$ the uniform kernel is the only reversible kernel with uniform stationary distribution and zero diagonal. Table 7 lists the mean risks. The order $\mathrm{P} < H(1,30) < \mathrm{RP}$ holds in all eight cells. Replacing T_{u} by T_{p} changes the order of the three rules in 0 and 2 of the 256 CIFAR dictionaries at $\rho = 1$ and $\rho = 4$, and in 2 and 0 of the 256 Omniglot dictionaries.

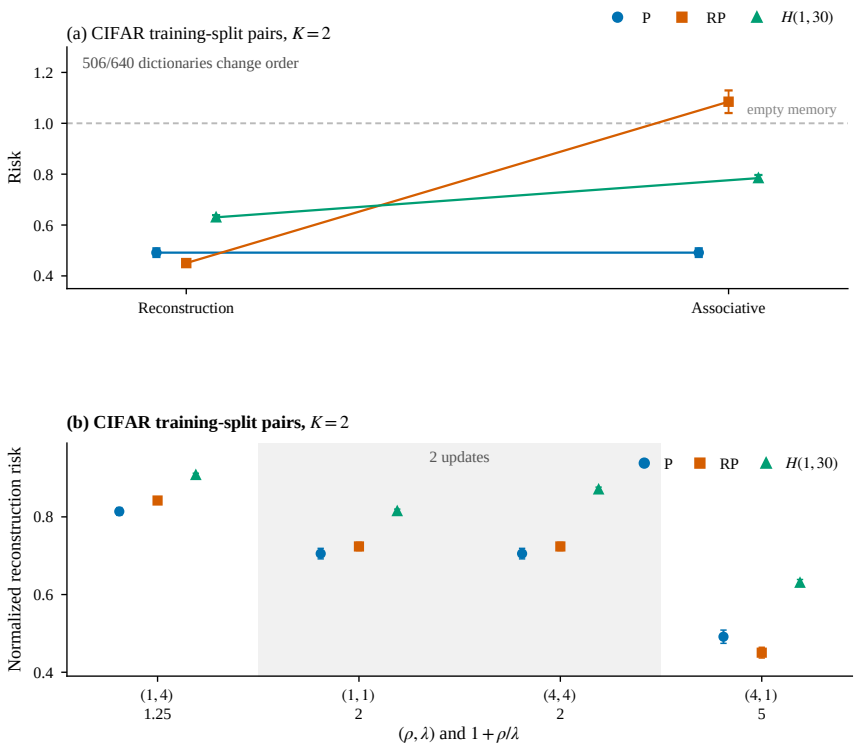

Figure 5: Query and time panels on the 640 CIFAR training-split pairs. (a) Mean risk at $(\rho, \lambda) = (4, 1)$ under the two queries, as in Figure 1(a). (b) Normalized reconstruction risk in the four rate cells, as in Figure 2(a). Error bars are 95% paired bootstrap intervals.

A.2.3 Replication on the CIFAR training-split bank

We repeat the query and time panels on the 640 CIFAR training-split pairs of Appendix A.1.2, with the protocols of Appendix A.2.1 and Appendix A.2.2. The encoder was trained on these images. At $(\rho, \lambda) = (4, 1)$ (Figure 5a), the mean risk orders $\mathrm{RP} < \mathrm{P} < H(1, 30)$ under reconstruction, with means 0.4504, 0.4915, and 0.6303, and $\mathrm{P} < H(1, 30) < \mathrm{RP}$ under the associative query, with means 0.4915, 0.7845, and 1.0848. The order changes with the query in 506 of the 640 dictionaries. In the cells $(1, 1)$ and $(4, 4)$ (Figure 5b), P and RP have equal risk, while the normalized reconstruction risk of $H(1, 30)$ changes from 0.8148 to 0.8706.

A.3 Tuned comparators

Figures 1 and 2 compare the fixed rules P, RP, and $H(1, 30)$. To test whether tuning the momentum rule changes this comparison, we select the parameters of $H(\alpha, \beta)$ and of the delta rule with a constant learning rate θ, applied at the first update and at every key change, on validation dictionaries, and we report risks on separate evaluation dictionaries. All other settings follow Appendix A.1.1, with $\rho = 4$, $\lambda = 1$, the uniform access kernel, and both queries.

Data and selection. For CIFAR ($K = 2$), the 740 dictionaries of Appendix A.2.1 serve for validation, and the separate 256-pair held-out sample of Appendix A.1.2 serves for evaluation. The two sets share no images but are drawn from the same 20 held-out classes. For Omniglot ($K = 3$), the 512 dictionaries of Appendix A.1.2 are split by a hash of the dictionary identifier into 128 validation and 384 evaluation dictionaries, which share no images and share classes. A candidate is admissible when, on every validation dictionary, the spectral abscissa s of its second-moment generator satisfies $s < \lambda - 0.05$, which makes its risk finite, and its linear solves pass the residual test of Appendix A.1.1. For each source and query we select the admissible candidate

Table 8: Mean risk on the evaluation dictionaries, with 95% paired bootstrap intervals below each mean. Lower is better within a column. Reconstruction risk is divided by the initial risk of each dictionary, and associative risk is the absolute expected squared retrieval error. CIFAR has 256 and Omniglot 384 evaluation dictionaries. Intervals describe variation across the dictionaries of each bank. Values above one are worse than the empty memory.

	CIFAR, $K = 2$		Omniglot, $K = 3$	
Rule	reconstruction	associative	reconstruction	associative
empty memory ($W_0 = 0$)	1	1	1	1
P	0.5150	0.5150	0.9939	0.9997
	[0.4881, 0.5420]	[0.4876, 0.5430]	[0.9932, 0.9945]	[0.9832, 1.0165]
RP	0.4688	1.1463	0.9919	2.2249
	[0.4459, 0.4924]	[1.0735, 1.2221]	[0.9910, 0.9928]	[2.1898, 2.2607]
$H(1, 30)$	0.6407	0.8010	0.9933	1.4740
	[0.6266, 0.6555]	[0.7808, 0.8222]	[0.9925, 0.9940]	[1.4651, 1.4829]
tuned $H(\alpha, \beta)$	0.5100	0.5087	0.9924	0.7749
	[0.4846, 0.5355]	[0.4870, 0.5307]	[0.9915, 0.9932]	[0.7699, 0.7802]
delta rule, tuned θ	**0.4406**	**0.4271**	**0.9909**	**0.6966**
	[0.4183, 0.4634]	[0.4122, 0.4421]	[0.9899, 0.9919]	[0.6906, 0.7029]

with the lowest mean validation risk. The learning rate ranges over $\theta \in \{0.05, 0.10, \ldots, 2\}$. An earlier version of this comparison used a 25-point grid $\alpha \in \{\frac{1}{4}, \ldots, 4\} \times \beta \in \{1, \ldots, 100\}$ and a 49-point extension, and every selected H lay on the boundary of both. We therefore start from $\alpha \in \{2^{-5}, \ldots, 2^5\} \times \beta \in \{\frac{1}{2}, 1, 3, 10, 30, 100, 200, 500, 10^3, 2 \cdot 10^3, 5 \cdot 10^3, 10^4\}$ and, while a selected H lies on an outer edge, extend that edge by one halving or doubling, for at most three rounds and using validation risks only. The final grid is $\alpha \in \{2^{-5}, \ldots, 2^8\} \times \beta \in \{\frac{1}{2}, \ldots, 10^4\}$, and Figure 6 shows the validation risk surfaces. Intervals are percentile 95% intervals from 10,000 paired bootstrap resamples of whole evaluation dictionaries.

Results. Validation selects $H(32, 5000)$ and $H(256, 5000)$ for CIFAR, and $H(4, 200)$ and $H(256, 500)$ for Omniglot, under the reconstruction and associative queries, and $\theta = 1.5$, 0.65, 1.65, and 0.4 in the column order of Table 8. The two reconstruction selections are interior to the final grid. Both associative selections lie on its $\alpha = 256$ edge, and the CIFAR column reached the three-round cap, so the comparison does not establish an optimum over the momentum family. The direction of that edge is informative: the damping ratios $\alpha/(2\sqrt{\beta})$ of the associative selections are 1.8 and 5.7, against 0.23 and 0.14 for the reconstruction selections. As $\alpha \to \infty$ with $\beta/\alpha = \gamma$ fixed, equation 4 reduces to the first-order flow $\dot{E}_t = -\gamma E_t P_t$, a member of $\mathcal{D}$. Under the associative query, validation therefore drives momentum toward the overdamped limit, in which it becomes a delta rule with gain below one. This agrees with the selected $\theta < 1$ for that query. The delta rule with the selected θ attains the lowest mean risk in every column. Its paired differences to the tuned H are -0.069, -0.082, -0.0014, and -0.078, with 95% intervals $[-0.074, -0.064]$, $[-0.089, -0.075]$, $[-0.0016, -0.0013]$, and $[-0.079, -0.077]$. RP attains lower mean risk than the tuned H under the reconstruction query and higher risk under the associative query, and the associative query ranks P ahead of RP, as Proposition B.1 predicts for two keys. The selected learning rate exceeds one under the reconstruction query and lies below one under the associative query. These intervals are conditional on the fixed banks, whose Omniglot validation and evaluation parts share character classes. Table 10 repeats the comparison on class-disjoint splits.

Class-disjoint splits. We repartition the same frozen image banks by complete Omniglot alphabets and complete CIFAR-100 fine classes before building new dictionaries. Omniglot has 127 validation dictionaries from six alphabets and 379 evaluation dictionaries from 14 alphabets, each three-key dictionary using three distinct characters of one alphabet. CIFAR-100 has 745 validation pairs from 15 fine classes and 250 evaluation pairs from five fine classes, each pair containing two images of one class. Selection uses the final grid above on the new validation dictionaries, separately for each source and query. The primary interval resamples whole alphabets or classes, keeping every dictionary of a sampled cluster and weighting the mean by dictionary count. Table 10 reports the results. The paired difference of the delta rule to the selected H lies below zero in all four columns, with cluster intervals $[-0.081, -0.077]$, $[-0.098, -0.084]$, $[-0.0012, -0.0007]$, and $[-0.079, -0.075]$

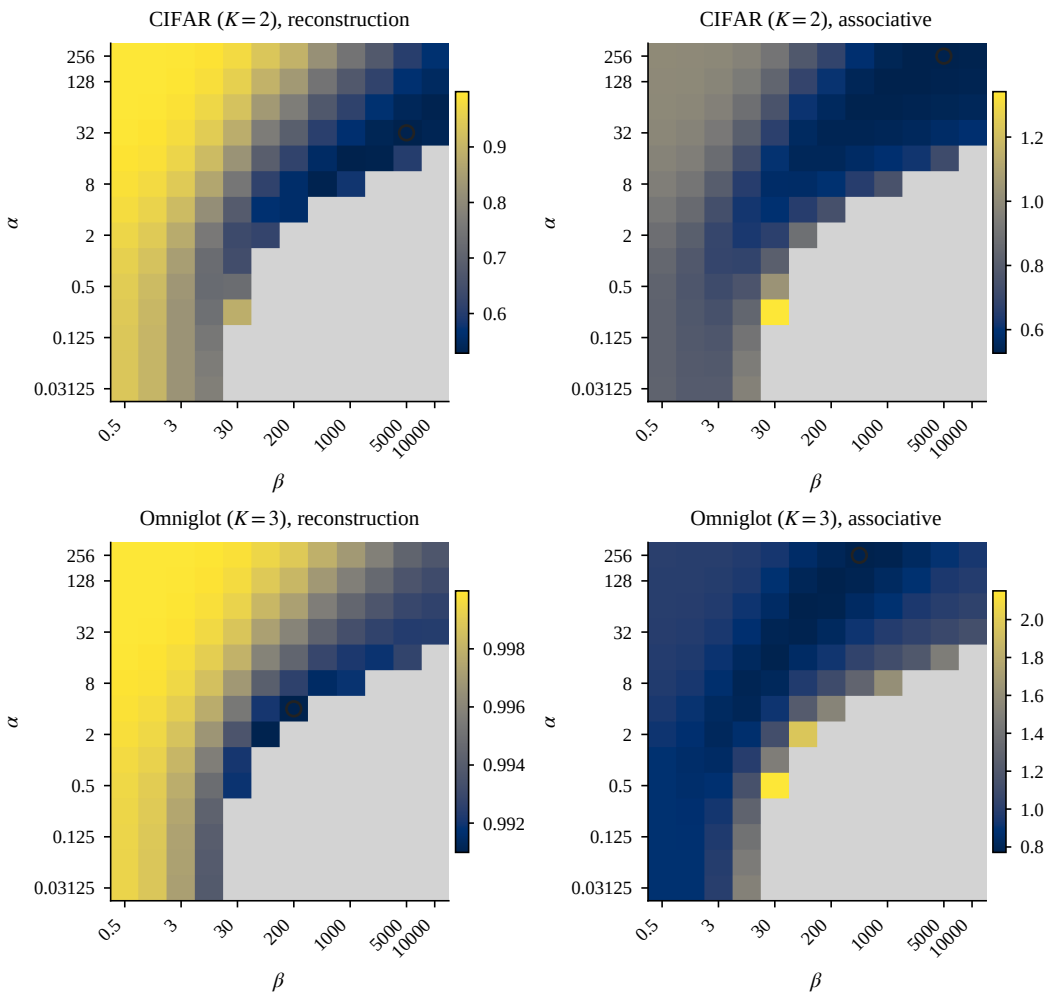

Figure 6: Mean validation risk of $H(\alpha, \beta)$ on the final grid, with a separate continuous color scale in each panel. Grey cells failed admission on at least one validation dictionary, through infinite risk or an unresolved spectral margin or numerical check. Open circles mark the selected $H(\alpha, \beta)$. The CIFAR panels use 740 and the Omniglot panels 128 validation dictionaries. No evaluation risk was used to draw these surfaces or to choose the parameters.

in the column order of the table. RP minus the selected H is -1.46×10^{-4} under the Omniglot reconstruction query, with cluster interval $[-3.55 \times 10^{-4}, 5.43 \times 10^{-5}]$, which crosses zero, and the CIFAR reconstruction interval of P minus the selected H also crosses zero. The images had been exposed to method development, so this is not a blind holdout. The split and the construction of dictionaries within a cluster change together, so differences from Table 8 cannot be attributed to class disjointness alone. CIFAR has only five evaluation clusters, which makes its cluster intervals coarse. The delta rule nevertheless has lower mean risk than the selected H in every cluster, 5 of 5 CIFAR classes and 14 of 14 Omniglot alphabets, under both queries. Table 9 lists the selections. Only the CIFAR associative selection lies on the $\alpha = 256$ edge. The other three are interior.

Selection-split tuning. The selections above use held-out validation dictionaries. We therefore repeat the selection on the selection splits of Appendix A.1.2 alone, with 740 CIFAR pairs and 128 Omniglot triples built by the same rules, the same grids and admission rule, and at most three edge extensions, and evaluate the selected rules on both frozen evaluation sets. The same code first reproduces the selections and risks of Tables 8 and 10. The delta rule attains the lowest mean risk in all eight columns, and every interval for the delta rule minus the tuned H lies below zero (Table 11). No dictionary is a tie. The reconstruction selections match the earlier ones up to θ, whereas the

Table 9: Class-disjoint validation selections. Edge refers to the full final grid: $\alpha \in [1/32, 256]$, $\beta \in [0.5, 10000]$, $\theta \in [0.05, 2]$ at their listed grid points.

Source	Query	α	β	θ	α edge	β edge	θ edge
Omniglot $K = 3$	recon.	2	100	1.65	interior	interior	interior
Omniglot $K = 3$	assoc.	128	200	0.35	interior	interior	interior
CIFAR $K = 2$	recon.	8	500	1.5	interior	interior	interior
CIFAR $K = 2$	assoc.	256	2000	0.6	upper	interior	interior

Table 10: Class-disjoint evaluation. Omniglot uses 379 three-key dictionaries in 14 evaluation alphabets and CIFAR-100 uses 250 two-key dictionaries in five evaluation fine classes. Each cell reports the mean risk and, below it, the 95% interval from 10,000 paired resamples of complete alphabets or classes. Lower is better. Values above one are worse than the empty memory.

	CIFAR, $K = 2$		Omniglot, $K = 3$	
Rule	reconstruction	associative	reconstruction	associative
empty memory ($W_0 = 0$)	1	1	1	1
P	0.5777	0.5777	0.9953	1.0316
	[0.5087, 0.6466]	[0.5087, 0.6466]	[0.9947, 0.9959]	[1.0045, 1.0644]
RP	0.5232	1.3168	0.9937	2.2759
	[0.4682, 0.5791]	[1.1389, 1.4947]	[0.9929, 0.9945]	[2.2147, 2.3502]
$H(1, 30)$	0.6743	0.8490	0.9947	1.4640
	[0.6427, 0.7077]	[0.8005, 0.8979]	[0.9940, 0.9954]	[1.4523, 1.4779]
tuned $H(\alpha, \beta)$	0.5714	0.5490	0.9939	0.7881
	[0.5189, 0.6238]	[0.5070, 0.5910]	[0.9930, 0.9946]	[0.7817, 0.7956]
delta rule, tuned θ	**0.4921**	**0.4584**	**0.9930**	**0.7114**
	[0.4384, 0.5461]	[0.4235, 0.4934]	[0.9920, 0.9938]	[0.7032, 0.7209]

associative selections move further in the overdamped direction and stop at the extension cap, so the comparison is again bounded by the grid.

Omniglot reconstruction plateau. The normalized reconstruction risks of the Omniglot three-key dictionaries stay near 0.99 for every rule, against 0.57 to 0.71 for Omniglot pairs (Table 4), and Table 12 traces this to the key geometry. Normalization divides by $\|W_\star\|_F^2 = \mathrm{tr}(G^{-1})$. Across the 384 evaluation triples, the median of $\lambda_{\min}(G)$ is 5.9×10^{-4}, and the weakest eigendirection of G carries a median 99.6% of the initial error but only a median 2.0×10^{-4} of the associative query weight. For the six reconstruction rules evaluated on the earlier grids, the mean normalized risk lies in $[0.9912, 0.9939]$ and the median fraction of the weakest-direction error that survives to the query lies in $[0.9959, 0.9975]$. The Omniglot and CIFAR pairs have median $\lambda_{\min}(G)$ of 0.069 and 0.111, although their median pairwise cosines, 0.93 and 0.89, are close to that of the triples, 0.90. The third, highly correlated key thus creates a much weaker direction that holds almost all of the normalized initial error and is barely corrected before a random query, while the associative query gives that direction little weight and separates the rules more. This is a geometric account. We did not decompose the associative risk by direction.

A.4 Update granularity: per-step and per-token rules

A.4.1 Discrete update rules

Setting. We reuse the 740 CIFAR pairs of Appendix A.1.2 in the four rate cells of Table 6, with both queries. For a step $h \in \{0.1, 0.05, 0.025, 0.0125\}$, the nth update happens at time nh. Between two updates, the active key moves by the transition matrix $e^{h\rho(T-I)}$, so the two keys switch with probability $(1 - e^{-2\rho h})/2$. After each update, the memory is queried with probability $1 - e^{-\lambda h}$. P projects at every update. RP projects at the first update and afterwards updates only at key changes, alternating reflection and projection. The momentum rules use the discrete update of Proposition C.4 with $\eta = 1 - \alpha h$, $\theta = \beta h^2$, and $S_0 = 0$. We compare $H(1, 30)$ with $H(4, 100)$ for the reconstruction query and $H(4, 30)$ for the associative query, the selections of the earlier 25-point grid of Appendix A.3. The final grid there selects different parameters. These parameters were

Table 11: Selection-split tuning. Both rules are selected only on the selection splits of Appendix A.1.2 (740 CIFAR pairs, 128 Omniglot triples) and evaluated on the frozen evaluation dictionaries of Table 8 and Table 10. The last two columns give the mean paired difference, delta rule minus tuned H, in units of 10^{-2}, with 95% intervals from 10,000 paired resamples of dictionaries (Table 8) or class clusters (Table 10). "Edge" marks selections that stop at the grid-extension cap.

Source	query	$H(\alpha, \beta)$	θ	grid	$\delta - H$, held-out	$\delta - H$, disjoint
CIFAR, $K = 2$	rec.	$(32, 5000)$	1.45	interior	-7.01 $[-7.41, -6.60]$	-7.56 $[-8.35, -6.78]$
	assoc.	$(2048, 80000)$	0.70	edge	-8.05 $[-8.88, -7.26]$	-9.92 $[-11.88, -7.97]$
Omniglot, $K = 3$	rec.	$(4, 200)$	1.60	interior	-0.14 $[-0.16, -0.13]$	-0.10 $[-0.12, -0.09]$
	assoc.	$(2048, 5000)$	0.45	edge	-7.73 $[-7.82, -7.64]$	-7.58 $[-7.73, -7.41]$

Table 12: Gram geometry of the evaluation banks: medians of the smallest eigenvalue of G, of $\mathrm{tr}(G^{-1})$, of the share w_{init} of the initial reconstruction error in the weakest eigendirection, of the associative query weight q_{weak} of that direction, and of the pairwise cosine.

Bank	N	$\lambda_{\min}(G)$	$\mathrm{tr}(G^{-1})$	w_{init}	q_{weak}	cosine
Omniglot, $K = 3$	384	0.000592	1.70×10^3	0.996	0.000197	0.897
CIFAR, $K = 2$	256	0.111	9.52	0.944	0.0556	0.889
Omniglot, $K = 2$	640	0.0693	15.0	0.965	0.0346	0.931

selected on the same 740 pairs, so the comparison with them is in-sample. Each risk is the fixed point of the exact-expectation recursion, and all 11,840 combinations of pair, cell, and step give finite risk for every rule. The intervals come from 10,000 paired bootstrap resamples of the pairs.

Results. Figure 7 shows the risk of each rule against h. Over rules, cells, and queries, the largest mean gap between a discrete risk and its continuous-time value is 0.173, 0.093, 0.048, and 0.025 at $h = 0.1, 0.05, 0.025$, and 0.0125. Under the reconstruction query, RP has higher risk than $H(4, 100)$ at $h = 0.1$ in all four cells, by 0.016 to 0.041, and lower risk at $h = 0.0125$ in all four, by 0.004 to 0.068, with every paired interval excluding zero. The sign changes at a cell-dependent step in between. Against $H(1, 30)$, RP attains lower reconstruction risk at every step and in every cell. Under the associative query, RP attains higher risk than both momentum rules at every step and in every cell. The final-grid selections of Appendix A.3 are stiff, and their direct discretization is unstable at most steps, so this comparison keeps the earlier selections. It only tests how fixed flows discretize. The per-step tuning below supersedes it as a comparison of rules, and Appendix A.4.2 tunes per token.

Common discrete protocol with per-step tuning. The comparison above maps a fixed flow to each step and evaluates in sample. We therefore repeat it under one discrete protocol for all rules. At every step, the key moves by $e^{h\rho(T-I)}$, every rule has one update opportunity, and the memory is queried after the step with probability $1 - e^{-\lambda h}$. The cell $(4, 4)$ at step h coincides with $(1, 1)$ at step $4h$, so the 16 cells contain 14 distinct settings. P and RP act at the first step and at observed key changes, the delta rule updates at every step with a gain from $\{0.05, 0.10, \ldots, 2\}$, and discrete momentum $S \leftarrow \eta S + \theta U$ updates at every step with $\eta \in \{0, 0.5, 0.8, 0.9, 0.95, 0.98, 0.99\}$ and θ from the same 40 values, tuned directly rather than through (α, β). For each step $h \in \{0.1, 0.05, 0.025, 0.0125\}$, rate cell of Table 6, and query, we select parameters on the 740 CIFAR pairs and report exact expected risks on the 256 disjoint evaluation pairs of Appendix A.3, with 10,000 paired bootstrap resamples of whole pairs. Candidates whose stability could not be resolved on every validation pair were excluded from selection.

Figure 8 shows RP minus the selected momentum rule. Under the reconstruction query, RP attains lower risk only at $(\rho, \lambda) = (4, 1)$ with $h = 0.025$ and 0.0125, by 0.017 $[0.015, 0.018]$ and 0.025 $[0.021, 0.029]$. Momentum attains lower risk in the other 14 cells, and in all 16 cells under the associative query. On this grid, the selected η is zero in all 16 associative cells and in 4 of the 16 reconstruction cells, so the winning rule is often a per-step delta rule. Two ablations probe the first step. If momentum also projects at its first step, RP leads in three reconstruction cells instead of two. If RP omits its first projection, it leads in none. The first update therefore matters for the ordering. The direct discrete grid, the first-step ablation, and the change of evaluation pairs alter the comparison

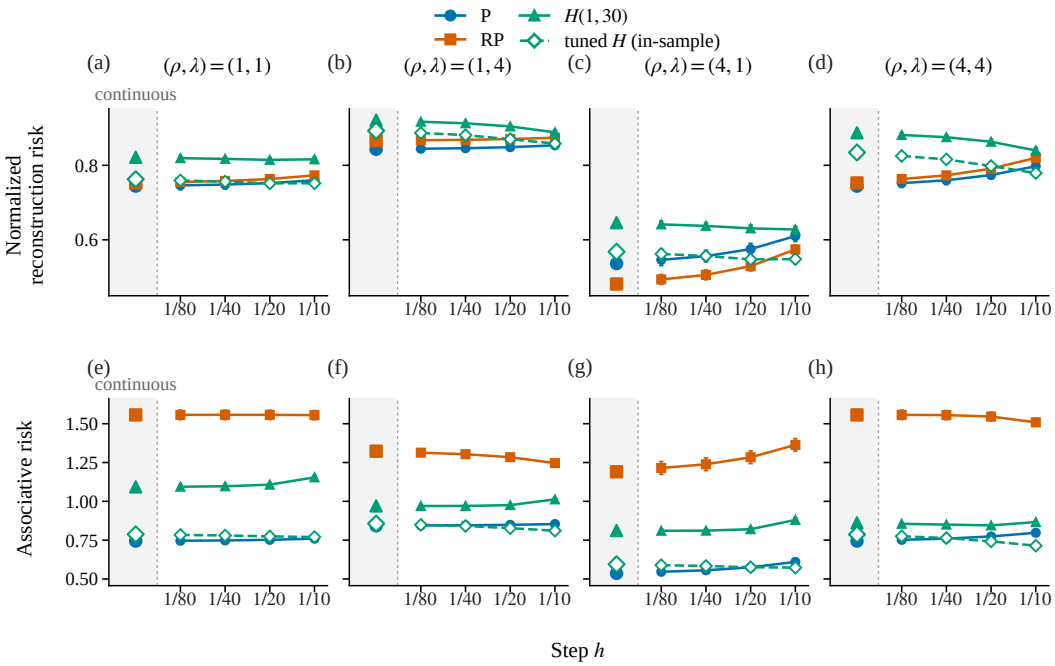

Figure 7: Risk of the discrete rules on the 740 CIFAR pairs against the step h, in the four rate cells of Table 6. The top row shows normalized reconstruction risk and the bottom row associative risk. The shaded column gives the continuous-time risk. The earlier grid selected $H(4, 100)$ in the top row and $H(4, 30)$ in the bottom row, selected on the same pairs. Error bars are 95% paired bootstrap intervals, most of them narrower than the markers.

together, so these results do not attribute the step-size reversal of the fixed-flow comparison above to the parameter mapping alone. The continuous-time advantage of RP in Section 3 is thus a property of the flow comparison and of event-triggered updates, and it does not survive per-step tuning at the step sizes tested here except in one rate cell.

Per-step delta rule. To separate momentum from the update granularity, we add the delta rule tuned per step on the same θ grid and validation pairs (Table 14). Under the reconstruction query and on this grid ($\eta = 0$ or $\eta \geq 0.5$), the selected momentum rule, with $\eta \in \{0.5, 0.8\}$, attains lower risk than the per-step delta rule in 12 of the 16 cells, by 0.002 to 0.041 with every interval above zero. In the four cells with $(\rho, \lambda) = (1, 4)$, validation selects $\eta = 0$ and the two rules coincide. Under the associative query, validation on this grid selects $\eta = 0$ in all 16 cells, and the best rule with $\eta \geq 0.5$ has higher risk than the per-step delta rule by 0.008 to 0.042. RP attains lower risk than the per-step delta rule only at $(\rho, \lambda) = (4, 1)$ under reconstruction, by 0.013 to 0.041. The cells $(4, 4)$ at step h and $(1, 1)$ at step $4h$ coincide, as the time scaling of Proposition B.2 predicts for a discrete protocol. Per-step updating therefore accounts for the associative advantage of the selected momentum rule over RP, and momentum adds a further gain only under reconstruction.

Small momentum. The grids above contain no η between 0 and 0.5. We therefore add $\eta \in \{0.05, 0.1, 0.2, 0.3\}$, with the same 40 values of θ, select the best rule M_+ with $\eta > 0$ on the same validation pairs, and evaluate only the selected rule on the 256 evaluation pairs (Table 13). The earlier candidates and their selections are reproduced exactly. Under the reconstruction query, the four $(\rho, \lambda) = (1, 4)$ cells now select $\eta \in \{0.2, 0.3\}$ and beat the per-step delta rule by 0.0005 to 0.0020, so momentum attains lower reconstruction risk than the per-step delta rule in all 16 cells. Under the associative query, validation selects $\eta \in \{0.05, 0.1\}$ with small θ, and the gap to the per-step delta rule shrinks to at most 0.004 in either direction: the delta rule is lower in 10 cells, momentum in 2, and 4 intervals contain zero. On these two-key CIFAR pairs, small momentum therefore does not improve associative risk over the per-step delta rule.

Table 13: Best momentum rule M_+ with $\eta \in \{0.05, 0.1, 0.2, 0.3, 0.5, 0.8, \ldots, 0.99\}$ on the 256 CIFAR evaluation pairs, against the per-step delta rule D_{step}. Differences are means with 95% paired bootstrap intervals. Positive values favor M_+. No evaluation pair has an absolute difference below 10^{-12}.

Query	(ρ, λ)	h	selected η	selected θ	$D_{\text{step}} - M_+$ [95% CI]	tie frac.
reconstruction	$(1, 1)$	0.1	0.5	0.9	0.0096 [0.0082, 0.0108]	0.000
associative	$(1, 1)$	0.1	0.05	0.2	-0.0010 [-0.0013, -0.0006]	0.000
reconstruction	$(1, 4)$	0.1	0.3	1.05	0.0020 [0.0016, 0.0023]	0.000
associative	$(1, 4)$	0.1	0.05	0.35	-0.0026 [-0.0028, -0.0023]	0.000
reconstruction	$(4, 1)$	0.1	0.5	1.1	0.0410 [0.0367, 0.0452]	0.000
associative	$(4, 1)$	0.1	0.05	0.4	0.0001 [-0.0004, 0.0005]	0.000
reconstruction	$(4, 4)$	0.1	0.5	1.1	0.0144 [0.0118, 0.0168]	0.000
associative	$(4, 4)$	0.1	0.05	0.4	-0.0024 [-0.0031, -0.0018]	0.000
reconstruction	$(1, 1)$	0.05	0.5	0.85	0.0058 [0.0050, 0.0065]	0.000
associative	$(1, 1)$	0.05	0.05	0.1	0.0000 [-0.0003, 0.0003]	0.000
reconstruction	$(1, 4)$	0.05	0.2	1.1	0.0015 [0.0013, 0.0016]	0.000
associative	$(1, 4)$	0.05	0.05	0.2	-0.0028 [-0.0037, -0.0019]	0.000
reconstruction	$(4, 1)$	0.05	0.5	1	0.0314 [0.0286, 0.0340]	0.000
associative	$(4, 1)$	0.05	0.05	0.3	-0.0012 [-0.0020, -0.0004]	0.000
reconstruction	$(4, 4)$	0.05	0.5	1	0.0135 [0.0114, 0.0154]	0.000
associative	$(4, 4)$	0.05	0.05	0.3	-0.0023 [-0.0031, -0.0015]	0.000
reconstruction	$(1, 1)$	0.025	0.5	0.8	0.0032 [0.0028, 0.0036]	0.000
associative	$(1, 1)$	0.025	0.05	0.05	0.0005 [0.0002, 0.0008]	0.000
reconstruction	$(1, 4)$	0.025	0.2	1.05	0.0009 [0.0008, 0.0010]	0.000
associative	$(1, 4)$	0.025	0.05	0.1	-0.0036 [-0.0053, -0.0019]	0.000
reconstruction	$(4, 1)$	0.025	0.8	0.25	0.0200 [0.0150, 0.0248]	0.000
associative	$(4, 1)$	0.025	0.05	0.2	-0.0014 [-0.0025, -0.0002]	0.000
reconstruction	$(4, 4)$	0.025	0.5	0.9	0.0096 [0.0082, 0.0108]	0.000
associative	$(4, 4)$	0.025	0.05	0.2	-0.0010 [-0.0013, -0.0006]	0.000
reconstruction	$(1, 1)$	0.0125	0.5	0.8	0.0017 [0.0015, 0.0019]	0.000
associative	$(1, 1)$	0.0125	0.05	0.05	-0.0011 [-0.0014, -0.0008]	0.000
reconstruction	$(1, 4)$	0.0125	0.2	1.05	0.0005 [0.0004, 0.0005]	0.000
associative	$(1, 4)$	0.0125	0.1	0.05	0.0013 [0.0008, 0.0017]	0.000
reconstruction	$(4, 1)$	0.0125	0.8	0.2	0.0162 [0.0133, 0.0190]	0.000
associative	$(4, 1)$	0.0125	0.05	0.15	-0.0002 [-0.0005, 0.0002]	0.000
reconstruction	$(4, 4)$	0.0125	0.5	0.85	0.0058 [0.0050, 0.0065]	0.000
associative	$(4, 4)$	0.0125	0.05	0.1	0.0000 [-0.0003, 0.0003]	0.000

A.4.2 TOKEN-LEVEL DISCRETE COMPARISON

Setting. This experiment removes the continuous-time dwell. Every token triggers exactly one update, and every rule sees the same token stream. Each dictionary has K unit keys $k_i \propto s\, e_1 + \sqrt{1 - s^2}\, g_i$ in $\mathbb{R}^{32}$, with $g_i \sim N(0, I/32)$, so s controls how closely the keys align. The next key is drawn uniformly from the other $K - 1$ keys (*move-on*), or the current key repeats with probability $1/2$ (*revisit*). The query follows each token with probability $p_Q = 1/(4K)$, so on average there are $4K$ updates before the query. The initial error is isotropic on the key span. The reconstruction risk $\|E\|_F^2$ is divided by the dimension of that span. The associative risk is $K^{-1} \sum_i \|Ek_i\|^2$. Both risks are 1 for the empty memory. All risks are exact expectations of the second-moment recursion over the geometric query time, and a Monte Carlo check with 2×10^4 paths agrees to three decimals.

Rules and selection. P is the delta rule with $\theta = 1$. RP projects at the first token, then alternates reflection and projection at key changes and holds on repeats. The delta rule tunes $\theta \in \{0.05, 0.10, \ldots, 2\}$. Momentum, $S \leftarrow \eta S + \theta U$ and $W \leftarrow W + S$, tunes $\eta \in \{0, 0.05, 0.1, 0.2, \ldots, 0.9, 0.95\}$ together with the same θ grid, so the momentum family contains the delta rule at $\eta = 0$. Each rule is then refined on the validation dictionaries over a window of ± 0.1 around its coarse optimum in steps of 0.01, in θ and, for momentum, in η. Each condition uses 64 validation and 192 test dictionaries. Selection minimizes the mean validation risk separately for each query, and the intervals are 95% paired bootstrap intervals over test dictionaries.

Results.

Table 14: Per-step comparison on the 256 CIFAR evaluation pairs. D_{step} is the delta rule tuned per step, H the selected momentum rule on the grid with $\eta = 0$, and M_+ the best momentum rule with $\eta \geq 0.5$. Differences are means in units of 10^{-2} with 95% paired bootstrap intervals. Positive values favor the right-hand rule, and 0 marks identical rules ($\eta_H = 0$). The last three columns give the selected θ of D_{step} and (η, θ) of H and M_+.

(ρ, λ)	h	RP $-$ D_{step}	$D_{\text{step}} - M_+$	$D_{\text{step}} - H$	θ_D	H (η, θ)	M_+ (η, θ)
Reconstruction							
$(1, 1)$	0.1	2.09 [1.65, 2.57]	0.96 [0.82, 1.08]	0.96 [0.82, 1.08]	1.25	(0.5, 0.9)	(0.5, 0.9)
	0.05	1.78 [1.31, 2.27]	0.58 [0.50, 0.65]	0.58 [0.50, 0.65]	1.25	(0.5, 0.85)	(0.5, 0.85)
	0.025	1.60 [1.12, 2.11]	0.32 [0.28, 0.36]	0.32 [0.28, 0.36]	1.25	(0.5, 0.8)	(0.5, 0.8)
	0.0125	1.50 [1.01, 2.02]	0.17 [0.15, 0.19]	0.17 [0.15, 0.19]	1.25	(0.5, 0.8)	(0.5, 0.8)
$(1, 4)$	0.1	2.30 [2.09, 2.51]	-0.17 [-0.28, -0.07]	0	1.1	(0.0, 1.1)	(0.5, 0.95)
	0.05	2.46 [2.23, 2.69]	-0.15 [-0.23, -0.08]	0	1.1	(0.0, 1.1)	(0.5, 0.85)
	0.025	2.52 [2.28, 2.77]	-0.10 [-0.15, -0.06]	0	1.1	(0.0, 1.1)	(0.5, 0.8)
	0.0125	2.55 [2.30, 2.80]	-0.06 [-0.09, -0.04]	0	1.1	(0.0, 1.1)	(0.5, 0.8)
$(4, 1)$	0.1	-1.33 [-1.87, -0.76]	4.10 [3.67, 4.52]	4.10 [3.67, 4.52]	1.35	(0.5, 1.1)	(0.5, 1.1)
	0.05	-2.80 [-3.37, -2.18]	3.14 [2.86, 3.40]	3.14 [2.86, 3.40]	1.4	(0.5, 1.0)	(0.5, 1.0)
	0.025	-3.65 [-4.28, -3.00]	2.00 [1.50, 2.48]	2.00 [1.50, 2.48]	1.4	(0.8, 0.25)	(0.8, 0.25)
	0.0125	-4.12 [-4.77, -3.44]	1.62 [1.33, 1.90]	1.62 [1.33, 1.90]	1.4	(0.8, 0.2)	(0.8, 0.2)
$(4, 4)$	0.1	3.10 [2.72, 3.49]	1.44 [1.18, 1.68]	1.44 [1.18, 1.68]	1.2	(0.5, 1.1)	(0.5, 1.1)
	0.05	2.58 [2.16, 3.01]	1.35 [1.14, 1.54]	1.35 [1.14, 1.54]	1.25	(0.5, 1.0)	(0.5, 1.0)
	0.025	2.09 [1.64, 2.56]	0.96 [0.82, 1.08]	0.96 [0.82, 1.08]	1.25	(0.5, 0.9)	(0.5, 0.9)
	0.0125	1.78 [1.31, 2.27]	0.58 [0.50, 0.65]	0.58 [0.50, 0.65]	1.25	(0.5, 0.85)	(0.5, 0.85)
Associative							
$(1, 1)$	0.1	85.36 [81.04, 89.78]	-2.26 [-2.27, -2.25]	0	0.2	(0.0, 0.2)	(0.5, 0.1)
	0.05	84.84 [80.26, 89.41]	-1.29 [-1.29, -1.28]	0	0.1	(0.0, 0.1)	(0.5, 0.05)
	0.025	84.42 [79.70, 89.03]	-1.03 [-1.45, -0.59]	0	0.05	(0.0, 0.05)	(0.5, 0.05)
	0.0125	84.14 [79.81, 88.38]	-2.18 [-2.56, -1.80]	0	0.05	(0.0, 0.05)	(0.5, 0.05)
$(1, 4)$	0.1	50.21 [47.98, 52.33]	-4.18 [-4.21, -4.15]	0	0.35	(0.0, 0.35)	(0.5, 0.2)
	0.05	52.73 [50.47, 54.97]	-3.27 [-3.42, -3.11]	0	0.25	(0.0, 0.25)	(0.5, 0.1)
	0.025	53.67 [51.25, 55.99]	-2.21 [-2.41, -2.01]	0	0.15	(0.0, 0.15)	(0.5, 0.05)
	0.0125	53.59 [50.93, 56.18]	-1.10 [-1.41, -0.78]	0	0.05	(0.0, 0.05)	(0.5, 0.05)
$(4, 1)$	0.1	79.21 [73.64, 84.81]	-1.61 [-1.76, -1.46]	0	0.4	(0.0, 0.4)	(0.5, 0.25)
	0.05	72.77 [67.28, 78.36]	-1.47 [-1.57, -1.36]	0	0.35	(0.0, 0.35)	(0.5, 0.15)
	0.025	68.77 [63.32, 74.35]	-1.00 [-1.11, -0.89]	0	0.25	(0.0, 0.25)	(0.5, 0.1)
	0.0125	66.49 [61.23, 72.01]	-0.78 [-0.99, -0.55]	0	0.15	(0.0, 0.15)	(0.5, 0.05)
$(4, 4)$	0.1	81.57 [77.96, 85.16]	-3.61 [-3.73, -3.49]	0	0.45	(0.0, 0.45)	(0.5, 0.25)
	0.05	85.11 [81.02, 89.11]	-3.40 [-3.59, -3.21]	0	0.35	(0.0, 0.35)	(0.5, 0.15)
	0.025	85.36 [80.97, 89.73]	-2.26 [-2.27, -2.25]	0	0.2	(0.0, 0.2)	(0.5, 0.1)
	0.0125	84.84 [80.27, 89.43]	-1.29 [-1.29, -1.28]	0	0.1	(0.0, 0.1)	(0.5, 0.05)

- *Highly aligned keys ($s = 0.9$, pairwise cosine ≈ 0.8).* Validation selects $\eta \in [0.09, 0.48]$, and tuned momentum attains lower risk than the tuned constant-gain delta rule on every test dictionary, under both queries and in all three conditions. The gap is largest under reconstruction and grows with K.
- *Intermediate alignment ($s \in \{0.7, 0.8\}$, median pairwise cosine 0.50 and 0.65).* Validation selects $\eta \in [0.17, 0.26]$, and the gap is 0.002 to 0.007 with every interval below zero. Momentum is lower on at least 98% of the test dictionaries.
- *Less aligned keys ($s = 0.5$).* Validation still selects $\eta > 0$ (0.08 and 0.11) although $\eta = 0$ is available, but the gap is below 0.001 and momentum is lower on 72% and 92% of the test dictionaries.
- *Queries.* As in Appendix A.3, the selected θ exceeds one under reconstruction and lies at or below one under the associative query, and RP is worse than P under the associative query in every condition.

On the two-key CIFAR pairs of Appendix A.4.1, momentum with small η also beats the per-step delta rule under reconstruction but not under the associative query, where the gaps stay below 0.004 (Table 13). On frozen natural features with $K = 4$ and 8 (below), momentum also wins under the associative query, so the difference from Appendix A.4.1 is not due to synthetic keys alone. It may

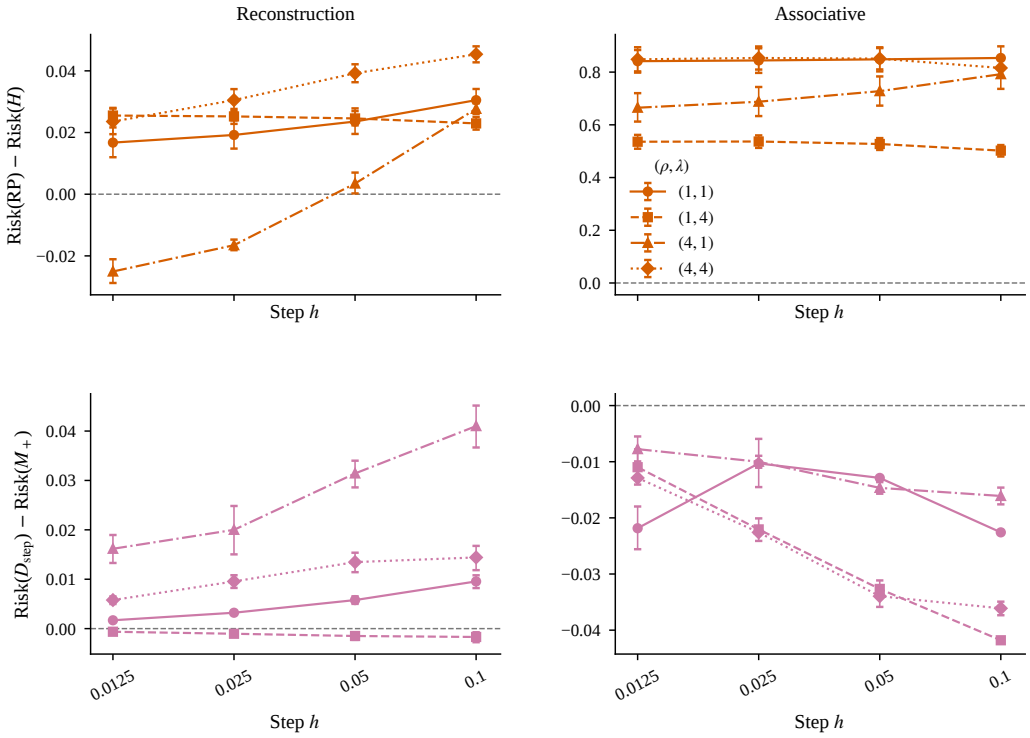

Figure 8: Paired evaluation risk of RP minus the validation-selected discrete momentum rule, by step h and rate cell (ρ, λ), under the reconstruction and associative queries. Momentum (η, θ) is selected separately for each h, cell, and query on 740 CIFAR validation pairs. Values and 95% intervals use 256 evaluation pairs and 10,000 paired bootstrap resamples. RP acts at the first step and at observed key changes, and both rules are queried after each step. Positive values favor momentum. The second row shows the per-step delta rule minus the best momentum rule with $\eta \geq 0.5$ (Table 14).

Table 15: Token-level mean test risks on synthetic keys after validation-only local refinement. Parameters are selected on 64 validation dictionaries, and risks and intervals use 192 test dictionaries. Parentheses give the selected θ of the constant-gain delta rule D and (η, θ) of momentum M. Validation selects $\eta > 0$ in every row. $M - D$ is in units of 10^{-3} with 95% paired bootstrap intervals (10,000 resamples), and Pr is the fraction of test dictionaries on which M is lower. Access is move-on unless marked revisit (rev.).

Condition	query	P	RP	D (θ)	M (η, θ)	$M - D$ [95% CI]	Pr
K=4, s=0.5	rec.	0.1424	0.2753	0.1419 (1.05)	0.1417 (0.08, 1.00)	−0.21 [−0.26, −0.17]	0.72
	assoc.	0.1215	0.2688	0.1215 (0.99)	0.1210 (0.11, 0.92)	−0.52 [−0.58, −0.46]	0.92
K=4, s=0.7	rec.	0.1729	0.2467	0.1697 (1.12)	0.1674 (0.17, 1.01)	−2.34 [−2.50, −2.19]	0.98
	assoc.	0.1154	0.2288	0.1153 (0.97)	0.1125 (0.18, 0.88)	−2.74 [−2.85, −2.62]	1.00
K=4, s=0.8	rec.	0.2095	0.2362	0.2009 (1.19)	0.1940 (0.26, 1.04)	−6.90 [−7.21, −6.57]	0.99
	assoc.	0.1081	0.1964	0.1076 (0.95)	0.1028 (0.23, 0.86)	−4.80 [−4.90, −4.70]	1.00
K=4, s=0.9	rec.	0.2980	0.2690	0.2728 (1.32)	0.2512 (0.37, 1.16)	−21.65 [−22.27, −21.04]	1.00
	assoc.	0.0904	0.1548	0.0888 (0.89)	0.0832 (0.25, 0.85)	−5.57 [−5.63, −5.51]	1.00
K=4, s=0.9, rev.	rec.	0.4228	0.3999	0.4125 (1.27)	0.3964 (0.33, 1.05)	−16.06 [−16.46, −15.65]	1.00
	assoc.	0.1309	0.2295	0.1287 (0.85)	0.1280 (0.09, 0.83)	−0.60 [−0.63, −0.58]	1.00
K=8, s=0.9	rec.	0.3927	0.3573	0.3612 (1.37)	0.3269 (0.48, 1.03)	−34.21 [−34.72, −33.69]	1.00
	assoc.	0.1089	0.1778	0.1060 (0.84)	0.0990 (0.34, 0.75)	−6.99 [−7.05, −6.93]	1.00

reflect the several updates each key receives per visit in Appendix A.4.1, the number of keys, or the access process, which we did not separate. The advantage of the tuned constant-gain delta rule in Appendix A.3 is therefore a property of the dwell model, in which the scalar rules act only at key

Table 16: Token-level comparison on frozen natural features. Parameters are selected only on 64 validation dictionaries per condition, and test columns use 192 dictionaries. Keys are the raw features (al.) or features centered by the bank mean and renormalized (cen.). Access is move-on unless marked revisit (rev.). The cosine column is the median signed pairwise cosine within a dictionary. Parentheses give the selected θ of the constant-gain delta rule D and (η, θ) of momentum M. Validation selects $\eta > 0$ in every row. $M - D$ is in units of 10^{-3} with 95% paired bootstrap intervals over whole test dictionaries (10,000 resamples). Pr is the fraction of test dictionaries on which M is lower, and no test dictionary has $|M - D| \leq 10^{-12}$.

K, keys	cos.	query	P	RP	D (θ)	M (η, θ)	$M - D$ [95% CI]	Pr
CIFAR								
4, al.	0.89	rec.	0.4652	0.4331	0.4398 (1.4)	0.4213 (0.3, 1.4)	-18.59 [-19.88, -17.26]	0.97
		assoc.	0.0570	0.1009	0.0559 (0.9)	0.0535 (0.2, 0.9)	-2.33 [-2.46, -2.19]	0.98
4, cen.	0.04	rec.	0.2910	0.3299	0.2794 (1.2)	0.2755 (0.2, 1.1)	-3.92 [-4.43, -3.40]	0.85
		assoc.	0.0918	0.1966	0.0915 (0.9)	0.0898 (0.1, 0.9)	-1.71 [-1.80, -1.62]	1.00
8, al.	0.89	rec.	0.6405	0.6160	0.6199 (1.5)	0.5990 (0.5, 1.2)	-20.97 [-21.81, -20.12]	1.00
		assoc.	0.0530	0.0881	0.0520 (0.9)	0.0496 (0.2, 0.8)	-2.36 [-2.42, -2.31]	1.00
8, cen.	0.00	rec.	0.4831	0.4936	0.4674 (1.3)	0.4613 (0.3, 1.1)	-6.03 [-6.53, -5.53]	0.95
		assoc.	0.0885	0.1747	0.0877 (0.9)	0.0866 (0.2, 0.8)	-1.09 [-1.18, -1.00]	0.99
4, al., rev.	0.89	rec.	0.5439	0.5229	0.5368 (1.3)	0.5252 (0.3, 1.2)	-11.66 [-12.30, -11.02]	1.00
		assoc.	0.0824	0.1480	0.0814 (0.9)	0.0811 (0.05, 0.9)	-0.23 [-0.25, -0.20]	0.85
Omniglot								
4, al.	0.92	rec.	0.5625	0.5563	0.5567 (1.3)	0.5534 (0.3, 1.2)	-3.28 [-4.03, -2.55]	0.69
		assoc.	0.0375	0.0666	0.0375 (1.0)	0.0364 (0.1, 0.9)	-1.10 [-1.17, -1.02]	0.97
4, cen.	-0.18	rec.	0.5519	0.5382	0.5374 (1.4)	0.5299 (0.3, 1.3)	-7.49 [-9.08, -5.89]	0.73
		assoc.	0.0396	0.0764	0.0393 (0.9)	0.0383 (0.1, 0.9)	-1.00 [-1.09, -0.91]	0.93
8, al.	0.93	rec.	0.7623	0.7585	0.7594 (1.3)	0.7569 (0.3, 1.2)	-2.55 [-2.82, -2.29]	0.96
		assoc.	0.0254	0.0418	0.0254 (1.0)	0.0248 (0.1, 1.0)	-0.68 [-0.72, -0.64]	0.98
8, cen.	0.03	rec.	0.7148	0.7018	0.7011 (1.5)	0.6930 (0.3, 1.3)	-8.17 [-8.84, -7.49]	0.94
		assoc.	0.0362	0.0650	0.0357 (0.9)	0.0350 (0.1, 0.9)	-0.74 [-0.79, -0.69]	0.98
4, al., rev.	0.92	rec.	0.5990	0.5972	0.5969 (1.2)	0.5938 (0.2, 1.1)	-3.16 [-3.50, -2.81]	0.88
		assoc.	0.0604	0.1078	0.0604 (0.9)	0.0601 (0.05, 0.9)	-0.35 [-0.38, -0.32]	0.88

changes, and it does not carry over to per-token updates. The comparison covers constant gains only. A gain schedule tuned per token is not tested.

Frozen natural features. We repeat the protocol on frozen features, with the same selection rule, risks, and numbers of validation and test dictionaries, on a coarser grid without local refinement. The delta rule tunes $\theta \in \{0.1, 0.2, \ldots, 2\}$, and momentum tunes $\eta \in \{0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.8, 0.9, 0.95\}$ together with the same θ grid. CIFAR dictionaries use held-out images outside the 512 evaluation images of Appendix A.3, and Omniglot dictionaries use held-out alphabets. Validation and test images are disjoint. For $K = 8$, CIFAR has too few such images, so test dictionaries reuse 1,106 distinct images in 1,536 positions within the test pool, and its intervals describe this fixed collection. Each source has five conditions: $K \in \{4, 8\}$ with move-on access on the raw features (*aligned*) or on features centered by the bank mean and renormalized (*centered*), and $K = 4$ with revisit access on the raw features. Empty-memory risks equal one for every dictionary. Table 16 reports the results, and Figure 9 plots the gaps against the within-dictionary absolute cosine. Validation selects $\eta > 0$ in all 20 query cells, and tuned momentum has lower mean test risk than the tuned constant-gain delta rule in every cell, with every 95% interval below zero: by 0.004 to 0.021 on CIFAR and 0.003 to 0.008 on Omniglot under reconstruction, and by 0.0002 to 0.0024 under the associative query. Momentum is lower on 69% to 100% of the test dictionaries. Centering lowers the absolute cosines of CIFAR dictionaries and the reconstruction gap with them, but after centering at least half of the Omniglot pairwise cosines still have absolute value above 0.94, so the centered condition does not isolate alignment on natural features.

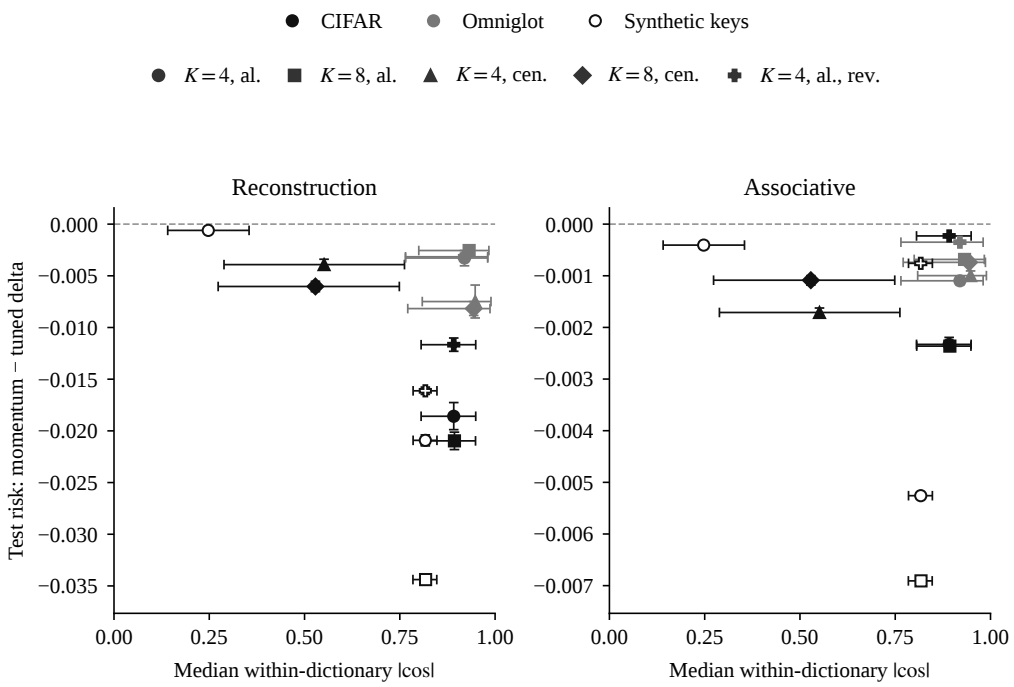

Figure 9: Token-level gap between tuned momentum and the tuned constant-gain delta rule against the median within-dictionary absolute cosine, on frozen CIFAR (black) and Omniglot (grey) features and on the synthetic keys of Table 15 (open markers). Marker shapes give the condition. Horizontal bars are interquartile ranges of the pairwise absolute cosines. Vertical bars are 95% paired bootstrap intervals. Negative values favor momentum. Synthetic points use the coarse grid without local refinement, so their gaps differ from Table 15.

A.5 Numerical checks of the expansions and the criterion

A.5.1 Numerical check of the small-ε expansions

Risk matrices. We evaluate the reconstruction risk for an arbitrary initial error row $f_0^\top$ through $R_a(f_0) = f_0^\top \mathcal{R}_a(\varepsilon) f_0$. Throughout this check, $C = I_d$, $Q \sim \mathrm{Exp}(1)$, and $X_0 \sim \pi$, where π is the stationary distribution of the chosen kernel. The momentum rule is $H(1, 30)$ with zero initial derivative. For the P and RP families, the first gain is $1 + y$. Subsequent gains are those of Appendix A.1.1. We compute the full risk matrices and compare their entries with the expansions in Propositions 3.2 and 3.4 and Lemma 4.1.

Exact-expectation equations. For H, let $\mathcal{V}_i \in \mathbb{R}^{2d \times 2d}$ be the matrix of expected quadratic cost conditional on the initial active key i. We solve the coupled equations

$$\mathcal{G}_i^\top \mathcal{V}_i + \mathcal{V}_i \mathcal{G}_i - \mathcal{V}_i + \rho \sum_{i'=1}^{K} T_{ii'} (\mathcal{V}_{i'} - \mathcal{V}_i) = -\operatorname{diag}(I_d, 0_d), \qquad \mathcal{G}_i = \begin{pmatrix} 0_d & I_d \\ -\beta P_{k_i} & -\alpha I_d \end{pmatrix},$$

where 0_d is the $d \times d$ zero matrix and $(\alpha, \beta) = (1, 30)$. With zero initial derivative, the risk matrix is

$$\mathcal{R}_H = \sum_{i=1}^{K} \pi_i (\mathcal{V}_i)_{11},$$

where $(\mathcal{V}_i)_{11}$ is the upper-left $d \times d$ block.

For the scalar update families, let $\mathcal{V}_{i,f}$ be the expected quadratic-cost matrix after the current update, conditional on active key i and next-update phase f. The projection family has one phase with gain

$\theta_f = 1$. Reflect–project has two phases with gains 2 and 1, in that order. Let f' denote the next phase and define $A_{i,f} = I_d - \theta_f P_{k_i}$. We iterate

$$\mathcal{V}_{i,f} = p_{\mathrm{q}} I_d + p_{\mathrm{c}} \sum_{i'=1}^{K} T_{ii'} A_{i',f} \mathcal{V}_{i',f'} A_{i',f}, \qquad p_{\mathrm{q}} = \frac{1}{1+\rho}, \qquad p_{\mathrm{c}} = \frac{\rho}{1+\rho}.$$

Starting from zero matrices, we stop when the maximum absolute entry change is at most 2×10^{-15}. The initial update gives

$$\mathcal{R}_a = \sum_{i=1}^{K} \pi_i A_{0,i} \mathcal{V}_{i,\mathrm{first}} A_{0,i}, \qquad A_{0,i} = I_d - (1+y) P_{k_i},$$

where the first phase is projection for P and reflection for RP. Direct linear solves of the same fixed-point equations provide a second evaluation.

Two keys. We use $k_1 = (1, \varepsilon)^\top / \sqrt{1+\varepsilon^2}$ and $k_2 = (1, -\varepsilon)^\top / \sqrt{1+\varepsilon^2}$, with deterministic alternation at key changes, $\pi = (1/2, 1/2)$, $\rho = 3$, and $y = 0$. The strong and weak directions are $e_{\mathrm{s}} = (1,0)^\top$ and $e_{\mathrm{w}} = (0,1)^\top$. At $\varepsilon \in \{0.02, 0.01, 0.005\}$, we report

$$\begin{aligned} \mathrm{err}_{H,\mathrm{w}} &= \left| \frac{1 - R_H^{\mathrm{w}}}{\varepsilon^2} - \bar{c}_{\mathrm{H}} \right|, \\ \mathrm{err}_{H,\mathrm{s}} &= \left| R_H(e_{\mathrm{s}}) - q_{\mathrm{s}} \right|, \\ \mathrm{err}_{\mathrm{RP},\mathrm{w}} &= \left| R_{\mathrm{RP}}^{\mathrm{w}} - (1 - \bar{c}_{\mathrm{RP}} \varepsilon^2) \right|. \end{aligned}$$

Here equation 8, equation 7, and equation 9 give $\bar{c}_{\mathrm{H}} = 31890/1763$, $q_{\mathrm{s}} = 11/41$, and $\bar{c}_{\mathrm{RP}} = 151/7$, respectively.

Four keys with reversible access. We fix $K = 4$, $d = 3$, $\rho = 1.7$, and $y = 0.4$. Using NumPy's PCG64 generator with a fixed seed, we draw six independent $\mathrm{Uniform}(0,1)$ weights a_{ij} in the order $(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)$. We set $a_{ji} = a_{ij}$ and $a_{ii} = 0$, then define

$$b_i = \sum_{i'=1}^{4} a_{ii'}, \qquad T_{ii'} = \frac{a_{ii'}}{b_i}, \qquad \pi_i = \frac{b_i}{\sum_{j=1}^{4} b_j}.$$

These positive off-diagonal weights give irreducibility and $\pi_i T_{ii'} = \pi_{i'} T_{i'i}$. Thus this kernel satisfies Assumption 4.1. Continuing the same generator, we draw the two coordinates of each $u_i \sim \mathcal{N}(0, I_2)$ in index order and construct the normalized keys using equation 12. The same T, π, and weak components serve all rules and all three values of ε.

Let ϕ_ℓ be the nonconstant π-orthonormal eigenvectors of T, with eigenvalues t_ℓ. We compute $\mu = \sum_i \pi_i u_i$ and $\mu_\ell = \sum_i \pi_i \phi_\ell(i) u_i$, and form

$$B_a = c_a(1) \mu \mu^\top + \sum_{\ell=1}^{3} c_a(t_\ell) \mu_\ell \mu_\ell^\top.$$

We use $s_H = q_{\mathrm{s}}$ from equation 7 and $c_a(t) = c_{\mathrm{H}}(\rho(1-t))$ for $a = H$, with c_{H} from equation 25. For $a = \mathrm{P}, \mathrm{RP}$, we use the strong coefficients s_a of Lemma 4.1 and the modal coefficients from equation 26 and equation 27. In each case, $c_a(1) = 1 - s_a$. Partitioning the risk matrix into its strong and weak coordinates, we measure

$$\begin{aligned} \mathrm{err}_{a,\mathrm{s}} &= |\mathcal{R}_{a,\mathrm{ss}} - s_a|, \\ \mathrm{err}_{a,\mathrm{c}} &= \|\varepsilon^{-1} \mathcal{R}_{a,\mathrm{ws}} - (s_a - 1)\mu\|_{\max}, \\ \mathrm{err}_{a,\mathrm{w}} &= \|\varepsilon^{-2}(I_2 - \mathcal{R}_{a,\mathrm{ww}}) - B_a\|_{\max}, \end{aligned}$$

where $\|\cdot\|_{\max}$ is the largest absolute entry. These errors compare the strong block, the scaled cross block, and the scaled weak-block reduction, respectively.

Numerical results. Table 17 reports the computed errors. Halving ε reduces the four-key errors by factors between 0.2502 and 0.2514. The two-key H errors have factors between 0.2505 and 0.2522. These values agree with the $O(\varepsilon^2)$ remainders of the corresponding scaled expansions. The unscaled two-key RP error has factors 0.0630 and 0.0626, consistent with the $O(\varepsilon^4)$ remainder in equation 9. Dividing that error by ε^2 gives factors 0.2519 and 0.2505.

Table 17: Absolute errors between the exact-expectation risk matrices evaluated in float64 and their small-ε expansions. The four-key errors use the maximum absolute entry in each block. Every occurrence of H denotes $H(1, 30)$.

Two keys, $\rho = 3$, $y = 0$			
ε	$\mathrm{err}_{H,\mathrm{w}}$	$\mathrm{err}_{H,\mathrm{s}}$	$\mathrm{err}_{\mathrm{RP},\mathrm{w}}$
0.020	2.12×10^{-1}	4.91×10^{-3}	9.00×10^{-5}
0.010	5.34×10^{-2}	1.24×10^{-3}	5.67×10^{-6}
0.005	1.34×10^{-2}	3.10×10^{-4}	3.55×10^{-7}
Four keys, $\rho = 1.7$, $y = 0.4$			
(a, ε)	$\mathrm{err}_{a,\mathrm{s}}$	$\mathrm{err}_{a,\mathrm{c}}$	$\mathrm{err}_{a,\mathrm{w}}$
$(\mathrm{P}, 0.020)$	1.06×10^{-3}	2.65×10^{-3}	1.31×10^{-2}
$(\mathrm{P}, 0.010)$	2.66×10^{-4}	6.65×10^{-4}	3.28×10^{-3}
$(\mathrm{P}, 0.005)$	6.65×10^{-5}	1.66×10^{-4}	8.21×10^{-4}
$(\mathrm{RP}, 0.020)$	1.12×10^{-3}	2.60×10^{-3}	2.45×10^{-2}
$(\mathrm{RP}, 0.010)$	2.81×10^{-4}	6.53×10^{-4}	6.15×10^{-3}
$(\mathrm{RP}, 0.005)$	7.02×10^{-5}	1.63×10^{-4}	1.54×10^{-3}
$(H, 0.020)$	3.86×10^{-3}	5.90×10^{-3}	3.48×10^{-2}
$(H, 0.010)$	9.70×10^{-4}	1.48×10^{-3}	8.74×10^{-3}
$(H, 0.005)$	2.43×10^{-4}	3.71×10^{-4}	2.19×10^{-3}

Solver diagnostics. The computation uses NumPy 2.2.6 and SciPy 1.15.3 in float64. Across the reported solves, the largest scaled residual is 1.15×10^{-15}. We scale each equation residual by one plus the maximum absolute right-hand-side entry. For the vectorized momentum system $\mathcal{L}x = b$, the posterior error estimate is $\|\mathcal{L}^{-1}\|_\infty \|\mathcal{L}x - b\|_\infty$. For the phase equations, we divide the largest spectral norm of a residual block by $1 - p_{\mathrm{c}}$. The largest estimates are 1.93×10^{-14} for H and 5.77×10^{-15} for the scalar families. We compute these diagnostics in floating-point arithmetic. The scalar fixed-point and direct solutions differ by at most 5.66×10^{-15} in their risk-matrix entries. For the two-key check, the closed-form P and RP risks and the main solver agree with the computed risks within 6.3×10^{-15}.

A.5.2 Sign check of the criterion on natural features

Dictionaries and key paths. The check uses the 256 CIFAR evaluation pairs of Appendix A.3 ($K = 2$) and the 512 Omniglot access triples of Appendix A.1.2 ($K = 3$), 768 dictionaries in total, all in their stored order. We turn each dictionary into a path of nearly collinear keys. The top eigenvector of the average projector $\frac{1}{K}\sum_i P_{k_i}$ gives the strong axis e_{s}. We flip the sign of each key so that its strong coordinate is nonnegative, which leaves P_{k_i} unchanged. A QR decomposition of the residuals, with fixed pivot and sign rules, gives an orthonormal basis of the weak span. With u_i the weak coordinates of k_i divided by its strong coordinate, the path is

$$k_i(\varepsilon) = \frac{e_{\mathrm{s}} + \varepsilon u_i}{\sqrt{1 + \varepsilon^2 \|u_i\|^2}}, \qquad \varepsilon \in \{1/200,\ 1/100,\ 1/50,\ 1/25,\ 1/10\},$$

and the original dictionary lies on it at $\varepsilon = 1$. The query uses $C = I$ on the span of e_{s} and the weak basis.

Access, rules, and sign rule. CIFAR pairs switch keys by deterministic alternation. Omniglot triples use the uniform kernel and two weighted reversible kernels,

$$T_a = \begin{pmatrix} 0 & 4/5 & 1/5 \\ 4/5 & 0 & 1/5 \\ 1/2 & 1/2 & 0 \end{pmatrix}, \qquad T_b = \begin{pmatrix} 0 & 1/5 & 4/5 \\ 1/2 & 0 & 1/2 \\ 4/5 & 1/5 & 0 \end{pmatrix},$$

with stationary distributions $(5, 5, 2)/12$ and $(5, 2, 5)/12$. We take $\rho \in \{4, 10\}$, $\lambda = 1$, $X_0 \sim \pi$, the momentum rule $H(1, 30)$ with zero initial derivative, and the rules P and RP with $y = 0$. For each combination we compute g_0 and M from Proposition 4.2, and the risk matrices $\mathcal{R}_H(\varepsilon)$ and $\mathcal{R}_j(\varepsilon)$ with the exact-expectation equations of Appendix A.5.1. The quantity

$$\Delta_j(\varepsilon) = \frac{\lambda_{\min}\big(\mathcal{R}_H(\varepsilon) - \mathcal{R}_j(\varepsilon)\big)}{\varepsilon^2}$$

is positive exactly when rule j attains lower risk than H for every initial error, and it tends to $\lambda_{\min}(M)$ as $\varepsilon \to 0$ when $g_0 > 0$. A sign counts as resolved when the value exceeds $\max(10^{-12}, 10^{-10}\max(1, \|\cdot\|_2))$ in absolute value. Every linear solve and eigenpair has relative residual below 10^{-10}.

Results. All 7,168 combinations of dictionary, kernel, rate, and rule have $g_0 > 0$. Within each kernel, rate, and rule, all dictionaries share the sign of $\lambda_{\min}(M)$ (Table 18). The rate switches the sign for P under every kernel. At $\rho = 4$, the weighted kernels switch it for RP on the same Omniglot dictionaries. Over the five values of ε, the check has 35,840 cases. In all 35,829 cases where both signs are resolved, the sign of $\Delta_j(\varepsilon)$ equals the sign of $\lambda_{\min}(M)$, including 2,660 cases with $|\lambda_{\min}(M)| \le 10^{-4}\|M\|_2$. The other 11 cases have $\lambda_{\min}(M) > 0$ and an unresolved Δ_j at $\varepsilon = 1/200$. All 11 are P on Omniglot at $\rho = 10$, seven under T_a and four under T_b, and each has $\Delta_j > 0$ at the four larger values of ε. No combination changes sign across ε. Figure 10 plots $\Delta_j(\varepsilon)$ against $\lambda_{\min}(M)$ for every case.

Table 18: Sign of $\lambda_{\min}(M)$ against $H(1, 30)$, shared by every dictionary in each cell. In every case with a resolved sign, $\Delta_j(\varepsilon)$ has the same sign at all five values of ε.

Source	Kernel	Dictionaries	P, $\rho = 4$	P, $\rho = 10$	RP, $\rho = 4$	RP, $\rho = 10$
CIFAR, $K = 2$	alternation	256	−	+	+	+
Omniglot, $K = 3$	uniform	512	−	+	+	+
Omniglot, $K = 3$	T_a	512	−	+	−	+
Omniglot, $K = 3$	T_b	512	−	+	−	+

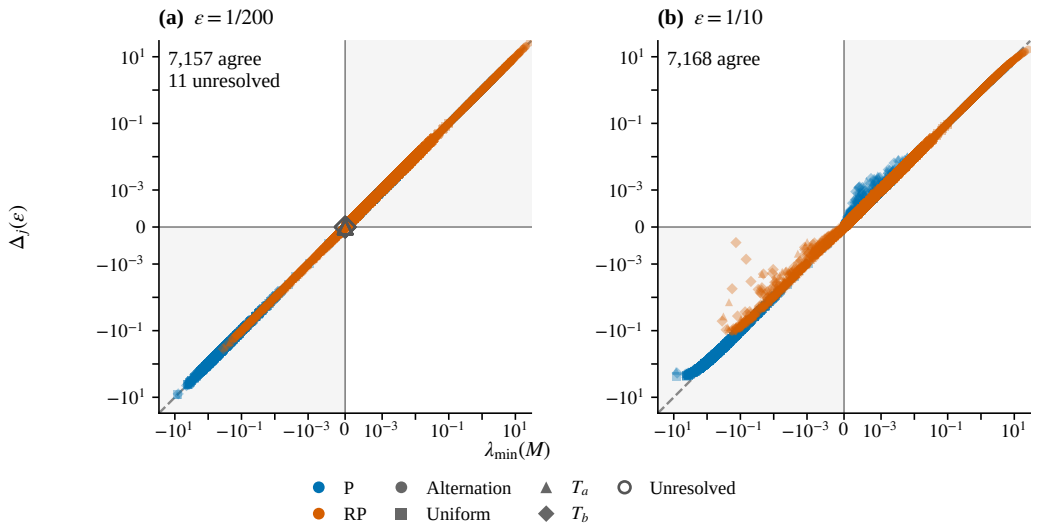

Figure 10: Sign check of the criterion. Each point is one combination of dictionary, kernel, rate, and rule, with $\lambda_{\min}(M)$ on the horizontal axis and $\Delta_j(\varepsilon)$ on the vertical axis, at $\varepsilon = 1/200$ (a) and $\varepsilon = 1/10$ (b). Colors mark the rule and marker shapes mark the kernel. Both axes use a symmetric logarithmic scale that is linear on $[-10^{-3}, 10^{-3}]$. The dashed line $y = x$ is the limit as $\varepsilon \to 0$ when $g_0 > 0$. The 11 cases with an unresolved sign at $\varepsilon = 1/200$ appear as open markers at $y = 0$.

Scope. Each CIFAR dictionary enters 20 cases and each Omniglot dictionary enters 60, so the counts describe these dependent dictionaries on a finite grid of ε in float64. Within each cell the dictionaries share one sign, so this check tests the rate, the kernel, and the rule, and Appendix A.2.1 tests the key geometry. Every combination has $g_0 > 0$, and the case $g_0 = 0$ remains untested. By Lemma E.3, this uniformity is expected: with $K - 1$ access modes spanning the weak span, the sign of $\lambda_{\min}(M)$ is that of $\min_\ell g_\ell$.

Extension to $K > d$ with mixed modal margins. The check above cannot separate the geometry from the margins, so we ran a second, separately frozen check in the regime where Lemma E.3 does not apply. A modal configuration is a choice of $K \in \{4, 5, 6, 8\}$, one of eight reversible access kernels per K, $\rho \in \{2, 4, 10\}$, a momentum rule ($H(1, 30)$ or the reconstruction-tuned $H(32, 5000)$ and $H(4, 200)$ of Appendix A.3), and a scalar rule (P or RP), giving $4 \times 8 \times 3 \times 3 \times 2 = 576$ configurations. A screen kept the 260 in which the margins g_ℓ take both signs. Only these configurations test the pairing of modes with weak directions: when all margins share one sign, $M = \sum_\ell g_\ell \mu_\ell \mu_\ell^\top$ is semidefinite with that sign for any geometry. In the screened-out configurations the margins share one sign, so $M = \sum_\ell g_\ell \mu_\ell \mu_\ell^\top$ is semidefinite with that sign whatever the geometry, and these configurations cannot test the pairing of modes with weak directions. This screen was seen during method development and before the risk computation was frozen. Each kept configuration is evaluated for every admissible $d \in \{3, 4\}$ with $d < K$ (465 cells), on 100 Gaussian weak-component samples and on 50 samples from each matching frozen feature source (both sources for $H(1, 30)$, the matching source for each tuned H, 667 cells in all), which gives $465 \times 100 + 667 \times 50 = 79{,}850$ comparison cases, all with $g_0 > 0$. Each case is evaluated at the ten values $\varepsilon \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1\}$, giving 798,500 case–$\varepsilon$ comparisons. The sign rule and the thresholds are those above. All risk matrices are finite and admissible, 50 of the 798,500 signs are unresolved.

Construction of the $K > d$ cases. Each case specifies weak coordinates $u_1, \dots, u_K \in \mathbb{R}^{d-1}$ and uses the keys $k_i(\varepsilon) = (1, \varepsilon u_i)/\sqrt{1 + \varepsilon^2 \|u_i\|^2}$ in $\mathbb{R}^d$. Gaussian cases draw the entries of u_i independently from $\mathcal{N}(0, 1)$, without centering or normalization. A natural case starts from K unit-normalized 32-dimensional frozen features $F_1, \dots, F_K$ with fixed identifiers. The leading eigenvector e_s of the uncentered mean projector $K^{-1} \sum_i F_i F_i^\top$ gives the strong direction, features are sign-flipped so that every strong coordinate $F_i^\top e_s$ is positive, and each feature gives the weak ratio $r_i = (F_i - (F_i^\top e_s) e_s)/(F_i^\top e_s)$. The top $d - 1$ principal axes of the centered ratios fix the weak coordinates, and u_i is the projection of the uncentered r_i onto these axes. Groups without a clear strong eigengap or with rank-deficient ratios are marked unresolved and are not replaced. For each K, the eight kernels are two random draws from each of four families. With $m = \lfloor K/2 \rfloor$, the base weight of a pair $i \neq j$ is 1 for every pair (dense), for cyclic neighbors (ring), for pairs within the same block $\{1, \dots, m\}$ or $\{m + 1, \dots, K\}$ (community), or for pairs across the two blocks (bipartite), and 0.02 otherwise. Each base weight is multiplied by a symmetric factor $\exp(0.4 Z_{ij})$ with $Z_{ij} \sim \mathcal{N}(0, 1)$. A Metropolis–Hastings chain with proposal proportional to these weights targets a random distribution π^{MH}, the softmax of standard normal logits, and removing its self-loops gives the jump kernel $T_{ij} = P_{ij}/(1 - P_{ii})$ for $j \neq i$ and $T_{ii} = 0$. Every kernel is irreducible and satisfies detailed balance with stationary law proportional to $\pi_i^{\mathrm{MH}}(1 - P_{ii})$.

Table 19 and Figure 11 report the agreement between $\operatorname{sign} \Delta_j(\varepsilon)$ and $\operatorname{sign} \lambda_{\min}(M)$. At $\varepsilon = 0.005$ the signs agree in 79,807 of 79,814 resolved cases. The seven exceptions are Gaussian samples with $|\lambda_{\min}(M)|/\|M\| \leq 3.7 \times 10^{-3}$, where the leading term is small and the expansion converges slowly. In 6,455 of the 73,350 cases where the majority sign of the g_ℓ is defined, M has the opposite sign, so a mode-by-mode vote would predict the wrong winner for every initial error. The matrix criterion is right in 99.9% of them at $\varepsilon = 0.005$ and in 98.1% at $\varepsilon = 0.02$. This is the effect of the key geometry that the natural-path check could not show.

Agreement falls as ε grows, to 93.1% at $\varepsilon = 0.1$ and 30.0% at $\varepsilon = 1$, and the failures are one-sided. When $\lambda_{\min}(M) > 0$, so that the scalar family wins for every initial error in the limit, the prediction still holds in 99.6% of cases at $\varepsilon = 0.1$ and 64.2% at $\varepsilon = 1$. When $\lambda_{\min}(M) < 0$, the predicted momentum advantage disappears at finite ε: agreement is 91.1% at $\varepsilon = 0.1$ and 19.3% at $\varepsilon = 1$, and 49,148 of the 55,931 disagreements at $\varepsilon = 1$ are of this kind. This matches Appendix D.3.1, where the weak-direction advantage of momentum requires $\beta\varepsilon^2$ to be small. The first disagreement (Figure 12) most often occurs at $\varepsilon = 0.2$ and has a similar distribution for every K, and 23,737 cases agree at all ten values. At $\varepsilon = 0.1$ agreement is 99.2% on the CIFAR and Omniglot samples and 88.7% on the Gaussian ones. The criterion is therefore a small-ε statement: it identifies the winner reliably for $\varepsilon \lesssim 0.05$, including where the geometry overturns the modal margins, but it does not predict the ranking at the scale of the original dictionaries, $\varepsilon = 1$. The samples reuse frozen geometries across kernels, rates, rules, and values of ε, so all counts are descriptive, and the case $g_0 = 0$ remains untested.

Table 19: Agreement between $\operatorname{sign}\Delta_j(\varepsilon)$ and $\operatorname{sign}\lambda_{\min}(M)$ in the $K > d$ extension, as a percentage of resolved cases. "M opposes" counts the 6,455 cases whose M has the sign opposite to the majority of the modal margins g_ℓ. Shading marks $\varepsilon \le 0.05$.

ε	0.005	0.01	0.02	0.05	0.1	0.2	0.3	0.5	0.7	1
all cases	99.99	99.96	99.79	98.59	93.08	70.36	57.54	44.52	36.89	29.95
$\lambda_{\min}(M) > 0$	100.00	100.00	99.98	99.94	99.59	95.61	85.15	73.79	69.26	64.22
$\lambda_{\min}(M) < 0$	99.99	99.95	99.73	98.17	91.06	62.50	48.95	35.41	26.82	19.29
M opposes	99.92	99.58	98.08	89.74	72.01	51.98	43.18	34.61	30.47	27.59

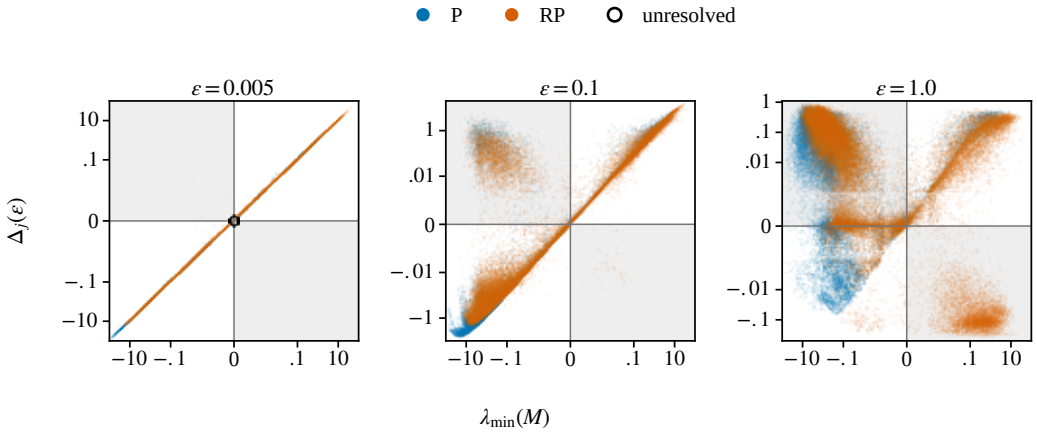

Figure 11: $\Delta_j(\varepsilon)$ against $\lambda_{\min}(M)$ for all 79,850 cases of the $K > d$ extension at $\varepsilon = 0.005$, 0.1, and 1. Both axes use a symmetric logarithmic scale. Blue points are P and orange points RP. Open circles mark unresolved signs. The shaded upper-left and lower-right quadrants are where the finite-ε comparison disagrees with the criterion.

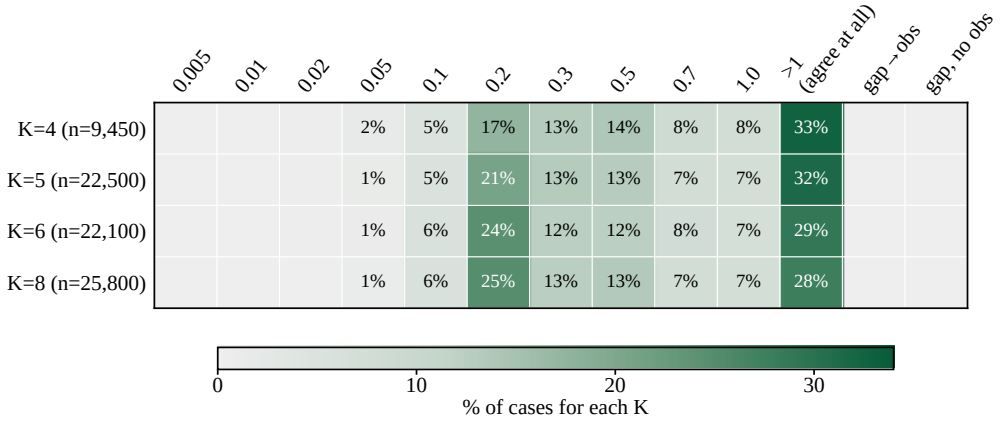

Figure 12: Grid value ε_* of the first resolved disagreement in the $K > d$ extension, as a percentage of the cases for each K. "> 1 (agree at all)" marks cases that agree at all ten values. "gap→obs" marks cases with an observed disagreement after an earlier unresolved sign, and "gap, no obs" cases with an unresolved sign and no observed disagreement. Empty cells are below 0.5%.

B RESULTS FOR THE FOUR FACTORS

B.1 RETRIEVAL QUERY AND ELAPSED TIME

B.1.1 RETRIEVAL QUERY

We return to the two-key setting of Section 3 with a general query rate λ, and compare P and RP (both with $y = 0$) under the reconstruction query $C = I_2$ and under the associative query at the stored keys, $CC^\top = \frac{1}{2}(P_{k_+} + P_{k_-})$. Write $c = \cos 2\varphi$ as in Section 3, $\chi = c^2$, and $p_c = \rho/(\rho + \lambda)$.

Proposition B.1 (The query reverses P and RP)**.** *For every $\rho, \lambda > 0$ and $\chi \in (0, 1)$, all four risk matrices are positive multiples of $B_0 = \mathrm{diag}(\sin^2\varphi, \cos^2\varphi)$, so each comparison below holds for every nonzero initial error.*

(i) Under reconstruction, RP *attains lower risk than* P *if and only if $\chi > (2\rho + \lambda)/(4\rho)$.*

(ii) Under the associative query, P *attains lower risk than* RP.

Hence for $\chi > (2\rho + \lambda)/(4\rho)$ the two queries rank P *and* RP *in opposite orders. At $\rho = 4$ and $\lambda = 1$ the threshold is $\chi > 9/16$, that is, $|\cos 2\varphi| > 3/4$.*

Proof. Step 1: products. By Steps 1 and 2 of the proof of Proposition 3.4, after n key changes the product of update matrices is $\Pi_n = c^n\, n_{\sigma_0} n_{\sigma_n}^\top$ for P. For RP it is $\Pi_n = (2\chi - 1)^m\, n_{\sigma_0} n_{\sigma_0}^\top$ for $n = 2m$ and $\Pi_n = (2\chi - 1)^m\, n_{\sigma_0} u^\top$ for $n = 2m + 1$, where $u = (I - 2P_{k_{-\sigma_0}})n_{\sigma_0}$ and we used $\cos 4\varphi = 2\chi - 1$. Hence $\Pi_n CC^\top \Pi_n^\top = \gamma_n\, n_{\sigma_0} n_{\sigma_0}^\top$ with a scalar γ_n, and averaging over σ_0 turns $n_{\sigma_0} n_{\sigma_0}^\top$ into B_0.

Step 2: the scalars. Under reconstruction, $\gamma_n = \chi^n$ for P and $\gamma_n = (2\chi - 1)^{2m}$ for RP, since $\|n_\sigma\| = \|u\| = 1$. Under the associative query we need the quadratic form of $\bar{P} = \frac{1}{2}(P_{k_+} + P_{k_-})$ at n_σ and at u. Since $k_\sigma^\top n_\sigma = 0$ and $(k_{-\sigma}^\top n_\sigma)^2 = \sin^2 2\varphi = 1 - \chi$,

$$n_\sigma^\top \bar{P} n_\sigma = \tfrac{1}{2}(1 - \chi).$$

For u, we have $k_{-\sigma_0}^\top u = -k_{-\sigma_0}^\top n_{\sigma_0}$ and $k_{\sigma_0}^\top u = -2(k_{-\sigma_0}^\top n_{\sigma_0})(k_{\sigma_0}^\top k_{-\sigma_0})$, so

$$u^\top \bar{P} u = \tfrac{1}{2}(1 - \chi)(1 + 4\chi).$$

Hence under the associative query $\gamma_n = \frac{1}{2}(1 - \chi)\chi^n$ for P, and $\gamma_n = \frac{1}{2}(1 - \chi)(2\chi - 1)^{2m}$ or $\frac{1}{2}(1 - \chi)(1 + 4\chi)(2\chi - 1)^{2m}$ for RP with $n = 2m$ or $n = 2m + 1$.

Step 3: expectations. By Lemma C.2, $\Pr(N(Q) = n) = (1 - p_c)p_c^n$. Summing the geometric series, the multiples of B_0 are

$$\tau_{\mathrm{P}}^{\mathrm{rec}} = \frac{1 - p_c}{1 - p_c\chi}, \qquad \tau_{\mathrm{RP}}^{\mathrm{rec}} = \frac{1 - p_c^2}{1 - p_c^2(2\chi - 1)^2},$$

$$\tau_{\mathrm{P}}^{\mathrm{assoc}} = \frac{1 - \chi}{2} \cdot \frac{1 - p_c}{1 - p_c\chi}, \qquad \tau_{\mathrm{RP}}^{\mathrm{assoc}} = \frac{1 - \chi}{2} \cdot \frac{(1 - p_c)(1 + p_c + 4p_c\chi)}{1 - p_c^2(2\chi - 1)^2}.$$

Step 4: comparisons. For (i), cross-multiplying,

$$(1 - p_c)\big(1 - p_c^2(2\chi - 1)^2\big) - (1 - p_c^2)(1 - p_c\chi) = p_c(1 - p_c)(1 - \chi)\big(4p_c\chi - 1 - p_c\big),$$

which is positive if and only if $\chi > (1 + p_c)/(4p_c) = (2\rho + \lambda)/(4\rho)$. For (ii),

$$(1 + p_c + 4p_c\chi)(1 - p_c\chi) - \big(1 - p_c^2(2\chi - 1)^2\big) = p_c\big[(1 + p_c) + (3 - 5p_c)\chi\big].$$

The bracket is linear in χ, equals $1 + p_c > 0$ at $\chi = 0$, and equals $4(1 - p_c) > 0$ at $\chi = 1$, so it is positive on $[0, 1]$. Hence $\tau_{\mathrm{P}}^{\mathrm{assoc}} < \tau_{\mathrm{RP}}^{\mathrm{assoc}}$. □

At $\lambda = 1$, condition (i) reads $2\rho\cos 4\varphi > 1$, which is the condition of Proposition 3.4.

B.1.2 ELAPSED TIME

Proposition B.2 (Time scaling). *Let the access process have rate ρ and let $Q \sim \mathrm{Exp}(\lambda)$.*

(i) For P *and* RP *with any first learning rate, and for any keys, access kernel T, query directions C, and initial error, the risk depends on (ρ, λ) only through ρ/λ. The expected number of updates before the query is $1 + \rho/\lambda$.*

(ii) For the momentum rule, $R\big(H(\alpha, \beta); \rho, \lambda\big) = R\big(H(\alpha/\lambda, \beta/\lambda^2); \rho/\lambda, 1\big)$. In particular, the strong-direction risk at $\varepsilon = 0$ is

$$q_{\mathrm{s}}(\lambda) = \frac{\lambda\big[(\alpha+\lambda)(2\alpha+\lambda)+2\beta\big]}{(\alpha+\lambda)(\lambda^2+2\alpha\lambda+4\beta)},$$

which does not depend on ρ, tends to 0 as $\lambda \to 0$, and tends to 1 as $\lambda \to \infty$.

Proof. (i) The error at the query is $E_0\Pi_{N(Q)}$, where Π_n is the product of the first $n+1$ update matrices, determined by the keys $Y_0, \dots, Y_n$. By Lemma C.2, $N(Q)$ is independent of (Y_n) and $\Pr(N(Q) = n) = \frac{1}{1+r}\big(\frac{r}{1+r}\big)^n$ with $r = \rho/\lambda$, and by Lemma E.1(iii) the law of (Y_n) depends only on T. Hence $R = \sum_{n\ge 0} \Pr(N(Q) = n)\, \mathbb{E}\|E_0\Pi_n C\|_F^2$ depends on (ρ, λ) only through r, and $\mathbb{E}[N(Q)] + 1 = 1 + r$.

(ii) Let $t' = \lambda t$, $\widetilde{E}(t') = E(t'/\lambda)$, and $\widetilde{X}_{t'} = X_{t'/\lambda}$, which has rate ρ/λ. Since $\dot{E} = \lambda\widetilde{E}'$ and $\ddot{E} = \lambda^2\widetilde{E}''$, dividing equation 4 by λ^2 gives $\widetilde{E}'' + (\alpha/\lambda)\widetilde{E}' + (\beta/\lambda^2)\widetilde{E}P = 0$ with $\widetilde{E}'(0) = 0$. The query time becomes $\lambda Q \sim \mathrm{Exp}(1)$ and $E(Q) = \widetilde{E}(\lambda Q)$, which proves the identity. Substituting α/λ and β/λ^2 into equation 7 and multiplying numerator and denominator by λ^3 gives $q_{\mathrm{s}}(\lambda)$. □

B.2 KEY GEOMETRY AND ACCESS ORDER

We give one construction in which the same four keys, with the same average similarity, lead to a positive definite M or to an indefinite M depending on how they are assigned to the states, and depending on the access kernel. Take $H(1, 30)$, $\rho = 10$, and RP with $y = 0$, so that $g_0 = q_{\mathrm{s}} = 11/41 > 0$ and the modal margin $g(t)$ is the function of Appendix E.3. It takes the values

$$g\big(-\tfrac{1}{3}\big) = \frac{38887199}{2177469} > 0, \qquad g\big(-\tfrac{2}{3}\big) = \frac{407509901}{12545631} > 0, \qquad g\big(\tfrac{1}{3}\big) = -\frac{46193}{34317} < 0.$$

Two kernels. Let $K = 4$, $d = 3$, and consider the uniform kernel T_{u} with $(T_{\mathrm{u}})_{ij} = \frac{1}{3}$ for $i \neq j$, and the pair kernel

$$T_{\mathrm{p}} = \begin{pmatrix} 0 & \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \\ \frac{2}{3} & 0 & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & 0 & \frac{2}{3} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} & 0 \end{pmatrix},$$

which mostly moves within the pairs $\{1, 2\}$ and $\{3, 4\}$. Both are symmetric, so Assumption 4.1 holds with uniform π. Every nonconstant eigenvalue of T_{u} equals $-\frac{1}{3}$. The kernel T_{p} has the eigenvector $\phi_1 = (1, 1, -1, -1)$ with eigenvalue $\frac{1}{3}$, which separates the pairs, and the eigenvectors $\phi_2 = \sqrt{2}(1, -1, 0, 0)$ and $\phi_3 = \sqrt{2}(0, 0, 1, -1)$ with eigenvalue $-\frac{2}{3}$, which alternate within a pair.

Two assignments of the same keys. Use the four weak components $(1, 1)$, $(1, -1)$, $(-1, 1)$, $(-1, -1)$, and assign them to the states $1, 2, 3, 4$ in two orders:

aligned: $(1, 1)$, $(1, -1)$, $(-1, 1)$, $(-1, -1)$, crossed: $(1, 1)$, $(-1, -1)$, $(1, -1)$, $(-1, 1)$.

Both assignments use the same four keys, so their Gram matrices differ by a permutation of the states and have the same average similarity. In both, $\mu = 0$ and $\Sigma = I$.

Under the aligned assignment, the first weak coordinate is $(1, 1, -1, -1) = \phi_1$ and the second is $(1, -1, 1, -1) = (\phi_2 + \phi_3)/\sqrt{2}$. Hence $\mu_1 = e_1$ and $\mu_2 = \mu_3 = e_2/\sqrt{2}$. Under the crossed assignment, the two coordinates are $(1, -1, 1, -1) = (\phi_2 + \phi_3)/\sqrt{2}$ and $(1, -1, -1, 1) = (\phi_2 - \phi_3)/\sqrt{2}$, so $\mu_1 = 0$, $\mu_2 = (e_1 + e_2)/\sqrt{2}$, and $\mu_3 = (e_1 - e_2)/\sqrt{2}$.

The compatibility matrices. Under T_{p},

$$M_{\mathrm{al}} = g\big(\tfrac{1}{3}\big)e_1e_1^\top + g\big(-\tfrac{2}{3}\big)e_2e_2^\top, \qquad M_{\mathrm{cr}} = g\big(-\tfrac{2}{3}\big)\, I.$$

Under T_{u}, every nonconstant eigenvalue is $-\frac{1}{3}$, so $M = g(-\frac{1}{3})\,\Sigma = g(-\frac{1}{3})\,I$ for either assignment. By Proposition 4.2, this gives two conclusions for all sufficiently small $\varepsilon > 0$.

(i) *Key geometry.* Under T_{p}, RP attains lower retrieval risk than $H(1, 30)$ for every nonzero initial error with the crossed assignment, while $H(1, 30)$ attains lower retrieval risk from the initial error $(0, e_1)$ with the aligned assignment.

(ii) *Access order.* With the aligned assignment, RP attains lower retrieval risk than $H(1, 30)$ for every nonzero initial error under T_{u}, while $H(1, 30)$ attains lower retrieval risk from $(0, e_1)$ under T_{p}.

In both cases the losing direction e_1 is the weak coordinate that varies between the pairs, which is the slow mode of T_{p}.

B.3 The online rule in the indefinite regime

Setting. Consider the slots $n = 0, 1, \ldots, \bar{n}$. At slot n a pair (k_n, v_n) with $v_n = W_\star k_n$ is revealed and each memory is updated. The query happens after the update at slot n with conditional probability p_n, given that it has not happened before, and $p_{\bar{n}} = 1$. Let $\mathcal{F}_n$ contain the current memory and the pairs revealed up to slot n, and assume that the law of the next pair, the probability p_n, and the conditional second moment $\mathcal{Q}_n$ below depend on $\mathcal{F}_n$ only through the revealed pairs. At the query slot, the query direction ψ has conditional second moment $\mathcal{Q}_n = \mathbb{E}[\psi\psi^\top \mid \mathcal{F}_n, \text{query at } n]$, and the target is $W_\star\psi$.

The baseline projects at every slot. The online rule reflects at slot 0, takes no update at the slots $n \geq 1$ before the first slot at which a fixed predicate of the revealed pair (k_n, v_n) holds, and projects at that slot and at every later slot. Write $E_n^{\mathrm{P}} = W_n^{\mathrm{P}} - W_\star$ and $\widetilde{E}_n = \widetilde{W}_n - W_\star$ for the two errors after slot n, and $\Delta_n = \widetilde{E}_n - E_n^{\mathrm{P}}$. Let n_{q} be the query slot. Since $\|W\psi - W_\star\psi\|^2 = \|(W - W_\star)\psi\|^2$, the conditional margin is

$$\Gamma(\mathcal{F}_0) = \mathbb{E}\big[\|E_{n_{\mathrm{q}}}^{\mathrm{P}}\psi\|^2 - \|\widetilde{E}_{n_{\mathrm{q}}}\psi\|^2 \;\big|\; \mathcal{F}_0\big].$$

Lemma B.3 (Gain at the query)**.** *For matrices E and Δ and a query with second moment $\mathcal{Q}$,*

$$\mathbb{E}\|E\psi\|^2 - \mathbb{E}\|(E+\Delta)\psi\|^2 = \mathrm{gain}(\Delta, E, \mathcal{Q}), \qquad \mathrm{gain}(\Delta, E, \mathcal{Q}) = -2\,\mathrm{tr}(\Delta\mathcal{Q}E^\top) - \mathrm{tr}(\Delta\mathcal{Q}\Delta^\top).$$

Proof. $\mathbb{E}\|X\psi\|^2 = \mathrm{tr}(X\mathcal{Q}X^\top)$ for every matrix X. Expanding $\mathrm{tr}\big((E+\Delta)\mathcal{Q}(E+\Delta)^\top\big)$ and using $\mathrm{tr}(E\mathcal{Q}\Delta^\top) = \mathrm{tr}(\Delta\mathcal{Q}E^\top)$ gives the claim. □

Proposition B.4 (Exact evaluation of Γ)**.** *For each row, stack the two error rows into $\zeta_n = (f_n^{\mathrm{P}}, \tilde{f}_n) \in \mathbb{R}^{2d}$, and let A_n^{P} and $\widetilde{A}_n$ be the update matrices of the two rules at slot n, so that $\zeta_n = \Xi_n\zeta_{n-1}$ with $\Xi_n = \mathrm{diag}(A_n^{\mathrm{P}}, \widetilde{A}_n)$. Define $\Lambda_n = \mathrm{diag}(\mathcal{Q}_n, -\mathcal{Q}_n)$ and the value matrices*

$$\Theta_{\bar{n}} = \Lambda_{\bar{n}}, \qquad \Theta_n = p_n\Lambda_n + (1-p_n)\,\mathbb{E}\big[\Xi_{n+1}^\top\Theta_{n+1}\Xi_{n+1} \;\big|\; \mathcal{F}_n,\ \textit{no query at } n\big].$$

Then Θ_n depends only on the revealed pairs, and $\Gamma(\mathcal{F}_0) = \sum_{i=1}^m \zeta_{0,i}^\top\Theta_0\,\zeta_{0,i}$, where $\zeta_{0,i}$ stacks the ith rows of E_0^{P} and $\widetilde{E}_0$.

Proof. The predicate uses only the revealed pair, so whether the online rule updates at slot n, and hence Ξ_n, is a function of the revealed pairs. Together with the assumption on the laws of the pairs and of the query, this shows by backward induction that Θ_n depends only on the revealed pairs.

Let Γ_n be the conditional expectation of the gain given $\mathcal{F}_n$ and no query before slot n. We show $\Gamma_n = \sum_i \zeta_{n,i}^\top\Theta_n\zeta_{n,i}$ by backward induction. At $n = \bar{n}$ the query happens, and Lemma B.3 applied row by row gives $\Gamma_{\bar{n}} = \sum_i \zeta_{n,i}^\top\Lambda_{\bar{n}}\zeta_{n,i}$. For $n < \bar{n}$, the query happens at slot n with probability p_n, contributing $\sum_i \zeta_{n,i}^\top\Lambda_n\zeta_{n,i}$. Otherwise the next slot is reached with $\zeta_{n+1,i} = \Xi_{n+1}\zeta_{n,i}$, and by the induction hypothesis the contribution is $\mathbb{E}\big[\sum_i \zeta_{n,i}^\top\Xi_{n+1}^\top\Theta_{n+1}\Xi_{n+1}\zeta_{n,i} \mid \mathcal{F}_n, \text{no query at } n\big]$. Adding the two contributions gives the recursion. Taking $n = 0$ proves the claim. □

Consequently, $\Gamma(\mathcal{F}_0) > 0$ means that the online rule attains lower expected retrieval error than the baseline at the query, conditionally on the current state. When the laws of the pairs and of the query are only known to lie in a set, the infimum of Γ over that set is a lower bound.

A nonempty region of states. Let $m = 1$, $d = 2$, and $W_\star = 0$, so that the error equals the memory. Let the current key be $k_0 = e_1$ and the current error be the row $(x, -b)$ with $x, b > 0$. At slots $n \geq 1$ the key equals e_1 until the first slot at which it equals some $k' = (c', s')$ with $c', s' > 0$. The predicate holds exactly when $k_n \neq e_1$, the query direction is $\psi = e_1$, and the query happens immediately after the hit if it has not happened before.

At slot 0, projection gives $E_0^{\mathrm{P}} = (0, -b)$ and reflection gives $\widetilde{E}_0 = (-x, -b)$, so $\Delta_0 = (-x, 0)$. While the key repeats, the baseline projects onto e_1 again, which leaves $(0, -b)$ unchanged, and the online rule takes no update. A query before the hit, or a forced query without a hit, therefore gives $\mathrm{gain}(\Delta_0, E_0^{\mathrm{P}}, e_1 e_1^\top) = 0 - x^2 = -x^2$. At a hit with key k', both rules project onto k'. With $I - P_{k'} = \left(\begin{smallmatrix} s'^2 & -c's' \\ -c's' & c'^2 \end{smallmatrix}\right)$, the query values are $E^{\mathrm{P}} e_1 = bc's'$ and $\Delta_0 e_1 = -xs'^2$, so the gain is

$$(bc's')^2 - (bc's' - xs'^2)^2 = xs'^3\,(2bc' - xs').$$

If a hit with key k' happens before the query with probability p_{hit}, then

$$\Gamma(\mathcal{F}_0) = p_{\mathrm{hit}}\, xs'^3(2bc' - xs') - (1 - p_{\mathrm{hit}})\, x^2.$$

For $x = 1$, $b = 3$, $k' = (1, 1)/\sqrt{2}$, and $p_{\mathrm{hit}} = \frac{1}{2}$, this equals $\frac{1}{2} \cdot \frac{5}{4} - \frac{1}{2} = \frac{1}{8} > 0$. Since Γ is continuous in (x, b), it is positive on an open neighborhood of this state.

C Model assumptions and auxiliary results

C.1 Model assumptions and auxiliary lemmas

C.1.1 Model assumptions

The analysis uses a realizable, noiseless target $W_\star$ shared by all key–value pairs. With noisy values, randomized Kaczmarz converges only to a neighborhood of the least-squares solution (Needell, 2010), and removing this neighborhood requires extended variants or tail averaging (Zouzias & Freris, 2013; Epperly et al., 2026). Momentum under a time-varying target carries an additional tracking error (Bernardi et al., 2026). The access kernel T is reversible. Gradient methods under non-reversible Markov sampling are analyzed by Sun et al. (2018), and a non-reversible T need not have a real eigenbasis orthonormal under π, so the access modes ϕ_ℓ and the compatibility matrix M would need a different construction.

C.1.2 Auxiliary lemmas

This subsection collects three facts used throughout the proofs. The first reduces every risk to a quadratic form in the initial error, the second gives the law of the number of key changes before the query, and the third solves a delta rule on an interval with a fixed key.

Lemma C.1 (Risk as a quadratic form)**.** *Fix a rule $a \in \mathcal{D} \cup \{H, \mathrm{P}, \mathrm{RP}\}$. For every realization of the access process X there is a matrix $\Phi_a(t) \in \mathbb{R}^{d\times d}$, which does not depend on the initial error E_0, such that $E_a(t) = E_0\Phi_a(t)$ for all $t \geq 0$. Consequently, if*

$$\mathcal{R}_a = \mathbb{E}\big[\Phi_a(Q)\, CC^\top\, \Phi_a(Q)^\top\big]$$

is finite, then

$$R(a;\rho) = \sum_{i=1}^{m} f_{0,i}^\top\, \mathcal{R}_a\, f_{0,i}, \tag{14}$$

where $f_{0,i}^\top$ is the ith row of E_0.

Proof. For P and RP, each update multiplies the error on the right by $I - P_k$ or $I - 2P_k$, and the error is constant between updates. Hence $\Phi_a(t)$ is the product of these matrices in the order of the updates up to time t.

For $d \in \mathcal{D}$, the learning rate γ_t depends on the keys and values only, and every value equals $v = W_\star k$. Hence γ_t does not depend on E_0. Let Φ solve $\dot\Phi_t = -\gamma_t\Phi_t P_t$ with $\Phi_0 = I$. Then $E_0\Phi_t$ solves equation 3 with the initial value E_0, and the solution of this linear equation with locally integrable coefficients is unique. Hence $E_t = E_0\Phi_t$.

For $H(\alpha,\beta)$, let Φ solve $\ddot\Phi_t + \alpha\dot\Phi_t + \beta\Phi_t P_t = 0$ with $\Phi_0 = I$ and $\dot\Phi_0 = 0$. On a bounded interval the path X has finitely many key changes almost surely, so P_t is piecewise constant and the solution is unique. The matrix $E_0\Phi_t$ solves equation 4 with $E_0\Phi_0 = E_0$ and $E_0\dot\Phi_0 = 0$. Hence $E_t = E_0\Phi_t$.

For the risk, the ith row of $E_a(Q)C$ is $f_{0,i}^\top\Phi_a(Q)C$. Summing the squared row norms gives

$$\big\|E_a(Q)C\big\|_F^2 = \sum_{i=1}^{m} f_{0,i}^\top\, \Phi_a(Q)CC^\top\Phi_a(Q)^\top\, f_{0,i}.$$

Taking the expectation over X and Q gives equation 14. □

By Lemma C.1, the rows of the memory do not interact, and it suffices to study one error row $f_0 \in \mathbb{R}^d$ with risk $R_a(f_0) = f_0^\top\mathcal{R}_a f_0$. This is the reduction used in Sections 3 and 4.

Lemma C.2 (Number of key changes before the query)**.** *Let the access process have generator $\rho(T - I)$ with $T_{ii} = 0$, and let $Q \sim \mathrm{Exp}(\lambda)$ be independent of it. Then the key-change times form a Poisson process of rate ρ that is independent of the visited keys. The number $N(Q)$ of key changes in $[0, Q)$ is independent of the sequence of visited keys and satisfies*

$$\Pr\big(N(Q) = n\big) = \frac{\lambda}{\lambda+\rho}\Big(\frac{\rho}{\lambda+\rho}\Big)^n, \qquad n \geq 0. \tag{15}$$

In particular, for $|x| \le 1$,

$$\mathbb{E}\big[x^{N(Q)}\big] = \frac{\lambda}{\lambda + \rho - \rho x}.$$

Proof. Since $T_{ii} = 0$, every jump of the chain changes the key, and the holding time in every state is $\mathrm{Exp}(\rho)$. The holding times are independent of each other and of the sequence of visited keys (Norris, 1997), so the key-change times $\tau_1 < \tau_2 < \cdots$ form a Poisson process of rate ρ that is independent of the visited keys. Because Q is independent of this process,

$$\Pr\big(N(Q) \ge n\big) = \Pr(Q > \tau_n) = \mathbb{E}\big[e^{-\lambda \tau_n}\big] = \Big(\frac{\rho}{\rho + \lambda}\Big)^n,$$

where the last equality uses that τ_n is a sum of n independent $\mathrm{Exp}(\rho)$ variables, each with Laplace transform $\rho/(\rho + \lambda)$ at λ. Subtracting $\Pr(N(Q) \ge n + 1)$ from $\Pr(N(Q) \ge n)$ gives equation 15. Independence of $N(Q)$ from the visited keys follows because $N(Q)$ is a function of the key-change times and Q alone. For $|x| \le 1$, the geometric series gives

$$\mathbb{E}\big[x^{N(Q)}\big] = \frac{\lambda}{\lambda + \rho} \sum_{n=0}^{\infty} \Big(\frac{\rho x}{\lambda + \rho}\Big)^n = \frac{\lambda}{\lambda + \rho - \rho x}.$$

□

Lemma C.3 (Delta rule on a fixed key)**.** *Let $d \in \mathcal{D}$ and suppose the key k is fixed on an interval $[t_0, t_1]$. Then for $t \in [t_0, t_1]$,*

$$E_t = E_{t_0}\big(e^{-\bar\gamma} I + (1 - e^{-\bar\gamma})(I - P_k)\big), \qquad \bar\gamma = \int_{t_0}^{t} \gamma_s \, ds \in [0, \infty),$$

so $E_t = E_{t_0} A_k (1 - e^{-\bar\gamma})$ with $1 - e^{-\bar\gamma} \in [0, 1)$.

Proof. On $[t_0, t_1]$, equation 3 reads $\dot E_t = -\gamma_t E_t P_k$ with a fixed matrix P_k. Its solution is $E_t = E_{t_0} \exp(-\bar\gamma P_k)$, which can be verified by differentiating in t. Since $P_k^2 = P_k$, we have $P_k^n = P_k$ for every $n \ge 1$, and

$$\exp(-\bar\gamma P_k) = I + \sum_{n=1}^{\infty} \frac{(-\bar\gamma)^n}{n!} P_k = I + (e^{-\bar\gamma} - 1) P_k = e^{-\bar\gamma} I + (1 - e^{-\bar\gamma})(I - P_k).$$

Local integrability of γ gives $\bar\gamma < \infty$, hence $e^{-\bar\gamma} \in (0, 1]$ and $1 - e^{-\bar\gamma} \in [0, 1)$. □

C.2 Discrete momentum and the momentum rule

The momentum rule $H(\alpha, \beta)$ of Definition 2.3 is a second-order flow, while the update equation 1 is discrete. They are related by the standard finite-difference form of heavy-ball momentum (Qian, 1999). Let $h > 0$ be the time between two updates, and identify the nth update with time nh.

Proposition C.4 (Correspondence)**.** *Let the discrete momentum rule of Section 2.1 use $\eta = 1 - \alpha h$ and $\theta = \beta h^2$, and let $E_n = W_n - W_\star$. Then, for $n \ge 1$,*

$$\frac{E_n - 2E_{n-1} + E_{n-2}}{h^2} + \alpha \, \frac{E_{n-1} - E_{n-2}}{h} + \beta \, E_{n-1} P_{k_n} = 0,$$

with $E_{-1} = E_0$. This is the explicit finite-difference scheme for $\ddot E_t + \alpha \dot E_t + \beta E_t P_t = 0$, and the condition $S_0 = 0$ corresponds to $\dot E_0 = 0$.

Proof. Since $v_n = W_\star k_n$, the residual correction is $U_n(W_{n-1}) = -E_{n-1} P_{k_n}$, and $S_n = W_n - W_{n-1} = E_n - E_{n-1}$. Setting $E_{-1} = E_0$ makes $S_0 = 0$. The momentum update then reads

$$E_n - E_{n-1} = \eta \, (E_{n-1} - E_{n-2}) - \theta \, E_{n-1} P_{k_n}.$$

Subtracting $E_{n-1} - E_{n-2}$ from both sides gives

$$(E_n - 2E_{n-1} + E_{n-2}) + (1 - \eta)(E_{n-1} - E_{n-2}) + \theta \, E_{n-1} P_{k_n} = 0.$$

Dividing by h^2 and inserting $1 - \eta = \alpha h$ and $\theta = \beta h^2$ gives the claim. □

For example, $H(1, 30)$ with $h = 0.1$ corresponds to $\eta = 0.9$ and $\theta = 0.3$. The proposition states a formal consistency. How closely the discrete rule follows the flow depends on h relative to the timescales $1/\rho$, $1/\alpha$, and $1/\sqrt{\beta}$.

Appendix A.4.1 measures this on the 740 CIFAR pairs at $h \in \{0.1, 0.05, 0.025, 0.0125\}$. The largest mean gap between a discrete risk and its continuous-time value roughly halves with h. At $h = 0.0125$, the discrete rules keep the continuous-time comparison between RP and momentum in all four rate cells. At $h = 0.1$, the momentum rule $H(4, 100)$ has $\theta = 1$ and attains lower reconstruction risk than RP in all four cells. These rules keep the flow parameters fixed as h changes. When every rule updates at every step and momentum is tuned separately for each h, the comparison changes, as reported in the second part of Appendix A.4.1.

D PROOFS FOR THE TWO-KEY SETTING

Throughout this section $C = I_2$ and $\lambda = 1$. By Lemma C.1 we follow one error row $f_0 \in \mathbb{R}^2$, which evolves as $f_t = \Phi(t)^\top f_0$, and write $R_a(f_0) = \mathbb{E}\|f_a(Q)\|_2^2$. Theorem 3.6 compares one momentum setting. For the whole family, see the frontier of Appendix D.3.1, the matched comparison of Appendix D.3.2, and the tuned fixed-ε comparison of Appendix D.4.2.

D.1 TWO-KEY GEOMETRY

Define the unit vectors orthogonal to the keys and the coordinate reflection

$$n_+ = (-\sin\varphi, \cos\varphi)^\top, \qquad n_- = (\sin\varphi, \cos\varphi)^\top, \qquad Z = \mathrm{diag}(1, -1).$$

Lemma D.1 (Two-key identities). *For $\sigma \in \{+, -\}$ the following identities hold.*

(i) $I - P_{k_\sigma} = n_\sigma n_\sigma^\top$.

(ii) $n_+^\top n_- = \cos 2\varphi = c$ *and* $n_\sigma^\top e_{\mathrm{w}} = \cos\varphi$.

(iii) $Zk_\sigma = k_{-\sigma}$, $ZP_{k_\sigma}Z = P_{k_{-\sigma}}$, *and* $Zn_\sigma = -n_{-\sigma}$.

Proof. For (i), $\{k_\sigma, n_\sigma\}$ is an orthonormal basis of $\mathbb{R}^2$, so $I = k_\sigma k_\sigma^\top + n_\sigma n_\sigma^\top = P_{k_\sigma} + n_\sigma n_\sigma^\top$. For (ii), $n_+^\top n_- = -\sin^2\varphi + \cos^2\varphi = \cos 2\varphi$, and the second entry of both n_+ and n_- is $\cos\varphi$. For (iii), Z flips the sign of the second coordinate, which maps k_+ to k_-, k_- to k_+, n_+ to $-n_-$, and n_- to $-n_+$. Then $ZP_{k_\sigma}Z = (Zk_\sigma)(Zk_\sigma)^\top = k_{-\sigma}k_{-\sigma}^\top = P_{k_{-\sigma}}$. □

Lemma D.2 (Sign symmetry). *For $a \in \{H, \mathrm{P}, \mathrm{RP}\}$ the risk matrix is diagonal, $\mathcal{R}_a = \mathrm{diag}\big(R_a(e_{\mathrm{s}}), R_a(e_{\mathrm{w}})\big)$. Hence for $f_0 = a_0 e_{\mathrm{s}} + b_0 e_{\mathrm{w}}$,*

$$R_a(f_0) = a_0^2 R_a(e_{\mathrm{s}}) + b_0^2 R_a(e_{\mathrm{w}}).$$

Proof. Let X' be the path obtained from X by exchanging k_+ and k_-. Since the chain is symmetric and π is uniform, X' has the same law as X, and it has the same key-change times. By Lemma D.1(iii), the projector along X' is ZP_tZ.

For P and RP, $\Phi_a[X]$ is a product of factors $I - \theta P_k$, and $Z(I - \theta P_k)Z = I - \theta ZP_kZ$. Inserting $Z^2 = I$ between consecutive factors gives $\Phi_a[X'] = Z\Phi_a[X]Z$. For H, the matrix $Z\Phi_H[X]Z$ satisfies the momentum equation with projector ZP_tZ and the initial values $ZIZ = I$ and 0. By uniqueness, $\Phi_H[X'] = Z\Phi_H[X]Z$.

Since X and X' have the same law,

$$\mathcal{R}_a = \mathbb{E}\big[\Phi_a[X'](Q)\Phi_a[X'](Q)^\top\big] = Z\,\mathbb{E}\big[\Phi_a[X](Q)\Phi_a[X](Q)^\top\big]\,Z = Z\mathcal{R}_aZ.$$

Conjugation by Z flips the sign of the off-diagonal entries and keeps the diagonal entries. Hence the off-diagonal entries of $\mathcal{R}_a$ vanish. The diagonal entries are $e_{\mathrm{s}}^\top\mathcal{R}_a e_{\mathrm{s}} = R_a(e_{\mathrm{s}})$ and $e_{\mathrm{w}}^\top\mathcal{R}_a e_{\mathrm{w}} = R_a(e_{\mathrm{w}})$. □

D.2 PROOFS OF THE RESULTS OF SECTION 3

D.2.1 PROOF OF THEOREM 3.1

Proof. Fix $d \in \mathcal{D}$ and $f_0 = e_{\mathrm{w}}$.

Step 1: path representation. Condition on a path with $n = N(Q)$ key changes before Q. Split $[0, Q)$ into the $n + 1$ intervals between key changes, and let $\sigma_i \in \{+, -\}$ be the key on the ith interval, $i = 0, \dots, n$, so that $\sigma_{i+1} = -\sigma_i$. Let $\bar\gamma_i$ be the integral of γ_t over the ith interval and $z_i = e^{-\bar\gamma_i} \in (0, 1]$. By Lemma C.3 and Lemma D.1(i), the error row at the query is

$$f_d(Q) = A_n A_{n-1} \cdots A_0\, e_{\mathrm{w}}, \qquad A_i = z_i I + (1 - z_i)\, n_{\sigma_i} n_{\sigma_i}^\top,$$

where we used that each A_i is symmetric.

Step 2: expansion of the product. Expanding each factor A_i into its two terms gives

$$A_n \cdots A_0 = \sum_{J \subseteq \{0,\dots,n\}} w_J \, B_J, \qquad w_J = \prod_{i \in J} (1 - z_i) \prod_{i \notin J} z_i \;\geq 0,$$

where $B_\varnothing = I$ and, for $J = \{i_1 < \cdots < i_j\}$, $B_J = n_{\sigma_{i_j}} n_{\sigma_{i_j}}^\top \cdots n_{\sigma_{i_1}} n_{\sigma_{i_1}}^\top$. The weights sum to $\prod_{i=0}^{n} \big(z_i + (1 - z_i)\big) = 1$.

Step 3: the component along n_{σ_n}. For $J = \{i_1 < \cdots < i_j\} \neq \varnothing$,

$$n_{\sigma_n}^\top B_J \, e_{\mathrm{w}} = \big(n_{\sigma_n}^\top n_{\sigma_{i_j}}\big) \prod_{a=1}^{j-1} \big(n_{\sigma_{i_{a+1}}}^\top n_{\sigma_{i_a}}\big) \big(n_{\sigma_{i_1}}^\top e_{\mathrm{w}}\big).$$

By Lemma D.1(ii), each inner product $n_\sigma^\top n_{\sigma'}$ equals 1 if $\sigma = \sigma'$ and c otherwise, and $n_{\sigma_{i_1}}^\top e_{\mathrm{w}} = \cos\varphi$. Hence $n_{\sigma_n}^\top B_J e_{\mathrm{w}} = \cos\varphi \, c^{q_J}$, where q_J is the number of sign changes in the sequence $(\sigma_{i_1}, \dots, \sigma_{i_j}, \sigma_n)$. This sequence is a subsequence of $(\sigma_0, \dots, \sigma_n)$, which has exactly n sign changes, so $q_J \leq n$. For $J = \varnothing$, $n_{\sigma_n}^\top e_{\mathrm{w}} = \cos\varphi$, so $q_\varnothing = 0$. Since $c \in (0, 1)$ and the weights are nonnegative and sum to one,

$$n_{\sigma_n}^\top f_d(Q) = \cos\varphi \sum_J w_J c^{q_J} \geq \cos\varphi \, c^n.$$

Because n_{σ_n} is a unit vector, the Cauchy–Schwarz inequality gives

$$\|f_d(Q)\|_2^2 \;\geq\; \big(n_{\sigma_n}^\top f_d(Q)\big)^2 \;\geq\; \cos^2\varphi \, c^{2N(Q)}. \tag{16}$$

Step 4: strict inequality. If $n \geq 1$, the empty set has weight $w_\varnothing = \prod_i z_i > 0$ and $c^{q_\varnothing} = 1 > c^n$, so the second inequality in equation 16 is strict. If $n = 0$, let $s_0 = k_{\sigma_0}^\top e_{\mathrm{w}}$, which equals $\sin\varphi$ for $\sigma_0 = +$ and $-\sin\varphi$ for $\sigma_0 = -$. Then $e_{\mathrm{w}} = \cos\varphi \, n_{\sigma_0} + s_0 \, k_{\sigma_0}$ and $A_0 e_{\mathrm{w}} = \cos\varphi \, n_{\sigma_0} + z_0 s_0 \, k_{\sigma_0}$, so

$$\|f_d(Q)\|_2^2 = \cos^2\varphi + z_0^2 \sin^2\varphi > \cos^2\varphi,$$

since $z_0 > 0$ and $\sin\varphi > 0$. Hence equation 16 is strict for almost every path, and taking expectations gives $R_d^{\mathrm{w}} > \cos^2\varphi \, \mathbb{E}[c^{2N(Q)}]$.

Step 5: evaluation. Lemma C.2 with $\lambda = 1$ and $x = c^2$ gives $\mathbb{E}[c^{2N(Q)}] = 1/(1 + \rho - \rho c^2)$. Substituting $\cos^2\varphi = 1/(1 + \varepsilon^2)$ and

$$1 - c^2 = \frac{(1 + \varepsilon^2)^2 - (1 - \varepsilon^2)^2}{(1 + \varepsilon^2)^2} = \frac{4\varepsilon^2}{(1 + \varepsilon^2)^2}$$

yields

$$\cos^2\varphi \, \mathbb{E}[c^{2N(Q)}] = \frac{1}{1 + \varepsilon^2} \cdot \frac{(1 + \varepsilon^2)^2}{(1 + \varepsilon^2)^2 + 4\rho\varepsilon^2} = \frac{1 + \varepsilon^2}{(1 + \varepsilon^2)^2 + 4\rho\varepsilon^2} = L_\varepsilon.$$

Since $(1 + \varepsilon^2)^2 + 4\rho\varepsilon^2 = 1 + (2 + 4\rho)\varepsilon^2 + O(\varepsilon^4)$,

$$L_\varepsilon = (1 + \varepsilon^2)\big(1 - (2 + 4\rho)\varepsilon^2 + O(\varepsilon^4)\big) = 1 - (1 + 4\rho)\varepsilon^2 + O(\varepsilon^4).$$

Step 6: the infimum is approached and not attained. Take the constant learning rate $\gamma_t = g$, which belongs to $\mathcal{D}$. Every interval between key changes has positive length almost surely, so $z_i \to 0$ as $g \to \infty$ and

$$A_n \cdots A_0 \to n_{\sigma_n} n_{\sigma_n}^\top \cdots n_{\sigma_0} n_{\sigma_0}^\top = c^n \, n_{\sigma_n} n_{\sigma_0}^\top,$$

where the equality uses Lemma D.1(ii) once for each key change. Hence $f_d(Q) \to c^n \cos\varphi \, n_{\sigma_n}$ and $\|f_d(Q)\|_2^2 \to \cos^2\varphi \, c^{2N(Q)}$. Each A_i has operator norm at most one, so $\|f_d(Q)\|_2^2 \leq 1$, and dominated convergence gives $R_d^{\mathrm{w}} \to L_\varepsilon$. Together with Step 4, this shows $\inf_{d \in \mathcal{D}} R_d^{\mathrm{w}} = L_\varepsilon$ and that no $d \in \mathcal{D}$ attains it. □

Adaptive learning rates. The proof uses only that $\gamma_t \geq 0$ and that γ_t is integrable on each interval between key changes, and it argues path by path. Hence the lower bound $R_d^{\mathrm{w}} > L_\varepsilon$ also holds when γ_t depends on the memory state or on past residuals, which covers adaptive learning rates. Such a rate may depend on E_0, so Lemma C.1 no longer gives a single matrix $\Phi_d(t)$ for all initial errors, and the risk need not be a quadratic form in f_0. The bound is then a statement for the weak initial error e_{w} only.

D.2.2 Proof of Proposition 3.2

Let $\xi_t \in \{+1, -1\}$ be the sign of the active key, so that $k_{\xi_t} = (\cos\varphi, \xi_t \sin\varphi)^\top$. Transposing equation 4 and using $P_k^\top = P_k$, one error row satisfies

$$\ddot{f}_t + \alpha \dot{f}_t + \beta P_{k_{\xi_t}} f_t = 0, \qquad \dot{f}_0 = 0.$$

With the state $x_t = (f_t^\top, \dot{f}_t^\top)^\top \in \mathbb{R}^4$ this reads $\dot{x}_t = \mathcal{G}_{\xi_t}(\varepsilon)\, x_t$, where

$$\mathcal{G}_\sigma(\varepsilon) = \begin{pmatrix} 0 & I_2 \\ -\beta P_{k_\sigma} & -\alpha I_2 \end{pmatrix}, \qquad \sigma \in \{+1, -1\}.$$

At $\varepsilon = 0$ both keys equal e_s, and we write $\mathcal{G}_0 = \mathcal{G}_\sigma(0)$.

Lemma D.3 (Discounted stability and smooth dependence on ε)**.** *Let V_0 be the positive definite solution of the Lyapunov equation*

$$\big(\mathcal{G}_0 - \tfrac{1}{2} I\big)^\top V_0 + V_0 \big(\mathcal{G}_0 - \tfrac{1}{2} I\big) = -I, \tag{17}$$

and let $\varepsilon_0 = \min\{1,\, 1/(4\beta\|V_0\|)\}$ and $\delta = 1/(2\|V_0\|)$. For $|\varepsilon| \le \varepsilon_0$ and every path of the access process, the following hold.

(i) $\|x_t\|^2 \le \mathrm{cond}(V_0)\, e^{(1-\delta)t} \|x_0\|^2$ *for all $t \ge 0$, where* $\mathrm{cond}(V_0) = \|V_0\|\, \|V_0^{-1}\|$.

(ii) For every integer $m \ge 0$, the derivative $\partial_\varepsilon^m x_t$ exists and $\|\partial_\varepsilon^m x_t\| \le \kappa_m (1+t)^m e^{(1-\delta)t/2} \|x_0\|$, where κ_m depends on m, α, β only.

(iii) The map $\varepsilon \mapsto R_H(f_0)$ is infinitely differentiable on $|\varepsilon| < \varepsilon_0$, and its derivatives are obtained by differentiating under $\mathbb{E} \int_0^\infty e^{-t} \|f_t\|_2^2\, dt$.

Proof. Existence of V_0. In the coordinates $(f_1, \dot{f}_1)$ and $(f_2, \dot{f}_2)$, the matrix $\mathcal{G}_0$ is block diagonal with blocks $\left(\begin{smallmatrix} 0 & 1 \\ -\beta & -\alpha \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & 1 \\ 0 & -\alpha \end{smallmatrix}\right)$. The eigenvalues of the first block are the roots of $s^2 + \alpha s + \beta$, which have negative real part since $\alpha, \beta > 0$. The eigenvalues of the second block are 0 and $-\alpha$. Hence every eigenvalue of $\mathcal{G}_0 - \frac{1}{2} I$ has real part at most $-\frac{1}{2}$, and equation 17 has a unique solution V_0, which is positive definite (Horn & Johnson, 1991, Chapter 2).

Perturbation bound. $\mathcal{G}_\sigma(\varepsilon) - \mathcal{G}_0$ has the single nonzero block $-\beta(P_{k_\sigma} - e_s e_s^\top)$. The difference of the orthogonal projectors onto two unit vectors at angle φ has spectral norm $\sin\varphi$, so

$$\|\mathcal{G}_\sigma(\varepsilon) - \mathcal{G}_0\| = \beta \sin\varphi \le \beta|\varepsilon|.$$

Proof of (i). Let $\nu_t = x_t^\top V_0 x_t$ and $\Omega_t = \mathcal{G}_{\xi_t}(\varepsilon) - \mathcal{G}_0$. By equation 17, $\mathcal{G}_0^\top V_0 + V_0 \mathcal{G}_0 = V_0 - I$. Hence, on each interval between key changes,

$$\dot{\nu}_t = x_t^\top \big(V_0 - I + \Omega_t^\top V_0 + V_0 \Omega_t\big) x_t \le \nu_t - \big(1 - 2\beta|\varepsilon|\|V_0\|\big)\|x_t\|^2 \le \nu_t - \tfrac{1}{2}\|x_t\|^2,$$

where the last inequality uses $|\varepsilon| \le \varepsilon_0$. Since $\nu_t \le \|V_0\|\|x_t\|^2$, we obtain $\dot{\nu}_t \le (1-\delta)\nu_t$. The function ν is continuous across key changes, so Grönwall's inequality gives $\nu_t \le e^{(1-\delta)t}\nu_0$. Combining this with $\|x_t\|^2 \le \|V_0^{-1}\|\nu_t$ and $\nu_0 \le \|V_0\|\|x_0\|^2$ gives (i). The same argument bounds the propagator $\Psi(t,s)$ of $\dot{y} = \mathcal{G}_{\xi_t}(\varepsilon) y$ by $\|\Psi(t,s)\|^2 \le \mathrm{cond}(V_0)\, e^{(1-\delta)(t-s)}$.

Proof of (ii). On a bounded interval the path has finitely many key changes almost surely, and $\mathcal{G}_\sigma(\varepsilon)$ is a rational function of ε without poles on $|\varepsilon| \le 1$. Hence x_t is infinitely differentiable in ε. Differentiating $\dot{x}_t = \mathcal{G}_{\xi_t}(\varepsilon) x_t$ $m \ge 1$ times and using that x_0 does not depend on ε gives, by variation of constants,

$$\partial_\varepsilon^m x_t = \int_0^t \Psi(t,s) \sum_{i=1}^{m} \binom{m}{i} \big(\partial_\varepsilon^i \mathcal{G}_{\xi_s}(\varepsilon)\big)\, \partial_\varepsilon^{m-i} x_s\, ds.$$

The derivatives $\partial_\varepsilon^i \mathcal{G}_\sigma$ are bounded on $|\varepsilon| \le 1$ by constants depending on i and β. Induction on m, with the bound on Ψ from (i), gives (ii).

Proof of (iii). Each derivative $\partial_\varepsilon^m \|f_t\|_2^2$ is a finite sum of inner products of $\partial_\varepsilon^i x_t$ and $\partial_\varepsilon^{m-i} x_t$. By (ii) it is bounded by $\kappa'_m (1+t)^m e^{(1-\delta)t} \|x_0\|^2$, so $e^{-t}|\partial_\varepsilon^m \|f_t\|_2^2| \le \kappa'_m (1+t)^m e^{-\delta t}\|x_0\|^2$, which is integrable in t and does not depend on ε or on the path. Differentiation under the expectation and the integral then follows from dominated convergence. □

Proof of Proposition 3.2. Throughout, let j solve $j'' + \alpha j' + \beta j = 0$ with $j(0) = 0$ and $j'(0) = 1$. Since the roots of $s^2 + \alpha s + \beta$ have negative real part, j and j' decay exponentially, so every boundary term at $t = \infty$ below vanishes. The Laplace transform of j is $\widehat{j}(s) = 1/(s^2 + \alpha s + \beta)$, and for a forcing g the solution of $a'' + \alpha a' + \beta a = g$ with $a(0) = a'(0) = 0$ is $a(t) = \int_0^t j(t-s)g(s)\,ds$. Define

$$I = \int_0^\infty e^{-t} j(t)^2\,dt, \qquad J = \int_0^\infty e^{-t} j'(t)^2\,dt.$$

Step 1: parity in ε. Replacing ε by $-\varepsilon$ along the same path replaces P_{k_σ} by $ZP_{k_\sigma}Z$. Hence $Zf_t(\varepsilon)$ solves the equation with $-\varepsilon$ and initial value Zf_0. Since $Ze_{\mathrm{s}} = e_{\mathrm{s}}$ and $Ze_{\mathrm{w}} = -e_{\mathrm{w}}$, the norm $\|f_t\|_2$ started from e_{s} or e_{w} is even in ε. By Lemma D.3(iii), $R_H(e_{\mathrm{s}})$ and $R_H(e_{\mathrm{w}})$ are smooth even functions of ε, so

$$R_H(f_0) = R_H(f_0)\big|_{\varepsilon=0} + \tfrac{1}{2}\,\partial_\varepsilon^2 R_H(f_0)\big|_{\varepsilon=0}\,\varepsilon^2 + O(\varepsilon^4), \qquad f_0 \in \{e_{\mathrm{s}}, e_{\mathrm{w}}\}.$$

By Lemma D.2, these two functions determine $R_H(f_0)$ for every f_0.

Step 2: the strong direction at $\varepsilon = 0$. At $\varepsilon = 0$ every key equals e_{s}. Starting from e_{s}, the second coordinate stays zero, and the first coordinate solves $\ddot{f}_1 + \alpha \dot{f}_1 + \beta f_1 = 0$ with $f_1(0) = 1$ and $\dot{f}_1(0) = 0$. This solution is $f_1 = j' + \alpha j$, as both sides satisfy the same equation and initial values. Hence

$$R_H(e_{\mathrm{s}})\big|_{\varepsilon=0} = \int_0^\infty e^{-t}(j' + \alpha j)^2\,dt = J + 2\alpha \int_0^\infty e^{-t} j j'\,dt + \alpha^2 I = J + \alpha I + \alpha^2 I,$$

where the last equality uses integration by parts, $\int_0^\infty e^{-t} j j'\,dt = \frac{1}{2}\int_0^\infty e^{-t}(j^2)'\,dt = \frac{1}{2}\int_0^\infty e^{-t} j^2\,dt = \frac{1}{2} I$.

Step 3: the values of I and J. Multiplying $j'' = -\alpha j' - \beta j$ by $e^{-t} j$ and integrating,

$$\int_0^\infty e^{-t} j j''\,dt = -\tfrac{\alpha}{2} I - \beta I.$$

Integration by parts gives $\int_0^\infty e^{-t} j j''\,dt = -\int_0^\infty (e^{-t} j)' j'\,dt = \frac{1}{2} I - J$. Hence

$$J = \Big(\beta + \frac{\alpha + 1}{2}\Big) I. \tag{18}$$

Multiplying $j'' = -\alpha j' - \beta j$ by $e^{-t} j'$ and integrating,

$$\int_0^\infty e^{-t} j' j''\,dt = -\alpha J - \tfrac{\beta}{2} I.$$

Integration by parts gives $\int_0^\infty e^{-t} j' j''\,dt = -\frac{1}{2} j'(0)^2 + \frac{1}{2} J = -\frac{1}{2} + \frac{1}{2} J$. Hence

$$\Big(\alpha + \frac{1}{2}\Big) J + \frac{\beta}{2} I = \frac{1}{2}. \tag{19}$$

Substituting equation 18 into equation 19 and using $(2\alpha+1)(\alpha+1+2\beta)+2\beta = (\alpha+1)(2\alpha+4\beta+1)$,

$$I = \frac{2}{(\alpha+1)(2\alpha+4\beta+1)}, \qquad J = \frac{2\beta+\alpha+1}{(\alpha+1)(2\alpha+4\beta+1)}.$$

Inserting these into Step 2 gives

$$R_H(e_{\mathrm{s}})\big|_{\varepsilon=0} = \frac{2\beta+\alpha+1+2\alpha+2\alpha^2}{(\alpha+1)(2\alpha+4\beta+1)} = \frac{(\alpha+1)(2\alpha+1)+2\beta}{(\alpha+1)(2\alpha+4\beta+1)} = q_{\mathrm{s}}.$$

With Step 1 this proves $R_H(e_{\mathrm{s}}) = q_{\mathrm{s}} + O(\varepsilon^2)$.

Step 4: expansion equations along the weak direction. Start from $f_0 = e_{\mathrm{w}}$. At $\varepsilon = 0$ the weak coordinate is not corrected, so $f_t = e_{\mathrm{w}}$ and $R_H(e_{\mathrm{w}})|_{\varepsilon=0} = 1$. In coordinates, the equation of motion reads

$$\ddot{f}_1 + \alpha \dot{f}_1 + \beta\big(\cos^2\varphi\, f_1 + \xi_t \cos\varphi \sin\varphi\, f_2\big) = 0, \qquad \ddot{f}_2 + \alpha \dot{f}_2 + \beta\big(\xi_t \cos\varphi \sin\varphi\, f_1 + \sin^2\varphi\, f_2\big) = 0,$$

with $\cos^2\varphi = 1 + O(\varepsilon^2)$, $\cos\varphi\sin\varphi = \varepsilon/(1+\varepsilon^2) = \varepsilon + O(\varepsilon^3)$, and $\sin^2\varphi = \varepsilon^2/(1+\varepsilon^2) = \varepsilon^2 + O(\varepsilon^4)$. By Step 1, f_1 is odd and f_2 is even in ε. Let $a = \partial_\varepsilon f_1|_{\varepsilon=0}$ and $b = \frac{1}{2}\partial_\varepsilon^2 f_2|_{\varepsilon=0}$. Differentiating the first equation once and the second equation twice at $\varepsilon = 0$ gives

$$a'' + \alpha a' + \beta a = -\beta\xi_t, \qquad b'' + \alpha b' = -\beta(\xi_t a + 1),$$

with $a(0) = a'(0) = b(0) = b'(0) = 0$, since f_0 and $\dot{f}_0$ do not depend on ε. Moreover $\frac{1}{2}\partial_\varepsilon^2\|f_t\|_2^2|_{\varepsilon=0} = a(t)^2 + 2b(t)$. By Step 1 and Lemma D.3(iii),

$$R_H^{\mathrm{w}} = 1 - \bar{c}_{\mathrm{H}}\varepsilon^2 + O(\varepsilon^4), \qquad \bar{c}_{\mathrm{H}} = -\mathbb{E}\int_0^\infty e^{-t}\big(a(t)^2 + 2b(t)\big)\,dt. \tag{20}$$

Step 5: the term in b. Multiplying the equation for b by e^{-t} and integrating, the zero initial values give $\int_0^\infty e^{-t}b''\,dt = \int_0^\infty e^{-t}b'\,dt = \int_0^\infty e^{-t}b\,dt$. Taking expectations,

$$(1+\alpha)\,\mathbb{E}\int_0^\infty e^{-t}b\,dt = -\beta\Big(1 + \mathbb{E}\int_0^\infty e^{-t}\xi_t a(t)\,dt\Big).$$

Step 6: the correlation of the key sign. By Lemma C.2, the number of key changes in $(s,t]$ is Poisson with mean $\rho|t-s|$, and $\xi_t\xi_s = (-1)^{\text{number of changes}}$. Hence $\mathbb{E}[\xi_t\xi_s] = e^{-2\rho|t-s|}$. Since $a(t) = -\beta\int_0^t j(t-s)\xi_s\,ds$,

$$\mathbb{E}[\xi_t a(t)] = -\beta\int_0^t j(u)e^{-2\rho u}\,du.$$

By Fubini's theorem, $\int_0^\infty e^{-t}\int_0^t g(u)\,du\,dt = \int_0^\infty e^{-u}g(u)\,du$ for integrable g. With $z = 1 + 2\rho$,

$$\mathbb{E}\int_0^\infty e^{-t}\xi_t a(t)\,dt = -\beta\int_0^\infty e^{-zu}j(u)\,du = -\beta\,\widehat{j}(z).$$

Combining with Step 5,

$$\mathbb{E}\int_0^\infty e^{-t}\,2b(t)\,dt = -\frac{2\beta}{1+\alpha}\big(1 - \beta\,\widehat{j}(z)\big). \tag{21}$$

Step 7: the term in a^2. From the representation of a and Step 6,

$$\mathbb{E}[a(t)^2] = \beta^2\int_0^t\int_0^t j(t-s)j(t-u)\,e^{-2\rho|s-u|}\,ds\,du.$$

Substitute $p = t-s$ and $q = t-u$, multiply by e^{-t}, and integrate over $t \ge \max(p,q)$ by Fubini's theorem:

$$\mathbb{E}\int_0^\infty e^{-t}a(t)^2\,dt = \beta^2\int_0^\infty\int_0^\infty j(p)j(q)\,e^{-2\rho|p-q|}\,e^{-\max(p,q)}\,dp\,dq.$$

The integrand is symmetric in (p,q). On $p \ge q$ write $p = q + s$ with $s \ge 0$, so that $e^{-2\rho|p-q|}e^{-\max(p,q)} = e^{-q}e^{-zs}$. Hence

$$\mathbb{E}\int_0^\infty e^{-t}a(t)^2\,dt = 2\beta^2\int_0^\infty e^{-zs}\mathcal{J}(s)\,ds, \qquad \mathcal{J}(s) = \int_0^\infty e^{-q}j(q+s)j(q)\,dq.$$

Differentiating under the integral, $\mathcal{J}'' + \alpha\mathcal{J}' + \beta\mathcal{J} = 0$. Its initial values are $\mathcal{J}(0) = I$ and $\mathcal{J}'(0) = \int_0^\infty e^{-q}j'(q)j(q)\,dq = \frac{1}{2}I$ by Step 2. The solutions $j' + \alpha j$ and j of the same equation have initial values $(1,0)$ and $(0,1)$, so

$$\mathcal{J} = I\,(j' + \alpha j) + \tfrac{1}{2}I\,j = I\big(j' + (\alpha + \tfrac{1}{2})j\big).$$

Since $j(0) = 0$, the Laplace transform of j' is $s\,\widehat{j}(s)$. Therefore

$$\mathbb{E}\int_0^\infty e^{-t}a(t)^2\,dt = 2\beta^2 I\big(z + \alpha + \tfrac{1}{2}\big)\widehat{j}(z). \tag{22}$$

Step 8: the coefficient $\bar{c}_{\mathrm{H}}$. Inserting equation 21 and equation 22 into equation 20,

$$\bar{c}_{\mathrm{H}} = \frac{2\beta}{\alpha+1} - 2\beta^2\,\widehat{j}(z)\Big[\frac{1}{\alpha+1} + I\big(z + \alpha + \tfrac{1}{2}\big)\Big].$$

With the value of I from Step 3, the bracket equals

$$\frac{(2\alpha + 4\beta + 1) + 2(z + \alpha + \frac{1}{2})}{(\alpha+1)(2\alpha+4\beta+1)} = \frac{2(z + 2\alpha + 2\beta + 1)}{(\alpha+1)(2\alpha+4\beta+1)}.$$

Substituting $\widehat{j}(z) = 1/(z^2 + \alpha z + \beta)$ gives equation 8. □

D.2.3 Proof of Proposition 3.4

Write $P_\sigma = P_{k_\sigma}$ for $\sigma \in \{+,-\}$.

Lemma D.4 (Block identities). *For $\sigma \in \{+,-\}$,*

$$(I-P_\sigma)(I-P_{-\sigma})(I-P_\sigma) = \cos^2 2\varphi\,(I-P_\sigma), \qquad (I-P_\sigma)(I-2P_{-\sigma})(I-P_\sigma) = \cos 4\varphi\,(I-P_\sigma).$$

Proof. By Lemma D.1(i), $I - P_\sigma = n_\sigma n_\sigma^\top$, so for every matrix X, $(I - P_\sigma)X(I - P_\sigma) = (n_\sigma^\top X n_\sigma)\,(I - P_\sigma)$. For $X = I - P_{-\sigma} = n_{-\sigma}n_{-\sigma}^\top$ the scalar is $(n_\sigma^\top n_{-\sigma})^2 = c^2 = \cos^2 2\varphi$. For $X = I - 2P_{-\sigma}$ the scalar is $1 - 2(k_{-\sigma}^\top n_\sigma)^2$. A direct computation gives $k_-^\top n_+ = -2\sin\varphi\cos\varphi$ and $k_+^\top n_- = 2\sin\varphi\cos\varphi$, so $(k_{-\sigma}^\top n_\sigma)^2 = \sin^2 2\varphi$ and the scalar is $1 - 2\sin^2 2\varphi = \cos 4\varphi$. □

Proof of Proposition 3.4. Condition on the first key σ_0 and on $n = N(Q)$. With two keys the key after the ith change is $\sigma_i = (-1)^i\sigma_0$, and by Lemma C.2 $N(Q)$ is independent of σ_0. Both rules hold the error fixed between key changes, so by Lemma C.1 the error row at the query is $f(Q)^\top = f_0^\top \Pi_n$, where $\Pi_n = A_0A_1\cdots A_n$ and $A_0 = I - P_{\sigma_0}$ is the first update. Every A_i is symmetric, so $\Pi_n\Pi_n^\top = A_0\cdots A_nA_n\cdots A_0$.

Step 1: projection. For P, $A_i = I - P_{\sigma_i}$ for all i. Since $A_n^2 = A_n$ and, by Lemma D.4, $A_{i-1}A_iA_{i-1} = c^2A_{i-1}$ for $i \geq 1$, collapsing the product from the middle outward gives

$$\Pi_n\Pi_n^\top = c^{2n}\,(I - P_{\sigma_0}).$$

Step 2: reflect–project. For RP, $A_i = I - 2P_{\sigma_i}$ for odd i and $A_i = I - P_{\sigma_i}$ for even i. For even $i \geq 2$, the keys satisfy $\sigma_{i-2} = \sigma_i = -\sigma_{i-1}$, so Lemma D.4 gives $A_{i-2}A_{i-1}A_i = \cos 4\varphi\, A_{i-2}$. Hence $\Pi_i = \cos 4\varphi\,\Pi_{i-2}$ for even $i \geq 2$, and by induction $\Pi_{2m} = \cos^m 4\varphi\,(I - P_{\sigma_0})$. For odd $n = 2m+1$, $\Pi_n = \Pi_{2m}A_n$ with $A_n^2 = I$. In both cases

$$\Pi_n\Pi_n^\top = \cos^{2\lfloor n/2\rfloor} 4\varphi\,(I - P_{\sigma_0}).$$

Step 3: expectation. Averaging over σ_0 uniform on $\{+,-\}$ gives $\frac{1}{2}\big((I - P_+) + (I - P_-)\big) = \mathrm{diag}(\sin^2\varphi, \cos^2\varphi)$. By Lemma C.2 with $\lambda = 1$, $\Pr(N(Q) = n) = p_{\mathrm{q}}\,p_{\mathrm{c}}^n$, where $p_{\mathrm{q}} = 1/(1+\rho)$ is the probability that the query arrives before the next key change and $p_{\mathrm{c}} = \rho/(1+\rho)$ is the probability that the key changes first. Hence $\mathcal{R}_{\mathrm{P}} = t_1\,\mathrm{diag}(\sin^2\varphi, \cos^2\varphi)$ with

$$t_1 = \mathbb{E}\big[c^{2N(Q)}\big] = \frac{1}{1 + \rho(1 - \cos^2 2\varphi)} = \frac{1}{1 + \rho\sin^2 2\varphi},$$

and $\mathcal{R}_{\mathrm{RP}} = t_2\,\mathrm{diag}(\sin^2\varphi, \cos^2\varphi)$ with

$$t_2 = \mathbb{E}\big[\cos^{2\lfloor N(Q)/2\rfloor} 4\varphi\big] = \sum_{m\geq 0} p_{\mathrm{q}}\big(p_{\mathrm{c}}^{2m} + p_{\mathrm{c}}^{2m+1}\big)\cos^{2m} 4\varphi$$
$$= \frac{p_{\mathrm{q}}(1 + p_{\mathrm{c}})}{1 - p_{\mathrm{c}}^2\cos^2 4\varphi} = \frac{1 + 2\rho}{1 + 2\rho + \rho^2\sin^2 4\varphi},$$

where the last equality multiplies numerator and denominator by $(1+\rho)^2$. The weak entries are $R_{\mathrm{P}}^{\mathrm{w}} = t_1\cos^2\varphi$ and $R_{\mathrm{RP}}^{\mathrm{w}} = t_2\cos^2\varphi$ with $\cos^2\varphi = 1/(1+\varepsilon^2)$. Since $t_1 = \mathbb{E}[c^{2N(Q)}]$, Step 5 of the proof of Theorem 3.1 gives $R_{\mathrm{P}}^{\mathrm{w}} = L_\varepsilon$.

Step 4: comparison. Both risk matrices equal a scalar times the same positive definite diagonal matrix, so $\mathcal{R}_{\mathrm{P}} - \mathcal{R}_{\mathrm{RP}}$ is positive definite if and only if $t_2 < t_1$. Cross-multiplying and using $\sin^2 4\varphi = 4\sin^2 2\varphi\cos^2 2\varphi$ and $2\cos^2 2\varphi = 1 + \cos 4\varphi$,

$$(1 + 2\rho + \rho^2\sin^2 4\varphi) - (1 + 2\rho)(1 + \rho\sin^2 2\varphi) = \rho\sin^2 2\varphi\,\big(2\rho\cos 4\varphi - 1\big).$$

Since $\sin 2\varphi \neq 0$ for $\varepsilon > 0$, we get $t_2 < t_1$ if and only if $2\rho\cos 4\varphi > 1$.

Step 5: expansion. From $\sin 2\varphi = 2\varepsilon/(1+\varepsilon^2)$ and $\cos 2\varphi = (1-\varepsilon^2)/(1+\varepsilon^2)$, $\sin^2 4\varphi = 16\varepsilon^2 + O(\varepsilon^4)$. Hence $t_2 = 1 - 16\rho^2\varepsilon^2/(1+2\rho) + O(\varepsilon^4)$ and

$$R_{\mathrm{RP}}^{\mathrm{w}} = \frac{t_2}{1+\varepsilon^2} = 1 - \Big(1 + \frac{16\rho^2}{1+2\rho}\Big)\varepsilon^2 + O(\varepsilon^4),$$

which is equation 9. □

D.2.4 Proof of Theorem 3.6

Proof. Step 1: the two weak inequalities. By Theorem 3.1 and Proposition 3.2, $L_\varepsilon - R_H^{\mathrm{w}} = (\bar{c}_{\mathrm{H}} - 1 - 4\rho)\varepsilon^2 + O(\varepsilon^4)$, and by Proposition 3.2 and Proposition 3.4, $R_H^{\mathrm{w}} - R_{\mathrm{RP}}^{\mathrm{w}} = (\bar{c}_{\mathrm{RP}} - \bar{c}_{\mathrm{H}})\varepsilon^2 + O(\varepsilon^4)$. Under equation 10 both leading coefficients are positive, so both differences are positive for all sufficiently small $\varepsilon > 0$. Together with $\inf_{d\in\mathcal{D}} R_d^{\mathrm{w}} = L_\varepsilon$ from Theorem 3.1, this gives equation 11.

Step 2: every nonzero initial error. Write $f_0 = a_0 e_{\mathrm{s}} + b_0 e_{\mathrm{w}}$. By Lemma D.2 applied to H and RP,

$$R_H(f_0) - R_{\mathrm{RP}}(f_0) = a_0^2\big(R_H(e_{\mathrm{s}}) - R_{\mathrm{RP}}(e_{\mathrm{s}})\big) + b_0^2\big(R_H^{\mathrm{w}} - R_{\mathrm{RP}}^{\mathrm{w}}\big).$$

By Proposition 3.2, $R_H(e_{\mathrm{s}}) = q_{\mathrm{s}} + O(\varepsilon^2)$, and $q_{\mathrm{s}} > 0$ since its numerator and denominator in equation 7 are positive. By Step 3 of the proof of Proposition 3.4, $R_{\mathrm{RP}}(e_{\mathrm{s}}) = t_2 \sin^2\varphi = O(\varepsilon^2)$. Hence the first coefficient equals $q_{\mathrm{s}} + O(\varepsilon^2)$, and the second is positive by Step 1. For all sufficiently small $\varepsilon > 0$ both coefficients are positive, and $f_0 \neq 0$ implies $a_0^2 + b_0^2 > 0$. Hence $R_{\mathrm{RP}}(f_0) < R_H(f_0)$.

Step 3: an explicit range of ρ. Take $\alpha = 1$ and $\beta = 30$. Substituting into equation 8 gives

$$\bar{c}_{\mathrm{H}} = \frac{30(82\rho^2 + 103\rho + 16)}{41(2\rho^2 + 3\rho + 16)},$$

and subtracting $1 + 4\rho$ and $\bar{c}_{\mathrm{RP}}$ from equation 9,

$$\bar{c}_{\mathrm{H}} - (1 + 4\rho) = \frac{-p_3(\rho)}{41(2\rho^2 + 3\rho + 16)}, \qquad \bar{c}_{\mathrm{RP}} - \bar{c}_{\mathrm{H}} = \frac{p_4(\rho)}{41(2\rho + 1)(2\rho^2 + 3\rho + 16)},$$

where

$$p_3(x) = 328x^3 - 1886x^2 - 343x + 176, \qquad p_4(x) = 1312x^4 - 2788x^3 + 2184x^2 - 2615x + 176.$$

The denominators are positive for $\rho > 0$, so it suffices to show $p_3 < 0$ and $p_4 > 0$ on $[2, 5]$. For $x \in [2, 5]$, $328x - 1886 \leq -246$, so

$$p_3(x) = x^2(328x - 1886) - 343x + 176 \leq -246 \cdot 4 - 343 \cdot 2 + 176 < 0.$$

For $x \geq 2$, the quadratic $1312x^2 - 2788x + 2184$ is increasing and at least 1856, so

$$\begin{aligned} p_4(x) &= x^2(1312x^2 - 2788x + 2184) - 2615x + 176 \\ &\geq 1856x^2 - 2615x + 176 \geq (3712 - 2615)\, x + 176 > 0. \end{aligned}$$

For $\rho > 0$, $p_3 < 0$ exactly on (r_3, ρ_3) and $p_4 > 0$ exactly on $(0, r_4) \cup (\rho_4, \infty)$, where $r_3 \approx 0.2311$ and $\rho_3 \approx 5.9115$ are the positive roots of p_3, and $r_4 \approx 0.0712$ and $\rho_4 \approx 1.7933$ are the real roots of p_4. Since $r_4 < r_3 < \rho_4$, the full set of $\rho > 0$ satisfying equation 10 is the interval (ρ_4, ρ_3). □

D.3 The momentum family and matched initialization

D.3.1 The momentum family

Proposition D.5 (Strong–weak frontier of the momentum family)**.** *Fix $\rho > 0$, and let q_{s}, $\bar{c}_{\mathrm{H}}$ and $\bar{c}_{\mathrm{RP}}$ be given by equation 7, equation 8 and equation 9. Put $D(\rho) = 16\rho^2 + 4\rho - 1$ and $[x]_+ = \max\{x, 0\}$. Then*

$$\sup_{\alpha,\beta>0} \bar{c}_{\mathrm{H}} = 1 + 4\rho + \tfrac{1}{2}[D(\rho)]_+, \tag{23}$$

$$C^*(q, \rho) := \sup_{\substack{\alpha,\beta>0 \\ q_{\mathrm{s}} \leq q}} \bar{c}_{\mathrm{H}} = 1 + 4\rho + [D(\rho)]_+ \min\{q, \tfrac{1}{2}\}, \qquad q > 0. \tag{24}$$

These are suprema. No finite (α, β) attains them. For every $\rho > 0$, the unconstrained supremum strictly exceeds $\bar{c}_{\mathrm{RP}}$. More precisely, for a given $q > 0$ there is a finite momentum rule with $q_{\mathrm{s}} \leq q$ and $\bar{c}_{\mathrm{H}} > \bar{c}_{\mathrm{RP}}$ if and only if $C^*(q, \rho) > \bar{c}_{\mathrm{RP}}$. This holds for all $q > 0$ when $0 < \rho \leq \frac{1}{2}$. When $\rho > \frac{1}{2}$ it is equivalent to

$$\min\{q, \tfrac{1}{2}\} > q_{\mathrm{crit}}(\rho) := \frac{4\rho(2\rho - 1)}{(1 + 2\rho)D(\rho)}.$$

Whenever $\bar{c}_{\mathrm{H}} > \bar{c}_{\mathrm{RP}}$, that fixed H has $R_H(e_{\mathrm{w}}) < R_{\mathrm{RP}}(e_{\mathrm{w}})$ for all sufficiently small positive ε. Conversely, for every fixed $\alpha, \beta > 0$, RP has $R_{\mathrm{RP}}(e_{\mathrm{s}}) < R_H(e_{\mathrm{s}})$ for all sufficiently small positive ε. The two "sufficiently small" ranges depend on the fixed parameters. No uniform range over the unbounded momentum family is asserted.

Proof. Write $a = \alpha$, $b = \beta$, $r = \rho$ and $z = 1 + 2r$. Straight substitution into equation 8 gives

$$\bar{c}_{\mathrm{H}} = \frac{2b(Kb + L)}{(a+1)(4b+d)(b+e)}, \qquad d = 2a + 1, \quad e = z^2 + az,$$

where

$$K = 8ar + 2a + 16r^2 + 12r + 1,$$
$$L = 4a^2 r + 2a^2 + 8ar^2 + 10ar + 3a + 4r^2 + 4r + 1.$$

All of d, e, K, L are positive. Apart from the positive factor $2/((a+1)(4b+d)^2(b+e)^2)$, the derivative with respect to b is

$$\{K(4e + d) - 4L\}b^2 + 2Kde\, b + Lde.$$

The coefficient of b^2 is

$$64a^2r^2 + 48a^2r + 4a^2 + 256ar^3 + 320ar^2 + 112ar + 4a$$
$$+ 256r^4 + 448r^3 + 272r^2 + 60r + 1 > 0.$$

Thus $\bar{c}_{\mathrm{H}}$ strictly increases with b, and

$$c_\infty(a, r) := \lim_{b\to\infty} \bar{c}_{\mathrm{H}} = 1 + 4r + \frac{D(r)}{2(a+1)}.$$

For $D(r) \geq 0$ the supremum over a is the limit $a \downarrow 0$, $(16r^2 + 12r + 1)/2$. For $D(r) < 0$, increasing a raises c_∞ and the supremum is $1 + 4r$. The turning rate is $r_0 = (\sqrt{5} - 1)/8$.

From equation 7,

$$\frac{\partial q_{\mathrm{s}}}{\partial b} = -\frac{2(2a+1)^2}{(a+1)(2a+4b+1)^2} < 0, \qquad \lim_{b\to\infty} q_{\mathrm{s}} = \frac{1}{2(a+1)}.$$

For a finite feasible pair with $q_{\mathrm{s}} \leq q$, necessarily $1/(2(a+1)) < q$. Conversely every $a > 0$ satisfying this strict inequality becomes feasible for all sufficiently large finite b. Since $\bar{c}_{\mathrm{H}}$ increases with b, the constrained supremum is the supremum of $c_\infty(a, r)$ over this set of a. If $D(r) \leq 0$, take $a \to \infty$. If $D(r) > 0$, take $a \downarrow \max\{0, 1/(2q) - 1\}$ from the feasible side. These limits yield equation 23 and equation 24. Strict increase in b excludes attainment.

For $0 < r < 1/2$, $1 + 4r - \bar{c}_{\mathrm{RP}} = 4r(1 - 2r)/(1 + 2r) > 0$. At $r = 1/2$, $D(r) > 0$, so the frontier exceeds $\bar{c}_{\mathrm{RP}}$. For $r > 1/2$, $D(r) > 0$, and direct subtraction gives

$$1 + 4r + \tfrac{1}{2}D(r) - \bar{c}_{\mathrm{RP}} = \frac{32r^3 + 8r^2 + 10r - 1}{2(1 + 2r)} > 0.$$

The numerator is increasing for $r > 0$ and positive at $r = 1/2$. For $r > 1/2$, subtracting $\bar{c}_{\mathrm{RP}}$ from equation 24 gives the displayed q_{crit} condition. Since no supremum is attained, strict comparison with the supremum is exactly the condition for a finite feasible rule with $\bar{c}_{\mathrm{H}} > \bar{c}_{\mathrm{RP}}$.

Finally, Lemma D.2 makes the strong–weak cross term *exactly zero* for every ε, not merely of order $O(\varepsilon)$. For any fixed finite parameters, Lemma D.3 gives finite second moments and smooth even expansions in a nonzero interval around $\varepsilon = 0$. By Propositions 3.2 and 3.4,

$$R_H(a_0 e_{\mathrm{s}} + b_0 e_{\mathrm{w}}) = a_0^2\{q_{\mathrm{s}} + O(\varepsilon^2)\} + b_0^2\{1 - \bar{c}_{\mathrm{H}}\varepsilon^2 + O(\varepsilon^4)\},$$
$$R_{\mathrm{RP}}(a_0 e_{\mathrm{s}} + b_0 e_{\mathrm{w}}) = a_0^2 O(\varepsilon^2) + b_0^2\{1 - \bar{c}_{\mathrm{RP}}\varepsilon^2 + O(\varepsilon^4)\}.$$

Taking e_{w} proves the weak comparison when $\bar{c}_{\mathrm{H}} > \bar{c}_{\mathrm{RP}}$. Taking e_{s} proves the opposite comparison for each fixed rule because $q_{\mathrm{s}} > 0$. The intervals need not be uniform as a or b diverges. The explicit finite-moment interval provided by Lemma D.3 is $|\varepsilon| < \min\{1, (4\beta\|V_0\|)^{-1}\}$. This is a sufficient stability bound, not a bound on the relative error of a truncated expansion. □

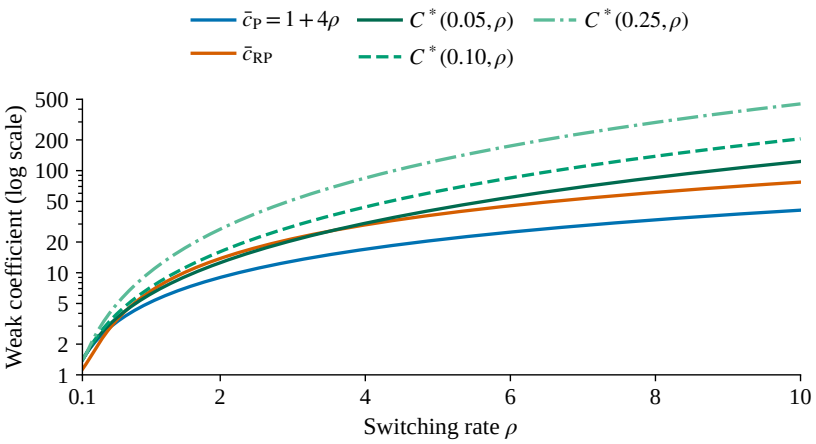

Figure 13: Leading small-ε weak-direction coefficients under the reconstruction query. The blue and orange curves show $\bar{c}_{\mathrm{P}} = 1 + 4\rho$ and $\bar{c}_{\mathrm{RP}}$. The green curves show the momentum frontier $C^*(q, \rho)$ of Proposition D.5 for strong-risk budgets $q \in \{0.05, 0.10, 0.25\}$. The ordinate is logarithmic. Each frontier value is a supremum over the momentum parameters and is not attained by a finite rule or at a common finite ε.

Finite-ε check for the momentum frontier. With $\alpha = 1$, $\rho = 3$, and the exact-expectation solver of Appendix A.5.1, we evaluated $\beta \in \{30, 300, 3000\}$ at seven prespecified values of ε from 0.005 to 0.1. The solver certified finite risk in 20 of the 21 cells. At $(\beta, \varepsilon) = (3000, 0.1)$ it returned an unresolved status, with second-moment spectral abscissa $1.376 > \lambda = 1$, which alone does not prove divergence for the specified initial error. Requiring both the weak coefficient and the strong intercept to have relative error at most 5%, the largest contiguous tested ε from 0.005 is 0.03 for $\beta = 30$, 0.005 for $\beta = 300$, and below the tested range for $\beta = 3000$. The omitted terms grow with $\beta\varepsilon^2$: at $(\beta, \varepsilon) = (3000, 0.05)$ the exact weak risk is 1.108, above its initial value. The predicted trade-off is nonetheless visible at finite ε. For $\beta = 3000$ and $\varepsilon \le 0.02$, H has lower weak-direction risk than RP, for example 0.99876 against 0.99946 at $\varepsilon = 0.005$, while its strong-direction risk is 0.324 against 2.5×10^{-5} for RP. These are empirical thresholds on one grid, not uniform guarantees over the family.

D.3.2 Matched initialization

P and RP project onto the first key at $t = 0$, while H starts from E_0 with $\dot{E}_0 = 0$. To separate this asymmetry from the effect of momentum, the *matched* rule $H^{\mathrm{m}}(\alpha, \beta)$ applies the same initial projection, $E_{0^+} = E_0(I - P_{k_{X_0}})$, and then follows equation 4 with $\dot{E}_{0^+} = 0$. All risks below are exact expectations from the linear second-moment equations of Appendix A.5.1. They use no sampling. The initial projection makes the strong-direction risk $O(\varepsilon^2)$ for every rule, so $g_0 = 0$ at leading order and Proposition 4.2 leaves the comparison open. We therefore evaluate the full risk matrices directly.

The ordering of Theorem 3.6. Table 20 gives the weak-direction coefficient of $H(1, 30)$ with and without the matched projection. We obtain it by Richardson extrapolation from $\varepsilon \in \{0.004, 0.002\}$. The projection lowers the coefficient by 0.2 to 1.1. On $\rho \in [2, 5]$ it still satisfies $1 + 4\rho < c_{\mathrm{H}}^{\mathrm{m}} < \bar{c}_{\mathrm{RP}}$. At $\varepsilon = 0.01$ and $\rho \in \{2, 4, 5\}$, $R_{H^{\mathrm{m}}} - R_{\mathrm{RP}}$ is positive definite and $R_{H^{\mathrm{m}}} - R_{\mathrm{P}}$ is negative definite. The strict ordering $\mathrm{RP} \prec H^{\mathrm{m}} \prec \mathrm{P}$ therefore holds for every initial error, so it does not depend on the strong margin. At $\rho = 1$ the matched coefficient falls below $\bar{c}_{\mathrm{RP}}$, and the unmatched one does not.

The momentum family. We evaluated 240 two-key cells, $\alpha \in \{0.1, 0.3, 1, 3, 10, 30\}$, $\beta \in \{1, 3, 10, 30, 10^2, 3 \cdot 10^2, 10^3, 10^4\}$ and $\rho \in \{0.5, 1, 2, 4, 8\}$, at $\varepsilon = 0.005$. We classify each cell by the eigenvalues of $R_H - R_{\mathrm{RP}}$.

Table 20: Weak-direction coefficients with and without the matched initial projection. The column $\bar{c}_{\mathrm{H}}$ reproduces equation 8 to three decimals.

ρ	$1+4\rho$	$\bar{c}_{\mathrm{H}}(1,30,\rho)$	$c^{\mathrm{m}}_{\mathrm{H}}(1,30,\rho)$	$\bar{c}_{\mathrm{RP}}(\rho)$
1	5	7.003	6.157	6.333
2	9	13.415	12.317	13.800
3	13	18.088	17.132	21.571
4	17	21.220	20.512	29.444
5	21	23.315	22.861	37.364
6	25	24.754	24.525	45.308

- *Unmatched.* RP wins for every initial error in 146 cells. In the other 94 cells the comparison is split, because H wins along e_{w} and loses along e_{s}.
- *Matched.* All 146 cells keep their outcome. The 94 split cells become definite: 90 favor H^{m} and 4 favor RP for every initial error. In all 240 cells the sign of the definite difference equals the sign of the weak-direction difference.

The initial projection therefore does not create the fixed-rule ordering, but it hides that stiff momentum dominates RP once both start from the same projection. Ten cells with $\beta = 10^4$ reverse between $\varepsilon = 0.01$ and 0.005, as expected when $\beta\varepsilon^2$ is not small.

Three keys. For $K = 3$ and $d = 3$, we drew 40 dictionaries of weak components $u_i \sim N(0, I_2)$ at $\varepsilon = 0.005$ in each of six cells. The cells cross a uniform kernel and two reversible weighted kernels, with pair weights $(10, 1, 1)$ and $(0.1, 1, 1)$, with $\rho \in \{1, 4\}$. Against $H(1, 30)$, the classification of $R_H - R_{\mathrm{RP}}$ is unchanged by matching in 222 of 240 dictionaries. The 18 changes all occur under the uniform kernel at $\rho = 1$ and turn a split comparison into a win for H^{m}.

D.4 Constant gains and tuned families at fixed ε

D.4.1 Constant over-relaxed gain

The momentum frontier of Proposition D.5 can also be compared with the simplest momentum-free rule beyond projection: one constant gain $\theta \in (1, 2)$ applied at $t = 0$ and at every key change. Gains $\theta \le 1$ are covered by Theorem 3.1, and RP, which alternates gains 2 and 1 after an initial projection, is not a constant-gain rule.

Proposition D.6 (Constant gain at key changes)**.** *Fix $\rho > 0$ and $1 < \theta < 2$, and apply $A_k(\theta) = I - \theta P_k$ at $t = 0$ and at every key change in the two-key model of Section 3, with $\lambda = 1$. Then*

$$R_\theta(e_{\mathrm{s}}) = s_\theta + O(\varepsilon^2), \qquad R_\theta(e_{\mathrm{w}}) = 1 - c_\theta(\rho)\,\varepsilon^2 + O(\varepsilon^4),$$

with

$$s_\theta = \frac{(1-\theta)^2}{1+\rho\theta(2-\theta)}, \qquad c_\theta(\rho) = \frac{\theta(1+\rho)(2-\theta)\big(4\rho^2\theta + 3\rho\theta + 2\rho + 1\big)}{\big[1+\rho(2-\theta)\big]\big[1+\rho\theta(2-\theta)\big]}.$$

At $\theta = 1$ these reduce to $s_1 = 0$ and $c_1 = 1 + 4\rho$, the values of P*. For each $\rho \in \{1, 2, 3, 4, 5\}$,*

$$c_\theta(\rho) > C^*(s_\theta, \rho) \quad \Longleftrightarrow \quad 1 < \theta < \theta_\star(\rho), \qquad \theta_\star(\rho) = \frac{2(2\rho+1)}{2\rho + \sqrt{2\rho+2}},$$

with $\theta_\star = 1.5000,\ 1.5505,\ 1.5858,\ 1.6126$, and 1.6340 for $\rho = 1, \dots, 5$.

Proof. Put $a = 1 - \theta$ and condition on $N(Q) = n$ key changes before the query, so that there are $m = n+1$ updates and $\Pr(N(Q) = n) = (1+\rho)^{-1}[\rho/(1+\rho)]^n$ (Lemma C.2). At $\varepsilon = 0$ each update multiplies the strong component by a, so the leading strong risk is $\mathbb{E}[a^{2m}] = a^2/(1 + \rho - \rho a^2) = s_\theta$.

From e_{w}, write the error after m updates as $f_m = (\sigma_0\varepsilon X_m,\ 1+\varepsilon^2 B_m) + O(\varepsilon^3)$, where $\sigma_0 \in \{-1, 1\}$ is the sign of the first key. The alternating keys give

$$X_{m+1} = aX_m - \theta(-1)^m, \qquad B_{m+1} = B_m - \theta(-1)^m X_m - \theta, \qquad X_0 = B_0 = 0,$$

with solutions

$$X_m = -\frac{\theta}{1+a}\big(a^m - (-1)^m\big), \qquad B_m = -\theta m + \frac{\theta^2}{1+a}\Big[\frac{1-(-a)^m}{1+a} - m\Big].$$

The pathwise weak coefficient $-2B_m - X_m^2$ is therefore

$$c_m = \frac{4\theta m}{1+a} - \frac{\theta^2}{(1+a)^2}\big[3 + a^{2m} - 4(-a)^m\big].$$

Taking expectations with $\mathbb{E}[m] = 1+\rho$, $\mathbb{E}[a^{2m}] = s_\theta$, and $\mathbb{E}[(-a)^m] = -a/(1+\rho+\rho a)$ gives c_θ. The expansion is even in ε, so the next weak term is $O(\varepsilon^4)$.

For the comparison, $s_\theta < \frac{1}{2}$ exactly for $\theta < \theta_{1/2} = 1 + \sqrt{(\rho+1)/(\rho+2)}$. There equation 24 gives

$$c_\theta - C^*(s_\theta, \rho) = \frac{4\rho(2\rho+1)(\theta-1)\big[\rho(\theta-2)^2 - \theta^2 + 2\big]}{\big[1+\rho(2-\theta)\big]\big[1+\rho\theta(2-\theta)\big]}.$$

The bracket decreases strictly on $(1,2)$ and vanishes at $\theta_\star(\rho)$, which lies below $\theta_{1/2}$. For $\theta \geq \theta_{1/2}$ the frontier is constant in q. Multiplying $C^* - c_\theta$ by the positive factor $2[1+\rho(2-\theta)][1+\rho\theta(2-\theta)]$ and substituting $x = \theta/2$ gives a quartic in $x \in [0,1]$ whose Bernstein coefficients are all positive for $\rho = 1, \ldots, 5$:

ρ	b_0	b_1	b_2	b_3	b_4
1	87	295/2	82	9/2	29
2	445	1216	763	62	89
3	1267	9481/2	9196/3	439/2	181
4	2745	13025	8525	525	305
5	5071	58283/2	19190	2053/2	461

so $c_\theta < C^*(s_\theta, \rho)$ on that region. □

Since $C^*(q,\rho)$ bounds $\bar{c}_{\mathrm{H}}$ over every momentum rule with $q_{\mathrm{s}} \leq q$, the proposition says that for $\theta \in (1, \theta_\star)$ no momentum rule attains both a strong cost at most s_θ and a weak coefficient at least c_θ: a constant over-relaxed gain lies beyond the momentum frontier at leading order. The frontier concerns momentum without the matched initial projection of Appendix D.3.2. With that projection the comparison is not covered here. For each fixed momentum rule with $q_{\mathrm{s}} \leq s_\theta$, it therefore attains lower risk from e_{w} at sufficiently small ε, with the same non-uniformity in ε as in Appendix D.3.1. Figure 14 shows the curves for $\rho = 2, 3, 4$. This is consistent with the dwell-model comparison of Appendix A.3, where the selected constant gain exceeds one under the reconstruction query and attains lower risk than the tuned momentum rule. We checked s_θ and c_θ against the exact solver at $\rho \in \{1, \ldots, 5\}$ and $\theta \in \{1.1, 1.2, \ldots, 1.9\}$ by Richardson extrapolation from $\varepsilon = 0.002$ and 0.001. The largest relative error is 2.0×10^{-7}, and an independent 6×6 moment solve at three (ρ, θ) pairs agrees to 4.7×10^{-8}.

Finite-ε survey against RP. We compared each rule with RP by exact second-moment solves on the grid $\rho \in \{0.5, 0.75, \ldots, 10\}$, $\alpha \in \{0.01, 0.03, \ldots, 30\}$, $\beta \in \{1, 3, \ldots, 10^4\}$, and 61 logarithmic values $\varepsilon \in [10^{-3}, 1]$. Evaluations without a positive spectral margin are not counted as finite. At every ρ, some H attains lower risk than RP from e_{w} up to $\varepsilon = 1$, the end of the searched range. Constant gains $\theta \in (1,2]$ can exceed $\bar{c}_{\mathrm{RP}}$ in the weak coefficient, but $s_\theta > 0$ while RP has zero leading strong cost, so no constant gain dominates RP for every initial error. Exact finite-ε risks at $\rho \in \{2,3,4,5\}$ confirm the strong-axis ordering, and the weak-axis ordering holds at sufficiently small ε, with the non-uniformity noted in Appendix D.3.1.

D.4.2 Tuned families at fixed ε

The results above fix the rule and let $\varepsilon \to 0$. Here we fix $\varepsilon \in \{0.01, 0.02, 0.03, 0.05, 0.07, 0.1, 0.15, 0.2, 0.3, 0.4, 0.6, 0.8, 1\}$ and $\rho \in \{1, 2, 3, 4, 5, 10\}$ and tune each family separately for each query and initial error, using exact second-moment solves. The flow $H(\alpha, \beta)$ is searched on a logarithmic grid over $\alpha \in [10^{-2}, 10^4]$ and $\beta \in [1, 10^7]$ with local refinement and edge extensions. The constant gain uses $\theta \in \{0.01, \ldots, 2\}$, and momentum with one update per key change uses $\eta \in \{0, \ldots, 0.99\}$ and the same θ grid, with the update count distributed as in the flow model. Figure 15 shows $\rho = 3$.

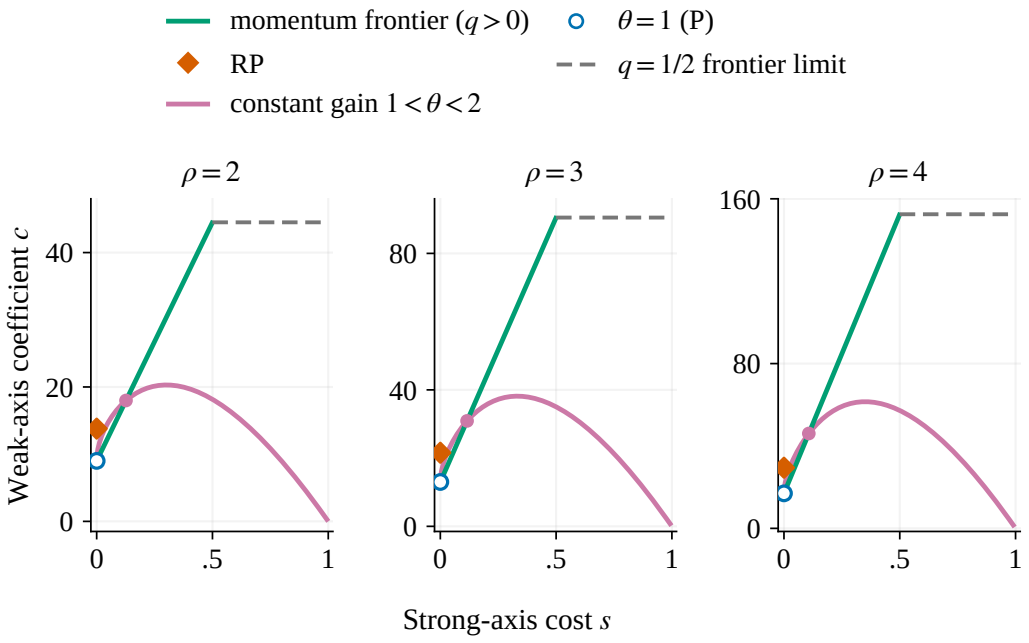

Figure 14: Leading strong-direction cost against weak-direction coefficient for the constant gain $\theta \in (1, 2)$, the momentum-family frontier $C^*(q, \rho)$ of Proposition D.5, and RP, at $\rho = 2, 3, 4$. The green line is the momentum frontier for $q < \frac{1}{2}$ and the grey dashed line its limit at $q = \frac{1}{2}$. The frontier is a supremum, not attained by a finite rule. The purple curve is the constant gain, the orange diamond is RP, the open blue circle is the $\theta = 1$ endpoint (P), outside the proposition's open range, and the purple dot marks $\theta_\star$, where the constant-gain curve meets the frontier.

From $E_0 = -W_\star$, flow momentum beats the best constant gain under reconstruction only below a crossover between adjacent grid values, $[0.07, 0.1]$ at $\rho = 1$, $[0.05, 0.07]$ at $\rho = 2, 3$, $[0.03, 0.05]$ at $\rho = 4, 5$, and $[0.02, 0.03]$ at $\rho = 10$, and never under the associative query. The best constant gain has lower risk than RP in every cell, and momentum with one update per key change beats or ties all other families. Ties occur at large ε, where it selects $\eta = 0$. Many flow selections lie on the extended search boundary, so the flow risks are the best found in the searched range. Restricting $\eta \leq 0.95$ changes no winner.

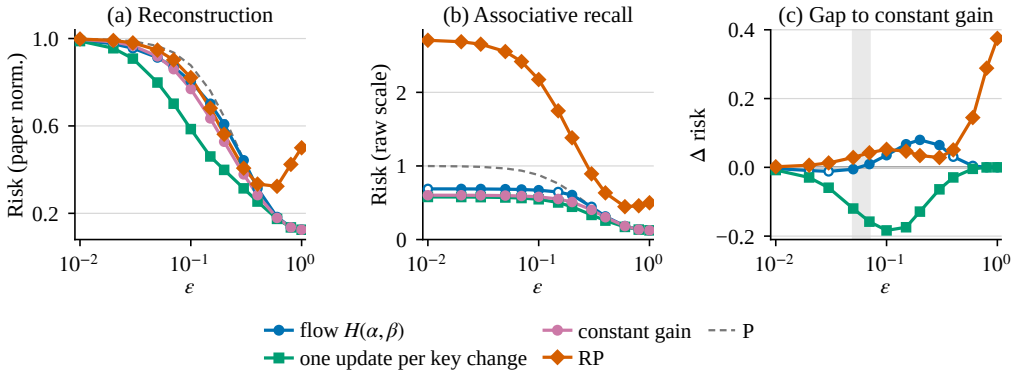

Figure 15: Best risk found for each tuned family against ε at $\rho = 3$ from $E_0 = -W_\star$. Each point is tuned separately at its ε. (a) Normalized reconstruction risk. (b) Associative risk on its raw scale. (c) Reconstruction risk minus that of the best constant gain. The grey band marks the grid bracket $[0.05, 0.07]$ of the flow–constant crossover. Open markers are selections on the search boundary.

E PROOFS FOR THE K-KEY SETTING

Throughout this section $C = I_d$ and $\lambda = 1$. We write an error row as $f = (x, w)$ with $x \in \mathbb{R}$ along e_{s} and $w \in \mathbb{R}^{d-1}$ in the weak subspace, and we let Y_n be the key after the nth key change, with $Y_0 = X_0$.

E.1 ACCESS MODES AND THE RISK MATRIX

E.1.1 ACCESS MODES

Lemma E.1 (Spectral decomposition of the access process)**.** *Under Assumption 4.1 the following hold.*

(i) *T has real eigenvalues $1 = t_0 > t_1 \geq \cdots \geq t_{K-1} \geq -1$ with eigenvectors $\phi_0 \equiv 1, \phi_1, \ldots, \phi_{K-1}$ that are orthonormal under π.*

(ii) *$u_i = \mu + \sum_{\ell \geq 1} \phi_\ell(i)\mu_\ell$ for every i, and $\Sigma = \sum_{\ell \geq 1} \mu_\ell \mu_\ell^\top$.*

(iii) *The sequence (Y_n) is a stationary Markov chain with transition matrix T and law π, and for $m, n \geq 0$,*

$$\mathbb{E}\big[u_{Y_m} u_{Y_n}^\top\big] = \mu\mu^\top + \sum_{\ell \geq 1} t_\ell^{|m-n|}\, \mu_\ell \mu_\ell^\top .$$

(iv) *For $s, t \geq 0$,*

$$\mathbb{E}\big[u_{X_s} u_{X_t}^\top\big] = \mu\mu^\top + \sum_{\ell \geq 1} e^{-\omega_\ell |t-s|}\, \mu_\ell \mu_\ell^\top .$$

Proof. (i) Let $\Pi = \mathrm{diag}(\pi)$. Reversibility $\pi_i T_{ij} = \pi_j T_{ji}$ says that ΠT is symmetric, so $\Pi^{1/2} T \Pi^{-1/2}$ is symmetric. Its eigenvalues are real and it has an orthonormal eigenbasis. Multiplying that basis by $\Pi^{-1/2}$ gives eigenvectors of T that are orthonormal under π (Levin & Peres, 2017). Since T is stochastic, $T\mathbf{1} = \mathbf{1}$ and every eigenvalue lies in $[-1, 1]$, and since T is irreducible, the eigenvalue 1 is simple (Levin & Peres, 2017). Hence $t_\ell < 1$ for $\ell \geq 1$.

(ii) Expanding each coordinate of $i \mapsto u_i$ in the π-orthonormal basis $\{\phi_\ell\}$ gives $u_i = \sum_{\ell \geq 0} \phi_\ell(i) \sum_h \pi_h \phi_\ell(h) u_h$. The coefficient for $\ell = 0$ is $\sum_h \pi_h u_h = \mu$, and for $\ell \geq 1$ it is μ_ℓ. Substituting this expansion into $\Sigma = \sum_i \pi_i (u_i - \mu)(u_i - \mu)^\top$ and using the orthonormality of the ϕ_ℓ gives $\Sigma = \sum_{\ell \geq 1} \mu_\ell \mu_\ell^\top$.

(iii) Since $T_{ii} = 0$, every jump of X changes the key, so (Y_n) is the jump chain of X and has transition matrix T. Since $X_0 \sim \pi$ and $\pi T = \pi$, every Y_n has law π. For $m \leq n$ and functions g, h on the keys, $\mathbb{E}[g(Y_m)h(Y_n)] = \sum_i \pi_i g(i)(T^{n-m}h)(i)$. With $h = \phi_\ell$ this equals $t_\ell^{n-m} \sum_i \pi_i g(i) \phi_\ell(i)$. Inserting the expansion (ii) for both factors and using orthonormality gives the claim.

(iv) The transition matrix of X over a time $\tau \geq 0$ is $e^{\rho(T-I)\tau}$, which maps ϕ_ℓ to $e^{-\rho(1-t_\ell)\tau}\phi_\ell$. The argument of (iii) with T^{n-m} replaced by this matrix gives the claim. □

E.1.2 FORM OF THE RISK MATRIX

Let $a \in \{H, \mathrm{P}, \mathrm{RP}\}$, where P and RP use the learning rate $1 + y$ at the first update. Let $s_a = R_a(e_{\mathrm{s}})$ at $\varepsilon = 0$.

Lemma E.2 (Structure of the expansion)**.** *There are $\varepsilon_1 > 0$ and a function $c_a : [-1, 1] \to \mathbb{R}$ that does not depend on the weak components u_i, such that for $|\varepsilon| < \varepsilon_1$*

$$\mathcal{R}_a(\varepsilon) = \begin{pmatrix} s_a + O(\varepsilon^2) & \varepsilon(s_a - 1)\mu^\top + O(\varepsilon^3) \\ \varepsilon(s_a - 1)\mu + O(\varepsilon^3) & I - \varepsilon^2\big[c_a(1)\,\mu\mu^\top + \sum_{\ell \geq 1} c_a(t_\ell)\,\mu_\ell \mu_\ell^\top\big] + O(\varepsilon^4) \end{pmatrix},$$

and $c_a(1) = 1 - s_a$.

Proof. Step 1: smoothness. For H, the proof of Lemma D.3 applies with β replaced by $\beta \max_i \|u_i\|$, since $\|P_{k_i} - e_{\mathrm{s}} e_{\mathrm{s}}^\top\| \leq \varepsilon \|u_i\|$. For P and RP, $\Phi_a(Q)$ is a product of $N(Q) + 1$ matrices that are

smooth in ε, have spectral norm at most one (the first factor has eigenvalues 1 and $-y$ with $|y| \le 1$), and have derivatives bounded uniformly in $|\varepsilon| \le 1$. The mth derivative of the product is therefore bounded by a constant times $(N(Q)+1)^m$, which is integrable by Lemma C.2. In both cases $\mathcal{R}_a$ is smooth near $\varepsilon = 0$ and can be differentiated under the expectation.

Step 2: parity. Let $Z = \mathrm{diag}(1, -I_{d-1})$. Then $k_i(-\varepsilon) = Z k_i(\varepsilon)$, so every projector along the path is conjugated by Z, and, as in the proof of Lemma D.2, $\Phi_a(-\varepsilon) = Z\Phi_a(\varepsilon)Z$ path by path. Hence $\mathcal{R}_a(-\varepsilon) = Z\mathcal{R}_a(\varepsilon)Z$: the diagonal blocks are even and the off-diagonal blocks are odd in ε. At $\varepsilon = 0$ every key equals e_{s}, the weak subspace is never corrected, and $\mathcal{R}_a(0) = \mathrm{diag}(s_a, I)$. This gives the orders of all remainders.

Step 3: the weak block is a quadratic form in the weak components. Write $P_{k_i} = e_{\mathrm{s}}e_{\mathrm{s}}^\top + \varepsilon P_i^{(1)} + \varepsilon^2 P_i^{(2)} + O(\varepsilon^3)$ with

$$P_i^{(1)} = \begin{pmatrix} 0 & u_i^\top \\ u_i & 0 \end{pmatrix}, \qquad P_i^{(2)} = \begin{pmatrix} -\|u_i\|^2 & 0 \\ 0 & u_i u_i^\top \end{pmatrix}.$$

Start from $f_0 = (0, w_0)$. The second derivative $\frac{1}{2}\partial_\varepsilon^2 \|f(Q)\|_2^2$ at $\varepsilon = 0$ collects terms with either two factors $P^{(1)}$ or one factor $P^{(2)}$, and its weak–weak part is therefore a quadratic form in the scalars $u_{Y_n}^\top w_0$ (for P and RP) or $u_{X_s}^\top w_0$ (for H). Its coefficients depend on the rule, on y, and on $N(Q)$ or the times of the path, but not on the weak components. Since these are independent of the visited keys by Lemma C.2, Lemma E.1(iii)–(iv) evaluates the expectation as $-w_0^\top \big[c_a(1)\mu\mu^\top + \sum_{\ell \ge 1} c_a(t_\ell)\mu_\ell\mu_\ell^\top\big] w_0$, where $c_a(t)$ is obtained by replacing every correlation $t_\ell^{|m-n|}$ or $e^{-\rho(1-t_\ell)|t-s|}$ by $t^{|m-n|}$ or $e^{-\rho(1-t)|t-s|}$.

Step 4: identical keys. Suppose $u_i = \mu$ for every i, so that $\mu_\ell = 0$ and every key equals $k = (1, \varepsilon\mu)^\top / \sqrt{1+\varepsilon^2\|\mu\|^2}$. Let O be an orthogonal matrix with $Oe_{\mathrm{s}} = k$. Then $P_k = Oe_{\mathrm{s}}e_{\mathrm{s}}^\top O^\top$, so $\Phi_a = O\,\Phi_a|_{\varepsilon=0}\,O^\top$ path by path and

$$\mathcal{R}_a = O\,\mathrm{diag}(s_a, I)O^\top = I - (1-s_a)\,kk^\top.$$

Since $kk^\top$ has off-diagonal block $\varepsilon\mu/(1+\varepsilon^2\|\mu\|^2)$ and weak block $\varepsilon^2\mu\mu^\top/(1+\varepsilon^2\|\mu\|^2)$, the off-diagonal block of $\mathcal{R}_a$ is $\varepsilon(s_a-1)\mu + O(\varepsilon^3)$ and the weak block is $I - \varepsilon^2(1-s_a)\mu\mu^\top + O(\varepsilon^4)$. Comparing with Step 3 gives $c_a(1) = 1 - s_a$.

Step 5: the off-diagonal block. The first derivative of the off-diagonal block at $\varepsilon = 0$ contains exactly one factor $P^{(1)}$, so it is linear in the weak components, with coefficients that do not depend on them. By Lemma E.1(iii)–(iv), each u_{Y_n} and each u_{X_s} has mean μ, and it is independent of those coefficients. Hence this derivative takes the same value as for identical keys, and Step 4 gives $\varepsilon(s_a-1)\mu + O(\varepsilon^3)$. □

Lemma E.2 reduces Lemma 4.1 to computing s_a and the function c_a for each rule, which we do next.

E.1.3 Modal coefficients

The modal coefficients of Lemma 4.1 are as follows. For momentum,

$$c_{\mathrm{H}}(\omega_\ell) = \frac{2\beta}{\alpha+1} - \frac{4\beta^2}{(\alpha+1)(2\alpha+4\beta+1)} \cdot \frac{\omega_\ell + 2\alpha + 2\beta + 2}{(1+\omega_\ell)^2 + \alpha(1+\omega_\ell) + \beta}. \tag{25}$$

For the projection and reflect–project families,

$$c_{\mathrm{P}}(t_\ell, y) = 1 + 2\rho(1-t_\ell) + \frac{2\rho(1-t_\ell)}{1+\rho}\,y - \frac{y^2}{1+\rho}, \tag{26}$$

$$c_{\mathrm{RP}}(t_\ell, y) = 1 + \frac{2\rho^2(1-t_\ell)(3-t_\ell)}{1+2\rho} + \frac{2\rho^2(1-t_\ell)^2}{(1+\rho)^2}\,y - \frac{(1+2\rho)y^2}{(1+\rho)^2}. \tag{27}$$

E.2 Proofs of Lemma 4.1 and Proposition 4.2

E.2.1 Proof of Lemma 4.1 for P and RP

By Lemma E.2, it suffices to compute s_j and $c_j(t, y) = c_j(t)$ for $j \in \{\mathrm{P}, \mathrm{RP}\}$. Label the updates by $n = 0, \dots, N(Q)$: update 0 happens at time 0 with key Y_0 and learning rate $\theta_0 = 1 + y$, and update $n \geq 1$ happens at the nth key change with key Y_n and learning rate θ_n. For P, $\theta_n = 1$ for $n \geq 1$. For RP, $\theta_n = 2$ for odd n and $\theta_n = 1$ for even $n \geq 2$. The error is constant between updates, so $f(Q)$ is the error after update $N(Q)$. Recall $p_{\mathrm{q}} = \Pr(N(Q) = 0) = 1/(1 + \rho)$ and $p_{\mathrm{c}} = \rho/(1 + \rho)$.

Strong coefficients. At $\varepsilon = 0$ update n multiplies x by $1 - \theta_n$ and leaves w unchanged, so $s_j = \mathbb{E} \prod_{n=0}^{N(Q)} (1 - \theta_n)^2$. For P, the factor at $n = 1$ is zero, so the product is $-y$ if $N(Q) = 0$ and 0 otherwise, and $s_{\mathrm{P}} = y^2 p_{\mathrm{q}} = y^2/(1 + \rho)$. For RP, the factors are $-y$ at $n = 0$, -1 at odd n, and 0 at even $n \geq 2$, so the product is $-y$, y, or 0 for $N(Q) = 0$, $N(Q) = 1$, or $N(Q) \geq 2$. Hence $s_{\mathrm{RP}} = y^2(p_{\mathrm{q}} + p_{\mathrm{q}} p_{\mathrm{c}}) = (1 + 2\rho)y^2/(1 + \rho)^2$.

Recursions along the weak direction. For a key $k = (1, \varepsilon u)^\top / \sqrt{1 + \varepsilon^2 \|u\|^2}$, an update with learning rate θ maps $f = (x, w)$ to

$$f - \theta k k^\top f = f - \theta \, \frac{x + \varepsilon u^\top w}{1 + \varepsilon^2 \|u\|^2} \begin{pmatrix} 1 \\ \varepsilon u \end{pmatrix}.$$

Start from $f_0 = (0, w_0)$. By Step 2 of the proof of Lemma E.2, x is odd and w is even in ε, so after update n we can write $x = \varepsilon \xi_n + O(\varepsilon^3)$ and $w_0^\top w = \|w_0\|^2 + \varepsilon^2 \zeta_n + O(\varepsilon^4)$. Let $\chi_n = u_{Y_n}^\top w_0$. Substituting into the update and collecting the coefficients of ε in x and of ε^2 in $w_0^\top w$ gives

$$\xi_n = (1 - \theta_n)\, \xi_{n-1} - \theta_n \chi_n, \qquad \zeta_n = \zeta_{n-1} - \theta_n \chi_n \big(\xi_{n-1} + \chi_n\big), \tag{28}$$

with $\xi_{-1} = \zeta_{-1} = 0$. Since $\|f\|_2^2 = x^2 + \|w\|^2$, the second-order coefficient of the risk started from f_0 is $\mathbb{E}[\xi_{N(Q)}^2 + 2\zeta_{N(Q)}]$. By Step 3 of the proof of Lemma E.2,

$$c_j(t) = -\mathbb{E}^{(t)} \big[\xi_{N(Q)}^2 + 2\zeta_{N(Q)}\big],$$

where $\mathbb{E}^{(t)}$ denotes the expectation in which every $\mathbb{E}[\chi_m \chi_n]$ is replaced by $t^{|m-n|}$. Let $F(N) = \mathbb{E}^{(t)}[\xi_N^2 + 2\zeta_N \mid N(Q) = N]$.

Projection. By equation 28, $\xi_0 = -(1+y)\chi_0$, $\zeta_0 = -(1+y)\chi_0^2$, and $\xi_n = -\chi_n$ for $n \geq 1$. Hence $F(0) = (1 + y)^2 - 2(1 + y) = y^2 - 1$, and for $N \geq 1$

$$\xi_N^2 + 2\zeta_N = \chi_N^2 - 2(1 + y)\chi_0^2 - 2\chi_1\big(\chi_1 - (1 + y)\chi_0\big) - 2\sum_{n=2}^{N} \chi_n(\chi_n - \chi_{n-1}),$$

so

$$F(N) = 1 - 2(1 + y) - 2\big(1 - (1 + y)t\big) - 2(N - 1)(1 - t) = -1 - 2(1 - t)(N + y).$$

By Lemma C.2, $\Pr(N(Q) \geq 1) = p_{\mathrm{c}}$ and $\mathbb{E}N(Q) = \rho$. Hence

$$-c_{\mathrm{P}}(t) = p_{\mathrm{q}}(y^2 - 1) - p_{\mathrm{c}} - 2(1 - t)\big(\rho + y p_{\mathrm{c}}\big),$$

and inserting p_{q} and p_{c} gives equation 26.

Reflect–project. By equation 28, $\xi_0 = -(1+y)\chi_0$, $\xi_n = -\xi_{n-1} - 2\chi_n$ for odd n, and $\xi_n = -\chi_n$ for even $n \geq 2$. Hence $\xi_1 = (1+y)\chi_0 - 2\chi_1$ and $\xi_n = \chi_{n-1} - 2\chi_n$ for odd $n \geq 3$. The increments $\kappa_n = \zeta_n - \zeta_{n-1}$ and their values under $\mathbb{E}^{(t)}$ are

$$\begin{aligned}
&\kappa_0 = -(1 + y)\chi_0^2, && \mathbb{E}^{(t)}\kappa_0 = -(1 + y),\\
&\kappa_1 = -2\chi_1\big(\chi_1 - (1 + y)\chi_0\big), && \mathbb{E}^{(t)}\kappa_1 = -2 + 2(1 + y)t,\\
&\kappa_2 = -\chi_2\big(\chi_2 - 2\chi_1 + (1 + y)\chi_0\big), && \mathbb{E}^{(t)}\kappa_2 = -1 + 2t - (1 + y)t^2,\\
&\kappa_n = -2\chi_n(\chi_n - \chi_{n-1}),\ n \geq 3 \text{ odd}, && \mathbb{E}^{(t)}\kappa_n = -2(1 - t),\\
&\kappa_n = -\chi_n(\chi_n - 2\chi_{n-1} + \chi_{n-2}),\ n \geq 4 \text{ even}, && \mathbb{E}^{(t)}\kappa_n = -(1 - t)^2,
\end{aligned}$$

and $\mathbb{E}^{(t)}\xi_N^2$ equals $(1+y)^2$ for $N=0$, $(1+y)^2-4(1+y)t+4$ for $N=1$, 1 for even $N\geq 2$, and $5-4t$ for odd $N\geq 3$. Summing,

$$F(0)=F(1)=y^2-1, \qquad F(2)=F(3)=-1-2y(1-t)^2-2(1-t)(3-t).$$

For $N\geq 4$, passing from $N-2$ to N adds one odd and one even increment and leaves $\mathbb{E}^{(t)}\xi_N^2$ unchanged, so $F(N)=F(N-2)-2\big(2(1-t)+(1-t)^2\big)=F(N-2)-2(1-t)(3-t)$. With $J=\lfloor N(Q)/2\rfloor$,

$$F(N)=\begin{cases} y^2-1, & J=0,\\ -1-2y(1-t)^2-2J(1-t)(3-t), & J\geq 1.\end{cases}$$

By Lemma C.2, $\Pr(J=m)=p_{\mathrm{q}}\big(p_{\mathrm{c}}^{2m}+p_{\mathrm{c}}^{2m+1}\big)=(1-\nu)\nu^m$ with $\nu=p_{\mathrm{c}}^2=\rho^2/(1+\rho)^2$, so $\Pr(J\geq 1)=\nu$, $1-\nu=(1+2\rho)/(1+\rho)^2$, and $\mathbb{E}J=\nu/(1-\nu)=\rho^2/(1+2\rho)$. Hence

$$-c_{\mathrm{RP}}(t)=-1+(1-\nu)y^2-2\nu y(1-t)^2-\frac{2\rho^2(1-t)(3-t)}{1+2\rho},$$

which is equation 27.

As a check, $t=1$ gives $c_j(1,y)=1-s_j$, as required by Lemma E.2, and $t=-1$, $y=0$ give $c_{\mathrm{P}}(-1,0)=1+4\rho$ and $c_{\mathrm{RP}}(-1,0)=\bar{c}_{\mathrm{RP}}$, the coefficients of Section 3.

E.2.2 Proof of Lemma 4.1 for $H(\alpha,\beta)$

At $\varepsilon=0$ every key equals e_{s}, so Steps 2 and 3 of the proof of Proposition 3.2 apply without change and give $s_H=q_{\mathrm{s}}$.

For the weak block, start from $f_0=(0,w_0)$ and let $\chi_t=u_{X_t}^{\top}w_0$. By Step 2 of the proof of Lemma E.2, we can write $x_t=\varepsilon a(t)+O(\varepsilon^3)$ and $w_0^{\top}w_t=\|w_0\|^2+\varepsilon^2 b(t)+O(\varepsilon^4)$. Since $P_kf=\frac{x+\varepsilon u^{\top}w}{1+\varepsilon^2\|u\|^2}(1,\varepsilon u)^{\top}$, collecting the coefficient of ε in the first coordinate and of ε^2 in $w_0^{\top}w$ gives

$$a''+\alpha a'+\beta a=-\beta\chi_t, \qquad b''+\alpha b'=-\beta\chi_t\big(a+\chi_t\big),$$

with zero initial values. These are the equations of Step 4 of the proof of Proposition 3.2 with ξ_t replaced by χ_t. Steps 5–7 there use the key process only through $\mathbb{E}[\xi_t^2]=1$ and the correlation $\mathbb{E}[\xi_s\xi_t]=e^{-2\rho|t-s|}$. By Step 3 of the proof of Lemma E.2, the coefficient c_{H} at an eigenvalue t of T is obtained by giving χ_s the correlation $e^{-\omega|s-r|}$ with $\omega=\rho(1-t)$, which also has value 1 at $s=r$. Repeating Steps 5–8 with 2ρ replaced by ω, that is, with $z=1+\omega$, gives

$$c_{\mathrm{H}}(\omega)=\frac{2\beta}{\alpha+1}-\frac{4\beta^2}{(\alpha+1)(2\alpha+4\beta+1)}\cdot\frac{\omega+2\alpha+2\beta+2}{(1+\omega)^2+\alpha(1+\omega)+\beta},$$

which is equation 25 at $\omega=\omega_\ell$. As a check, $t=1$ gives $\omega=0$ and $c_{\mathrm{H}}(0)=1-q_{\mathrm{s}}$, as required by Lemma E.2, and $t=-1$ gives $\omega=2\rho$ and $c_{\mathrm{H}}(2\rho)=\bar{c}_{\mathrm{H}}$. Together with Appendix E.2.1, this proves Lemma 4.1.

E.2.3 Proof of Proposition 4.2

Proof. Step 1: the blocks of D. Subtracting the expansions of Lemma 4.1 block by block, the strong entry of $D(\varepsilon)$ is $q_{\mathrm{s}}-s_j+O(\varepsilon^2)=g_0+O(\varepsilon^2)$. The off-diagonal block is $\varepsilon\big[(q_{\mathrm{s}}-1)-(s_j-1)\big]\mu+O(\varepsilon^3)=\varepsilon g_0\mu+O(\varepsilon^3)$. The weak block is

$$\varepsilon^2\Big[(1-s_j)-(1-q_{\mathrm{s}})\Big]\mu\mu^{\top}+\varepsilon^2\sum_{\ell\geq 1}\big(c_j(t_\ell,y)-c_{\mathrm{H}}(\omega_\ell)\big)\mu_\ell\mu_\ell^{\top}+O(\varepsilon^4)=\varepsilon^2\big(g_0\mu\mu^{\top}+M\big)+O(\varepsilon^4).$$

This is equation 13.

Step 2: the Schur complement. Write D_{ss}, $D_{\mathrm{sw}}=D_{\mathrm{ws}}^{\top}$, and D_{ww} for the blocks of $D(\varepsilon)$. For $g_0>0$ and small ε, $D_{\mathrm{ss}}>0$, and for every $f_0=(s,w)$,

$$f_0^{\top}D(\varepsilon)f_0=D_{\mathrm{ss}}\big(s+D_{\mathrm{ss}}^{-1}D_{\mathrm{sw}}w\big)^2+w^{\top}\mathcal{S}(\varepsilon)w, \qquad \mathcal{S}(\varepsilon)=D_{\mathrm{ww}}-D_{\mathrm{ws}}D_{\mathrm{ss}}^{-1}D_{\mathrm{sw}}, \qquad (29)$$

which follows by expanding the square. By Step 1, $D_{\mathrm{ws}}D_{\mathrm{ss}}^{-1}D_{\mathrm{sw}} = \varepsilon^2 g_0^2 \mu\mu^\top/(g_0 + O(\varepsilon^2)) + O(\varepsilon^4) = \varepsilon^2 g_0 \mu\mu^\top + O(\varepsilon^4)$, so $\mathcal{S}(\varepsilon) = \varepsilon^2 M + O(\varepsilon^4)$.

Step 3: the case $\lambda_{\min}(M) > 0$. Then $\varepsilon^{-2}\mathcal{S}(\varepsilon) = M + O(\varepsilon^2)$ has smallest eigenvalue at least $\frac{1}{2}\lambda_{\min}(M) > 0$ for small ε. By equation 29, $f_0^\top D(\varepsilon) f_0 > 0$ whenever $w \neq 0$, and $f_0^\top D(\varepsilon) f_0 = D_{\mathrm{ss}} s^2 > 0$ whenever $w = 0$ and $s \neq 0$. Hence $D(\varepsilon)$ is positive definite.

Step 4: the case $w^\top M w < 0$. Choose $s = -D_{\mathrm{ss}}^{-1} D_{\mathrm{sw}} w = -\varepsilon \mu^\top w + O(\varepsilon^3)$. Then the square in equation 29 vanishes and $f_0^\top D(\varepsilon) f_0 = \varepsilon^2 w^\top M w + O(\varepsilon^4) < 0$ for small ε. □

E.3 Positive and indefinite compatibility matrices

This subsection shows that both cases of Proposition 4.2 occur for $K > 2$, and that M can be indefinite while its trace and its diagonal are positive.

Lemma E.3 (Minimal number of keys). *Suppose $K = d$ and $\lambda_{\min}(\Sigma) > 0$. Then $\lambda_{\min}(M) > 0$ if and only if $g_\ell > 0$ for every ℓ. If $g_\ell < 0$ for some ℓ, then some $w \neq 0$ satisfies $w^\top M w = g_\ell < 0$.*

Proof. Let $\mathcal{U} = [\mu_1, \ldots, \mu_{K-1}]$, a square matrix of size $d - 1$. By Lemma E.1(ii), $\Sigma = \mathcal{U}\mathcal{U}^\top$, so $\lambda_{\min}(\Sigma) > 0$ makes $\mathcal{U}$ invertible. Then $M = \mathcal{U}\,\mathrm{diag}(g_1, \ldots, g_{K-1})\mathcal{U}^\top$ is congruent to $\mathrm{diag}(g_1, \ldots, g_{K-1})$, and Sylvester's law of inertia gives the first claim. For the second, take $w = \mathcal{U}^{-\top} e_\ell$, so that $\mathcal{U}^\top w = e_\ell$ and $w^\top M w = g_\ell$. □

For $K > d$ the vectors μ_ℓ are linearly dependent, a negative g_ℓ can be compensated by the other terms, and the full matrix M must be checked.

An indefinite M with positive trace and positive diagonal. Let $K = d = 3$ and

$$T = \begin{pmatrix} 0 & \frac{1}{10} & \frac{9}{10} \\ \frac{1}{10} & 0 & \frac{9}{10} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}, \qquad \pi = \tfrac{1}{19}(5, 5, 9).$$

One checks $\pi_i T_{ij} = \pi_j T_{ji}$, so Assumption 4.1 holds. The nonconstant eigenvalues are $t_1 = -\frac{1}{10}$ and $t_2 = -\frac{9}{10}$, with π-orthonormal eigenvectors

$$\phi_1 = \sqrt{\tfrac{19}{10}}\,(1, -1, 0), \qquad \phi_2 = \left(\tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, -\tfrac{\sqrt{10}}{3}\right).$$

Let $O = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ and $u_i = O\,(\phi_1(i), \phi_2(i))^\top$. By orthonormality, $\mu = 0$, $\mu_1 = Oe_1 = (1, 1)^\top/\sqrt{2}$, $\mu_2 = Oe_2 = (-1, 1)^\top/\sqrt{2}$, and $\Sigma = I$. Take $H(1, 30)$, $\rho = 4$, and RP with $y = 0$. Then $g_0 = q_{\mathrm{s}} = 11/41 > 0$, and equation 25 and equation 27 give

$$g_1 = -\frac{3402113}{2481525} < 0, \qquad g_2 = \frac{9598429}{1442175} > 0.$$

Hence

$$M = g_1 \mu_1 \mu_1^\top + g_2 \mu_2 \mu_2^\top = \frac{1}{2}\begin{pmatrix} g_1 + g_2 & g_1 - g_2 \\ g_1 - g_2 & g_1 + g_2 \end{pmatrix}.$$

Both diagonal entries equal $(g_1 + g_2)/2 > 0$, so the trace and the diagonal are positive, while $\mu_1^\top M \mu_1 = g_1 < 0$. By Proposition 4.2(ii), $H(1, 30)$ attains lower retrieval risk than RP from the initial error $(0, \mu_1)$ for all sufficiently small $\varepsilon > 0$, since $\mu = 0$.

A positive M for every number of keys. Take $H(1, 30)$, $\rho = 10$, and RP with $y = 0$, so that $g_0 = q_{\mathrm{s}} = 11/41 > 0$. By equation 25 and equation 27, the modal margin at an eigenvalue t is

$$g(t) = c_{\mathrm{RP}}(t, 0) - c_{\mathrm{H}}\big(10(1 - t)\big) = \frac{410000t^4 - 2583000t^3 + 4417750t^2 - 2677365t + 436311}{861\,(50t^2 - 115t + 81)}.$$

For $t \in [-1, 0]$, put $x = -t \in [0, 1]$. The numerator becomes $410000x^4 + 2583000x^3 + 4417750x^2 + 2677365x + 436311 \geq 436311$, and the denominator becomes $861(50x^2 + 115x + 81) \leq 861 \cdot 246$. Hence $g(t) \geq g_* = 436311/211806 > 0$ on $[-1, 0]$. Consequently, for every K, every T

satisfying Assumption 4.1 whose nonconstant eigenvalues lie in $[-1, 0]$, and every choice of keys with $\lambda_{\min}(\Sigma) > 0$,

$$M - g_*\Sigma = \sum_{\ell \geq 1} \big(g(t_\ell) - g_*\big)\, \mu_\ell \mu_\ell^\top$$

is positive semidefinite, so $\lambda_{\min}(M) \geq g_* \lambda_{\min}(\Sigma) > 0$. By Proposition 4.2(i), RP then attains lower retrieval risk than $H(1, 30)$ for every nonzero initial error and all sufficiently small $\varepsilon > 0$. Uniform access, $T_{ij} = 1/(K-1)$ for $i \neq j$, is an example: its nonconstant eigenvalues all equal $-1/(K-1)$.

F EXTENDED RELATED WORK

The main text cites prior work at the point where it enters the argument, so that each comparison sits next to the result it qualifies. This appendix collects those citations into one account. Each part below traces a line of work, states the question it answers, and ends with what this paper adds. Table 21 summarizes these differences.

Table 21: Lines of related work and the increment of this paper in each.

Thread	Typical question in prior work	What this paper adds
Fast-weight and delta-rule memories	Which architecture or update performs well when trained on a task?	A controlled comparison of adding momentum state and widening the gain range, at fixed target and prescribed rules (Section 3).
Momentum in stochastic optimization	Does momentum improve convergence rate or sample complexity on one objective?	An unattained weak-direction lower bound for all adaptive delta rules with gain below one, beaten by a fixed momentum rule, which is beaten in turn by reflect–project. Neither side dominates the momentum family (Theorems 3.1 and 3.6 and proposition D.5).
Projections and reflections	Do the iterates or their averages converge?	Reflection inside an online schedule, evaluated at a query that can fall between a reflection and its projection (Proposition 3.4).
Markovian and reshuffled access	How does access order change the convergence of one method?	Access modes enter the compatibility matrix M and can reverse which rule wins (Proposition 4.2).
Jump linear systems and risk geometry	What is the quadratic cost, or how does risk depend on the covariance?	Retrieval risk at an exponential query time along query directions C. The query, key geometry, access order, and elapsed time each reorder the rules (Section 5).

From fast weights to delta-rule sequence models. A matrix memory that stores associations as a sum of key–value outer products goes back to correlation matrix memories (Kohonen, 1972) and to Hebbian learning (Hebb, 1949). Fast-weight memories change such a matrix during processing, so that it holds the recent context (Schmidhuber, 1992; Ba et al., 2016). Later work uses the same idea for fast parametric adaptation and for associative inference (Rae et al., 2018; Schlag et al., 2021b). Linear attention is a recurrent model whose state is such a matrix (Katharopoulos et al., 2020), and Schlag et al. (2021a) show that it is a fast-weight programmer. They replace its additive Hebbian write by the delta rule, and Irie et al. (2021) add recurrence to the programmer. The delta rule itself is the least-mean-squares update of adaptive filtering (Widrow & Hoff, 1960; Sayed, 2008). DeltaNet parallelizes it over the sequence length (Yang et al., 2024), and Gated DeltaNet combines it with a decay gate (Yang et al., 2025). Test-time training treats the recurrent state as a model trained online by a learning rule (Sun et al., 2025), and test-time regression classifies associative memories by their regression objective, their parameterization, and their optimizer (Wang et al., 2026). This view makes the memory update an online optimizer, and several recent models import optimizer features. Titans adds momentum and forgetting to the delta rule (Behrouz et al., 2025b), ATLAS optimizes the memory over a window of past tokens (Behrouz et al., 2025a), and MDN parallelizes a stepwise momentum delta rule (Huang et al., 2026). A second group changes the update matrix instead of adding state. Negative eigenvalues (Grazzi et al., 2025) and products of Householder reflections (Siems et al., 2025) widen the range of the delta update to increase expressivity for state tracking. We take gains above one from this line of work. Our question is how such gains change the retrieval risk at a random query time for a fixed target, and in which regimes this competes with momentum. Preconditioning adds curvature information to the delta rule (Tumma et al., 2026), and further variants separate the erase direction from the write direction (Gupta et al., 2026; Li et al., 2026). Related work also connects test-time training with key–value binding to linear attention (Liu et al., 2026) and

applies fast-weight attention to continual learning (Zhang et al., 2026). Retrieval of recently stored associations is measured by memory-augmented meta-learning and by associative-recall benchmarks (Santoro et al., 2016; Arora et al., 2024).

These works design and train architectures, and they evaluate learned representations on tasks. *Our increment* is a controlled comparison of the two design directions above, adding momentum state versus widening the update itself. We fix a realizable target, prescribe the update rules, and compare their expected retrieval error at a random query time. In this setting a momentum-free rule with gains in $[0, 2]$, the reflect–project rule, attains lower risk than a momentum rule that already beats every delta rule with gains in $[0, 1)$ (Theorems 3.1 and 3.6).

Does momentum help? Heavy-ball momentum accelerates gradient descent on deterministic strongly convex problems (Polyak, 1964; Qian, 1999). Its convergence has been analyzed by global arguments, integral quadratic constraints, and Lyapunov functions (Ghadimi et al., 2015; Lessard et al., 2016; Wilson et al., 2021), and it is a standard ingredient of deep network training (Sutskever et al., 2013). With stochastic gradients the advantage is less clear. For small constant step sizes, SGD with momentum behaves like SGD with a rescaled step size (Yuan et al., 2016; Paquette & Paquette, 2021; Wang et al., 2024). For least squares and quadratics, standard momentum schemes do not improve on SGD in the worst case, and the sample complexity of SGD and SGD with momentum coincide once bias and variance are both counted (Kidambi et al., 2018; Ganesh et al., 2023). Acceleration becomes possible with modified schemes, averaging, or favorable noise structure (Flammarion & Bach, 2015; Jain et al., 2018; Dieuleveut et al., 2017; Pan et al., 2024). Further analyses give convergence rates for stochastic heavy-ball in continuous and discrete time (Gadat et al., 2018; Loizou & Richtárik, 2020; Sebbouh et al., 2021), study the memory that momentum keeps (Gitman et al., 2019; Orvieto et al., 2020), cover the overparameterized and minibatch regimes (Gess & Kassing, 2026; Bollapragada et al., 2025), and treat time-varying objectives (Bernardi et al., 2026). In adaptive filtering, momentum LMS has been analyzed for its transient and steady-state behavior (Roy & Shynk, 1990).

These results compare convergence rates or sample complexity for a single objective. *Our increment* is to pose the question for online associative memory, where each arriving pair defines its own objective and the memory is read at an unknown time. We give an exact lower bound for the whole family of adaptive continuous-time delta rules, which no rule attains (Theorem 3.1), show that a fixed momentum rule goes below it, and show that the reflect–project rule goes below that momentum rule (Theorem 3.6). The comparison is between fixed rules and does not rank the whole momentum family (Section 6).

Projections, relaxation, and reflection. A delta update with unit gain is a Kaczmarz projection. Projection methods for convex feasibility and linear systems are surveyed by Bauschke & Borwein (1996), and randomized Kaczmarz converges linearly in expectation (Strohmer & Vershynin, 2009; Gower & Richtárik, 2015). Its rate improves with weighted sampling (Needell et al., 2016; Steinerberger, 2021c), with scheduled learning rates (Marshall & Mickelin, 2023), and with several constraints per step (Needell & Ward, 2013; Necoara, 2019). Momentum and acceleration have been added to Kaczmarz as well (Liu & Wright, 2016; Loizou & Richtárik, 2020; Alderman et al., 2024; Zeng et al., 2024). The closest of these to our setting is Alderman et al. (2024), who study how momentum acts on the error along the singular directions of a least-squares system. We instead compare update rules by the squared risk at a random query time under Markovian access, include reflect–project, and give a criterion over all initial errors. Sketched methods can orthogonalize against past directions (Patel et al., 2021). Exact mean-square analyses and the observation that the error concentrates along small singular directions separate the total error from its directional parts (Agaskar et al., 2014; Steinerberger, 2021a). Noisy and inconsistent systems require extended iterations or averaging (Needell, 2010; Zouzias & Freris, 2013; Epperly et al., 2026). The gain range $(0, 2)$ is the classical relaxation range, and reflections, which use gain 2, appear in reflection–projection methods and in relaxed alternating projections (Bauschke & Kruk, 2004; Bauschke et al., 2016). Randomized reflections preserve the error norm at each step, yet their averages converge (Steinerberger, 2021b).

These analyses measure convergence of the iterates or of a returned average. *Our increment* is to use gain 2 inside an online schedule and to evaluate it at a random query time. The reflect–project rule sees only the active key, waits while the key repeats, and can be queried between a reflection and the projection that completes it. The contraction of a completed reflect–project block therefore

does not decide the comparison. The expected risk at the query does, and we compute it exactly (Proposition 3.4).

Dependent access and access order. Stochastic optimization under Markovian sampling has been analyzed for incremental subgradient methods (Johansson et al., 2009), for Markov chain gradient descent (Sun et al., 2018), for least squares with Markovian data (Nagaraj et al., 2020), and for SGD in general (Even, 2023), using spectral and mixing tools from Markov chain theory (Norris, 1997; Levin & Peres, 2017). Access order also separates random reshuffling from sampling with replacement. Upper bounds show when reshuffling helps (Gürbüzbalaban et al., 2021; HaoChen & Sra, 2019; Mishchenko et al., 2020), and lower bounds show its limits (Safran & Shamir, 2020; 2021). For momentum specifically, periodic covariates can drive SGD with momentum to diverge through resonance (Banman et al., 2022).

In these works the access order affects how fast an optimizer converges. *Our increment* is to show that it can decide which update rule wins. The eigenvalues t_ℓ of the access kernel enter every modal margin g_ℓ of the compatibility matrix M, so the same keys under different reversible kernels can give $\lambda_{\min}(M)$ of either sign (Proposition 4.2, Appendix B.2), and the finite-ε comparison on natural features follows this sign (Appendix A.5.2).

Risk at a random time, query directions, and key geometry. The expected quadratic cost of a linear system whose dynamics switch with a Markov chain is the subject of the theory of Markov jump linear systems (Fang & Loparo, 2002; Costa et al., 2005; 2011; 2013), including quadratic costs over horizons given by stopping times (Nespoli et al., 2004). In least-squares learning, the excess risk of SGD depends on the full covariance spectrum and on the initialization, not only on the average conditioning (Zou et al., 2021; 2023).

We use the jump-linear-system machinery as a tool. *Our increment* is to make the query an explicit part of the comparison. The query time $Q \sim \mathrm{Exp}(\lambda)$ turns the retrieval risk into a discounted quadratic cost with closed forms, and the query matrix C fixes which directions are evaluated. The reconstruction and the associative query can rank the same rules in opposite orders (Appendix B.1.1), dictionaries with equal mean similarity can fall on opposite sides of the criterion (Appendix B.2), and the elapsed time moves momentum while it leaves the scalar families unchanged (Appendix B.1.2).