TRAJECTORY REPLAY IN SEQUENTIAL FLOW MATCHING: CONCENTRATION & COVERAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work on sequential flow matching shows that the infinite-capacity empirical velocity field is a kernel regressor over stored transitions, approaching nearest-neighbour trajectory replay as its weights concentrate. We characterise when this occurs and identify which modelling choices control it. For Gaussian conditional paths, we characterise the onset of replay, its behaviour in high dimensions, and how it depends on the geometry of the stored transitions and the choice of conditional path. We further show that, under constant-in-flow-time covariance, the empirical field remains within the convex hull of the stored velocity targets, imposing a representation floor that cannot be overcome by reweighting. Experiments on chaotic systems validate the predicted concentration and coverage behaviour, while finite-capacity experiments show that neural networks can depart from the empirical minimiser and thereby mitigate replay. Together, these results distinguish replay concentration from target coverage, showing that controlling replay alone does not guarantee accurate prediction.

1 INTRODUCTION

Flow matching and stochastic-interpolant formulations provide a regression-based approach to learning flow and diffusion models (Lipman et al., 2023; Tong et al., 2024; Albergo et al., 2025; Pooladian et al., 2023). Their use for generating data with temporal structure can adopt diverse paradigms, with growing attention to modelling the evolution of sequences generated by dynamical systems (Zhang et al., 2024; Chen et al., 2024). When conditioning on multiple consecutive states, multi-marginal approaches learn consistent paths across timestamps (Lee et al., 2025; Islam et al., 2025; Rohbeck et al., 2025; Kviman et al., 2026). In these settings, this work is concerned with the optimal velocity field obtained when the class of admissible models for such a field is unrestricted.

For canonical Gaussian conditional paths between states and a training set of state transitions, Lim et al. (2026) show that the empirical minimiser, denoted $v_t^\star$, has a memory-based structure given by a weighted kernel estimator. This Nadaraya-Watson estimator is known to suffer from the curse of dimensionality, with convergence rates deteriorating with the state dimension (Stone, 1980). What remains unclear in the sequential setting is *if/when* this estimator enters the nearest-neighbour regime (aka trajectory replay) and which modelling choices can alter that phenomenon.

Building on the multi-marginal extension of Lim et al. (2026), we formulate and address two distinct questions; **Concentration:** How does the flow-model specification control the transition of $v_t^\star$ from kernel averaging to nearest-neighbour replay? **Target coverage:** Can convex combinations of the stored conditional targets represent the required velocity at a given query? This distinction, illustrated geometrically in Fig. 1, separates the mechanism producing replay from the question of whether replay, smoothing or any intermediate weighting scheme can be accurate.

Our contributions are as follows: (i) We introduce a logarithmic discrepancy measure for the kernel-estimator weights and derive its law. (ii) We define a concentration parameter determined by the training set and probability path, and show that replay emerges when such a parameter grows above a threshold of order $\log M$, where M is the number of stored state transitions. (iii) We characterise the effect of the probability path on replay concentration. (iv) For constant-covariance paths, we show that $v_t^\star$ lies in the convex hull of the stored velocity targets, establishing which velocities are unattainable by reweighting. (v) We experimentally validate these theoretical claims.

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.

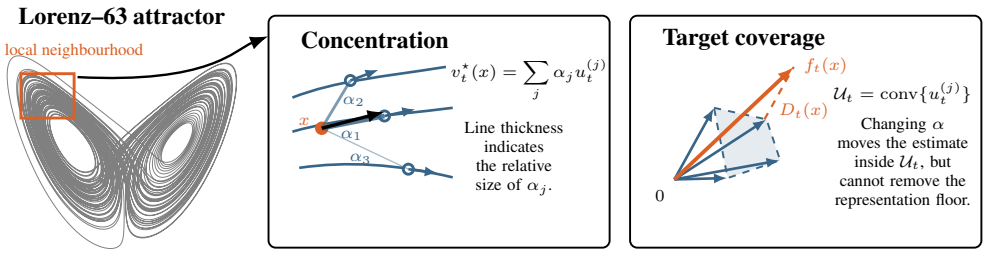


Figure 1: Schematic distinction between *concentration* and *target coverage* questions.

All theoretical claims concern the infinite-capacity (unrestricted) minimiser $v_t^\star$ of the empirical flow-matching objective for a fixed trajectory dataset. Accordingly, our experiments evaluate this closed-form field directly, with an additional finite-capacity experiment assessing the extent to which trained neural networks inherit its behaviour. We use chaotic systems as a primary testbed because their trajectories repeatedly visit the same regions without exactly repeating, making the distinction between trajectory replay and gaps in the stored trajectory bank particularly apparent.

2 Preliminaries

Sequential flow matching and notation. Flow matching specifies a conditional probability path $p_t(\cdot \mid Z)$ and velocity target $u_t^\circ(x \mid Z)$, to learn a velocity field by minimising the loss given by

$$\mathcal{L}_v(v) = \mathbb{E}_{t,Z,X_t}\left[\|v_t(X_t) - u_t^\circ(X_t \mid Z)\|^2\right], \qquad X_t \sim p_t(\cdot \mid Z).$$

For an unrestricted square-integrable class of velocity-field functions, the minimiser is the usual squared-loss regression solution (Lipman et al., 2023),

$$v_t^*(x) = \mathbb{E}[u_t^\circ(X_t \mid Z) \mid X_t = x].$$

In sequential flow matching, the conditioning variable Z is constructed from observed states $x_0, \ldots, x_T \in \mathbb{R}^d$. We define a window, of order $k \geq 1$, by $W_\tau = (x_\tau, \ldots, x_{\tau+k})$, and use the empirical distribution over the M stored windows as the multi-marginal coupling, $\hat{\pi} = \frac{1}{M}\sum_{j=1}^M \delta_{W_j}$. Notice that τ denotes physical or observation time, while $t \in [0,1]$ denotes flow time.

For an isotropic Gaussian conditional path $p_t(\cdot \mid W) = \mathcal{N}(\mu_t(W), \sigma_t^2 I_d)$, where $\mu_t(W)$ is specified by the chosen interpolant, the conditional velocity target is

$$u_t^\circ(x \mid W) = \dot{\mu}_t(W) + \frac{\sigma_t'}{\sigma_t}\big(x - \mu_t(W)\big).$$

For constant variance, $\sigma_t' = 0$, and the target reduces to $\dot{\mu}_t(W)$.

Empirical minimiser and the *trajectory replay* phenomenon. We state the empirical minimiser presented by Lim et al. (2026)(Appendix B, Proposition 4).

Lemma 1 (Empirical velocity minimiser (Lim et al., 2026)). *Fix $t \in (0,1)$ and let $\hat{\pi} = \frac{1}{M}\sum_{j=1}^M \delta_{Z_j}$ be the uniform empirical coupling over the stored conditioning variables $Z_1, \ldots, Z_M$. Suppose each conditional path $p_t(\cdot \mid Z_j)$ has a density that is strictly positive on $\mathbb{R}^d$, and $\int p_t(x \mid Z_j)\,\|u_t^\circ(x \mid Z_j)\|^2\,dx < \infty$ for each j. Then, over square-integrable measurable fields, $\mathcal{L}_v$ has the unique a.e. minimiser*

$$v_t^\star(x) = \sum_{j=1}^M \alpha_j(t,x)\, u_t^\circ(x \mid Z_j), \qquad \alpha_j(t,x) = \frac{p_t(x \mid Z_j)}{\sum_{l=1}^M p_t(x \mid Z_l)},$$

that is, $v_t^\star(x) = \mathbb{E}[u_t^\circ(X_t \mid Z) \mid X_t = x]$ with $Z \sim \hat{\pi}$ and $X_t \mid Z \sim p_t(\cdot \mid Z)$.

Remark. *For a Gaussian path p_t, the minimiser in Lemma 1 takes the form of an explicit Gaussian kernel smoother. Depending on the dispersion of the α_j, its weights can concentrate on a single point, showing that replay is intrinsically linked to the choice of conditional path.*

Explicitly, for a Gaussian path $p_t(x \mid Z_j) \propto \exp\Big(-\|x-\mu_t^{(j)}\|^2/(2\sigma_t^2)\Big)$, where $\mu_t^{(j)} = \mu_t(Z_j)$, these weights become

$$\alpha_j(t,x) = \frac{\exp\Big(-\|x-\mu_t^{(j)}\|^2/(2\sigma_t^2)\Big)}{\sum_l \exp\Big(-\|x-\mu_t^{(l)}\|^2/(2\sigma_t^2)\Big)}. \tag{1}$$

The kernel bandwidth is $2\sigma_t^2$. The conditioning variable Z_j may be a pairwise transition or a higher-order window. Critically, window order and interpolation can change the interpolated path centres and conditional targets, while the similarity weight form of the estimator remains invariant.

Proofs of all theorems and propositions in Sections 3 to 5 are provided in Section A.

3 When Trajectory Replay occurs

We now ask when the weights in eq. (1) become concentrated enough for trajectory replay. Consider a query point $x \in \mathbb{R}^d$ (i.e., a state at which the empirical velocity field is evaluated) drawn from conditional path associated with stored window i as $x = \mu_t^{(i)} + \sigma_t\varepsilon, i \sim \mathrm{Unif}\{1,\dots,M\}, \varepsilon \sim \mathcal{N}(0, I_d)$. By concentration, we mean that a single weight in eq. (1) carries nearly all of the mass, meaning that replay corresponds to $\alpha_l(t,x) \approx 1$ for some window l. For the shared isotropic Gaussian path, and a pair of windows i and j, define the dimensionless concentration ratio

$$\kappa_{ij} := \frac{\|\mu_t^{(i)} - \mu_t^{(j)}\|^2}{\sigma_t^2},$$

which compares the squared separation between two path centres with the conditional path variance, and will be the basic quantity controlling concentration in our work.

We first provide a pairwise comparison between windows, and then consider the full memory bank. We proceed by defining and analysing the log-weight gap as follows.

Theorem 1 (The log-weight gap is Gaussian)**.** *Fix two distinct stored windows i and j, and let $x = \mu_t^{(i)} + \sigma_t\varepsilon$ and $\Delta_{ij} := \mu_t^{(i)} - \mu_t^{(j)}$. Conditionally, on the stored path centres $\mu_t^{(i)}$ and $\mu_t^{(j)}$ the log-weight gap satisfies*

$$G_{ij} := \log\alpha_i - \log\alpha_j = \frac{\|\Delta_{ij}\|^2}{2\sigma_t^2} + \frac{\langle\Delta_{ij},\varepsilon\rangle}{\sigma_t} \;\sim\; \mathcal{N}(\kappa_{ij}/2,\ \kappa_{ij}).$$

Therefore, $\mathbb{P}(G_{ij} < 0) = \Phi(-\frac{1}{2}\sqrt{\kappa_{ij}})$ when $\Delta_{ij} \neq 0$.

κ_{ij} can now be identified as a **concentration parameter**. When $\kappa_{ij} \ll 1$, the two windows receive comparable weights; when $\kappa_{ij} \gg 1$, the generating window i is likely to dominate. The corresponding signal-to-noise ratio is $\frac{\mathbb{E}[G_{ij}]}{\sqrt{\mathrm{Var}(G_{ij})}} = \frac{1}{2}\sqrt{\kappa_{ij}}$.

3.1 Replay onset and memory bank size

We can now answer the question of when replay can be expected to occur. Theorem 1 characterises the competition between generating window i and a single alternative j, but trajectory replay is a statement about the full collection of M posterior weights. To connect pairwise gaps to the full set of weights, fix the generating window i and define

$$S := \sum_{j\neq i} e^{-G_j} \geq 0, \qquad S_2 := \sum_{j\neq i} e^{-2G_j} \geq 0.$$

Since $\alpha_j = \alpha_i e^{-G_j}$, we can express the ith weight as $\alpha_i = \frac{1}{1+S}$ and the effective sample size as

$$\mathrm{ESS} = \frac{1}{\sum_j \alpha_j^2} = \frac{(1+S)^2}{1+S_2} \leq (1+S)^2.$$

Thus replay corresponds to $S \ll 1$: in that case, the combined weight of all competitors $\{\alpha_j\}_{j\neq i}$ is small, giving $\alpha_i \approx 1$ and $\mathrm{ESS} \approx 1$.

Let $n := M - 1$, and write $\bar{\kappa} := \frac{1}{n}\sum_{j\neq i}\kappa_{ij}$ and $\kappa_{\min} := \min_{j\neq i}\kappa_{ij}$ for the mean and minimum concentration over competitors.

Proposition 1 (Replay onset is logarithmic in M)**.** *Under the assumptions of Theorem 1 and any $\delta \in (0, 1/2]$:*

(i) (necessity) *Denote $z_\delta := \Phi^{-1}(1-\delta)$, with probability at least $1-\delta$,*

$$\log S \;\geq\; \log n - \frac{\bar{\kappa}}{2} - z_\delta\sqrt{\bar{\kappa}}.$$

In particular, if $\bar{\kappa} \leq 2(1-\eta)\log n$ for a fixed $\eta \in (0,1)$, then $\alpha_i \xrightarrow{\mathbb{P}} 0$ as $M \to \infty$;

(ii) (sufficiency) *if $\kappa_{\min} \geq 32\log(n/\delta)$, then with probability at least $1-\delta$,*

$$1 - \alpha_i \leq \frac{\delta^8}{n^7}, \qquad \mathrm{ESS} \leq \left(1 + \frac{\delta^8}{n^7}\right)^2.$$

When pairwise separations concentrate, $\kappa_{\min} \approx \bar{\kappa}$, which we simply call κ. Proposition 1 shows that replay requires κ to grow on the order of $\log M$ as the number of competing windows increases.

3.2 High dimensional scaling

We now identify how κ_{ij} changes with the ambient dimension d. For independent windows W, W' with the same law, and their respective interpolant means, define

$$V_{k,d}(t) := \frac{1}{d}\mathbb{E}\left[\|\mu_t(W) - \mu_t(W')\|^2\right] = \frac{2}{d}\operatorname{tr}\operatorname{Cov}(\mu_t(W)). \tag{2}$$

Thus $V_{k,d}(t)$ is the average squared separation per coordinate of the interpolated window cloud, dependent on window order k, flow time t, window distribution and interpolant. The quantity may change as the ambient dimension changes.

The high-dimensional replay result in Theorem 2 requires two properties of these separations. First, the interpolated window cloud should retain non-vanishing spread per coordinate as dimension increases, that is, $\liminf_{d\to\infty} V_{k,d}(t) > 0$. Second, pairwise squared separations should concentrate around their population scale $dV_{k,d}(t)$. We assume

$$\|\mu_t(W) - \mu_t(W')\|^2 = dV_{k,d}(t) + O_p(\sqrt{d}). \tag{3}$$

where O_p denotes stochastic order in probability. Equivalently, whenever $V_{k,d}(t)$ remains bounded away from zero,

$$\frac{\|\mu_t(W) - \mu_t(W')\|^2}{dV_{k,d}(t)} = 1 + O_p(d^{-1/2}).$$

A sufficient condition for Eq. (3) is that the coordinate-wise squared differences have bounded moments and sufficiently weak cross-coordinate dependence, so that $\mathrm{Var}\big(\|\mu_t(W) - \mu_t(W')\|^2\big) = O(d)$.

For stored windows i and j whose separation obeys Eq. (3), it follows that

$$\kappa_{ij} = \frac{\|\mu_t^{(i)} - \mu_t^{(j)}\|^2}{\sigma_t^2} = \frac{dV_{k,d}(t)}{\sigma_t^2} + O_p\left(\frac{\sqrt{d}}{\sigma_t^2}\right). \tag{4}$$

Thus the leading concentration scale is $dV_{k,d}(t)/\sigma_t^2$. In particular, if $V_{k,d}(t) = \Theta(1)$ and $\sigma_t^2 = \Theta(1)$, then κ_{ij} grows linearly with d to leading order.

Theorem 2 (High-dimensional replay)**.** *Consider the generated query $x = \mu_t^{(i)} + \sigma_t\varepsilon, \varepsilon \sim \mathcal{N}(0, I_d)$, and assume that $\liminf_{d\to\infty} V_{k,d}(t) > 0$. Assume further that the squared window separations concentrate around $dV_{k,d}(t)$ as in Eq.* (3)*, and that $\sigma_t^2 \leq \bar{\sigma}^2$ for some $\bar{\sigma} < \infty$ independent of d. If M is fixed, then as $d \to \infty$, $\alpha_i(t,x) \xrightarrow{\mathbb{P}} 1$, $\mathrm{ESS}(t,x) \xrightarrow{\mathbb{P}} 1$.*

From Eq. (4), when pairwise separations are comparable, we use κ to denote their typical concentration scale:

$$\kappa \asymp \frac{dV_{k,d}(t)}{\sigma_t^2}. \tag{5}$$

Combining the order-$\log M$ replay onset from Proposition 1 with Eq. (5) gives $d_\star \asymp \sigma_t^2 \log M / V_{k,d}(t)$.

Remark. *Replay can thus be said to arise when the path-centre separation grows with d faster than the chosen conditional-path variance. Section A.1.6 extends this to component-dependent variances.*

4 What controls replay concentration

The analysis above identifies the replay concentration scale as $\kappa \asymp \frac{dV_k(t)}{\sigma_t^2}$. We use this decomposition of κ to ask which common design choices in sequential flow matching can alter the dimensional onset of replay: coupling and interpolation through $V_k(t)$, the addition of a score objective, kernel bandwidth and covariance geometry.

Multi-marginal and windowed flow-matching methods replace pairwise transitions by higher order couplings and richer interpolants. We study their effects on concentration below.

Proposition 2 (Coupling and interpolant do not alter dimensional scaling)**.** *Under the assumptions of Eq.* (3)*, for fixed window order k with $V_k(t) = \Theta(1)$, coupling and interpolation affect replay concentration only through $V_k(t)$; the leading d/σ_t^2 scaling in Eq.* (4) *is unchanged.*

Coupling and interpolation can change *what* is being replayed, but they do not change the concentration dimensional dependence directly.

A number of sequential and multi-marginal methods augment the velocity regression with a score-matching objective. We study whether the joint score objective changes replay mechanism.

Proposition 3 (The score estimator inherits the replay weights)**.** *Under the shared isotropic Gaussian conditional path $p_t(\cdot \mid W) = \mathcal{N}(\mu_t(W), \sigma_t^2 I_d)$, the infinite-capacity empirical minimizer $\mathcal{L}_s(s) = \mathbb{E}[\|s_t(X_t) - \nabla_x \log p_t(X_t \mid W)\|^2]$ of the score objective is $s_t^\star(x) = \sum_j \alpha_j(t,x) \frac{\mu_t^{(j)} - x}{\sigma_t^2}$, where $\alpha_j(t,x)$ are the same posterior weights appearing as in Eq.* (1)*.*

Consequently, adding a score objective does not change the infinite-capacity empirical concentration governing trajectory replay. This does not rule out a finite-capacity benefit from score matching.

4.1 Bandwidth regime

Within the isotropic family, the remaining direct control on concentration is the bandwidth σ_t^2. Suppose $\sigma_t^2 \asymp d^a$. Since $\kappa \asymp \frac{dV_k(t)}{d^a} = d^{1-a} V_k(t)$, the asymptotic behaviour is determined by exponent a

Theorem 3 (Bandwidth regimes)**.** *Fix M, k and $t \in (0,1)$. Suppose $V_{k,d}(t) = \Theta(1)$, pairwise separations satisfy Eq.* (3)*, and $\sigma_t^2 \asymp d^a$. Then, as $d \to \infty$,*

$$\begin{aligned}
a < 1 \quad &\Longrightarrow \quad \alpha_i(t,x) \xrightarrow{\mathbb{P}} 1, \quad \mathrm{ESS}(t,x) \xrightarrow{\mathbb{P}} 1, \\
a = 1 \quad &\Longrightarrow \quad \kappa_{ij} \text{ remains order one in probability}, \\
a > 1 \quad &\Longrightarrow \quad \alpha_j(t,x) \xrightarrow{\mathbb{P}} 1/M \ \forall j, \quad \mathrm{ESS}(t,x) \xrightarrow{\mathbb{P}} M.
\end{aligned}$$

The case $a = 1$ represents neither one-hot nor uniform averaging. The usual dimension-independent bandwidth lies in the replay regime. Linear variance scaling $\sigma_t^2 = \Theta(d)$ keeps the weights non-degenerate, while faster growth drives them toward uniformity.

Remark (Concentration is not a performance objective)**.** *We note here that preventing nearest-neighbour replay is not an indicator of performance improvement; sufficiently increasing the bandwidth can remove replay by replacing it with uniform averaging. We thus treat* ESS *as a useful description of an estimator's weighting behaviour, but not a sufficient quantity for determining performance. We defer that work to the empirical analysis.*

4.2 SHARED ANISOTROPIC COVARIANCE

In high-dimensional nonparametric regression, structural assumptions such as anisotropic bandwidths or low dimensional manifold structure can replace dependence on ambient dimension by a lower-dimensional quantity (Conn & Li, 2019; McRae et al., 2020). This motivates asking whether anisotropy changes the replay mechanism itself, or only how distances between stored windows are measured. The following result shows that the same log-weight law holds, with Euclidean distance replaced by a covariance-weighted Mahalanobis distance.

Proposition 4 (Anisotropic log-weight law). *Let* $x = \mu_t^{(i)} + A\varepsilon, \qquad AA^\top = \Sigma_t, \qquad \varepsilon \sim \mathcal{N}(0, I_d)$, *and define* $\Delta_{ij} = \mu_t^{(i)} - \mu_t^{(j)}$, $\kappa_{ij}^\Sigma = \Delta_{ij}^\top \Sigma_t^{-1} \Delta_{ij}$. *Conditionally on the interpolated window locations,*

$$G_{ij} = \log \alpha_i - \log \alpha_j \sim \mathcal{N}\left(\frac{\kappa_{ij}^\Sigma}{2}, \kappa_{ij}^\Sigma\right).$$

Thus, anisotropy changes which stored windows are considered to be close without changing the concentration mechanism. We can therefore ask how the covariance should be shaped to reduce concentration as much as possible without increasing its total variance.

Let $\Gamma \coloneqq \mathbb{E}[\Delta_{ij}\Delta_{ij}^\top]$ be the covariance of the displacements between stored path centres.

Proposition 5 (Concentration-optimal shared covariance). *Suppose* $\Gamma \succ 0$. *Among all shared covariances* $\Sigma_t \succ 0$ *satisfying* $\operatorname{tr}(\Sigma_t) = d\sigma_t^2$, *the expected concentration* $\mathbb{E}[\kappa_{ij}^\Sigma] = \operatorname{tr}(\Sigma_t^{-1}\Gamma)$ *is minimised by* $\Sigma_t^\star = \frac{d\sigma_t^2}{\operatorname{tr}(\Gamma^{1/2})}\Gamma^{1/2}$. *At this optimum,* d *is replaced by* $r_{\mathrm{eff}}(\Gamma)$

$$\mathbb{E}[\kappa_{ij}^{\Sigma^\star}] = \frac{r_{\mathrm{eff}}(\Gamma) V_k(t)}{\sigma_t^2}, \qquad r_{\mathrm{eff}}(\Gamma) = \frac{\left(\operatorname{tr}\Gamma^{1/2}\right)^2}{\operatorname{tr}\Gamma}.$$

The optimal covariance therefore allocates more variance in directions where the stored windows vary more. Under this choice, the ambient dimension d is replaced by $r_{\mathrm{eff}}(\Gamma)$, which is small when the displacement cloud is concentrated in relatively few directions and equals d in the isotropic case.

For completeness, Section A.2.6 treats heavy-tailed scale-mixture paths. These alter the rate of concentration but do not remove the high-dimensional replay limit.

5 TARGET-SPACE COVERAGE

The concentration analysis above describes how the infinite-capacity empirical estimator distributes weight across stored components, but whether the stored targets can represent the field being estimated is an open question. Throughout this section, covariance is constant in flow time $\Sigma_t' = 0$, so that the conditional targets $u_t^\circ(x \mid Z_j) = \dot{\mu}_t^{(j)}(Z_j)$ are independent of query x.

Definition 1 (Target cloud and coverage functional). *Let* $u_t^{(j)} \coloneqq \dot{\mu}_t(W_j)$, $j = 1, \dots, M$, *denote the stored conditional targets, and let* $\mathcal{U}_t \coloneqq \operatorname{conv}\{u_t^{(1)}, \dots, u_t^{(M)}\} \subset \mathbb{R}^d$ *be their convex hull. Given a reference velocity field* $f_t : \mathbb{R}^d \to \mathbb{R}^d$, *representing the true target field against which the empirical estimator is evaluated, define the target-space coverage functional*

$$D_t(x) \coloneqq \operatorname{dist}(f_t(x), \mathcal{U}_t) = \min_{w \in \mathcal{U}_t} \|f_t(x) - w\|.$$

$\mathcal{U}_t$ is determined by the memory bank, coupling order and interpolant. It is independent of the query x and, under the constant-covariance assumption, of the shared bandwidth or covariance. Since $v_t^\star(x)$ is a convex weighting of these targets, this produces a representation constraint, formalised in Theorem 4.

Theorem 4 (Convex-hull representation floor). *Fix* $t \in (0, 1)$ *and let* $v_t^\star$ *be the empirical minimizer of Lemma 1 under* $p_t(\cdot \mid W) = \mathcal{N}(\mu_t(W), \Sigma_t)$, *with* $\Sigma_t \succ 0$ *and* $\Sigma_t' = 0$. *Then, for every* $x \in \mathbb{R}^d$,

$$v_t^\star(x) \in \mathcal{U}_t, \qquad \|v_t^\star(x) - f_t(x)\| \geq D_t(x),$$

where the latter bound is tight over all convex weightings of the stored targets. Consequently, for any query law Q *held fixed as the shared covariance varies,*

$$\inf_{\Sigma_t \succ 0} \mathbb{E}_{x \sim Q} \|v_t^\star(x) - f_t(x)\|^2 \geq \mathbb{E}_{x \sim Q}[D_t(x)^2].$$

For a fixed target bank and reference field, changing the shared covariance only reweights the same stored targets and therefore leaves $D_t(x)$ unchanged. Even the optimal convex weighting cannot cross this floor. Thus, reducing replay can improve accuracy only when the target bank is already rich enough to represent the required field. A corresponding necessary condition based on the span of the target bank is given in Section A.3.2. We also address a deterministic-interpolant limit in Section A.4.

6 Empirical Validation

The theory separates two properties of the estimator. We test these two mechanisms directly on the closed-form empirical minimiser in Sections 6.1 to 6.3, before briefly examining their transfer to finite-capacity networks in Section 6.4. Experimental details are deferred to the appendix.

6.1 Replay onset is governed by concentration

We test whether changes in dimension, window order and interpolant are absorbed by the predicted concentration parameter $\widehat{\kappa} = d\widehat{V}_k(t)/\sigma_t^2$.

Figure 2 shows that the ESS curves collapse when expressed in terms of $\widehat{\kappa}$, pairwise separations concentrate at the predicted $d^{-1/2}$ relative rate, and the concentration required for replay grows logarithmically with M. These observations support κ, rather than ambient dimension alone, as the variable controlling the replay–smoothing transition. The measured memory-bank scaling also lies within the constant-factor bounds of Proposition 1.

6.2 Coverage constrains the achievable error

We vary the distance of query states from the stored bank while sweeping concentration. As an observable state-space proxy for support coverage, define $\rho = \frac{d_{\mathrm{nn}}}{\sqrt{d\widehat{V}}}$, where d_{nn} is the distance from a query to its nearest stored state. For each query family we report e_{best}, the lowest normalized error observed over the prescribed concentration grid. We separately compute the exact target-space hull floor D_t from Theorem 4.

Figure 3 shows that error increases strongly as state-space coverage worsens, while a positive D_t remains an uncrossable lower bound regardless of bandwidth or covariance. Thus, concentration controls how the available targets are combined, whereas coverage constrains what those targets can represent. Reducing replay can improve the weighting of a sufficiently rich target bank, but cannot remove a representation error caused by missing target directions.

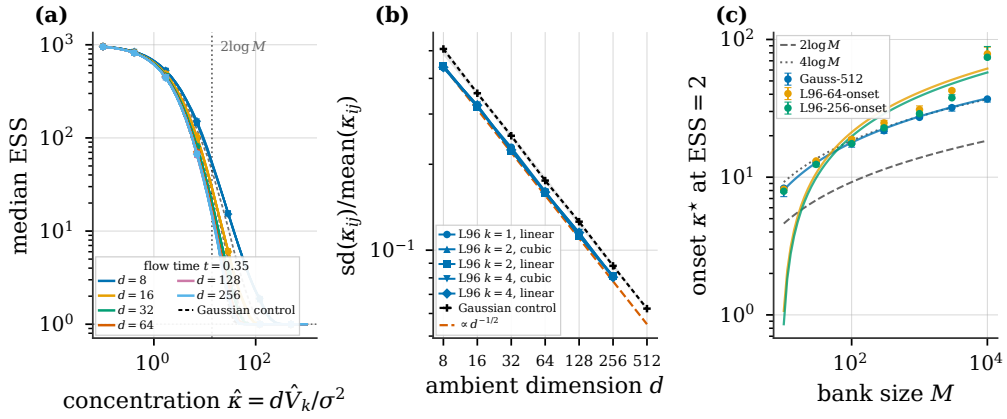


Figure 2: **Replay onset is controlled by the concentration parameter.** (a) Median effective sample size as a function of $\widehat{\kappa} = d\widehat{V}_k/\sigma^2$ for Lorenz–96 across ambient dimensions, together with an iid Gaussian control. (b) Relative spread of pairwise concentrations. (c) The concentration $\kappa^\star$ at which ESS $= 2$ grows logarithmically with memory-bank size M.

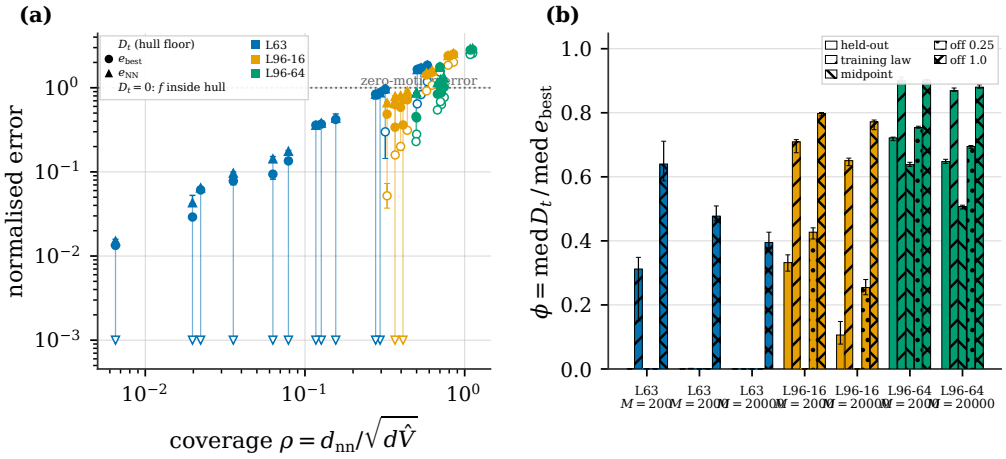


Figure 3: **Support coverage constrains estimator error.** (a) Best closed-form and nearest-neighbour error versus normalized query-to-bank distance ρ. Stems show the hull floor D_t; open markers indicate $D_t = 0$, and the dashed line is the zero-motion error. (b) Fraction $\phi = \mathrm{med}\, D_t / \mathrm{med}\, e_{\mathrm{best}}$ across query families and bank sizes.

6.3 Concentration is not itself a prediction objective

To separate covariance shape from its effect on concentration, we compare anisotropic and isotropic kernels at realised $\bar{\kappa}^{\Sigma} = \mathrm{tr}(\Sigma^{-1}\widehat{\Gamma})$.

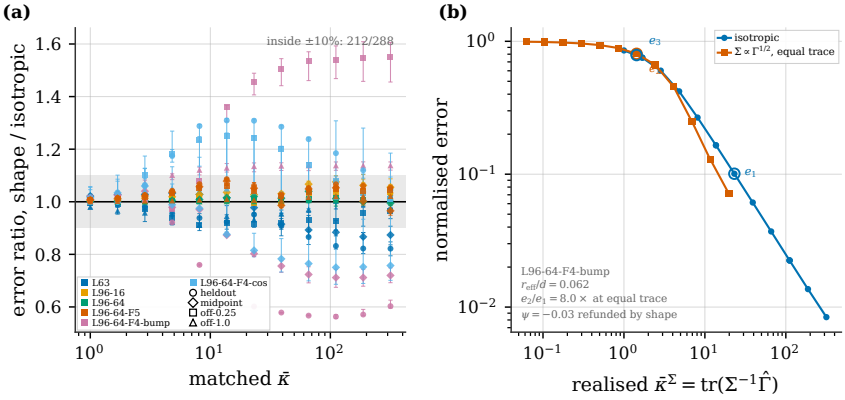


Figure 4: **Covariance geometry and concentration have distinct effects.** (a) Anisotropic/isotropic error ratio at matched realised concentration; 212 of 288 comparisons lie within $\pm 10\%$. (b) Equal-trace half-whitening and an isotropic kernel retuned to the same concentration on L96-64-F4-bump.

Figure 4 shows that matching concentration removes most of the prediction-error difference in most configurations, although residual shape effects remain in deliberately low-effective-rank cases.

The effective-rank result should therefore be interpreted as a statement about how strongly covariance geometry can change replay concentration under a fixed variance budget, not as a prediction-error optimum. Prediction additionally depends on target coverage and, in some cases, covariance geometry beyond its induced concentration.

6.4 Finite capacity networks partially inherit replay

Finally, we train MLPs of widths $W \in \{128, 512\}$ on a fixed Lorenz–96 problem and vary concentration relative to the $\Theta(\log M)$ replay scale. Figure 5 shows that both networks track $v^{\star}$ at low concentration, while the wider network inherits substantially more of its replay behaviour as concentration increases.

The narrower network departs further from the empirical minimiser and can achieve lower off-support error, where it is not constrained to the convex hull of the stored targets. This is consistent

with finite capacity acting as an implicit regulariser in this experiment, rather than with replay being inherited automatically by every trained model (capacity and optimisation effects are not separated here).

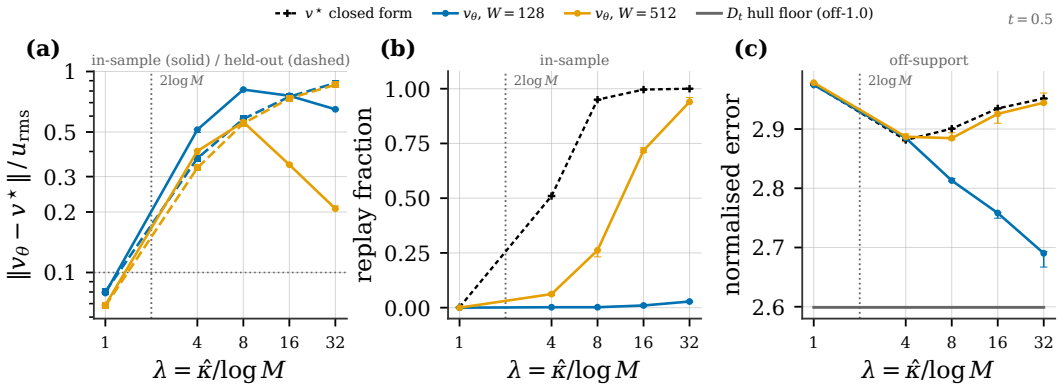


Figure 5: **Finite-capacity networks need not inherit the empirical minimiser.** (a) Network–closed-form discrepancy on in-sample and held-out queries. (b) In-sample single-target replay fraction. (c) Off-support error; the horizontal line is the hull floor D_t for $v^\star$, and the dotted line marks $\widehat{\kappa} = 2\log M$.

7 Related Work

Sequential flow matching and stochastic interpolants have been developed for forecasting, conditional trajectories and multi-marginal temporal modelling (Zhang et al., 2024; Chen et al., 2024; Tamir et al., 2024; Lim et al., 2025; Lee et al., 2025; Islam et al., 2025; Rohbeck et al., 2025; Kviman et al., 2026), while coupling design has been used to simplify or straighten learned flows (Pooladian et al., 2023; Tong et al., 2024). Our analysis instead studies the unrestricted empirical regression solution and when its posterior weights concentrate onto stored sequential targets. This connects the problem to classical high-dimensional nonparametric regression and distance concentration (Beyer et al., 1999; Bickel & Li, 2007; Kpotufe, 2011), as well as to recent work on memorisation and finite-sample behaviour in diffusion models (Carlini et al., 2023; Somepalli et al., 2023; Biroli et al., 2024; Achilli et al., 2024; Kadkhodaie et al., 2024). A fuller comparison, including finite-capacity generalisation, is given in Section C.

8 Discussion and Limitations

Our results revealed two distinct behaviours in the empirical sequential flow-matching estimator. Specifically, we identified replay concentration, which concerns how strongly the posterior weights select individual stored windows, and target coverage, which concerns whether the stored conditional targets can represent the required velocity at all. These mechanisms respond differently to modelling choices: window order and interpolation change the geometry of the stored centres and targets, while bandwidth and covariance geometry control concentration. We also showed that, under constant-in-flow-time covariance, changing these weights cannot move the empirical field outside the convex hull of the stored targets, showing that reducing replay is not by itself a prescription for improving prediction. Through a finite-capacity experiment, we also showed that wider networks inherit substantially more of the closed-form replay behaviour, while narrower networks can depart from this and achieve lower off-support error.

Our analysis gives a starting point for two natural extensions. First, our concentration results characterise queries generated from a stored conditional component, which isolates the mechanism responsible for replay. Extending this analysis to off-support queries would connect the proposed weight-gap framework more directly to inference, where the query need not lie near any stored trajectory. Second, our coverage results show that, for constant-in-flow-time covariance, reweighting alone cannot overcome deficiencies in the stored target set. This motivates the study of more general probability paths. Such extensions could clarify when modifying the path can improve not only concentration, but also the representational capacity of the empirical field.

AI USE STATEMENT

In this work, we used generative AI tools for literature review, formalisation of notations and definitions and polishing up the writing. We have not used generative AI tools for unsupervised generation of any math/proofs done here. Generative AI tools were used for coding tasks. We have reviewed all AI-assisted code and figures. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The closed form experiments in Sections 6.1 to 6.3 evaluate the empirical flow-matching field directly; no network is trained. Randomness is controlled via fixed seeds defining the stored banks and evaluation queries, and comparisons within an experiment use the same banks and query sets.

The finite capacity experiment in Section 6.4 additionally fixes the training protocol and reports results across independent training seeds. Experiment-specific systems, sample counters, hyperparameters, aggregation procedures and implementation details are provided in Section B.

Code, configurations and scripts required to reproduce the experiments and figures will be made available upon request.

REFERENCES

Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion. *arXiv preprint arXiv:2410.08727*, 2024.

Beatrice Achilli, Luca Ambrogioni, Carlo Lucibello, Marc Mézard, and Enrico Ventura. Memorization and generalization in generative diffusion under the manifold hypothesis, 2025. URL https://arxiv.org/abs/2502.09578.

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Database Theory—ICDT '99*, volume 1540 of *Lecture Notes in Computer Science*, pp. 217–235. Springer, 1999.

Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pp. 177–186, 2007.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15:9957, 2024. arXiv:2402.18491.

Nicholas M. Boffi, Arthur Jacot, Stephen Tu, and Ingvar Ziemann. Shallow diffusion networks provably learn hidden low-dimensional structure. In *International Conference on Learning Representations (ICLR)*, 2025.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training, 2025. URL https://arxiv.org/abs/2505.17638.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning (ICML)*, 2023.

Yifan Chen, Mark Goldstein, Mengjian Hua, Michael Samuel Albergo, Nicholas Matthew Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and föllmer processes.

In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6728–6756. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24n.html.

Daniel Conn and Gang Li. An oracle property of the nadaraya–watson kernel estimator for high-dimensional nonparametric regression. *Scandinavian Journal of Statistics*, 46(3):735–764, 2019.

Mohammad Mohaiminul Islam, Thijs P. Kuipers, Sharvaree Vadgama, Coen de Vente, Afsana Khan, Clara I. Sánchez, and Erik J. Bekkers. Longitudinal flow matching for trajectory modeling, 2025.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations (ICLR)*, 2024.

Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. *Advances in neural information processing systems*, 24, 2011.

Oskar Kviman, Kirill Tamogashev, Nicola Branchini, Víctor Elvira, Jens Lagergren, and Nikolay Malkin. Multi-marginal flow matching with adversarially learnt interpolants. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 125197–125217, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cb0ce861adaf6f8a93069c064733f402-Paper-Conference.pdf.

Hugo Latourelle-Vigeant, Sinho Chewi, Aram-Alexandre Pooladian, John Sous, and Theodor Misiakiewicz. Generalization, memorization, and overfitting for diffusion models trained in the lazy high-dimensional regime. 2026. URL https://arxiv.org/abs/2608.23938.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Justin Lee, B. Moradijamei, and H. Shakeri. Multi-marginal stochastic flow matching for high-dimensional snapshot data at irregular time points. *arXiv preprint arXiv:2508.04351*, 2025.

Soon Hoe Lim, Yijin Wang, Annan Yu, Emma Hart, Michael W. Mahoney, Xiaoye S. Li, and N. Benjamin Erichson. Elucidating the design choice of probability paths in flow matching for forecasting. *Transactions on Machine Learning Research*, 2025. arXiv:2410.03229.

Soon Hoe Lim, Shizheng Lin, Michael W. Mahoney, and N. Benjamin Erichson. Is flow matching just trajectory replay for sequential data? *arXiv preprint arXiv:2602.08318*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Andreas Maurer. A bound on the deviation probability for sums of non-negative random variables. *Journal of Inequalities in Pure & Applied Mathematics*, 4(1):Paper No. 15, 6 p., 2003. URL http://eudml.org/doc/122913.

Andrew McRae, Justin Romberg, and Mark Davenport. Sample complexity and effective dimension for regression on manifolds. *Advances in Neural Information Processing Systems*, 33:12993–13004, 2020.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 28100–28127, 2023.

Martin Rohbeck, Edward De Brouwer, Charlotte Bunne, Jan-Christian Huetter, Anne Biton, Kelvin Y. Chen, Aviv Regev, and Romain Lopez. Modeling complex system dynamics with flow matching across time and conditions. In *International Conference on Learning Representations*, 2025.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058. IEEE, 2023.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2240947.

Ella Tamir, Najwa Laabid, Markus Heinonen, Vikas Garg, and Arno Solin. Conditional flow matching for time series modelling. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024. URL https://openreview.net/forum?id=Hqn4Aj7xrQ.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Xi Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis L. Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modeling. In *Advances in Neural Information Processing Systems*, 2024.

A Appendix: Proofs

The appendix is structured in the following way:

1. Section A contains the proofs for every theorem and proposition stated in the main text, as well as a few addition remarks and corollaries. Each section has its proofs in Sections A.1 to A.3.
2. Section B contains detailed description of the empirical tests conducted.
3. Section C contains a more detailed discussion of the project at large, including related works and future directions.

A.1 Proofs from Section 3

A.1.1 Proof of Theorem 1

Proof. By the Gaussian form of the posterior weights,

$$\alpha_j \propto \exp\left(-\frac{\|x-\mu_t^{(j)}\|^2}{2\sigma_t^2}\right),$$

with a normalising constant independent of j. Hence

$$G_{ij} = \log\alpha_i - \log\alpha_j = \frac{\|x-\mu_t^{(j)}\|^2 - \|x-\mu_t^{(i)}\|^2}{2\sigma_t^2}.$$

Since

$$x-\mu_t^{(j)} = \Delta_{ij} + \sigma_t\varepsilon, \qquad x-\mu_t^{(i)} = \sigma_t\varepsilon,$$

we have

$$\|x-\mu_t^{(j)}\|^2 = \|\Delta_{ij}\|^2 + 2\sigma_t\langle\Delta_{ij},\varepsilon\rangle + \sigma_t^2\|\varepsilon\|^2,$$

and

$$\|x-\mu_t^{(i)}\|^2 = \sigma_t^2\|\varepsilon\|^2.$$

The common quadratic noise term therefore cancels exactly, giving

$$G_{ij} = \frac{\|\Delta_{ij}\|^2}{2\sigma_t^2} + \frac{\langle\Delta_{ij},\varepsilon\rangle}{\sigma_t}.$$

Conditionally on $\{\mu_t^{(l)}\}$, the vector Δ_{ij} is deterministic and $\varepsilon\sim\mathcal{N}(0,I_d)$. Therefore

$$\langle\Delta_{ij},\varepsilon\rangle \sim \mathcal{N}(0,\|\Delta_{ij}\|^2),$$

and hence

$$\frac{\langle\Delta_{ij},\varepsilon\rangle}{\sigma_t} \sim \mathcal{N}(0,\kappa_{ij}).$$

Since

$$\frac{\|\Delta_{ij}\|^2}{2\sigma_t^2} = \frac{\kappa_{ij}}{2},$$

it follows that

$$G_{ij} \sim \mathcal{N}\left(\frac{\kappa_{ij}}{2},\kappa_{ij}\right).$$

For $\Delta_{ij}\neq 0$, standardising gives

$$\mathbb{P}(G_{ij}<0) = \Phi\left(\frac{-\kappa_{ij}/2}{\sqrt{\kappa_{ij}}}\right) = \Phi\left(-\frac{1}{2}\sqrt{\kappa_{ij}}\right).$$

□

A.1.2 Proof of Proposition 6

Proposition 6. *Let*

$$\kappa_{\min} \coloneqq \min_{j \neq i} \kappa_{ij}.$$

Under the assumptions of Theorem 1, with probability at least

$$1 - (M-1)e^{-\kappa_{\min}/32},$$

the effective sample size satisfies

$$\mathrm{ESS} \leq \left(1 + (M-1)e^{-\kappa_{\min}/4}\right)^2.$$

Proof. For $j \neq i$,

$$G_{ij} = \log \alpha_i - \log \alpha_j,$$

so

$$\alpha_j = \alpha_i e^{-G_{ij}}.$$

Set $G_i \equiv 0$. Using $\sum_j \alpha_j = 1$,

$$1 = \sum_j \alpha_j = \alpha_i \left(1 + \sum_{j \neq i} e^{-G_j}\right) = \alpha_i (1 + S),$$

and therefore

$$\alpha_i = \frac{1}{1+S}.$$

Similarly,

$$\sum_j \alpha_j^2 = \alpha_i^2 \left(1 + \sum_{j \neq i} e^{-2G_j}\right) = \frac{1 + S_2}{(1+S)^2}.$$

Thus

$$\mathrm{ESS} = \frac{(1+S)^2}{1+S_2} \leq (1+S)^2.$$

Fix $\theta = 1/4$. By Theorem 1,

$$\mathbb{P}\left(G_{ij} \leq \frac{1}{4}\kappa_{ij}\right) = \Phi\left(-\frac{1}{4}\sqrt{\kappa_{ij}}\right).$$

Using the Gaussian tail bound

$$\Phi(-z) \leq e^{-z^2/2}, \qquad z > 0,$$

gives

$$\mathbb{P}\left(G_{ij} \leq \frac{1}{4}\kappa_{ij}\right) \leq e^{-\kappa_{ij}/32} \leq e^{-\kappa_{\min}/32}.$$

Define the event

$$E \coloneqq \left\{G_{ij} > \frac{1}{4}\kappa_{ij} \quad \forall j \neq i\right\}.$$

By a union bound,

$$\mathbb{P}(E^c) \leq (M-1)e^{-\kappa_{\min}/32},$$

so

$$\mathbb{P}(E) \geq 1 - (M-1)e^{-\kappa_{\min}/32}.$$

On E,

$$G_{ij} > \frac{1}{4}\kappa_{ij} \geq \frac{1}{4}\kappa_{\min},$$

and therefore

$$e^{-G_j} < e^{-\kappa_{\min}/4}.$$

Summing over the $M-1$ competitors,

$$S < (M-1)e^{-\kappa_{\min}/4}.$$

Substituting this into the bound on ESS yields

$$\mathrm{ESS} \le \left(1 + (M-1)e^{-\kappa_{\min}/4}\right)^2$$

on an event of probability at least

$$1 - (M-1)e^{-\kappa_{\min}/32}.$$

□

A.1.3 Proof of Proposition 1

Proof. (i) By convexity of exp, $S = \sum_{j\neq i} e^{-G_j} \ge n\exp(-\bar{G})$ with $\bar{G} \coloneqq \frac{1}{n}\sum_{j\neq i} G_j$. By Theorem 1,

$$\bar{G} = \frac{\bar{\kappa}}{2} + \xi, \qquad \xi \coloneqq \frac{1}{n}\sum_{j\neq i} \frac{\langle \Delta_{ij}, \varepsilon\rangle}{\sigma_t}.$$

Conditionally on the path centres, ξ is a centred Gaussian. By the triangle inequality in L^2 and Jensen's inequality,

$$\mathrm{sd}(\xi) \le \frac{1}{n}\sum_{j\neq i}\sqrt{\kappa_{ij}} \le \sqrt{\bar{\kappa}}.$$

Since $z_\delta \ge 0$, $\mathbb{P}(\xi > z_\delta\sqrt{\bar{\kappa}}) \le \Phi(-z_\delta) = \delta$, which gives the bound on $\log S$. If $\bar{\kappa} \le 2(1-\eta)\log n$, the right-hand side is at least $\eta\log n - z_\delta\sqrt{2\log n} \to \infty$ for every fixed δ, so $S \xrightarrow{\mathbb{P}} \infty$ and $\alpha_i = 1/(1+S) \xrightarrow{\mathbb{P}} 0$.

(ii) Apply Proposition 6. Since $\kappa_{\min} \ge 32\log(n/\delta)$, the failure probability is at most $ne^{-\kappa_{\min}/32} \le \delta$. On the complementary event, $S < ne^{-\kappa_{\min}/4} \le n(\delta/n)^8 = \delta^8/n^7$, and then $1-\alpha_i = S/(1+S) \le S$ and $\mathrm{ESS} \le (1+S)^2$. □

Remark (Interpretation of the replay-onset bounds)**.** *Proposition 1 establishes the scale of the replay transition, not a sharp threshold. The necessity and sufficiency statements involve different summaries of the pairwise concentrations, $\bar{\kappa}$ and $\kappa_{\min}$, and the constant 32 in the sufficient condition comes from a Gaussian tail bound followed by a union bound. When pairwise separations themselves concentrate so that $\kappa_{\min}$ and $\bar{\kappa}$ are comparable, both statements place the transition at constant multiples of $\log M$. We therefore use $\Theta(\log M)$ only as an onset scale, rather than identifying a universal critical constant.*

A.1.4 Proof for Eq. (2)

Proof. Let $m \coloneqq \mathbb{E}[\mu_t(W)]$. For i.i.d. $\mu_t(W)$,$\mu_t(W')$

$$\mathbb{E}\|\mu_t(W) - \mu_t(W')\|^2 = \mathbb{E}\|(\mu_t(W) - m) - (\mu_t(W') - m)\|^2$$

$$= \mathbb{E}\|\mu_t(W) - m\|^2 + \mathbb{E}\|\mu_t(W') - m\|^2 - 2\langle \mathbb{E}[\mu_t(W) - m], \mathbb{E}[\mu_t(W') - m]\rangle$$

with the cross term factorising by independence and vanishing because each term is zero. Hence

$$\mathbb{E}\|\mu_t(W) - \mu_t(W')\|^2 = 2\mathbb{E}\|\mu_t(W) - m\|^2 = 2\,\mathrm{tr}\,\mathrm{Cov}(\mu_t(W))$$

□

A.1.5 PROOF OF THEOREM 2

The qualitative statement of Theorem 2 uses only the concentration assumption Eq. (3). For an explicit rate we additionally assume a sub-Gaussian lower tail: there is $c > 0$, independent of d, such that for every $j \neq i$ and every $0 \leq u \leq \sqrt{d}\, V_k(t)$,

$$\mathbb{P}\Big(\|\Delta_{ij}\|^2 \leq dV_k(t) - u\sqrt{d} \Big) \leq e^{-cu^2}. \tag{6}$$

This holds, for example, when $\|\Delta_{ij}\|^2$ is a sum of d independent nonnegative coordinate contributions with second moments bounded by B, with $c = 1/(2B)$ (Maurer, 2003).

Proof. The noise ε is independent of the stored windows, and $\sigma_t^2 \leq \bar{\sigma}^2$ for all d.

Define

$$A_d := \Big\{ \min_{j \neq i} \|\Delta_{ij}\|^2 \geq \tfrac{1}{2}\, dV_k(t) \Big\}, \qquad Y_j := \frac{\|\Delta_{ij}\|^2 - dV_k(t)}{\sqrt{d}}.$$

By Eq. (3), each Y_j is $\mathrm{O}_p(1)$. Since M is fixed, $\max_{j \neq i} |Y_j| = \mathrm{O}_p(1)$. As $V_k(t) \geq V_{\min}$,

$$\mathbb{P}(A_d^c) \leq \mathbb{P}\Big(\max_{j \neq i} |Y_j| > \tfrac{1}{2}\sqrt{d}\, V_{\min} \Big) \longrightarrow 0.$$

Under Eq. (6), take $u = \frac{1}{2}\sqrt{d}\, V_{\min}$. Since $dV_k(t) - \frac{1}{2} dV_{\min} \geq \frac{1}{2} dV_k(t)$, a union bound gives

$$\mathbb{P}(A_d^c) \leq (M-1) \exp\big(-\tfrac{c}{4} V_{\min}^2\, d \big).$$

On A_d,

$$\kappa_{\min} \geq \frac{dV_k(t)}{2\sigma_t^2} \geq \frac{dV_{\min}}{2\bar{\sigma}^2} =: \kappa_d.$$

Conditionally on the stored windows, Proposition 6 applies. Hence, on A_d and with conditional probability at least $1 - (M-1)e^{-\kappa_d/32}$,

$$S < s_d := (M-1)e^{-\kappa_d/4}, \qquad \mathrm{ESS} \leq (1+s_d)^2, \qquad 1 - \alpha_i = \frac{S}{1+S} \leq s_d.$$

Let B_d be the event on which the three bounds as earlier hold. Then

$$\mathbb{P}(B_d^c) \leq \mathbb{P}(A_d^c) + (M-1)e^{-\kappa_d/32} \longrightarrow 0.$$

On B_d, $1 \leq \mathrm{ESS} \leq (1+s_d)^2$ and $1 - s_d \leq \alpha_i \leq 1$. Since $s_d \to 0$, $\mathrm{ESS} \xrightarrow{\mathbb{P}} 1$ and $\alpha_i \xrightarrow{\mathbb{P}} 1$.

Rate. Under Eq. (6), with probability at least

$$1 - (M-1)\Big[\exp\big(-\tfrac{c}{4} V_{\min}^2 d \big) + \exp\Big(-\frac{V_{\min} d}{64\bar{\sigma}^2} \Big) \Big],$$

we have

$$0 \leq \mathrm{ESS} - 1 \leq 2s_d + s_d^2, \qquad 0 \leq 1 - \alpha_i \leq s_d, \qquad s_d = (M-1)\exp\Big(-\frac{V_{\min} d}{8\bar{\sigma}^2} \Big).$$

□

Remark (When the high-dimensional assumptions hold)**.** *The assumption*

$$\|\Delta_{ij}\|^2 = dV_{k,d}(t) + O_p(\sqrt{d})$$

requires the squared coordinate contributions to have total variance of order d. A sufficient condition is uniformly bounded fourth moments together with sufficiently weak cross-coordinate dependence, for example when

$$\sup_r \sum_s \big| \mathrm{Cov}(\Delta_{ij,r}^2, \Delta_{ij,s}^2) \big| < \infty.$$

This allows dependent coordinates, but excludes the case in which an $O(d)$ collection of coordinates fluctuates coherently, for which the variance of $\|\Delta_{ij}\|^2$ can instead be $O(d^2)$.

The independent windows W, W' used to define $V_{k,d}(t)$ specify a population separation scale. Rolling windows in the stored bank may overlap and need not be independent; the theorem therefore assumes concentration of the relevant bank separations around this scale rather than deriving it from temporal independence. Figure 2(b) checks this assumption empirically for the systems used in our experiments.

A.1.6 COMPONENT-DEPENDENT VARIANCES

Proposition 7 (Log-weight gap with component-dependent variances)**.** *Fix t and let $p_t(\cdot \mid Z_l) = \mathcal{N}(\mu_t^{(l)}, \sigma_l^2 I_d)$ with $\sigma_l > 0$ allowed to differ across components, under equal prior weights. Let $x = \mu_t^{(i)} + \sigma_i \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I_d)$, $\Delta_{ij} = \mu_t^{(i)} - \mu_t^{(j)} \neq 0$, $\kappa_{ij} = \|\Delta_{ij}\|^2/\sigma_j^2$ and $\rho = \sigma_i/\sigma_j$. Conditionally on the path centres,*

$$G_{ij} = \mathrm{KL}(p_i \,\|\, p_j) + \rho\sqrt{\kappa_{ij}}\,\xi + \tfrac{\rho^2-1}{2}(\xi^2 - 1) + \tfrac{\rho^2-1}{2}\big(Q - (d-1)\big),$$

where $\xi \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2_{d-1}$ are independent, and

$$\mathrm{KL}(p_i \,\|\, p_j) = \frac{\kappa_{ij}}{2} + d\,h(\rho), \qquad h(\rho) = \tfrac{1}{2}\big(\rho^2 - 1 - \log\rho^2\big) \geq 0,$$

with $h(\rho) = 0$ iff $\rho = 1$. Moreover $\mathrm{Var}(G_{ij}) = \rho^2\kappa_{ij} + \frac{d}{2}(\rho^2-1)^2$. For $\rho = 1$ this reduces exactly to Theorem 1.

Proof. The normalisers no longer cancel: $G_{ij} = d\log(\sigma_j/\sigma_i) - \frac{\|x-\mu_t^{(i)}\|^2}{2\sigma_i^2} + \frac{\|x-\mu_t^{(j)}\|^2}{2\sigma_j^2}$. Substituting $x - \mu_t^{(i)} = \sigma_i\varepsilon$ and $x - \mu_t^{(j)} = \Delta_{ij} + \sigma_i\varepsilon$,

$$G_{ij} = \frac{\|\Delta_{ij}\|^2}{2\sigma_j^2} + \frac{\sigma_i}{\sigma_j^2}\langle\Delta_{ij}, \varepsilon\rangle + \frac{\rho^2-1}{2}\|\varepsilon\|^2 - d\log\rho.$$

Write $\xi = \langle\Delta_{ij}, \varepsilon\rangle/\|\Delta_{ij}\|$ and decompose $\|\varepsilon\|^2 = \xi^2 + Q$, where Q is the squared norm of the component of ε orthogonal to Δ_{ij}; then $\xi \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2_{d-1}$ are independent. Since $\frac{\sigma_i}{\sigma_j^2}\|\Delta_{ij}\| = \rho\sqrt{\kappa_{ij}}$, collecting the constant terms gives the representation, and the constant equals $\mathrm{KL}(p_i\|p_j)$ by the closed form for isotropic Gaussians. Nonnegativity of h follows from $u - 1 - \log u \geq 0$ with $u = \rho^2$. For the variance, ξ and ξ^2 are uncorrelated, $\mathrm{Var}(\xi^2) = 2$ and $\mathrm{Var}(Q) = 2(d-1)$. □

Corollary 1 (High-dimensional replay, component-dependent variances)**.** *Let M be fixed and suppose the variance ratios ρ_{ij} lie in a compact subset of $(0,\infty)$ uniformly in d. If for every $j \neq i$ either $\kappa_{ij} \geq c\,d$ with probability tending to one, or $|\log\rho_{ij}| \geq c$, for some $c > 0$ independent of d, then $\alpha_i(t,x) \xrightarrow{P} 1$ and $\mathrm{ESS}(t,x) \xrightarrow{P} 1$ as $d \to \infty$.*

Proof. Under either condition $\mathbb{E}[G_{ij}] \geq c'd$ for some $c' > 0$, while $\mathrm{Var}(G_{ij}) = O(\kappa_{ij} + d)$. By Chebyshev, $P\big(G_{ij} \leq \frac{1}{2}\mathbb{E}[G_{ij}]\big) \leq 4\,\mathrm{Var}(G_{ij})/\mathbb{E}[G_{ij}]^2 \to 0$, so $G_{ij} \to \infty$ in probability for each of the finitely many $j \neq i$. Hence $S = \sum_{j\neq i} e^{-G_{ij}} \to 0$ in probability, and $\alpha_i = 1/(1+S) \to 1$ and $\mathrm{ESS} \leq (1+S)^2 \to 1$ as in Proposition 6 □

A.2 PROOFS FROM SECTION 4

A.2.1 PROOF OF PROPOSITION 2

Proof. Under the shared isotropic Gaussian conditional path,

$$p_t(x \mid W_j) = (2\pi\sigma_t^2)^{-d/2}\exp\left(-\frac{\|x-\mu_t^{(j)}\|^2}{2\sigma_t^2}\right).$$

The Gaussian prefactor is independent of j and therefore cancels in the normalised posterior weights:

$$\alpha_j(t,x) = \frac{\exp\left(-\|x-\mu_t^{(j)}\|^2/(2\sigma_t^2)\right)}{\sum_l \exp\left(-\|x-\mu_t^{(l)}\|^2/(2\sigma_t^2)\right)}.$$

Thus the interpolant and the window order enter the weights only through the locations $\{\mu_t^{(j)}\}$.

Taking

$$\Delta_{ij} = \mu_t^{(i)} - \mu_t^{(j)},$$

the calculation of Theorem 1 applies unchanged and gives

$$G_j \sim \mathcal{N}\left(\frac{\kappa_{ij}}{2}, \kappa_{ij}\right), \qquad \kappa_{ij} = \frac{\|\Delta_{ij}\|^2}{\sigma_t^2}.$$

For the scaling statement, use the distance-concentration assumption

$$\|\Delta_{ij}\|^2 = dV_k(t) + \mathrm{O}_p(\sqrt{d}).$$

Then

$$\kappa_{ij} = \frac{dV_k(t)}{\sigma_t^2} + \mathrm{O}_p\left(\frac{\sqrt{d}}{\sigma_t^2}\right).$$

When the window order k is fixed and

$$V_k(t) = \Theta(1),$$

changing the window order or interpolant may change the value of $V_k(t)$, but does not change the leading d/σ_t^2 dimensional scaling. □

A.2.2 Proof of Proposition 3

Proof. For the shared isotropic Gaussian conditional path,

$$p_t(x \mid W_j) = \mathcal{N}(\mu_t^{(j)}, \sigma_t^2 I_d),$$

the conditional score is

$$\nabla_x \log p_t(x \mid W_j) = \frac{\mu_t^{(j)} - x}{\sigma_t^2}.$$

For fixed x, the empirical squared-loss objective is a weighted least-squares problem over the conditional targets

$$\frac{\mu_t^{(j)} - x}{\sigma_t^2},$$

with weights proportional to the conditional densities $p_t(x \mid W_j)$. By the same conditional-regression argument as Lemma 1, its minimiser is therefore

$$s_t^\star(x) = \sum_j \alpha_j(t,x) \frac{\mu_t^{(j)} - x}{\sigma_t^2},$$

where

$$\alpha_j(t,x) = \frac{p_t(x \mid W_j)}{\sum_l p_t(x \mid W_l)}.$$

These are exactly the posterior responsibilities appearing in the empirical velocity minimiser. □

A.2.3 Proof of Theorem 3

Proof. From Eq. (4),

$$\kappa_{ij} = \frac{dV_{k,d}(t)}{\sigma_t^2}\left(1 + O_p(d^{-1/2})\right).$$

Since $V_{k,d}(t) = \Theta(1)$ and $\sigma_t^2 \asymp d^a$,

$$\kappa_{ij} \asymp d^{1-a}\left(1 + O_p(d^{-1/2})\right).$$

If $a < 1$, then $\kappa_{ij} \to \infty$ in probability. By Theorem 1,

$$G_{ij} = \frac{\kappa_{ij}}{2} + \sqrt{\kappa_{ij}}\, Z_{ij}, \qquad Z_{ij} \sim \mathcal{N}(0,1),$$

and hence

$$\frac{G_{ij}}{\kappa_{ij}} = \frac{1}{2} + \frac{Z_{ij}}{\sqrt{\kappa_{ij}}} \xrightarrow{\mathbb{P}} \frac{1}{2}.$$

Thus $G_{ij} \to \infty$ in probability for every $j \neq i$. Since M is fixed,

$$S = \sum_{j \neq i} e^{-G_{ij}} \xrightarrow{\mathbb{P}} 0,$$

which gives

$$\alpha_i = \frac{1}{1+S} \xrightarrow{\mathbb{P}} 1, \qquad \mathrm{ESS} \xrightarrow{\mathbb{P}} 1.$$

If $a = 1$, then κ_{ij} remains on an order-one scale. Therefore the mean and variance of each log-weight gap, $\kappa_{ij}/2$ and κ_{ij}, remain order one, so increasing dimension alone drives neither one-hot concentration nor uniform averaging.

If $a > 1$, then $\kappa_{ij} \to 0$ in probability. Theorem 1 therefore gives

$$G_{ij} \xrightarrow{\mathbb{P}} 0$$

for every $j \neq i$, and hence

$$\frac{\alpha_j}{\alpha_i} = e^{-G_{ij}} \xrightarrow{\mathbb{P}} 1.$$

Since M is fixed and the posterior weights sum to one,

$$\alpha_j \xrightarrow{\mathbb{P}} \frac{1}{M}$$

for every j, and consequently

$$\mathrm{ESS} = \left(\sum_j \alpha_j^2 \right)^{-1} \xrightarrow{\mathbb{P}} M.$$

□

Remark (Scope of the bandwidth regimes)**.** *The three regimes are taken with fixed memory-bank size M, fixed window order k and fixed flow time t, while the ambient dimension increases. The assumption $V_{k,d}(t) = \Theta(1)$ rules out a simultaneous collapse of the interpolated window cloud itself. The result should therefore not be read as a statement about arbitrary joint growth of M, k and d; joint M–d asymptotics would additionally require controlling the extremes of the pairwise separation distribution.*

A.2.4 Proof of Proposition 4

Proof. The anisotropic Gaussian density is

$$p_t(x \mid W_j) = (2\pi)^{-d/2} |\Sigma_t|^{-1/2} \exp\left(-\frac{1}{2} (x - \mu_t^{(j)})^\top \Sigma_t^{-1} (x - \mu_t^{(j)}) \right).$$

Since Σ_t is shared across windows, the prefactor is independent of j and cancels in the posterior-weight ratio. Hence

$$G_j = \frac{1}{2} \left[(x - \mu_t^{(j)})^\top \Sigma_t^{-1} (x - \mu_t^{(j)}) - (x - \mu_t^{(i)})^\top \Sigma_t^{-1} (x - \mu_t^{(i)}) \right].$$

Write

$$x = \mu_t^{(i)} + A\varepsilon, \qquad AA^\top = \Sigma_t.$$

Then

$$x - \mu_t^{(j)} = \Delta_{ij} + A\varepsilon, \qquad x - \mu_t^{(i)} = A\varepsilon.$$

Expanding,

$$(x - \mu_t^{(j)})^\top \Sigma_t^{-1} (x - \mu_t^{(j)}) = \Delta_{ij}^\top \Sigma_t^{-1} \Delta_{ij} + 2\Delta_{ij}^\top \Sigma_t^{-1} A\varepsilon + \varepsilon^\top A^\top \Sigma_t^{-1} A\varepsilon,$$

whereas

$$(x-\mu_t^{(i)})^\top \Sigma_t^{-1}(x-\mu_t^{(i)}) = \varepsilon^\top A^\top \Sigma_t^{-1} A\varepsilon.$$

The quadratic noise terms cancel exactly, giving

$$G_j = \frac{1}{2}\kappa_{ij}^{\Sigma} + \Delta_{ij}^\top \Sigma_t^{-1} A\varepsilon.$$

Conditionally on the interpolated window locations, the stochastic term is a linear functional of ε. Its variance is

$$\Delta_{ij}^\top \Sigma_t^{-1} A A^\top \Sigma_t^{-1} \Delta_{ij} = \Delta_{ij}^\top \Sigma_t^{-1} \Delta_{ij} = \kappa_{ij}^{\Sigma}.$$

Therefore

$$\Delta_{ij}^\top \Sigma_t^{-1} A\varepsilon \sim \mathcal{N}(0, \kappa_{ij}^{\Sigma}),$$

and hence

$$G_j \sim \mathcal{N}\left(\frac{\kappa_{ij}^{\Sigma}}{2}, \kappa_{ij}^{\Sigma}\right).$$

□

A.2.5 Proof of Proposition 5

Proof. For a shared covariance Σ_t,

$$\mathbb{E}[\kappa_{ij}^{\Sigma}] = \mathbb{E}[\Delta_{ij}^\top \Sigma_t^{-1} \Delta_{ij}] = \operatorname{tr}(\Sigma_t^{-1}\Gamma).$$

We minimise this subject to

$$\operatorname{tr}(\Sigma_t) = d\sigma_t^2.$$

The Lagrangian is

$$\mathcal{L}(\Sigma_t, \eta) = \operatorname{tr}(\Sigma_t^{-1}\Gamma) + \eta\left(\operatorname{tr}(\Sigma_t) - d\sigma_t^2\right).$$

Using

$$\frac{\partial}{\partial \Sigma_t}\operatorname{tr}(\Sigma_t^{-1}\Gamma) = -\Sigma_t^{-1}\Gamma\Sigma_t^{-1},$$

stationarity gives

$$-\Sigma_t^{-1}\Gamma\Sigma_t^{-1} + \eta I = 0.$$

Multiplying on the left and right by Σ_t,

$$\Gamma = \eta\Sigma_t^2.$$

Hence

$$\Sigma_t = \eta^{-1/2}\Gamma^{1/2}.$$

The trace constraint gives

$$\eta^{-1/2} = \frac{d\sigma_t^2}{\operatorname{tr}(\Gamma^{1/2})},$$

and therefore

$$\Sigma_t^\star = \frac{d\sigma_t^2}{\operatorname{tr}(\Gamma^{1/2})}\Gamma^{1/2}.$$

Substitution gives

$$\operatorname{tr}((\Sigma_t^\star)^{-1}\Gamma) = \frac{\left(\operatorname{tr}\Gamma^{1/2}\right)^2}{d\sigma_t^2}.$$

Since

$$r_{\text{eff}}(\Gamma) = \frac{(\operatorname{tr}\Gamma^{1/2})^2}{\operatorname{tr}\Gamma}$$

and

$$\operatorname{tr}\Gamma = dV_k(t),$$

we obtain

$$\mathbb{E}[\kappa_{ij}^{\Sigma^\star}] = \frac{r_{\text{eff}}(\Gamma)V_k(t)}{\sigma_t^2}.$$

The map

$$\Sigma_t \mapsto \operatorname{tr}(\Sigma_t^{-1}\Gamma)$$

is convex on the positive-definite cone, so the stationary point is the global minimum. □

Remark (Covariance optimisation assumptions)**.** *The constraint* $\operatorname{tr}(\Sigma_t) = d\sigma_t^2$ *fixes the total conditional variance, so the optimisation changes covariance shape rather than simply increasing the bandwidth. The assumption* $\Gamma \succ 0$ *ensures that the closed-form optimizer is strictly positive definite. If* Γ *is singular, the same calculation applies on its support; a strictly positive covariance on the full ambient space then requires a variance floor in the null directions. Thus the proposition is best interpreted as a statement about how a fixed variance budget should be allocated across directions in which the interpolated window cloud varies.*

A.2.6 Proof of the elliptic scale-mixture claim

Proposition 8 (Scale-mixture paths)**.** *Let* P_ω *be a probability law on* $(0,\infty)$ *and let every window share the conditional path*

$$p_t(x \mid W) = \int_0^\infty \mathcal{N}\big(x;\, \mu_t(W),\, \Sigma_t/\omega\big)\, P_\omega(d\omega), \qquad \Sigma_t \succ 0.$$

Generate the query as $x = \mu_t^{(i)} + \omega_0^{-1/2} A\varepsilon$ *with* $AA^\top = \Sigma_t$, $\varepsilon \sim \mathcal{N}(0, I_d)$ *and* $\omega_0 \sim P_\omega$, *independent of each other and of the stored windows. Let* $\alpha_j \propto p_t(x \mid W_j)$, $G_j = \log\alpha_i - \log\alpha_j$ *and* $\kappa_{ij}^\Sigma = \Delta_{ij}^\top \Sigma_t^{-1} \Delta_{ij}$. *Conditionally on the interpolated window locations:*

(i) $\mathbb{P}(G_j < 0 \mid \omega_0) = \Phi\big(-\tfrac{1}{2}\sqrt{\omega_0 \kappa_{ij}^\Sigma}\big)$;

(ii) $\mathbb{P}(G_j < 0) \le \mathbb{E}\big[e^{-\omega_0 \kappa_{ij}^\Sigma/8}\big]$. *For the multivariate* t_ν *path,* $\omega_0 \sim \mathrm{Gamma}(\nu/2, \nu/2)$, *this bound equals* $\big(1 + \kappa_{ij}^\Sigma/(4\nu)\big)^{-\nu/2}$, *and* $\mathbb{P}(G_j < 0) = \Theta\big((\kappa_{ij}^\Sigma)^{-\nu/2}\big)$ *as* $\kappa_{ij}^\Sigma \to \infty$;

(iii) for the t_ν *path with* M *and* ν *fixed, if* $\min_{j \ne i} \kappa_{ij}^\Sigma \ge rd$ *for some* $r > 0$ *independent of* d, *then* $\alpha_i \xrightarrow{\mathbb{P}} 1$ *and* $\mathrm{ESS} \xrightarrow{\mathbb{P}} 1$ *as* $d \to \infty$.

Proof. Write $Q_l := (x - \mu_t^{(l)})^\top \Sigma_t^{-1} (x - \mu_t^{(l)})$. Because Σ_t is shared, $p_t(x \mid W_l) = h(Q_l)$ with

$$h(q) = (2\pi)^{-d/2} \det(\Sigma_t)^{-1/2} \int_0^\infty \omega^{d/2} e^{-\omega q/2}\, P_\omega(d\omega),$$

which is strictly decreasing in q. Hence $G_j = \log h(Q_i) - \log h(Q_j)$, and $G_j < 0$ if and only if $Q_j < Q_i$. The weights depend on x only through the marginal density h, not on the realised ω_0.

Substituting the query,

$$Q_i = \frac{\|\varepsilon\|^2}{\omega_0}, \qquad Q_j - Q_i = \kappa_{ij}^\Sigma + 2\omega_0^{-1/2}\sqrt{\kappa_{ij}^\Sigma}\,\eta_j, \qquad \eta_j := \frac{\Delta_{ij}^\top \Sigma_t^{-1} A\varepsilon}{\sqrt{\kappa_{ij}^\Sigma}} \sim \mathcal{N}(0,1),$$

using $\mathrm{Var}(\Delta_{ij}^\top \Sigma_t^{-1} A\varepsilon) = \kappa_{ij}^\Sigma$ as in Proposition 4.

(i). $\mathbb{P}(Q_j < Q_i \mid \omega_0) = \mathbb{P}\big(\eta_j < -\tfrac{1}{2}\sqrt{\omega_0 \kappa_{ij}^\Sigma}\big) = \Phi\big(-\tfrac{1}{2}\sqrt{\omega_0 \kappa_{ij}^\Sigma}\big)$.

(ii). From $\Phi(-z) \le e^{-z^2/2}$, $\mathbb{P}(G_j < 0 \mid \omega_0) \le e^{-\omega_0 \kappa_{ij}^\Sigma/8}$; take expectations over ω_0. For $\omega_0 \sim \mathrm{Gamma}(\nu/2, \nu/2)$, $\mathbb{E}[e^{-s\omega_0}] = (1 + 2s/\nu)^{-\nu/2}$; set $s = \kappa_{ij}^\Sigma/8$. For the matching lower bound, write $\kappa = \kappa_{ij}^\Sigma \ge 1$. If $\omega_0 \le 1/\kappa$ then $\Phi(-\tfrac{1}{2}\sqrt{\omega_0\kappa}) \ge \Phi(-\tfrac{1}{2})$. The Gamma density is at least $c_\nu\, \omega^{\nu/2-1}$ on $(0,1]$ with $c_\nu = (\nu/2)^{\nu/2} e^{-\nu/2}/\Gamma(\nu/2)$, so

$$\mathbb{P}(G_j < 0) \ge \Phi(-\tfrac{1}{2})\, \mathbb{P}(\omega_0 \le 1/\kappa) \ge \Phi(-\tfrac{1}{2})\, \frac{2c_\nu}{\nu}\, \kappa^{-\nu/2}.$$

(iii). For the t_ν path, $h(q) \propto (1 + q/\nu)^{-(\nu+d)/2}$, so

$$G_j = \frac{\nu + d}{2} \log\Big(1 + \frac{Q_j - Q_i}{\nu + Q_i}\Big).$$

Fix $\omega_0 > 0$ and let

$$E \coloneqq \Big\{\eta_j \geq -\tfrac{1}{4}\sqrt{\omega_0 \kappa_{ij}^{\Sigma}}\ \ \forall j \neq i\Big\} \cap \big\{\|\varepsilon\|^2 \leq 2d\big\}.$$

By $\Phi(-z) \leq e^{-z^2/2}$, a union bound, and the chi-square tail $\mathbb{P}(\|\varepsilon\|^2 \geq d + 2\sqrt{dx} + 2x) \leq e^{-x}$ at $x = d/8$ (Laurent & Massart, 2000),

$$\mathbb{P}(E^c \mid \omega_0) \leq (M-1)e^{-\omega_0 r d/32} + e^{-d/8}.$$

On E, $Q_j - Q_i \geq \kappa_{ij}^{\Sigma}/2 \geq rd/2$ and $Q_i \leq 2d/\omega_0$. For $d \geq \nu\omega_0$,

$$\frac{Q_j - Q_i}{\nu + Q_i} \geq \frac{r\omega_0 d}{2(\nu\omega_0 + 2d)} \geq \frac{r\omega_0}{6}, \qquad \text{so} \qquad G_j \geq \frac{d}{2}\log\Big(1 + \frac{r\omega_0}{6}\Big).$$

Therefore, on E, $S = \sum_{j \neq i} e^{-G_j} \leq (M-1)(1 + r\omega_0/6)^{-d/2} \to 0$, giving $\alpha_i = 1/(1+S) \to 1$ and $\mathrm{ESS} \leq (1+S)^2 \to 1$ with conditional probability tending to one for every fixed $\omega_0 > 0$. Since $\omega_0 > 0$ almost surely, dominated convergence over ω_0 gives the unconditional statement. □

Remark. *Taking $P_\omega = \delta_1$ in (i) recovers the Gaussian misranking probability $\Phi(-\frac{1}{2}\sqrt{\kappa_{ij}})$ of Theorem 1. Heavy tails therefore do not change the direction of the high-dimensional limit; they convert exponential misranking decay into polynomial decay. Under $\kappa_{ij}^{\Sigma} = \Theta(d)$, a fraction $\Theta(d^{-\nu/2})$ of queries avoids concentration on the generating window.*

A.3 Proofs from Section 5

A.3.1 Proof of Theorem 4

For this result we require the shared covariance to be constant in flow time, so that the conditional velocity targets are

$$u_t^{(j)} = \dot{\mu}_t(W_j)$$

and contain no covariance-derivative term. In the isotropic case this is the condition $\sigma_t' = 0$; for a general shared covariance it corresponds to $\dot{\Sigma}_t = 0$.

Proof. By Lemma 1,

$$v_t^\star(x) = \sum_j \alpha_j(t,x) u_t^{(j)},$$

where

$$\alpha_j(t,x) = \frac{p_t(x \mid W_j)}{\sum_l p_t(x \mid W_l)}.$$

Since $\Sigma_t \succ 0$, every Gaussian density is strictly positive, so

$$\alpha_j(t,x) \geq 0, \qquad \sum_j \alpha_j(t,x) = 1.$$

Hence

$$v_t^\star(x) \in \mathrm{conv}\{u_t^{(1)}, \ldots, u_t^{(M)}\} = \mathcal{U}_t,$$

proving convex-hull inclusion.

By definition,

$$D_t(x) = \min_{w \in \mathcal{U}_t} \|f_t(x) - w\|.$$

Since

$$v_t^\star(x) \in \mathcal{U}_t,$$

it follows immediately that

$$\|v_t^\star(x) - f_t(x)\| \geq D_t(x),$$

proving the pointwise lower bound.

For the infimum result, the map

$$\alpha \mapsto \sum_j \alpha_j u_t^{(j)}$$

maps the simplex

$$\mathcal{S}_{M-1} = \left\{ \alpha \in \mathbb{R}^M_{\geq 0} : \sum_j \alpha_j = 1 \right\}$$

onto the convex hull $\mathcal{U}_t$. Therefore

$$\left\{ \sum_j \alpha_j u_t^{(j)} : \alpha \in \mathcal{S}_{M-1} \right\} = \mathcal{U}_t,$$

and hence

$$\inf_{\alpha \in \mathcal{S}_{M-1}} \left\| f_t(x) - \sum_j \alpha_j u_t^{(j)} \right\| = \min_{w \in \mathcal{U}_t} \|f_t(x) - w\| = D_t(x).$$

Additionally, under the constant-in-flow-time covariance assumption,

$$u_t^{(j)} = \dot{\mu}_t(W_j).$$

Thus the target cloud

$$\mathcal{U}_t = \mathrm{conv}\{u_t^{(1)}, \ldots, u_t^{(M)}\}$$

depends on the memory bank and interpolant but not on the shared covariance. Changing Σ_t changes the posterior weights but not $\mathcal{U}_t$, and therefore does not change $D_t(x)$.

Finally, fix a query law Q independently of the covariance being varied. From the pointwise lower bound,

$$\|v_t^\star(x) - f_t(x)\|^2 \geq D_t(x)^2.$$

Taking expectation under Q gives

$$\mathbb{E}_{x \sim Q} \|v_t^\star(x) - f_t(x)\|^2 \geq \mathbb{E}_{x \sim Q} D_t(x)^2.$$

Since the right-hand side is independent of Σ_t,

$$\inf_{\Sigma_t \succ 0} \mathbb{E}_{x \sim Q} \|v_t^\star(x) - f_t(x)\|^2 \geq \mathbb{E}_{x \sim Q} D_t(x)^2.$$

□

Remark (Why constant covariance is required)**.** *The constant-in-flow-time covariance assumption is essential for the fixed convex-hull statement. When $\dot{\Sigma}_t = 0$, the conditional targets are $u_t^{(j)} = \dot{\mu}_t(W_j)$ and changing the shared covariance changes only the posterior weights, not the set of available targets. If $\dot{\Sigma}_t \neq 0$, the conditional velocity contains an additional covariance-derivative term depending on both x and the path covariance, so changing the covariance also changes the targets themselves. In that case the fixed target cloud $\mathcal{U}_t$ used in Theorem 4 is no longer invariant to covariance choice.*

A.3.2 Target Span Deficiency and Proof

Corollary 2 (Target-span deficiency)**.** *Let $T_t := \mathrm{span}\{u_t^{(j)} - u_t^{(1)} : j = 2, \ldots, M\}$ and let $P^\perp$ denote orthogonal projection onto $T_t^\perp$. Then for every x,*

$$D_t(x) \;\geq\; \left\| P^\perp \big(f_t(x) - u_t^{(1)} \big) \right\|.$$

In particular, if $\dim T_t < d$ and $f_t(x) - u_t^{(1)}$ has a nonzero component orthogonal to T_t, the floor is strictly positive regardless of the bandwidth, shared covariance, or choice of convex weights.

Proof. For any

$$\alpha \in \mathcal{S}_{M-1},$$

we have

$$\sum_j \alpha_j u_t^{(j)} - u_t^{(1)} = \sum_j \alpha_j \left(u_t^{(j)} - u_t^{(1)} \right) \in T_t,$$

using

$$\sum_j \alpha_j = 1.$$

Therefore

$$P^\perp\left(\sum_j \alpha_j u_t^{(j)} - u_t^{(1)}\right) = 0.$$

It follows that

$$P^\perp\left(f_t(x) - \sum_j \alpha_j u_t^{(j)}\right) = P^\perp\left(f_t(x) - u_t^{(1)}\right).$$

Since $P^\perp$ is an orthogonal projection,

$$\|P^\perp z\| \le \|z\|$$

for every $z \in \mathbb{R}^d$. Thus

$$\left\| f_t(x) - \sum_j \alpha_j u_t^{(j)} \right\| \ge \left\| P^\perp\left(f_t(x) - u_t^{(1)}\right)\right\|$$

for every

$$\alpha \in \mathcal{S}_{M-1}.$$

Taking the infimum over α and taking the infimum over all convex weights as in Theorem 4 gives

$$D_t(x) \ge \left\| P^\perp\left(f_t(x) - u_t^{(1)}\right)\right\|.$$

□

A.4 Deterministic-interpolant limit and proof

Corollary 3 (Deterministic-interpolant limit)**.** *Fix $t \in (0,1)$ and a shared isotropic path $p_t(\cdot \mid Z) = \mathcal{N}(\mu_t(Z), \sigma^2 I_d)$ with σ constant in flow time, and let the empirical coupling place weights $\pi_j > 0$ on conditioning variables Z_j. For every x with a unique nearest centre $i(x) = \arg\min_j \|x - \mu_t^{(j)}\|$, which holds for Lebesgue-almost every x,*

$$\lim_{\sigma\downarrow 0} \alpha_j(t,x) = \mathbb{1}\{j = i(x)\}, \qquad \lim_{\sigma\downarrow 0} v_t^\star(x) = u_t^{(i(x))}.$$

Consequently the limiting field equals a single stored target, and

$$\left\|v_t^\star(x) - f_t(x)\right\| \;\to\; \left\|u_t^{(i(x))} - f_t(x)\right\| \;\ge\; \min_j \left\|u_t^{(j)} - f_t(x)\right\| \;\ge\; D_t(x).$$

Proof of Corollary 3. With weights π_j, Lemma 1 generalises to $\alpha_j \propto \pi_j\, p_t(x \mid Z_j)$. For $j \ne i := i(x)$,

$$\log\alpha_j - \log\alpha_i = \log\frac{\pi_j}{\pi_i} - \frac{\|x-\mu_t^{(j)}\|^2 - \|x-\mu_t^{(i)}\|^2}{2\sigma^2}.$$

The first term is finite and independent of σ; the second tends to $-\infty$ because the bracket is strictly positive by uniqueness of the nearest centre. Hence $\alpha_j \to 0$ for $j \ne i$ and $\alpha_i \to 1$. Since σ is constant in flow time, $u_t^{(j)} = \dot\mu_t(Z_j)$ is independent of x and σ, so $v_t^\star(x) = \sum_j \alpha_j u_t^{(j)} \to u_t^{(i)}$. The first inequality is immediate; the second holds because every stored target lies in $\mathcal{U}_t$. Points with a non-unique nearest centre lie on finitely many hyperplanes and have Lebesgue measure zero.

Some multi-marginal methods train the vector field on deterministic interpolants $\sigma_t \equiv 0$, particularly Kviman et al. (2026) whose learned interpolant is regressed without added noise in its released implementation. For this case, the conditional paths have no density and Lemma 1 does not apply directly: the population minimiser is determined only on the finite set $\{\mu_t^{(j)}\}$ and is arbitrary everywhere else. Corollary 3 identifies the natural limiting object, nearest-centre replay which is the regime these methods occupy.

B Appendix: Empirical Details

Common closed-form protocol. The experiments in Sections 6.1 to 6.3 evaluate the infinite-capacity empirical flow-matching estimator directly; no network is trained. Posterior weights and the corresponding velocity field are evaluated in closed form. For the dynamical systems, ground-truth targets are exact one-step displacements obtained by numerically integrating the underlying ODE.

Unless otherwise stated, every experimental cell contains 500 evaluation queries and is repeated over five seeds, $0, \dots, 4$. Random draws are deterministically keyed by system, seed and stream, so results are independent of execution order. Banks of different sizes are nested subsets of a common random permutation within each system and seed. Query families receive their own deterministic streams and are therefore identical when reused across bank sizes, concentration values or covariance geometries.

Within each seed we first take the median over evaluation queries. Across seeds we report the median of these seed-level values, with the seed minimum–maximum as the error bar. Paired quantities, including error ratios and differences, are formed within each seed before aggregation. All concentration grids and matching tolerances are specified in the experiment configuration before evaluation.

Dynamical systems. We use Lorenz–63 with the standard parameters $(\sigma, \rho, \beta) = (10, 28, 8/3)$, integrated with step size 0.005 and recorded every 10 integration steps, and Lorenz–96

$$\dot{x}_i = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F$$

with integration step 0.01 and observations every 5 integration steps. Thus all dynamical-system experiments use observation interval $\Delta\tau = 0.05$.

The standard Lorenz–96 systems use $F = 8$. To probe covariance geometries with different displacement spectra, the anisotropy experiment additionally uses a $d = 64$, $F = 5$ system and two $d = 64$, $F = 4$ systems. The latter are initialised in two different structured regimes: `F4-bump` starts from the constant forcing state with a 0.01 perturbation at one coordinate, whereas `F4-cos` starts from

$$x_i(0) = F + 0.5\cos(2\pi i/d).$$

These variants use the same Lorenz–96 equations and numerical integration; only the forcing and initial condition differ.

The replay-onset experiment additionally includes iid $\mathcal{N}(0, I_d)$ banks. These have no dynamics or prediction target and are used solely as a control whose pairwise-separation geometry is known independently of an attractor.

Replay onset and dimensional concentration. For Fig. 2, we test three predictions: whether weight behaviour collapses onto the concentration parameter $\widehat{\kappa} = d\widehat{V}_k/\sigma_t^2$ across dimension, window order and interpolant; whether pairwise separations concentrate at relative rate $d^{-1/2}$; and whether the concentration required for replay grows logarithmically with memory-bank size.

For each interpolated window cloud we compute

$$\widehat{V}_k(t) = \frac{2}{d}\,\mathrm{tr}\,\widehat{\mathrm{Cov}}(\mu_t(W)), \qquad \widehat{\kappa} = \frac{d\widehat{V}_k(t)}{\sigma_t^2},$$

and evaluate the exact posterior weights. Weight spread is summarized by

$$\mathrm{ESS} = \left(\sum_j \alpha_j^2\right)^{-1}.$$

The principal sweep uses Lorenz–96 dimensions $d \in \{8, 16, 32, 64, 128, 256\}$, window orders $k \in \{1, 2, 4\}$, and $M = 1000$ stored windows. We evaluate flow times $t \in \{0.5, 0.35\}$. The second value is included because $t = 0.5$ is an interpolation knot for the even window orders, at which

the linear and cubic interpolants coincide; $t = 0.35$ therefore provides a nontrivial comparison of interpolation rules. The concentration sweep contains 40 values over $\widehat{\kappa} \in [0.1, 1000]$.

The same measurements are repeated on iid Gaussian banks at $d \in \{8, 16, 32, 64, 128, 256, 512\}$. In Fig. 2(b) we report

$$\frac{\mathrm{sd}(\kappa_{ij})}{\mathrm{mean}(\kappa_{ij})}$$

and compare it with the $d^{-1/2}$ relative-concentration prediction.

For the memory-bank sweep, $M \in \{10, 30, 100, 300, 1000, 3000, 10000\}$ on `Gauss-512`, `L96-64-onset`, and `L96-256-onset`. Banks are nested across M. For each bank size, the empirical replay onset $\kappa^\star$ is defined as the concentration at which the median effective sample size reaches $\mathrm{ESS} = 2$.

Coverage and attainable prediction error. For Fig. 3, we vary the distance of a query from the stored state bank while independently sweeping kernel concentration. Let

$$d_{\mathrm{nn}}(x) = \min_j \|x - X_j\|$$

be the Euclidean distance to the nearest stored state. We normalize this by the typical bank separation,

$$\rho(x) = \frac{d_{\mathrm{nn}}(x)}{\sqrt{d\widehat{V}}},$$

and use ρ as an observable state-space coverage proxy.

Five query families are used. `heldout` consists of states sampled without replacement from an independently simulated test trajectory. `isotropic-noise` samples a stored centre and adds independent $\mathcal{N}(0, I_d)$ noise, corresponding to the Gaussian flow-matching training query law. The two off-support families start from a held-out state and add a perturbation in a uniformly random direction of norm

$$0.25\sqrt{d\widehat{V}} \quad \text{or} \quad 1.0\sqrt{d\widehat{V}},$$

respectively. Finally, `midpoint` takes the midpoint of two independently sampled stored centres. Each named family is generated once per system and seed and then reused across all bank sizes and concentration values.

The experiment uses Lorenz–63 with $M \in \{200, 2000, 20000\}$ and Lorenz–96 at $d \in \{16, 64\}$ with $M \in \{2000, 20000\}$. For each bank, concentration is swept over 20 values beginning at $\kappa = 0.3$ and extending to $64 \times 2 \log M$, so all memory-bank sizes are explored to the same depth relative to their predicted replay onset.

For each family we report e_{best}, the lowest normalized closed-form prediction error observed over this pre-specified concentration grid, and the nearest-neighbour error e_{NN}. We separately compute the target-space representation floor

$$D_t(x) = \mathrm{dist}(f_t(x), \mathcal{U}_t)\,,$$

by projecting the reference target onto the convex hull of stored targets using non-negative least squares with an explicit simplex constraint. Errors are normalized by the median magnitude of the true one-step field on held-out trajectory states.

Matched-concentration covariance experiment. For Fig. 4, we separate the effect of covariance shape from the effect of changing overall posterior concentration. For a shared covariance Σ, realised concentration is summarized by

$$\bar{\kappa}^{\Sigma} = \mathrm{tr}(\Sigma^{-1}\widehat{\Gamma}),$$

where $\widehat{\Gamma}$ is the empirical covariance of pairwise interpolated centre displacements.

All experiments use $M = 20000$ and the query families `heldout`, `off-0.25`, `off-1.0`, and `midpoint`. We evaluate Lorenz–63, standard Lorenz–96 at $d = 16$ and $d = 64$, and the three controlled $d = 64$ Lorenz–96 variants `F5`, `F4-bump`, and `F4-cos`. The concentration grid contains 12 operating points.

For Fig. 4(a), isotropic and anisotropic kernels are compared only when their realised concentrations agree within the pre-specified relative tolerance of 10%. The same stored bank and query family are used on both sides of every comparison, and the error ratio is formed within seed before aggregation. Any residual difference at matched $\bar{\kappa}^{\Sigma}$ therefore reflects covariance shape rather than merely a different degree of replay concentration.

In Fig. 4(b), we additionally compare the equal-trace concentration-optimal covariance

$$\Sigma \propto \widehat{\Gamma}^{1/2}$$

with an isotropic covariance under the same total variance budget. The isotropic bandwidth is then independently retuned to match the realised concentration of the anisotropic kernel. This separates the reduction in error obtained by moving along the concentration axis from any residual effect attributable to covariance geometry itself.

Finite-capacity experiment. For Fig. 5, we use Lorenz–96 with $d = 64$, $F = 8$, $\Delta\tau = 0.05$ and a fixed bank of $M = 2000$ pairwise transitions. We vary

$$\lambda = \frac{\widehat{\kappa}}{\log M} \in \{1, 4, 8, 16, 32\},$$

and train velocity MLPs of widths $W \in \{128, 512\}$ for $50{,}000$ steps over three seeds. The network input is $[x, \mathrm{emb}(t)]$, where $\mathrm{emb}(t)$ contains sine and cosine Fourier features with frequencies $1, \ldots, 16$; the network has three hidden SiLU layers and a linear d-dimensional output, with no normalization or residual layers.

Each training batch samples a stored transition j, flow time $t \sim \mathrm{Unif}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, I_d)$, forms

$$x = (1 - t)X_0^{(j)} + tX_1^{(j)} + \sigma\varepsilon,$$

and minimizes $\|v_\theta(x, t) - u_j\|^2$. We use Adam with batch size 512, $\beta = (0.9, 0.999)$ and no weight decay. The learning rate warms linearly for 500 steps to 10^{-3} and then follows cosine decay to 10^{-5}; gradients are clipped to norm 1. The same hyperparameters are used for every (λ, W) cell, with no early stopping, EMA or augmentation.

The memory bank, held-out data, evaluation queries and closed-form targets are fixed across training seeds, so seed variation reflects training randomness alone. We evaluate at $t \in \{0.25, 0.5, 0.75\}$ and checkpoints $1{,}000$, $3{,}000$, $10{,}000$, $30{,}000$ and $50{,}000$; the main figure reports the final checkpoint at $t = 0.5$. Each cell contains 500 evaluation queries. We report discrepancy from the closed-form minimizer, single-target replay fraction and prediction error on held-out and off-support queries.

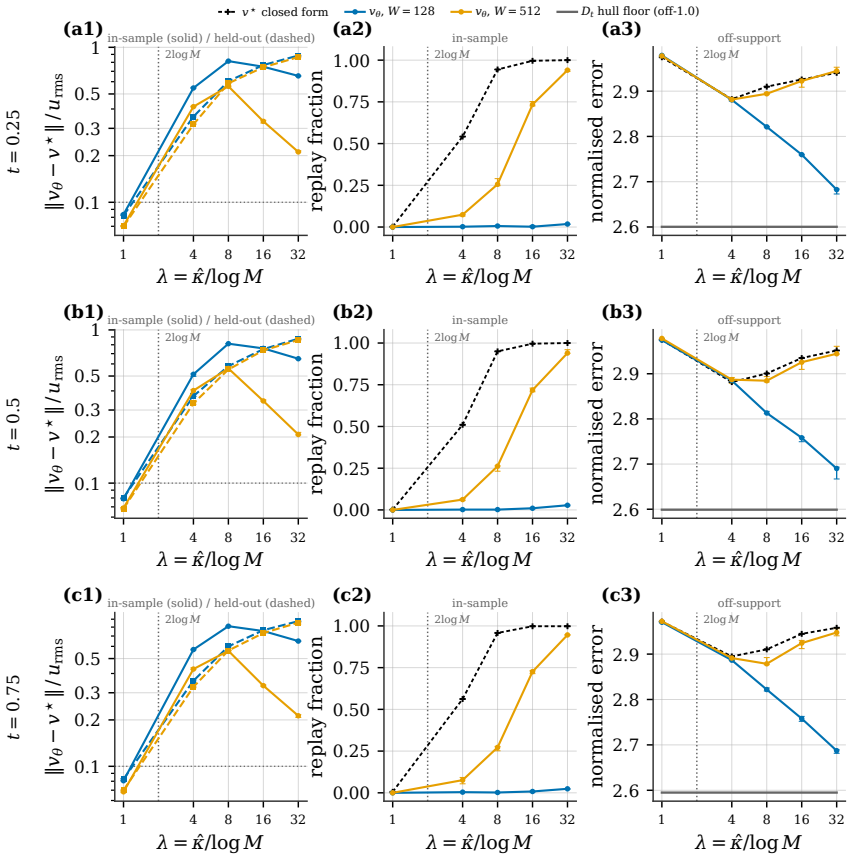
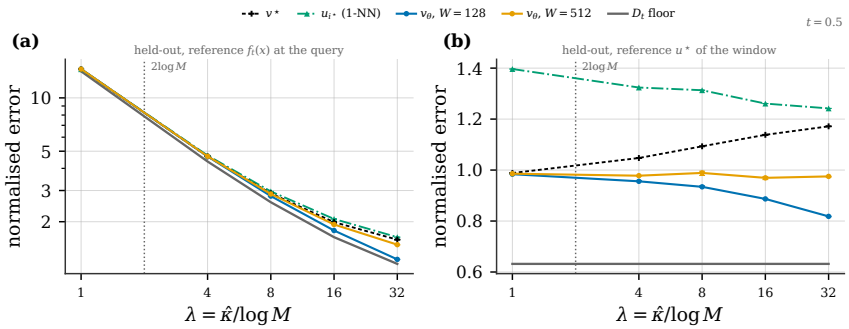
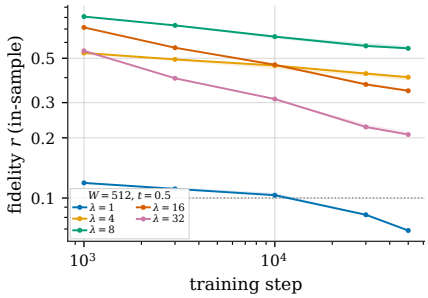


Figure 6: **Finite-capacity replay behaviour is consistent across flow time.** Rows correspond to $t \in \{0.25, 0.5, 0.75\}$. The first column shows network–closed-form discrepancy on in-sample (solid) and held-out (dashed) queries; the second shows the in-sample single-target replay fraction; and the third shows off-support prediction error for the `off-1.0` query family. The wider network increasingly inherits the replay behaviour of $v^\star$ as concentration grows, whereas the narrower network remains further from the empirical minimiser and can attain lower off-support error. The vertical dotted line marks $\widehat{\kappa} = 2\log M$.



(a) **Held-out prediction error under two reference fields.** At $t = 0.5$, we compare the closed-form field $v^\star$, the trained networks, and the nearest stored target across concentration $\lambda = \widehat{\kappa}/\log M$. (i) Error is measured against the exact true field $f_t(x)$ at the query; because the held-out query contains path noise whose scale varies with λ, the corresponding hull floor D_t also varies. (ii) Error is instead measured against the held-out window's own displacement $u^\star$, for which the target bank and hull floor are fixed across λ. This separates approximation of the stored-window target from the effect of moving the query off the attractor.



(b) **Training-time convergence toward the closed-form minimiser.** Median in-sample fidelity $r = \|v_\theta - v^\star\|/u_{\mathrm{rms}}$ for the width-512 network at $t = 0.5$, shown across training checkpoints and concentration levels $\lambda = \widehat{\kappa}/\log M$. Fidelity continues to improve through the final 50,000-step checkpoint, so the reported finite-capacity results should not be interpreted as fully converged network limits.

Figure 7: Additional finite-capacity results. Panel separates off-attractor query effects from approximation of the stored window targets. Panel (b) shows fidelity of the trained models continuing to improve above 50000 steps, so the finite-capacity results should not be interpreted as fully converged network limits

C Appendix: Related Works and Future Directions

Sequential flow matching Flow matching and stochastic interpolants have been extended to temporal and forecasting settings through conditional trajectory models, probabilistic forecasting constructions and multi-marginal paths (Zhang et al., 2024; Chen et al., 2024; Tamir et al., 2024; Lim et al., 2025; Lee et al., 2025; Islam et al., 2025; Rohbeck et al., 2025; Kviman et al., 2026). Coupling design has also been studied as a means of simplifying or straightening learned flows (Pooladian et al., 2023; Tong et al., 2024). Our question is different: we hold the empirical regression objective itself under examination and ask when its unrestricted minimiser concentrates onto stored sequential targets

High dimensional nonparametric estimation The closed-form sequential estimator is a Nadaraya-Watson regression estimator, placing replay within the classical curse of dimensionality for local regression. High-dimensional distance concentration makes nearest neighbour geometry degenerate (Beyer et al., 1999), while nonparametric methods can instead adapt to intrinsic dimension when the data possess suitable low-dimensional structure (Bickel & Li, 2007; Kpotufe, 2011). Our covariance result gives an analogous replacement of ambient dimension by a spectral effective dimension for the replay-concentration mechanism.

Memorisation and finite-capacity generalisation Empirical and theoretical work on diffusion models documents both training-data replication (Carlini et al., 2023; Somepalli et al., 2023) and collapse or memorisation phenomena in the exact generative dynamics (Biroli et al., 2024; Achilli et al., 2024; 2025). Conversely, learned score networks can exploit geometric or low-dimensional structure and depart substantially from the corresponding empirical estimator (Kadkhodaie et al., 2024; Chen et al., 2023; Boffi et al., 2025). Recent work also emphasises that finite neural networks need not realise the exact finite-data regression solution (Latourelle-Vigeant et al., 2026; Bonnaire et al., 2025). Our finite-capacity experiment is consistent with this distinction: the replay theory characterises the empirical target, while trained networks may approximate that target only partially.

Future Work The work done here naturally leads to a few open questions. First, a systematic finite-capacity forecasting study could compare the modelling choices analysed here including window order, interpolation, bandwidth and covariance geometry.

Another direction is in the theoretical aspects, as discussed in the Discussion and Limitations section where extending the theory for coverage results which do not depend on constant-in-flow-time covariance as well as off-support query log weight gap generalisations for inference could be extremely useful.

Finally, the finite-capacity results motivate studying how replay emerges during optimisation. Inspired by Latourelle-Vigeant et al. (2026), one possible approach is to analyse network training through a continuous optimisation-time or gradient-flow limit, with the empirical infinite-capacity solution representing the limiting regression target. Such an analysis could help characterise when, during training and as a function of model capacity, a learned network begins to inherit the replay behaviour of the empirical minimiser.