LANGEVIN TEMPORAL PREDICTIVE CODING TURNS POINT ESTIMATES INTO POSTERIOR DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal predictive coding (tPC) is a model of sequential memory with local inference and Hebbian learning. Its inference, however, is deterministic. At each time step it propagates a point estimate of the latent state rather than a posterior distribution. We add Langevin noise whose scale is set by the fluctuation–dissipation theorem (FDT). The resulting Langevin tPC samples the conditional posterior at each time step. A Langevin step has two hyperparameters, the preconditioner M and the step size η. We compare two preconditioners, the identity matrix I and the transition covariance Σ_H, and relate them through the effective step size. Because latent directions relax at different rates, we also propose a per-direction step size that shrinks as curvature grows. On synthetic targets, the energy and the KL divergence decrease as the FDT predicts. Under anisotropy, the per-direction step size reaches lower KL divergence than identity preconditioning. Langevin tPC also represents multimodal posteriors, which MAP inference collapses to a single mode. We then evaluate our method on twelve regression datasets with Gaussian-mixture priors, where the exact posteriors are known and ten of them are multimodal. Across all twelve datasets, Langevin tPC approximates the exact posterior more closely than an ensemble of MAP estimates. On five video benchmarks, where posteriors are unimodal, it matches standard tPC. Adding noise thus extends tPC from point estimates to posterior distributions, and from sequential memory to probabilistic inference.

1 INTRODUCTION

Predictive coding models perception as inference. Latent units predict the activity of the layer below, and prediction errors drive both the latent state and the synaptic weights, with every update using only locally available signals (Rao & Ballard, 1999; Friston, 2005; Bogacz, 2017; Millidge et al., 2021). Temporal predictive coding (tPC) extends this to sequences. At each position, recurrent weights W_H predict the current latent state from the previous one, and sensory weights W_F predict the input from the current latent state (Figure 1a) (Tang et al., 2023; Millidge et al., 2024). Its inference is deterministic descent on the prediction-error energy, so each position propagates a single latent point (Figure 1c). Occlusion and ambiguous dynamics allow multiple futures; we show that local Langevin inference lets tPC represent posterior mass across such alternatives (Figure 1d), which dispersed ensembles of deterministic tPC do not.

Not only modes, but also mass. When the posterior has several modes, a prediction needs to know both *which* alternatives exist (multimodality) and how *probable* each one is, its *posterior mass*. Running deterministic tPC from many dispersed initial states (a *dispersed ensemble*) can find every mode, but how often it lands in each reflects where it started and how large each mode's basin is, not the mode's probability (Proposition 4 in App. E.1). Here a mode's *basin* $\mathcal{B}_k$ is the set of states from which noise-free descent reaches it, and its posterior mass is $\pi(\mathcal{B}_k)$ (the component weight w_k for a well-separated mixture).

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.

From descent to sampling. Fluctuation–dissipation (FDT)-matched Langevin noise makes tPC a per-step conditional posterior sampler, a classical mechanism (Kubo, 1966; Jordan et al., 1998); finite-budget accuracy remains unresolved, because larger steps contract faster but add bias.

Three questions guide the paper.

- **Q1.** How can tPC be lifted from point estimates to posterior distributions, with local updates only, and how should the preconditioner and step size be chosen? (§3–4, §5.2)
- **Q2.** Does it capture multimodality, not only the modes but also their mass? (§5.3–5.4)
- **Q3.** Does it work on real data, both multimodal and unimodal? (§5.4–5.6)

Contributions.

- We turn tPC's deterministic sequential inference into local sampling of the per-step conditional, and show that dispersed ensembles report initialisation basin probabilities (Proposition 4 in App. E.1) whereas chains recover posterior mass on exact synthetic and UCI posteriors; oracle controls attribute 75% of the gap to weights, not shapes (Q1, Q2).
- We measure the scope, namely a 3–6 posterior-standard-deviation window at the tested budget, matched point prediction on unimodal video, and the gap to filtering (Q2, Q3).
- We test against exact posteriors on real and temporal data. On twelve regression inverse problems, Langevin tPC is closer to the exact posterior than the dispersed ensemble on every dataset (§5.4); on a switching system and a real hourly series with exact per-step posteriors, it keeps the posterior width and calibration that deterministic tPC loses (§5.5) (Q2, Q3).

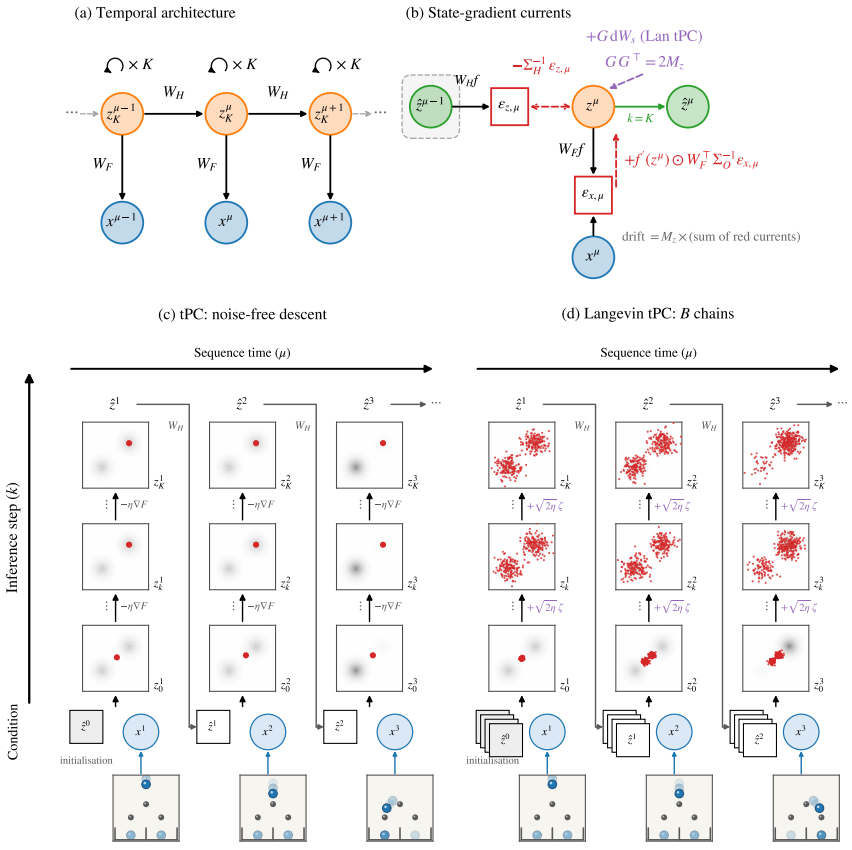

Figure 1: **Langevin tPC. (a)** z^μ predicts $z^{\mu+1}$ via W_H and x^μ via W_F. **(b)** Drift is M_z times the two error currents (4); Langevin adds $G\,dW_s$. **(c,d)** Inference over sequence time μ and inference step k on a two-mode posterior, for tPC and for B Langevin chains.

2 Related work

Predictive coding and sampling-based extensions. Predictive coding uses local prediction errors (Rao & Ballard, 1999; Friston, 2005; Bogacz, 2017; Millidge et al., 2021). tPC adds learned transitions but retains deterministic per-position inference (Tang et al., 2023; Millidge et al., 2024). Monte Carlo predictive coding (MCPC) adds Langevin noise to static inference (Oliviers et al., 2024); Sample-as-you-Infer adds an Adam-style preconditioner (Zahid et al., 2024), drawing on Riemannian Langevin methods and adaptive optimisation (Girolami & Calderhead, 2011; Li et al., 2016). *What is new.* FDT stationarity is classical (Proposition 1; Kubo, 1966; Jordan et al., 1998). The per-position update is known as MCPC's E-step (Oliviers et al., 2024), and Proposition 3 is a leading-order specialisation of Dalalyan (2017); Vempala & Wibisono (2019). Basin versus mass (Proposition 4 in App. E.1) is elementary but new for tPC. The direction-wise $\mathcal{W}_2$ bound (Theorem 1) is a Gaussian ULA calculation, new in this use. New are the local per-direction schedule (15), the measured recovery window and collapse analysis (Appendices F.1, G), and the exact-posterior benchmarks (§5.4–5.5).

Langevin dynamics and neural sampling. FDT pairs gradient drift and diffusion (Einstein, 1905; Kubo, 1966); any constant symmetric positive definite (SPD) preconditioner preserves the stationary target (Girolami & Calderhead, 2011). Euler–Maruyama gives the unadjusted Langevin algorithm (Roberts & Tweedie, 1996), trading contraction against discretisation bias (Dalalyan, 2017; Vempala & Wibisono, 2019). Neural sampling reads trial-to-trial variability as samples of uncertainty (Hoyer & Hyvärinen, 2003; Berkes et al., 2011; Orbán et al., 2016).

3 Setup

3.1 Gaussian generative model and filtering target

The model uses Gaussian transitions and emissions (Millidge et al., 2024), in the two-layer notation of Tang et al. (2023) (Figure 1). Here $z^\mu \in \mathbb{R}^m$ is the latent population, $x^\mu \in \mathbb{R}^n$ the sensory input, W_H the recurrent weights, and W_F the sensory weights, so that

$$z^\mu = W_H f(z^{\mu-1}) + \xi_{z,\mu}, \qquad \xi_{z,\mu} \sim \mathcal{N}(0, \Sigma_H), \tag{1}$$

$$x^\mu = W_F f(z^\mu) + \xi_{x,\mu}, \qquad \xi_{x,\mu} \sim \mathcal{N}(0, \Sigma_O), \tag{2}$$

The function f acts elementwise ($f = \tanh$ in every trained model), and Σ_H, Σ_O are SPD covariances. All finite-step results assume $f = \mathrm{id}$, the identity $f(u) = u$, under which (1)–(2) have an exact Kalman posterior (Kalman, 1960).

Inference targets $p(z^\mu \mid x^\mu, \hat{z}^{\mu-1})$ with parent $\hat{z}^{\mu-1}$ fixed, not the smoothing or filtering posterior $p(z^\mu \mid x^{1:\mu})$. The reason is locality, since this is the target a chain can reach from its own prediction errors, without normalising or resampling across chains. Define $\varepsilon_{z,\mu} := z^\mu - W_H f(\hat{z}^{\mu-1})$ and $\varepsilon_{x,\mu} := x^\mu - W_F f(z^\mu)$. Without parent weights $p(x^\mu \mid \hat{z}^{\mu-1})$, the B chains form an unweighted particle system; its gap to exact Kalman filtering (Kalman, 1960; Millidge et al., 2024) shows no sustained growth after the transient, and a bootstrap filter (Gordon et al., 1993) is four times further at equal B but better at matched *compute* (Table 5, App. C). The per-step energy is

$$F_\mu := \tfrac{1}{2}\varepsilon_{z,\mu}^\top \Sigma_H^{-1} \varepsilon_{z,\mu} + \tfrac{1}{2}\varepsilon_{x,\mu}^\top \Sigma_O^{-1} \varepsilon_{x,\mu} = -\log p(x^\mu, z^\mu \mid \hat{z}^{\mu-1}) + C, \tag{3}$$

with C independent of z^μ; hence $\exp(-F_\mu) \propto p(z^\mu \mid x^\mu, \hat{z}^{\mu-1})$. As in tPC (Tang et al., 2023), inference (in pseudo-time s) and learning (§3.3) both descend F_μ; with $\odot$ elementwise multiplication and f' the elementwise derivative, inference follows

$$\frac{dz^\mu}{ds} \propto -\frac{\partial F_\mu}{\partial z^\mu} = -\Sigma_H^{-1}\varepsilon_{z,\mu} + f'(z^\mu) \odot \left(W_F^\top \Sigma_O^{-1} \varepsilon_{x,\mu}\right). \tag{4}$$

3.2 Langevin inference with drift and diffusion

Langevin inference replaces descent on F_μ with the diffusion at pseudo-time $s \geq 0$

$$dz_s^\mu = -M_z \nabla_{z^\mu} F_\mu(z_s^\mu)\, ds + G\, dW_s, \qquad GG^\top = 2M_z, \tag{5}$$

where M_z is constant SPD, W_s is standard Brownian motion, G satisfies the fluctuation–dissipation relation, and the drift is the tPC update (4) with gain M_z. Euler–Maruyama with step size $\eta > 0$ gives

$$z^\mu_{k+1} = z^\mu_k - \eta M_z \frac{\partial F_\mu}{\partial z^\mu}(z^\mu_k) + \sqrt{\eta}\, G\zeta_k, \qquad \zeta_k \sim \mathcal{N}(0, I), \tag{6}$$

Each chain propagates its own final iterate $\hat{z}^\mu = z^\mu_K$, not a time average or chain mean; the B chains remain independent. Both errors use z^μ_k; inference starts at $z^\mu_0 = W_H f(\hat{z}^{\mu-1})$ and runs K steps. Removing noise recovers deterministic tPC. We call the method *Langevin tPC* and its B independent runs *chains*; the baseline, the *dispersed ensemble*, is deterministic tPC run from many initial states spread over the prior. §4.3 lists the preconditioners' roles.

3.3 Hebbian learning

After inference produces $\hat{z}^\mu$, the weights descend the same energy, errors evaluated at $\hat{z}^\mu$ (learning rate in App. A), giving

$$\begin{aligned} \Delta W_H &\propto -\frac{\partial F_\mu(\hat{z}^\mu, W_H, W_F)}{\partial W_H} = \Sigma_H^{-1} \varepsilon_{z,\mu} f(\hat{z}^{\mu-1})^\top; \\ \Delta W_F &\propto -\frac{\partial F_\mu(\hat{z}^\mu, W_H, W_F)}{\partial W_F} = \Sigma_O^{-1} \varepsilon_{x,\mu} f(\hat{z}^\mu)^\top. \end{aligned} \tag{7}$$

Each update reads only the prediction error and activity at the two ends of a synapse (Hebbian).

With B chains, $\Delta W \propto B^{-1} \sum_{b=1}^{B} \varepsilon(z_b) f(z_b)^\top$; each chain propagates its own $\hat{z}^{\mu-1}_b$.

Circuit reading, and what is not local. A chain is one realisation of the latent population z^μ, with error units computing $\varepsilon_{z,\mu}, \varepsilon_{x,\mu}$; B chains are B copies (or trials) of the circuit. The noise is intrinsic variability of variance $2\eta M_z$; a diagonal M_z is a per-unit gain on the error current, so each unit reads only its own errors (Appendix I.4). A read-out averaging over chains (the posterior mean of §5.6) and the $B^{-1}\sum_b$ of the weight update are not local, because they pool copies.

4 Theory

4.1 Per-step conditional posterior as Langevin equilibrium

Proposition 1 (Per-step conditional posterior as Langevin equilibrium)**.** *For each position $\mu \geq 1$ and fixed $\hat{z}^{\mu-1}$, the SDE* (5) *with any constant SPD preconditioner M_z admits the unique stationary density*

$$q_\infty(z^\mu) = Z_\mu^{-1} e^{-F_\mu(z^\mu)} = p(z^\mu \mid x^\mu, \hat{z}^{\mu-1}). \tag{8}$$

Proof. Appendix B.

4.2 Free-energy dissipation and KL contraction

Here $\mathbb{H}[q]$ denotes differential entropy and D_{KL} the Kullback–Leibler divergence.

Proposition 2 (Joint descent)**.** *Let q_s be the law of z^μ_s under* (5)*. With $\mathcal{F}[q] := \mathbb{E}_q[F_\mu] - \mathbb{H}[q]$ and the evidence lower bound $\mathrm{ELBO}[q] := \mathbb{E}_q[\log p(x^\mu, z^\mu \mid \hat{z}^{\mu-1})] + \mathbb{H}[q]$,*

$$\frac{d}{ds}\mathrm{ELBO}[q_s] = -\frac{d}{ds}\mathcal{F}[q_s] = -\frac{d}{ds} D_{\mathrm{KL}}(q_s \| q_\infty) = \mathbb{E}_{q_s}\left[\left\| \nabla \log \frac{q_s}{q_\infty} \right\|^2_{M_z} \right] \geq 0, \tag{9}$$

with equality iff $q_s = q_\infty$; the common rate is the entropy-production rate (Seifert, 2012).

4.3 Choosing the preconditioner by direction-wise finite-step analysis

The four preconditioners are $M_z{=}I$ (baseline; all mass experiments), $M_z{=}\Sigma_H$ (covariance matching) and the per-direction schedule below (a non-local diagnostic), with its local within-chain version.

They share the stationary target and differ only at a finite budget. For $f = \mathrm{id}$ (linear-Gaussian), write $M = M_z$ and

$$H := \Sigma_H^{-1} + W_F^\top \Sigma_O^{-1} W_F \tag{10}$$

for the posterior precision; this refines finite-step ULA analyses (Dalalyan, 2017; Vempala & Wibisono, 2019). Write $\bar{z}_i$ for the posterior mean in direction i and $\mathcal{W}_2$ for Wasserstein-2 distance.

Theorem 1 (Anisotropic Wasserstein-2 bound for preconditioned ULA)**.** *Take the linear-Gaussian model ($f = \mathrm{id}$) and a constant SPD preconditioner M that commutes with H, so both are diagonal in one orthonormal eigenbasis. In that common eigenbasis write $H = \mathrm{diag}(h_i)$ and $M = \mathrm{diag}(m_i)$, and set the effective step $\eta_i := \eta m_i$ and contraction factor $a_i := 1 - \eta_i h_i \in (0, 1)$. Then for* every $K \geq 0$*,*

$$\mathcal{W}_2^2(q_K^{\eta,M}, q_\infty) \;\leq\; \underbrace{\sum_i a_i^{2K}\, \mathbb{E}_{q_0}\big[(z_i - \bar{z}_i)^2\big]}_{\textit{initial-state transient}} \;+\; \underbrace{\sum_i \left(\sqrt{v_i\,(1 - a_i^{2K})} - h_i^{-1/2}\right)^2}_{\textit{width deficit}}, \tag{11}$$

where $v_i := [\, h_i(1 - \eta_i h_i/2)\,]^{-1}$ is the stationary variance of fixed-step ULA in direction i, with equality when q_0 is a point mass — the initialisation Algorithm 1 uses. Appendix D.1 states the hypotheses in full and proves the bound.

It covers $M_z{=}I$ and the schedule below; for the other arms §5.2 is empirical.

Summed over directions, the same analysis gives an aggregate criterion for any constant preconditioner.

Proposition 3 (Finite-step KL, to leading order in η)**.** *For $f = \mathrm{id}$, any constant SPD M, $\eta\, \lambda_{\max}(M^{1/2} H M^{1/2}) < 1$ and large $K\eta$,*

$$D_{\mathrm{KL}}(q_K^{\eta,M} \| q_\infty) \approx C_M\, \eta^2 + e^{-2\lambda_M K \eta} D_{\mathrm{KL}}(q_0 \| q_\infty), \qquad C_M = \tfrac{1}{16}\mathrm{tr}\big((MH)^2\big) + O(\eta), \tag{12}$$

with $\lambda_M = \lambda_{\min}(M^{1/2} H M^{1/2})$ (leading order, not a bound; App. D.2).

Scalar $M_z = \Sigma_H = \sigma_H^2 I$ is $M_z{=}I$ at step $\sigma_H^2 \eta$, so §5.2 tunes each arm's step.

(i) Each direction has its own relaxation budget. The relaxation number $u_i := \eta_i h_i K$ counts contraction time-constants. Initialisation bias persists when $u_i \ll 1$; both transients decay when $u_i \gg 1$.

(ii) Equal stationary width bias fixes the per-direction schedule. The chain tends to width $\sqrt{v_i}$, not $h_i^{-1/2}$. Its bias is $b_i = \sqrt{v_i} - h_i^{-1/2} = \eta_i \sqrt{h_i}/4 + O(\eta_i^2 h_i^{3/2})$ (Appendix D.1). Equalising the leading term gives, for a tunable constant c,

$$\eta_i^\star = \frac{c}{\sqrt{h_i K}}, \qquad \text{under which} \quad b_i^2 = \frac{c^2}{16K}\Big[1 + O\big(c\sqrt{h_i/K}\big)\Big] \;\text{ in every direction,} \tag{13}$$

This *equal-bias* choice (stable when $c^2 h_{\max} < K$, so $\eta_i h_i < 1$) balances leading stationary width errors; higher orders still depend on curvature.

(iii) The per-direction schedule is a constant preconditioner. With U the eigenvectors of H, (13) and FDT-paired noise reproduce (6) exactly with

$$\eta M_z = U\, \mathrm{diag}(\eta_i^\star)\, U^\top = \frac{c}{\sqrt{K}}\, H^{-1/2} \tag{14}$$

so Theorem 1 applies to it directly (App. D.1).

Equal bias is not optimality. KL bias is equalised by $M_z \propto H^{-1}$ (Remark 1, App. D.2).

A local per-direction schedule. The per-direction schedule (14) needs the eigendecomposition of H, so it is non-local, but each chain already estimates its own units' standard deviations. Let s_i be

the running standard deviation of unit i's own activity within a K-step $M_z{=}I$ pilot of its own chain (App. D). The *local per-direction schedule* is

$$\eta_i^{\mathrm{loc}} = \frac{c}{\sqrt{K}}\, s_i, \tag{15}$$

i.e. a constant diagonal M_z, so Proposition 1 applies unchanged. Since $h_i^{-1/2}$ is the posterior standard deviation along eigendirection i, (13) and (15) coincide when the posterior is axis-aligned. It recovers a fraction of the per-direction schedule's KL gain over $M_z{=}I$ that grows with chain length (§5.2; per-K fractions and step scaling in App. D), and none on trained Moving MNIST (Table 7 in App. D.3), whose anisotropy is off-axis.

5 Experiments

Error bars are one standard error of the mean (s.e.m.) over the unit named in each caption; Table 19 (App. I.1) alone reports a standard deviation. Synthetic tests use exact calculations where available; video uses Moving MNIST, SM-MNIST, KTH, UCF101 and BAIR (Srivastava et al., 2015; Denton & Fergus, 2018; Schüldt et al., 2004; Soomro et al., 2012; Ebert et al., 2017). Anisotropy is how many orders of magnitude the prior variances span across latent directions. Seeds, sizes and all other settings are in App. A.

5.1 Stationary target

We give the sampler a nonlinear two-dimensional model, scored against a NUTS reference (Hoffman & Gelman, 2014), and a linear-Gaussian model with an exact posterior, and track free energy, ELBO and KL to the reference along the chain for Langevin and deterministic tPC. **Once the chain is stationary, sampled KL tracks free energy, and only Langevin inference recovers posterior width** (Figure 2; Figure 5 in App. D, which also gives the numbers).

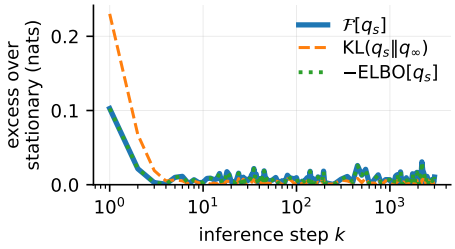

Figure 2: **Free energy, ELBO and KL along the chain.** Excess free energy, negated ELBO and KL to a NUTS reference on the nonlinear 2D target (Proposition 2); Figure 5 in App. D gives the sweep.

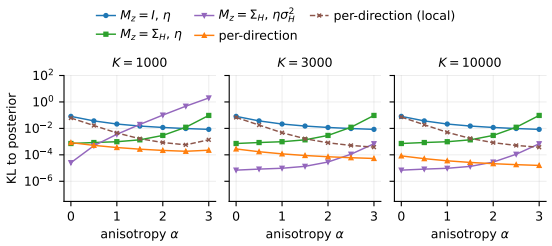

Figure 3: **Finite-K behaviour across anisotropy.** Exact KL to the closed-form posterior for three chain lengths. Dashed lines are per-direction (local), (15).

5.2 Finite-step anisotropy

We give linear-Gaussian tPC priors of increasing anisotropy and several chain budgets, and compare the identity preconditioner, the prior-covariance preconditioner and the per-direction schedule by exact KL to the closed-form posterior, first at one shared step-size constant (Figure 3), then with each at its own best step size. **No arm wins everywhere. With each at its own best step size, the prior-covariance preconditioner wins at moderate anisotropy and the per-direction schedule at the highest, and the schedule's advantage over the identity grows with anisotropy** (Figure 7 in App. D.3). A within-chain local version of the schedule recovers most of its gain, but only for long chains, and the aggregate criterion of Proposition 3 ranks the arms correctly only at each arm's own step size (Table 6 and Figure 6 in App. D.3, with the numbers).

5.3 MULTIMODALITY BY PACHINKO

Sampler versus ensemble. Descent from many initial states reports how much initial mass lands in each basin; a diffusion reports each basin's posterior mass (Proposition 4 in App. E.1).

Known weights. We give two-basin targets with known, unequal weights and read out the weights each method recovers. **Langevin tPC recovers them; the dispersed ensemble follows its initialisation** (Figure 8 in App. E.1).

Informative observation. We feed observations that shift the exact posterior weights. **The chains track the weights; deterministic protocols do not** (Table 1).

Occluded pachinko drop. We feed videos of a ball bouncing left or right off a hidden peg. **Each chain commits to one branch, the chains together cover both, and they score better than deterministic forecasts** (Figure 4; Table 10 in App. E.5).

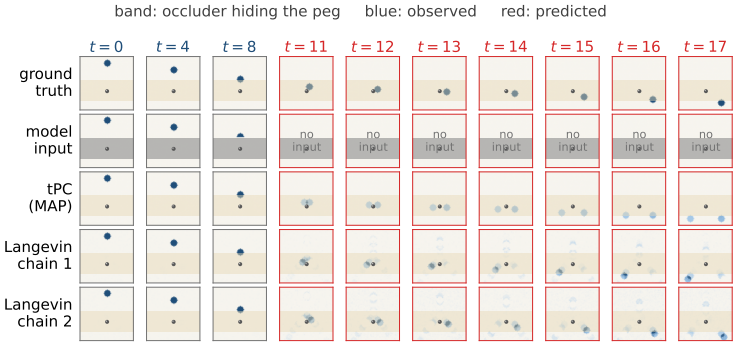

Figure 4: **Occluded pachinko drop rollouts per method.** Rows show ground truth; model input (occluder band hiding the peg masked; no input after the conditioning window); tPC (MAP) with the single-Gaussian transition (1); two Langevin chains with the mixture transition (19). Blue marks observed and red predicted frames.

Table 1: **Recovered posterior weight per protocol.** 1-D mixture, analytic posterior, 20 seeds, B=100, K=3000.

exact posterior mass	0.20	0.50	0.80
Langevin chains	**0.188**	**0.504**	**0.806**
dispersed ensemble (equal-weight initial states)	0.487	0.502	0.511
branch drawn from the prior, then filtered	0.500	0.500	0.500

5.4 REAL MULTIMODAL POSTERIORS BY UCI

We give the models twelve regression inverse problems (ten UCI datasets (Kelly et al., 2026), diabetes and California housing (Efron et al., 2004; Pace & Barry, 1997)). The latent is a clean input and the observation its noisy copy, and a mixture prior fitted on training inputs only makes the exact posterior a known, mostly multimodal mixture. We compare Langevin tPC with a dispersed MAP ensemble given far more compute, by distance to the exact posterior, interval coverage and mode-mass error (Table 2). **Langevin tPC is closer to the exact posterior on every dataset and better calibrated on all but one, though it still under-covers.** The gain is mode mass, not mode discovery, and it holds when modes are moderately separated, as for most UCI inputs (App. K.1, K.4, F.1); this test is static, and baselines, read-outs, numbers and controls for mode count, prior family, leakage and splits are in App. K.

Table 2: **Distance to the exact posterior on twelve regression datasets.** 15,000 returned points per method; *tPC ensemble* is the dispersed ensemble; bold marks the better (for coverage, closer to 90%). MMD^2 in Table 24 (App. K); servo and yacht have one supported mode (Table 26 in App. K.3).

	$\text{SW}_2^2 \downarrow$		cov_{90} (%), target 90		mass $L_1 \downarrow$	
dataset	mean over Langevin chains	tPC ensemble	mean over Langevin chains	tPC ensemble	mean over Langevin chains	tPC ensemble
auto MPG	**0.071**	0.485	**74**	68	**0.28**	0.95
wine quality	**0.174**	0.376	**45**	30	**0.64**	1.28
energy efficiency	**0.134**	0.420	**94**	54	**0.49**	1.14
concrete	**0.042**	0.255	**72**	45	**0.35**	1.12
power plant	**0.023**	0.337	**93**	52	**0.16**	0.80
bike sharing	**0.147**	0.416	**60**	50	**0.47**	1.19
diabetes	**0.040**	0.244	**66**	36	**0.38**	1.00
california housing	**0.060**	1.504	**84**	20	**0.19**	1.13
abalone	**0.207**	0.501	**94**	59	**0.72**	1.30
yacht	**0.050**	0.261	**88**	5	**0.27**	1.00
real estate	**0.074**	0.440	57	**68**	**0.24**	0.99
servo	**0.090**	0.258	**81**	7	**0.59**	1.38
pooled	**0.093**	0.458	**76**	41	**0.40**	1.11

5.5 Temporal posteriors

We give the models a synthetic switching linear system and a real hourly bike-sharing series with hidden day-type regimes, each at several observation-noise levels; both keep the transition, and each chain carries its own state forward, so every per-step posterior is an exact two-component mixture. We compare Langevin tPC with the dispersed ensemble by distance, mode-mass error and interval coverage against that posterior (Table 3; full Tables 11 and 12 in App. F.2 and F.3, with the numbers). **Langevin tPC is far closer at every noise level on both; its mass advantage on the switching system holds only at intermediate noise, and on the real series its intervals stay near nominal while the ensemble's cover nothing.** The real series tests width and calibration, not mass, because its regimes overlap.

Table 3: **Temporal benchmarks with exact per-step posteriors.** Mean over Langevin chains vs. tPC (dispersed) ensemble per observation-noise level σ_O; full results in Tables 11 and 12 in App. F.

	$\text{SW}_2^2 \downarrow$		mass $L_1 \downarrow$		cov_{90} (%), target 90	
σ_O	mean over Langevin chains	tPC ensemble	mean over Langevin chains	tPC ensemble	mean over Langevin chains	tPC ensemble
switching linear system (synthetic)						
0.25	**0.0003**	0.0538	0.146	**0.121**	–	–
0.50	**0.0008**	0.0780	**0.071**	0.212	–	–
1.00	**0.0023**	0.1128	**0.086**	0.167	–	–
2.00	**0.0032**	0.1211	0.078	**0.037**	–	–
4.00	**0.0031**	0.1221	0.069	**0.010**	–	–
hourly bike-sharing series (real)						
0.25	**0.0002**	0.0477	**0.0017**	0.0050	**90.0**	0.0
0.50	**0.0004**	0.1113	**0.0042**	0.0120	**85.7**	0.0
1.00	**0.0010**	0.1665	**0.0063**	0.0096	**82.7**	0.0
2.00	**0.0017**	0.1900	**0.0074**	0.0085	**82.1**	0.0

5.6 Unimodal posteriors by video

This is a negative control. For linear-Gaussian models the posterior mean and MAP coincide exactly, and for the nonlinear video emission we observe it empirically, so the two should predict alike. We give deterministic and Langevin tPC the five video datasets under a matched protocol and compare their rollouts by *PSNR* (peak signal-to-noise ratio, a log-scale pixel error) and *SSIM* (structural similarity (Wang et al., 2004) of local brightness, contrast and structure); higher is better. **The two predict alike on every dataset, both from the same weights and when trained separately** (Table 4); model-free predictors such as persistence match or beat both, which limits this check's resolution (App. J, with the numbers).

Table 4: **Matched video benchmarks.** Scores average 10 predicted frames and 20 held-out sequences; **one training seed per cell**, shared by both arms, so differences carry no error bar. Model-free baselines are in Table 22 (App. J.1). †public 405-clip, 10-class subset. ‡greyscale.

	tPC		Langevin tPC		difference	
Dataset	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR	SSIM
Moving MNIST	14.83	0.616	14.84	0.618	+0.01	+0.002
SM-MNIST	16.64	0.746	16.66	0.748	+0.02	+0.003
KTH	24.13	0.534	24.28	0.542	+0.15	+0.008
UCF101†	21.79	0.550	21.85	0.553	+0.06	+0.003
BAIR‡	16.89	0.441	16.88	0.441	−0.01	+0.000
average (5)	18.85	0.577	18.90	0.580	+0.04	+0.003

6 Discussion

(i) Q1. How to lift tPC from point estimates to posterior distributions? Add Langevin noise matched to the descent by the fluctuation–dissipation relation. Each position then samples its per-step conditional (Proposition 1) with the same local error currents. Within a finite budget, the step should be set per direction (Theorem 1). *Scope.* The target is the per-step conditional, not the filtering marginal, for which sequential Monte Carlo is better (App. C); the finite-step theory is linear-Gaussian.

(ii) Q2. Does it capture not only the modes but also their mass? Yes. Chains weigh the modes by their posterior mass, while the dispersed ensemble weighs them by where it started (Proposition 4 in App. E.1). This holds when the modes are close enough for the chains to cross between them within the budget (App. F.1); the number of modes itself is not the limit (App. H).

(iii) Q3. Does it work on real data, multimodal and unimodal? Yes. On every multimodal real dataset Langevin tPC is closer to the exact posterior than the dispersed ensemble; on the real hourly series it is better calibrated; on unimodal video it matches deterministic tPC.

Conclusion. Q1. Fluctuation–dissipation-matched Langevin noise lifts tPC to local posterior sampling, with the step set per direction. **Q2.** Yes, it recovers both the modes and their mass. **Q3.** Yes, it wins on multimodal real data and matches on unimodal data. **Limitations.** Current tPC cannot yet predict very complex dynamics. Its transition is a single learned linear map with one nonlinearity, and on video trivial persistence predictors beat it. Temporal inference does not yet hold over long horizons. Each position conditions on its own propagated parent, so errors can accumulate along long sequences, and the per-step target differs from the filtering marginal. Both call for future work, such as richer transitions (mixtures or hierarchies) and local mechanisms that correct accumulated error or reweight chains.

Reproducibility statement

Algorithms 1–2 specify inference and learning; appendices give proofs and finite-step derivations. Experiments fix seeds, write machine-readable outputs and state seed counts and single-seed results.

AI use statement

LLM assistants were used to implement and run experiments, to draft and check derivations and proofs, to analyse results, to search for and check related work and references, and to draft and edit the text. The authors verified the retained mathematical claims and empirical numbers against the released code and take full responsibility for the content; the two reference gaps in Table 23 are quoted from an earlier, uncontrolled run.

REFERENCES

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX*, volume 1123 of *Lecture Notes in Mathematics*, pp. 177–206. Springer, 1985.

Thomas Bengtsson, Peter Bickel, and Bo Li. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 2 of *IMS Collections*, pp. 316–334. Institute of Mathematical Statistics, 2008. arXiv:0805.3034.

Pietro Berkes, Gergő Orbán, Máté Lengyel, and József Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, 2011.

Rafal Bogacz. A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, 76:198–211, 2017.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes I: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004. doi: 10.4171/JEMS/14.

Yifan Chen, Xiaoou Cheng, Jonathan Niles-Weed, and Jonathan Weare. Convergence of unadjusted Langevin in high dimensions: Delocalization of bias. *Communications on Pure and Applied Mathematics*, 79(6):1467–1491, 2026. doi: 10.1002/cpa.70032.

Xiwei Cheng, Kexin Fu, and Farzan Farnia. On the hardness of sampling from mixture distributions via Langevin dynamics. *arXiv preprint arXiv:2406.02017*, 2024.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2980–2988, 2015.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B*, 79(3):651–676, 2017.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1174–1183, 2018.

Arnaud Doucet, Nando de Freitas, and Neil Gordon (eds.). *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

Frederik Ebert, Chelsea Finn, Alex X. Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *Conference on Robot Learning (CoRL)*, volume 78 of *Proceedings of Machine Learning Research*, pp. 344–356, 2017.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–451, 2004. doi: 10.1214/009053604000000067.

Albert Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322(8):549–560, 1905.

Hadi Fanaee-T and João Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2–3):113–127, 2014. doi: 10.1007/s13748-013-0040-3.

Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, 2005.

Zoubin Ghahramani and Michael I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29: 245–273, 1997. doi: 10.1023/A:1007425814087.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 73(2):123–214, 2011.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113, 1993. doi: 10.1049/ip-f-2.1993.0015.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

Patrik O. Hoyer and Aapo Hyvärinen. Interpreting neural response variability as Monte Carlo sampling of the posterior. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, 2003.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, 2016.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. `https://archive.ics.uci.edu`, 2026. Accessed September 2026.

Ryogo Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255–284, 1966.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *AAAI Conference on Artificial Intelligence*, volume 30, 2016.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992. doi: 10.1162/neco.1992.4.3.448.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

Georg Menz and André Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, 42(5):1809–1884, 2014. doi: 10.1214/14-AOP908.

Sean P. Meyn and R. L. Tweedie. Stability of Markovian processes II: continuous-time processes and sampled chains. *Advances in Applied Probability*, 25(3):487–517, 1993a. doi: 10.2307/1427521.

Sean P. Meyn and R. L. Tweedie. Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548, 1993b. doi: 10.2307/1427522.

Beren Millidge, Anil Seth, and Christopher L. Buckley. Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979*, 2021.

Beren Millidge, Mufeng Tang, Mahyar Osanlouy, Nicol S. Harper, and Rafal Bogacz. Predictive coding networks for temporal prediction. *PLOS Computational Biology*, 20(4):e1011183, 2024. doi: 10.1371/journal.pcbi.1011183.

Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity. *Bernoulli*, 28(3):1577–1601, 2022. doi: 10.3150/21-BEJ1343.

Gaspard Oliviers, Rafal Bogacz, and Alexander Meulemans. Learning probability distributions of sensory inputs with Monte Carlo predictive coding. *PLOS Computational Biology*, 20(10): e1012532, 2024.

Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33 (3):291–297, 1997. doi: 10.1016/S0167-7152(96)00140-X.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

Patrick Rebeschini and Ramon van Handel. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866, 2015. doi: 10.1214/14-AAP1061.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local SVM approach. In *International Conference on Pattern Recognition (ICPR)*, volume 3, pp. 32–36, 2004.

Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, 2012.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 843–852, 2015.

Mufeng Tang, Helen Barron, and Rafal Bogacz. Sequential memory with temporal predictive coding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44341–44355, 2023.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. doi: 10.1007/b13794.

Santosh S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Umais Zahid, Qinghai Guo, and Zafeirios Fountas. Sample as you infer: Predictive coding with Langevin dynamics. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58105–58121, 2024.

APPENDIX CONTENTS

A Notation and algorithms **15**

B Continuous-time proofs **16**

C The per-step conditional versus the filtering marginal as inference target **17**

D Finite-step theory and anisotropy controls **17**

D.1 Direction-wise derivation of Theorem 1 18

D.2 Order of the stationary bias in Proposition 3 18

D.3 Oracle-tuned preconditioners 19

E Multimodal calibration and occlusion controls **22**

E.1 Proof of Proposition 4 22

E.2 Occluded pachinko drop as a representational control 23

E.3 Why a single-Gaussian transition cannot carry the branch 23

E.4 Appearance transfer 24

E.5 A proper scoring rule on a learned nonlinear temporal model 24

F Scope of the mass claim, and two benchmarks it is tested on **24**

F.1 Between-mode mass is recovered inside a window, not everywhere 25

F.2 A temporal benchmark whose per-step posterior is exactly a mixture 25

F.3 Real data, kept temporal, with the per-step posterior still in closed form 26

F.4 A probabilistic baseline that has width, and the read-out we should have used 27

G When does the branch representation collapse? **27**

H Scaling with the number of independent ambiguities **28**

H.1 Results 29

H.2 Coupled ambiguities 30

H.3 Proof of Proposition 5 31

H.4 Setup 32

I Equilibration, criterion, cost and locality **33**

I.1 The chains equilibrate rather than remember their initialisation 33

I.2 On a unimodal posterior the criterion, not the modality, decides 33

I.3 What a chain actually costs 33

I.4 What each variant requires a unit to read 34

J Video controls and protocol audit **35**

J.1 How much dynamic range do rollout PSNR and SSIM have here? 35

J.2 Controlled audit of the unstabilised-versus-stabilised protocol 35

K UCI posterior-recovery controls **35**

K.1 Is the advantage between-mode mass, or within-mode shape? 37
K.2 The thirteenth dataset, and why it is not in the pool 37
K.3 Mode count and sensitivity to the prior 37
K.4 Mode reach versus mode weight 38
K.5 Prior-leakage controls 38
K.6 Training-split check 38

L How the latent dimension acts on learning, with three clocks checked on an LDS 40

Where the main text points. §3 draws on App. A (notation, algorithms, hyperparameters) and C (target); §4 on B, D; §5.3 on E; §5.4 on F, K; §5.5 on F.2, F.3, G; §5.6 on J; scaling, cost and locality are covered in H, I; latent dimension m and learning in L.

A NOTATION AND ALGORITHMS

Notation.

z^μ, x^μ	latent / observation at position μ	η, η_i, η_w	global / effective directional / learning step
$\hat{z}^{\mu-1}$	previous inferred latent (frozen)	K, B	chain length / number of chains
m, n, P	latent dim. / obs. dim. / sequence length	q_s, q_∞	iterate law / stationary density
W_H, W_F	transition / emission weights	H	Hessian $\Sigma_H^{-1} + W_F^\top \Sigma_O^{-1} W_F$ in the linear case
Σ_H, Σ_O	transition / observation covariances	h_i	Hessian curvature in the common eigenbasis
M_z	constant SPD preconditioner	G	diffusion factor, $GG^\top = 2M_z$
α	controlled anisotropy of Σ_H	F_μ	per-step energy in (3)

Hyperparameters and settings. The main text names five inference knobs, namely step size η, preconditioner M_z, chain length K, number of chains B and schedule constant c. Every other setting it uses is listed here; section-specific appendices give the rest.

experiment	settings
learning (§3.3)	learning rate η_w; optional spectral cap κ (Algorithm 2)
synthetic tests (§5)	5–30 seeds
anisotropy (§5.2)	$\Sigma_H = \mathrm{diag}(0.1 \cdot 10^{\alpha i/m})_{i=0}^{m-1}$ (α=3 spans three orders of magnitude); $m = 32$, $n = 64$; shared constant $c = 0.2$
local schedule (§4.3)	EMA window $\tau = K/8$ over the second half of a K-step $M_z = I$ pilot (App. D)
known weights (§5.3)	ten weight/separation settings, B=100, 20 seeds
informative observation (§5.3)	prior $\frac{1}{2}\mathcal{N}(-2, 1) + \frac{1}{2}\mathcal{N}(2, 1)$, likelihood $x \mid z \sim \mathcal{N}(z, 4)$; B=100, K=3000, 20 seeds
UCI (§5.4)	$M_z = I$, $\eta = 0.01$, $K = 300$, $B = 100$, first half of each chain discarded as burn-in; prior by k-means, J=4 components; σ_x=1; coverage read-out $\hat{y}^{(b)} = w_{\mathrm{ro}}^\top z^{(b)} + \beta_0$ fitted by least squares on training data (shared by both arms, so coverage supports comparison, not certified calibration); 5 seeds, 50 held-out inputs each

Algorithm 1 Langevin tPC filtering (E-step), one sequence

Require: observations $x^{1:P}$; weights W_H, W_F; covariances Σ_H, Σ_O; preconditioner M_z with diffusion factor G, $GG^\top = 2M_z$; step η; steps per position K; chains B

$z_{(b)}^0 \leftarrow 0$ for all b
for $\mu = 1$ **to** P **do**
  $z_{(b)}^\mu \leftarrow W_H f(z_{(b)}^{\mu-1})$ // each chain from its own transition mean
  for $k = 1$ **to** K **do**
    $\varepsilon_z \leftarrow z_{(b)}^\mu - W_H f(z_{(b)}^{\mu-1}), \quad \varepsilon_x \leftarrow x^\mu - W_F f(z_{(b)}^\mu)$
    $\xi \sim \mathcal{N}(0, I_m)$
    $z_{(b)}^\mu \leftarrow z_{(b)}^\mu - \eta\, M_z \Big[\Sigma_H^{-1}\varepsilon_z - f'(z_{(b)}^\mu) \odot \big(W_F^\top \Sigma_O^{-1} \varepsilon_x\big)\Big] + \sqrt{\eta}\, G\, \xi$
  end for
end for
return chains $\{z_{(b)}^\mu\}_{b=1}^B$ for every position μ

Algorithm 2 Monte Carlo EM training, one sequence

Require: observations $x^{1:P}$; weights W_H, W_F; learning rate η_w; optional spectral cap κ; everything Algorithm 1 requires

$\{z^{\mu}_{(b)}\} \leftarrow$ Algorithm 1$(x^{1:P}, \dots)$ // E-step; chains never merged
$\Delta W_H \leftarrow 0; \quad \Delta W_F \leftarrow 0$
for $\mu = 1$ **to** P **do**
  $\varepsilon^{(b)}_{z,\mu} \leftarrow z^{\mu}_{(b)} - W_H f(z^{\mu-1}_{(b)}), \quad \varepsilon^{(b)}_{x,\mu} \leftarrow x^{\mu} - W_F f(z^{\mu}_{(b)})$ // recomputed here, at the pre-sweep weights
  $\Delta W_H \leftarrow \Delta W_H + \frac{1}{B}\sum_b \Sigma_H^{-1} \varepsilon^{(b)}_{z,\mu} f(z^{\mu-1}_{(b)})^\top$
  $\Delta W_F \leftarrow \Delta W_F + \frac{1}{B}\sum_b \Sigma_O^{-1} \varepsilon^{(b)}_{x,\mu} f(z^{\mu}_{(b)})^\top$ // M-step; one update per chain, averaged after
end for
$W_H \leftarrow W_H + \eta_w \Delta W_H; \quad W_F \leftarrow W_F + \eta_w \Delta W_F$ // one update per sweep; W_H, W_F frozen inside the loop
if κ is given and $\|W_H\|_2 > \kappa$ **then**
  $W_H \leftarrow \kappa\, W_H / \|W_H\|_2$ // power iteration
end if
return W_H, W_F

B CONTINUOUS-TIME PROOFS

Standing assumptions. Write τ for the target temperature, with $\tau = 1$ in the main text. Assume (A1) $F_\mu \in C^2(\mathbb{R}^m)$ is bounded below and $\int e^{-F_\mu/\tau}\, dz < \infty$; (A2) F_μ is coercive; (A3) the Itô SDE has a unique non-explosive strong solution; and (A4) q_0 has finite $D_{\mathrm{KL}}(q_0 \| q_\infty)$.

Fokker–Planck equation. For $dZ_s = b(Z_s)\, ds + G\, dW_s$ with constant G,

$$\partial_s q_s = -\nabla \cdot (b q_s) + \tfrac{1}{2} \nabla \cdot (G G^\top \nabla q_s).$$

For $b = -M_z \nabla F_\mu$, the FDT choice $GG^\top = 2\tau M_z$ gives

$$\partial_s q_s = \nabla \cdot \left[M_z q_s \nabla \log \frac{q_s}{q_\infty} \right], \qquad q_\infty \propto e^{-F_\mu/\tau},$$

so q_∞ has zero stationary probability current. Conversely, this pairing is required for detailed-balance equilibrium with this drift.

Free-energy identity. With $\mathcal{F}[q] = \mathbb{E}_q[F_\mu] - \tau \mathbb{H}[q]$,

$$\mathcal{F}[q] - \mathcal{F}[q_\infty] = \tau D_{\mathrm{KL}}(q \| q_\infty), \qquad \mathcal{F}[q_\infty] = -\tau \log Z_\mu.$$

At $\tau = 1$, differentiating the KL along the Fokker–Planck flow and integrating by parts gives

$$\frac{d}{ds} D_{\mathrm{KL}}(q_s \| q_\infty) = -\mathbb{E}_{q_s}\left[\left\| \nabla \log \frac{q_s}{q_\infty} \right\|^2_{M_z} \right] \le 0.$$

Equality holds only at $q_s = q_\infty$ under the recurrence assumptions.

Proof of Proposition 1. The FDT pairing above makes $q_\infty \propto e^{-F_\mu}$ stationary. Equation (3) and Bayes' rule identify it with $p(z^\mu \mid x^\mu, \hat{z}^{\mu-1})$. The KL identity is a strict Lyapunov function, and Harris recurrence under the standing assumptions gives uniqueness (Meyn & Tweedie, 1993a;b). □

Proof of Proposition 2. The free-energy identity gives $d\mathcal{F}[q_s]/ds = dD_{\mathrm{KL}}(q_s \| q_\infty)/ds$, while $\mathrm{ELBO}[q] = -\mathcal{F}[q] + C_\mu$ at $\tau = 1$. Combining these with the KL derivative above yields all equalities in (9). □

C THE PER-STEP CONDITIONAL VERSUS THE FILTERING MARGINAL AS INFERENCE TARGET

Algorithm 1 is an unweighted particle system that omits the bootstrap filter's predictive-likelihood weight $Z(z') = p(x^\mu \mid z')$. This analytically explains the gap between its frozen-parent target $p(z^\mu \mid x^\mu, \hat{z}^{\mu-1})$ (Proposition 1) and the filtering marginal $p(z^\mu \mid x^{1:\mu})$.

Setup. Linear-Gaussian tPC with m=16, n=32, P=20, σ_H=σ_O=0.3, $\|W_H\|_2 = 0.9$; B=2000, K=400, 5 seeds. We report moment-based Gaussian KL to the exact Kalman marginal for Algorithm 1 and a multinomial-resampling bootstrap filter (Gordon et al., 1993).

Table 5: **Gaussian KL to the exact Kalman filtering marginal by position**, for Algorithm 1 and a bootstrap particle filter; B=2000, 5 seeds.

position μ	Algorithm 1	bootstrap PF
1	0.131	1.307
5	0.307	1.441
10	0.260	1.761
15	0.257	1.694
20	0.275	2.400

The bias is a per-step offset rather than cumulative. Algorithm 1 gives 0.13 nats initially and 0.26 averaged over $\mu > 10$, without a positional trend.

The same-B comparison is not a compute match, and we say so. A separate m=16, n=32, B=2000, K=400 sweep gives chain KL 0.286 nats (mean 0.251 in Table 5). The bootstrap filter gives 1.18 nats, four times worse through particle degeneracy in 16 dimensions. Langevin costs $2BKmn$ multiply–accumulates per position versus $B(m^2 + nm)$, a factor 533. Matched compute buys 1.07×10^6 particles and 0.0089 nats, 32× better; a locally-optimal proposal at B=2000 gives 0.114 nats, 2.5× better at negligible cost. The chain's advantage is local, Hebbian-compatible updates without resampling or global normalisation, which neither filter provides. Particle filters and latent-variable sequence models represent uncertainty over time (Doucet et al., 2001; Chung et al., 2015); our diagonal chains instead use local, Hebbian-compatible updates without population-wide normalisation or resampling.

D FINITE-STEP THEORY AND ANISOTROPY CONTROLS

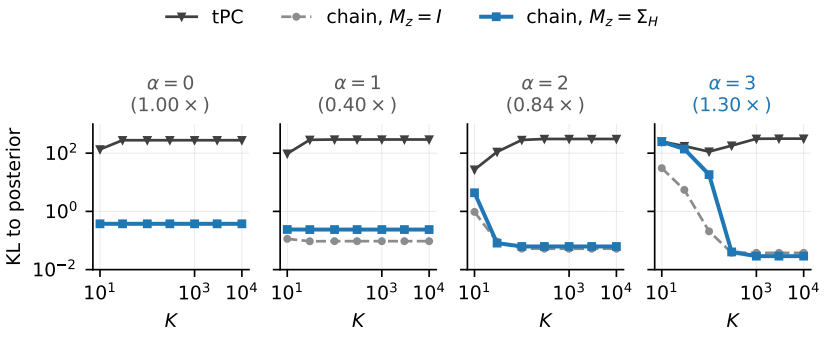

Figure 5: **Exact KL to the linear-Gaussian posterior against chain length** at four anisotropies, for deterministic tPC and Langevin with M_z=I and M_z=Σ_H. Brackets give $\mathrm{KL}_I/\mathrm{KL}_{\Sigma_H}$ at the largest budget.

In Figure 5 deterministic tPC sits at a floor set by its collapsed covariance; the preconditioner effect is small and does not order monotonically in α, which Figure 6 isolates. As reported in §5.1, after

stationarity sampled KL tracks free energy to within 0.037 nats (Figure 2); deterministic tPC stays at the jitter floor (278–316; the true value is unbounded) because its covariance collapses, whereas both samplers fall below 0.4 (Figure 5).

The pilot of the local per-direction schedule. Let s_i be the running standard deviation of unit i's own activity within its own chain, computed from exponential moving averages of z_i and z_i^2 (window $\tau{=}K/8$ steps) over the second half of a K-step pilot run of that chain with $M_z{=}I$, then held fixed; nothing is shared across chains. The pilot doubles the per-position cost (synthetic runs start it from a posterior draw; on video the chain's own run at the previous frame serves as pilot). In the eigenbasis $h_i^{-1/2}$ is the posterior standard deviation along direction i. Thus (13) uses the standard deviation along each eigendirection, and (15) each unit's own estimate s_i along its own axis.

D.1 Direction-wise derivation of Theorem 1

The schedule in the family $M_z \propto H^{-p}$. It is the $p{=}\frac{1}{2}$ member of $M_z \propto H^{-p}$; $M_z{=}I$ is $p{=}0$ and $M_z{=}\Sigma_H$ need not belong (at isotropy it rescales $M_z{=}I$).

Hypotheses in full. Theorem 1 assumes the linear-Gaussian model with $f = \mathrm{id}$ and a constant SPD M with $MH = HM$, so that H and M share an orthonormal eigenbasis; nothing further is required of Σ_H, Σ_O or W_F. Assume $0 < \eta_i h_i < 1$, so $a_i = 1 - \eta_i h_i \in (0, 1)$. Also assume q_0 has finite second moments independent of the injected noise. Here $\bar{z}$ is the posterior mean and m_i is the i-th eigenvalue of M — unrelated to the latent dimension m.

A shared-noise coupling proves the bound. Summing $Y_{i,k+1} = a_i Y_{i,k} + \sqrt{2\eta_i}\,\xi_{i,k}$ with $a_i = 1{-}\eta_i h_i$ gives $Y_{i,K} \overset{d}{=} a_i^K Y_{i,0} + \sqrt{v_i(1 - a_i^{2K})}\,\zeta_i$, with standard normal ζ_i independent of Y_0. Coupling to $h_i^{-1/2}\zeta_i$ cancels cross terms and gives (11), since $\mathcal{W}_2^2$ minimises coupling cost; equality holds for point-mass q_0. The split $(b-d)^2 \le 2b^2 + 2d^2$ separates stationary bias $b_i = \sqrt{v_i} - h_i^{-1/2}$ from the transient, as in §4.3.

In the common eigenbasis of H and M, write $H = \mathrm{diag}(h_i)$ and $M = \mathrm{diag}(m_i)$, and define the effective step $\eta_i := \eta m_i$. For the centred coordinate $y_{i,k} := z_{i,k} - \bar{z}_i$, Euler–Maruyama becomes

$$y_{i,k+1} = (1 - \eta_i h_i) y_{i,k} + \sqrt{2\eta_i}\,\xi_{i,k}, \qquad \xi_{i,k} \sim \mathcal{N}(0, 1). \tag{16}$$

Thus

$$\mathrm{Var}(y_{i,K}) = (1 - \eta_i h_i)^{2K} \mathrm{Var}(y_{i,0}) + 2\eta_i \frac{1 - (1 - \eta_i h_i)^{2K}}{1 - (1 - \eta_i h_i)^2}. \tag{17}$$

The target variance is h_i^{-1}, and (17) is exactly the variance appearing inside the width deficit of (11), since $\mathrm{Var}(y_{i,K}) = v_i(1 - a_i^{2K})$ when $\mathrm{Var}(y_{i,0}) = 0$.

D.2 Order of the stationary bias in Proposition 3

In particular $C_I = \frac{1}{16}\mathrm{tr}(H^2)$ and, for $\Sigma_H = \sigma_H^2 I$, $C_{\Sigma_H} = \frac{1}{16}\sigma_H^4 \mathrm{tr}(H^2)$, a ratio that reflects step size, not geometry. For a point-mass start $D_{\mathrm{KL}}(q_0 \| q_\infty) = \infty$; the transient term of (12) then applies from the first noisy step, whose law is Gaussian with covariance $2\eta M$. The chain's own stationary law is Gaussian with variance v_i per direction, so with $\delta_i := \eta_i h_i / 2$ and $h_i v_i = (1 - \delta_i)^{-1}$,

$$D_{\mathrm{KL}}\big(q_\infty^{\eta, M} \,\|\, q_\infty\big) = \tfrac{1}{2}\sum_i \big[h_i v_i - 1 - \log(h_i v_i)\big] = \tfrac{1}{2}\sum_i \Big[\tfrac{\delta_i}{1-\delta_i} + \log(1-\delta_i)\Big] = \tfrac{1}{4}\sum_i \delta_i^2 + O(\delta_i^3),$$

The leading Gaussian stationary bias is $O(\eta^2)$, namely $\frac{1}{16}\sum_i \eta_i^2 h_i^2 = \frac{\eta^2}{16}\mathrm{tr}\big((MH)^2\big)$; the $O(\eta)$ terms cancel. The same term holds for any constant SPD M, commuting with H or not. The stationary covariance solves $V = AVA^\top + 2\eta M$ with $A = I - \eta MH$, whose expansion is $V = H^{-1} + \frac{\eta}{2}M + O(\eta^2)$, so $E := HV - I = \frac{\eta}{2}HM + O(\eta^2)$ and $D_{\mathrm{KL}} = \frac{1}{2}[\mathrm{tr}E - \log\det(I{+}E)] = \frac{1}{4}\mathrm{tr}(E^2) + O(\eta^3) = \frac{\eta^2}{16}\mathrm{tr}((MH)^2) + O(\eta^3)$. For $h = (0.3, 1, 4, 17)$, $M = I$, exact/predicted KL is 1.934×10^{-5}/1.913×10^{-5} at $\eta = 10^{-3}$ (1.1%), and 1.9151×10^{-7}/1.9131×10^{-7} at $\eta = 10^{-4}$ (0.1%). The update depends only on ηM_z, so $(M_z, \eta) \mapsto (\gamma M_z, \eta/\gamma)$ preserves iterates and this expression to eight significant figures for $\gamma \in \{1, 2, 10, 100\}$. A linear-in-$\eta$ form instead varies 100-fold.

Small-step expansion of the stationary bias. Writing $b_i := \sqrt{v_i} - h_i^{-1/2}$ for the stationary part of the width deficit,

$$b_i^2 = \frac{1}{h_i}\left[(1-\eta_i h_i/2)^{-1/2} - 1\right]^2 = \frac{\eta_i^2 h_i}{16}\big[1 + O(\eta_i h_i)\big], \tag{18}$$

i.e. $b_i = \eta_i\sqrt{h_i}/4 + O(\eta_i^2 h_i^{3/2})$, which is the form quoted in §4.3. Setting $\eta_i = \eta_i^\star = c/\sqrt{h_i K}$ makes $b_i^2 = c^2/(16K)$ in every direction, up to a relative factor $1 + O(c\sqrt{h_i/K})$, provided $c^2 h_{\max} < K$ (so $\eta_i h_i < 1$).

The $K^{-1/2}$ choice makes it $O(K^{-1})$; it is a design choice rather than implied by equalisation, which fixes only $\eta_i \propto h_i^{-1/2}$.

Remark 1 (Equal bias is not optimality)**.** Equation (13) equalises the leading *stationary* $\mathcal{W}_2$ width bias; it does not minimise (11), whose transients a smaller step makes worse, nor KL, whose per-direction bias is equalised by $M_z \propto H^{-1}$.

Why equal width bias is not equal KL bias (Remark 1). The schedule equalises $\mathcal{W}_2$ width bias, but the per-direction KL bias $(\eta_i h_i)^2/16$ becomes $c^2 h_i/(16K) \propto h_i$ under (13); equalising it needs $\eta_i h_i$ constant, i.e. $M_z \propto H^{-1}$.

The per-direction schedule is a constant preconditioner. Running (13) in the eigenbasis of H with its FDT-paired noise gives, in the original basis,

$$z^+ = z - U\mathrm{diag}(\eta_i^\star)U^\top \nabla F_\mu + U\mathrm{diag}(\sqrt{2\eta_i^\star})\,\xi,$$

This is (6) with $\eta M_z = U\mathrm{diag}(\eta_i^\star)U^\top = (c/\sqrt{K})H^{-1/2}$, hence (14). These parameterisations define the same chain, rather than merely agreeing numerically; closed-form moments match to $< 10^{-15}$.

D.3 Oracle-tuned preconditioners

Remark 2 (What an aggregate constant cannot say)**.** Evaluated at each arm's own step size, $C_M\eta^2$ tracks the measured ordering closely (§5.2). What it cannot express is *where* a finite budget was spent. It sums directions with different contraction and discretisation scales into one number, so it is silent about which direction has not yet left its initialisation. That is the quantity Theorem 1 isolates, and it is what a per-direction step size has to be chosen against.

At a shared constant c=0.4 the per-direction schedule appears to win at every anisotropy by one to three orders of magnitude; oracle tuning removes essentially all of that apparent advantage (Figure 7; Figure 3 uses c=0.2).

Table 6: **Aggregate criterion versus measured KL ratio.** Proposition 3 evaluated at common η (C_M) and at each arm's own step $\eta_M = 0.4/\lambda_{\max}(M^{1/2}HM^{1/2})$ ($\eta_M^2 C_M$); m=32, n=64, 5 seeds; measured column from the K=10^4 sweep of Figure 6.

α	measured $\mathrm{KL}_I/\mathrm{KL}_{\Sigma_H}$	C_I/C_{Σ_H} (common η)	$\eta_I^2 C_I/\eta_{\Sigma_H}^2 C_{\Sigma_H}$	error
0.0	1.000	100	**1.000**	0.0%
0.5	0.505	39.6	**0.521**	3.2%
1.0	0.412	20.1	**0.402**	2.4%
1.5	0.536	10.2	**0.486**	9.3%
2.0	0.825	3.79	**0.751**	8.9%
2.5	1.055	0.818	**1.021**	3.2%
3.0	1.122	0.118	**1.143**	1.9%
correlation with the measurement		0.029	**0.993**	

Scored at each arm's own step, the aggregate criterion predicts the measurement (Table 6); the apparent mis-ranking at common η is a step-size artifact. The measured ratio is U-shaped in α, and its three K curves coincide because the contraction transient has decayed by K=10^3 (Figure 6).

The common-η ratio 100 is not an aggregate mis-ranking. At α=0, $\Sigma_H = \sigma_H^2 I$, so M_z=Σ_H equals M_z=I at ten times the step, giving ratio 1. Remark 2 survives, because an accurate aggregate constant cannot identify which direction exhausts its budget.

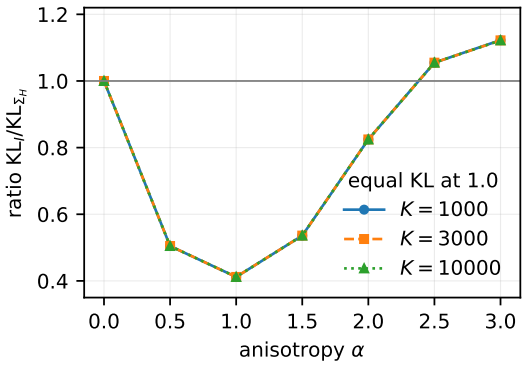

Figure 6: **Measured** $\mathrm{KL}_I/\mathrm{KL}_{\Sigma_H}$ **against anisotropy** for three chain lengths K; m=32, 30 seeds, exact KL.

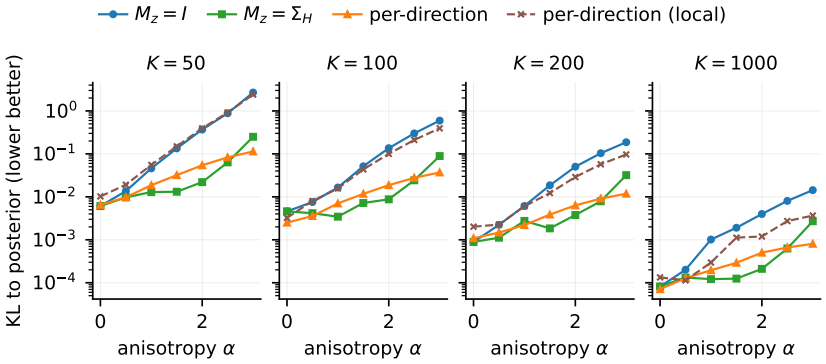

Figure 7: **Each arm at its own best step-size constant.** Exact KL to the closed-form posterior against α, one panel per budget K; each arm takes the best c of an 18-point grid in $[10^{-3}, 0.99]$ (m=32, n=64, 5 seeds). The dashed line is the local per-direction schedule (15) with 32 chains, scored by the mean per-chain KL.

Figure 7, not Figure 3, ranks the methods, because Figure 3 fixes one constant for all arms so each comparison moves a single knob, whereas here the constant is free. No arm wins everywhere. Tuned M_z=Σ_H is best in 19 of the 28 cells (the per-direction schedule in 8, its local version in 1) and beats the schedule at every $1.5 \le \alpha \le 2.5$ and every budget, by 1.1–2.5×; at α=3 the schedule is best at every budget, by 2.2–3.4× over M_z=Σ_H. The per-direction schedule has the monotone dependence on α consistent with Theorem 1 (the flattest curve in every panel) and an advantage over M_z=I growing from 0.8–1.8× at α=0 to 15.8–23.6× at α=3. The local version (dashed) uses each chain's own units' running standard deviations from a K-step M_z=I pilot (window $K/8$). Windows $\tau \in [K/8, K/2]$ form a plateau (τ=$K/32$ fails at K=100), a $4K$ pilot lifts K=100 to 71–72% of the schedule's gain, and at $\alpha \le 0.5$ the recovered fraction is erratic. As reported in §5.2, tuned M_z=Σ_H wins 19 of 28 cells and beats the per-direction schedule at every $1.5 \le \alpha \le 2.5$; the schedule wins at α=3 for every budget; its advantage over M_z=I grows monotonically from 0.8–1.8× at α=0 to 15.8–23.6× at α=3, so the schedule is diagnostic, identifying $M_z \propto H^{-1/2}$; its local version recovers 77–96% of its gain over M_z=I at $\alpha \ge 1$ (long chains). The criterion $\eta_M^2 C_M$ has median error 3.2% and correlation 0.993 with the measurement, versus 0.029 at common η (Table 6).

The Gaussian moment KL of Table 7 is $D_{\mathrm{KL}}(\hat{q} \,\|\, \pi_{\mathrm{GN}})$, with $\hat{q}$ the Gaussian carrying the arm's post-burn-in sample mean and covariance; it is blind to any discrepancy beyond second moments. Only KL is emboldened because an all-−1 frame already scores 14.40 dB and SSIM 0.788 here

Table 7: **Finite-step arms on a trained Moving MNIST model.** m=480, n=4096, B=32, c=0.2, one model, one seed. KL is the Gaussian moment KL (nats) to a fixed Gauss–Newton Gaussian at the deterministic filter's mode; PSNR/SSIM average ten rolled-out frames; bold marks the lowest KL.

	$K = 10^3$			$K = 3\times10^3$			$K = 10^4$		
arm	KL↓	PSNR↑	SSIM↑	KL↓	PSNR↑	SSIM↑	KL↓	PSNR↑	SSIM↑
$M_z{=}I$, η	324	14.56	0.617	231	14.56	0.616	147	14.56	0.616
$M_z{=}\Sigma_H$, η	16041	14.36	0.630	9129	14.47	0.634	2502	14.60	0.632
$M_z{=}\Sigma_H$ rescaled, η	324	14.56	0.617	231	14.56	0.616	147	14.56	0.616
$M_z{=}\Sigma_H$, $\eta\sigma_H^2$	134022	7.72	0.275	73201	10.46	0.556	26447	14.31	0.659
per-direction, $\eta_i{=}c/\sqrt{h_i K}$	**122**	14.56	0.616	**108**	14.56	0.616	**103**	14.56	0.616
per-direction (local), η	575	14.57	0.622	301	14.56	0.617	189	14.56	0.617

(Table 22), so no PSNR/SSIM difference is a quality ordering. Row 3 is row 1 to every printed digit. Since $\Sigma_H = \sigma_H^2 I$ is a scalar, rescaling $M_z{=}\Sigma_H$ to the stability limit of $M_z{=}I$ reproduces it to float32 rounding (relative 10^{-8}); rows 2 and 4 differ from row 1 only in effective step size ($\times10^{-2}$ and $\times10^{-4}$). Row 5 is the one genuine geometric change; row 6, the local per-direction schedule (15) (each unit's running standard deviation within its own chain at the previous frame, window $K/8$), is diagonal and worse than row 1 at every K. Over $c \in \{0.05, 0.2, 0.8\}$ its best KL ($274/174/120$) never beats the best $M_z{=}I$ ($205/139/112$), which $M_z{=}\Sigma_H$ equals exactly, being a scalar here; the per-direction schedule's best is $122/108/103$. Table 8 decomposes the KL column.

The geometry is exploitable. The Hessian $H = \Sigma_H^{-1} + J_F^\top \Sigma_O^{-1} J_F$, with J_F the emission Jacobian at the mode, has emission-induced condition number 144. Thus the per-direction schedule has the lowest KL at every K (Table 7); Σ_H misses the anisotropy in W_F. Yet per-unit posterior standard deviations differ by under $2\times$, which is why the local per-direction schedule recovers none of the schedule's gain here (-4 to -9% at $c{=}0.2$), nor do the exact per-unit posterior standard deviations (-4 to -8%). A diagonal rule needs axis-aligned anisotropy, and off-axis anisotropy needs cross-unit information.

Scaling and chain-length dependence of the local per-direction schedule. Our runs keep $\eta_i \propto s_i$ but set the scale by stability, like the other arms ($M_z = \mathrm{diag}(s)/s_{\max}$, $\eta = c/\lambda_{\max}(H)$ at shared c; $c/\lambda_{\max}(M_z H)$ when tuned). Scored by the mean of per-chain KLs, it recovers, tuned at $\alpha \geq 1$, 49–88% of the per-direction schedule's KL gain over $M_z{=}I$ at $K{=}10^3$ but 3–51% at $K \in \{100, 200\}$ and -34 to 11% at $K{=}50$; at fixed $c{=}0.2$ and $K \in [10^3, 10^4]$, 77–96% (Figures 7, 3). Short pilots give too noisy a variance estimate; a $4K$ pilot lifts $K{=}100$ to about 72%.

Table 8: **Decomposition of the Gaussian moment KL on a target with known law.** Nats; B, burn-in and estimator as Table 7. *cold* is the run as performed, *stat.* a stationary start, *iid* the estimation floor at that sample count, and *anal.* the exact stationary discretisation KL of fixed-step ULA.

	$K = 10^3$				$K = 3\times10^3$				$K = 10^4$			
arm	cold	stat.	iid	anal.	cold	stat.	iid	anal.	cold	stat.	iid	anal.
$M_z{=}I$, η	291	293	3.7	0.01	137	137	1.2	0.01	47	47	0.4	0.01
$M_z{=}\Sigma_H$, η	1110	1265	3.7	< 0.01	899	1023	1.2	< 0.01	691	754	0.4	< 0.01
$M_z{=}\Sigma_H$, $\eta\sigma_H^2$	2111	2233	3.7	< 0.01	1890	2029	1.2	< 0.01	1622	1775	0.4	< 0.01
per-direction ($M_z \propto H^{-1/2}$)	**22**	22	3.7	15.6	**8.2**	8.2	1.2	4.1	**3.4**	3.4	0.4	1.1

What the KL column is made of. Neither estimation nor transient explains the KL column (Table 8). The iid floor is 0.4–3.7 nats, the cold and stationary runs agree to within 1% for $M_z{=}I$ and exactly for the per-direction schedule, and the analytic bias is negligible except for the per-direction schedule. What is left is autocorrelation, i.e. effective sample size. The KL column measures mixing efficiency, not posterior distance. An iid floor of 3.7 nats at 16,000 draws makes identity's 291 nats roughly 200 effective samples. Its autocorrelation time is order 10^2, consistent with the identity arm's smallest relaxation number $u_{\min} = \min_i \eta_i h_i K = 1.39$ at $K{=}10^3$; the per-direction schedule stays within 6–9 times the iid floor.

The Gaussian control reproduces no video arm. Identity gives $291/137/47$ versus $324/231/147$ (90%, 59%, 32%), and per-direction schedule $22/8.2/3.4$ versus $122/108/103$. The residual suggests

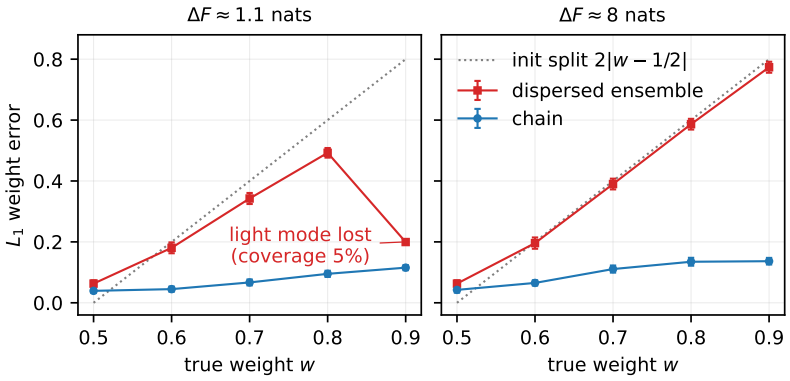

Figure 8: **Mode-weight error against true weight.** L_1 error of the recovered mode weights for the dispersed ensemble (red) and the chain (blue), at barrier heights $\Delta F \approx 1.1$ and ≈ 8 nats; the dotted line is the init-split line $2|w - 1/2|$. 20 seeds; error bars are standard errors.

nonlinear-posterior versus fixed-Gauss–Newton mismatch, greatest for the best-mixing chain; it is not identified because KL differences are nonadditive and no nonlinear reference is available.

E Multimodal calibration and occlusion controls

E.1 Proof of Proposition 4

Proposition 4 (Weight calibration separates samplers from ensembles)**.** *For a mixture posterior $\pi = \sum_k w_k \phi_k$ with component weights w_k and basins $\mathcal{B}_k$, deterministic descent from initialisations ρ assigns basin mass $\rho(\mathcal{B}_k)$, whereas the FDT-paired* diffusion*'s occupation mass converges to $\pi(\mathcal{B}_k) = w_k + \epsilon_k \approx w_k$, where the tail term ϵ_k is the component mass lying across basin boundaries (at most of order $\Phi(-\Delta/2)$ for two equal-covariance components Δ posterior standard deviations apart). The discretised chain adds fixed-step ULA bias.*

Proof.

Let $\pi(z) \propto \sum_k w_k \phi_k(z)$ with basins $\mathcal{B}_k$ (noise-free descent from $\mathcal{B}_k$ reaches mode k), and let ρ be the initialisation law.

Deterministic ensembles. Every start in $\mathcal{B}_k$ descends to mode k, so the ensemble reports $\rho(\mathcal{B}_k)$ exactly (boundaries have measure zero); w_k enters only through where the boundaries lie. More initial states refine initialisation mass, not w_k, unless $\rho(\mathcal{B}_k) = w_k$ already.

Langevin chains. Proposition 1 and ergodicity give occupation fraction $\pi(\mathcal{B}_k)$ as $K \to \infty$, and $\pi(\mathcal{B}_k) = \sum_j w_j \phi_j(\mathcal{B}_k) = w_k + \epsilon_k$ with tail term $\epsilon_k = \sum_{j \neq k} w_j \phi_j(\mathcal{B}_k) - w_k \phi_k(\mathcal{B}_k^c)$, the component mass across basin boundaries; for two equal-covariance Gaussian components Δ posterior standard deviations apart it is at most of order $\Phi(-\Delta/2)$ (0.02 at Δ=4). Sampler mass is thus consistent, unlike dispersed-ensemble mass. □

Finite-budget interpretation. The claim is asymptotic mass, not faster discovery. The target must be representable and barriers crossable within K. Otherwise first passage and initialisation dominate; the synthetic experiment therefore starts near the saddle and reports dispersed starts separately.

In Figure 8 the dispersed ensemble lands on the init-split line, so its mass reports the initialisation, not the posterior. The chain stays near zero at both barriers, with residual bias growing mildly with the barrier. At the low barrier and $w = 0.9$ the light mode's deterministic basin vanishes and MAP reaches it in only 5% of runs (its L_1 point reflects mass 1 on the heavy mode). As reported in §5.3, with B=200 chains, Langevin tPC reduces mixture-weight error by 1.5–5.7×, to 0.04–0.14; it finds both basins in all ten settings, the dispersed ensemble in nine.

E.2 Occluded pachinko drop as a representational control

Committing to one branch (Figure 4) needs a transition that represents the exact $\mathrm{Bernoulli}(\frac{1}{2})$ predictive law. In the pachinko drop, a ball falls under gravity onto a hidden peg, leaves it left or right with probability $\frac{1}{2}$ at 50° from vertical, and follows a parabola; the peg sits behind a horizontal bar excluded from observations and likelihood (observations stop at $t{=}9$; the bounce at $t{=}10$ is never observed; $\sigma_H{=}\sigma_O{=}0.05$, 60 epochs on 240 sequences). A single Gaussian cannot represent both futures, so we also test

$$z^\mu = W_H f(z^{\mu-1}) + b_r + \xi_{z,\mu}, \qquad r \sim \mathrm{Bernoulli}(\tfrac{1}{2}), \tag{19}$$

Branch biases b_r, indexed by r, are co-trained with W_H. The B chains together cover both branches (a $50/50$ split in Figure 4).

The *lobe-mass ratio* of a rollout, weaker over stronger rendered intensity of the two post-deflection regions (App. E.3), is near 0 for one committed continuation and near 1 for both superposed. The shape-transfer and capacity controls below were run on an earlier linear-motion version of this task (a ball moving right at constant speed, deflecting up or down; $\sigma_H{=}\sigma_O{=}0.1$). Shape transfer and capacity controls are in Appendices E.4 and E.3.

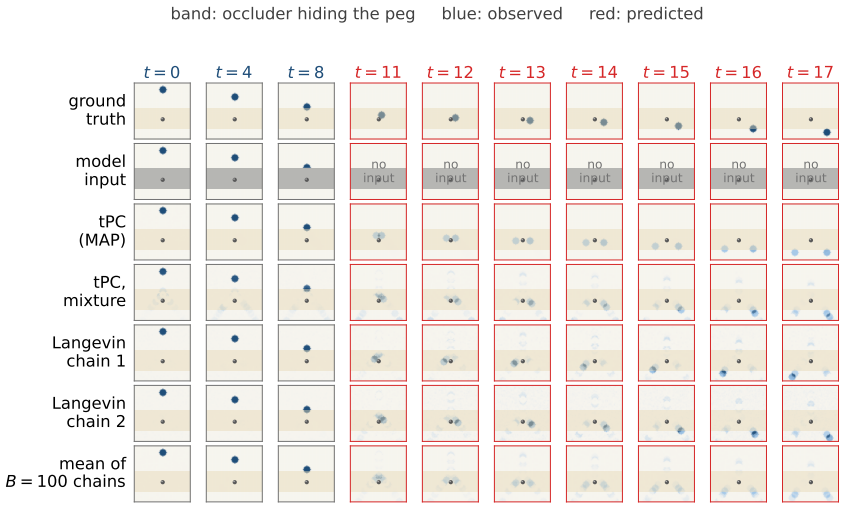

Figure 9: **Occluded pachinko drop with both deterministic baselines.** Rows show ground truth; model input (the grey band is the occluder, zeroed from the observation and dropped from the likelihood; no input after $t{=}9$); tPC with the single-Gaussian transition (1); tPC with the mixture transition (19), filtered deterministically; two Langevin chains; mean of 100 chains.

In Figure 9 the ball is entirely behind the bar in the last observed frame, so no arm can know which way it went. With the single-Gaussian transition, W_H learns the average of the two dynamics and the averaged latent decodes to both balls at half brightness. The deterministic mixture filter neither averages fully nor commits (lobe-mass ratio 0.462, versus 0.395 for a single Langevin chain and 0.913 for the mean of 100 chains). At $t{=}14$ the 100 chains split $50/50$ between peak columns 20 and 44 of 64. As reported in §5.3, the lobe-mass ratio (weaker over stronger rendered post-deflection intensity; 0 committed, 1 superposed) is 0.395 per chain and 0.913 for the chain mean, and the deterministic filter's 0.462 holds on all 20 sequences; on the energy score, a strictly proper rule, chains beat dispersed and point forecasts (7.22 versus 11.97 and 12.51; Table 10).

E.3 Why a single-Gaussian transition cannot carry the branch

Lobe-mass ratio. The rendered intensity of a predicted rollout in the weaker of the two post-deflection regions (above and below the pre-deflection path, right of the bar) divided by that in the stronger.

The single-Gaussian transition, not sampling, limits branch retention ((1)), since commitment remains 6.9–7.7% across rerun scales. Residual clustering identifies deflection with 95% accuracy at 14.6σ, but one learned transition reduces the unoccluded-future decoded branch score from 0.55 to 0.02.

Sampling can allocate mass only to representable alternatives, motivating the mixture transition of §5.3.

E.4 Appearance transfer

Training uses circles; transfer changes only appearance to squares and triangles, with trajectory and occluder fixed.

Table 9: **Appearance transfer from the trained circle to unseen shapes.** 20 sequences × 100 chains per shape, occluded protocol of §E.2. Branch agreement is the fraction of chains whose rendered continuation matches their sampled branch; recon PSNR is the filtered reconstruction of the conditioning frames against the unoccluded truth.

test shape	branch agreement ↑	KL to $\mathrm{Bern}(\frac{1}{2})$ ↓	both branches	recon PSNR (dB)
circle *(trained on)*	0.853	0.0068	100%	51.2
square *(unseen)*	0.841	0.0081	100%	22.5
triangle *(unseen)*	0.825	0.0091	100%	23.1

Branch agreement is not 1, because the occluder removes the evidence and the rollout carries transition noise, so some chains are pushed across. Recon PSNR scores the bar columns against the unoccluded truth, so it is an inpainting test, which is why the trained circle and the unseen shapes are far apart. The emission does not transfer appearance.

E.5 A proper scoring rule on a learned nonlinear temporal model

The occluded pachinko drop tests temporal sampling with $\tanh$ emission, learned W_H and branch biases, and ten rollout frames; the roughly fivefold UCI gain uses $W_H = 0$. Branch frequency alone relies on the designed $\mathrm{Bernoulli}(\frac{1}{2})$ law.

Energy score $\mathrm{ES}(F, y) = \mathbb{E}\|X - y\| - \frac{1}{2}\mathbb{E}\|X - X'\|$ and per-pixel CRPS (Matheson & Winkler, 1976) test predictive laws without knowing the truth's distribution (Gneiting & Raftery, 2007). Held-out arms use $B{=}100$, namely sampled branches, deterministic starts dispersed at $10\sigma_H$ (Proposition 4), and the point rollout replicated B times.

Table 10: **Energy score and CRPS on the occluded pachinko drop** at the last rollout frame; 20 held-out sequences per shape, $B{=}100$ per arm, lower is better. Trained on the circle only (Appendix E.4).

	energy score ↓			CRPS ↓		
test shape	Langevin	dispersed	point	Langevin	dispersed	point
circle	**6.02**	11.16	11.28	**0.0372**	0.1471	0.0730
square	**8.60**	13.85	13.84	**0.0444**	0.1568	0.0700
triangle	**7.05**	10.89	12.41	**0.0458**	0.1431	0.1084
pooled	**7.22**	11.97	12.51	**0.0425**	0.1490	0.0838

Branching is built in, not discovered (§5.3); sampling cannot recover unrepresentable modes. The dispersed baseline shares support, so the proper-score comparison tests branch weights (Proposition 4) without presupposing them.

F Scope of the mass claim, and two benchmarks it is tested on

The twelve-for-twelve UCI mass win (§5.4) and equilibration (Appendix I.1) hold within their tested regimes, not universally.

F.1 Between-mode mass is recovered inside a window, not everywhere

At fixed K, exponentially slower barrier crossings eventually leave chain mass determined by initialisation, the same failure as deterministic ensembles.

Setup. Two isotropic Gaussian components with m=8, sd $= 0.25$, separation $\Delta \cdot$ sd, weights $(w_0, 1 - w_0)$. Both arms return 20,000 points from branches initialised at $1/2$, as a mixture transition supplies Algorithm 1; chains use B=200, K=2000, $\eta = 0.004$. Eight seeds.

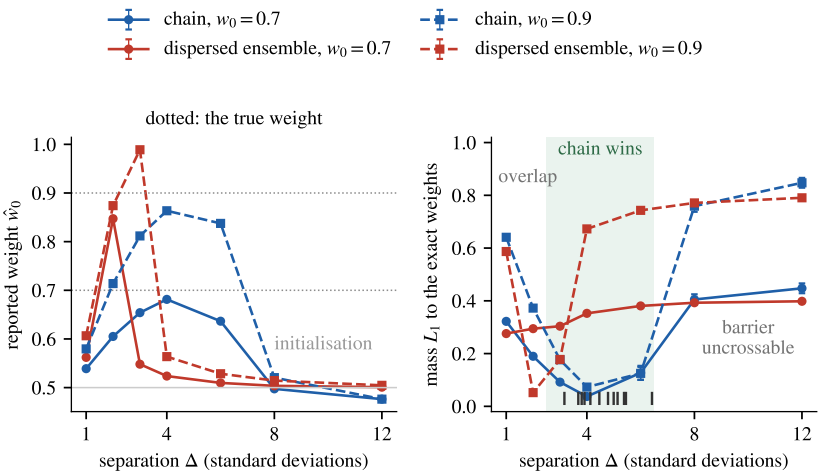

Figure 10: **Between-mode mass recovery against mode separation.** The left panel shows the reported weight of the heavy component (dotted line the truth, grey the $1/2$ both arms start from). The right panel shows mass L_1 to the exact weights, $\pm$ s.e.m. over 8 seeds; foot ticks mark median separations of the twelve UCI posteriors of Table 2.

The chain recovers between-mode mass in a window of 3–6 posterior standard deviations and outside it fails as the ensemble does. Below $\Delta \approx 2$ the components overlap and the responsibilities are not identified; above $\Delta \approx 8$ the barrier is uncrossable at this budget and the chain returns its own initialisation, 0.497 and 0.521 at Δ=8 against true weights of 0.70 and 0.90.

Recovery peaks at 4 within $\Delta \in [3, 6]$ sd (Figure 10). At w_0=0.70, Δ=4, chain/ensemble weights are 0.681/0.524, mass L_1 0.037/0.353 (9.5×); at w_0=0.90, Δ=4, 0.863/0.564, 0.073/0.673 (9.2×). At Δ=12, both retain initialisation and are uninformative.

UCI mostly lies in the recoverable window, since separations of the two heaviest posterior components have dataset medians 3.16–6.42, pooled median 4.63 versus optimum 4. Of held-out inputs, 84% lie in $[3, 6]$; nine of twelve datasets exceed 0.8. Abalone's median 6.42 and 30% inside accompany the smallest gain in Table 2.

Budget moves the upper edge only as $\sqrt{\log K}$. A Kramers estimate places it near $\Delta \approx \sqrt{8 \ln(K\eta/\sigma^2)} \approx 6.2$ here (Appendix G); recovery requires identified, crossable modes. The switching benchmark's chain mass error at σ_O=0.25, where the chain loses to the ensemble (§F.2), is flat in budget (mass L_1 0.161, 0.167, 0.168, 0.160 at $K \in \{300, 10^3, 3 \times 10^3, 10^4\}$, $\pm$0.026; 5 seeds, 5 positions), so that loss is not a finite-K transient that longer chains repair; we have not measured the per-chain separation or an i.i.d. floor for this statistic, so we do not attribute it further. Deterministic mass instead tracks basin volume. At w_0=0.90, it gives 0.874 for Δ=2 and 0.529 for Δ=6.

F.2 A temporal benchmark whose per-step posterior is exactly a mixture

Unlike UCI (W_H=0), this branching temporal model retains an exact conditional target for Algorithm 1. The switching linear system is $z^\mu = A_{s_\mu} z^{\mu-1} + \xi_z$ with $s_\mu \sim$ Bernoulli$(1/2)$ over two rotations, and $x^\mu = Cz^\mu + \xi_x$; m=8, n=16, P=12, σ_H=0.35, ten seeds. Freezing the parent as Algorithm 1 does makes $p(z^\mu \mid x^\mu, \hat{z}^{\mu-1})$ a two-component Gaussian mixture with $S =$

$(\Sigma_H^{-1} + C^\top \Sigma_O^{-1} C)^{-1}$, $m_s = S(\Sigma_H^{-1} A_s \hat{z} + C^\top \Sigma_O^{-1} x)$ and $w_s \propto \pi_s N(x; C A_s \hat{z}, C \Sigma_H C^\top + \Sigma_O)$. Varying σ_O moves w away from π.

Table 11: **Chain versus dispersed ensemble on a switching linear system across observation noise.** Exact two-component posterior, 10 seeds $\times$ 11 positions, equal numbers of returned points. Each chain and descent propagates its own final iterate; the reference is the mixture of per-chain exact conditionals; $\max_s w_s$ averaged over chains.

σ_O	$\max_s w_s$	SW_2^2 chain	SW_2^2 ens.	ratio	mass L_1 chain	mass L_1 ens.	ratio
0.25	0.919	**0.0003**	0.0538	194×	0.146	**0.121**	0.82×
0.50	0.848	**0.0008**	0.0780	98×	**0.071**	0.212	2.99×
1.00	0.749	**0.0023**	0.1128	48×	**0.086**	0.167	1.95×
2.00	0.651	**0.0032**	0.1211	38×	0.078	**0.037**	0.47×
4.00	0.581	**0.0031**	0.1221	40×	0.069	**0.010**	0.14×

As reported in §5.5, the chain is 38–194× closer in SW_2^2 at every noise level, and closer in mass L_1 only at $\sigma_O \in \{0.5, 1\}$. The reference is the equal-weight mixture of the per-chain exact conditionals.

Table 11 shows a width advantage, not uniform mass recovery. Deterministic within-mode spread is 0.000, making the distributional gap large at every σ_O. Chain mass wins only at $\sigma_O \in \{0.5, 1\}$, where the observation moves the weights partly away from $1/2$. At σ_O=0.25 the observation nearly picks the branch, and at $\sigma_O \geq 2$ the per-chain weights approach $1/2$; in both regimes the ensemble is closer. This is consistent with each chain committing to one branch from its start and not hopping within K (Appendix F.1; at σ_O=0.25 mass L_1 is flat from K=300 to 10^4). The chain's branch mass is then an average over only B=100 near-binary chains, whereas the ensemble averages 15,000 independent descents.

F.3 Real data, kept temporal, with the per-step posterior still in closed form

This benchmark combines real data with temporal transitions, unlike synthetic §F.2 or static UCI (W_H=0, §5.4).

Setup. Bike-sharing (Fanaee-T & Gama, 2014) covers 17,379 consecutive hours in 2011–2012, with standardised casual/registered rentals, temperature, humidity, and wind. Working days have two commute peaks, non-working days one midday peak; `workingday` labels regimes. Ridge-fit A_0, A_1 use consecutive within-regime pairs in the first 70% (12,165 hours); pooled residual $\sigma_H = 0.445$, working-day prior $\pi = 0.685$. Evaluation hides regime labels on the last 30% and sweeps Gaussian emission noise in $x^\mu = z^\mu + \xi_x$. The frozen-parent target of Algorithm 1 is an exact Gaussian mixture. The series, transitions, noise scale, prior, and regime ambiguity are empirical; the Gaussian emission is imposed.

Table 12: **Bike-sharing series with an exact per-step mixture posterior.** 25 held-out windows $\times$ 12 positions $\times$ 5 seeds; propagation and reference as in Table 11. Coverage (%) of a nominal 90% interval on the first channel.

σ_O	$\max_s w_s$	SW_2^2 chain	SW_2^2 ens.	ratio	mass L_1 chain	mass L_1 ens.	cov_{90} chain	cov_{90} ens.
0.25	0.733	**0.0002**	0.0477	215×	**0.0017**	0.0050	**90.0**	0.0
0.50	0.710	**0.0004**	0.1113	274×	**0.0042**	0.0120	**85.7**	0.0
1.00	0.688	**0.0010**	0.1665	160×	**0.0063**	0.0096	**82.7**	0.0
2.00	0.687	**0.0017**	0.1900	115×	**0.0074**	0.0085	**82.1**	0.0

This is not a mass result. Here $\|A_0 - A_1\|_F = 0.278$, and within each chain the two components of its exact conditional lie a median 0.17, 0.25, 0.30 and 0.33 posterior standard deviations apart at the four noise levels (median over chains and positions of $\|m_{b,0} - m_{b,1}\|/\sqrt{S_{11}}$), far below §F.1. As reported in §5.5, the chain is 115–274× closer in SW_2^2, and its 90% intervals cover 90.0% at the lowest noise (82.1% at the highest) versus 0.0% for the ensemble. The 1.15–2.98× mass ratio compares near-zero errors (0.0017/0.0050 at σ_O=0.25); overlapping components barely identify

weights despite visibly different raw regimes. This limitation is in the data, not either method; mass evidence remains UCI (§5.4), median separation 4.63.

We report worsening coverage ($90.0\% \to 82.1\%$) with noise, because the target conditions on each chain's frozen parent, so parent error accumulated along the series lies outside the conditional's width. This is §C's conditional-versus-filtering gap expressed as miscalibration rather than KL. The ensemble's coverage is 0.0 at every noise level because its within-mode width is zero.

F.4 A PROBABILISTIC BASELINE THAT HAS WIDTH, AND THE READ-OUT WE SHOULD HAVE USED

Static UCI ($W_H{=}0$, §5.4) also admits sampling-based predictive coding. A width-bearing baseline tests whether gains exceed restored variance; read-out and preconditioner controls test the cited methods' differences from Algorithm 1.

Arms. Algorithm 1 with $M_z{=}I$ supplies pooled post-burn-in *chain* states (Table 2) or one final state per chain, *chain-final* (one sample per chain, as in MCPC's learning rule; MCPC's learning rule averages steady-state samples and its posterior histograms pool a long trajectory (Oliviers et al., 2024)). *Laplace* (MacKay, 1992) is $\mathcal{N}(z^*, H(z^*)^{-1})$ at MAP, using the exact Hessian of $-\log p(z \mid x)$, including responsibility covariance rather than Gauss–Newton truncation. *Ensemble* is dispersed deterministic descent; all four return $15{,}000$ points, scored as in §5.4, including responsibility term $\log w_k$.

Table 13: **Probabilistic baselines on the twelve regression datasets.** 5 seeds, 50 held-out inputs each; mean $\pm$ s.e.m. across datasets; coverage (%) targets 90.

	SW_2^2	MMD^2	mass L_1	cov_{90} (%)
chain (Algorithm 1, $M_z{=}I$)	0.094 ± 0.017	0.062 ± 0.013	0.402 ± 0.053	$\mathbf{75.7 \pm 4.7}$
chain, final-state read-out	$\mathbf{0.091 \pm 0.015}$	$\mathbf{0.048 \pm 0.010}$	$\mathbf{0.333 \pm 0.048}$	74.1 ± 4.5
Laplace at the MAP	0.354 ± 0.032	0.310 ± 0.025	1.106 ± 0.050	60.5 ± 5.9
dispersed ensemble	0.458 ± 0.098	0.214 ± 0.015	1.108 ± 0.048	41.2 ± 6.3

A single Gaussian with the right width in the wrong place (Laplace) is no better on mass than a collapsed ensemble, and both chain read-outs beat it on every dataset (Table 13). Our pooled read-out is worse, and we say so. Final states improve MMD^2 (0.048/0.062) and mass L_1 (0.333/0.402) on all twelve datasets, SW_2^2 on five. Pooling 150 consecutive states inflates counts, because $15{,}000$ states contain less information than $15{,}000$ independent draws. Table 2 retains the released-code read-out used throughout; implementations should take one sample per chain and slightly outperform what we report.

The Adam-style per-chain diagonal second-moment preconditioner (Zahid et al., 2024), in our re-implementation normalised to $\lambda_{\max}(M_z) = 1$ at matched η, K, and chain count, loses on all twelve datasets and all three distributional metrics (SW_2^2 0.123/0.094). Nearly isotropic components offer little geometry; adaptation costs mixing.

G WHEN DOES THE BRANCH REPRESENTATION COLLAPSE?

We say the representation *collapses* at position μ when the B chains carry one alternative, or a split other than the posterior's. Five mechanisms produce it; they differ in whether the model, the data, the sampler or the read-out is responsible. We use the switching notation of Appendix F.2, with branch means $A_s\hat{z}$, emission C, posterior covariance S, predictive covariance $S_x = C\Sigma_H C^\top + \Sigma_O$, and separation Δ between the two conditional components in posterior standard deviations. New numbers come from two analytic diagnostics included in the supplementary code.

(i) Representational collapse, when the model has only one branch. *Condition.* A single-Gaussian transition (1). *Why.* With $f = \mathrm{id}$ the per-step conditional is the Kalman Gaussian for every parent, so no sampler can find a second mode. With learned $\tanh$ dynamics, the Hebbian fixed point of (7) places $W_H f(\hat{z})$ at the least-squares average of the continuations, between the branches.

(ii) Unobserved directions, where nothing collapses the split and nothing corrects it. *Condition.* The branch difference is invisible to the observation, i.e. $\delta := C(A_0 - A_1)\hat{z} = 0$. This holds for occluded pixels ($\Sigma_O^{-1} = 0$) and, when $m > n$, for latent directions in the null space V of W_F. *Why.* $w_s \propto \pi_s \mathcal{N}(x; CA_s\hat{z}, S_x)$ has equal likelihood factors, so $w = \pi$. For $\Sigma_H = \sigma_H^2 I$, V is invariant under $H = \Sigma_H^{-1} + W_F^\top \Sigma_O^{-1} W_F$, and on V the posterior equals the prior (numerically to 2.3×10^{-6} at m=32, n=8; `exp_verify_surplus`). The split in these directions is therefore the transition's own. It is kept when the transition can represent it (KL to $\mathrm{Bernoulli}(\frac{1}{2})$ of 0.0068–0.0091, §E.2), and it is lost when the transition cannot, as in (i). Appendix L covers how m acts on learning.

(iii) Observational collapse, when the data pick a branch. *Condition.* $D^2 := \delta^\top S_x^{-1} \delta \gg 1$. *Why.* When branch 0 generated x, the log-odds $\log(w_0/w_1) = \log(\pi_0/\pi_1) + \delta^\top S_x^{-1} r_0 + \frac{1}{2} D^2$, with $r_0 = x - CA_0\hat{z} \sim \mathcal{N}(0, S_x)$, has mean $D^2/2$ and variance D^2. The posterior becomes unimodal as a matter of fact. *Evidence.* Along the switching benchmark run with chain-mean parents, the median D^2 is 10.1, 2.8, 0.22 and 0.012 at $\sigma_O = 0.25$, 0.5, 1 and 2, and the mean $\max_s w_s$ is 0.912, 0.802, 0.622 and 0.525; with per-chain parents (Table 11) it is 0.919, 0.848, 0.749 and 0.651. The chain follows this reweighting to within 0.012 of the exact mass (Table 1); the deterministic protocols keep their initial split. Here collapse is the correct behaviour, and keeping both branches would be the error.

(iv) Mixing failure, when the sampler cannot cross the barrier. *Condition.* Too few barrier crossings within the budget. For two equal-covariance components with equal weights, the barrier along the separating axis is $\Delta F \approx \Delta^2/8 - \log 2$ for large Δ. With $M_z = I$, Kramers' rate gives an expected $N \approx (K\eta/\sigma^2)(\omega/2\pi)e^{-\Delta F}$ crossings, where σ is the posterior standard deviation and ω the geometric mean of the minimum and saddle curvatures in standard-deviation units. Mass stays at its initial split when $N \lesssim 1$, that is, when $\Delta \gtrsim \sqrt{8\ln(K\eta/\sigma^2)}$ up to an $O(1)$ term inside the logarithm. The edge therefore grows only as $\sqrt{\log K}$. *Evidence.* Figure 10 uses $K\eta/\sigma^2 = 128$, so the bound is $\Delta \approx 6.2$. At $w_0 \in \{0.7, 0.9\}$ we find $N = 2.3$–6.0 at Δ=4 (mass L_1 0.037–0.073), $N = 0.39$–0.83 at Δ=6 (0.125–0.127) and $N = 0.017$–0.034 at Δ=8. At Δ=8 the chain returns 0.497 and 0.521 against true weights of 0.70 and 0.90. In one dimension, essentially no chain crosses at K=10, and opposite starts agree within 0.015 at K=10^3 (Table 19). Of the five mechanisms, this is the only failure of the sampler, and deterministic ensembles share it (Proposition 4).

(v) Propagation collapse, when a single parent erases the split. *Condition.* One parent (the chain mean or the MAP point) is fed forward in place of each chain's own iterate. *Why.* The branch predictions differ by $(A_0 - A_1)\hat{z}$. The mean of a split ensemble lies between the branches. Under weak observations it also contracts towards the prior mean, and there the branch predictions coincide. *Evidence.* On the bounce, weaker-to-stronger lobe mass is 0.395 for a single chain and 0.913, a superposition, for the mean over 100 chains (§E.2). Run with chain-mean parents, the switching benchmark shows it. With the true parent the median separation is 4.7–4.9 at $\sigma_O \in [1, 4]$; with the mean parent it falls to 1.45, 0.68 and 0.35, and both arms' mass errors vanish at high σ_O. The reported switching and bike-sharing benchmarks (Appendices F.2 and F.3) propagate each chain's own iterate. Algorithm 1 propagates each chain's own iterate and so avoids this mechanism.

Summary. Langevin tPC is worth using when the transition can represent the alternatives, the observations leave them only partly resolved (D^2 of order one), and their separation lies below $\sqrt{8\ln(K\eta/\sigma^2)}$ posterior standard deviations. Outside that range the remedy is a richer transition or nonlocal moves, not a larger K.

H Scaling with the number of independent ambiguities

A scene with N occluded objects, each of which may or may not have deflected, has 2^N explanations. We ask how each inference rule scales as N, the latent dimension m, the budget K and the number of chains B grow, and when the ambiguities are coupled, on a per-step conditional whose 2^N mode weights are known exactly. Write $\mathrm{BC}(\rho, w) = \sqrt{\rho w} + \sqrt{(1-\rho)(1-w)}$ for the Bhattacharyya coefficient of two Bernoulli laws.

Proposition 5 (Independent ambiguities)**.** *Let the per-step conditional factorise over N blocks, $p = \bigotimes_i p_i$, each p_i with two separated basins of weights $(w_i, 1-w_i)$, and let M_z be block-diagonal.*

Let q be the law a method reports over the 2^N basin configurations and ρ_i its mass on block i's first basin.

(i) Chains. *The chain law at every step factorises, and* $\mathrm{TV}(q,p) \le \sqrt{1-\prod_i \mathrm{BC}(\rho_i,w_i)^2} \le \sqrt{2\sum_i h_i^2}$ *with* $h_i^2 = 1-\mathrm{BC}(\rho_i,w_i)$*, so the joint error grows as* $\sqrt{N}$ *times the per-block error, and the relaxation time is set by the slowest block, so the steps needed grow as* $\log N$.

(ii) Deterministic ensembles. *Descent from a product initialisation ρ reports the initialisation masses* $\rho_i = \rho(B_i)$ *(Proposition 4), and* $\mathrm{TV}(q,p) \ge 1-\prod_i \mathrm{BC}(\rho_i,w_i) \to 1$ *exponentially in N unless* $\rho_i = w_i$*, for any number of initial states.*

(iii) One deterministic run *commits to one configuration $\hat{c}$, so* $\mathrm{TV}(q,p) = 1-p(\hat{c}) \ge 1-\prod_i \max(w_i, 1-w_i)$.

Both bounds are product identities for the Bhattacharyya coefficient (Tsybakov, 2009, §2.4); the content is that the two classes move in opposite directions as N grows. Chains inherit the difficulty of the hardest block, which the window of Appendix F.1 bounds (Bovier et al., 2004; Menz & Schlichting, 2014), not of the 2^N configurations. Chen et al. (2026) show that ULA's bias delocalises across the coordinates of Gaussian and sparsely interacting strongly log-concave targets, though not universally; (i) is the corresponding statement for the joint law of a multimodal product. Multi-start deterministic methods fail on mixtures with a probability that grows with the number of components (Jin et al., 2016); (ii) and (iii) state the same for mass, and Cheng et al. (2024) show that plain Langevin can miss components of general mixtures. Exact inference in factorial HMMs, whose chains the observation couples, is exponential in N (Ghahramani & Jordan, 1997), and weighted particle methods need ensembles exponential in dimension (Bengtsson et al., 2008); local particle filters avoid this under decay of correlations (Rebeschini & van Handel, 2015), and unweighted local chains avoid weights altogether when the posterior factorises (Proposition 5(i)) and, empirically, under moderate coupling (Appendix H.2).

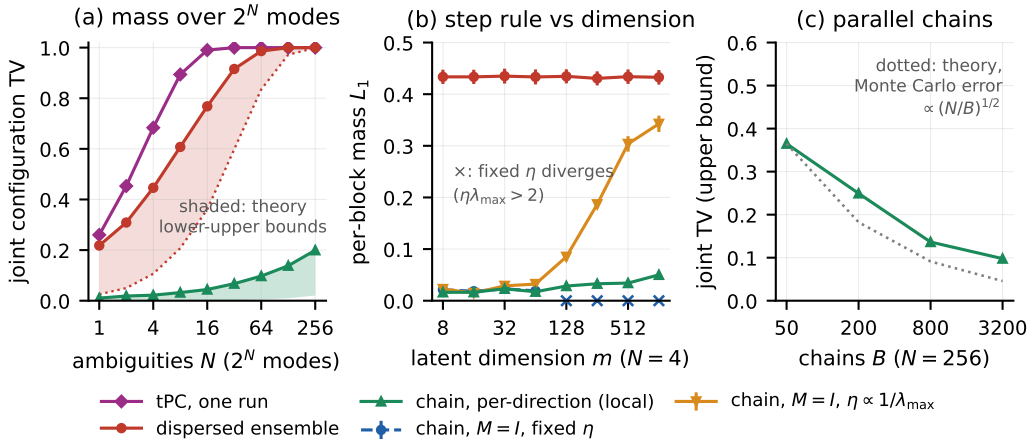

Figure 11: **Scaling with the number of ambiguities.** Product of N two-basin blocks with exact weights $w_i \in \{0.2, 0.3, 0.7, 0.8\}$, separation 4 posterior sd, 10 seeds. (a) Joint configuration error; the shaded band shows the bounds $[1-\prod_i \mathrm{BC}_i, \sqrt{1-\prod_i \mathrm{BC}_i^2}]$ of Proposition 5 at each method's per-block masses. (b) Per-block mass L_1 at N=4 against latent dimension, for a fixed step, a $1/\lambda_{\max}$-scaled step and the local per-direction schedule. (c) Joint-error bound at N=256, K=10^3 against the number of parallel chains (dotted line $B^{-1/2}$).

H.1 Results

The chain's per-block error does not grow with N; the ensemble's joint error goes to its maximum (Figure 11a, Table 15). From N=1 to 256 (2^{256} modes, m=1024), the chain's per-block mass L_1 stays at 0.015–0.021 with the local per-direction schedule (B=200, K=2000). The

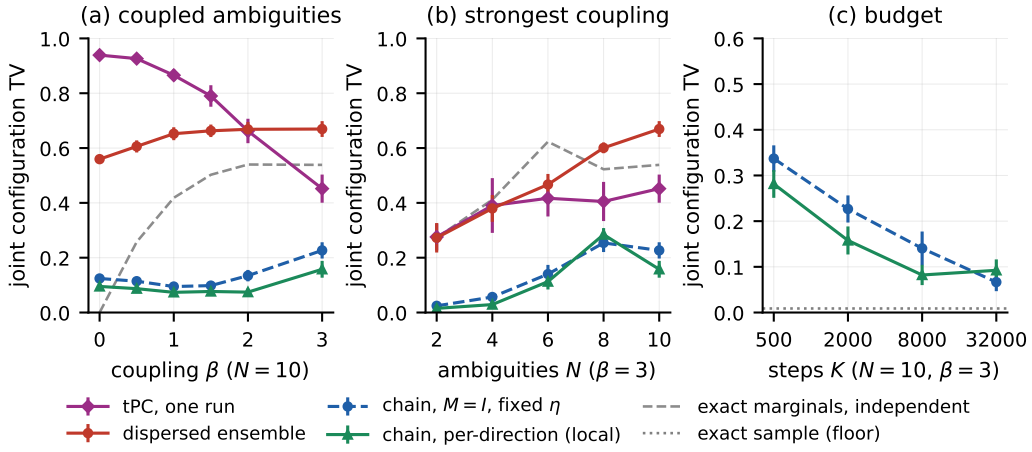

Figure 12: **Coupled ambiguities.** 10 seeds, B=200. (a) Joint configuration TV at N=10 (1024 modes), K=2000, against the coupling β. Dashed lines show the exact per-ambiguity marginals treated as independent. (b) At β=3 against the number of ambiguities. (c) At N=10, β=3 against the number of steps.

dispersed ensemble stays at 0.42–0.44 per block, so its joint error rises to ≥ 0.999, although it returns 20,000 points and receives 15× the chain's gradient evaluations. A single deterministic run reached the MAP configuration in all 170 runs; its probability is 2^{-107} at N=256. In all 120 ensemble runs with $N \leq 8$, where the joint error can be measured directly, it lies above the lower bound (ii).

Dimension is handled by the step rule of §4.3, not by more steps (Figure 11b, Table 16). With M_z=I, a fixed step diverges from m=128 on, and rescaling it by $1/\lambda_{\max}$ keeps it stable but raises mass L_1 from 0.02 to 0.34 at m=1024, because the stiff nuisance directions set the step for the flat branch direction. The local per-direction schedule, which needs no curvature, stays at 0.016–0.034 up to m=512 and reaches 0.050 at m=1024, seven times below the $1/\lambda_{\max}$ step, because its own pilot runs at that step, so its estimate of the branch spread degrades as the nuisance directions stiffen.

Accuracy is bought with parallel chains (Figure 11c, Tables 17 and 18). For $K \geq 10^3$ the per-block error does not grow with N from 1 to 256 (its spread across N is at most 0.016), so the joint error at large N is Monte Carlo error; at K=10^2 the 100-step pilot is too short and N=256 reaches 0.165. At N=256 it falls from 0.37 at B=50 to 0.10 at B=3200, as $B^{-0.33}$, slower than the Monte Carlo rate $B^{-1/2}$; the step constant is tuned per cell and K is fixed, so step bias and incomplete mixing contribute a floor. Chains are independent and local, so B grows with N without communication; one run with B=3200 chains, pilot included, took 3.2 s on one GPU in float64.

Scope. Factorisation at a separation of 4 sd is the favourable case, because it is where Proposition 5 applies exactly. Strong coupling costs steps and can exceed the budget (Appendix H.2), wider separations are limited by barrier crossing as in Appendix F.1, and the local schedule's pilot is not charged to the budget K.

H.2 Coupled ambiguities

Proposition 5 assumes independent ambiguities. We couple them, so the configuration weights become $\pi(c) \propto \prod_i w_{i,c_i} \exp(\beta \sum_{i<j} J_{ij} s_i s_j)$ with $s = 2c - 1$ and $J_{ij} \sim \mathcal{N}(0, 1/N)$, while every mode keeps its location and width, so β=0 is the product case. The posterior no longer factorises, and its 2^N weights are enumerated exactly for $N \leq 10$. Joint configuration TV is computed from each returned state's posterior configuration probabilities, which is unbiased on exact samples (an exact sample of equal size gives 0.000–0.056).

Table 14: Joint configuration TV for coupled ambiguities (mean ± s.e.m., 10 seeds). "Indep." is the error of the exact marginals treated as independent.

	indep.	tPC, one run	ensemble	$M_z{=}I$	local
coupling β, $N{=}10$, $K{=}2000$					
0	0.00	0.94 ± 0.00	0.56 ± 0.01	0.12 ± 0.00	0.10 ± 0.00
0.5	0.26	0.93 ± 0.01	0.61 ± 0.02	0.11 ± 0.01	0.09 ± 0.00
1	0.42	0.87 ± 0.02	0.65 ± 0.02	0.09 ± 0.01	0.07 ± 0.00
1.5	0.50	0.79 ± 0.04	0.66 ± 0.02	0.10 ± 0.01	0.08 ± 0.01
2	0.54	0.66 ± 0.04	0.67 ± 0.03	0.13 ± 0.02	0.07 ± 0.01
3	0.54	0.45 ± 0.05	0.67 ± 0.03	0.23 ± 0.03	0.16 ± 0.03
ambiguities N, $\beta{=}3$, $K{=}2000$					
2	0.27	0.28 ± 0.05	0.27 ± 0.05	0.02 ± 0.01	0.02 ± 0.00
4	0.41	0.39 ± 0.10	0.38 ± 0.05	0.06 ± 0.02	0.03 ± 0.01
6	0.62	0.42 ± 0.07	0.47 ± 0.04	0.14 ± 0.03	0.11 ± 0.03
8	0.52	0.41 ± 0.07	0.60 ± 0.02	0.25 ± 0.03	0.28 ± 0.02
10	0.54	0.45 ± 0.05	0.67 ± 0.03	0.23 ± 0.03	0.16 ± 0.03
steps K, $N{=}10$, $\beta{=}3$					
500				0.34 ± 0.03	0.28 ± 0.03
2000				0.23 ± 0.03	0.16 ± 0.03
8000				0.14 ± 0.04	0.08 ± 0.02
32000				0.07 ± 0.02	0.09 ± 0.02

Chains represent the coupling, not only each ambiguity's mass (Figure 12a, Table 14). At $N{=}10$ and $K{=}2000$, the chain's joint error with the local per-direction schedule is 0.07–0.10 for every $\beta \leq 2$, as without coupling. The exact marginals treated as independent err by 0.26–0.54; over $N{=}2$–10 at $\beta{=}3$, the chain's mean error is below both theirs and the dispersed ensemble's at every N (47 and 48 of 50 per run). The ensemble stays at 0.56–0.67. Coupling concentrates the mass, so one deterministic run improves with β, from 0.94 to 0.45, but still misses about half of it.

Strong coupling costs steps (Figure 12b,c). At $\beta{=}3$ the coupled configurations form groups separated by barriers, and the chain's error at $K{=}2000$ grows with N, from 0.02 at $N{=}2$ to 0.16 at $N{=}10$ (not monotonically, with 0.28 at $N{=}8$). At $N{=}10$, more steps lower the mean error from 0.28 at $K{=}500$ to 0.08–0.09 at $K{=}8000$–32000. On 4 of the 10 targets it reaches 0.01–0.04 (exact-sample floor 0.01), while on 4 it stays at 0.12–0.22 from $K{=}8000$ to 32000. This is the barrier-crossing limit of Appendix F.1, now between groups of configurations rather than between the two modes of one ambiguity.

H.3 Proof of Proposition 5

For two laws P, Q on a finite set, with $\mathrm{BC}(P,Q) = \sum_c \sqrt{P(c)Q(c)}$,

$$1 - \mathrm{BC}(P,Q) \;\leq\; \mathrm{TV}(P,Q) \;\leq\; \sqrt{1 - \mathrm{BC}(P,Q)^2}, \tag{20}$$

and $\mathrm{BC}(\bigotimes_i P_i, \bigotimes_i Q_i) = \prod_i \mathrm{BC}(P_i, Q_i)$ (Tsybakov, 2009, §2.4).

(i) Chains. The energy is separable, $F_\mu(z) = \sum_i F_i(z_i)$, M_z is block-diagonal and the noise is independent across blocks, so the update (6) of block i reads only z_i. Algorithm 1 draws each block's initial branch independently, so the law of a chain at every step k is a product $\bigotimes_i q_{i,k}$, and its configuration law is $\bigotimes_i \mathrm{Bern}(\rho_{i,k})$. The upper bound in (20) and multiplicativity give $\mathrm{TV}(q_k, p) \leq \sqrt{1 - \prod_i \mathrm{BC}(\rho_{i,k}, w_i)^2}$. Since $1 - \prod_i x_i^2 \leq \sum_i (1 - x_i^2) \leq 2\sum_i (1 - x_i)$ for $x_i \in [0,1]$, this is at most $\sqrt{2\sum_i h_i^2}$. A pooled read-out averages q_k over retained steps; TV is convex, so its error is at most the average of the per-step bounds. For a small deviation $\delta_i = |\rho_i - w_i|$, $h_i^2 = \delta_i^2/(8w_i(1-w_i)) + O(\delta_i^3)$, so the joint error is $\sqrt{N}$ times the per-block error.

Relaxation. For the diffusion, if each p_i satisfies a log-Sobolev inequality with constant α_i, the product satisfies one with $\alpha = \min_i \alpha_i$ (Bakry & Émery, 1985), so $D_{\mathrm{KL}}(q_t \| p) \leq e^{-2\alpha t} D_{\mathrm{KL}}(q_0 \| p)$ (Vempala & Wibisono, 2019), with $D_{\mathrm{KL}}(q_0 \| p) = \sum_i D_{\mathrm{KL}}(q_{0,i} \| p_i) \leq N\kappa_0$ and $\kappa_0 = \max_i D_{\mathrm{KL}}(q_{0,i} \| p_i)$. With Pinsker, $\mathrm{TV} \leq \epsilon$ once $t \geq \log(N\kappa_0/2\epsilon^2)/(2\alpha)$, which is logarithmic in N with the slowest block's constant. That constant falls exponentially in the barrier height (Bovier et al., 2004; Menz & Schlichting, 2014), which is the window of Appendix F.1. For the discretised chain the stationary bias is additive in KL over blocks; with an $O(\eta^2)$ per-block bias (Theorem 3 for Gaussian blocks,

under the smoothness conditions of Mou et al. 2022 in general) the step must shrink as $N^{-1/2}$ to hold it fixed. The Monte Carlo error of $\hat{\rho}_i$ from B independent chains is $O(B^{-1/2})$ per block and therefore $O((N/B)^{1/2})$ jointly.

(ii) Deterministic ensembles. Descent is separable in the same way, so each block ends in basin B_i with probability $\rho(B_i) + O(e^{-\Delta F})$ (Proposition 4), independently across blocks. The law of the ensemble's configurations is $\bigotimes_i \mathrm{Bern}(\rho_i)$, and the lower bound in (20) gives $\mathrm{TV} \geq 1 - \prod_i \mathrm{BC}(\rho_i, w_i)$. If $|\rho_i - w_i| \geq \delta$ and $w_i \in [w_{\min}, 1 - w_{\min}]$, then $\mathrm{BC}(\rho_i, w_i) \leq 1 - c\,\delta^2$ for a constant $c > 0$ that depends on $w_{\min}$, so the bound is at least $1 - e^{-cN\delta^2}$. More initial states estimate $\prod_i \rho_i$ more precisely; they do not move it towards $\prod_i w_i$.

(iii) One deterministic run. Its configuration law is a point mass at $\hat{c}$, and $\mathrm{TV}(\delta_{\hat{c}}, p) = 1 - p(\hat{c}) \geq 1 - \max_c p(c) = 1 - \prod_i \max(w_i, 1 - w_i)$. □

H.4 Setup

Block i has a branch coordinate with posterior $w_i\mathcal{N}(-\frac{\Delta s}{2}, s^2) + (1 - w_i)\mathcal{N}(\frac{\Delta s}{2}, s^2)$, $s = 0.25$, $\Delta = 4$ (inside the window of Appendix F.1), and $d - 1$ nuisance coordinates with Gaussian posterior precision $I/\sigma_H^2 + W_i^\top W_i/\sigma_O^2$, where $W_i \in \mathbb{R}^{2(d-1)\times(d-1)}$ has iid standard normal entries and $\sigma_H = \sigma_O = 0.5$; $m = Nd$. The weights w_i are drawn from $\{0.2, 0.3, 0.7, 0.8\}$, so an initialisation split of $\frac{1}{2}$ is visibly wrong. Exact samples and exact configuration weights are closed form. Chains follow Algorithm 1 with $B = 200$ and $K = 2000$ unless swept, discard half as burn-in, and return 20,000 thinned states. Three step rules are compared, namely $M_z = I$ with the fixed $\eta = 0.004$ of Appendix F.1; $M_z = I$ with $\eta = 0.4/\lambda_{\max}(H)$; and the local per-direction schedule (15), in which chain b uses $M_z^{(b)} = \mathrm{diag}(s_1^{(b)}, \ldots, s_m^{(b)})$ with step $c/\sqrt{K}$, $s_i^{(b)}$ the running standard deviation of unit i within chain b's own K-step $M_z{=}I$ pilot (step $0.4/\lambda_{\max}(H)$, same initialisation as the other chains, exponential moving averages of z_i and z_i^2 with window $K/8$ over the pilot's second half; nothing shared across chains, pilot not charged to K), and c tuned by SW_2^2 on a held-out seed over $\{0.05, \ldots, 3.2\}$. A run counts as diverged once any coordinate exceeds 10^3. Ensembles start each block at a branch mean drawn with probability $\frac{1}{2}$ plus $4s$ jitter and descend with per-unit steps $1/H_{ii}$ (the branch coordinate uses s^2) for 300 steps to convergence (largest preconditioned gradient below 10^{-15} in every run). The single deterministic run starts every block at the saddle. Per-block mass uses soft responsibilities as in §5.4; joint configurations use the sign of each branch coordinate. Ten seeds per cell. The coupled targets reuse these blocks with $d{=}4$ and J drawn per seed. The whole appendix (26 factorized shards, 13 coupled cells) reruns in about ten minutes on two GPUs.

Table 15: **Sweep over the number of ambiguities** N ($d = 4$). Per-block mass L_1 (mean $\pm$ s.e.m., 10 seeds) and the joint-error bracket of Proposition 5; the last column is exact.

		per-block mass L_1		joint TV bracket		tPC, one run
N	m	local	ensemble	local	ensemble	$1 - p(\hat{c})$
1	4	0.018 ± 0.003	0.416 ± 0.027	$[0.000, 0.010]$	$[0.025, 0.218]$	0.260
2	8	0.021 ± 0.004	0.415 ± 0.017	$[0.000, 0.019]$	$[0.050, 0.309]$	0.453
4	16	0.016 ± 0.002	0.434 ± 0.014	$[0.000, 0.022]$	$[0.106, 0.446]$	0.684
8	32	0.016 ± 0.001	0.440 ± 0.011	$[0.001, 0.032]$	$[0.207, 0.608]$	0.894
16	64	0.015 ± 0.001	0.433 ± 0.006	$[0.001, 0.044]$	$[0.361, 0.768]$	0.990
32	128	0.017 ± 0.001	0.437 ± 0.003	$[0.002, 0.067]$	$[0.598, 0.916]$	$1 - 2^{-13}$
64	256	0.017 ± 0.000	0.434 ± 0.002	$[0.005, 0.097]$	$[0.835, 0.986]$	$1 - 2^{-27}$
128	512	0.017 ± 0.000	0.432 ± 0.002	$[0.010, 0.138]$	$[0.972, 1.000]$	$1 - 2^{-54}$
256	1024	0.017 ± 0.000	0.434 ± 0.001	$[0.020, 0.199]$	$[0.999, 1.000]$	$1 - 2^{-107}$

Table 16: **Sweep over the latent dimension** ($N = 4$, $m = 4d$). Per-block mass L_1; "div." means all 10 seeds diverged.

m	$\eta\lambda_{\max}$	$M_z{=}I$, fixed η	$M_z{=}I$, $\eta \propto 1/\lambda_{\max}$	local	ensemble
8	0.09	0.021 ± 0.003	0.023 ± 0.002	0.016 ± 0.002	0.434 ± 0.014
16	0.25	0.018 ± 0.002	0.015 ± 0.002	0.016 ± 0.002	0.434 ± 0.014
32	0.61	0.023 ± 0.002	0.029 ± 0.004	0.023 ± 0.003	0.435 ± 0.014
64	1.39	0.019 ± 0.002	0.032 ± 0.004	0.017 ± 0.002	0.433 ± 0.015
128	2.93	div.	0.085 ± 0.007	0.028 ± 0.002	0.435 ± 0.014
256	5.75	div.	0.187 ± 0.007	0.033 ± 0.003	0.431 ± 0.014
512	11.83	div.	0.304 ± 0.015	0.034 ± 0.004	0.434 ± 0.014
1024	23.88	div.	0.343 ± 0.016	0.050 ± 0.005	0.433 ± 0.014

Table 17: **Per-block mass L_1 against N and budget K** (local per-direction schedule, $B = 200$, $d = 4$; at $K{=}10^2$ its pilot is 100 steps).

N	$K = 10^2$	3×10^2	10^3	3×10^3	10^4
1	0.046	0.070	0.038	0.019	0.007
4	0.058	0.042	0.027	0.018	0.012
16	0.064	0.031	0.022	0.015	0.011
64	0.069	0.033	0.022	0.015	0.010
256	0.165	0.036	0.022	0.015	0.009

Table 18: **Parallel chains at $N = 256$** (2^{256} modes, $m = 1024$, $K = 10^3$, $100B$ returned states). Per-block mass L_1, the joint-TV upper bound of Proposition 5(i), and wall-clock per run (pilot included, float64) on one GPU.

	local per-direction		
B	mass L_1	joint TV $\leq$	seconds
50	0.0330 ± 0.0003	0.365	0.20
200	0.0219 ± 0.0003	0.249	0.29
800	0.0119 ± 0.0001	0.136	0.80
3200	0.0085 ± 0.0001	0.098	3.15

I EQUILIBRATION, CRITERION, COST AND LOCALITY

I.1 THE CHAINS EQUILIBRATE RATHER THAN REMEMBER THEIR INITIALISATION

Starting every chain deep in either basin tests equilibration (Table 19); halving the step in §5.3 tests only discretisation sensitivity and could pass without crossings.

As reported in §5.3, halving η changes the first column of Table 1 by 0.013.

I.2 ON A UNIMODAL POSTERIOR THE CRITERION, NOT THE MODALITY, DECIDES

Table 20 compares replicated MAP and B Langevin samples on a unimodal linear-Gaussian target across σ_O, under point loss and a strictly proper rule.

The chain mean equals the MAP here because the posterior is Gaussian, which is why the point column cannot separate them. Sampling improves energy score 29.4–29.5% at every width while MAE differs by at most 0.001, so on this symmetric unimodal target the criterion determines the benefit. Likewise, PSNR rollouts (§5.6) cannot establish a distributional gain, even with multimodality.

I.3 WHAT A CHAIN ACTUALLY COSTS

Equal gradient counts omit Bm Gaussian draws per step and preconditioner construction/application; Table 21 measures these costs, with setup paid per conditioning frame.

Table 19: **Recovered mass from opposite initialisations.** B=200 chains, 20 seeds, $\pm$ one standard deviation across seeds. Left/right is the basin every chain starts in, crossed the fraction ending in the other, and estimate the mean posterior responsibility of the right component.

		estimated mass		chains crossed	
exact mass	K	from left	from right	from left	from right
0.20	10	0.004 ± 0.001	0.992 ± 0.003	0%	0%
	10^2	0.081 ± 0.016	0.740 ± 0.025	7%	25%
	10^3	0.196 ± 0.024	0.203 ± 0.025	19%	80%
	10^4	$\mathbf{0.195 \pm 0.029}$	$\mathbf{0.195 \pm 0.029}$	19%	81%
0.50	10	0.006 ± 0.002	0.994 ± 0.002	0%	0%
	10^2	0.155 ± 0.022	0.848 ± 0.022	14%	14%
	10^3	0.481 ± 0.033	0.496 ± 0.034	48%	50%
	10^4	$\mathbf{0.497 \pm 0.036}$	$\mathbf{0.497 \pm 0.036}$	50%	50%
0.80	10	0.007 ± 0.002	0.995 ± 0.002	0%	0%
	10^2	0.278 ± 0.025	0.920 ± 0.018	27%	7%
	10^3	0.794 ± 0.030	0.801 ± 0.030	80%	19%
	10^4	$\mathbf{0.795 \pm 0.033}$	$\mathbf{0.795 \pm 0.033}$	80%	20%

Table 20: **MAP versus chain on a unimodal linear-Gaussian target across σ_O.** m=16, n=32, B=100, K=2000, 5 seeds; MAE and energy score against the true latent.

		MAE ↓		energy score ↓		
σ_O	posterior sd	MAP	chain	MAP	chain	ES gain
0.1	0.170	0.138	0.138	0.684	**0.482**	29.5%
0.2	0.241	0.195	0.195	0.944	**0.666**	29.4%
0.5	0.297	0.236	0.235	1.148	**0.810**	29.4%
1.0	0.311	0.246	0.245	1.199	**0.846**	29.5%
2.0	0.315	0.244	0.244	1.214	**0.856**	29.5%
5.0	0.316	0.248	0.247	1.218	**0.859**	29.5%

The $0.87\times$ of $M_z{=}I$ is within run-to-run variation of parity. Peak memory differs by at most 6%, because all arms store B states and draw noise in place.

I.4 What each variant requires a unit to read

Local connectivity is preserved only by some variants.

variant	a unit reads	verdict
$M_z{=}I$	its own two prediction errors	local
$M_z{=}\Sigma_H$ diagonal	the above, scaled by its own $\sigma^2_{H,i}$	local
per-direction, $\eta_i = c/\sqrt{h_i K}$	a projection onto every eigenvector of H	**not local**
per-direction (local), (15)	its own two prediction errors, scaled by the running spread of its own activity within its own chain	local
mixture transition (19)	its errors under each branch, plus a shared responsibility	local given one broadcast scalar per branch
dense SPD M_z	all m prediction errors	**not local**

Only the first two rows and the local per-direction schedule use each unit's own errors, although Proposition 1 holds for any constant SPD M_z.

The branch responsibility in (19), shared across units, is the sole non-pointwise quantity inside a chain's update required by any reported experiment. The chain-mean read-out and the $B^{-1}\sum_b$ average of the weight update are not local in any variant, because they pool copies of the circuit.

Table 21: **Wall-clock and memory at matched B and K.** One conditioning frame at the video geometry (m=480, n=4096, B=32, K=200), single GPU, 3 repeats after a warm-up.

arm	setup (s)	run (s)	× det.	peak (MiB)	setup needs
deterministic tPC	0	0.0194	1.00	17.6	—
Langevin, $M_z{=}I$	0	0.0170	0.87	17.8	—
Langevin, $M_z{=}\Sigma_H$	0	0.0234	1.21	17.8	—
Langevin, per-direction	0.075	0.0268	1.38	18.7	eigendecomposition of H

J VIDEO CONTROLS AND PROTOCOL AUDIT

J.1 HOW MUCH DYNAMIC RANGE DO ROLLOUT PSNR AND SSIM HAVE HERE?

Model-free controls test whether the agreement in Table 4 is informative. Persistence and the training-mean image use the same 20 held-out sequences and 10 rollout frames.

Table 22: **Model-free predictors against both trained models.** Same 20 held-out sequences, 10 predicted frames and read-out as Table 4. Persistence repeats the last conditioning frame; constant is the training-set mean image; black is the uniform -1 frame; bold marks the best in each column pair.

	tPC		Langevin tPC		persistence		constant		black	
dataset	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Moving MNIST	14.83	0.616	14.84	0.618	12.21	0.673	**14.86**	0.669	14.40	**0.788**
SM-MNIST	16.64	0.746	16.66	0.748	13.93	0.780	**16.71**	0.827	16.46	**0.873**
KTH	24.13	0.534	24.28	0.542	**26.96**	**0.721**	14.81	-0.059	5.97	-0.003
UCF101[†]	21.79	0.550	21.85	0.553	**23.78**	**0.725**	12.32	0.091	7.51	0.089
BAIR[‡]	16.89	0.441	16.88	0.441	**19.79**	**0.794**	14.71	0.287	3.63	-0.171

In §5.6, PSNR is $10\log_{10}(\mathrm{MAX}^2/\mathrm{MSE})$ decibels, and SSIM is computed from local means, variances and covariances. On the same weights, inference-only posterior-mean and MAP read-outs agree within 0.02 dB on all five datasets (at most 0.016 dB, KTH); the trained models differ by at most 0.15 dB and 0.008 SSIM (Table 4); persistence or constant-image predictors match or beat both by up to 2.9 dB (Table 22).

On all five datasets a model-free predictor matches or beats both trained models on both PSNR and SSIM (Table 22), so these models do not beat model-free predictors at this horizon; foreground- or motion-weighted evaluation is needed but not attempted.

No positive claim rests on video quality. Table 4 does not establish general posterior-mean/MAP agreement; the separate same-weights read-out comparison (at most 0.016 dB) is the evidence (§5.6).

J.2 CONTROLLED AUDIT OF THE UNSTABILISED-VERSUS-STABILISED PROTOCOL

An uncontrolled comparison that changes the training protocol and the E-step together produces large video gaps; retraining the protocol/E-step 2×2 isolates them. Table 4 reports only matched comparisons.

The uncontrolled comparison changes the protocol and the E-step at once, so its gap is not attributable to either; run separately (Table 23), the inference rule contributes $+0.02$ dB on SM-MNIST and $+0.14$ dB on KTH under a matched protocol, and neither dataset reproduces the carried-over gap. On Moving MNIST all four {cap}×{E-step} cells reach PSNR 14.56, each capped cell bitwise identical to its uncapped twin. Under the $0.5I$ initialisation the cap never binds, so that factor is untested at this scale.

K UCI POSTERIOR-RECOVERY CONTROLS

Numbers reported in §5.4. The dispersed MAP ensemble gets $150\times$ the compute (15,000 returned points). Langevin tPC is closer to the exact posterior on all twelve datasets in SW_2^2, MMD^2 and

Table 23: **Controlled 2×2 audit of training protocol and E-step.** Same data, architecture, 100 epochs and seed; rollout PSNR on 20 held-out sequences with deterministic read-out, final $\|W_H\|_2$ beside each.

	det E-step (tPC)		Langevin E-step		
training protocol	PSNR	$\|W_H\|_2$	PSNR	$\|W_H\|_2$	Δ
SM-MNIST — an earlier, uncontrolled run reports a $+10.29$ dB gap					
unstabilised (no cap, random W_H, σ_O=1.0)	16.66	5.5	16.63	53.4	-0.03
stabilised (cap 2.0, W_H=0.5I, σ_O=0.1)	16.64	1.4	16.66	1.4	$+0.02$
protocol effect at fixed rule	-0.03 dB (det)		$+0.02$ dB (Lan.)		
KTH — an earlier, uncontrolled run reports a $+13.11$ dB gap					
unstabilised (no cap, random W_H, σ_O=1.0)	20.75	3.7	18.05	21.0	-2.70
stabilised (cap 2.0, W_H=0.5I, σ_O=0.1)	24.13	1.4	24.27	1.4	$+0.14$
protocol effect at fixed rule	$+3.38$ dB (det)		$+6.22$ dB (Lan.)		

mass L_1, and on eleven in coverage (Table 2). Pooled SW_2^2 is $4.9\times$ smaller ($2.2\times$ in distance), mass L_1 falls from 1.11 to 0.40, and 90% coverage rises from 41.2% to 75.6%, still below target. Exact weights close 75% of the gap and exact shapes 11% (Table 25), and an ensemble that reaches every mode still has mass error 1.133 (Table 28), as Proposition 4 predicts. Laplace at the MAP (1.106; ensemble 1.108) and an Adam-style preconditioner (Zahid et al., 2024) lose on all twelve; a one-sample-per-chain read-out as in MCPC (Oliviers et al., 2024) beats our pooled one (Table 13). The test is static (W_H=0); mass is recovered at mode separations of 3–6 posterior standard deviations (peak $9.5\times$, none beyond 8), where 84% of UCI inputs lie (App. F.1).

Energy and drift. Energy is $-\log \sum_k \pi_k \mathcal{N}(z; \nu_k, \Sigma_k) + \|x - z\|^2 / 2\sigma_x^2$; drift averages component gradients by responsibility, and the posterior is an exact Gaussian mixture.

Metric definitions. Three metrics compare returned samples S with exact-posterior samples E (lower is better). *Squared sliced Wasserstein-2* (Bonneel et al., 2015), SW_2^2, averages squared one-dimensional Wasserstein-2 distances over 128 random unit projections; ratio r means $\sqrt{r}$ in distance. *Squared maximum mean discrepancy* (Gretton et al., 2012), MMD^2, uses an RBF kernel with median-heuristic bandwidth. *Mass* L_1, $\sum_k |\hat{w}_k - w_k|$, compares exact weights w_k with mean responsibilities $\hat{w}_k$ over S. Soft assignments handle overlap; the measure excludes within-basin shape.

Table 24: **Per-dataset MMD^2 to the exact posterior**, the column omitted from Table 2; RBF kernel, median-heuristic bandwidth.

dataset	chain ↓	ens.
auto MPG	**0.0428**	0.2166
wine quality	**0.1350**	0.2485
energy efficiency	**0.0635**	0.2398
concrete	**0.0290**	0.1827
power plant	**0.0096**	0.1618
bike sharing	**0.0916**	0.2005
diabetes	**0.0335**	0.1748
california housing	**0.0175**	0.3282
abalone	**0.1542**	0.2693
yacht	**0.0325**	0.1517
real estate	**0.0520**	0.1706
servo	**0.0815**	0.2291
pooled	**0.0619**	0.2145

MMD^2 agrees with SW_2^2 on all twelve datasets (Table 24), which is why only one of the two is in the main text.

K.1 Is the advantage between-mode mass, or within-mode shape?

SW_2^2, MMD^2, and responsibilities confound within-mode shape with between-mode mass. Table 25 separates them using fitted responsibilities from samples S. Shape oracles retain estimated weights, weight oracles retain estimated shapes, and both return B samples.

Table 25: **Shape and weight oracles.** 12 datasets, 3 seeds, 30 held-out inputs each, B=100; *exact* denotes B draws from the true posterior.

arm	$SW_2^2 \downarrow$	$MMD^2 \downarrow$	mass $L_1 \downarrow$	what it is given
Langevin chain	0.102	0.063	0.401	—
dispersed ensemble	0.462	0.217	1.106	—
shape oracle	0.424	0.168	1.061	exact shapes, *ensemble's* weights
weight oracle	**0.192**	0.131	**0.072**	exact weights, *ensemble's* shapes
exact, B=100	0.022	0.007	0.062	the B-sample floor

Weight errors dominate as Proposition 4 predicts. Exact shapes change $0.462 \to 0.424$ (11% gap closure), exact weights $0.462 \to 0.192$ (75%), with mass error $1.106 \to 0.072$. The chain's 0.102 remains $4.7\times$ the B=100 floor, so mass recovery does not imply exact shape recovery or exact sampling.

K.2 The thirteenth dataset, and why it is not in the pool

Naval propulsion is excluded for an ill-posed read-out, with design condition number 8.4×10^{18} versus 7.4×10^{15} at worst among the twelve included datasets, nine below 25. Its results are chain/ensemble SW_2^2 0.197/0.430, coverage 100%/9%, MAE 0.861/0.783 versus exact-mean 0.781. Covering everything is not calibration; including this thirteenth dataset changes pooled distance ratio $4.9\times \to 4.5\times$ and coverage $75.6\% \to 77.5\%$.

K.3 Mode count and sensitivity to the prior

Supported modes must have non-negligible mass, a separating density dip, and bootstrap stability after component selection.

Table 26: **Supported modes per dataset at two mass thresholds.** $J^\star$ is the plateau selection; brackets give min–max over 3 seeds.

		modes, τ=1%		modes, τ=5%	
dataset	$J^\star$	count	range	count	range
auto_mpg	5	4	[4,8]	4	[4,5]
wine_quality	6	4	[2,7]	4	[2,4]
energy_efficiency	6	5	[4,7]	4	[1,5]
concrete	30	12	[5,16]	3	[2,5]
power_plant	2	2	[2,2]	2	[2,2]
bike_sharing	24	9	[1,10]	1	[1,3]
diabetes	16	3	[3,3]	2	[2,3]
california_housing	3	2	[2,3]	2	[2,2]
abalone	5	4	[3,9]	3	[3,5]
yacht	1	1	[1,9]	1	[1,1]
real_estate	6	6	[2,13]	6	[2,8]
servo	1	1	[1,1]	1	[1,1]
median	—	4	—	2.5	—

Counting every bump of a criterion-selected mixture instead gives 3 to 36, with two criteria disagreeing threefold; the filters are what make the count meaningful. At 1% mass, counts span 1–12 (median

Table 27: **The UCI experiment under seven priors**, pooled over 12 datasets × 3 seeds; chain and budget-matched dispersed ensemble fixed, only the prior changes.

		SW_2^2		mass $L_1 \downarrow$		cov_{90}		
prior	J	chain ↓	ratio ↑	chain	ens.	chain	ens.	MAE ↓
k-means, J=2	2.0	0.085	3.80×	0.258	0.772	71.4%	29.6%	0.623
k-means, J=4 *(released)*	4.0	0.091	5.11×	0.392	1.108	75.8%	40.7%	0.584
k-means, J=8	8.0	0.094	5.47×	0.456	1.232	79.3%	53.4%	0.564
k-means, J=16	16.0	0.079	5.03×	0.425	1.229	81.3%	62.3%	0.550
GMM, J by held-out likelihood	11.0	0.069	4.90×	0.365	1.099	79.0%	49.2%	0.569
Dirichlet process, J inferred	12.9	0.053	5.63×	0.345	1.117	79.9%	48.0%	0.558
mean-shift, no J anywhere	2.8	0.061	9.16×	0.088	0.447	64.9%	21.3%	0.670

4); ten of twelve datasets are multimodal, servo and yacht unimodal, so $J = 4$ is not universally correct.

Mean-shift has no J at any point in Table 27, because it takes a bandwidth and the components fall out of it. The SW_2^2 advantage persists across priors, at 3.8–5.5× for $J \in \{2, 4, 8, 16\}$ and comparable or larger with inferred counts, despite changed posterior targets.

K.4 Mode reach versus mode weight

Mode-matched initial states at every component mean guarantee initial occupancy.

Table 28: **Dispersed versus mode-matched ensembles**, pooled over 12 datasets × 3 seeds at the chain's sample budget. *Dispersed* is the released baseline (initial states from the prior, spread at 4σ); *mode-matched* places initial states at every component mean; *uniform limit* is $\sum_k |1/J - w_k|$ (Proposition 4, one initial state per mode).

		mass $L_1 \downarrow$				$SW_2^2 \downarrow$		
prior	J	chain	disp.	mode-m.	unif. lim.	chain	disp.	mode-m.
k-means, J=4	4	**0.394**	1.108	1.133	1.186	0.090	0.464	1.830
k-means, J=8	8	**0.455**	1.232	1.215	1.285	0.091	0.513	1.430
Dirichlet process	13	**0.340**	1.117	1.093	1.325	0.051	0.296	0.279

Weighting, rather than support access, limits ensembles. Guaranteed reach changes mass error by at most 3%, while chains remain 2.7–3.3× better.

K.5 Prior-leakage controls

Table 29 varies prior-fitting data for fixed and Dirichlet-process priors; the released protocol uses training inputs only.

The disjoint split shares no point with the read-out training set or the test set, so no information path exists; fitting on test features bounds what leakage is worth; shuffling preserves every marginal and destroys all joint structure. Disjoint data reproduce the released ratio and MAE; deliberate test leakage does not improve them. Shuffling improves posterior matching but worsens MAE, so distance ratios require predictive metrics alongside them.

K.6 Training-split check

UCI has no EM ($W_H = 0$, identity emission), allowing unchanged inference on training data.

The two rows of Table 30 agree to 0.001 in MAE and 0.3 points in coverage, so nothing here is retrieval.

Table 29: **Prior-leakage controls under a fixed-J and a Dirichlet-process prior.** Within a block only the prior's fitting data change, which are *disjoint* (a third split), *test* (test features) or *shuffled* (training columns permuted independently); J is the resulting component count. 12 datasets $\times$ 3 seeds.

prior fitted on	J	SW_2^2 chain $\downarrow$	ratio vs. ens. $\uparrow$	mass L_1 $\downarrow$	MAE $\downarrow$	cov_{90}
k-means prior family, J=4 fixed						
X_{train} (released protocol)	4.0	0.103	4.99$\times$	0.407	0.590	76.6%
a disjoint third split	4.0	0.098	5.00$\times$	0.418	0.591	77.0%
X_{test} (deliberate leakage)	4.0	0.106	3.77$\times$	0.428	0.593	77.9%
X_{train}, columns permuted	4.0	0.023	10.31$\times$	0.165	0.647	72.2%
Dirichlet-process prior family, J inferred						
X_{train} (released protocol)	11.0	0.046	6.66$\times$	0.297	0.564	80.7%
a disjoint third split	9.3	0.045	6.08$\times$	0.292	0.566	79.8%
X_{test} (deliberate leakage)	7.8	0.049	6.24$\times$	0.289	0.573	79.2%
X_{train}, columns permuted	5.9	0.013	14.20$\times$	0.130	0.632	72.3%

Table 30: **UCI on the training versus the held-out split**, pooled over 12 datasets $\times$ 5 seeds, equal numbers of points; *Bayes* is the exact posterior-mean estimator.

	MAE $\downarrow$			SW_2^2 $\downarrow$		MMD^2 $\downarrow$		cov_{90}	
split	chain	ens.	Bayes	chain	ens.	chain	ens.	chain	ens.
train	0.599	0.679	0.565	0.099	0.462	0.062	0.213	75.8%	38.8%
test	0.599	0.678	0.568	0.093	0.461	0.061	0.214	75.5%	41.2%

L How the latent dimension acts on learning, with three clocks checked on an LDS

This appendix gives a learning-side account of how the latent dimension m changes training under the Hebbian M-step (7), and checks its signs on a synthetic linear dynamical system (LDS). *Scope.* No main-text claim uses it. It concerns weight learning, not the inference rule, and it is derived for $f = \mathrm{id}$ and $\Sigma_H = \sigma_H^2 I$, $\Sigma_O = \sigma_O^2 I$. "Memorisation" means only a train/held-out gap in one-step prediction error (onset is when held-out error first exceeds twice the training error), not storage or retrieval of sequences. The account predicts signs and orderings; its exponents are not confirmed (below).

Three clocks. Averaged over data, (7) with $f = \mathrm{id}$ is a linear flow driven by the latent second moment $\hat{\Sigma}_{zz}$; the eigendirection with eigenvalue λ_i relaxes with time constant $\tau_i = \sigma_O^2/(\eta_w \lambda_i)$. Split the spectrum into a rank-r signal part and a Marchenko–Pastur bulk of ratio $\gamma = (m - r)/(N_{\mathrm{tr}} P)$, for N_{tr} training sequences of P positions. The same flow then has a *generalisation* clock $\tau_{\mathrm{gen}} = \sigma_O^2/(\eta_w \lambda_{\mathrm{sig}})$, independent of m, and a *memorisation* clock $\tau_{\mathrm{mem}} \propto \sigma_O^2 N_{\mathrm{tr}} P/(\eta_w m)$. A third clock, different in kind, governs W_H. In the eigenbasis of the lagged second moment (eigenvalues ν_i) singular value i approaches a limit $\propto 1/\nu_i$ with time constant $\sigma_H^2/(\eta_w \nu_i)$, so the direction that overshoots most is the slowest, $\tau_{\mathrm{spec}} = \sigma_H^2/(\eta_w \nu_{\min})$. Early growth is ν-independent, so spectral runaway appears only late, which is why Algorithm 2's optional cap κ bounds $\|W_H\|_2$ instead of treating it as a tuned quantity. All of (N_{tr}, m, t) then enters through $\gamma = m_{\mathrm{eff}}/(N_{\mathrm{tr}} P)$ and $s = t/\tau_{\mathrm{mem}}$, with $m_{\mathrm{eff}} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2 \le m$ the participation ratio of $\hat{\Sigma}_{zz}$. Here $\gamma < 1$ is safe at any t, $\gamma \ge 1$ with $s < 1$ is safe by early stopping, and $\gamma \ge 1$, $s \ge 1$ memorises. Once m_{eff} stops growing, a larger m moves neither γ nor the filter. By contrast M_z never changes the target (Proposition 1); m changes the model.

LDS check. The ground truth is a 16-dimensional latent with a rank-8 rotation ($0.9\times$ orthogonal), $n = 256$, $P = 20$, $\sigma_H = \sigma_O = 0.1$, $\eta_w = 2 \times 10^{-3}$; exact filtering E-step, averaged Hebbian M-step, 64 held-out sequences, 3 seeds (`exp_verify_clocks`, `exp_clocks_exponents`). (a) m_{eff} *stays far below* m *but is not flat* (Table 31). At $m = 512$ it is 16.8–23.3; the last doubling adds 0.9–1.8, the one from 128 to 256 adds 2.5–7.6. After 3000 epochs only $m \in \{256, 512\}$ at $N_{\mathrm{tr}} = 8$ have reached memorisation onset (1693 and 955 epochs). (b) *Onset tracks the interpolation threshold.* Over $m \in \{64, \ldots, 512\} \times N_{\mathrm{tr}} \in \{8, 16, 32\}$ and up to 6×10^4 epochs, every cell with $m \ge 0.8\, N_{\mathrm{tr}} P$ reaches onset and no cell with $m \le 0.4\, N_{\mathrm{tr}} P$ does. The latest onset is the marginal cell $m = 512$, $N_{\mathrm{tr}} = 32$ (2.4–3.8 $\times 10^4$ epochs). (c) *Signs hold, magnitudes do not.* Log-log fits of onset time give $\partial \log \tau_{\mathrm{mem}}/\partial \log N_{\mathrm{tr}} = +1.9$ ($m = 256$, two points) and $+2.3$ ($m = 512$), against a predicted $+1$, and $\partial \log \tau_{\mathrm{mem}}/\partial \log m = -1.1$ ($N_{\mathrm{tr}} = 8$) and -0.66 ($N_{\mathrm{tr}} = 16$, two points), against -1. Onsets fall on a $\times 1.58$ checkpoint grid, so a two-point slope over a factor of 2 is uncertain by up to ± 0.66.

Table 31: **Participation ratio m_{eff} of the inferred latents** after 3000 epochs on the LDS (true latent dimension 16, signal rank 8), mean over 3 seeds.

$N_{\mathrm{tr}} \backslash m$	4	8	16	32	64	128	256	512
8	3.8	7.2	8.1	8.6	10.3	13.3	15.9	16.8
32	3.7	7.6	8.7	9.4	11.1	14.9	21.3	23.2
128	3.8	7.7	8.9	9.6	10.9	14.3	21.9	23.3