# WHY MCPC GENERALIZES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Monte Carlo predictive coding (MCPC) infers latent causes by Langevin sampling, which makes it a generative model. We ask when an MCPC network generalizes during training and when it begins to memorize. Its local Hebbian updates average over the sampled cloud of latent states, so the cloud's spread acts as a ridge. For a linear–Gaussian coder we derive this ridge in closed form. It sets a threshold on the data spectrum, and directions farther above it are learned sooner. Population directions lie far above and set the generalization onset. Directions that exist only because the training set is finite lie just above and set a later memorization onset. The network generalizes without memorizing in the window between the two onsets. More data widens this window, and past a critical data size memorization never begins. Longer or more chains move the memorization onset but cannot prevent it. A larger step size can prevent it, but only by also losing the signal. The theory's learning flow reproduces both measured onsets. A pre-registered causal test shows that the ridge, not the sampling noise, sets the memorization onset. On CelebA faces more data likewise delays memorization while sample quality peaks at about the same step. To make MCPC generalize rather than memorize, stop training between the two onsets or train on more data than the critical size.

## 1 INTRODUCTION

Predictive coding models cortex as a top-down generative model. Each layer predicts the activity of the layer below and passes on only the mismatch, the prediction error (Rao and Ballard, 1999; Friston, 2005). The same local errors drive inference and learning (Bogacz, 2017; Millidge et al., 2021; Whittington and Bogacz, 2017). In inference, latent activity relaxes until the errors are as small as possible, which settles at the most probable (MAP) latent. In learning, each weight changes by a local Hebbian product of pre-synaptic activity and post-synaptic error. Monte Carlo predictive coding (MCPC; Oliviers et al., 2024) adds noise to inference, which turns it into a Langevin chain. Activity then samples the posterior over causes instead of settling at its mode, the network becomes a generative model, and the unchanged Hebbian update averages over a cloud of latent states (Fig. 2).

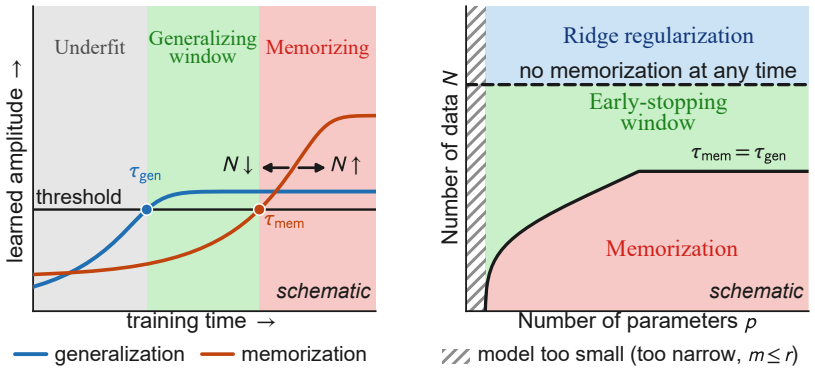
Data, capacity and training time decide whether a network learns structure or fits its training set (Zhang et al., 2017; Belkin et al., 2019), and whether a generative model copies its training examples (Gu et al., 2025; Yoon et al., 2023; Carlini et al., 2023); for diffusion models, Bonnaire et al. (2025) found that memorization sets in only after a training time that grows with the data size. We ask the same of MCPC. When does an MCPC network generalize, when does it memorize, and which part of the sampled inference decides? Here memorization means learning the directions of the training data that exist only because the training set is finite; it is not copying. Fig. 1 sketches the answer for model size, data and training time.

**Three questions.**
**Q1** What does sampled inference change in what the Hebbian update learns? (§§3.2–3.3, 4.1.1)
**Q2** Which directions does the network keep, and can the sampler's parameters prevent memorization? (§§3.5–3.6, 4.1.2–4.1.3)
**Q3** How do data size and model width decide whether those directions are kept? (§§3.7, 4.2)

**The idea.** The Hebbian update does not see a single latent state but the average over the sampled cloud, which is the posterior mean's contribution plus the spread of the cloud. The update inverts this average, so the spread acts as a ridge (Hastie et al., 2009; Bishop, 1995). A larger ridge suppresses



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.




(a) two timescales (b) parameters vs. data

Figure 1: **Schematic of the analysed model**, drawn from its formulas at schematic constants, not measured (App. B.8). (a) Learned amplitude of a generalization (blue) and a memorization (red) direction against training time, with the threshold, the onsets $\tau_{\text{gen}}$ and $\tau_{\text{mem}}$, and the shift at larger $N$. (b) The $(p, N)$ plane ($p = nm$ decoder parameters; axes as in App. Fig. 13), with memorization first (red), an early-stopping window (green) and no memorization above $N^\star_\infty$ (blue).

weak directions, those carried only by finite-sample fluctuation, before strong population ones. We call the spread the ridge term, and the scatter of a single sampled state around it the draw.

**Our contribution.** (i) We identify the ridge that a finite Langevin chain hands to the Hebbian update in closed form. It is the posterior covariance of exact expectation maximization, inflated by the discretized chain and truncated by the window of averaged states (Q1). (ii) The ridge sets a retention threshold, computable before training. Exact posterior sampling keeps finite-sample directions at every data size; the chain's larger ridge suppresses them, and a pre-registered causal test shows that the ridge itself, not the sampling noise, is the cause (Q2). (iii) Of the sampler's parameters, only the step size can lift the threshold past the noise level; burn-in, window and chain count change when memorization happens, not whether (Q2). (iv) More data delays memorization and, past a crossing size, removes it; more capacity raises the data needed within the training budget (Q3). Drawn from its prior, the trained decoder produces faces that are not copies.

**Known, and new.** Injected noise acts as a Tikhonov ridge (Bishop, 1995), and weight-space Langevin noise bounds generalization (Mou et al., 2018). Sampled E-steps (Wei and Tanner, 1990; Neal and Hinton, 1998; Oliviers et al., 2024; Zahid et al., 2024) have known asymptotics (Nielsen, 2000), and the unadjusted Langevin chain has a known bias (Roberts and Tweedie, 1996; Durmus and Moulines, 2017; Dalalyan, 2017; Vempala and Wibisono, 2019). Under spiked spectra the finite-sample fluctuation bulk reaches the Marchenko–Pastur edge (Baik et al., 2005; Johnstone, 2001), above the cut of exact EM (Tipping and Bishop, 1999), and noise-calibrated thresholds just above the bulk edge select singular values (Gavish and Donoho, 2014) and variational-PCA components (Nakajima et al., 2015). New here are the closed-form ridge of a finite latent Langevin E-step, its retention threshold, and the one sampler parameter that moves it (Cor. 3.5).

## 2 Model and Algorithm

### 2.1 A two-layer linear–Gaussian predictive coder

We analyze a multivariate version of the linear model of Oliviers et al. (2024), probabilistic PCA (Tipping and Bishop, 1999; Roweis and Ghahramani, 1999) trained by a local rule, with latent $z \in \mathbb{R}^m$ (capacity $m$), observation $y \in \mathbb{R}^n$, emission matrix $W$ and $N$ training points $Y = \{y_a\}_{a=1}^N$,

$$z \sim \mathcal{N}(0, \sigma_H^2 I), \;\; y = Wz + \xi, \;\; \xi \sim \mathcal{N}(0, \sigma_O^2 I), \qquad z \mid y \sim \mathcal{N}(m_y, H^{-1}),$$

$$H = \sigma_H^{-2} I + \sigma_O^{-2} W^\top W, \qquad m_y = H^{-1} \sigma_O^{-2} W^\top y.$$

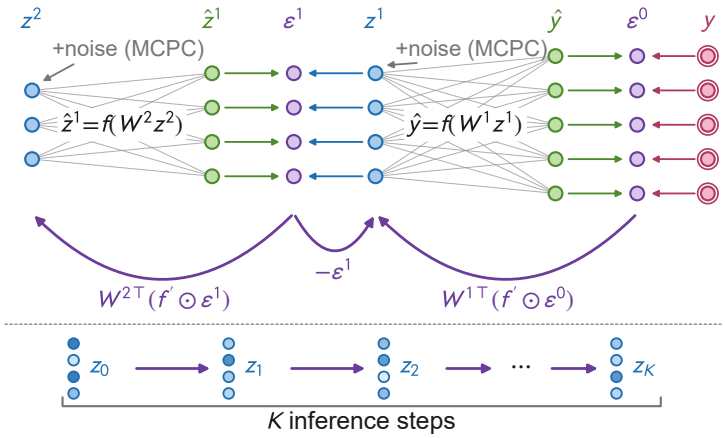



Figure 2: **Network shared by PC and MCPC.** Each layer predicts the one below, and each latent layer updates from its neighbours' prediction errors for $K$ inference steps, noise-free under PC (MAP) and with Langevin noise under MCPC. The analysed model has one latent layer ($z = z^1$) and $f$ the identity; §4.3's decoder is $W_2 \tanh(W_1 z)$.

PC and MCPC share the energy, the negative log joint up to a constant,

$$F(z, W; y) = \frac{1}{2\sigma_O^2}\|y - Wz\|^2 + \frac{1}{2\sigma_H^2}\|z\|^2, \qquad \nabla_z F = H(z - m_y), \qquad -\nabla_W F = \sigma_O^{-2}(y - Wz)z^\top. \tag{1}$$

PC descends $F$ in $z$ to the MAP latent $m_y$, and MCPC samples $p(z \mid y) \propto e^{-F}$. Directions the likelihood ignores have precision $\sigma_H^{-2}$, and memorization happens there (App. A). Learning is a local Hebbian M-step, averaged over the batch and, under MCPC, over the sampled states of the E-step; PC evaluates it once, at the MAP latent,

$$W \leftarrow W + \frac{\eta_w}{\sigma_O^2}\left(\widehat{\Sigma}_{yz} - W\widehat{\Sigma}_{zz}\right), \qquad \Delta W \propto \langle (y - Wz)z^\top \rangle, \tag{2}$$

with $\widehat{\Sigma}_{yz}$ the data–latent cross moment and $\widehat{\Sigma}_{zz}$ the latent second moment, $\widehat{\Sigma}_{mm} + C$ under MCPC (Eq. 5) and $\widehat{\Sigma}_{mm}$ under PC.

### 2.2 Sampling the inference phase, and two ways to summarize the samples

MCPC replaces MAP inference by discretized Langevin sampling at effective step size $\bar{\eta}$,

$$z_{k+1} = z_k - \bar{\eta} H(z_k - m_y) + \sqrt{2\bar{\eta}}\,\xi_k, \qquad \xi_k \sim \mathcal{N}(0, I), \tag{3}$$

started at $m_y$, with $b$ burn-in steps, a window of $S$ states and $B$ chains per datum. The accumulating rule of the original MCPC code, studied here, averages $z_k z_k^\top$ over the window and so keeps the covariance; the averaging rule takes $\bar{z}\bar{z}^\top$ of the window mean (App. C.2, App. Fig. 10).

### 2.3 What generalization and memorization mean here

The generalization onset $\tau_{\text{gen}}$ is the first checkpoint whose held-out likelihood is within a fixed tolerance of the population optimum's. Memorization is a move of the model covariance $\Sigma_{\text{model}} = \sigma_H^2 W W^\top + \sigma_O^2 I$ toward the empirical covariance $\Sigma_{\text{emp}}$ rather than the population one $\Sigma_{\text{pop}}$, read from the sign of the fit fraction $\phi_{\text{fit}} = \|\Sigma_{\text{model}} - \Sigma_{\text{pop}}\|_F / \|\Sigma_{\text{model}} - \Sigma_{\text{emp}}\|_F$; the memorization onset $\tau_{\text{mem}}$ is the first checkpoint with $\phi_{\text{fit}}$ above a fixed bar (App. C.3).

# 3 THE ANALYSIS IN THE EIGENBASIS OF $H$

## 3.1 THREE ASSUMPTIONS

(i) Identity activations and one matrix for emission and inference. (ii) The chain starts at the posterior mean. (iii) The data spectrum's signal part is separated from its fluctuation part (Baik et al., 2005). Their cost is in App. Table 5, the schematic in App. Fig. 11 and the proofs in App. B.

## 3.2 STEP ONE — ANY SAMPLED INFERENCE PHASE IS A RIDGE

**Proposition 3.1** (Any sampled E-step is a ridge)**.** Let the E-step deliver a cross moment unbiased for $\widehat{\Sigma}_{ym} = \frac{1}{N}\sum_a y_a m_{y_a}^\top$ and a latent second moment whose covariance $C \succeq 0$ does not depend on the datum, and let $\widehat{\Sigma}_{mm} = \frac{1}{N}\sum_a m_{y_a} m_{y_a}^\top$ (App. Eq. 28). Then the M-step's frozen target is

$$\mathbb{E}[\widehat{\Sigma}_{zz} \mid Y] = \widehat{\Sigma}_{mm} + C, \qquad W_C = \widehat{\Sigma}_{ym}\big(\widehat{\Sigma}_{mm} + C\big)^{-1}, \tag{4}$$

against the interpolant $W^{\mathrm{MAP}} = \widehat{\Sigma}_{ym}\widehat{\Sigma}_{mm}^{\dagger}$ at $C = 0$.

## 3.3 STEP TWO — WHICH RIDGE

**Theorem 3.2** (The realized ridge is the posterior covariance, twice corrected)**.** Run the chain (3) from $m_y$ with $\bar{\eta}\lambda_{\max}(H) < 2$, discard $b$ states and average the outer product over the next $S$. Then $C$ is diagonal in $H$'s eigenbasis,

$$C = H^{-1}\Big(I - \tfrac{\bar{\eta}}{2}H\Big)^{-1}\Phi(H), \qquad \Phi(\lambda) = 1 - \frac{\rho_\lambda^{2(b+1)}\big(1-\rho_\lambda^{2S}\big)}{S\big(1-\rho_\lambda^2\big)}, \qquad \rho_\lambda = 1 - \bar{\eta}\lambda. \tag{5}$$

Sampling thus changes learning through $C$, not through the posterior mean (proofs in App. B.2 and B.3).

## 3.4 STEP THREE — WHAT THE RIDGE DOES TO LEARNING

**Theorem 3.3** (The flow decouples, and each mode has a threshold)**.** Take centred data and the M-step flow at rate $\eta_w$, in expectation over the sampler, with $H$ and the inferred means depending on $W$, and let

$$\widehat{\Sigma}_y u_j = c_j u_j, \qquad W = \sum_j s_j u_j v_j^\top, \qquad C = \phi(H),$$

with $\widehat{\Sigma}_y = \frac{1}{N}\sum_a y_a y_a^\top$ (eigenvalues $c_1 \geq c_2 \geq \cdots$), $v_j$ the right singular vectors of $W$ and $\phi$ acting on $H$'s eigenvalues.

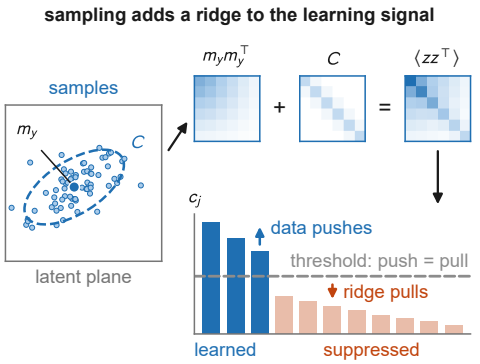
**(i) Decoupling, for every $\phi$.** This aligned set is invariant, and on it each singular mode follows, with $\alpha = \eta_w/\sigma_O^2$,

$$\dot{s}_j = \alpha\, s_j \left[\frac{c_j}{\sigma_O^2\sigma_H^2\lambda_j^2} - \phi(\lambda_j)\right], \qquad \lambda_j = \sigma_H^{-2} + \frac{s_j^2}{\sigma_O^2}. \tag{6}$$

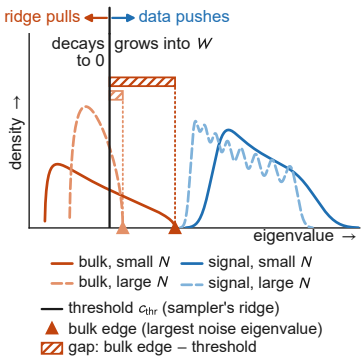
**(ii) Fixed point.** Under (H1) $\lambda^2\phi(\lambda)$ non-decreasing on the stable range (true for the accumulating rule, App. Lemma B.2) and (H2) $\lambda^2\phi(\lambda)$ exceeding $c_j/(\sigma_O^2\sigma_H^2)$ somewhere on it, at $\bar{\eta}$ frozen at each update's value (App. B.5, B.4), mode $j$ converges from every $s_j > 0$ with $\lambda_j < 2/\bar{\eta}$ to an equilibrium, interior iff $c_j > c_{\mathrm{thr}} = \sigma_O^2 C_\perp/\sigma_H^2$, with $C_\perp = \phi(\sigma_H^{-2})$ the ridge in the directions $W$ ignores. It is unique and globally attracting when $\lambda^2\phi$ is strictly increasing (the accumulating rule); under (H1) alone the equilibria form an interval.

**(iii) Destination.** For the fully mixed accumulating rule, as $\bar{\eta} \to 0$ after mixing, each mode's amplitude is the probabilistic-PCA maximum-likelihood amplitude for its $c_j$; that the modes pair with the top eigenvectors is not shown (App. B.5).

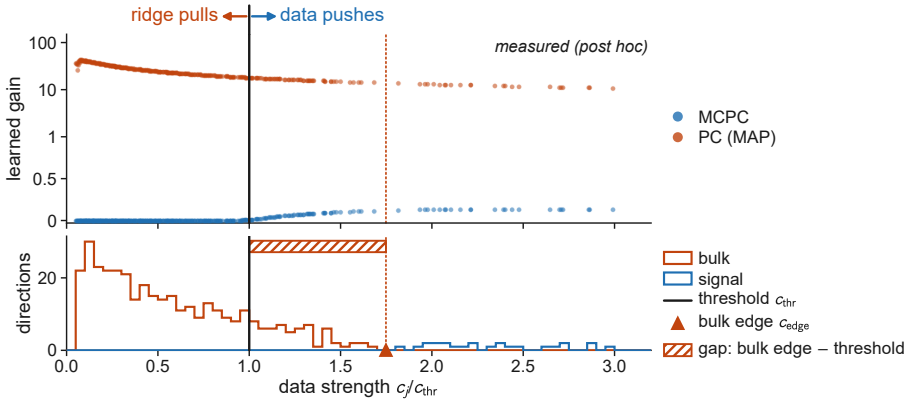
One flow gives both phases. Population modes cross the threshold first, finite-sample ones later or never.




(a) sampling to spectrum




(b) spectrum




(c) measured gain on (b)'s axis

Figure 3: **Sampling, ridge and data spectrum, schematic (a,b) and measured gain (c).** (a) The sampled latent cloud, covariance $C$, adds a ridge to the M-step's second moment (Prop. 3.1, Thm. 3.2); data directions above (blue) and below (red) the threshold $c_{\text{thr}}$. (b) The same threshold on the data spectrum, bulk and signal at small and large $N$. (c) Measured, post hoc, solvable tier (App. D), against $c_j/c_{\text{thr}}$. Top, learned gain $\sigma_H^2 \|u_j^\top W\|^2/c_j$ per data direction, MCPC and PC (MAP). Bottom, the directions' eigenvalues (bulk, bulk edge, signal outliers).

### 3.5 Step four — when a fluctuation direction is kept

**Theorem 3.4** (Memorization criterion, in the algorithm's own parameters)**.** Take the accumulating rule under Thm. 3.3's hypotheses without (H2), the model's noise variance equal to the generator's, a rank-$r$ signal with $r = o(n)$, so that with $\gamma = n/N$ the fluctuation bulk of $\widehat{\Sigma}_y$ has the Marchenko–Pastur edge $c_{\text{edge}} = \sigma_O^2(1+\sqrt{\gamma})^2$, and a free latent slot, $m > r$, aligned with the top fluctuation eigenvector of $\widehat{\Sigma}_y$. As $n, N \to \infty$ at fixed $\gamma$ (Baik and Silverstein, 2006), the top fluctuation direction (kept at the fixed point when (H2) holds, otherwise growing without bound; App. B.6) satisfies, eventually almost surely and away from equality,

$$\text{fluctuation retained} \iff C_\perp < \sigma_H^2\big(1+\sqrt{n/N}\big)^2, \qquad \text{where} \qquad C_\perp = \frac{\sigma_H^2\,\Phi(\sigma_H^{-2})}{1-\bar{\eta}/(2\sigma_H^2)}. \tag{7}$$

MAP ($C_\perp = 0$) and exact EM ($C_\perp = \sigma_H^2$) sit below the bar at every finite $N$. The finite chain lifts $C_\perp$ above $\sigma_H^2$ where inflation outweighs truncation, which at the experiments' step needs $\lambda_{\max}(H) \lesssim 31$ (App. D). Below $N_\infty^\star$ (Eq. 9) the chain still retains and its larger ridge only delays memorization; past $N_\infty^\star$ the ridge removes the direction (§4.1).

### 3.6 Step five — what the sampler's parameters can and cannot buy

**Corollary 3.5** (The prevention regime)**.** For the accumulating rule and every $\bar{\eta} < 2\sigma_H^2$, $\Phi \leq 1$ gives $C_\perp \leq \sigma_H^2/(1 - \bar{\eta}/(2\sigma_H^2))$ over every $b$, $S$ and $B$, so breaking (7) needs $\bar{\eta} \geq 2\sigma_H^2[1 - (1 + \sqrt{\gamma})^{-2}]$, while stability requires $\bar{\eta} < 2/\lambda_{\max}(H)$, $\lambda_{\max}(H) = \sigma_H^{-2} + s_{\max}^2/\sigma_O^2$ ($s_{\max}$ the top singular value of $W$), so

$$\exists \text{ stable prevention step} \iff \frac{\sigma_H^2 s_{\max}^2}{\sigma_O^2} < \frac{1}{2\sqrt{n/N} + n/N}. \tag{8}$$

Below the required step $C_\perp$ stays under the bar for every $b$, $S$, $B$ (a necessary condition, not an impossibility). As $s_{\max}$ grows the regime closes. At the fully mixed fixed point, preventing while keeping the top signal mode needs $c_1 > c_{\text{edge}}$ (App. B.7).

**Beyond Gaussian samplers.** The identity $\mathbb{E}[zz^\top] = m_y m_y^\top + C$ needs only a finite second moment, and for a multimodal posterior $C$ gains a between-mode ridge that any point-estimate E-step sets to zero (App. B.2, E.12).

### 3.7 Step six — data size

By Eq. (6) a mode grows at a rate set by its data eigenvalue minus $c_{\text{thr}}$. The fastest fluctuation mode sits at the Marchenko–Pastur edge, so its gap to the threshold, and the data size that closes it, are

$$\Delta_N = \sigma_O^2\left(1 + \sqrt{n/N}\right)^2 - c_{\text{thr}}, \qquad N_\infty^\star = \frac{n}{\left(\sqrt{c_{\text{thr}}/\sigma_O^2} - 1\right)^2} \quad (c_{\text{thr}} > \sigma_O^2). \tag{9}$$

More data slows that mode and, past $N_\infty^\star$, stops it; for exact EM ($c_{\text{thr}} = \sigma_O^2$) the gap never closes. Width $m$ only sets how many fluctuation directions can be held, $(m - r)_+$ (App. B.8).

## 4 Experiments

**Two tiers, one integrator.** The solvable tier draws data from a known linear–Gaussian generator, so the population optimum is known; MAP is the no-ridge arm. The sampled E-step runs one chain of $K$ steps per datum and hands its final state to the M-step (a window of one), so its ridge $C_K$ is Eq. (5) at $S = 1$. The image tier is MNIST; §4.3 leaves both tiers. Integrated rows run no chain; they integrate the theory's learning flow with each cell's constants. A cell reaches the bar or memorizes when that onset fires within the training budget (settings in App. D).

**Scoring.** A scripted rule scores every prediction the paper makes. The statements of §3 mostly hold; the misses concentrate in the asymptotic power laws (App. E.2).

### 4.1 Tests of Section 3

#### 4.1.1 Does the sampler realize the ridge? (Eq. 5)

We run the sampled chain on the solvable tier, next to exact posterior moments and MAP on the same cells. We measure the ridge the chain leaves in the Hebbian update, and how often each arm reaches the bar or memorizes. The chain leaves the predicted ridge with no free coefficient, and its larger ridge memorizes less, and later, than exact moments; MAP, with no ridge, is only a divergence control (Fig. 4, Table 1; draw, onsets and registered miss in App. E.6).

#### 4.1.2 Is the ridge the cause? (Eq. 7)

On the solvable linear tier we give the M-step the chain's ridge without any sampling, or the chain's noise without the ridge, and scale the ridge around its predicted switching dose. We measure whether the model keeps the finite-sample fluctuation directions, and which latent directions carry the contrast of Table 1. Past the data threshold the ridge alone reproduces MCPC, noise alone behaves like MAP, and retention switches at the predicted dose, with stiff directions carrying most of the contrast; below the threshold the ridge only delays memorization (Fig. 5; counts and the two registered misses, exact EM at MCPC's threshold and the step-size flip, are in App. E.6).

Table 1: **Arms by the ridge their M-step receives** (solvable tier; App. D). Exact moments carry the ridge $H^{-1}$, the chain its own ridge $C_K$ with or without its draw, and MAP none, as a divergence control (App. B.5). Integrated rows run one theory flow per cell and no chain; measured rows run the same cells, seeds and bars (App. E.3). Counts are cells where $\tau_{\text{gen}}$ (reaches the bar) or $\tau_{\text{mem}}$ (memorizes) fires. Excess NLL is held-out NLL over the oracle's, lower is better, read as an ordering. The $\Delta$ gap (post hoc) is the median final held-out minus training NLL per sample (App. E.8).

| arm | reaches the bar | memorizes | excess NLL (nats) | $\Delta$ NLL gap (nats) |
|---|---|---|---|---|
| integrated flow, no chain run | | | | |
| exact moments | 78/90 | 89/90 | 6.0 | 28.8 |
| chain without draw | 90/90 | 16/90 | 2.8 | 23.9 |
| chain | 90/90 | 33/90 | 3.3 | 24.8 |
| MAP | 0/90 | 90/90 | 42.6 | 31.4 |
| measured | | | | |
| exact moments | 70/90 | 89/90 | — | 28.8 |
| chain | 90/90 | 33/90 | 4.5 | 24.7 |
| MAP | 0/90 | 90/90 | 43 | 31.4 |

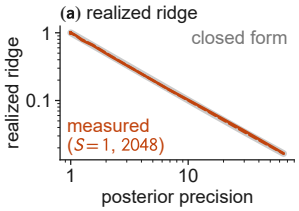



Figure 4: **Realized ridge against its closed form** (a fixed-weight probe, not a training run; App. E.6). (a) Realized ridge in $H$'s eigenbasis against posterior precision, measured at two windows $S$ (red; the two overlap), and Eq. (5) (grey band). Onsets are in App. Fig. 17 (App. E.5).

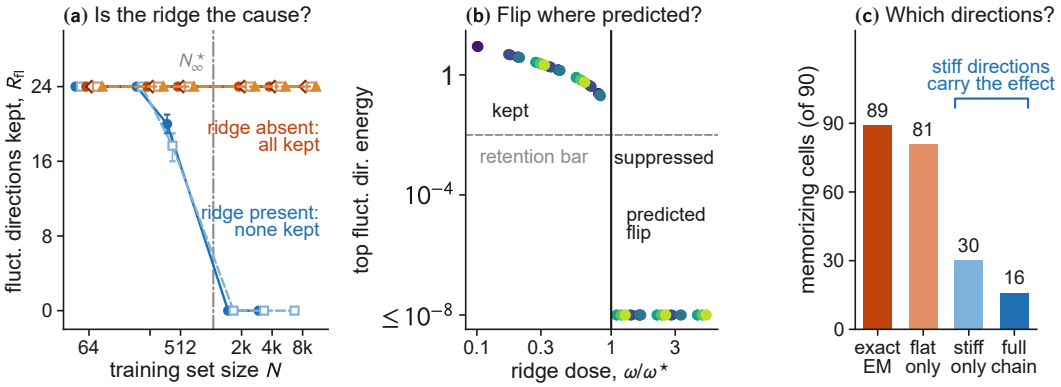



Figure 5: **Causal intervention, arms with and without the chain's ridge $C_K$ in the M-step statistic** (solvable tier; App. D; (a,b) pre-registered, (c) post hoc). (a) Fluctuation directions kept, $R_{\text{fl}}$, against $N$. With the ridge (blue), MCPC (filled), ridge only (open; posterior-mean moments plus $C_K$, no sampling). Without (warm), noise only (the chain's statistic minus $C_K$), exact posterior draws, MAP, compute-matched MAP; the four coincide. Dash-dot, $N_\infty^\star$. (b) Ridge-only arm at dose $\omega C_K$, energy of the top fluctuation direction against $\omega/\omega^\star$, $\omega^\star$ from Eq. (7); colour, $N$ from 64 (dark) to 8k (light). (c) Memorizing cells of Table 1's no-draw flow under exact EM's ridge, flat-direction inflation only, stiff-direction inflation only, and the full chain's ridge.

### 4.1.3 Can the sampler's parameters prevent memorization? (Eq. 8)

At a fixed training budget we vary the chain length, the number of chains and the step size, one at a time. We measure whether any setting prevents memorization, and whether it still keeps the signal. Only the step size prevents memorization, and only by also losing the signal, as Cor. 3.5 predicts (App. Table 9, block C; settings in App. D).

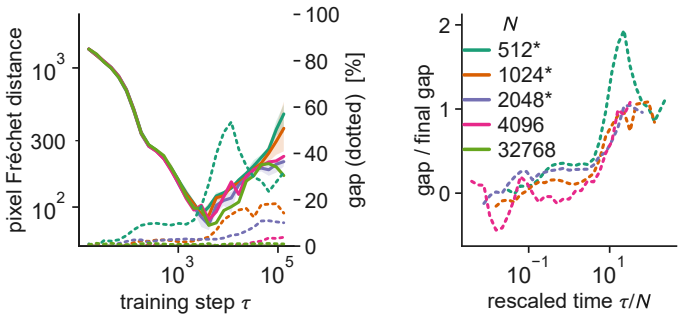



Figure 6: **Data-size curves, MCPC**, after Bonnaire et al. (2025, Fig. 2, left panel and inset) (CelebA, $N$ faces from the full train split, $2^{17}$ steps; $*$ marks three seeds, band their range; App. E.10.1). Left, pixel Fréchet distance of 5000 prior samples (positive members) to the test faces (solid) and the MAP-scored train/held-out gap (dotted) against training step $\tau$. Right, each $N$'s gap over its final value against $\tau/N$ (plateaus below 2.5% left off).

### 4.2 Data size and width (Eq. 9)

We vary the number of training samples on the strong-signal generator and the model width, and, post hoc, step up the chain's ridge. We measure the delay between reaching the bar and memorizing, the gap between held-out and training likelihood, and how many fluctuation directions the model holds. More data delays memorization, in sign; more ridge lowers the likelihood gap; width changes no direction's growth rate, only how many directions can be held (App. Fig. 12; the registered misses, a generalization onset flat in data size and the occupancy law's unit slope, in App. B.8.1; the likelihood gap in App. E.8). On CelebA, training on more faces delays the rise of the train/held-out gap while sample quality is best at about the same step (Fig. 6).

### 4.3 MCPC is also a generative model (a post-hoc finding)

We find, post hoc, that like a diffusion model a trained MCPC network maps noise to images. Given a latent $z$ drawn from the prior $\mathcal{N}(0, I)$, the tanh decoder trained on CelebA faces generates in one decoder pass, with no iterative reverse process, and its samples are faces that are not copies of training faces (Fig. 7; App. E.13.1). It needs little data. A model trained on a few hundred faces already generates faces (App. Fig. 19), and sample quality is best at about the same step for every data size (Fig. 6). It also interpolates smoothly between prior draws (Fig. 8). The same decoder trained as PC gives saturated, off-range samples from this prior, because its latents do not fill $\mathcal{N}(0, I)$ (Fig. 9).

## 5 Related work, discussion and conclusions

**Related work.** Generative-model memorization is read by nearest-neighbour replication (Carlini et al., 2023; Somepalli et al., 2023; Gu et al., 2025; Kadkhodaie et al., 2024); the ideal score machine can only reproduce its training data (Kamb and Ganguli, 2025), which a linear–Gaussian coder does not allow (App. C.3). Bidirectional PC (Oliviers et al., 2026) and PC–diffusion hybrids (Chen et al., 2024) may keep a ridge (conjecture). Non-Gaussian PC (Pinchetti et al., 2022) infers a point estimate, so no posterior covariance enters its weight update (§3.6). Concurrent anonymized temporal-PC work shares only Prop. 3.1.

**Beyond the linear model.** The theory of §3 does not apply to the tanh decoder of §4.3. There the gap between held-out and training likelihood falls as the training set grows, for both MCPC and PC (post hoc, App. E.10.2).

**Relation to diffusion models.** Bonnaire et al. (2025) explain diffusion's non-memorization by a timescale, $\tau_{\mathrm{mem}} \propto N$, with none at infinite time above a model-dependent $N^\star(p)$. Here a mechanism sits under the timescale. The ridge sets the threshold $c_{\mathrm{thr}}$, the directions it removes and a data size $N^\star_\infty$ past which memorization never occurs, at any width. Unlike noise-blind diffusion, whose noise-level posterior concentrates (Sahraee-Ardakan et al., 2026, §5), the ridge needs a spread latent posterior (post hoc, App. E.13.2). The learned variance is cut at $c_{\mathrm{thr}}$, where probabilistic PCA cuts at $\sigma_O^2$ (Tipping and Bishop, 1999) and a Wiener filter only shrinks (post hoc, App. Fig. 25), so the band $(\sigma_O^2, c_{\mathrm{thr}})$ stays unlearned (App. E.13).

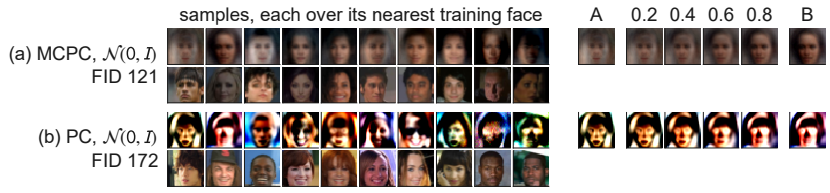



Figure 7: **Prior samples, (a) MCPC and (b) PC.** Latents $z \sim \mathcal{N}(0, I)$, one decoder pass. Below each sample, its nearest training face; right, interpolation between two draws (post hoc; App. E.13.1). The decoder is odd, so each $\pm$ pair is shown by its member with $\mu^\top \hat{x} > 0$, $\mu$ the mean training face.

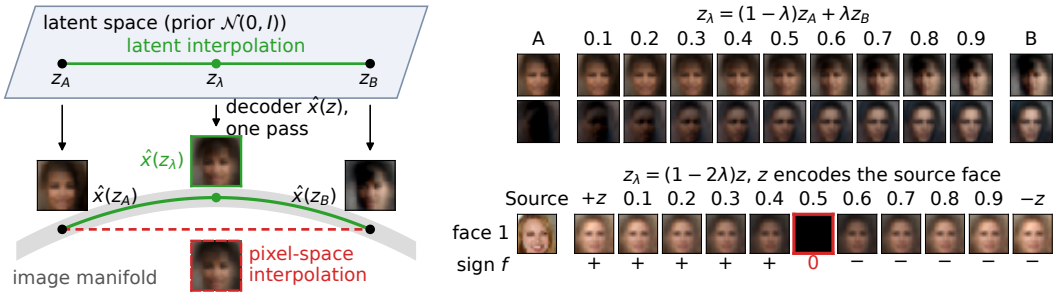



Figure 8: **Latent paths, MCPC** (layout of Ho et al. (2020, Fig. 8); App. E.11). Top, two pairs of prior draws with endpoints $sz$, $s = \text{sign}\, \mu^\top \hat{x}(z)$. Bottom, a test face's latent $z$ to $-z$, tiles $\hat{x}(|1 - 2\lambda|z)$, with the sign of $f$ below.

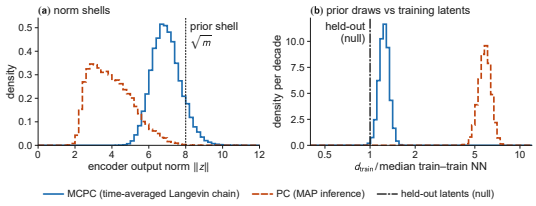



Figure 9: **Latents of the CelebA coders against the prior** $\mathcal{N}(0, I)$, MCPC and PC (post hoc; App. E.13.2). (a) Density of the latent norm, with the prior shell dotted. (b) Distance of each prior draw to the nearest training latent, over the typical train–train distance.

**Conclusions, answering Q1–Q3.** For the analysed variant, **A1.** The ridge has a closed form, exact EM's $H^{-1}$ times integrator inflation and window truncation. The inflation suppresses memorization within the training budget through the stiff directions (Table 1; post hoc, App. E.7), and past $N^\star_\infty$ through the flat direction (causal, §4.1.2). **A2.** At the fixed point, fluctuation is kept only while the unconstrained directions' ridge is below the Marchenko–Pastur scale (dose test held, §4.1.2; step-size flip within the registered tolerance at each $N$, missed on the pooled interval). At the default step the ridge reaches it only past $N^\star_\infty$, so below $N^\star_\infty$ the chain retains. Only the step size, not burn-in, chain length or chain count, can prevent memorization at the tested training budget, and only by also losing the signal. **A3.** Width adds $(m - r)_+$ fluctuation slots and changes no mode's rate. Once the ridge exceeds exact EM's, data pull the fluctuation edge below the threshold at $N^\star_\infty$ (Eq. 9). Measured, more data delays memorization, in sign.

**Recommendation.** Report how a sampled inference phase is reduced to a statistic for the weight update, and treat that reduction as a regularization hyperparameter. Keep the chain finite, and do not tune chain length or chain count against memorization.

**Limitations.** The analysis covers a two-layer predictive coder with one linear–Gaussian latent layer and tied maps (App. Table 5). Deep and nonlinear predictive coding (Salvatori et al., 2026; Pinchetti et al., 2025) are not analysed, the nonlinear case only briefly (App. E.10). Results are for the accumulating rule and the posterior-mean start of the original MCPC code. Below a data threshold MCPC only delays memorization, and preventing it costs likelihood relative to exact EM or a tuned ridge (App. E.4). Natural images largely lack the two-bulk premise, so there we test the mechanism's sign and use constants only where it holds (App. E.9). Image experiments use one seed, except the starred points of Fig. 6, and one FID implementation.

## ETHICS STATEMENT

This work studies when a generative model retains finite-sample structure of its training data, on public benchmark image datasets; the only personal content is CelebA's research release of public figures' faces, shown solely as nearest-neighbour references, and none of the models' samples copies a training face (§4.3). We claim no privacy guarantee, only what was measured: under the copy metric of App. E.13.1, no sample here copies a training image. The reduction of a sampled inference procedure to a statistic can instead shrink an implicit regularizer and move a model toward its training set, which is why we report it as a hyperparameter to control rather than as a safeguard.

## AI USE STATEMENT

We used generative AI assistance in two ways. It was used to aid and polish the writing, including drafting and editing prose and improving readability. It was also used for proving mathematical claims, in formulating and deriving the statements and proofs of Prop. 3.1, Theorems 3.2–3.4, Cor. 3.5 and App. Prop. B.4. The authors checked every proof, and every closed form stated in the paper is checked against direct numerical simulation, with the checks that failed reported in §4, App. E.2 and App. Table 8. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

All theoretical claims are stated with their assumptions inline and proved in Appendix B. The assumptions are measured in App. C.4 and E.9. The experimental setup, grid and order parameters are in Appendices C and D. The commands that rebuild the paper are in App. D.1. Every measured number in the main text is generated from the experiment output files by a script, so no value there can drift from the run that produced it; the appendix additionally quotes figures from the derivations and run records it cites.

## REFERENCES

Bai, Z.-D. and Y.-Q. Yin. "Limit of the smallest eigenvalue of a large dimensional sample covariance matrix." Annals of Probability 21 (1993): 1275–1294.

Baik, J., G. Ben Arous and S. Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices." Annals of Probability 33 (2005): 1643–1697.

Baik, J. and J. W. Silverstein. "Eigenvalues of large sample covariance matrices of spiked population models." Journal of Multivariate Analysis 97 (2006): 1382–1408.

Belkin, Mikhail, Daniel J. Hsu, Siyuan Ma and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." Proceedings of the National Academy of Sciences 116 (2019): 15849–15854.

Bishop, Christopher M. "Training with noise is equivalent to Tikhonov regularization." Neural Computation 7 (1995): 108–116.

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

Bonnaire, Tony, Raphaël Urfin, Giulio Biroli and Marc Mézard. "Why diffusion models don't memorize: The role of implicit dynamical regularization in training." Advances in Neural Information Processing Systems (2025).

Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito and Eric Wallace. "Extracting training data from diffusion models." USENIX Security Symposium (2023): 5253–5270.

Chen, Kaiyuan, Xingzhuo Guo, Yu Zhang, Jianmin Wang and Mingsheng Long. "CogDPM: Diffusion probabilistic models via cognitive predictive coding." International Conference on Machine Learning, PMLR 235 (2024): 6757–6775.

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

Durmus, A. and É. Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm." Annals of Applied Probability 27 (2017): 1551–1587.

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

Gavish, M. and D. Donoho. "The optimal hard threshold for singular values is $4/\sqrt{3}$." IEEE Transactions on Information Theory 60 (2014): 5040–5053.

Geyer, C. "Practical Markov chain Monte Carlo." Statistical Science 7 (1992): 473–483.

Gu, Xiangming, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin and Ye Wang. "On memorization in diffusion models." Transactions on Machine Learning Research (2025): n. pag.

Hastie, T., R. Tibshirani and J. Friedman. "The Elements of Statistical Learning." Springer, 2nd edition (2009).

Ho, Jonathan, Ajay Jain and P. Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020).

Johnstone, I. "On the distribution of the largest eigenvalue in principal components analysis." Annals of Statistics 29 (2001): 295–327.

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

Kamb, M. and Surya Ganguli. "An analytic theory of creativity in convolutional diffusion models." International Conference on Machine Learning, PMLR 267 (2025): 28795–28831.

Marchenko, V. A. and L. A. Pastur. "Distribution of eigenvalues for some sets of random matrices." Matematicheskii Sbornik 72 (1967): 507–536.

Millidge, Beren, Anil K. Seth and C. Buckley. "Predictive coding: A theoretical and experimental review." ArXiv abs/2107.12979 (2021): n. pag.

Mou, Wenlong, Liwei Wang, Xiyu Zhai and Kai Zheng. "Generalization bounds of SGLD for non-convex learning: two theoretical viewpoints." Conference on Learning Theory, PMLR 75 (2018): 605–638.

Nakajima, Shinichi, Ryota Tomioka, Masashi Sugiyama and S. Derin Babacan. "Condition for perfect dimensionality recovery by variational Bayesian PCA." Journal of Machine Learning Research 16 (2015): 3757–3811.

Neal, Radford M. and Geoffrey E. Hinton. "A view of the EM algorithm that justifies incremental, sparse, and other variants." Learning in Graphical Models (1998): 355–368.

Nielsen, Søren Feodor. "The stochastic EM algorithm: estimation and asymptotic results." Bernoulli 6 (2000): 457–489.

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

Oliviers, Gaspard, Mufeng Tang and Rafal Bogacz. "Bidirectional predictive coding." International Conference on Learning Representations (2026).

Pinchetti, Luca, Tommaso Salvatori, Yordan Yordanov, Beren Millidge, Yuhang Song and Thomas Lukasiewicz. "Predictive coding beyond Gaussian distributions." Advances in Neural Information Processing Systems (2022).

Pinchetti, Luca, C. Qi, Oleh Lokshyn, Gaspard Oliviers, Cornelius Emde, Mufeng Tang, Amine M'Charrak, Simon Frieder, Bayar Menzat, Rafal Bogacz, Thomas Lukasiewicz and Tommaso Salvatori. "Benchmarking predictive coding networks – made simple." International Conference on Learning Representations (2025).

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

Roberts, G. and R. Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations." Bernoulli 2 (1996): 341–363.

Roweis, S. and Zoubin Ghahramani. “A unifying review of linear Gaussian models.” Neural Computation 11 (1999): 305–345.

Sahraee-Ardakan, Mojtaba, M. Delbracio and P. Milanfar. “The geometry of noise: why diffusion models don’t need noise conditioning.” ArXiv abs/2602.18428 (2026): n. pag.

Salvatori, Tommaso, A. Mali, C. Buckley, Thomas Lukasiewicz, Rajesh P. N. Rao, K. Friston and Alexander G. Ororbia. “A survey on neuro-mimetic deep learning via predictive coding.” Neural Networks 195 (2026): 108161.

Sokal, Alan D. “Monte Carlo methods in statistical mechanics: foundations and new algorithms.” Functional Integration (1997): 131–192.

Somepalli, Gowthami, Vasu Singla, Micah Goldblum, Jonas Geiping and T. Goldstein. “Diffusion art or digital forgery? Investigating data replication in diffusion models.” IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023): 6048–6058.

Tipping, Michael E. and Charles M. Bishop. “Probabilistic principal component analysis.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (1999): 611–622.

Vempala, S. and Andre Wibisono. “Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices.” Advances in Neural Information Processing Systems 32 (2019): 8092–8104.

Wei, Greg C. G. and M. Tanner. “A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms.” Journal of the American Statistical Association 85 (1990): 699–704.

Whittington, James C. R. and R. Bogacz. “An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity.” Neural Computation 29 (2017): 1229–1262.

Yoon, Taehun, Jin Young Choi, Sehyun Kwon and Ernest K. Ryu. “Diffusion probabilistic models generalize when they fail to memorize.” ICML Workshop on Structured Probabilistic Inference & Generative Modeling (2023).

Zahid, Umais, Qinghai Guo and Z. Fountas. “Sample as you infer: Predictive coding with Langevin dynamics.” International Conference on Machine Learning, PMLR 235 (2024): 58105–58121.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, B. Recht and O. Vinyals. “Understanding deep learning requires rethinking generalization.” International Conference on Learning Representations (2017).

# APPENDIX CONTENTS


**A The algorithm, in pseudocode and in one picture; notation** **14**

**B The theory, derived** **14**

B.1 Posterior geometry . . . 15
B.2 The ridge identity . . . 15
B.3 Langevin covariance and the realized ridge . . . 19
B.4 Spectral M-step flow . . . 20
B.5 Fixed point, threshold, attraction . . . 21
B.6 Memorization criterion . . . 24
B.7 Prevention regime . . . 25
B.8 Capacity, dataset size, training time: the occupancy model . . . 26
B.9 Averaging rule and effective samples . . . 30

**C The setup in full** **31**

C.1 The coder's energy, and the closed forms inference descends on . . . 31
C.2 Sampling the inference phase, and the two ways to reduce it . . . 32
C.3 What generalization and memorization mean for this model class . . . 33
C.4 The three assumptions, priced . . . 33

**D Experimental setup** **33**

D.1 Experiment index and reproduction . . . 35

**E The results in full** **35**

E.1 What did not survive, and what stands in its place . . . 35
E.2 The scoreboard . . . 36
E.3 Pre-registered direct tests . . . 39
E.4 Causal ridge intervention . . . 40
E.5 One flow, both onsets . . . 42
E.6 Details of the tests in §4.1 . . . 43
E.7 Where the inflation acts (post hoc) . . . 44
E.8 Likelihood gap on the solvable tier (post hoc) . . . 44
E.9 Supporting readings . . . 46
E.10 Nonlinear coder . . . 46
E.11 The sign of the mean-face axis . . . 49
E.12 Mixture latent . . . 51
E.13 Correspondence with diffusion models . . . 52

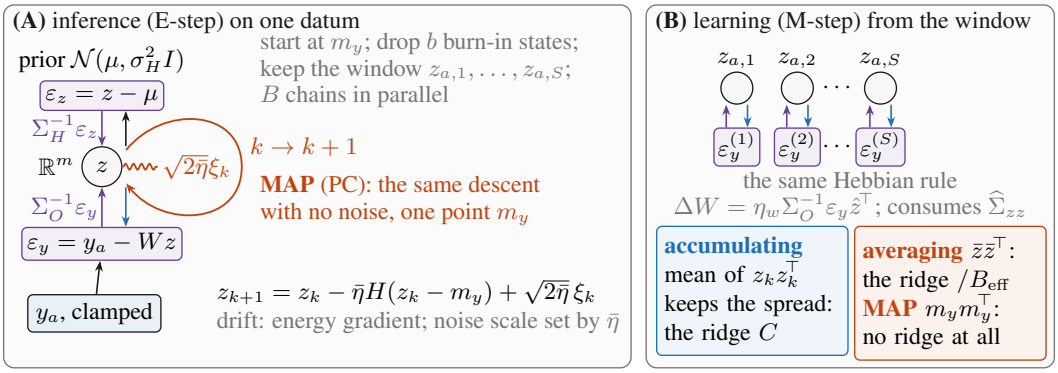



Figure 10: **Monte Carlo predictive coding as analysed here** (schematic). (A) The two-layer coder of §2.1: one latent layer $z$ over a clamped observation, error units at each level (purple); inference is the Langevin chain (3) on the latent, noise in vermillion; MAP is the same descent without noise. (B) The Hebbian M-step and its statistic, the latent second moment (27), under the accumulating rule, the averaging rule (App. B.9) and MAP. In App. E.10 the emission becomes $W_2 \tanh(W_1 z)$, its hidden layer without error units or noise.

## A THE ALGORITHM, IN PSEUDOCODE AND IN ONE PICTURE; NOTATION

This section gives the algorithm as pseudocode and as a picture, and collects the notation. Algorithms 1 and 2 are the loop that the code `pcgen` runs (`samplers.py`, `train.py`). Table 2 lists the symbols. The solvable grid runs $B = 1$, $b = K - 1$, $S = 1$ (Table 6). The averaging rule changes only line 4 of Alg. 2, which becomes $\widehat\Sigma_{zz} \leftarrow \frac{1}{N}\sum_a \bar z_a \bar z_a^\top$. It starts its chains the same way, at $m_y$ and with the full burn-in $b$ at every update. The PC/MAP baseline drops the noise. Started from $m_y$, its descent then does nothing, so $\bar z_a = m_{y_a}$ and $\widehat\Sigma_a = m_{y_a} m_{y_a}^\top$, which give $\widehat\Sigma_{mm}, \widehat\Sigma_{ym}$ (28). App. E.10 gives the nonlinear variant.

**Algorithm 1** MCPC inference (Langevin E-step)

**Require:** datum $y$; $W, \mu$; $\sigma_H, \sigma_O$; $\eta_{\text{rel}}$; $b, S, B$
$H \leftarrow \sigma_H^{-2} I + \sigma_O^{-2} W^\top W$
$\bar\eta \leftarrow 2\eta_{\text{rel}}/\lambda_{\max}(H)$
$m_y \leftarrow H^{-1}(\sigma_H^{-2}\mu + \sigma_O^{-2} W^\top y)$
$z^{(\iota)} \leftarrow m_y$ for $\iota = 1, \ldots, B$; $\bar z, \widehat\Sigma \leftarrow 0$
**for** $k = 1, \ldots, b + S$ **do**
  **for** $\iota = 1, \ldots, B$ **do**
    $g \leftarrow \Sigma_H^{-1}(z^{(\iota)} - \mu) - W^\top \Sigma_O^{-1}(y - W z^{(\iota)})$
    $z^{(\iota)} \leftarrow z^{(\iota)} - \bar\eta\, g + \sqrt{2\bar\eta}\,\xi_k^{(\iota)}$, $\xi_k^{(\iota)} \sim \mathcal{N}(0, I)$
    **if** $k > b$: $\bar z \mathrel{+}= \frac{z^{(\iota)}}{BS}$, $\widehat\Sigma \mathrel{+}= \frac{z^{(\iota)} z^{(\iota)\top}}{BS}$
  **end for**
**end for**
**return** $\bar z$, $\widehat\Sigma$

**Algorithm 2** MCPC training (accumulating rule + Hebbian M-step)

**Require:** $\{y_a\}_{a=1}^N$; $m$; $\eta_w$; updates $T$
$W_{ij} \sim \mathcal{N}(0, 1/m)$, $\mu \leftarrow 0$
**for** $t = 0, \ldots, T - 1$ **do**
  $(\bar z_a, \widehat\Sigma_a) \leftarrow$ Alg. 1$(y_a; W, \mu)$, all $a$
  $\widehat\Sigma_{zz} \leftarrow \frac{1}{N}\sum_a \widehat\Sigma_a$ ▷ accumulating
  $\widehat\Sigma_{yz} \leftarrow \frac{1}{N}\sum_a y_a \bar z_a^\top$
  $W \leftarrow W + \frac{\eta_w}{\sigma_O^2}\big(\widehat\Sigma_{yz} - W\widehat\Sigma_{zz}\big)$
  $\mu \leftarrow \mu + \frac{\eta_w}{\sigma_H^2}\big(\frac{1}{N}\sum_a \bar z_a - \mu\big)$
**end for**
**return** $W, \mu$

## B THE THEORY, DERIVED

This section proves the statements of §3 and the occupancy model. The proofs follow the order of the argument. Each proof names a hypothesis at the point where it uses it. The argument runs as one chain,

$$H^{-1} \Rightarrow C(H) \Rightarrow \dot s_j \Rightarrow c_{\text{thr}} \Rightarrow \text{retention} \Rightarrow \text{prevention},$$

and the occupancy model has its own chain,

$$\text{linearized flow} + \text{Marchenko–Pastur spectrum} + \operatorname{rank} W \le m \Rightarrow \tau_Q(m, N, t).$$

Table 3 lists each link with its statement, where it is proved and what it assumes.

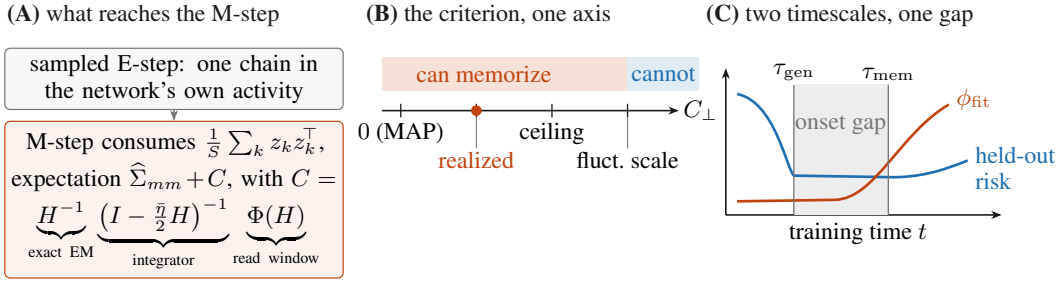



Figure 11: **Schematic of the thesis** (nothing here is a measurement). (A) The second moment a sampling E-step hands the Hebbian M-step: $\widehat{\Sigma}_{mm}$ plus the datum-independent term $C$ (Thm. 3.2). (B) One axis: the realized ridge $C_\perp$ (dot) against the bar $\sigma_H^2(1+\sqrt{n/N})^2$ (Thm. 3.4), with MAP at zero, the chain's ceiling and the fluctuation scale (Cor. 3.5). (C) Held-out risk (blue) and $\phi_{\text{fit}}$ (vermillion) over training for one $N$, with the onsets $\tau_{\text{gen}}$ and $\tau_{\text{mem}}$ (§4).

The experiments depart from the analysed protocol in two ways. The sampler may start away from the posterior mean, which is treated at the end of App. B.3. The M-step may use the averaging rule instead of the accumulating rule, which App. B.9 treats.

### B.1 Posterior geometry

This subsection shows that one matrix $H$ controls the posterior, the energy and the inference dynamics. For the isotropic model of §2.1 the energy (25) of App. C.1 is

$$\mathcal{F}(z,y)=\frac{1}{2\sigma_H^2}\|z-\mu\|^2+\frac{1}{2\sigma_O^2}\|y-Wz\|^2.$$

Its part quadratic in $z$ is $\frac{1}{2}z^\top Hz$ and its part linear in $z$ is $-z^\top h_y$, with

$$H:=\sigma_H^{-2}I+\sigma_O^{-2}W^\top W,\qquad h_y:=\sigma_H^{-2}\mu+\sigma_O^{-2}W^\top y,\qquad m_y:=H^{-1}h_y.$$

$H\succeq\sigma_H^{-2}I\succ 0$ is invertible for every $W$. Using $Hm_y=h_y$ to complete the square gives

$$\mathcal{F}=\tfrac{1}{2}(z-m_y)^\top H(z-m_y)+\text{const}(y),$$

so $p(z\mid y)\propto e^{-\mathcal{F}}$ is $\mathcal{N}(m_y,H^{-1})$, the posterior of §2.1. Since $\nabla_z\mathcal{F}=H(z-m_y)$ and $\nabla_z^2\mathcal{F}=H$, a single matrix that does not depend on $y$ plays three roles. It is the energy's Hessian, the posterior precision, and the rate at which inference relaxes along each direction. With $W=U\,\text{diag}(s_j)V^\top$,

$$H=\sum_j(\sigma_H^{-2}+s_j^2/\sigma_O^2)v_jv_j^\top+\lambda_\perp P_\perp$$

with $\lambda_\perp=\sigma_H^{-2}$, the sum over $s_j>0$ and $P_\perp$ the projector onto the null space of $W$. So every direction that $W$ does not see sits at the floor $\lambda_\perp$.

Suppose each datum has $B$ chains, indexed by $\iota$. Each chain runs $b$ burn-in steps and then records $S$ states. The accumulating rule then passes these statistics to the M-step,

$$\widehat{\Sigma}_{zz}=\frac{1}{NBS}\sum_{a=1}^{N}\sum_{\iota=1}^{B}\sum_{k=b+1}^{b+S}z_{a,\iota,k}z_{a,\iota,k}^\top,\qquad \widehat{\Sigma}_{yz}=\frac{1}{NBS}\sum_{a=1}^{N}\sum_{\iota=1}^{B}\sum_{k=b+1}^{b+S}y_a z_{a,\iota,k}^\top,\tag{10}$$

This is the rightmost expression of (27) with the chain index written out. It is compared against the reference statistics (28).

### B.2 The ridge identity

This subsection proves Prop. 3.1, that sampling adds a ridge $C$ to the M-step. The proposition's rate needs the exact law of the sampled state, so we derive that law first.

Table 2: **Notation.**

| symbol | meaning | symbol | meaning |
|---|---|---|---|
| Model (§2.1, §2.3) | | | |
| $z, y$ | latent in $\mathbb{R}^m$, observation in $\mathbb{R}^n$ | $n, m, N$ | observation dim., capacity, training-set size |
| $W$ | emission matrix, singular values $s_i$ | $Y$ | training set $\{y_a\}_{a=1}^N$ |
| $\xi$ | observation noise, $\mathcal{N}(0, \sigma_O^2 I)$ | $\sigma_H^2, \sigma_O^2$ | prior and likelihood variances |
| $H$ | posterior precision $\sigma_H^{-2} I + \sigma_O^{-2} W^\top W$ | $m_y$ | posterior mean $H^{-1}\sigma_O^{-2} W^\top y$ |
| $\lambda_i(H)$ | $\sigma_H^{-2} + s_i^2/\sigma_O^2$ | $\lambda_\perp$ | $\sigma_H^{-2}$, directions the likelihood ignores |
| $\Sigma_{\text{model}}$ | $\sigma_H^2 W W^\top + \sigma_O^2 I$ | $\Sigma_{\text{pop}}, \Sigma_{\text{emp}}$ | population and empirical covariance |
| $\phi_{\text{fit}}$ | fit fraction, the memorization order parameter (`f_mem` in the code), read as a sign, (29) | $\tau_{\text{gen}}, \tau_{\text{mem}}$ | generalization and memorization onsets |
| $\hat{y}$ | reconstruction $Wz$ (Fig. 2) | $t$ | training time, in weight updates |
| Sampler (§2.2) | | | |
| $\bar{\eta}$ | effective E-step step size | $\xi_k$ | chain noise at step $k$, $\mathcal{N}(0, I)$, (3) |
| $b, S, B$ | burn-in, window, chains per datum | $K$ | chain steps per datum, §4 |
| $\Sigma_\infty$ | chain stationary covariance $H^{-1}(I - \frac{\bar{\eta}}{2}H)^{-1}$ | $\bar{z}$ | window mean, averaging rule |
| $\widehat{\Sigma}_{zz}$ | latent second moment the M-step receives | $\widehat{\Sigma}_{yz}$ | observation–latent cross moment it receives |
| Statistics and ridge (§§3.2–3.3) | | | |
| $\widehat{\Sigma}_{mm}$ | $\frac{1}{N}\sum_a m_{y_a} m_{y_a}^\top$ | $\widehat{\Sigma}_{ym}$ | $\frac{1}{N}\sum_a y_a m_{y_a}^\top$ |
| $C$ | the realized ridge, $\mathbb{E}[\widehat{\Sigma}_{zz} \mid Y] - \widehat{\Sigma}_{mm}$, (5) | $W_C, W^{\text{MAP}}$ | frozen M-step target; the interpolant at $C = 0$ |
| $\Phi$ | window factor, (5) | $\rho_\lambda$ | $1 - \bar{\eta}\lambda$ |
| $\lambda_{\max}(H)$ | largest eigenvalue of $H$ | | |
| Spectral flow (§3.4) | | | |
| $\widehat{\Sigma}_y$ | $\frac{1}{N}\sum_a y_a y_a^\top$ | $c_j, u_j$ | eigenvalues ($c_1 \geq c_2 \geq \cdots$) and eigenvectors of $\widehat{\Sigma}_y$ |
| $s_j, v_j$ | singular values and right singular vectors of $W$ | $\lambda_j$ | $\sigma_H^{-2} + s_j^2/\sigma_O^2$ |
| $\phi$ | ridge shape, $C = \phi(H)$ | $\alpha, \eta_w$ | $\eta_w/\sigma_O^2$; M-step rate |
| $C_\perp$ | $\phi(\sigma_H^{-2})$, the ridge in the directions $W$ ignores | $c_{\text{thr}}$ | $\sigma_O^2 C_\perp/\sigma_H^2$, survival threshold of a mode |
| Criterion and prevention (§§3.5–3.6) | | | |
| $\gamma$ | aspect ratio $n/N$ | $c_{\text{edge}}$ | $\sigma_O^2(1+\sqrt{\gamma})^2$, upper Marchenko–Pastur edge |
| $r$ | signal rank | $s_{\max}$ | largest singular value of $W$, so $\lambda_{\max}(H) = \sigma_H^{-2} + s_{\max}^2/\sigma_O^2$ |
| Occupancy model (App. B.8) | | | |
| $\beta$ | $\alpha\sigma_H^2/\sigma_O^2$ | $s_0(m)$ | initial singular value of $W$ along a mode (the initialization scale) |
| $c_{\text{sig}}, s_G$ | signal eigenvalue; level it reaches at $\tau_{\text{gen}}$ | $\Delta_N$ | fluctuation-edge gap $c_{\text{edge}} - c_{\text{thr}}$ |
| $\tau_{\text{edge}}$ | time the edge mode reaches $s_M$ | $L_M(m)$ | $\log[s_M/s_{0,\text{edge}}(m)]$ |
| $N_\infty^\star$ | size at which $\Delta_N$ closes | $A^*, x^*$ | slope and $\sqrt{n/N_\infty^\star}$, (18) |
| $p_\gamma, c_-$ | Marchenko–Pastur density; its lower edge | $T_\gamma$ | tail mass of $p_\gamma$ |
| $R_{\text{fl}}(m, N)$ | fluctuation eigenmodes above threshold, capped by $m - r$ | $q_{m,N}$ | lower end of the retained band |
| $Q_{\text{mem}}, \mathcal{R}$ | fluctuation energy $\sum_{j\in\mathcal{R}} s_j^2$ over the retained set | $Q^*, \tau_Q$ | level; time $Q_{\text{mem}}$ reaches it |
| $A_\gamma$ | edge amplitude of $p_\gamma$, (22) | $N^\star(m)$ | boundary in the $(N, m)$ plane |

Table 3: **Where each statement is proved, and what it assumes.** "Aligned" is the invariant set of Thm. 3.3(i); that the flow reaches it is not proved (App. B.4).

| statement | proved in | hypotheses |
|---|---|---|
| Prop. 3.1 (ridge identity) | App. B.2, Lemma B.1 | datum-independent $C$; the $O_P(N^{-1/2})$ rate needs Lemma B.1's posterior-mean start |
| Thm. 3.2 (realized ridge) | App. B.3 | Assumption (ii); accumulating rule; $\bar{\eta}\lambda_{\max}(H) < 2$ |
| Thm. 3.3(i) (decoupling) | App. B.4 | centred data; expected flow; $C = \phi(H)$ at a frozen step |
| Thm. 3.3(ii) (fixed point) | App. B.5, Lemmas B.2, B.3 | (i) and the aligned set; (H1), (H2); $\bar{\eta}$ frozen; $\lambda_j < 2/\bar{\eta}$ |
| Thm. 3.3(iii) (destination) | App. B.5 | $\Phi \equiv 1$ first, then $\bar{\eta} \to 0$; per-mode amplitudes only |
| Thm. 3.4 (criterion) | App. B.6 | (H1); matched noise; fixed number of spikes $r$; $m > r$ and alignment; proportional limit |
| Cor. 3.5 (prevention) | App. B.7 | $\bar{\eta} < 2\sigma_H^2$; both reduction rules (App. B.9) |
| Prop. B.4 (occupancy) | App. B.8 | (T1)–(T3); linear regime $\sigma_H^2 s^2 \ll \sigma_O^2$ |

**Lemma B.1** (The remainder and its law)**.** Fix $W$, hence $H$ and every $m_{y_a}$, and run (3) for each datum $a$ and chain $\iota$ from $z_{a,\iota,0} = m_{y_a}$, the chains independent given $Y$. Then, conditional on $Y$,

$$z_{a,\iota,k} = m_{y_a} + d_{a,\iota,k}, \qquad d_{a,\iota,k} \sim \mathcal{N}(0, V_k), \qquad V_k = 2\bar{\eta}\sum_{j=0}^{k-1} P^{2j}, \qquad P := I - \bar{\eta}H,$$

exactly. The statistics (10) satisfy

$$\widehat{\Sigma}_{zz} = \widehat{\Sigma}_{mm} + C + R_N \quad \text{and} \quad \widehat{\Sigma}_{yz} = \widehat{\Sigma}_{ym} + R'_N \quad \text{with} \quad C = \frac{1}{S}\sum_{k=b+1}^{b+S} V_k$$

independent of $a$, $\mathbb{E}[R_N \mid Y] = \mathbb{E}[R'_N \mid Y] = 0$, and $R_N, R'_N = O_P\big((NB)^{-1/2}\big)$ whenever $\frac{1}{N}\sum_a \|y_a\|^2$ is bounded in probability. With one chain and one read state per datum ($B = S = 1$, the solvable tier's protocol) the law is $z_a = m_{y_a} + d_a$ with $d_a \sim \mathcal{N}(0, C)$, the single-draw law the integrator of §4 draws from.

*Proof.* Subtracting $m_{y_a}$ from (3) gives

$$d_{k+1} = Pd_k + \sqrt{2\bar{\eta}}\,\xi_k, \quad d_0 = 0, \qquad \text{so} \qquad d_k = \sqrt{2\bar{\eta}}\sum_{j<k} P^{k-1-j}\xi_j$$

is Gaussian with mean zero and, $P$ being symmetric, covariance $2\bar{\eta}\sum_{j<k} P^{2j} = V_k$. Expanding $z_{a,\iota,k}z_{a,\iota,k}^\top$ and averaging over $a, \iota, k$ gives $\widehat{\Sigma}_{zz} = \widehat{\Sigma}_{mm} + C + R_N$ with

$$R_N = \frac{1}{NB}\sum_{a,\iota} X_{a,\iota}, \qquad X_{a,\iota} = \frac{1}{S}\sum_{k=b+1}^{b+S}\Big[m_{y_a}d_{a,\iota,k}^\top + d_{a,\iota,k}m_{y_a}^\top + \big(d_{a,\iota,k}d_{a,\iota,k}^\top - V_k\big)\Big].$$

The $X_{a,\iota}$ have conditional mean zero and are conditionally independent across $(a, \iota)$, so

$$\mathbb{E}[\|R_N\|_F^2 \mid Y] = (NB)^{-2}\sum_{a,\iota}\mathbb{E}[\|X_{a,\iota}\|_F^2 \mid Y].$$

Since $V_{k+1} - V_k = 2\bar{\eta}P^{2k} \succeq 0$ at any step size, $V_k \preceq V_{b+S}$ for $k \le b+S$, and the Cauchy–Schwarz inequality over the $S$ terms with Gaussian fourth moments gives

$$\mathbb{E}[\|X_{a,\iota}\|_F^2 \mid Y] \le c_{\text{abs}}\big(\|m_{y_a}\|^2 \operatorname{tr} V_{b+S} + (\operatorname{tr} V_{b+S})^2\big),$$

$c_{\text{abs}}$ an absolute constant. Hence

$$\mathbb{E}[\|R_N\|_F^2 \mid Y] \le c_{\text{abs}}(NB)^{-1}\big(\tfrac{1}{N}\sum_a \|m_{y_a}\|^2 \operatorname{tr} V_{b+S} + (\operatorname{tr} V_{b+S})^2\big);$$

as $\|m_{y_a}\| \le \sigma_O^{-2}\|H^{-1}\|\|W\|\|y_a\|$ the bracket is bounded in probability, and Chebyshev's inequality gives $R_N = O_P((NB)^{-1/2})$. The same computation with $y_a d_{a,\iota,k}^\top$ gives $R'_N$. At $B = S = 1$ there is one term per datum and $C = V_{b+1}$. □

*Proof of Proposition 3.1, with the rate of the realized maximiser.* Beyond the main-text statement we show $\widehat{W} = W_C + O_P(N^{-1/2})$ when $C \succ 0$ and the remainders are $O_P(N^{-1/2})$, which for the sampler analysed here is Lemma B.1.

Write $\widehat{\Sigma}_{zz} = \frac{1}{N}\sum_a T_a$ with $\mathbb{E}[T_a \mid y_a] = m_{y_a} m_{y_a}^\top + C$. Because $C$ does not depend on $a$, averaging gives $\mathbb{E}[\widehat{\Sigma}_{zz} \mid Y] = \widehat{\Sigma}_{mm} + C$ exactly; with $R_N := \widehat{\Sigma}_{zz} - \mathbb{E}[\widehat{\Sigma}_{zz} \mid Y]$, $\mathbb{E}[R_N \mid Y] = 0$ and $\widehat{\Sigma}_{zz} = \widehat{\Sigma}_{mm} + C + R_N$. The frozen M-step minimizes

$$\mathcal{L}(W) = \frac{1}{2N}\sum_a \mathbb{E}[\|y_a - W z_a\|^2_{\Sigma_O^{-1}} \mid y_a],$$

whose gradient $-\Sigma_O^{-1}(\widehat{\Sigma}_{yz} - W\widehat{\Sigma}_{zz})$ the Hebbian update of (27) descends; the normal equation is $W\widehat{\Sigma}_{zz} = \widehat{\Sigma}_{yz}$.

For $C \succ 0$, $\widehat{\Sigma}_{mm} + C \succeq C$ gives $\|(\widehat{\Sigma}_{mm} + C)^{-1}\| \le \|C^{-1}\|$, and for $\|R_N\| \le 1/(2\|C^{-1}\|)$ the matrix $\widehat{\Sigma}_{zz}$ is invertible with $\|\widehat{\Sigma}_{zz}^{-1}\| \le 2\|C^{-1}\|$, so the resolvent identity

$$A^{-1} - A'^{-1} = A^{-1}(A' - A)A'^{-1} \quad \text{gives} \quad \widehat{\Sigma}_{zz}^{-1} = (\widehat{\Sigma}_{mm} + C)^{-1} + O(\|R_N\|).$$

With $\widehat{\Sigma}_{yz} = \widehat{\Sigma}_{ym} + O_P(N^{-1/2})$ and $R_N = O_P(N^{-1/2})$ (unbiasedness alone does not give the rate; the concentration of Lemma B.1 does),

$$\widehat{W} = \widehat{\Sigma}_{yz}\widehat{\Sigma}_{zz}^{-1} = W_C + O_P(N^{-1/2}).$$

For $C = 0$ the statistics are exact, the normal equation $W\widehat{\Sigma}_{mm} = \widehat{\Sigma}_{ym}$ may be singular, and its minimum-norm solution $\widehat{\Sigma}_{ym}\widehat{\Sigma}_{mm}^\dagger$ is the maximum-a-posteriori interpolant. □

**The identity is a target, not a destination.** (4) is the maximiser of one M-step with $H$ and the $m_{y_a}$ frozen at the current weights. It is not where the flow ends. Both statistics move with $W$, through $H$ and through the inferred means. Where the flow ends is given by Thm. 3.3. That theorem sends modes below threshold to exactly zero, not to the shrunken value a frozen ridge would give. The identity is also exact only when chains start at the posterior mean, as in Assumption (ii). From any other start the statistic gains the datum-dependent terms of App. B.3.

**Beyond Gaussian sampled laws.** Nothing in the identity uses the Gaussian law of Lemma B.1. For any sampled law $q_a$ of $z$ given $y_a$ with mean $m_{y_a}$ and finite covariance $C_a$,

$$\mathbb{E}_{q_a}[zz^\top] = m_{y_a} m_{y_a}^\top + C_a,$$

so the M-step target is $\widehat{\Sigma}_{mm} + \bar{C}$ with $\bar{C} = \frac{1}{N}\sum_a C_a$; Prop. 3.1's datum-independent $C$ is the case $C_a = C$. If $q_a = \sum_k \pi_{ak} q_{ak}$ is a mixture with component means $\mu_{ak}$ and covariances $C_{ak}$, so that $m_{y_a} = \sum_k \pi_{ak}\mu_{ak}$, the law of total covariance gives

$$C_a = \underbrace{\sum_k \pi_{ak} C_{ak}}_{\text{within-mode}} + \underbrace{\sum_k \pi_{ak}\big(\mu_{ak} - m_{y_a}\big)\big(\mu_{ak} - m_{y_a}\big)^\top}_{\text{between-mode}}.$$

The between-mode term is a ridge along the directions that join the modes. Its rank is at most one less than the number of modes. An E-step that returns a single point (MAP, or a Dirac-delta variational posterior) sets both terms to zero. Gaussianity enters only through the closed form of $C$ (Thm. 3.2), and through it the threshold $c_{\text{thr}}$ and criterion (7). A finite Langevin window that never crosses between modes realizes only the within-mode part. App. E.12 tests this in an experiment.

Table 4: **Four objects called a ridge, and what each changes.** The fixed point is where the expected M-step flow ends. The expected flow is the flow averaged over the sampler. The realized trajectory is the path that one run of updates takes. The last column says where each is measured.

| E-step | fixed point | expected flow | realized trajectory | measured in |
|---|---|---|---|---|
| MAP ($C = 0$) | no interior fixed point (App. B.5) | no ridge, so it drives toward the interpolant | deterministic, equals the expected flow | Table 1, zero-ridge rows |
| exact EM ($C = H^{-1}$) | ridge at the posterior covariance | ridge $H^{-1}$ | deterministic, equals the expected flow | Table 1, exact-moment row; §E.5, flip |
| finite-chain expected moments ($C = C_K$) | ridge moved by inflation and truncation (Eq. 5) | ridge $C_K$; sets the retention threshold | deterministic, equals the expected flow | Fig. 4 left; Table 1; §4.1.3; Fig. 15 |
| sampled draws ($C_K$ plus scatter) | as $C_K$, in expectation | as $C_K$ | adds zero-mean scatter, which adds memorization in the integrated flow at small $N$ (not analysed; App. E.7) | Table 1, with-draw and measured rows; Fig. 17, rows C, D |

### B.3 Langevin covariance and the realized ridge

This subsection proves Thm. 3.2, the closed form of the ridge that a finite Langevin chain produces. It then treats chains that do not start at the posterior mean.

*Proof of Theorem 3.2.* By Lemma B.1 the deviation $d_k = z_k - m_y$ has covariance $V_k = 2\bar{\eta}\sum_{j<k} P^{2j}$, $P = I - \bar{\eta}H$. The series converges iff $|1 - \bar{\eta}\lambda_i| < 1$ for every eigenvalue of $H$, i.e. $0 < \bar{\eta} < 2/\lambda_{\max}$; outside it $V_k$ grows without bound along $\lambda_{\max}$.

Inside, $V_\infty = 2\bar{\eta}(I - P^2)^{-1}$, and $I - P^2 = 2\bar{\eta}H(I - \frac{\bar{\eta}}{2}H)$ gives

$$V_\infty = H^{-1}(I - \tfrac{\bar{\eta}}{2}H)^{-1},$$

the $\Sigma_\infty$ of (3). Summing from $k$ onward, $V_k = (I - P^{2k})V_\infty$. All of $P$, $V_\infty$ are functions of $H$ and commute, so the window average is

$$C = \frac{1}{S}\sum_{k=b+1}^{b+S} V_k = V_\infty\Big[I - \frac{1}{S}\sum_{k=b+1}^{b+S} P^{2k}\Big] = H^{-1}\Big(I - \frac{\bar{\eta}}{2}H\Big)^{-1}\Big[I - \frac{P^{2(b+1)}(I - P^{2S})}{S(I - P^2)}\Big],$$

by the geometric sum; in the eigen-direction of $\lambda$, with $\rho = 1 - \bar{\eta}\lambda$, this is (5), and every factor is diagonal in the eigenbasis of $H$. □

For $|\rho| < 1$,

$$\Phi = 1 - \frac{1}{S}\sum_{j<S} \rho^{2(b+1+j)}$$

is one minus an average of terms in $[0, 1)$, so $0 < \Phi \le 1$; $\Phi \to 1$ as $b \to \infty$ at any $S$, and at finite burn-in the ceiling is attained only at $\rho = 0$.

**Start off the posterior mean.** For $z_0 = m_y + \delta$ the recursion is unchanged, so $\mathbb{E}[d_k \mid y] = P^k\delta$ with covariance still $V_k$, the window mean of $P^k\delta$ is $\bar{G}\delta$ with $\bar{G} = P^{b+1}(I - P^S)/(S(I - P))$, and with datum-dependent offsets $\delta_a$

$$\mathbb{E}\big[\widehat{\Sigma}_{zz} \mid Y\big] = \widehat{\Sigma}_{mm} + C + \widehat{\Sigma}_{m\delta}\bar{G}^\top + \bar{G}\widehat{\Sigma}_{\delta m} + \frac{1}{NS}\sum_a \sum_{k=b+1}^{b+S} P^k\delta_a\delta_a^\top P^k,$$

$$\widehat{\Sigma}_{m\delta} = \frac{1}{N}\sum_a m_{y_a}\delta_a^\top = \widehat{\Sigma}_{\delta m}^\top.$$

The cross term can have either sign. The last term is positive semi-definite. Both depend on the datum, so neither is a ridge. Hence Prop. 3.1 is exact only when chains start at the posterior mean, as in Assumption (ii). The two terms decay as $P^{b+1}$ and $P^{2(b+1)}$, slowest in the unconstrained directions, where $|\rho_\perp| = |1 - \bar{\eta}/\sigma_H^2|$ is closest to one on the branch $\bar{\eta}\lambda \le 1$.

Now suppose each chain starts where the previous M-step left it (a warm start), and the mean moves by $\Delta m_a$ per step. The offset then settles at

$$\delta_{a,\infty} = -(I - P^{b+S})^{-1}\Delta m_a, \qquad \text{which at } b = 0 \text{ gives} \qquad \bar{G}\delta_{a,\infty} = -P\,\Delta m_a/(S\bar{\eta}H)$$

per direction. It is largest in the unconstrained directions. The analysed protocol restarts every chain at the posterior mean at each M-step, so none of this enters its ridge.

### B.4 Spectral M-step flow

This subsection proves Thm. 3.3(i), that the M-step flow splits into independent scalar equations, one per data eigen-direction.

*Proof of Theorem 3.3(i).* With $\mu = 0$,

$$\begin{aligned} m_{y_a} &= \sigma_O^{-2} H^{-1} W^\top y_a, \\ \text{so} \quad \widehat{\Sigma}_{ym} &= \sigma_O^{-2} \widehat{\Sigma}_y W H^{-1} \\ \text{and} \quad \widehat{\Sigma}_{mm} &= \sigma_O^{-4} H^{-1} W^\top \widehat{\Sigma}_y W H^{-1}. \end{aligned}$$

The expected Hebbian flow at small M-step rate is

$$\dot{W} = \eta_w \sigma_O^{-2} (\mathbb{E}\widehat{\Sigma}_{yz} - W \mathbb{E}\widehat{\Sigma}_{zz}),$$

and Lemma B.1 gives

$$\dot{W} = \alpha\big[\widehat{\Sigma}_{ym} - W\big(\widehat{\Sigma}_{mm} + C\big)\big], \qquad C = \phi(H).$$

Let $\widehat{\Sigma}_y = \sum_j c_j u_j u_j^\top$ and $W = \sum_j s_j u_j v_j^\top$ with orthonormal $\{v_j\}$ in $\mathbb{R}^m$, over at most $m$ eigenvectors of $\widehat{\Sigma}_y$. Then

$$H = \sum_j \lambda_j v_j v_j^\top + \lambda_\perp P_\perp, \qquad \lambda_j = \sigma_H^{-2} + s_j^2/\sigma_O^2, \qquad Cv_j = \phi(\lambda_j) v_j,$$

and term by term

$$\widehat{\Sigma}_{ym} = \sum_j \frac{c_j s_j}{\sigma_O^2 \lambda_j} u_j v_j^\top, \qquad W\widehat{\Sigma}_{mm} = \sum_j \frac{c_j s_j^3}{\sigma_O^4 \lambda_j^2} u_j v_j^\top, \qquad WC = \sum_j s_j \phi(\lambda_j) u_j v_j^\top.$$

Every term of $\dot{W}$ is a multiple of some $u_j v_j^\top$, so the vector field is tangent to the set of such matrices with the $u_j, v_j$ fixed, and by uniqueness of ODE solutions the set is invariant. The coefficient of $u_j v_j^\top$ is

$$\dot{s}_j = \alpha s_j \Big[\frac{c_j}{\sigma_O^2 \lambda_j}\Big(1 - \frac{s_j^2}{\sigma_O^2 \lambda_j}\Big) - \phi(\lambda_j)\Big],$$

and $\sigma_O^2 \lambda_j = \sigma_O^2/\sigma_H^2 + s_j^2$ gives $1 - s_j^2/(\sigma_O^2\lambda_j) = 1/(\sigma_H^2 \lambda_j)$, which is (6). No property of $\phi$ was used beyond $C$ being a function of $H$. □

The proof does not assume that the statistics are frozen (App. B.2). This is why the result still holds although the driving operator is measured to drift by a factor of 5.26 over training.

**Invariance is not attraction.** The theorem makes the aligned set invariant, not attracting. Every per-mode statement below assumes that the flow reaches that set from its initialization. The only evidence that it does is the alignment measured on the independent instrument (App. E.9). That is a measurement, not a proof.

**The step rule the experiments run.** Both tiers recompute $\bar{\eta}$ at every update from $\lambda_{\max}(H)$, the largest posterior-precision eigenvalue at the current $W$ (Table 6). So along a trajectory $\phi$ is not one fixed function, and $\bar{\eta}\lambda \leq 1$ throughout. The theorem then describes each update's flow with $\phi$ frozen at that update's step. The fixed-step edge of (H2) is never approached. With $\bar{\eta}$ frozen at each update's value, the form of (H2) that applies is the reachable one of App. B.5. We do not extend the proof to this adaptive family. We apply its closed forms to it as an approximation, and App. E.9 measures how accurate that approximation is.

### B.5 Fixed point, threshold, attraction

This subsection proves Thm. 3.3(ii) and (iii). Each mode has one stable amplitude above a threshold and is erased below it, and the amplitude tends to the probabilistic-PCA value in the right limit.

**The hypotheses on the range the chain can visit.** A precision is admissible if $\lambda \in [\lambda_\perp, \lambda_e)$, $\lambda_e = 2/\bar{\eta}$, at a fixed step, and $\lambda \in [\lambda_\perp, 1/\bar{\eta}]$ under the relative step rule $\bar{\eta} = 2\eta_{\text{rel}}/\lambda_{\max}(H)$, $\eta_{\text{rel}} \leq \frac{1}{2}$, with $\bar{\eta}$ frozen at the current update's value, which gives $\bar{\eta}\lambda_j \leq 1$ for every mode. The hypotheses of Thm. 3.3 are

- (H1) $\lambda^2\phi(\lambda)$ non-decreasing on the admissible range, and
- (H2) $\sup_{\text{admissible}} \lambda^2\phi(\lambda) > c_j/(\sigma_O^2\sigma_H^2)$.

For the accumulating rule, (H1) holds on the whole stable range at every $b, S$ (Lemma B.2), and under the step rule (H2) reads $c_j/(\sigma_O^2\sigma_H^2) < 2/\bar{\eta}$, independent of $b$ and $S$ (Lemma B.3(c)). The fixed-step supremum $4(2b+S+1)/\bar{\eta}$ (Lemma B.3(b)) is approached only at the oscillating edge, which the step rule never visits. We keep it as a remark.

Throughout, $u := \bar{\eta}\lambda \in (0,2)$, $\rho = 1-u$, and $G(u) := \bar{\eta}\,\lambda^2 C(\lambda)$ for the accumulating rule. Since, with $v_\infty := 1/(\lambda(1-\bar{\eta}\lambda/2))$, $\lambda^2 v_\infty = \bar{\eta}^{-1}\,2u/(2-u)$,

$$G(u) = \frac{2u}{2-u}\,\Phi(u), \qquad \Phi(u) = \frac{1}{S}\sum_{j=b+1}^{b+S}\left(1-\rho^{2j}\right) \in [0,1], \tag{11}$$

and at fixed $\bar{\eta}$ (H1) is equivalent to $G$ non-decreasing on $(0,2)$.

**Lemma B.2** (Monotonicity of the accumulated ridge)**.** For all integers $b \geq 0$, $S \geq 1$,

$$G(u) = \frac{2u^2}{S}\sum_{j=b+1}^{b+S}\sum_{i=0}^{j-1}(u-1)^{2i}, \tag{12}$$

and $G$ is strictly increasing on $(0,2)$. Hence (H1) holds on the whole stable range.

*Proof.* **Identity.** Since $2-u = 1+\rho$ and $u = 1-\rho$,

$$1-\rho^{2j} = (1-\rho)(1+\rho)\sum_{i<j}\rho^{2i} = u(2-u)\sum_{i<j}\rho^{2i};$$

substitute into (11).

- **Case** $u \in (0,1]$ ($\rho \in [0,1)$). Here $\rho^2 = (1-u)^2$ is non-increasing in $u$, so $\Phi$ is non-decreasing, and $\Phi > 0$ because $\rho^2 < 1$; $2u/(2-u)$ is positive and strictly increasing, so the product is strictly increasing.
- **Case** $u \in [1,2)$ ($\rho \in (-1,0]$). With $t := u-1 \in [0,1)$, $G = \frac{2}{S}(1+t)^2\sum_j\sum_{i<j}t^{2i}$ is a non-constant polynomial in $t$ with non-negative coefficients, hence strictly increasing.

Both cases contain $u=1$ and $G$ is continuous, so $G$ is strictly increasing on $(0,2)$. □

**Lemma B.3** (End values and the reachable bound)**.** For all $b \geq 0$, $S \geq 1$:

(a) $0 < \lambda^2 C \leq (2b+S+1)\,\bar{\eta}\lambda^2 \cdot \frac{2}{2-u}$, so $\lambda^2 C \to 0$ as $u \to 0^+$; more precisely $\lambda^2 C = (2b+S+1)\bar{\eta}\lambda^2\,(1+O(u))$.

(b) $\lim_{u\to 2^-} \lambda^2 C(\lambda) = 4(2b+S+1)/\bar{\eta}$. This is the supremum over $(0,2)$ and it is not attained.

(c) $\sup_{u\in(0,1]} \lambda^2 C(\lambda) = \lambda^2 C\big|_{u=1} = 2/\bar{\eta}$, independent of $b$ and $S$, and $u=1$ is the unique maximiser on $(0,1]$.

*Proof.* **(a)** By the even-exponent Bernoulli inequality $(1-u)^{2j} \geq 1-2ju$ for all real $u$, so

$$\Phi \leq \frac{u}{S} \sum_{j=b+1}^{b+S} 2j = u\,(2b+S+1);$$

insert into (11). The asymptotic form follows from (12), since $\sum_{i<j}(u-1)^{2i} \to j$ as $u \to 0$.

**(b)** At $u=2$ in (12), $(u-1)^{2i}=1$ and

$$G(2^-) = \frac{8}{S} \sum_{j=b+1}^{b+S} j = \frac{8}{S}\big(S(b+1) + \tbinom{S}{2}\big) = 4(2b+S+1);$$

by Lemma B.2 this is the supremum and it is not attained on $(0,2)$.

**(c)** At $u=1$, $\rho=0$ and every $\rho^{2j}$ with $j \geq b+1 \geq 1$ vanishes, so $\Phi(1)=1$ and $G(1)=2$; strict monotonicity gives $G(u)<2$ on $(0,1)$. □

*Proof of Thm. 3.3(ii).* **What is shown.** On the aligned invariant set of part (i), with $\bar{\eta}$ frozen (so $\phi$ is one fixed function) and for initial conditions with $s_j > 0$ and $\lambda_j < 2/\bar{\eta}$, each scalar mode equation (6) has the stated equilibria and attraction. Reaching the aligned set, and the adaptive step, are not covered (App. B.4).

For $s_j > 0$ let

$$g_j(\lambda) := c_j/(\sigma_O^2\sigma_H^2) - \lambda^2\phi(\lambda).$$

Multiplying the bracket of (6) by $\lambda_j^2 > 0$ gives $\operatorname{sign} \dot{s}_j = \operatorname{sign} g_j(\lambda_j)$, a function of $\lambda_j$ alone since $\lambda_j$ is strictly increasing in $s_j > 0$; by (H1) $g_j$ is non-increasing. At the floor $\lambda_\perp^2 = \sigma_H^{-4}$, so

$$g_j(\lambda_\perp) > 0 \iff c_j > \sigma_O^2\phi(\lambda_\perp)/\sigma_H^2 = c_{\mathrm{thr}}.$$

**Existence.** If $c_j > c_{\mathrm{thr}}$, $g_j(\lambda_\perp) > 0$ while (H2) gives an admissible $\lambda$ with $g_j(\lambda) < 0$; $g_j$ is continuous, so by the intermediate value theorem it has a root $\lambda_j^\star$ strictly between.

**Uniqueness.** For the accumulating rule $g_j$ is strictly decreasing (Lemma B.2), so the root is unique; under (H1) alone the zero set is an interval and the argument below applies to its end points.

**Global attraction.** $\dot{s}_j > 0$ for $\lambda_j < \lambda_j^\star$ and $\dot{s}_j < 0$ for $\lambda_j > \lambda_j^\star$, so $s_j$ moves monotonically toward the $s_j^\star$ with $\lambda(s_j^\star) = \lambda_j^\star$ from every $s_j > 0$ with $\lambda_j < 2/\bar{\eta}$ at the frozen step (outside that range $C$ is undefined), and cannot cross it, because solutions of a scalar autonomous ODE with locally Lipschitz right-hand side do not cross equilibria. A monotone bounded trajectory converges to an equilibrium, and the only one in $s_j > 0$ is $s_j^\star$.

**Below the threshold.** If $c_j \leq c_{\mathrm{thr}}$ then $g_j \leq 0$ on the admissible range by (H1), strictly for $\lambda > \lambda_\perp$ under the accumulating rule. Then $s_j$ decreases monotonically to an equilibrium, and the only one is $s_j = 0$. So the mode is erased. □

**Where the bound fails there is no destination.** Where $\lambda^2\phi(\lambda)$ is instead bounded above on the admissible range by a constant below $c_j/(\sigma_O^2\sigma_H^2)$ – which includes $\phi = 0$ exactly, and, at a step frozen at $\bar{\eta}$, the averaging rule on the branch $\bar{\eta}\lambda \leq 1$ whenever $S\bar{\eta} > 2\sigma_O^2\sigma_H^2/c_j$, since there $\lambda^2\phi_{\mathrm{avg}} \leq 2/(S\bar{\eta})$ (Lemma B.5) – $g_j$ stays positive at every admissible $\lambda$. Then there is no equilibrium to converge to, and the mode's amplitude grows without bound. This is what the zero-ridge control variant measures. In 90 of 90 settings $(N, m, \text{seed})$ it never reaches its generalization bar (§4). Its $\phi_{\mathrm{fit}}$ onset is a divergence time, set partly by how $\phi_{\mathrm{fit}}$ is defined.

The same holds, at a step frozen at $\bar{\eta}$, for the accumulating rule itself. It applies to a mode with $c_j/(\sigma_O^2\sigma_H^2) > 2/\bar{\eta}$ on the branch $\bar{\eta}\lambda \le 1$. So on that branch the two rules differ only by the factor $S$ between Lemma B.3(c) and Lemma B.5, not in kind. Under the step rule the bound does not stay frozen. As the growing mode raises $\lambda_{\max}(H)$, $\bar{\eta}$ falls. So a fluctuation mode grows only until it overtakes the top mode, and there equilibrium returns.

**The ridge enters with a minus sign.** Above the threshold a mode's initial growth rate is the amount by which its drive eigenvalue exceeds the realized penalty,

$$\nu(c) = \alpha\,\sigma_H^2\left[\frac{c}{\sigma_O^2} - \frac{C_\perp}{\sigma_H^2}\right], \qquad C_\perp := C \text{ in the directions the likelihood is silent about.} \tag{13}$$

So the ridge enters the rate with a minus sign. It slows growth and speeds decay. It moves both where the mode ends and how fast it gets there. It can also push the rate through zero. That changes the mode's fate, not just its amplitude.

With $\beta := \alpha\sigma_H^2/\sigma_O^2$ the same rate reads $\nu(c) = \beta(c - c_{\text{thr}})$. It is the exponential rate of the amplitude $s_j$. (It is the linearization of (6) at $s_j = 0$, derived as (14) in App. B.8.) Any squared amplitude, such as $\sigma_H^2\|u^\top W\|^2$, grows at $2\nu(c)$. The rate does not depend on capacity. So if an onset depends on capacity, it does so only through the initial amplitude.

**Maximum-likelihood limit, proof of Thm. 3.3(iii).** We show the following. Take each mode on the aligned set, with whichever eigenvalue $c_j$ of $\widehat{\Sigma}_y$ it carries. Its stationary amplitude equals the probabilistic-PCA maximum-likelihood amplitude for that $c_j$. We do not show that the $m$ modes pair with the top $m$ eigenvectors, which probabilistic PCA's global maximum also requires.

In the fully-mixed limit $\Phi \equiv 1$,

$$\phi_\infty = v_\infty = 1/(\lambda(1 - \bar{\eta}\lambda/2))$$

and $\lambda^2\phi_\infty = \lambda/(1 - \bar{\eta}\lambda/2)$ is strictly increasing and unbounded as $\lambda \to 2/\bar{\eta}$, so (H1) and the fixed-step (H2) hold for every $c_j > 0$. The fixed-point condition

$$c_j/(\sigma_O^2\sigma_H^2\lambda^2) = 1/(\lambda(1 - \bar{\eta}\lambda/2)),$$

multiplied by $\lambda^2(1 - \bar{\eta}\lambda/2)$, is $c_j(1 - \bar{\eta}\lambda/2) = \sigma_O^2\sigma_H^2\lambda$, so

$$\lambda_j^\star = \frac{c_j}{\sigma_O^2\sigma_H^2 + \frac{\bar{\eta}}{2}c_j}, \qquad s_j^{\star 2} = \sigma_O^2\left(\lambda_j^\star - \sigma_H^{-2}\right),$$

in the stable range since $\bar{\eta}\lambda_j^\star < 2$, and interior iff $\lambda_j^\star > \sigma_H^{-2}$. Then $\bar{\eta} \to 0$ gives $\lambda_j^\star \to c_j/(\sigma_O^2\sigma_H^2)$ and $\sigma_H^2 s_j^{\star 2} \to (c_j - \sigma_O^2)_+$. The positive part appears because a mode with $c_j \le \sigma_O^2$ has no interior root and is erased. This limit is the probabilistic-PCA maximum-likelihood amplitude for eigenvalue $c_j$ of $\widehat{\Sigma}_y$ at the assumed noise variance (Tipping and Bishop, 1999).

The two limits do not commute. The order above is the fully-mixed limit first ($b \to \infty$ at fixed $\bar{\eta}$, or $S \to \infty$) and then $\bar{\eta} \to 0$. Taken the other way, $\bar{\eta} \to 0$ at fixed $b$ and $S$ sends $\rho \to 1$ and, by Lemma B.3(a), $\Phi(\lambda) \to 0$ and $C(\lambda) = (2b + S + 1)\bar{\eta}\,(1 + O(\bar{\eta})) \to 0$ in every direction. The chain then never leaves the posterior mean and the ridge vanishes. The flow tends to the zero-ridge case above, which has no interior fixed point, and not to probabilistic PCA. At finite window the root equation carries $\Phi(\lambda^\star)$ and is not explicit.

**Effective degrees of freedom.** On the invariant set $\widehat{\Sigma}_{mm}$ and $C$ are diagonal in $\{v_j\}$ with $(\widehat{\Sigma}_{mm})_{jj} = c_j s_j^2/(\sigma_O^4\lambda_j^2)$ and $C_j := \phi(\lambda_j)$. At an interior equilibrium $c_j/\lambda_j^2 = \sigma_O^2\sigma_H^2 C_j$, so

$$(\widehat{\Sigma}_{mm})_{jj} = \sigma_H^2 s_j^2 C_j/\sigma_O^2$$

and, the ridge cancelling,

$$\begin{aligned}(\widehat{\Sigma}_{mm})_{jj}/((\widehat{\Sigma}_{mm})_{jj} + C_j) &= (\sigma_H^2 s_j^2/\sigma_O^2)/(1 + \sigma_H^2 s_j^2/\sigma_O^2) \\ &= 1 - 1/(\sigma_H^2\lambda_j).\end{aligned}$$

Erased modes and the null space of $W$ contribute zero, so

$$d_{\text{eff}} = \operatorname{tr}[\widehat{\Sigma}_{mm}(\widehat{\Sigma}_{mm} + C)^{-1}] = \sum_j (1 - 1/(\sigma_H^2\lambda_j^\star))_+,$$

a graded functional of the data spectrum, not a count.

### B.6 MEMORIZATION CRITERION

This subsection proves Theorem 3.4, the rule that says when a noise direction is memorized.

*Proof of Theorem 3.4.* **What is used.** Only (H1) of Thm. 3.3, which holds for the accumulating rule by Lemma B.2. (H2) is not needed. Two further hypotheses, both in the statement of Thm. 3.4, are needed for a fluctuation direction to be carried at all.

- a free latent slot, $m > r$, since on the aligned set with the strongest eigenvectors paired first the $r$ signal directions take $r$ of the $m$ slots and $m \le r$ leaves none for fluctuation; and
- alignment, a slot paired with the top fluctuation eigenvector $u_{r+1}$ of $\widehat{\Sigma}_y$, which is assumed and not proved (App. B.4).

**Per mode.** If $c_j > c_{\mathrm{thr}}$ then $g_j(\lambda_\perp) > 0$ and the mode grows away from zero, to the interior equilibrium of Thm. 3.3(ii) when (H2) holds and without bound when it fails; either way it is retained. If $c_j \le c_{\mathrm{thr}}$ it is erased (App. B.5). Retention is therefore monotone in $c_j$, and some fluctuation direction is retained iff the largest fluctuation eigenvalue $c_{r+1}$ exceeds $c_{\mathrm{thr}} = \sigma_O^2 C_\perp/\sigma_H^2$, $C_\perp := \phi(\lambda_\perp)$.

**The edge.** The generator is a finite-rank spiked model,

$$\Sigma_{\mathrm{pop}} = \sigma_O^2 I + \sum_{\ell \le r} w_{\mathrm{scale}}^2 \sigma_H^2 u_\ell u_\ell^\top$$

with $r$ fixed, the case the cited results cover. In the proportional limit $n, N \to \infty$ at $n/N \to \gamma$, the $(r{+}1)$-th sample eigenvalue converges almost surely to the upper Marchenko–Pastur edge $c_{\mathrm{edge}} = \sigma_O^2(1+\sqrt{\gamma})^2$ (Marchenko and Pastur, 1967), whatever the spike strengths (Baik and Silverstein, 2006); for the null case see Bai and Yin (1993). So $c_{r+1} \to c_{\mathrm{edge}}$ a.s. Two cases follow.

- if $c_{\mathrm{edge}} > c_{\mathrm{thr}}$, then $c_{r+1} > c_{\mathrm{thr}}$ for all large $n$ and the direction is retained;
- if $c_{\mathrm{edge}} < c_{\mathrm{thr}}$, every fluctuation eigenvalue is eventually below $c_{\mathrm{thr}}$ and all are erased.

The equivalence

$$\text{retained} \iff C_\perp < \sigma_H^2(1+\sqrt{\gamma})^2$$

is therefore a statement about the limit, eventually almost surely, with the boundary case $c_{\mathrm{edge}} = c_{\mathrm{thr}}$ undecided; it is not a finite-$N$ guarantee (next paragraph). For the accumulating rule (5) at $\lambda_\perp = \sigma_H^{-2}$ gives

$$C_\perp = \sigma_H^2 \Phi(\lambda_\perp)/(1 - \bar{\eta}/(2\sigma_H^2)),$$

which is (7). □

The edge is a limit object. It holds as $n$ and $N$ grow together. At finite $n, N$ the top noise eigenvalue fluctuates around the edge on the Tracy–Widom scale, of relative order $N^{-2/3}$ (Johnstone, 2001), with the spiked case in Baik et al. (2005). Its right tail can place it above a bar that the limit puts it below. So near the bar (7) is an asymptotic statement, not a finite-$N$ guarantee.

Maximum-a-posteriori inference has $C_\perp^{\mathrm{MAP}} = 0$. Exact EM has $C_\perp^{\mathrm{EM}} = \sigma_H^2$. This lies on the retaining side by $\sigma_H^2(2\sqrt{\gamma}+\gamma) > 0$ at finite $N$, and the margin vanishes only as $N \to \infty$. So only the inflation of the finite chain can carry $C_\perp$ across the bar. Take the solvable tier's dimension $n = 128$ and a fixed absolute step $\bar{\eta} = 0.05$. (This is not the tier's relative step rule, under which $N_\infty^\star \approx 1078$–$1326$, App. E.4, E.7.) Then the condition would first bite at a dataset of order $7.9 \times 10^5$. This is the edge boundary $N_\infty^\star$ of Eq. (9), derived at a chosen step.

Nothing diverges at $\gamma = 1$. The onset is set by the bulk edge, which is smooth through $\gamma = 1$. Modes near the lower edge sit below threshold and are erased. What diverges is the memorization time, at the point where the realized ridge equals the fluctuation scale. Over this grid the exact-moment $C_\perp$ never reaches that scale. Under it the top noise direction is retained at every window, so any boundary a finite run finds there is a contour of the memorization time. The ridge of the finite chain does cross, between $N = 1024$ and 2048 at the protocol's step (integrated). Measured past $N_\infty^\star$, MCPC retains the top fluctuation direction in 0 of 6 settings $(N, m, \text{seed})$ (App. E.4).

### B.7 PREVENTION REGIME

This subsection proves Corollary 3.5, which says when some step size prevents memorization.

*Proof of Corollary 3.5.* **Ceiling.** Stability in the unconstrained direction is $\bar{\eta} < 2\sigma_H^2$, where $0 < \Phi(\lambda_\perp) \le 1$ (App. B.3). The chain count does not enter $C$ (Lemma B.1) and $b, S$ enter only through $\Phi$, so

$$C_\perp \le \sigma_H^2/(1 - \bar{\eta}/(2\sigma_H^2)),$$

approached as $b$ or $S$ grows and attained where $\Phi(\lambda_\perp) = 1$, at $\bar{\eta} = \sigma_H^2$ ($\rho_\perp = 0$).

**Necessary step.** Prevention is $C_\perp \ge \sigma_H^2(1 + \sqrt{\gamma})^2$; by the ceiling it requires

$$1 - \bar{\eta}/(2\sigma_H^2) \le (1 + \sqrt{\gamma})^{-2}, \quad \text{i.e.} \quad \bar{\eta} \ge \bar{\eta}_{\min} := 2\sigma_H^2[1 - (1 + \sqrt{\gamma})^{-2}] < 2\sigma_H^2.$$

**Stability.** Euler–Maruyama stability requires

$$\bar{\eta} < \bar{\eta}_{\max} := 2/\lambda_{\max} = 2\sigma_H^2/(1 + \sigma_H^2 s_{\max}^2/\sigma_O^2).$$

**Feasibility.** A preventing step exists iff $\bar{\eta}_{\min} < \bar{\eta}_{\max}$. Necessity is the two previous steps. For sufficiency, any $\bar{\eta} \in (\bar{\eta}_{\min}, \bar{\eta}_{\max})$ has a ceiling strictly above the bar and $C_\perp$ tends to it as $b \to \infty$, so a finite $b$ violates (7). With $A := (1 + \sqrt{\gamma})^2$ and $q := \sigma_H^2 s_{\max}^2/\sigma_O^2$ the condition is

$$1 - A^{-1} < 1/(1 + q), \quad \text{i.e.} \quad (1 + q)(A - 1) < A, \quad \text{i.e.} \quad q(A - 1) < 1,$$

and $A - 1 = 2\sqrt{\gamma} + \gamma$ gives (8).

**Averaging rule.** Its residual is the window-mean variance

$$v_{\text{avg}} = \frac{v_\infty}{S^2} \sum_{i,k} \rho^{|i-k|}$$

at stationarity (App. B.9). For $|\rho_\perp| < 1$, the whole stable range, every term is at most one in absolute value, so $\sum_{i,k} \rho^{|i-k|} \le S^2$, $v_{\text{avg}} \le v_\infty(\lambda_\perp)$ and $B_{\text{eff}} \ge 1$, on the oscillating branch $\rho_\perp < 0$ as well. After a finite burn-in the variance of a mean is at most the mean of the variances, $\frac{1}{S}\sum_k V_k = C_\perp^{\text{acc}} \le v_\infty$. The same ceiling therefore holds for both reduction rules, and since retention needs only $g_j(\lambda_\perp) > 0$, not (H1) (App. B.6), the necessary step bound carries to the averaging rule. □

The bound contains no $b$, $S$ or $B$. So they cannot create a prevention window when (8) fails. When it holds, they still move $C_\perp$ across the bar through $\Phi$. A dense search over settings tests the corollary (Table 7; Table 9, block C).

**The bound closed at the fixed point, in terms of $c_1$.** Here we restate the bound using only the data's top eigenvalue $c_1$. (8) is stated at the current $s_{\max}$, which learning moves. We close it at the destination of the fully mixed accumulating rule ($\Phi \equiv 1$, $\bar{\eta}$ fixed), where App. B.5 gives the top mode $\lambda_1^\star = c_1/(\sigma_O^2\sigma_H^2 + \frac{\bar{\eta}}{2}c_1)$. Then

$$\bar{\eta}\lambda_{\max} = \bar{\eta}c_1/(\sigma_O^2\sigma_H^2 + \tfrac{\bar{\eta}}{2}c_1) < 2$$

for every $\bar{\eta}$ (and $\lambda_{\max} = \sigma_H^{-2}$ if the mode is erased). So the stability half of the corollary never binds at the fixed point, because the equilibrium amplitude adjusts to the step.

What binds instead is the survival of the signal itself. By Thm. 3.3(ii) mode $j$ is interior iff $c_j > c_{\text{thr}} = \sigma_O^2 C_\perp/\sigma_H^2$, and the ceiling gives $c_{\text{thr}} \le \sigma_O^2/(1 - \bar{\eta}/(2\sigma_H^2))$, with equality where $\Phi(\lambda_\perp) = 1$; so a signal mode can be erased only at

$$\bar{\eta} \ge \bar{\eta}_j^\star := 2\sigma_H^2\Big(1 - \frac{\sigma_O^2}{c_j}\Big),$$

the step at which the ceiling reaches $c_j$, sharp in the fully mixed limit. Prevention needs $\bar{\eta} \ge \bar{\eta}_{\min} = 2\sigma_H^2(1 - A^{-1})$, $A = (1 + \sqrt{\gamma})^2$, so a preventing step that keeps the top mode exists iff $\bar{\eta}_{\min} < \bar{\eta}_1^\star$, i.e.

$$c_1 > \sigma_O^2(1 + \sqrt{\gamma})^2 = c_{\text{edge}}.$$

In words, the top signal eigenvalue must itself clear the fluctuation edge. Otherwise every step that prevents memorization also erases the top signal mode. Each further signal mode $j$ is then lost at its own $\bar{\eta}_j^\star$. This is the step against which the sharp drop on the prevention grid is read. At a finite window this is only a necessary condition, since the root equation carries $\Phi(\lambda^\star)$ (App. B.5).

### B.8 CAPACITY, DATASET SIZE, TRAINING TIME: THE OCCUPANCY MODEL

This subsection derives a simple model of how capacity, dataset size and training time together decide memorization. Step six (Eq. (9)) follows from Thm. 3.3 and the Marchenko–Pastur edge alone. The occupancy model adds the time course. With it come three assumptions, (T1)–(T3) of Prop. B.4. It is a reduced model, not one of the theorems of §3. Memorization needs two things. There must be free latent slots beyond the signal rank, and the fluctuation edge must lie above the ridge threshold. More parameters give more slots. More data pulls the edge below the threshold.

**Setting.** The data covariance is $\Sigma_{\text{pop}} = \sigma_O^2 I + \sum_{\ell=1}^{r} w_{\text{scale}}^2 \sigma_H^2 \, u_\ell u_\ell^\top$, with $r \le m \le n$. The decoder has $p = nm$ parameters. The default initialization violates (T2) of Prop. B.4 past $N \approx 801$ (App. E.9). No onset is seen there, and the occupancy model says nothing about that protocol. Take $W(0) = G/\sqrt{m}$ with $G$ standard Gaussian. Then the initial scale $s_0(m) = \|u^\top W(0)\|$ has $s_0^2$ of mean one, independent of $m$, while $s_{\max}^2(0) \approx n/m$ for $m \ll n$. Which of the two scales plays $s_0$ is a choice of observable.

**Linearized rate.** Near $W = 0$, $H = \sigma_H^{-2}(I + O(\|W\|^2))$, so

$$\begin{aligned} \widehat{\Sigma}_{ym} &= (\sigma_H^2/\sigma_O^2)\widehat{\Sigma}_y W + O(\|W\|^3), \\ W\widehat{\Sigma}_{mm} &= O(\|W\|^3) \qquad \text{and} \\ WC &= C_\perp W + O(\|W\|^3) \qquad (\phi \text{ smooth at } \lambda_\perp); \end{aligned}$$

with $\alpha C_\perp = \beta c_{\text{thr}}$,

$$\dot{W} = \beta\big(\widehat{\Sigma}_y - c_{\text{thr}} I\big)W + O(\|W\|^3), \qquad \beta := \alpha\frac{\sigma_H^2}{\sigma_O^2}, \qquad c_{\text{thr}} = \frac{\sigma_O^2}{\sigma_H^2} C_\perp. \tag{14}$$

Dropping the cubic term, $u_j^\top W(t) = e^{\nu(c_j)t} u_j^\top W(0)$ for every $W(0)$ with the rate (13), which has no $m$ (valid while $\sigma_H^2 s^2 \ll \sigma_O^2$). This holds per mode. For a signal mode $c_{\text{sig}} > c_{\text{thr}}$ that starts at $s_{0,\text{sig}}(m)$ and must reach $s_G$,

$$\dot{s} \simeq \beta\,(c - c_{\text{thr}})\,s, \tag{15}$$

$$\tau_{\text{gen}}(m) \approx \frac{\log\big[s_G/s_{0,\text{sig}}(m)\big]}{\beta\,(c_{\text{sig}} - c_{\text{thr}})}, \tag{16}$$

So width enters $\tau_{\text{gen}}$ only through $\log s_{0,\text{sig}}(m)$. The proposition below counts the occupied fluctuation modes.

**Proposition B.4** (Occupancy model; it adds (T1)–(T3) to §3 and is not one of its theorems)**.** Assume

- (T1) the decoder stays on the aligned invariant set, $r$ of its $m$ directions paired to the signal and the remaining $(m-r)_+$ available to fluctuation eigenvectors of $\widehat{\Sigma}_y$, strongest first;
- (T2) each fluctuation mode starts below its fixed point at the initial singular value $s_0(m)$ of $W$ along it;
- (T3) $r$ is fixed, so the Marchenko–Pastur law of Thm. 3.4 applies (App. B.6).

The fastest fluctuation mode sits at the bulk edge; its fluctuation-edge gap and the time $\tau_{\text{edge}}$ at which it reaches a level $s_M$, with $L_M(m) = \log[s_M/s_{0,\text{edge}}(m)]$, are

$$\Delta_N := \sigma_O^2\big(1 + \sqrt{n/N}\big)^2 - c_{\text{thr}}, \qquad \tau_{\text{edge}} = \frac{L_M(m)}{\beta\Delta_N} \;\; (\Delta_N > 0;\ \infty \text{ if } \Delta_N \le 0), \tag{17}$$

If $c_{\text{thr}} > \sigma_O^2$ ($C_\perp > \sigma_H^2$) the gap closes at a finite size $N_\infty^\star$, otherwise (exact EM included) $N_\infty^\star = \infty$; with $A^* = \sigma_O^2 x^*(1+x^*)/N_\infty^\star$, $x^* = \sqrt{n/N_\infty^\star}$,

$$N_\infty^\star = \frac{n}{\big(\sqrt{c_{\text{thr}}/\sigma_O^2} - 1\big)^2}, \qquad \tau_{\text{edge}} \sim \frac{L_M(m)}{\beta A^*(N_\infty^\star - N)} \quad (N \to N_\infty^{\star-}). \tag{18}$$

The bulk has the Marchenko–Pastur density $p_\gamma$, $c_- = \sigma_O^2(1-\sqrt{\gamma})^2$ (its atom at zero when $\gamma > 1$ carries no mass above a positive threshold); with its tail mass $T_\gamma$, the number $R_{\text{fl}}(m, N)$ of fluctuation

eigenmodes above threshold, capped by the latent slots $m - r$, and the retained band $[q_{m,N}, c_{\text{edge}}]$ are

$$p_\gamma(c) = \frac{\sqrt{(c_{\text{edge}} - c)(c - c_-)}}{2\pi\gamma\sigma_O^2 c}\,\mathbf{1}_{[c_-, c_{\text{edge}}]}(c), \qquad T_\gamma(x) = \int_{\max\{x, c_-\}}^{c_{\text{edge}}} p_\gamma,$$
$$R_{\text{fl}}(m, N) = \min\big\{(m - r)_+,\ nT_\gamma(c_{\text{thr}})\big\}, \qquad T_\gamma(q_{m,N}) = R_{\text{fl}}(m, N)/n. \tag{19}$$

A coarse aggregate model ($R_{\text{fl}}$ edge modes with a common $s_0(m)$ in place of the occupied tail; the band integral is (21)) then gives, for the fluctuation energy $Q_{\text{mem}}(t) := \sum_{j\in\mathcal{R}} s_j(t)^2$ over the retained set $\mathcal{R}$ and the time $\tau_Q$ at which it reaches a level $Q^*$,

$$Q_{\text{mem}}(t) \approx R_{\text{fl}}(m, N)\, s_0(m)^2\, e^{2\beta\Delta_N t}, \qquad \tau_Q \approx \frac{\log\big[Q^*/\big(R_{\text{fl}}(m, N)\, s_0(m)^2\big)\big]}{2\beta\Delta_N}. \tag{20}$$

**Derivation. Occupancy.** For $\gamma = n/N$ the fluctuation bulk has the density $p_\gamma$ of (19), $c_\pm = \sigma_O^2(1 \pm \sqrt{\gamma})^2$, $c_{\text{edge}} := c_+$ (an atom at zero when $\gamma > 1$ carries no mass above a positive threshold). The data supply $nT_\gamma(c_{\text{thr}})$ eligible modes and (T1) supplies $(m - r)_+$ slots, which gives $R_{\text{fl}}$. The rate increases in $c$, so the retained band is $[q_{m,N}, c_{\text{edge}}]$. Width sets how many modes are held, not the rate of any. The occupancy model applies linear growth and the rank cap together and does not derive when the alignment (T1) happens.

**Aggregate.** With the spectrum replaced by its limit,

$$Q_{\text{mem}}(t) \approx n \int_{q_{m,N}}^{c_{\text{edge}}} s_0(c, m)^2\, e^{2\beta(c - c_{\text{thr}})t}\, p_\gamma(c)\, dc, \tag{21}$$

Here $N$ acts through $p_\gamma$, and $m$ through $q_{m,N}$ and $s_0$. The fastest rate is $\beta\Delta_N$, which decreases in $N$. Replacing every retained mode by the edge rate and a common $s_0(m)$ gives (20). Its level crossing $\tau_Q$ plays the role of $\tau_{\text{mem}}$. This coarse form needs the band to be narrow on the onset's time scale, $2\beta\tau_Q(c_{\text{edge}} - q_{m,N}) \lesssim 1$. There, with width-independent $s_0$, capacity advances memorization only logarithmically (first line of (22)), and only until $m - r \geq nT_\gamma(c_{\text{thr}})$. If $s_0^2 \propto m^{-1}$ where $R_{\text{fl}} \propto m$, the width effect cancels. So no universal exponent in $m$ follows. In the edge-dominated regime (still $\sigma_H^2 s_j^2 \ll \sigma_O^2$),

$$p_\gamma \sim A_\gamma\sqrt{c_{\text{edge}} - c} \quad \text{with} \quad A_\gamma = \frac{\gamma^{-3/4}}{\pi\sigma_O c_{\text{edge}}} \qquad (\text{as } c_{\text{edge}} - c_- = 4\sigma_O^2\sqrt{\gamma});$$

for $s_0(\cdot, m)$ continuous and non-zero at $c_{\text{edge}}$, Watson's lemma with

$$\int_0^\infty x^{1/2} e^{-ax}\, dx = \Gamma(3/2)\, a^{-3/2}$$

gives the second line of

$$2\beta\tau_Q(c_{\text{edge}} - q_{m,N}) \lesssim 1: \quad \tau_Q \approx \text{const} - \frac{\log(m - r)}{2\beta\Delta_N},$$
$$\beta t(c_{\text{edge}} - q_{m,N}) \gg 1: \quad Q_{\text{mem}}(t) \sim n\, s_0(c_{\text{edge}}, m)^2 A_\gamma\, \Gamma(\tfrac{3}{2})\, (2\beta t)^{-3/2}\, e^{2\beta\Delta_N t}, \tag{22}$$

with no $q_{m,N}$ and no $\log(m - r)$ dependence.

**Dataset boundary.** With $\kappa := C_\perp/\sigma_H^2 = c_{\text{thr}}/\sigma_O^2$ the gap closes where $\sigma_O^2(1 + \sqrt{n/N})^2 = c_{\text{thr}}$, at the $N_\infty^\star$ of (18) if $\kappa > 1$ and never if $\kappa \leq 1$ (exact EM, $\kappa = 1$). With $x(N) = \sqrt{n/N}$,

$$\frac{dc_{\text{edge}}}{dN} = -\sigma_O^2\,\frac{x(1 + x)}{N} < 0, \qquad \text{so}$$
$$\Delta_N = A^*(N_\infty^\star - N)\big[1 + O\big((N_\infty^\star - N)/N_\infty^\star\big)\big] \qquad \text{with}$$
$$A^* = \frac{\sigma_O^2 x^*(1 + x^*)}{N_\infty^\star} = \frac{\sigma_O^2\sqrt{\kappa}(\sqrt{\kappa} - 1)^3}{n} > 0 \qquad \text{at } x^* = \sqrt{\kappa} - 1;$$

inserted in $\tau_{\text{edge}} = L_M/(\beta\Delta_N)$ this gives the simple pole of (18), residue $L_M(m)/(\beta A^*)$, width entering only through $L_M(m)$.

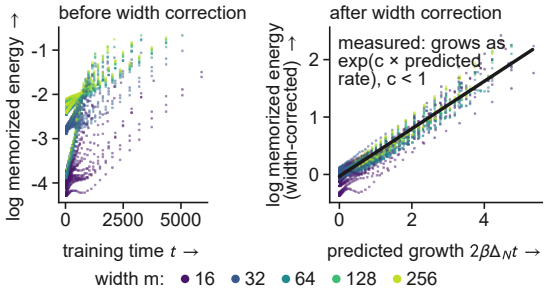



Figure 12: **Occupancy collapse test** (pre-registered). The left panel shows log memorized energy against training time, one line per width $m$. The right panel shows the same with the width correction $R_{\text{fl}} s_0^2$, against the predicted growth $2\beta\Delta_N t$. The line is $e^{c\cdot 2\beta\Delta_N t}$ with fitted $c = 0.41$.

**Finite training time.** At a budget $t$ the edge mode crosses its level iff $\Delta_N \geq L_M(m)/(\beta t)$, i.e. for

$$N \leq N_t^\star(m) = \frac{n}{\left[\sqrt{\kappa + L_M(m)/(\sigma_O^2 \beta t)} - 1\right]^2}$$

(at every $N$ when the root's argument is $\leq 1$). $N_t^\star(m)$ increases in $t$, decreases in $L_M(m)$ and tends to $N_\infty^\star$ from below, with $N_\infty^\star - N_t^\star(m) \simeq L_M(m)/(\beta A^* t)$. So the asymptotic boundary carries no width, and the finite-time one carries it only through $L_M(m)$.

**Phases, and what is untested.** Training has three phases. First $t < \tau_{\text{gen}}$, then a generalizing window $\tau_{\text{gen}} < t < \tau_Q$, then memorization $t > \tau_Q$. More data lowers $\Delta_N$. This delays memorization and widens the window ($\tau_Q \to \infty$ as $N \to N_\infty^{\star-}$). Width acts on occupancy and initial scale, not on the per-mode rate. So no asymptotic law $N^\star(m) \propto m^a$ should be assumed. The $(p, N)$ plane (Fig. 1b) splits into three regions. These are memorization before generalization, the window $\tau_{\text{gen}} < t < \tau_Q$, and ridge regularization above $N_\infty^\star$. Read at a finite training budget, the boundary between the first two is a budget contour, not a universal power law in $p$. The tests are in App. B.8.1.

### B.8.1 Tests of the occupancy model

This subsection tests the occupancy model against measured runs. **The occupancy law holds as a scaling, up to a constant rate factor, and the width collapses.**

**Collapse** (Fig. 12; pre-registered). We use the solvable tier at a fixed step, so $c_{\text{thr}}$ is constant, and a small initialization, so (T2) holds. Once $\log R_{\text{fl}} + 2\log s_0$ is subtracted, the residual $\log R_{\text{fl}}$ coefficient is -0.01 $[-0.03, 0.03]$ (held). The curves are linear in $2\beta\Delta_N t$ with pooled slope $c = 0.41$ $[0.40, 0.43]$. A constant factor $c$ on the growth rate enters the exponent ($e^{c\cdot 2\beta\Delta_N t}$ has slope $c$). So the law holds as the order statement $\log[Q_{\text{mem}}/(R_{\text{fl}} s_0^2)] = \Theta(2\beta\Delta_N t)$. The pre-registered unit slope itself was missed (missed, no setting within tolerance). A later, unregistered check explains the factor. The strongest retained direction grows at 0.93 of its own linear rate. And $c$ matches the band-averaged growth rate, which is below the edge rate, times the finite-$n$ edge gap, 0.86 of $\Delta_N$ (protocol in App. E.3).

**Data delays memorization.** On the strong-signal generator (premise in 224/231 settings; original grid 0/150), at small initialization

$$\tau_{\text{gen}} \propto N^{-0.06}, \qquad \tau_{\text{mem}} \propto N^{+0.74}, \qquad \text{gap} \propto N^{+0.80}\ [+0.76, +0.84].$$

All 42 fitted settings reach both onsets. The 22 $\tau_{\text{gen}}$-censored settings lie below $N = 256$. (A censored AFT refit gives $\tau_{\text{mem}} \propto N^{+0.74}$, unchanged, App. D.) This agrees in sign with $N \uparrow \Rightarrow \Delta_N \downarrow \Rightarrow \tau_Q \uparrow$ (Fig. 1a, Fig. 3b). At the default initialization only the sign survives, and Thm. 3.3's $N^0$ for $\tau_{\text{gen}}$ is scored a miss (Table 9, block B).

**Width moves the onset, not the rate.** We use a tied random-features instrument. Its readout is tied to its learnable feature map, it is coded independently of the theory, and it lies outside the base model. On it $\tau_{\text{gen}} \propto m^{-0.21}$, while the underlying rate goes as $m^{-0.010}$ (App. E.9). This is the split that (16) predicts.

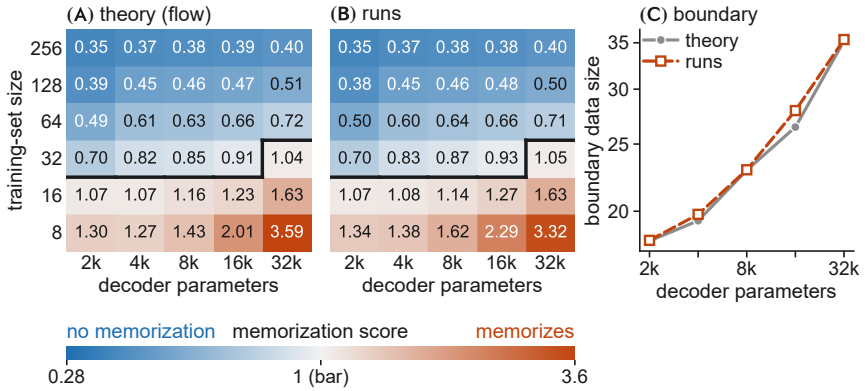



Figure 13: **The capacity plane** (solvable tier; 30 settings; axes as in Fig. 1b, with $p = nm$, $n = 128$, across and $N$ up). (A) Final fit fraction $\phi_{\text{fit}}$ of the integrated flow and (B) of the runs. Each box prints $\phi_{\text{fit}}$, and its colour shows the log (blue below the bar, red above). Black steps mark the bar $\phi_{\text{fit}} = 1$. (C) The boundary from each, theory (grey) and runs (red). The integration uses no measured input.

**A budget contour.** The integrated flow reproduces the measured $(N, m)$ plane (Fig. 13) to a median 0.6% (worst 12%). It reproduces $N^\star(m)$ at a median ratio 1.00. So it reproduces the form, but it does not test the model. The measured contour's exponent moves from +0.547 to +0.274 between two budgets.

**Untested.** Three predictions are not tested. They are the band integral (21), the $t^{-3/2}$ prefactor, and the $(N^\star_\infty - N)^{-1}$ divergence. The capacity plane of Fig. 13 integrates the full flow, not (20). So it motivates the occupancy law rather than tests it.

### B.8.2 The interpolation line $p = nN$ and the retention threshold

This subsection asks whether the classical interpolation line $p = nN$ matters here, and shows that the retention threshold matters instead. **Derivation.** For $N < n$ the sample second moment has rank $N$. The Marchenko–Pastur law carries an atom $1 - 1/\gamma$ at zero and mass $1/\gamma$ on $[c_-, c_+]$, so $nT_\gamma(0) = N$, and $N - r$ of the nonzero eigenvalues are noise. The retained fluctuation directions and the occupancy of (19), with this rank cap made explicit, are

$$D(N) := \#\{i > r : c_i > c_{\text{thr}}\} \simeq \min\{nT_\gamma(c_{\text{thr}}), N - r\}, \quad R_{\text{fl}} = \min\{(m - r)_+, D\}. \tag{23}$$

- $D = 0$ iff $c_+ \le c_{\text{thr}}$, i.e. $N \ge N^\star_\infty$ of (18);
- $D = N - r$ (every nonzero noise eigenvalue retained, $r \ll N$) iff $c_- \ge c_{\text{thr}}$, i.e. $N \le N_{\text{all}} := n/(\sqrt{\kappa} + 1)^2$.

Width stops mattering at the crossover $m_c(N) := r + D(N) \le N$, with equality iff $N \le N_{\text{all}}$. Since $\tau_Q$ in (20) depends on $m$ only through $R_{\text{fl}}$, $\partial_m \tau_Q = 0$ for $m \ge m_c(N)$. So the whole half-plane $p \ge nN$ is width-independent. The $\tau_Q = \tau_{\text{gen}}$ boundary flattens at $m_c \le N$, and it lies on the line $p = nN$ only for $N \le N_{\text{all}}$. The point $(N^\star_\infty, N^\star_\infty)$ is not critical. There $D = 0$, the line $m = N$ crosses $N^\star_\infty$ inside the data-limited phase, and the width-limited wedge $r < m < m_c(N)$ closes at the triple point $(N, m) = (N^\star_\infty, r)$ (Fig. 14a). So there are three regimes for the occupancy $R_{\text{fl}}$.

(i) $r < m < m_c$ (under-parameterized), $R_{\text{fl}} = m - r$;

(ii) $m \ge m_c$ (all $m \ge N$ included), $R_{\text{fl}} = D(N)$ independent of $m$, and for $N \le N_{\text{all}}$, $R_{\text{fl}} = N - r$, so the learned covariance reproduces the sample's fluctuation on its whole range (interpolation);

(iii) $N \ge N^\star_\infty$, $R_{\text{fl}} = 0$ at every $m$, however over-parameterized.

The effective interpolation threshold is $m_c$, a hard-threshold effective dimension like ridge's $\sum_i c_i/(c_i + \lambda)$, not $p = nN$. At a fixed step $\bar\eta$ ($\sigma_H = \sigma_O = 1$) the fixed point $\lambda^2 v_\infty(\lambda) = c$ gives

$$s^{\star 2}(c) = \frac{(2 - \bar\eta)c - 2}{2 + \bar\eta c},$$

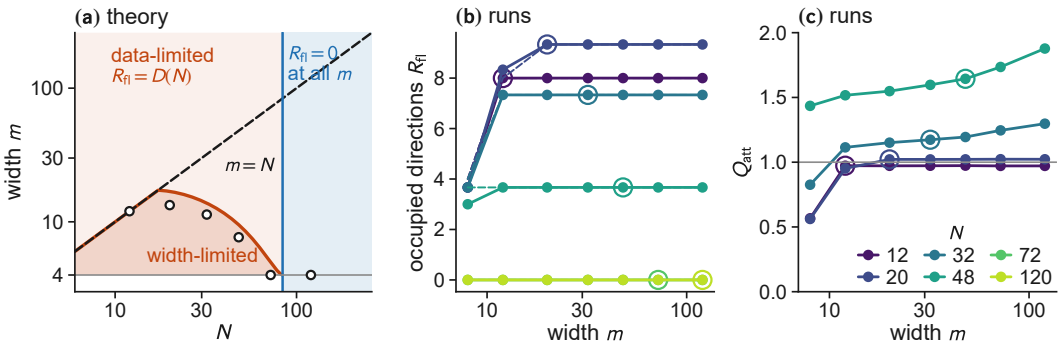



Figure 14: **The interpolation line and the retention threshold** ($n = 128$, $r = 4$, $\kappa = 5$). (a) Occupancy phases from (23). The width-limited wedge is dark, the data-limited phase is light, and the blue region past $N^\star_\infty$ retains nothing. The dashed line is $m = N$, and circles mark the runs' crossover $r + D$. (b) Occupied fluctuation directions against width, as seed means (solid) and $\min(m - r, D)$ (dashed). Rings mark $m = N$. (c) Fraction of the attainable fluctuation energy held.

which vanishes at $c = \kappa$ and is capped by $2/\bar{\eta} - 1$. Every fixed point is stable, but the margin

$$2/\bar{\eta} - \lambda^\star = \frac{4}{\bar{\eta}(2 + \bar{\eta}c)}$$

is small for the large sample eigenvalues of small $N$.

**Test** (pre-registered). We use the solvable tier with the accumulating rule and a cold start, with $n = 128$, $r = 4$, spike $\theta = 8$ and fixed $\bar{\eta} = 1.6$. This gives $\kappa = 5$, $N^\star_\infty = 83.8$ and $N_{\text{all}} = 12.2$. The grid is $N \in \{12, 20, 32, 48, 72, 120\} \times m \in \{8, 12, 20, 32, 48, 72, 120\} \times 3$ seeds (126 settings $(N, m, \text{seed})$). $W(0)$ is the default draw scaled by 0.01. The rows $N \le 20$ run $\eta_w = 0.005$ for 12000 updates with 16 chains per datum. The other rows run 0.02 for 3000 updates with one chain, which gives the same horizon $\eta_w t$. In the sample eigenbasis $u_i$, let $e_i = u_i^\top W W^\top u_i$. Then $R_{\text{fl}}$ counts noise directions with $c_i > c_{\text{thr}}$ and $e_i > s^{\star 2}(c_i)/2$. The subspace metric

$$Q_{\text{frac}} = \frac{\sum_{i \in \mathcal{B}} e_i}{\sum_{i \in \mathcal{B}} (c_i - \sigma_O^2)_+}$$

over the $N - r$ nonzero noise directions $\mathcal{B}$ is the fraction of the sample's excess fluctuation energy that $W$ holds. $Q_{\text{att}}$ replaces the denominator by the attainable $\sum_{\mathcal{B}} s^{\star 2}(c_i)$.

**Results** (Fig. 14b,c). $R_{\text{fl}} = \min(m - r, D)$ to within 0.7 of a direction in every setting (P1 HELD). $m_c \le N$ in every row and seed (P2a HELD). Past $m_c$, width never changes $R_{\text{fl}}$ by more than one. But the retained energy spreads by up to 16% against the registered 15% (P2b MISSED). In the one-chain rows it creeps up with $m$. At $N = 48$, $m = 8$ the slowest near-threshold mode has not reached $s^{\star 2}/2$ by the end. In the stabilized rows the spread is 0.2% (post hoc). At $N = 12$ ($D = N - r$ in every seed) all $N - r$ nonzero noise directions are occupied at every $m \ge N$, with $Q_{\text{att}} \ge 0.97$. P3 is MISSED on its last clause, $|Q_{\text{frac}}/Q_{\text{pred}} - 1| \le 20\%$ wherever $D > 0$, $m \ge m_c$. That deviation reaches 83% in the one-chain rows $N > 20$. It is a sub-threshold energy floor that grows with $m$. It is absent in the stabilized rows, where the deviation is at most 3% (post hoc). This fits the Monte Carlo jitter of the M-step, which the stabilized protocol reduces. $Q_{\text{frac}}$ itself stays at or below 0.011 even at full interpolation, because the cap $2/\bar{\eta} - 1$ keeps it far from its exact-EM value 1. For $N \ge 72$ ($D = 0$), $R_{\text{fl}} = 0$ at every $m$ including $m \ge N$, with $Q_{\text{frac}} \le 9 \times 10^{-4}$ (P4 HELD). None of the 126 settings failed the stability check (P5 HELD). This includes the $N = 12$ row, every setting of which diverged when run with $\eta_w = 0.02$ and one chain (the other rows' settings). Here $\bar{\eta}\lambda_{\max} \le 1.92 < 2$ at every checkpoint. $\phi_{\text{fit}}$ registers none of this. At $N = 12$ it starts at 0.34 and moves by at most $3 \times 10^{-4}$ at any $m$.

### B.9 Averaging rule and effective samples

This subsection works out the averaging rule, which the main text excludes, and shows why the theory rules it out. The averaging rule hands over the window mean $\bar{a}$ of $S$ states of a stationary

eigen-direction $a_{k+1} = \rho a_k + \sqrt{2\bar{\eta}}\,\xi_k$. Its autocovariance is $v_\infty \rho^{|h|}$. With

$$\sum_{h=1}^{S-1}(S-h)\rho^h = \frac{S\rho(1-\rho) - \rho(1-\rho^S)}{(1-\rho)^2},$$

$$B_{\text{eff}} := \frac{v_\infty}{\text{Var}(\bar{a})} = \left[\frac{1}{S}\left(\frac{1+\rho}{1-\rho} - \frac{2\rho(1-\rho^S)}{S(1-\rho)^2}\right)\right]^{-1} \qquad \text{(Geyer, 1992; Sokal, 1997),}$$

$B_{\text{eff}} \to S\bar{\eta}\lambda/(2-\bar{\eta}\lambda)$ for $S$ large against the relaxation time, and $B_{\text{eff}} \geq 1$ on the whole stable range $|\rho| < 1$, since $\sum_{i,k} \rho^{|i-k|} \leq S^2$ (App. B.7). The accumulating rule sits at $B_{\text{eff}} = 1$ wherever $\Phi = 1$. The mean of $B$ independent mixed chains sits at $B_{\text{eff}} = B$ exactly. $B_{\text{eff}}$ is a ratio of biases, not a variance-based effective sample size. Since it grows with $\lambda$, $\phi_{\text{avg}} = v_\infty/B_{\text{eff}} \sim 2/(S\bar{\eta}\lambda^2)$. So the exponent moves from $\lambda^{-1}$ to $\lambda^{-2}$ and cancels the drive's own $\lambda^{-2}$ in (6). The exact statement follows.

**Lemma B.5** (Averaging rule)**.** Let $v_{\text{avg}} = \frac{v_\infty}{S}\left[1 + 2\sum_{k=1}^{S-1}(1 - \frac{k}{S})\rho^k\right]$. Then

$$S\,\bar{\eta}\,\lambda^2 v_{\text{avg}} = 2 - \frac{4\rho}{S}\,\frac{1-\rho^S}{1-\rho^2}. \tag{24}$$

Consequently,

- on $u \in (0,1]$, $\lambda^2 v_{\text{avg}} \leq 2/(S\bar{\eta})$ with equality iff $u = 1$;
- on $u \in (1,2)$, $\lambda^2 v_{\text{avg}} > 2/(S\bar{\eta})$; and
- as $u \to 2^-$, $\lambda^2 v_{\text{avg}} \to 4/(S\bar{\eta})$ for even $S$ while $\lambda^2 v_{\text{avg}} \to +\infty$ for odd $S$.

*Proof.* By the sum above,

$$S + 2\sum_{k=1}^{S-1}(S-k)\rho^k = \frac{S(1+\rho)}{1-\rho} - \frac{2\rho(1-\rho^S)}{(1-\rho)^2};$$

multiplying by $\bar{\eta}\lambda^2 v_\infty/S = \frac{2(1-\rho)}{S(1+\rho)}$ gives (24). For $\rho \in [0,1)$ the subtracted term is non-negative and vanishes iff $\rho = 0$ ($u = 1$); for $\rho \in (-1,0)$ it is negative, since $-\rho > 0$ and $1 - \rho^S > 0$. As $\rho \to -1^+$ with $S$ even, L'Hôpital gives

$$\frac{1-\rho^S}{1-\rho^2} \to \frac{S}{2}\rho^{S-2}\Big|_{\rho=-1} = \frac{S}{2}$$

and the right side tends to 4; with $S$ odd, $1 - \rho^S \to 2$, $1 - \rho^2 \to 0^+$ and $-4\rho/S \to 4/S > 0$, so it tends to $+\infty$. □

So $2/(S\bar{\eta})$ is an upper bound on $\lambda^2\phi_{\text{avg}}$ for $\bar{\eta}\lambda \leq 1$, attained only at $\bar{\eta}\lambda = 1$, and (H1) holds below it. Boundedness negates the reachable (H2). So Thm. 3.3 excludes the rule at a frozen step exactly when $2/(S\bar{\eta}) < c_j/(\sigma_O^2\sigma_H^2)$ (App. B.5). This differs from the accumulating rule's $2/\bar{\eta}$ by a factor $S$, not in kind. Measured, $\text{Var}(\bar{z}_S)/v_\infty$ spans a factor 3.16 across the directions of $H$ and falls as $1/S$. So "averaging first is MAP" is the infinite-window limit of a continuum.

# C THE SETUP IN FULL

This section writes out the model, the sampler, the two learning rules and the two scores that the rest of the paper uses.

## C.1 THE CODER'S ENERGY, AND THE CLOSED FORMS INFERENCE DESCENDS ON

Here we give the energy that inference minimises and the closed forms that follow from it. The network holds the two-layer linear–Gaussian coder of §2.1 as an energy. Each of its two terms is held by one population of error units, one population per level.

$$\mathcal{F}(z,y) = \tfrac{1}{2}\|z-\mu\|^2_{\Sigma_H^{-1}} + \tfrac{1}{2}\|y - Wz\|^2_{\Sigma_O^{-1}}, \qquad \nabla_z\mathcal{F} = H\big(z - m_y\big), \qquad \lambda_\perp = \frac{1}{\sigma_H^2}, \tag{25}$$

so inference is descent in $z$ on a quadratic. Its gradient is $H(z - m_y)$, with $H$ and $m_y$ as in §2.1. The flat floor $\lambda_\perp$ is the prior precision alone. It is the eigenvalue that $H$ takes in every direction the emission matrix says nothing about. Whenever the rest of the paper talks about slow directions, it means this floor. The code learns $\mu$ by the second rule of (27). The analysis is written at $\mu = 0$. It applies to $z - \mu$ and $y - W\mu$, since $m_y = \mu + H^{-1}\sigma_O^{-2}W^\top(y - W\mu)$.

### C.2 Sampling the inference phase, and the two ways to reduce it

Here we write down the noisy inference chain and the two ways of turning its samples into a learning signal. In the language of expectation–maximization, the inference phase is the E-step and the learning phase is the M-step. We use both pairs of names. Monte Carlo predictive coding replaces the descent by a noisy descent. It is discretised by the Euler–Maruyama scheme, which takes a gradient step and adds Gaussian noise at every iteration. The effective step size $\bar\eta$ folds in any preconditioner the implementation applies. Because the energy is quadratic, the iteration, its contraction factor in each direction and its stationary variance are again explicit.

$$\begin{aligned} z_{k+1} &= z_k - \bar\eta H(z_k - m_y) + \sqrt{2\bar\eta}\,\xi_k, \qquad \rho_i = 1 - \bar\eta\lambda_i, \\ (\Sigma_\infty)_{ii} &= \frac{2\bar\eta}{1-\rho_i^2} = \frac{1}{\lambda_i}\cdot\frac{1}{1-\bar\eta\lambda_i/2}, \end{aligned} \tag{26}$$

where $\xi_k \sim \mathcal{N}(0, I)$ is drawn afresh at every step. It is not the observation noise $\xi$ of the model. The last expression already shows the shape of everything that follows. The stationary variance of the discrete chain is the posterior variance in that direction, times a factor. That factor depends only on the step size and on that direction's own precision. Alg. 1 gives the chain's schedule (start at $m_y$, burn-in, window).

The learning phase is the same as in the deterministic algorithm. It is the Hebbian pair on the left of (27), applied to whatever activity the inference phase leaves in the network. For the emission weights it uses the statistic on the right (Oliviers et al., 2024, Eq. 6).

$$\Delta W = \eta_w \Sigma_O^{-1}\varepsilon^y\hat z^\top, \qquad \Delta\mu = \eta_w\Sigma_H^{-1}(\hat z - \mu), \qquad \widehat\Sigma_{zz} = \frac{1}{NS}\sum_{a,k} z_{a,k}z_{a,k}^\top. \tag{27}$$

Here $\varepsilon^y = y - W\hat z$ is the prediction error at the observation level. $\hat z$ is the sampled activity at which the update is evaluated, and $\eta_w$ is the learning rate of the M-step. $S$ is the number of states read per datum after the burn-in. The sum runs over the $N$ training examples $a$ and the $S$ read states $k$. Under the averaging rule the last expression is replaced by

$$\frac{1}{N}\sum_a \bar z_a\bar z_a^\top \quad \text{with} \quad \bar z_a = \frac{1}{S}\sum_k z_{a,k},$$

which is the outer product of the averaged state. Under the accumulating rule $\hat z$ runs over each read state in turn. With $B$ chains the accumulation also runs over the chains.

**The rightmost expression is the whole story of this paper in one place.** It is an average of outer products of sampled states. It is not the outer product of their average. These two reductions of the same chain are different estimators. They differ by exactly the leftover covariance of the chain, and that difference is what the rest of the paper is about. We call the first the accumulating rule, which is the one the reference implementation runs. We call the second the averaging rule. In the code both rules restart at $m_y$ with the full burn-in $b$ at every M-step. So no chain state carries over between M-steps, and comparing the rules compares only the reductions.

As a reference for both rules we hold fixed the two statistics that a network solving its inference phase exactly would supply.

$$\widehat\Sigma_{mm} = \frac{1}{N}\sum_a m_{y_a}m_{y_a}^\top, \qquad \widehat\Sigma_{ym} = \frac{1}{N}\sum_a y_a m_{y_a}^\top, \tag{28}$$

where the first is the second moment that a maximum-a-posteriori inference phase would deliver. The second is the cross moment between observations and inferred means. The cross moment is linear in the latent activity, so it is the same under every reduction rule we consider. This is why the whole difference between the rules sits in the second moment.

### C.3 What generalization and memorization mean for this model class

**This subsection fixes the two numbers from which every onset in the paper is read.** Each is an order parameter in the statistical-physics sense, which means one number per checkpoint whose crossing of a bar marks a phase. One measures generalization and one measures memorization. We fix them once, before either is used, because for this model class the choice is forced rather than a convention. A static linear–Gaussian predictive coder is exactly its marginal covariance. So everything the trained network can ever generate is a function of that one matrix. Generalization is therefore measured as the exact held-out negative log-likelihood under the model's own marginal. Memorization is measured as where that marginal sits between the distribution the data came from and the training sample itself.

$$\Sigma_{\text{model}} = \sigma_H^2 W W^\top + \sigma_O^2 I, \qquad \phi_{\text{fit}} = \frac{\left\|\Sigma_{\text{model}} - \Sigma_{\text{pop}}\right\|_F}{\left\|\Sigma_{\text{model}} - \Sigma_{\text{emp}}\right\|_F}, \tag{29}$$

where the model's marginal is compared against two matrices, the population covariance and the empirical covariance of the training set. The memorization order parameter exceeds one exactly when the model sits closer to its own training set than to the distribution that set was drawn from.

Two properties of this score matter later. We state them here so that reading a later figure never depends on noticing them. First, it is not monotone in time. The model overshoots its empirical target, so the score takes the same value twice. Second, it describes a covariance, not an individual example. So the replication-based order parameter of the diffusion literature cannot be used for this model class at all, and this is not just a matter of it being weak here. The tanh coder lies outside the theory, and there we do use replication (§4.3).

### C.4 The three assumptions, priced

This subsection lists what it costs to break each of the three assumptions. Table 5 gives the measured cost of breaking each hypothesis of §3.1. Untying the maps changes a conclusion, not just a constant. The separation in (iii) is a property of the generator. At our original constant it is out of reach before any update is taken, so it is untested there rather than failed (App. E.9).

Table 5: **The three assumptions and the measured cost of breaking each.** The sources are M11 (tied and untied), M14 (edge test) and App. E.9.

| assumption | diagnostic | held | broken |
|---|---|---|---|
| (i) tied maps | escape rate vs. capacity | $m^{-0.010}$ | $m^{+0.325}$ $[+0.26, +0.39]$ |
| (i) tied maps | both onsets, and their gap | reference | $-0.649$, $-0.728$ dex (gap $-0.057$) |
| (i) tied maps | ridge on the mode it acts on | falls $18.4\times$ | frozen at initialization ($1.000\times$) |
| (ii) cold start | extra term in (4) | absent | datum-dependent, sign-indefinite |
| (iii) separation | settings passing the edge test (M14) | 70 of 90 | 0 of 90 as originally set |
| (iii) separation | directions above any bulk, images | $\geq$ 119 of 128 on three of the four datasets | |
| (iii) separation | an admissible bulk split at all, MNIST | none in 4 of 9 runs | |

## D Experimental setup

This section describes how the experiments are run and measured, so that any number in the paper can be traced to a protocol.

**Words used below.** A setting is one choice of dataset size $N$, capacity $m$ and seed. Three variants run on the same settings. The first is MCPC, the sampled inference phase. Its ridge is Thm. 3.2's

Table 6: **The two measurement tiers.** The solvable tier has a known generator, population solution and interpolant, and is the primary test of Thms. 3.2 and 3.3. The image tier uses the same generalization bar per dimension, 900 updates over 36 checkpoints and the memorization bar $\phi_{\text{fit}} \geq 1.05$. Its population covariance and optimum come from a disjoint pool. Both onset bars are absolute and fixed across the grid. The E-step budgets differ by design (Lemma B.1). Experiment codes are M9 for the solvable grid, M12 for the image grid, M11 for the independent instrument (App. E.9), M15 for the prevention grid and M16 for the observable comparison (App. E.9). All are listed in Table 7.

| | solvable tier (M9) | natural-image tier (M12) |
|---|---|---|
| generator | linear–Gaussian, rank 8, $m^\star = 16$ | MNIST |
| observation dimension $n$ | 128 | 784 |
| dataset size $N$ | 8−256 | 64 to 2048 |
| capacity $m$ | 16−256 | 128 |
| seeds | 3 | 3 |
| settings, runs | 90, 180 | 18 |
| updates, checkpoints | 10000, 40 log-spaced | 900, 36 |
| M-step rate $\eta_w$ | 0.02 | identical |
| prior and noise scales $\sigma_H, \sigma_O$ | 1, 1 | 1, 0.3 |
| E-step step size $\bar{\eta}$ | 0.5 of the stability edge $2/\lambda_{\max}(H)$, reset every update, so the realized ridge is a measurement | identical |
| E-step reduction | one chain ($B = 1$) of $K = 64$ steps per datum; the M-step takes its final state's outer product | window of $S = 64$ states after $b = 800$ burn-in steps, outer products accumulated |
| held-out set | 4096 draws | 2048 images |
| generalization onset $\tau_{\text{gen}}$ | within 7 nats of the population optimum | 43 nats, the same bar per dimension, from the rank-$m$ fit to a disjoint pool of 32768 images that stands in for the population |
| memorization onset $\tau_{\text{mem}}$ | first checkpoint with $\phi_{\text{fit}} \geq 1$ | first with $\phi_{\text{fit}} \geq 1.05$ |

product at the protocol's chain length, written $C_K$. The second is MAP, the zero-ridge point $C = 0$. The third uses exact posterior moments, $C = H^{-1}$. This variant is run directly in App. E.3 and App. E.4, and elsewhere only in the integrator of §E.5. The zero-mean part $R_N$ of Prop. 3.1 is called the draw. The integrator can keep it or drop it (§4.1.1). The horizon of a run is its update budget. An onset that is not reached within the horizon is censored, which means it is counted but not fitted. Every run records two quantities at every checkpoint, the held-out excess likelihood against the oracle and $\phi_{\text{fit}}$. The onsets are read from them against bars that are fixed across the grid. "Original", said of a setting or a grid, means the generator scale and held-out convention of the reference implementation. We keep those grids as controls. On the solvable tier one chain of $K$ steps per datum (Table 6; ridge $C_K$) hands its final state to the M-step. This is a window of one, where the two rules coincide.

**The protocol.** This paragraph gives the training and inference protocol on each tier. **The primary tier is a linear–Gaussian generator.** It is swept over dataset size crossed with capacity. Checkpoints are log-spaced in updates, not in epochs. The reason is that one epoch is one pass over the training set, and the size of that set is itself swept. The inference phase runs one Langevin chain of $K = 64$ steps from $m_y$. The chain is restarted every M-step (the code's ensemble estimator at $B = 1$), and its final state is handed to the M-step. Thm. 3.2 gives this chain the iterate form of the ridge. At $S = 1$, $b = K - 1$ that form is identical to the window form, and no mixing is needed (App. B.3). Reading a single state is a window of one. There the accumulating and averaging rules coincide (§2.2), so this grid cannot tell them apart. That difference is tested on the independent instrument, on the images (App. E.9) and in the criterion-flip test ($b = 10$, $S = 40$). On both tiers $\sigma_H^2, \sigma_O^2$ are fixed at their generator values and never learned. Every constant that contains them is therefore conditional on that oracle. The ablation variant replaces the inference phase with a maximum-a-posteriori one, which is the zero-ridge point. Started at $m_y$, where the gradient vanishes, its descent does nothing.

The causal test (App. E.4) compares likelihoods at $N = 256$, below $N^{\star}_{\infty}$, and at $N = 4096$, above it.

**Both onsets use bars fixed across the whole grid.** This matters. A bar normalised per setting measures the shape of a trajectory, not the time of an onset. The quantity such a bar would divide by spans a factor of 3.5 across this grid. A per-setting bar would therefore have turned a difference in initialization into a false dependence on dataset size.

**Uncertainty, stated once for every figure and table in this paper.** Values are seed means. Bands and stated spreads are the spread over seeds. Bracketed intervals are 95% bootstrap intervals over settings, on an exponent fitted per setting. A setting whose onset is never reached is censored, not dropped. It is counted and reported, but excluded from exponent fits. Censored settings are the slow ones, so excluding them biases a rising exponent downward in expectation. This is a bias, not a bound. It does not arise where no fitted setting is censored, as in the onset-gap fits of §4.2. The censored refits quoted in the main text keep those settings. They fit a log-normal accelerated-failure-time model on the same settings and axes. Each censored setting is right-, left- or interval-censored by the crossing rule of its source, and intervals are percentile-bootstrap intervals over settings. When no setting in range is censored the fit equals least squares. On the small-initialization grid over its merged $N$ range (the observable comparison and its large-$N$ extension) the escape-rate exponent is +0.68 with censored settings dropped and +0.82 refitted. On the onset-gap settings, none of which is censored, it is +0.69 (Table 9, block B).

**Held-out sets.** This paragraph explains which held-out sets the likelihoods are read on. A held-out set built by calling the generator a second time re-draws its parameters. Every likelihood read off such a set therefore measures transfer to a new generator. The axes grid's held-out set is of this kind, so its oracle likelihood is inflated by a near-constant 1.9 nats. Every onset bar is thus 1.9 nats looser than stated. Exponents survive this uniform shift, but "nats saved" do not. The strong-signal grid shares the training draw by design. The train–test gap beside the prevention counterexample is read on a held-out set from the training draw. Its median is 0.666, against 0.959 on a transfer set (inflated by 44%). The prevention grid and every gap quoted here use the former. A held-out set must come from the same generator draw as the training set.

### D.1 Experiment index and reproduction

This subsection says how to find each experiment and how to reproduce its numbers. Pre-registered experiments are labelled where they are reported. Post-hoc analyses are kept apart from registered ones and labelled. Every number is a macro emitted from stored outputs, and there are two ways to regenerate them. (a) The paper takes minutes on a CPU. Run the macro emitters, then `python -m paper_mcpc.paper_addenda.build` and `...verify`. (b) The experiments take hours on one GPU. Run each code's `run.py` (Table 7).

## E The results in full

This section gives every result in full. It first lists the predictions that did not survive and then scores every prediction against its measurement (App. E.2). The pre-registered direct tests (App. E.3), the readings the main text quotes (App. E.9) and the nonlinear coder (App. E.10) follow. Throughout, a cell is one setting of a run grid, such as one $(N, m, \text{seed})$, and an arm is one variant of the training rule. An onset fires in a cell when the run reaches its bar within the training budget.

### E.1 What did not survive, and what stands in its place

This subsection records the predictions that did not survive. **Every prediction that was retracted, withdrawn or replaced is listed once, in Table 8, with what stands in its place.** The misses that Table 9 scores are not repeated here. The misses listed here are the ones Table 9 does not score. Of the table's 12 entries, 2 were errors in our own derivations rather than in the theory, 1 were retracted outright, 5 were withdrawn because they should not have been printed, 3 were genuine misses, and 1 were untested rather than failed. The verdict words mean the same thing throughout. Retracted means a clause is withdrawn from the theory's statement. Withdrawn means a printed prediction is removed and its measurement is kept. Our error means the failure was in our own derivation or

Table 7: **Experiment index.** Every experiment the paper reads, listed by code. Each is one `run.py` in the code release.

| code | experiment | grid and protocol | read in |
|---|---|---|---|
| M1 | ridge law | realized E-step residual in $H$'s eigenbasis against (5), windows to $S = 2048$ | Fig. 4; Tab. 9 C |
| M2 | effective samples | window $S$ against chain count $B$; effective capacity read on $\widehat{\Sigma}_{mm}$ | Tab. 9 C |
| M8 | separation check | edge test of the two-bulk premise on the original grid | App. E.9 |
| M9 | solvable grid | Table 6 | §4 |
| M10 | disjoint-data models | two models on disjoint halves of one draw, ratio $R_{\text{indep}}$; pre-registered | Tab. 9 C |
| M11 | independent instrument | tied random-features network with a learnable feature map, own code path; registered roster | App. E.9 |
| M12 | image grid | Table 6 | §4 |
| M14 | strong-signal grid | M9's protocol at generator scale 6, 360 settings; pre-registered | App. E.9 |
| M15 | prevention grid | 96 settings, 3 seeds, 2500 updates; chain length, chain count and step size each swept, the other two at defaults | §4.1.3 |
| M16 | observable comparison | M14's protocol, $N$ to 2048, an initialization-scale variant and a ridge sweep | App. E.9, §4.2 |
| M17 | crossing runs | M9's protocol at $N = 512$ to 16384 | §E.5 |
| M18 | images to $N = 16384$ | M12's setting, population pool moved clear of the largest training set | §E.5 |
| M19 | observable, large $N$ | M16's grids at $N = 4096, 8192, 16384$. With M16, the onset-gap grid: $w = 6$, $n = 128$, $\sigma_H = \sigma_O = 1$, 3 seeds, $N = 16$ to 16384, $m \in \{32, 128\}$, 6000 updates | §4.2 |
| M22 | nonlinear coder | MNIST, Fashion-MNIST, RGB CIFAR-10 and CelebA (`celeba_full/`); sampled, MAP and compute-matched MAP variants | §4.3, App. E.10 |
| M23 | occupancy collapse | pre-registered; protocol in App. E.3 | App. B.8.1 |
| M24 | criterion flip, exact-moment variant | pre-registered; protocol in App. E.3 | §E.5 |
| M28 | causal ridge intervention | ridge in the M-step statistic against sampling noise, with dose, compute-matched, exact-sampling and explicit-regularization variants; pre-registered | App. E.4 |
| M31 | likelihood gap $\Delta$ | M9's and M24's settings re-run with a held-out/train NLL probe; post hoc | App. E.8 |
| M30 | where the inflation acts | $C_\perp$ along Table 1's trajectories; flat/stiff channel split of its no-draw flow; post hoc | App. E.7 |

design, and the theorem it concerned stands. Missed means the data refuse the prediction. Untested means its premise was never met. Held in sign means the prediction was a direction and the measured interval has that sign.

### E.2 THE SCOREBOARD

This subsection scores every prediction against its measurement. Table 9 collects every prediction this paper makes or made. The one exception is the 23-check roster that the independent instrument registered for itself, which App. E.9 scores. The table has three blocks, because the three are different kinds of claim. Block A integrates the statements of §3 at each tier's own constants and compares them cell by cell. Block B fits a single power law over a finite range, which is how the classical picture and our own power-law predictions are stated. Block C holds searches, instruments and controls, where no exponent is at stake. Each row gives its scope or its budget and a verdict in one of five words, namely held, held in sign, missed, withdrawn or untested. The intervals set the standard. A row is held only if its 95% interval contains the predicted value. A row whose interval excludes that value is missed, however close it comes, and the scope column says how close. A fit that leaves out censored (slow) cells is biased downward, so it cannot be scored missed from below. For this reason the $N^{+1}$ row below is scored on the onset-gap fit, in which every fitted cell fires both onsets. Rows without an interval (a count, a bar test) are held when the measurement matches the prediction's statement. A per-cell prediction is held when its median ratio to the measurement lies within 20% and it fires in the same cells in at least nine of ten. Rows marked simulated use the theory's own draw taken at every update (§E.5). They are a Monte Carlo evaluation of the theory rather than a closed form. The one row that is a search rather than a prediction carries no verdict.

**Read block A as four integrations against two experiments, not as thirteen independent scores.** It is one integrator in four configurations (expected flow, draw kept, $\phi = 0$, and the capacity plane read off the second), run against the solvable grid and the image tier. The two capacity rows summarise the same 90 trajectories as the draw-kept rows. A withdrawn row keeps its measurement and drops a prediction we should not have printed. A missed row is one the data refuse. By this rule, of 34 predictions 17 hold, 2 hold in sign only, 12 miss, 2 are withdrawn and 1 untested. The statements

Table 8: **Predictions that failed, were retracted or were withdrawn.** One row per prediction. The last column gives the verdict in the scoreboard's terms (Table 9), plus retracted and our error.

| prediction | original claim | what stands now | verdict |
|---|---|---|---|
| two-bulk premise | the separation the theory presupposes holds on our grid | unreachable at our original generator constant; holds in 70 of 90 cells at the strong-signal generator scale (App. E.9) | UNTESTED |
| memorization-time law | $\tau_{\mathrm{mem}}$ rises as $N^{+1}$ | the rate law tracks the measurement in slope, level and prefactor (its exponent interval excludes the closed form's); the sign reversal we measured is the amplitude channel of our observable, which has no onset past $N \approx 801$; Table 9 withdraws that observable's exponent as a timescale (App. E.9) | RETRACTED |
| time-rescaling collapse | one curve under $tm/N$ | budget-dependent, ranking last of five rescalings at the default budget, tightest at a longer horizon, second on the strong-signal generator; reported with its budget | MISSED |
| capacity-boundary exponent | $m^{+1}$, then our saturating replacement | the measured object is a budget contour, not an asymptote; both exponents are rejected (App. B.8.1) | OUR ERROR |
| mode-resolved rate law | one rate per mode | fails once the driving operator drifts during training | MISSED |
| independence magnitude law | a parameter-free level for $R_{\mathrm{indep}}$ | withdrawn; the exponent is kept | WITHDRAWN |
| nearest-neighbour test | replication of training images | mis-designed, and its repair refuted | WITHDRAWN |
| early-stopping differentiator | the onset gap is a stopping rule here and not in the temporal model | two different statistics were compared; like for like it is the wrong rule in both | WITHDRAWN |
| our retraction of the capacity claim | effective capacity loses to a chain summary | scored on the ridge-inflated matrix; on the matrix Thm. 3.3 names it ties | WITHDRAWN |
| unconditional no-prevention claim | the sampler's parameters cannot prevent memorization at all | true once the model has learned anything; false at small emission weights, by our own counterexample (§3.6, App. B.7) | OUR ERROR |
| registered prediction on the correction | the over-count dominates | it is a median, not a domination | MISSED |
| nonlinear between-rule orderings | the sampled rule memorizes later, and to a lower plateau, than MAP | read on a relative reconstruction gap that does not track the likelihood; the AIS gap falls with $n$ for both rules (App. E.10.2) | WITHDRAWN |

of §3 hold in 10 of 13 (block A), and the misses concentrate in the asymptotic power laws (7 of 11, block B).

Table 9: **Predicted against measured, in three groups, with each row's scope or budget and its verdict.** Memorization-family exponents name their noise convention, which sets the observation variance either to the generator's total noise (well-specified) or to the nominal corruption level. Withdrawn rows are removed, not counted as missed.

| quantity | predicted | measured (95% CI) | scope or budget | verdict |
|---|---|---|---|---|
| **A. The theory's own statements, integrated at each tier's constants** (§E.5; nothing measured enters a prediction) | | | | |
| $\tau_{\text{gen}}$ per cell, expected flow | the flow's own onset | median ratio 0.91, 61% within 20% | fires in the same 90/90 cells; nothing measured enters (§E.5) | HELD |
| $\tau_{\text{gen}}$ vs. $N$, expected flow | $N^{-0.08}$ above $N = 32$ | $N^{-0.15}$ $[-0.19, -0.11]$ | the remainder sits at $N \leq 16$, where the runs are slower than the flow | MISSED |
| $\tau_{\text{gen}}$ vs. $N$, draw kept (simulated) | $N^{-0.13}$ above $N = 32$ | $N^{-0.15}$ $[-0.19, -0.11]$ | the theory's own draw at every update (§E.5) | HELD |
| $\tau_{\text{mem}}$, expected flow | fires in 16 cells | fires in 33 cells | the dropped draw is the difference (§E.5) | MISSED |
| $\tau_{\text{mem}}$ per cell, draw kept (simulated) | fires in 33 cells | median ratio 0.99, 70% within 20% | agrees on firing in 90/90 cells | HELD |
| $\tau_{\text{mem}}$ vs. $N$, draw kept (simulated) | $N^{+0.18}$; $N^{+0.57}$ non-degenerate | $N^{+0.19}$ $[-0.15, +0.44]$; $N^{+0.87}$ $[+0.34, +1.33]$ | inside both intervals | HELD |
| $\tau_{\text{gen}}$ per cell, MNIST, flow | the flow's own onset | median ratio 0.995, worst 4.2% | all 18 cells within 5%; the data and the step rule, nothing measured (§E.5) | HELD |
| $\tau_{\text{gen}}$ vs. $N$, MNIST, flow | $N^{-0.087}$ | $N^{-0.096}$ $[-0.13, -0.05]$ | $N^{-0.011}$ over the full-rank sizes against a measured -0.012 | HELD |
| $\tau_{\text{mem}}$ per cell, MNIST, flow | the flow's own onset | median ratio 0.997, worst 2.2% | the ratio holds in every cell where both fire, but the two agree on whether it fires in only 15 of 18, below the nine-in-ten this table requires, so the row is scored missed however close the 3 disagreeing cells come (a peak margin of half a per cent) | MISSED |
| $\tau_{\text{mem}}$ vs. $N$, MNIST, flow | $N^{+0.140}$ | $N^{+0.123}$ $[+0.09, +0.18]$ | inside the measured interval (§E.5) | HELD |
| capacity plane per cell (simulated) | the flow's own $\phi_{\text{fit}}$ | median 0.6%, worst 12% | 30/30 points on the same side of the bar | HELD |
| boundary $N^{\star}(m)$ (simulated) | $m^{+0.24}$ at this budget | $m^{+0.24}$, ratio 1.00 | a contour of this budget, not an asymptote; worst 5% | HELD |
| zero-ridge arm, flow at $\phi = 0$ | no generalizing phase, every cell memorizes | 90/90 and 90/90 agree | deterministic arm, so a control on the integration, not on the sampling | HELD |
| **B. Power laws, the classical picture's and our own** (asymptotic exponents, fitted over a finite range) | | | | |
| $\tau_{\text{gen}}$ vs. $N$ | $N^{0}$ | $N^{-0.30}$ $[-0.33, -0.26]$ | as a power law over any range; $N^{-0.15}$ $[-0.19, -0.11]$ above $N = 32$ still excludes zero; a saturating curve fits ($R^2 \geq 0.9994$), which the theorem does not forbid | MISSED |
| $\tau_{\text{mem}}$ vs. $N$, low noise | $N^{+1}$ | $N^{+0.74}$ $[+0.72, +0.76]$ | on the strong-signal generator, where the premise holds, on the amplitude observable; the escape rate gives $N^{+0.69}$ $[+0.67, +0.71]$, both excluding $+1$. The original grid's premise fails (App. E.9) and its own fit, $N^{+0.19}$ $[-0.15, +0.44]$ over the grid and $N^{+0.87}$ $[+0.34, +1.33]$ at $R^2 = 0.26$ over its 18 non-degenerate cells, contained it | MISSED |
| $\tau_{\text{mem}}$ vs. $N$, on $\phi_{\text{fit}}$ | withdrawn | $N^{-1.069}$ $[-1.53, -0.36]$ | amplitude channel $N^{-0.916}$; not a timescale (App. E.9) | WITHDRAWN |
| $\tau_{\text{mem}}$ vs. $N$, weight amplitude | $N^{+1}$ to $N^{+1/2}$ | $N^{+0.931}$ $[+0.87, +0.99]$ | channel flat; on the extended grid, since at the default horizon 48/48 cells were censored | HELD |
| $\tau_{\text{mem}}$ vs. $N$, well specified | $N^{+1/2}$ | $N^{+0.545}$ $[+0.537, +0.554]$ | within a tenth of the prediction, but the interval excludes it (independent instrument) | MISSED |
| $\tau_{\text{gen}}$ vs. $N$, MNIST | $N^{0}$ | $N^{-0.096}$ $[-0.13, -0.05]$ | as a power law over the whole range; flat between the 2 sizes with $N \geq 785$ (local exponent -0.012, no interval), which carry the theorem's premise | MISSED |
| $\tau_{\text{mem}}$ vs. $N$, MNIST, closed form | rising | $N^{+0.123}$ $[+0.09, +0.18]$ | the closed form needs a bulk edge these images lack, so it gives a sign and no magnitude | HELD IN SIGN |

Table 9, continued from the previous page.

| quantity | predicted | measured (95% CI) | scope or budget | verdict |
|---|---|---|---|---|
| memorization at the largest $N$, MNIST (registered) | none | memorizes in every cell | the registered count prediction fails; it implies memorization here under Cor. 3.5's ceiling, so its "sign" is a consistency check, not a test | MISSED |
| boundary $N^\star(m)$, counting | $m^{+1}$ | running slope; no single exponent | a contour of a fixed training budget; no exponent printed | MISSED |
| boundary $N^\star(m)$, ours | $m^{+0.000}$, $\leq 16$ (its constants read off the strong-signal grid) | levels 30 to 68 | budget-dependent object | WITHDRAWN |
| risk minimum inside the onset gap | inside | 0/680 static, 0/162 temporal | the wrong rule in both; gap open in 33/33 | MISSED |
| **C. Searches, instruments, controls and direct tests** (no exponent at stake) | | | | |
| criterion flip, Eq. (7) (registered) | at $C_\perp$ over the bar $= 1$ | 0.962, 0.971 at the two $N$; pooled 0.966 $[0.937, 0.996]$ | signal-free centred data, $m = 1$, 3 seeds per $N$; the registered tolerance (0.05 about 1 at each $N$) is met, but the pooled interval excludes 1; it contains the Tracy–Widom value 0.970, tabulated a priori, not registered (App. E.3) | MISSED |
| occupancy collapse slope, Eq. (20) (registered) | 1 | 0.41 $[0.40, 0.43]$ | 75 cells, small initialization; the per-mode rate holds (0.93 of it, post hoc), the edge-rate aggregate fails (App. E.3) | MISSED |
| occupancy collapse, width dependence (registered) | removed | $\log R_{\mathrm{fl}}$ coefficient -0.01 $[-0.03, 0.03]$ | the same cells, after subtracting $\log R_{\mathrm{fl}} + 2 \log s_0$ | HELD |
| $\tau_{\mathrm{gen}}$ bar sensitivity | flat | survives $[5, 9]$ nats | and a change of estimator | HELD |
| prevention by chain length, chain count | none | 0 of 5, 0 of 3 settings | Cor. 3.5; M15 | HELD |
| useful prevention by step size | none stated | 0 of 24 settings | prevents in 8 of 24; registered bar unreachable; 0 at the reachable one; a search, not a prediction | — |
| $\lambda_{\mathrm{sig}}/\lambda_{\mathrm{mem}}$, original grid | $\gg 1$ | 1.98 (med.) | clears the edge in 0/150 cells (M8, both arms), so on this grid the two groups the ratio compares are not there to be compared (App. E.9) | UNTESTED |
| $\lambda_{\mathrm{sig}}/\lambda_{\mathrm{mem}}$, strong-signal generator | $\gg 1$ | separated in 224 of 231 cells | at the strong-signal generator scale, the grid the widening claim is read on (App. E.9) | HELD |
| $C^{\mathrm{acc}}$ vs. $\lambda$ and $S$ | $\lambda^{-1}$, flat | $\lambda^{-0.99}$, within 4.5% | a window average, not an iterate variance | HELD |
| effective capacity vs. $S$ | flat | 58 to 59 of 128 | on $\widehat{\Sigma}_{mm}$; on $\widehat{\Sigma}_{zz}$ it climbs 80 to 85 | HELD |
| $R_{\mathrm{indep}}$, accumulating | falls | $N^{-0.17}$ $[-0.21, -0.12]$ | magnitude law withdrawn, exponent kept; one scoring rule changed during the run | HELD IN SIGN |

### E.3 Pre-registered direct tests

This subsection gives the two direct tests we registered before running them, the criterion flip and the occupancy collapse, and one extra measured arm.

**Criterion flip.** This test checks whether a noise direction switches from retained to suppressed where (7) says it should. It runs on the solvable tier ($n = 128$, with the solvable grid's generator and seeds). The signal is off and the data are centred, with $m = 1$ and small initialization. Training uses the accumulating rule with $b = 10$, $S = 40$ and a cold start at every M-step, with $\eta_w = 0.01$ and 5000 updates. At a fixed absolute step $C_\perp$ is constant. The grid is therefore laid in $x = C_\perp/[\sigma_H^2(1 + \sqrt{n/N})^2]$, eleven points in $[0.85, 1.10]$, and each point is converted to $\bar{\eta}$ by inverting (7). With a signal, no stable step breaks the bar (8). A direction is retained when its mean energy over the last 30% of training exceeds three times its initial scale. The flip is the interpolated sign change. Three predictions were registered. (P1) The flip lies within 0.05 of $x = 1$ at each $N$. (P2) It lies within 3% of each draw's own top sample eigenvalue. (P3) Exact EM and MAP retain everywhere. 234 cells ran and 0 failed. The value 0.970 is the expected finite-$n$ position of the top sample eigenvalue, computed a priori beside P1, not the prediction. For real Gaussian noise at unit

variances the top eigenvalue of $Y^\top Y$ has Tracy–Widom centring and scale (Johnstone, 2001)

$$\mu_N = \left(\sqrt{N-1}+\sqrt{n}\right)^2, \qquad s_N = \left(\sqrt{N-1}+\sqrt{n}\right)\left(1/\sqrt{N-1}+1/\sqrt{n}\right)^{1/3},$$
$$x_{\mathrm{TW}}(N) = \frac{\mu_N + \mathbb{E}[\mathrm{TW}_1]\, s_N}{N\left(1+\sqrt{n/N}\right)^2},$$

with $\mathbb{E}[\mathrm{TW}_1] = -1.2065$ the mean of the real Tracy–Widom law; $x_{\mathrm{TW}} = 0.966$ at $N = 128$ and 0.974 at $N = 256$, mean 0.970. P1 was registered against $x = 1$ with tolerance 0.05 at each $N$, and it is held (0.962, 0.971). The scoreboard's rule (App. E.2) scores the pooled interval, $[0.937, 0.996]$. That interval excludes 1, so Table 9 marks the row missed.

**Result.** Fig. 15 shows the measurement. The test of (7) holds signal-free centred data fixed ($m = 1$) and increases only the step size $\bar{\eta}$. This raises the ridge $C_\perp$ across the bar. The top fluctuation direction switches from retained to suppressed at $C_\perp$ over the bar $= 0.966$ $[0.937, 0.996]$. This is within the registered tolerance of 1 at each $N$. The pooled interval excludes 1, however, so the interval rule scores it missed. The interval contains the finite-$n$ Tracy–Widom value 0.970, which was tabulated a priori but not registered. Every draw flips within 2.0% of its top sample eigenvalue. Exact EM and MAP retain the direction in 6 of 6 draws.

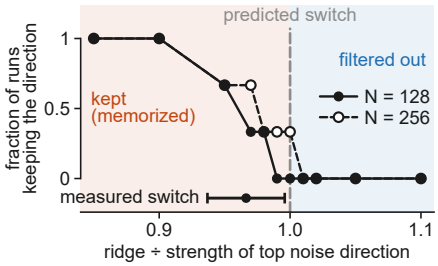



Figure 15: **Criterion flip** (measured). The data are fixed and the step size varies. Each curve is the fraction of 3 seeds keeping the top noise direction at one $N$, against $x = C_\perp/[\sigma_H^2(1+\sqrt{n/N})^2]$. The switch is at 0.966 $[0.937, 0.996]$ (Table 9, block C).

**Exact-moment arm, measured.** This arm reruns the solvable grid's protocol unchanged on its 90 cells, but with no chain. Its E-step (coded only in `M24_criterion_flip/run.py`) returns $(m_y,\ \overline{mm^\top} + H^{-1})$. Held-out likelihood is scored under the learned prior mean, as on the solvable grid, while the integration uses a zero mean. That convention, not the dynamics, is what separates it from Table 1's integrated row.

**Occupancy collapse.** This test checks the collapse law (20) directly. It runs on the solvable tier ($n = 128$, $r = 8$, unit variances) with the strong-signal generator at a fixed step $\bar{\eta} = 0.05$, so $c_{\mathrm{thr}}$ is constant. Training uses the accumulating rule with $b = S = 100$ and one chain started cold at every M-step. The initial weights are $W(0) = s_{\mathrm{init}}\,\mathcal{N}(0, 1/m)$ with $s_{\mathrm{init}} = 0.05$, so $s_0^2 = s_{\mathrm{init}}^2$ and (T2) holds. The learning rate is $\eta_w = 2 \times 10^{-4}$ and the horizon is $6/(2\beta\Delta_N)$. The grid is five widths by six sizes by three seeds, and 90 of 90 cells ran. $Q_{\mathrm{mem}}$ is summed over the registered retained set, the $R_{\mathrm{fl}}$ directions after the signal, rather than over all $n - r$ fluctuation directions as (20) is worded. A post hoc diagnosis explains the miss. The per-mode rate holds and the aggregation fails, because the occupied band lies below the edge and grows more slowly than it. At finite $n$ the top fluctuation eigenvalue also sits below the Marchenko–Pastur edge. The band integral of App. B.8 is the form to test next.

### E.4 Causal ridge intervention

Here we test whether the ridge alone, and not the sampling noise, is what keeps MCPC from memorizing.

**Design.** The test runs on the solvable tier at the default step. It uses the solvable grid's generator ($n = 128$, $r = 8$, unit variances) with the training data centred (Thm. 3.4's hypothesis) and width $m = 32$. The E-step runs one chain of $K = 64$ steps per datum from $m_y$ at $\bar{\eta} = 0.5 \cdot 2/\lambda_{\max}(H)$ and hands its final state to the M-step. The M-step uses $\eta_w = 0.02$ for 10,000 updates. The sizes are $N \in \{64, 256, 512, 2048, 4096, 8192\}$, with three seeds and a held-out set from the training

Table 10: **Causal ridge intervention, by arm.** The top block is the $2 \times 2$ of ridge in the statistic against sampling noise. The dose and compute controls, exact sampling and the explicit baselines follow. Retention and $R_{\text{fl}}$ ($R_{\text{fl}}$ of $m - r$) over the $(N, \text{seed})$ cells with $N^\star_\infty < N \leq 4096$; likelihood and $\phi_{\text{fit}}$ at $N = 4096$ (for small-step MALA, $N = 2048$, its largest); seed means.

| arm | statistic $\widehat{\Sigma}_{zz}$ handed to the M-step | noise | retains $u_{r+1}$ | $R_{\text{fl}}$ | $\phi_{\text{fit}}$ | NLL$-$oracle |
|---|---|---|---|---|---|---|
| MCPC (R+,N+) | chain's $\overline{zz^\top}$ (mean $\widehat{\Sigma}_{mm} + C_K$) | yes | 0 of 6 | 0 | 0.79 | 1.02 |
| ridge only (R+,N$-$) | $\widehat{\Sigma}_{mm} + C_K$ | no | 0 of 6 | 0 | 0.79 | 1.02 |
| noise only (R$-$,N+) | chain's $\overline{zz^\top} - C_K$ | yes | 6 of 6 | 24 | 1.01 | 37.68 |
| MAP (R$-$,N$-$) | $\widehat{\Sigma}_{mm}$ | no | 6 of 6 | 24 | 1.01 | 37.67 |
| dose | $\widehat{\Sigma}_{mm} + \omega C_K, \omega \in \{0.25, 0.5, 1, 2, 4\}$ | no | flips at $\omega^\star$ (P3) | | | |
| MAP, compute-matched | $\overline{\hat{z}\hat{z}^\top}$, 64 descent steps from $\mu$ | no | 6 of 6 | 24 | 1.01 | 37.67 |
| exact posterior draw | $\overline{zz^\top}$, $z \sim \mathcal{N}(m_y, H^{-1})$ | yes | 6 of 6 | 24 | 1.12 | 0.39 |
| small-step MALA | $\overline{zz^\top}$, 64 adjusted steps | yes | 3 of 3 | 24 | 1.13 | 0.72 |
| exact EM | $\widehat{\Sigma}_{mm} + H^{-1}$ | no | 6 of 6 | 24 | 1.12 | 0.39 |
| EM at $c_{\text{thr}}$ | $\widehat{\Sigma}_{mm} + H^{-1}$, $\sigma'^2_O = c_{\text{thr}}$ | no | 0 of 6 | 0 | 0.98 | 9.22 |
| EM, $\sigma'^2_O$ tuned | the same, $\sigma'^2_O$ by validation likelihood | no | 6 of 6 | 24 | 1.12 | 0.39 |
| explicit ridge, tuned | $\widehat{\Sigma}_{mm} + \lambda I$, $\lambda$ by validation likelihood | no | 6 of 6 | 24 | 0.77 | 0.39 |
| hard threshold | none, Gavish and Donoho (2014)'s cut on $\widehat{\Sigma}_y$ | — | 0 of 6 | 0 | 0.80 | 1.14 |

generator. Each arm changes only what the E-step hands the M-step (Table 10). The ridge $C_K$ is (5) at $S = 1$, $b = K - 1$, recomputed from the current $H$ and $\bar{\eta}$ at every update. A direction $u_i$ of the sample covariance is retained when its energy $\|u_i^\top W\|^2$, averaged over the last 30% of training, exceeds 0.01. $R_{\text{fl}}$ counts the retained directions among the $m - r$ fluctuation slots. The closed form sets $\bar{\eta}$ at the top mode's fixed point ($\bar{\eta}\lambda_1 = 2 \cdot 0.5$). It then places the crossing at $N^\star_\infty = 1078$. For each draw it places the dose flip at the $\omega^\star$ where $\omega C_\perp$ reaches that draw's top fluctuation eigenvalue $c_{r+1}$. 436 cells ran and 0 failed. The chain arms were not run at $N = 8192$ (compute), so P1 and P2 are scored at $N \leq 4096$. The registered MALA arm, at MCPC's own step, accepted at most 1.1% of its proposals. It stayed at $m_y$ and so ran as MAP. It ran on its first cells only, and a small-step MALA ($\bar{\eta}\lambda_{\max} = 0.2$, acceptance at least 90%), with its own pre-registered prediction P5b, was run at $N \in \{256, 2048\}$ in its place.

**Verdicts.** Each registered prediction is stated below, followed by its verdict. P1 says the ridge-only arm matches MCPC in $\phi_{\text{fit}}$ and $\tau_{\text{mem}}$ within seed spread at every $N$. It is held (5 of 5 sizes). P2 says the noise-only arm memorizes like MAP, with $\tau_{\text{mem}}$ within a factor 1.5. It is held (15 of 15 cells, median ratio 1.00). P3 says the ridge-only arm's retention flips inside the grid interval that contains the closed-form $\omega^\star$. Grid points within 5% of $\omega^\star$ are not scored. It is held (90 of 90 points). P4 says MAP still memorizes when given the chain's 64 gradient evaluations per datum and update. It is held (18 of 18 cells). P5 says exact posterior draws (and the registered MALA) retain the top fluctuation direction in every cell, with $R_{\text{fl}}$ at least MCPC's. It is held (22 of 22). The registered MALA's pass is uninformative, since it ran as MAP. P5b says the same for small-step MALA, and it is held (6 of 6). P6 says exact EM with its noise variance set to MCPC's threshold, $\sigma'^2_O = c_{\text{thr}}$, retains $u_{r+1}$ exactly where MCPC does, with $|\Delta R_{\text{fl}}| \leq 2$. It is missed (14 of 15). P7 was registered as the expected outcome. It says exact EM with $\sigma'^2_O$ and an explicit ridge $\lambda I$, each tuned on a separate validation draw, reach a held-out likelihood no worse than MCPC's at every $N$. It is missed (2 of 5 sizes).

**Reading.** Past $N^\star_\infty$ the arms split by whether the M-step statistic carries $C_K$, not by whether the E-step samples (Table 10, Fig. 16). MCPC and the ridge-only arm keep no fluctuation direction. The noise-only arm, MAP, compute-matched MAP, exact draws, small-step MALA and exact EM keep the top one in every cell, with $R_{\text{fl}}$ near $m - r$. Subtracting $C_K$ from the chain's statistic turns MCPC into MAP (median $\tau_{\text{mem}}$ ratio 1.00). Adding it to the posterior-mean moments turns MAP into MCPC. MAP's extra 64 gradient steps change nothing. Sampling without the integrator's bias memorizes as exact EM does, cell for cell. So the protection is the inflation of $C$, not the sampling. A post hoc reading is that at these sizes the draw adds no measurable memorization (P1, and exact draws against exact EM). The threshold-matched EM misses P6 in a single row (14 of 15 held), where its $R_{\text{fl}}$ falls just outside the tolerance. On likelihood, P7 is missed, as registered. Validation tuning of $\sigma'^2_O$ picked $\sigma'^2_O = 1$, which is plain exact EM, in 18 of 18 cells. That choice is worse than MCPC below the crossing (3.49 against 1.08 nats at $N = 256$) and better above it (0.39 against 1.02 at $N = 4096$). The tuned ridge is never worse than MCPC by more than 0.02 nats, and it is

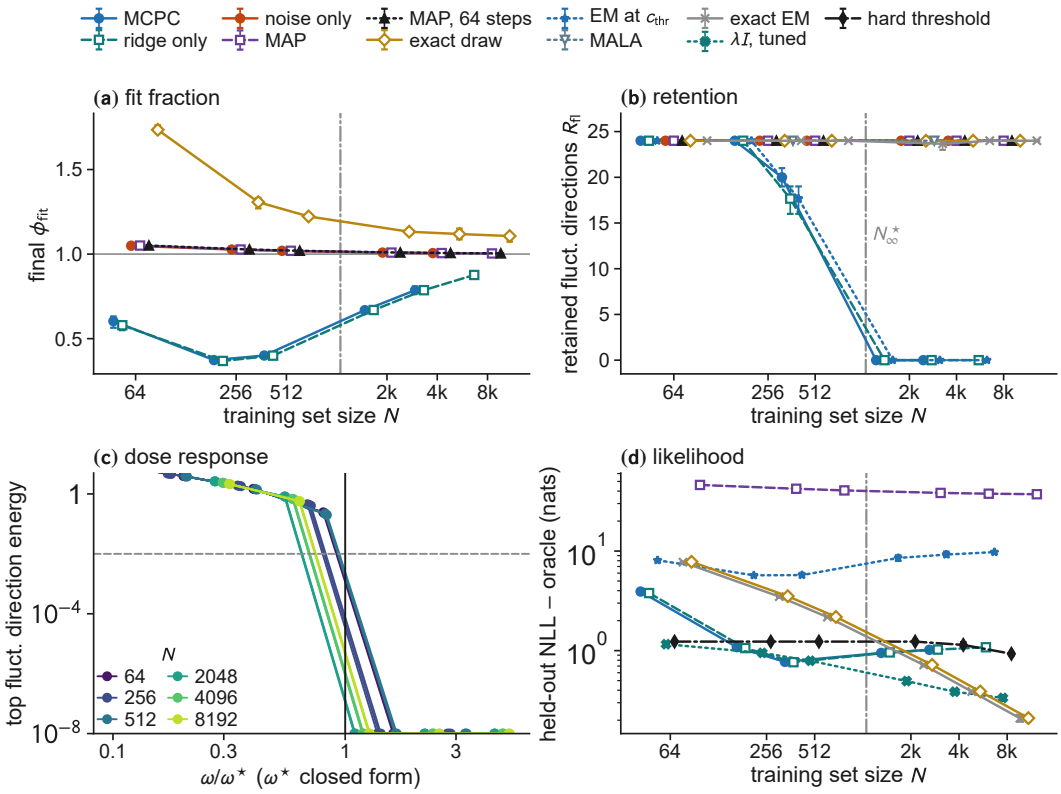



Figure 16: **Causal ridge intervention.** (a) Final $\phi_{\text{fit}}$ and (b) retained fluctuation directions $R_{\text{fl}}$ against $N$ per arm (seed means, bars the seed range). Arms are offset in $N$ in (a), (b) and (d) so that coincident points stay visible. MCPC and ridge only coincide, as do noise only and MAP. The dash-dot line is $N_\infty^\star$. (c) Dose response of the ridge-only arm, showing the energy of the top fluctuation direction against $\omega/\omega^\star$, one line per $(N, \text{seed})$ coloured by $N$. The dashed line is the retention bar and the solid line is $\omega = \omega^\star$. (d) Held-out NLL minus the oracle's, per arm. EM with tuned noise variance is exact EM in every cell and is not drawn.

better past the crossing (0.39). The threshold-matched EM pays 9.22 nats, because its inflated noise variance is also its model's noise floor. So the sampler sets the retention threshold for free, as a function of $\bar{\eta}$. It does not beat a tuned ridge, and the fluctuation directions that exact EM keeps past the crossing cost it little likelihood.

### E.5 One flow, both onsets

This subsection compares the onsets the theory predicts with the onsets the runs measure. We integrate the theory's M-step flow (Eqs. 6–7) at the constants of each setting $(N, m, \text{seed})$, with no chain run, and compare it with the runs' onsets. **With its draw kept, one integrated flow reproduces both onsets setting by setting. On images it misses some calls of whether memorization is reached.** Fig. 17 shows the onsets. The median integrated/measured ratio is 0.97 for $\tau_{\text{gen}}$ (90 settings) and 0.99 for $\tau_{\text{mem}}$ (33 settings where memorization is reached). Flow and runs agree on whether memorization is reached in 90/90 settings. Without the draw the two ratios are 0.91 and 2.14. So the ridge sets the threshold of the expected flow, and at small $N$ the draw adds memorization beyond it in this integrated flow. No such extra memorization is measured at $N \geq 64$ (App. E.4), and §3 does not analyse it. On MNIST the ratios are 0.995 and 0.997 over 18 settings. There flow and runs disagree on whether memorization is reached in 3 settings, which are scored as missed (Table 9). Crossing. In the integrated flow the chain's ridge stops retaining the flat direction of (7) between $N = 1024$ and 2048. With exact moments it never stops. Runs at the default initialization never memorize (0/6 below, 0/12 beyond). On the solvable tier, retention (Thm. 3.4) and the $\phi_{\text{fit}}$ onset separate for $N \gtrsim 512$. The integrated flow retains up to $N_\infty^\star \approx 1326$, yet no run memorizes. So retention is necessary for the $\phi_{\text{fit}}$ onset but not sufficient. On MNIST, which has no bulk edge (App. E.9), no setting at $N \geq 4096$ memorizes (0/9).

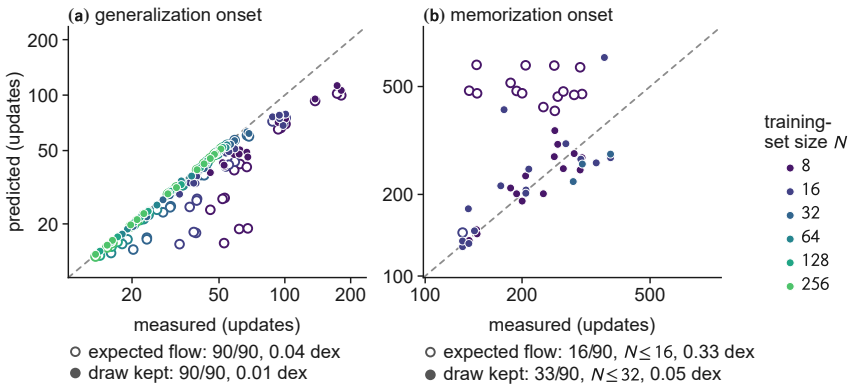



Figure 17: **Predicted against measured onsets, for the expected flow and for the flow with its draw.** (a) $\tau_{\text{gen}}$, (b) $\tau_{\text{mem}}$. Each dot is one setting $(N, m, \text{seed})$ where both onsets are reached, coloured by $N$. Open dots show the expected flow, filled dots the flow with its draw, and the dashed grey line is equality. The keys give the number of settings per variant and the median $|\log_{10}(\text{predicted}/\text{measured})|$ in dex. $\tau_{\text{mem}}$ is reached only at small $N$ (runs $N \leq 32$, expected flow $N \leq 16$, flow with the draw $N \leq 32$). $\tau_{\text{gen}}$ is read against the original, offset held-out set.

### E.6 Details of the tests in §4.1

This subsection gives the details behind the tests of §4.1.

**The realized ridge and Table 1 (§4.1.1).** Fig. 4 shows the ridge. In $H$'s eigenbasis over $[1.00, 65.0]$ the realized ridge follows (5) within 4.5%. It stays flat in the window up to $S = 2048$, with no free coefficient (Table 9, block C). The first two rows of Table 1 compare exact moments with the finite chain, both without the draw. On the measured rows the held-out−train NLL gap $\Delta$ (post hoc) ranks exact moments above the sampled E-step, as $\phi_{\text{fit}}$ does (82/90 settings). But it ranks MAP below MCPC where $m \geq n$ (34/36; App. E.8). In the integrated flow the draw adds memorization (17 settings, a factor of 2.1, all at $N \leq 32$). Runs at $N \geq 64$, checked post hoc, show none (App. E.7). The draw also moves generalization (paired ratio up to 2.65). MAP is a divergence control. It cannot reach the bar because $\phi = 0$ has no interior fixed point (App. B.5). With the draw dropped in both variants, the generalization contrast belongs to the term (90 against 0 settings). Measured post hoc, MCPC's learned gain is zero for every direction left of $c_{\text{thr}}$ (Fig. 3c). App. E.5 compares the theory's onsets with the runs' (Fig. 17) and lists the settings on MNIST where it misses whether memorization is reached. In Fig. 4 the realized ridge is measured at windows $S = 1$ and 2048 on a fixed decoder ($n = 64$, $m = 32$, $r = 12$, $\sigma_H = 1$, $\sigma_O = 0.5$, step $2 \times 10^{-3}$, $N = 20{,}000$; M1).

**Which directions carry the contrast (Fig. 5c, post hoc).** A latent direction is stiff if the posterior pins it tightly (high precision, a large eigenvalue of $H$) and flat if it does not (low precision; the flattest, $\lambda = \sigma_H^{-2}$, is the one Eq. (7) reads). We re-integrate the no-draw flow of Table 1 with the inflation of the flat direction alone. It then memorizes in 81/90 settings, near exact EM's 89. With the inflation of the stiff directions alone it memorizes in 30, most of the way to the full chain's 16 (Fig. 5). Along these trajectories the flat direction's ridge stays at $C_\perp/\sigma_H^2 \in [1.01, 1.44]$. This is below the Marchenko–Pastur bar ($\geq 2.91$), so by Eq. (7) the chain would still retain at the fixed point (inferred, not run). The contrast of Table 1 is therefore a finite-time effect of stiff-direction inflation, which shrinks learned amplitudes (App. E.7). It is not evidence for the threshold.

**The causal test (Fig. 5a,b, pre-registered).** This test is scored against the pre-registered predictions P1–P7 (App. E.4). Past $N_\infty^\star = 1078$ (Fig. 5a), MCPC and the ridge-only variant retain the top fluctuation direction in 0 of 6 and 0 of 6 settings. The noise-only variant, MAP, compute-matched MAP and exact posterior draws retain it in every setting (6 of 6, 6 of 6, 6 of 6, 6 of 6), with $R_{\text{fl}} = 24$. Subtracting $C_K$ makes MCPC match MAP in $R_{\text{fl}}$ and $\tau_{\text{mem}}$ (median $\tau_{\text{mem}}$ ratio 1.00). Adding it makes MAP match MCPC. MAP's extra gradient steps change nothing (P1, P2, P4 held). Exact EM with its noise variance set to MCPC's threshold matches MCPC within the registered $|\Delta R_{\text{fl}}| \leq 2$ in 14 of 15 settings, not all (P6 missed). When the ridge is scaled to $\omega C_K$, the ridge-only variant's

Table 11: **Channel split of the no-draw contrast of Table 1** (integrated, post hoc). Settings, bars and budget are those of Table 1 (90 settings), and only the ridge differs. The last column counts memorizing settings at $N = 8, 16, 32, 64, 128, 256$.

| variant | ridge | reaches the bar | memorizes | memorizing, by $N$ |
|---|---|---|---|---|
| exact posterior moments | $H^{-1}$ | 78/90 | 89/90 | 15, 15, 15, 15, 15, 14 |
| flat only | $C_\perp H^{-1}$ | 81/90 | 81/90 | 15, 15, 15, 15, 15, 6 |
| stiff only | $C_K/C_\perp$ | 90/90 | 30/90 | 15, 15, 0, 0, 0, 0 |
| finite chain, no draw | $C_K$ | 90/90 | 16/90 | 15, 1, 0, 0, 0, 0 |

retention switches on the predicted side of the closed-form $\omega^\star$ at 90 of 90 scored points (Fig. 5b; P3 held). The step-size flip on signal-free data is a registered miss on its pooled interval (App. E.3). Below $N^\star_\infty$ MCPC also retains the top fluctuation direction at convergence, so there the ridge only delays memorization (Table 1). In Fig. 5, compute-matched MAP takes the chain's 64 gradient steps per datum, and the chain variants are not run at $N = 8192$. Points are seed means and bars the seed range, and variants are offset in $N$ for legibility. In (b) there is one point per dose, $N$ and seed.

### E.7 Where the inflation acts (post hoc)

This subsection locates the latent directions in which the chain's inflated ridge changes the outcome. **Scope of the step rule.** On the solvable tier $\bar\eta = 0.5 \cdot 2/\lambda_{\max}(H_t)$ is recomputed at every update. The chain hands its final state to the M-step ($K = 64$, $S = 1$, $b = K - 1$). So at unit variances (5) gives $C_\perp/\sigma_H^2 = [1-(1-\bar\eta)^{128}]/(1-\bar\eta/2)$, which is above 1 iff $\lambda_{\max}(H) \lesssim 31$ (image tier, $S = 64$, $b = 800$, gives $\lesssim 265$; closed forms). Along every trajectory behind Table 1 $\lambda_{\max} \le 26.3$. These are the 48 measured settings whose top singular value was saved, and every integrated setting of both chain rows. So $C_\perp/\sigma_H^2 \in [1.01, 1.44]$, and it is 1.04–1.10 at the chain's own $\tau_{\mathrm{mem}}$ (Fig. 18). The sign is right and the margin small. The bar $(1 + \sqrt{n/N})^2$ is at least 2.91 for $N \le 256$, above every $C_\perp$. So at the fixed point the chain retains fluctuation in every setting of Table 1, and its "memorizes" is a finite-time $\phi_{\mathrm{fit}}$-onset effect. The threshold of (7) decides retention only for $N \gtrsim N^\star_\infty$ (1078 to 1326 at this step). App. E.4 tests it causally there. **Channel split.** We re-integrate the no-draw flow of Table 1 with two ridges built from $C_K$ (Table 11). The flat-only ridge, $C = C_\perp H^{-1}$, is exact EM's ridge raised uniformly by the flat direction's factor. The stiff-only ridge, $C = C_K/C_\perp$, is the chain's ridge with that factor divided out. The flat direction then sits at exact EM's $\sigma_H^2$, and the stiffer directions of $H$ keep their larger inflation. Flat only memorizes nearly as exact moments do, and stiff only nearly as the chain does. So the contrast of Table 1 comes from inflation in the stiff directions, which shrinks the learned amplitudes, and not from $c_{\mathrm{thr}}$. **The draw.** The integrated with-draw row's extra memorizing settings (33 against 16) all lie at $N \le 32$. In measured runs at $N \ge 64$ (App. E.4) the draw adds no measurable memorization. MCPC matches the ridge-only variant (P1), and exact draws match exact EM. No measured run at those sizes reproduces the integrated contrast.

### E.8 Likelihood gap on the solvable tier (post hoc)

This subsection checks whether a second measure of memorization, the likelihood gap $\Delta$, agrees with $\phi_{\mathrm{fit}}$. **Definition.** $\Delta(t)$ is the mean held-out minus the mean training negative log-likelihood, in nats per sample, under the model's exact marginal $\mathcal{N}(0, \Sigma_{\mathrm{model}})$ of (29). It is computed on 4096 held-out points from the training generator ($W^*$ fixed, fresh latents and noise). We re-run every setting of the measured variants of Table 1 (the sampled E-step, MAP and the exact-moment variant) with the producing code. The only addition is one probe that consumes no randomness, and $\phi_{\mathrm{fit}}$ reproduces the stored runs to 3e-15. The integrated rows score the flows of App. B.4 (no run) the same way on the same held-out draw. **Noise floor.** A model equal to $\Sigma_{\mathrm{pop}}$ has per-sample NLL $\frac{1}{2}\chi^2_n$ plus a constant, whatever $\Sigma_{\mathrm{pop}}$ is. So its gap is exactly $\frac{1}{2}(\chi^2_{nN_{\mathrm{te}}}/N_{\mathrm{te}} - \chi^2_{nN}/N)$. The bar is its 95th percentile, from 4.6 nats at $N = 8$ to 0.85 at $N = 256$. $R_{\mathrm{fl}}$ uses the retention rule of App. E.4 in the sample eigenbasis. **$\Delta$ ranks exact moments above the chain, as $\phi_{\mathrm{fit}}$ does. It ranks MAP below MCPC where $m \ge n$, which $\phi_{\mathrm{fit}}$ does not.** Every setting of every variant ends above its noise floor (Table 12). At $N \le 256$ and $p = nm \ge 2048$ any fitted coder's training likelihood beats its held-out one by many nats. So $\Delta$ above the bar separates no variant here. Across settings

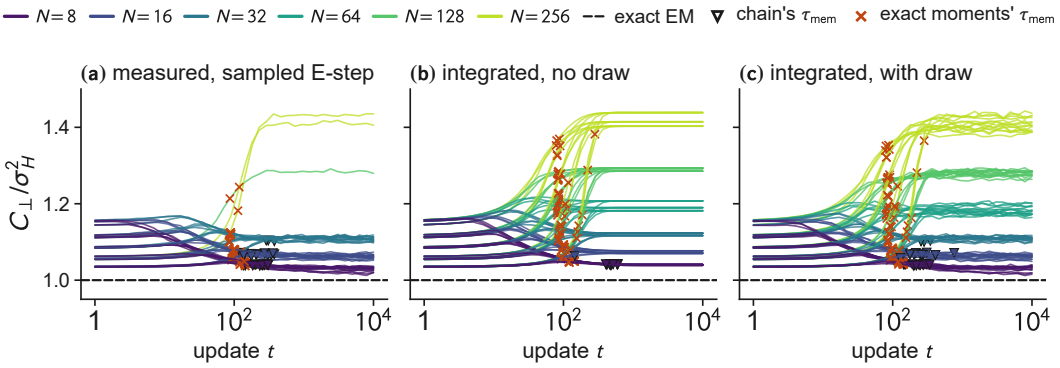



Figure 18: **How much the chain enlarges the ridge over training, by data size** (post hoc). The curves show $C_\perp/\sigma_H^2$ from (5) at each checkpoint's step, one line per setting of Table 1, coloured by $N$. Panel (a) is measured (the 48 settings whose top singular value was saved), and (b), (c) are the theory's integrated flow without and with the draw. The dashed line is exact EM. Every curve stays above exact EM, so the chain's ridge is always the larger. The enlargement is set by the step, which shrinks as the learned weights grow. With few data the weights also fit the inflated finite-sample spectrum, so the step is small and the ridge is barely enlarged. With more data it is enlarged more. Memorization onsets (triangles for the chain, crosses for exact moments) fall at small $N$, where the enlargement is a few percent. So the contrast of Table 1 comes mainly from the other latent directions (App. E.7).

Table 12: **The likelihood gap $\Delta$ on the variants of Table 1** (post hoc; 90 settings per variant, final checkpoint). Peak is the trajectory's largest $\Delta$, and $\rho$ is the Spearman correlation across the variant's settings.

| variant | median $\Delta$ (nats) | peak | above bar | median $\phi_{\text{fit}}$ | median $R_{\text{fl}}$ | $\rho(\Delta, \phi_{\text{fit}})$ | $\rho(\Delta, R_{\text{fl}})$ |
|---|---|---|---|---|---|---|---|
| integrated, no chain run | | | | | | | |
| exact moments, no draw | 28.8 | | 90/90 | | | +0.96 | −0.70 |
| finite chain, no draw | 23.9 | | 90/90 | | | +0.98 | −0.61 |
| chain, with the draw | 24.8 | | 90/90 | | | +0.99 | −0.66 |
| no ridge | 31.4 | | 90/90 | | | +0.86 | −0.36 |
| measured | | | | | | | |
| exact moments | 28.8 [25.8, 38.9] | 29.1 | 90/90 | 1.77 | 24 | +0.96 | −0.70 |
| MCPC | 24.7 [20.4, 35.0] | 25.1 | 90/90 | 0.70 | 24 | +0.99 | −0.68 |
| zero ridge, MAP | 31.4 [25.4, 34.8] | 31.4 | 90/90 | 1.06 | 56 | +0.86 | −0.36 |

$\Delta$ tracks $\phi_{\text{fit}}$ and falls with $N$. With a ridge, $R_{\text{fl}}$ rises with $N$ (a sample of $N$ points has at most $N - r$ fluctuation directions), hence $\rho(\Delta, R_{\text{fl}}) < 0$, while MAP holds every slot it has at every $N$. Within a setting, the sampled E-step's $\Delta$ is below exact moments' in 82/90 settings ($\phi_{\text{fit}}$ gives 90). Against MAP the two measures disagree. With $m < n$ MCPC's $\Delta$ is the lower in 45/54 settings ($\phi_{\text{fit}}$ gives 40). But with $m \geq n$ MAP's is the lower in 34/36 ($\phi_{\text{fit}}$ ranks MCPC lower in 23). There the zero-ridge flow, which has no interior fixed point (App. B.5), inflates $W$ so that training and held-out NLL rise together. Their difference shrinks, so $\Delta$ reads MAP's divergence as less memorization. By median, $\Delta$ orders the variants MAP, exact moments, MCPC, and $\phi_{\text{fit}}$ orders them exact moments, MAP, MCPC. Both put MCPC lowest. The integrated flows repeat the pattern. The chain's $\Delta$ is below exact moments' in 90/90 settings ($\phi_{\text{fit}}$ gives 90), the draw raises it in 87/90 ($\phi_{\text{fit}}$ gives 90), and with $m \geq n$ MAP's is below the chain's in 33/36 ($\phi_{\text{fit}}$ gives 0).

**Ridge dose on the onset-gap grid.** The onset-gap grid has four ridge variants ($w = 6$, $m = 32$, $\kappa = 0.05$, $N = 16$ to 16384, three seeds). They are read from their stored checkpoints. The grid's held-out set is drawn from the training generator, and the runs record both likelihoods at every checkpoint, under the learned prior mean. On the solvable tier that convention raises $\Delta$ by +1.48 nats in the median. Median final $\Delta$ is 10.5, 8.4, 6.9 and 6.0 nats for MAP and the $B = 2, 8, 64$ chains (1.24 to 0.30 at $N = 16384$). Each step up in ridge lowers $\Delta$ within a setting in 31, 33 and 31 of 33 settings. $\phi_{\text{fit}}$, which ends near one in every variant, does so in 15, 28 and 26. Every variant holds all $m - r$ slots (median $R_{\text{fl}} = 24$), so $R_{\text{fl}}$ separates none of them. Here $m < n$, the regime in

which $\Delta$ and $\phi_{\text{fit}}$ agree on the solvable tier. The counts above the floor are 33, 33, 33 and 32 of 33. Over the three steps, more ridge gives the lower $\Delta$ in at least 31 of 33 settings, and the lower $\phi_{\text{fit}}$ in 15 to 28 (§4.2).

**Criterion flip.** We re-run part A of the criterion-flip test (App. E.3) with a held-out set from its generator (observation noise only; $\phi_{\text{fit}}$ reproduces exactly). With $m = 1$ and no signal, $\Delta$ sits at the finite-sample floor. The chain variant's mean over its grid is 1.11 nats at $N = 128$ and 0.54 at $N = 256$, against bars of 1.18 and 0.85. 23/66 of its settings exceed the bar, as set by the training draw rather than the step. Across the flip, from $C_\perp/\text{bar} = 0.85$ to 1.10, $\Delta$ falls in 6 of 6 $(N, \text{seed})$ rows, by a median 0.047 nats. This is the criterion's sign, far below the bar. Exact EM and MAP, which retain (P3), sit at 2.15 and 2.43 nats ($N = 128$), and at 1.11 and 1.35 ($N = 256$).

### E.9 Supporting readings

This subsection collects three further readings that support the main claims.

**Ridge law and the independent instrument.** The realized ridge follows (5) within 4.5% and is flat in the window up to $S = 2048$. The independent instrument is a variant of the random-features score network analysed by Bonnaire et al. (2025) (frozen random first layer, trained readout, one noise level), with the readout tied to a learnable feature map. It shares no code path with the theory. Its mode decoupling reproduces to a relative error of $3.04\times10^{-5}$ on the signal mode and $5.60\times10^{-4}$ on the top fluctuation mode, with no fitted parameter. The alignment that Thm. 3.3 presupposes holds at 1.0000. The instrument splits capacity as §3.7 and (16) do, with $\tau_{\text{gen}} \propto m^{-0.21}$ against a rate $\propto m^{-0.010}$. Of the 23 pre-registered checks, 17 hold. The slope above the survival hinge and the hinge's location are among the misses, so only its dead-mode side is claimed.

**Which observable the memorization time is about.** A bar-crossing time splits into an amplitude channel (first term) and a rate channel (second),

$$\frac{d\ln\tau}{d\ln N} = \frac{d\ln\ln(D_*/D_0)}{d\ln N} - \frac{d\ln\nu_{\text{esc}}}{d\ln N}, \tag{30}$$

with $D_0$, $D_*$ the curve's start and bar and $\nu_{\text{esc}}$ its escape rate. So it reproduces Thm. 3.3 only if its amplitude channel is flat in $N$. We use one pre-registered grid (48 settings, identical trajectories). There the weight projection on the top fluctuation direction starts where the initialization put it, and it crosses its bar as $N^{+0.931}$. $\phi_{\text{fit}}$ starts a sampling error from its bar, and it crosses as $N^{-1.069}$. We regress the measured escape timescale, setting by setting, on the closed form's, which is built from each setting's own bulk edge and realized ridge. This is a test of form, not a prediction. The measured timescale tracks the closed form at slope +1.081, offset by the factor of two that a squared amplitude implies. At the default initialization the memorization mode starts above its own fixed point past $N \approx 801$ and descends. So there is no onset to time, and the $\phi_{\text{fit}}$ exponent is withdrawn as a timescale (Table 9).

**Two-bulk premise.** This paragraph checks the premise that the spectrum splits into a signal part and a fluctuation bulk. On the original grid the signal eigenvalue clears the edge in 0 of 150 settings. The premise is arithmetically unreachable at $n = 128$ for every $N < n$ before any update, so it is untested there, not failed. At the strong-signal generator scale it holds in 70 of 90 settings. On natural images no admissible split of the latent spectrum exists in 4 of 9 MNIST runs. On the data second moment $\widehat{\Sigma}_y$, the matrix Thm. 3.4 names, none is admissible on 4 of 4 datasets over the scanned range. There the leading remaining eigenvalue stands 1.54 to 1.85 times above its own bar.

### E.10 Nonlinear coder

This section tests the theory's predictions on a nonlinear decoder trained on real images.

**Protocol.** The energy is $F(z;x) = \frac{1}{2}\|x - W_2\tanh(W_1 z)\|^2 + \frac{1}{2}\|z\|^2$ on pixels $x \in [0,1]^D$ ($D = 784$ for MNIST and Fashion-MNIST, 3072 for $32 \times 32$ RGB). The decoder is a tanh network of width $(h, m) = (256, 64)$, with $z \in \mathbb{R}^m$. The sampled rule averages $\nabla_W F$ over an Euler–Maruyama chain on $z$ that starts at $z = 0$. The chain uses step 0.01, 50 burn-in steps, and 16

states thinned by 2 (steps $52, \ldots, 82$). The MAP variant instead uses the state after 50 unconverged gradient steps from $z = 0$. Both variants train by Adam at $10^{-3}$ (no Hebbian rate), with minibatch 100, for 131072 steps. Data, initialization and order are shared by seed. $N$ runs from 128 to 50000, with 5 seeds, giving 94 finished settings $(N, \text{seed})$ (Fashion-MNIST has 46, a partial grid). $\tau_{\text{gen}}$ is the step at which the pixel-space Fréchet distance of 5000 prior samples to the test set completes 90% of its drop. Memorization is the MAP-scored relative gap between the energy on training data and on a disjoint held-out part of the training split (null zero). The plateau is its mean over steps $[2^{16}, 2^{17}]$. The bar crossing is the first checkpoint above 5% that stays above at the next one. The half-plateau step is read only where the plateau is $\geq$2.5% of the held-out energy. The compute-matched MAP variant runs 82 inference steps. Every variant is scored by a 50-step MAP inference, and 6 compute-matched settings were rescored by a 400-step one. On RGB CIFAR-10 at this width neither rule learns image-like samples. The pixel Fréchet distance is 833 (sampled), 2028 (MAP) and 190 for the mean image.

CelebA (`experiments/M22_model_data/celeba_full`). Faces are centre-cropped to $140 \times 140$ and resized to $32 \times 32$ (bilinear, RGB). The full-data decoders train on the whole official train split (162770 faces). Their held-out pool is the official valid split (19867 faces). They use seed 0, $2^{16}$ steps, and the E-steps and optimizer above. The $N = 512$ decoders draw their training set from that split and a size-matched held-out pool from its remainder. Nearest neighbours are searched over the full train split (pixel $L_2$). For prior samples (Fig. 7, App. E.13.1) we take 10 000 draws per set, each decoded in one pass. The decoder has no bias, so $x(-z) = -x(z)$. A sample is therefore shown as $x(sz)$, $s = \text{sign}(x(z) \cdot \mu)$ with $\mu$ the mean training face, clipped to $[0, 1]$. Every distance is computed on this displayed image. Weights and run outputs stay out of the repository. The scripts there regenerate them, and from them every figure and number.

**Readings (MNIST unless named).** Memorization is read as the relative train/held-out gap above a 5% bar. The copy criterion is reached in 0/470 settings $(N, \text{seed})$, and over training the samples never become copies (Fig. 19). This gap does not track the likelihood (App. E.10.2). On the pixel-space Fréchet distance $\tau_{\text{gen}}$ is flat in $N$ (a transient of the MAP variant). On held-out energy it moves as $N^{+0.13}$ and $N^{+0.30}$. On this gap the intervals of the half-plateau slopes contain zero, while the plateau falls with $N$. The bar crossing moves as $N^{+2.54}$ and $N^{+1.33}$, and at the smallest $N$ both variants cross it before $\tau_{\text{gen}}$. On Fashion-MNIST (a partial grid) the crossing moves as $N^{+2.69}$ and $N^{+1.48}$ (30/50 and 22/50 censored). The ridge is measurable. The chain's window-mean second moment about the mode is 1.01 times (5) $[0.91, 1.37]$, against 0.75 for $H^{-1}$ alone, except in the stiff directions (1.52). On CelebA both rules reconstruct with 400 inference steps from $z = 0$ (Fig. 20). MCPC reconstructs toward a typical face, and PC is sharper. We also crossed both full-data decoders with three encoders (the chain's window mean, its last state, and MAP; post hoc) on 200 held-out faces. Sharpness (reconstruction over face pixel spread) follows the decoder, 0.79–0.84 under MCPC's and 0.94–1.09 under PC's.

**Inference from a noise input (not generation).** Here we check what the E-step does with noisy inputs. Fig. 21 feeds pure noise to the E-step, and Fig. 22 (post hoc) mixes noise and a face in the input. This is inference on a corrupted input, not the generative path. Each variant's E-step runs 400 steps from $z = 0$ on $y_\lambda = \lambda x + (1 - \lambda)\varepsilon$. MCPC reports the time average of the last 200 Langevin states and PC the MAP state. Here $x$ is one of 1 000 held-out faces and $\varepsilon \sim \mathcal{N}(\bar{x}\mathbf{1}, \sigma^2 I)$ as above. The outputs of both variants improve monotonically as the input carries more of the face. Inception FID (against 9 000 other test faces, with clean faces at 14) falls from 248 (MCPC) and 259 (PC) at $\lambda = 0$ to 112 and 89 at $\lambda = 1$. Likewise $\cos(\hat{x} - \mu, x - \mu)$ rises from $\approx$0 to 0.84 and 0.88. Neither variant copies the noise. Up to $\lambda = 0.5$ both outputs are closer to the clean face than the input is (MSE 0.073 and 0.083 against 0.139 at $\lambda = 0$). Above it the decoder's own reconstruction error bounds them. MCPC's outputs have the lower FID below $\lambda = 0.6$, and PC's from there on.

### E.10.1 Data-size curves on CelebA (extra measurement)

This subsection repeats the data-size measurement of Bonnaire et al. (2025) on CelebA (Fig. 6). It is not a test of §3, and its fits are post hoc. The gap is $(E_{\text{held-out}} - E_{\text{train}})/E_{\text{held-out}}$. Here $E(y) = \frac{1}{2}\|y - \hat{x}(z_{50})\|^2$ is the reconstruction term of $F$ alone at $z_{50}$, the state after 50 noise-free gradient steps on $F$ (step 0.01) from $z = 0$. It is averaged over the run's training faces and over a size-matched held-out pool from the rest of the train split. The copy fraction against the full train

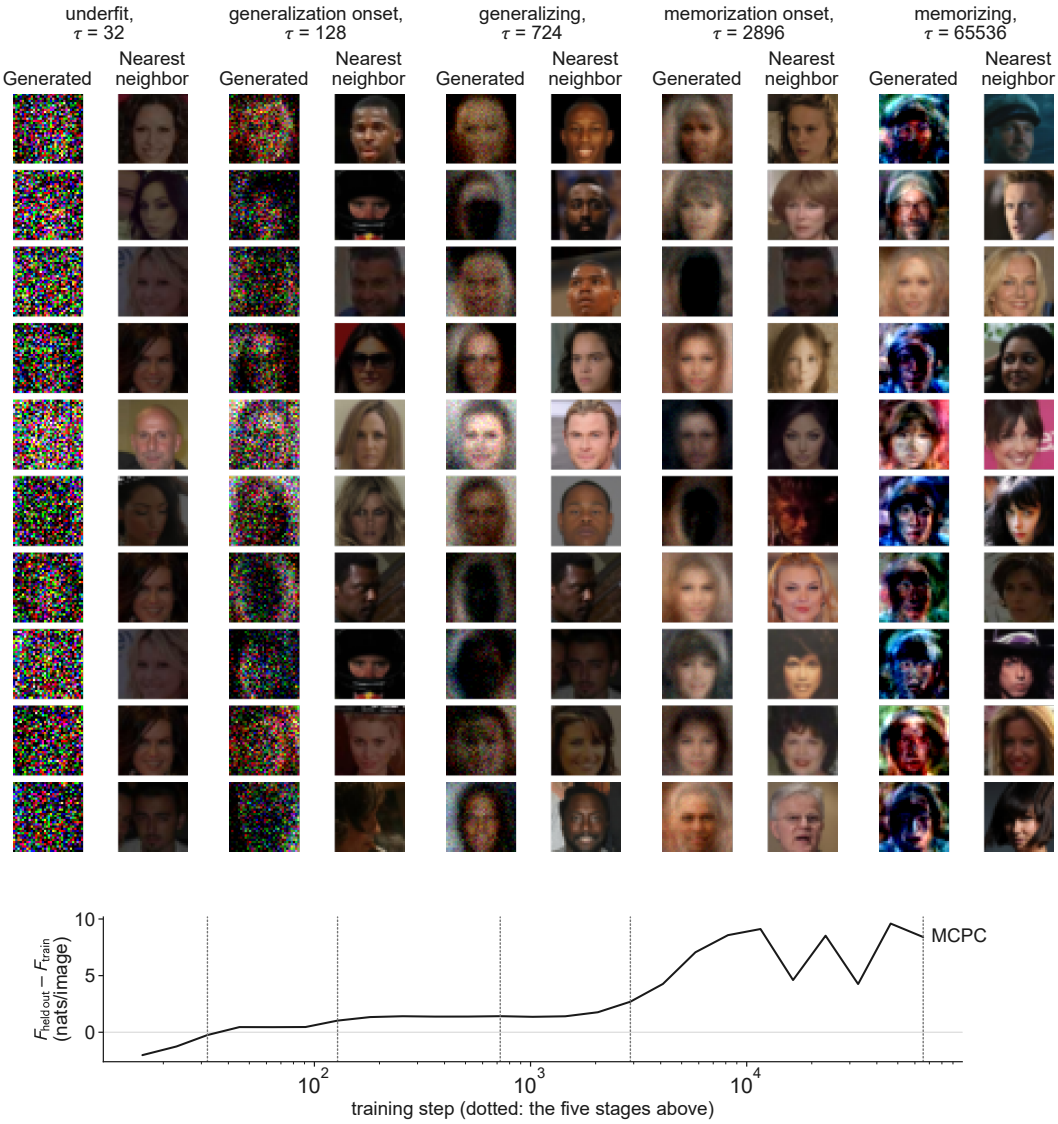



Figure 19: **Over training, samples never become copies** (CelebA, $N = 512$ training faces drawn from the full train split; one seed). Ten unselected prior samples of the sampled E-step, each with its pixel-nearest face of the full train split, at five checkpoints. The decoder is odd, so samples come in $\pm$ pairs, and we show the member with $\mu^\top \hat{x} > 0$. The copy fraction (nearest/second-nearest distance below $1/3$) is zero at every checkpoint for both rules. The bottom panel shows the train/held-out gap $F_{\text{ho}} - F_{\text{tr}}$ in nats/image, with $F = \frac{1}{2}\|y - \hat{x}(z^\star)\|^2 + \frac{1}{2}\|z^\star\|^2$ at the converged MAP latent $z^\star$ (one encoder for both rules, on the run's 512 training faces and 512 held-out faces). This is a likelihood proxy that tracks the AIS gap (Pearson 0.994, App. E.10.2). It is drawn for the sampled E-step (MCPC) at every other saved checkpoint (solid), and dotted lines mark the five stages above.

split stays at zero and is not drawn. We run the protocol above at $N = 512$ to 32768 faces drawn from the full train split, for $2^{17}$ steps (sampled variant with three seeds for $N \leq 2048$ and one above, MAP variant with one seed). The copy fraction is 0 against the full split at every checkpoint, size and variant (at most 0.02% against a run's own training set). Memorization is therefore read on the MAP-scored gap. Under MCPC its half-plateau onset grows as $N^{1.35}$ for $N \leq 4096$ (2896 to 46341 steps), whereas Bonnaire et al. (2025) find $\tau_{\text{mem}} \propto N$. The curves roughly collapse on $\tau/N$ ($\tau_{\text{mem}}/N$ from 5.7 to 11.3), and there is no onset at $N = 32768$. The gap's plateau shrinks from 28% to 1% as $N$ grows. $\tau_{\text{gen}}$ (the step at which the sample FD settles) stays at 362–512 steps, roughly independent of $N$. With the one-bit sign flip the final sample FD is 168–467 over $N$, against

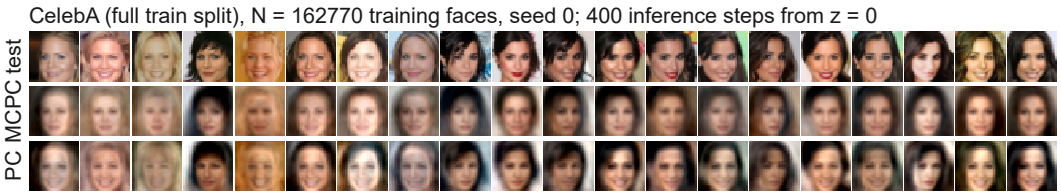



Figure 20: **Reconstruction** (CelebA, the decoders of Fig. 7, $N = 162770$ training faces, one seed, final weights). Held-out test images (top) and their MCPC and PC reconstructions, with 400 inference steps from $z = 0$ for both. MCPC shows the last state of the Langevin chain and PC the result of MAP descent.

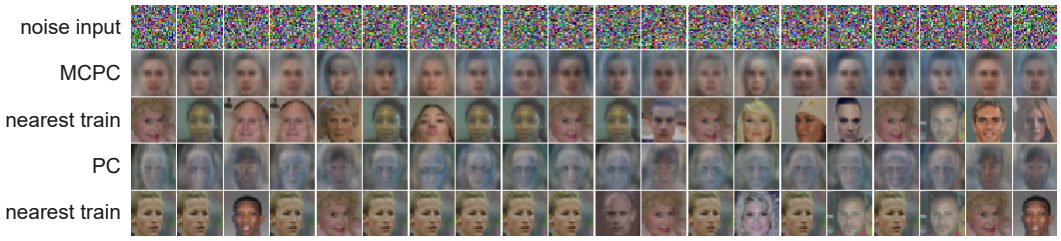



Figure 21: **Inference from a noise input, not generation** (CelebA, decoders trained on the full train split, final weights). Twenty inputs $y \sim \mathcal{N}(\bar{x}\mathbf{1}, \sigma^2 I)$, where $\bar{x}$ and $\sigma$ are the mean and standard deviation of the training pixels. Each input is inferred from $z_0 \sim \mathcal{N}(0, I)$ by the variant's own rule (MCPC uses one late state of the training-time Langevin chain, PC uses 400 MAP steps) and decoded. Under each output is its pixel-nearest face of the full train split.

711–907 for the raw samples $\hat{x}(z)$. Under PC the FD curve is not monotone, so no settle time is reported, and the onsets are not monotone in $N$ (1448, 23170, 4096, 512 steps, in increasing $N$).

### E.10.2 Likelihood gap (post hoc)

This subsection measures the train/held-out gap in true likelihood, which the reconstruction gap above does not track. We estimate $-\log p(y)$, with $p(y) = \int \mathcal{N}(y; W_2 \tanh(W_1 z), I)\, \mathcal{N}(z; 0, I)\, dz$, by annealed importance sampling from the Laplace Gaussian at the converged MAP. Quadrupling the temperatures or the chains moves the gap by at most 0.01 nats. We use up to 1000 training images and as many held-out images (`M22_model_data/nll_ais`; Table 13). The gap falls with $n$ for both rules and vanishes at full data. The relative reconstruction gap used above does not track it (Pearson 0.59 over 14 CelebA decoders), so its orderings between the rules are withdrawn (Table 8). The absolute gap of $F$ at the converged MAP does track it (0.994).

Table 13: **Held-out minus train negative log-likelihood**, nats/image, final weights (AIS; standard error at most 1.1 on CelebA). The CelebA columns give [a] the $N = 512$ decoders above ($2^{16}$ steps), then the data-size runs ($2^{17}$ steps) and the full train split. The MNIST columns use $(h, m) = (256, 64)$, mean of two seeds.

| | CelebA, $N$ | | | | | | | MNIST, $N$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| rule | 512[a] | 512 | 1024 | 2048 | 4096 | 32768 | full | 256 | 1024 | 4096 |
| MCPC | 8.15 | 8.72 | 4.39 | 2.58 | 1.00 | −0.35 | −0.18 | 5.85 | 1.08 | 0.58 |
| PC (MAP) | 5.43 | 8.23 | 4.00 | 1.24 | 0.03 | −0.02 | −0.37 | 3.53 | 0.62 | 0.26 |
| MAP, 82 steps | | | | | | | | 4.02 | 1.26 | 0.26 |

### E.11 The sign of the mean-face axis

This section explains why each decoded face has a sign, and how we choose which member of a $\pm$ pair to show. The analysis is post hoc, on the full-data CelebA decoders (`celeba_full/axis_compute.py`). $\mu$ is the mean training face, and $v_1$ is the top eigenvector of the uncentred data second moment $\frac{1}{N}\sum_a y_a y_a^\top$. We write $f(z) = \mu^\top \hat{x}(z)$ with

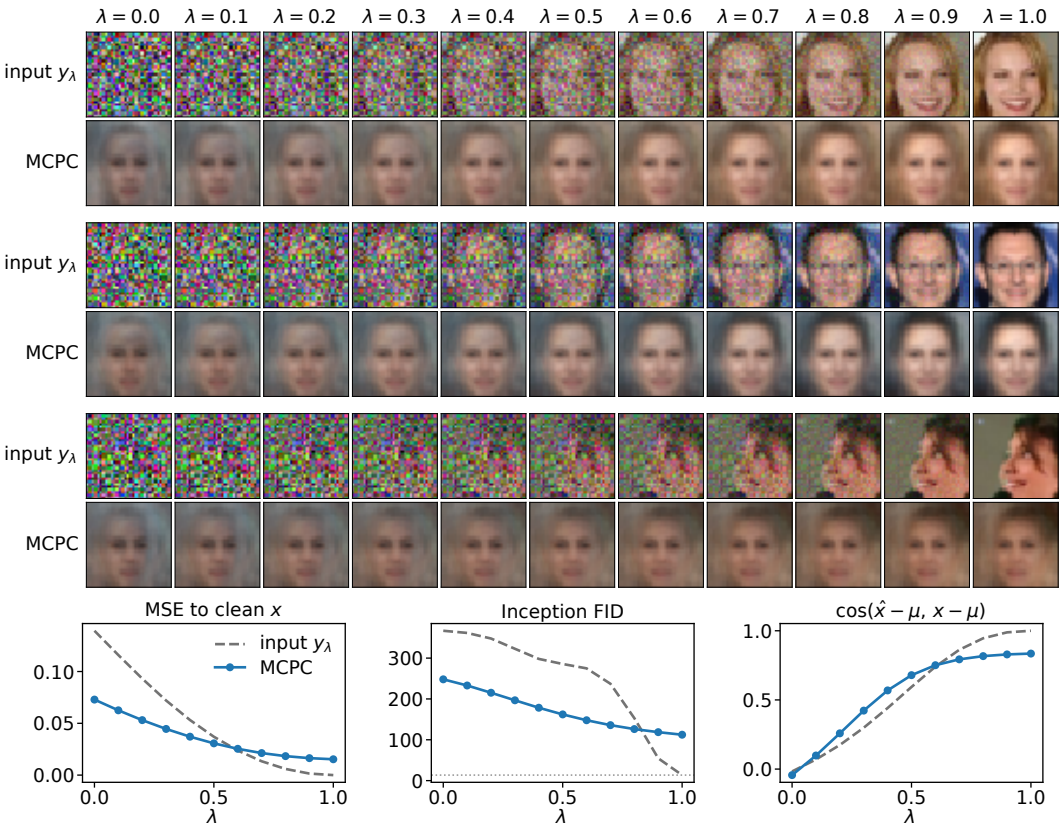



Figure 22: **Inference from noise→data inputs, MCPC** (post hoc; interpolation in input space, not in $z$, and not generation; CelebA, full-data decoders, final weights). The top rows take three held-out faces. For each we show the input $y_\lambda = \lambda x + (1-\lambda)\varepsilon$ (one noise draw per face), then the MCPC output $\hat{x}(z)$ of the E-step on $y_\lambda$ (400 steps from $z = 0$, the time average of the last 200 Langevin states), shown as its positive member. The bottom panels average over 1 000 faces. They show the mean-pixel MSE to the clean face, the Inception FID against 9 000 other test faces (dotted, the clean faces), and $\cos(\hat{x} - \mu, x - \mu)$. Dashed lines give the input itself.

$\hat{x}(z) = W_2 \tanh(W_1 z)$. The latents $z$ of 2000 training faces come from 400 inference steps from $z = 0$ (the chain's last state for MCPC, MAP descent for PC). $\bar{z}$ is their mean, $u_1$ is the top eigenvector of their second moment (signed by $\bar{z}$), and $a = W_1^\top W_2^\top \mu = \nabla f(0)$. The sign rules classify $\operatorname{sign} f$ on 5000 prior draws. In the data $\cos(v_1, \mu) = 0.9995$, and $v_1$ carries 82% of the uncentred trace (29% centred). Over the training latents $\cos(\bar{z}, u_1)$ is 0.996 (MCPC) and 0.997 (PC), with $f > 0$ for a fraction 1.000 and 1.000. On prior draws $\operatorname{sign}(a^\top z)$ gives $\operatorname{sign} f$ with accuracy 0.990 and 0.976, and $\operatorname{sign}(u_1^\top z)$ with 0.857 and 0.665. The rule $\operatorname{sign}(a^\top z)$ errs under MCPC only where $|\hat{a}^\top z| \le 0.041$ ($\hat{a} = a/\|a\|$). As a control, note that in the full-data MCPC decoder most latent directions also decode to $\pm\mu$. Over 1000 random unit directions $d$ at the training latents' median radius along $u_1$, the median $|\cos(\hat{x}(\varrho d), \mu)|$ is 0.94 (68% above 0.9, and for the $N = 512$ decoder 0.65 and 3%). What singles out $u_1$ is that the training latents agree in sign along it, not the image it decodes to ($\cos(\hat{x}(\varrho u_1), \mu) = 0.998$). Flipping the whole latent of a face turns it into its negative, and $f$ changes sign at $\lambda = 0.5$ (Fig. 8). Linear paths between the latents of two held-out test faces stay positive (`celeba_full/interp_ddpm.py`), as do those between the positive members of two prior draws (Fig. 8). A path from a positive to a negative prior draw changes sign once (`celeba_full/axis_figs.py`).

For display, every decoded image is shown as its positive member $\operatorname{clip}(s\hat{x}) = \operatorname{clip}(\hat{x}(sz))$, $s = \operatorname{sign} f$ (Figs. 7, 19, 8). Nearest neighbours and copies are computed on the same image. Reconstructions need no such rule, since 0 of the 24 per decoder have $f < 0$, in each of the four decoders (MCPC and PC, full data and $N = 512$; Fig. 23).

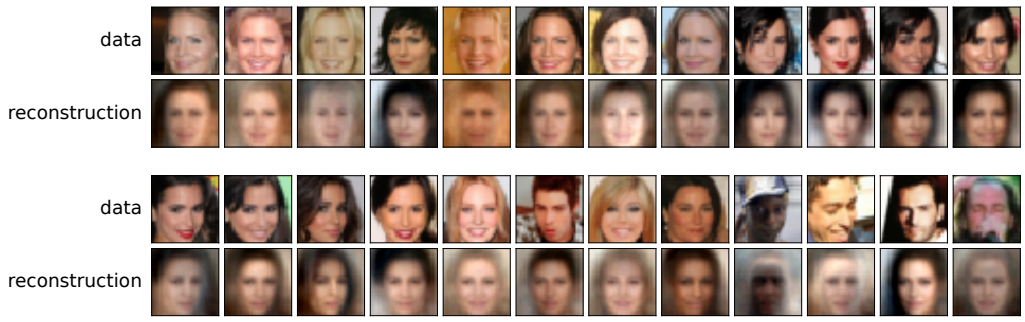



Figure 23: **Reconstructions** (MCPC, full-data decoder; 16 held-out test and 8 training faces, two blocks; 400 inference steps from $z = 0$, the chain's last state). In each block the data sit above their reconstructions. Every reconstruction has $f > 0$.

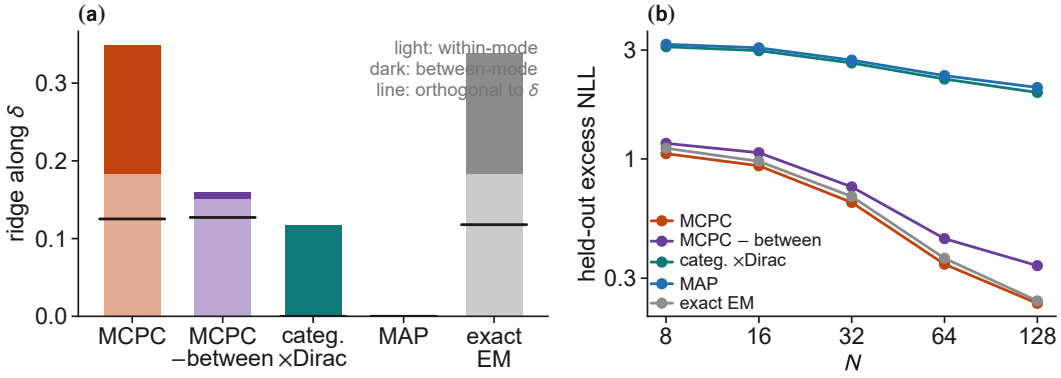



Figure 24: **The ridge beyond a Gaussian posterior** (bimodal latent prior; median over settings). (a) Effective ridge along the inter-mode direction $\delta$ per E-step. Light bars give the exact within-mode ridge at the variant's own $W$, dark bars the remainder (between-mode), and the line the ridge orthogonal to $\delta$. (b) Held-out excess NLL against $N$.

### E.12 Mixture latent

Here we test whether the ridge still acts when the posterior is not Gaussian but bimodal. The identity $\mathbb{E}[zz^\top] = m_y m_y^\top + C$ needs only a finite second moment. For a mixture posterior, the law of total covariance splits $C$ into a within-mode part and a between-mode part $\pi(1-\pi)\delta\delta^\top$. The between-mode part is a rank-one ridge along the inter-mode direction $\delta$, and a MAP E-step from a fixed start sets it to zero. We test this on a latent $z \in \mathbb{R}^4$ with the fixed prior $\frac{1}{2}\mathcal{N}(\pm a e_1, s^2 I)$ and a linear decoder whose true $e_1$ column is weak, so most posteriors are bimodal. We compare six E-steps. They are MCPC, MCPC with the between-mode part projected out of every draw, a categorical×Dirac node in the spirit of Pinchetti et al. (2022), MAP from $z = 0$, MAP from a fresh prior draw, and exact EM. The grid is $N = 8$–$128$, with 90 settings, pre-registered (Fig. 24). The MCPC chain hops between modes (at least 169 crossings per datum in every setting). Its measured split matches that of the exact posterior (to 1.8% between, 5.1% within), and the between part is rank one (0.999 of its trace along $\delta$). Along $\delta$ the between part roughly doubles the ridge (total over within $= 1.98$, settings 1.43–2.57). Orthogonal to $\delta$ the ratio is 1.04, which is the integrator's bias (post hoc). For MCPC, dropping the between part moves the learned $W$ away from its ridge form by a factor of 7 (at least 2.7). Projecting the between part out of every draw leaves the within-mode ridge alone (1.04 of the analytic within-mode ridge; post hoc). It removes the extra shrinkage along $\delta$ ($\|W\delta\|$ ratio 0.78, with the orthogonal block within 1.7%), and it costs held-out likelihood at every $N$ (0.34 against 0.23 nats at $N = 128$). MCPC tracks exact EM (0.24). A MAP E-step has no ridge orthogonal to $\delta$ either. Its $\|W\|_F$ grows to 11.7 (true 3.5), with held-out excess 2.06 nats. The categorical×Dirac node keeps only the between part and behaves the same ($\|W\|_F = 11.7$). MAP from a random start is not ridge-free. Its re-drawn basins give a between ridge 5.90 times that of MCPC, which collapses the inter-mode column ($\|W\delta\| = 0.009$; post hoc). The registered

predictions are P1a–c (hops, split, rank one), P1d (between share), P2a (stationarity identity), P2b (between part needed), P3a,b (anisotropy), P4 (MAP likelihood), P5 (projection) and P6 (random-start MAP). Their verdicts, in this order, are HELD, HELD, HELD, MISSED, MISSED, MISSED, HELD, HELD, HELD, MISSED, HELD. There are three misses. First, the between share along $\delta$ is 0.48 against 0.75 at the true $W$ (P1d), because the learned $\|W\delta\|$ is about twice the truth's, exact EM included. Second, the stationarity identity and the between-part ratio fail on the categorical×Dirac variant (P2a, relative distance up to $1.2 \times 10^{-1}$; P2b, ratio 1.0). That variant has no ridge orthogonal to $\delta$ and so no fixed point (a case the predictions did not exclude). On MCPC both hold. Third, the registered measure of the projection (P5, 21.9% of MCPC's between ridge left) is confounded, since splitting the projected, unimodal cloud by responsibility manufactures a between term. The direct check of P5 above is post hoc.

### E.13 Correspondence with diffusion models

This section maps each object of a diffusion model to its counterpart in our setting, and checks one of them, the learned gain, directly. Table 14 pairs each diffusion object with its counterpart here.

Gain check (post hoc). This check runs on the solvable tier, using the criterion-flip protocol of

Table 14: **Diffusion objects and their counterparts here** (§5). "Same form" means the same Gaussian formula, in latent rather than data space. "Analogous" means the same role played by a different object. "None" means there is no counterpart.

| Diffusion | This paper | Status |
|---|---|---|
| MMSE denoiser $\mathbb{E}[x \mid u, t]$; Gaussian case $\Sigma(\Sigma + \sigma_t^2 I)^{-1}u$ | posterior mean $m_y = \sigma_H^2 W^\top(\sigma_H^2 WW^\top + \sigma_O^2 I)^{-1}y$, in latent space, at the one level $\sigma_O^2$ | same form |
| denoiser error covariance $\mathrm{Cov}[x \mid u, t]$ | exact EM's ridge $H^{-1}$; MCPC's $H^{-1}(I - \frac{\bar\eta}{2}H)^{-1}\Phi(H)$, biased by $\bar\eta$, $b$, $S$ (Thm. 3.2) | same form (EM) |
| denoising training, where noisy inputs act as a ridge (Bishop, 1995) | the M-step inverts $\widehat{\Sigma}_{mm} + C$ (Prop. 3.1) | analogous |
| noise schedule $\sigma_t$, annealed over $t \in [0, 1]$ | one fixed level; fixed step $\bar\eta$, burn-in $b$, window $S$ | analogous |
| time input $t$ to the network | – | none |
| noise-level posterior $p(t \mid u)$ concentrates (Sahraee-Ardakan et al., 2026, §5) | latent posterior (a different one) must stay spread, $C \neq 0$; MAP ($C = 0$) memorizes | none |
| reverse-time sampler with a schedule | decoder push-forward $\hat{x}(z)$, $z \sim \mathcal{N}(0, I)$, as in any latent-variable model | none |
| denoiser gain $c/(c + \sigma_t^2) > 0$ in every direction | learned variance 0 for $c_j < c_{\mathrm{thr}}$ (Thm. 3.3; Fig. 3c); in probabilistic PCA, for $c_j < \sigma_O^2$ (Tipping and Bishop, 1999) | none |
| $\tau_{\mathrm{gen}}$ flat in $n$, $\tau_{\mathrm{mem}} \propto n$ (Bonnaire et al., 2025) | both onsets from one flow, Eq. (6); the onset gap grows with $N$ in sign (§4.2) | analogous |
| no memorization at infinite time above a model-dependent $n^\star(p)$ | $N \geq N_\infty^\star$ (Eq. 9), set by the ridge, not the parameter count; never for exact EM | analogous |
| a frozen random-feature score network at one noise level (Bonnaire et al., 2025) | ridge on a learnable map (M11, App. E.9) | analogous |

App. E.3 with the signal on. We use $N = 256$ and $n = m = 128$ (a latent slot for every direction), a small initialization and 8000 updates. The three variants see the same data, and MCPC runs at the fixed step $\bar\eta = 0.8$, so $c_{\mathrm{thr}} = 1.67\,\sigma_O^2$ (Fig. 3c; the gain is averaged over the last fifth of the updates, seeds pooled). MCPC's gain is at most 0.008 in all 300 directions below $c_{\mathrm{thr}}$. Above $1.2\,c_{\mathrm{thr}}$ it lies within 0.0002 (median) of the fixed point of Eq. (6). Exact EM cuts at $\sigma_O^2$. Of the 84 directions between $\sigma_O^2$ and $c_{\mathrm{thr}}$ it holds 76 (gain $> 0.05$), and the rest, just above $\sigma_O^2$, are still growing. MAP has no cutoff. Its gain is 10–41 in every direction and still growing, since $C = 0$ has no fixed point (Thm. 3.3). Exact EM's gain $(1 - \sigma_O^2/c)_+$ is the Wiener gain written in the observed variance and clipped at zero (probabilistic PCA, Tipping and Bishop, 1999). The sampler moves the cut from

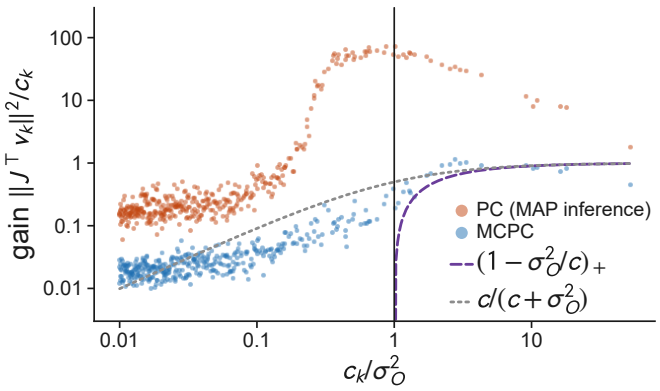



Figure 25: **Learned gain per data direction, full CelebA** (post hoc; Fig. 3c shows the solvable tier). Tanh decoders of §4.3, MCPC and PC (MAP inference). We plot the Jacobian gain $\|J^\top v_k\|^2/c_k$, with $J = W_2\,\mathrm{diag}(1-\tanh^2(W_1 z))\,W_1$ at the posterior-mean latent, averaged over 2000 training faces, against the data eigenvalue $c_k$. The dashed line is $(1-\sigma_O^2/c)_+$, and the dotted line is the Wiener gain $c/(c+\sigma_O^2)$.

$\sigma_O^2$ to $c_{\mathrm{thr}}$. On full CelebA (Fig. 25) no $c_{\mathrm{thr}}$ is derived. There the median gain over the 79 data directions with $0.1 \le c_k < \sigma_O^2$ is 0.06 (MCPC) and 1.9 (PC). Above $2\sigma_O^2$ it is 0.91 and 18. This agrees in sign only. The gain also explains blur. A direction below $c_{\mathrm{thr}}$ carries no learned variance, so $\hat{x}(z)$ lacks it. Probabilistic PCA already drops the directions below $\sigma_O^2$, and the sampler adds the band $(\sigma_O^2, c_{\mathrm{thr}})$. For images, whose variance falls with spatial frequency, that removes fine detail. The encoder–decoder test of App. E.10 places the blur in the decoder, not the encoder (post hoc). That the missing detail is exactly this band is untested.

### E.13.1 Generation from the prior (post hoc)

This subsection scores samples drawn from the prior and decoded once, to check that they are faces and not copies. Under $\mathcal{N}(0, I)$ the FID is 121 for MCPC and 2.8 for real faces. Fig. 7 shows

Table 15: **Generation from the prior** (post hoc; §4.3; full-data CelebA decoder, one seed, final weights; 10 000 draws, one decoder pass each). FID uses Inception features from torchvision's ImageNet InceptionV3, not the canonical FID-Inception weights, against 10 000 test faces. A redraw of $z$ moves it by at most 0.7. Pixel FD is the Fréchet distance on pixels. Flipped is the share of draws shown as $\hat{x}(-z)$. Copies count draws with $d_1/d_2 < 1/3$ against the full train split. Ratio is the median distance to the nearest of 10 000 training faces over that to the nearest test face.

| set | prior of $z$ | FID | pixel FD | flipped (%) | copies | ratio |
|---|---|---|---|---|---|---|
| MCPC | $\mathcal{N}(0, I)$ | 121 | 175 | 49.2 | 0 | 1.019 |
| real valid faces | — | 2.8 | 1.6 | — | — | 1.001 |

the first 10 MCPC draws, each over its pixel-nearest face of the full train split (162770 faces), clipped to $[0, 1]$. To the right of the block, draws 1 and 2 (positive members) are the endpoints A and B of a linear path in $z$, with one decoder pass per tile. Fig. 8 gives the full interpolation figure, both pairs and a schematic. The $\pm$ display flips 49.2% of the draws. The protocol is in `celeba_full/prior_gen/prior_gen.py`, and no model is trained. Generation is one decoder pass $\hat{x}(z)$, and no inference runs on a sample. Every sample is scored as it is shown, as its positive member clipped to $[0, 1]$. In Table 15 the redraw bound covers sampling noise in $z$ only. MCPC's ratio, 1.019, lies close to the real-face null 1.001 (no interval computed).

### E.13.2 Latent geometry (post hoc)

This subsection asks where the latents of each coder sit, compared with the prior and with each other. We repeat the geometric readings of Sahraee-Ardakan et al. (2026) in the latent space of these coders (Fig. 26). PC means MAP inference, and MCPC means the time average of the Langevin chain. (i) Norm shell (Fig. 26a; their §5.2 and App. C, where the noise level is read off the

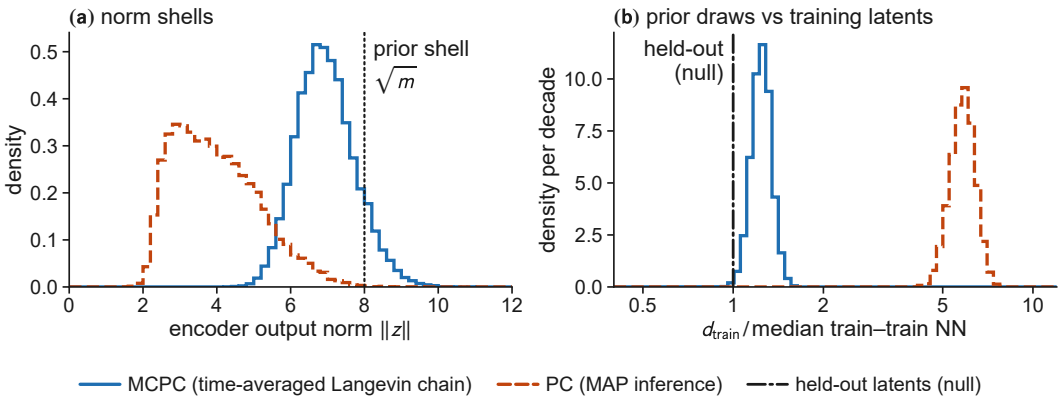



Figure 26: **Latent geometry of the CelebA tanh coders, MCPC and PC** (post hoc; the decoders of §4.3, one seed, final weights, full training split). The MCPC encoder output is the time average of the last 200 of 400 Langevin states, and the PC output is the MAP latent. (a) Density of $\|z\|$ over the training encoder outputs (held-out medians 6.9 and 3.9). The dotted line is the prior shell $\sqrt{m}$. (b) Density, per decade, of the distance from each of 5000 prior draws (positive member) to the nearest training latent, divided by the median train–train nearest-neighbour distance. The dash-dotted line is the held-out null at 1 (2000 held-out latents, median 0.99–1.00).

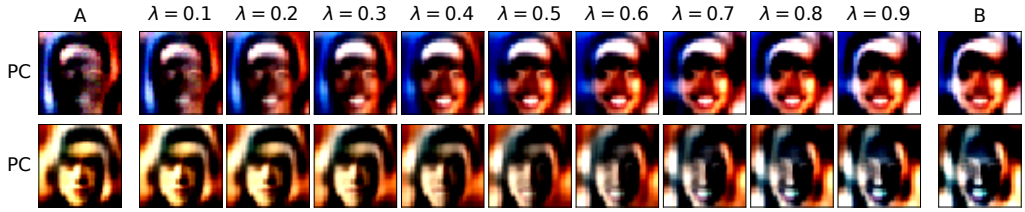



Figure 27: **PC decoder, the same prior draws and interpolation** as Fig. 8 (full-data decoder; endpoints $sz$, $s = \operatorname{sign} \mu^\top \hat{x}(z)$; $z_\lambda = (1-\lambda)z_A + \lambda z_B$, one decoder pass). Samples are off-range because PC's latents do not match the $\mathcal{N}(0, I)$ prior (prior mismatch; Fig. 26).

norm, $\hat{\sigma} = r/\sqrt{D-d}$). Prior draws and MCPC's chain states share the shell $\sqrt{m} = 8.0$ of the prior $\mathcal{N}(0, I)$ (the states have rms 8.7). Posterior means sit inside (MCPC 4.7, its encoder average 6.9, PC's MAP 3.8), and held-out latents behave like training ones. (ii) Spread (not drawn). This is the opposite of their concentrating $p(t \mid u)$ (§5.2, Prop. 1). MCPC's $\operatorname{tr} C_y/m = 0.82$ varies across faces by 1.3% beyond Monte Carlo noise, so it is nearly independent of the datum, as Prop. 3.1 assumes. On the solvable tier the realized $C$ matches Eq. (5) at the learned $H$ to 0.2%, with $C_\perp = 1.67$ below the bar 2.91 of Eq. (7). MAP has $C = 0$ by construction. (iii) Aggregate posterior against the prior (not drawn; their marginal over a hidden variable, Eq. 8). A likelihood maximum in $W_1$ has $\mathbb{E}_y\mathbb{E}[zz^\top \mid y] = I$. MCPC's long-chain means give $\operatorname{tr}\mathbb{E}[m_y m_y^\top]/m = 0.34$, and 1.16 with $\bar{C}$ (Gaussian KL to the prior 79 and 5.2 nats). PC's MAP latents give 0.27 (KL 80). Both fill one half-space (0.998 of training latents; App. E.11). (iv) Proximity (Fig. 26b; their §5.3). Prior draws lie a median 1.24 (MCPC, or 1.21 at $N = 512$) and 5.8 (PC, or 3.6) train–train spacings from the nearest training latent. Held-out latents lie 0.99–1.00 spacings away, and no draw has $d_1/d_2 < 1/3$. So neither rule's samples sit on training latents, and PC's fall in the gap between the prior shell and its shrunken aggregate.