RETRIEVAL METASTABILITY AND CENTROID HOPPING IN CONTINUOUS LOG-SUM-RELU DENSE ASSOCIATIVE MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the retrieval state of Dense Associative Memory with the log-sum-ReLU (LSR) energy on the N-sphere as a function of the load α and temperature T. The compact support of the LSR kernel partitions the retrieval region into three regimes. At low α, even the nearest spurious pattern is separated from the target by a free-energy barrier above NT, and retrieval is stable on any accessible timescale. Above a threshold $\alpha_{\rm th}$, the state thermally wandering near the retrieval minimum already fills a spurious pattern basin; the two basins merge into a centroid and retrieval is lost. Between these limits lies the metastable regime: spurious pattern basins exist nearby but are separated from the retrieval basin by an entropic barrier. To escape, the state must drift beyond its typical thermal range and reach a spurious basin, a rare event whose rate we identify as a multi-channel Kramers process. After escape, the system hops between two-pattern centroids. Monte Carlo simulations validate this mechanism. The disorder-averaged escape rate, with a single calibrated prefactor, reproduces the measured escape probabilities across the simulated range of load and temperature. At large N the exponential growth of the number of escape channels is exactly compensated by their entropic barriers, so the escape rate falls as $e^{-N(\alpha_{\rm th}-\alpha)}$ at every temperature, and the metastable regime extends to $\alpha_{\rm th}$. Simulations at fixed load and increasing system size show this suppression. The apparent phase boundary observed in finite-N simulations lies inside the metastable regime: it is a kinetic crossover, reached where the escape time equals the simulation time, not a thermodynamic transition. Since the escape time grows exponentially with N, the metastable state becomes indistinguishable from thermodynamic stability at system sizes of practical interest. The kinetic boundary is relevant only for the moderate N accessible to Monte Carlo simulations.

1 INTRODUCTION

Associative memory networks store patterns such that the correct memory can be retrieved from a noisy or partial cue. The classical Hopfield network (Hopfield, 1982) can store 0.138 patterns per neuron (Amit et al., 1987). Dense Associative Memory (DAM) energy models replace pairwise interactions with highly nonlinear energy functions, achieving exponentially large storage capacity: the number of patterns M scales as $e^{\alpha N}$ for N neurons or for continuous states on the N-sphere. The first such model used the log-sum-exp (LSE) energy (Demircigil et al., 2017; Ramsauer et al., 2021).

Hoover et al. (2025) introduced the log-sum-ReLU (LSR) energy, based on the compactly supported Epanechnikov kernel. For continuous neurons on the N-sphere, the LSR energy takes the form

$$E(\boldsymbol{x}) = -\frac{N}{b}\ln\left(\sum_{\mu=1}^{M}\max\Big(0,\ 1-b+b\,\phi_\mu\Big)\right), \tag{1}$$

where $\boldsymbol{x}$ is the system state with $|\boldsymbol{x}|^2 = N$; $b > 1$ is the rescaled inverse variance that controls the kernel sharpness; $\boldsymbol{\xi}^\mu$, $\mu = 1, \ldots, M$, are the stored patterns on $S^{N-1}(\sqrt{N})$; and $\phi_\mu \equiv (\boldsymbol{x}\cdot\boldsymbol{\xi}^\mu)/N$ is the alignment of state $\boldsymbol{x}$ with pattern μ. The compact support introduces a hard wall at $\varphi_c = (b-1)/b$: pattern μ contributes to the energy only when $\phi_\mu > \varphi_c$.

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.

Petrova et al. (2026) investigated the retrieval stability of the LSE and LSR models at finite temperatures, which model the persistent noise to which a physical memory is exposed. Following Ramsauer et al. (2021) in assuming that the inter-pattern alignment $\varphi_{1\mu} = (\boldsymbol{\xi}^1 \cdot \boldsymbol{\xi}^\mu)/N$ between a fixed retrieving pattern $\boldsymbol{\xi}^1$ and a random spurious pattern $\boldsymbol{\xi}^\mu$ is approximately distributed as $\mathcal{N}(0, N^{-1})$ (throughout, ϕ denotes an alignment of the state with a pattern and φ an alignment between two patterns), they showed that for the LSR kernel there exists a support threshold, $\alpha_{\rm th} = \varphi_c^2/2$, below which no spurious pattern falls within the support and retrieval is perfect at any temperature. For $b = 2 + \sqrt{2}$, the value used throughout both their work and the present paper, $\varphi_c \approx 0.707$ and $\alpha_{\rm th} = 0.25$. The choice is not restrictive: b enters the formulas only through φ_c, and the regimes found below persist for any $b > 1$, with boundaries given by the same formulas. The exact distribution

$$f(\varphi) = C_N \, (1 - \varphi^2)^{(N-3)/2} \,, \qquad C_N = \frac{\Gamma(N/2)}{\Gamma\big((N-1)/2\big)\,\sqrt{\pi}} \,, \tag{2}$$

shares the same mean and variance with the Gaussian, but has lighter tails: it vanishes at $|\varphi| = 1$, whereas the Gaussian remains nonzero. When weighted by the exponential factor $e^{\alpha N}$ that counts the stored patterns, this lighter tail shifts the threshold to (in the limit $N \to \infty$)

$$\alpha_{\rm th} = -\frac{1}{2} \ln(1 - \varphi_c^2) \approx 0.347 \,. \tag{3}$$

The Monte Carlo simulations (MCS) reported by Petrova (2026), however, place the LSR retrieval boundary much closer to $\alpha \approx 0.25$ than to 0.347. We repeated these simulations to investigate this apparent discrepancy in more detail. Since the MCS indicate a lack of retrieval within the theoretically predicted retrieving zone, we focused primarily on this region.

The simulation protocol follows Petrova (2026): M patterns, ramped from $20{,}000$ at low load to $500{,}000$ at high load to keep $N = \ln M/\alpha$ from falling too low (the targeted runs of Section 3 use fixed $M = 20{,}000$, except the finite-size test of Appendix B), Metropolis dynamics with step size $\sigma = 2.4T/\sqrt{N}$, and two independent replicas per disorder sample initialized near the target with $\phi_{\rm init} \sim \mathrm{U}(0.75, 1.0)$. We doubled the equilibration and sampling times ($N_{\rm eq} = 2^{15}$, $N_{\rm samp} = 2^{13}$, total $T_{\rm MC} \approx 4 \times 10^4$ steps) and refined the load grid to $\Delta\alpha = 0.005$. The number of independent trials per (α, T) point decreases linearly from 512 at $\alpha = 0.005$ to 64 at $\alpha = 0.55$ ($n_{\rm dis} = 256$ to 32 disorder samples, two replicas each). We recorded per-sample observables rather than disorder averages: the time-averaged target alignment ϕ_A, ϕ_B of each replica, and the time-averaged maximum alignment with any non-target pattern,

$$\phi_{\rm max}^{\rm other} = \frac{1}{N_{\rm samp}} \sum_t \max_{\mu \neq 1} \phi_\mu(\boldsymbol{x}_A(t)) \,. \tag{4}$$

The observable $\phi_{\rm max}^{\rm other}$ distinguishes retrieval ($\phi_{\rm max}^{\rm other} \ll \varphi_c$, no spurious pattern in support) from centroid states ($\phi_{\rm max}^{\rm other} \approx 0.82$, a spurious pattern shares the basin with the target). Throughout the paper, a trial *retrieves* when two conditions hold over the sampling window: the time-averaged target alignment stays at its thermal equilibrium value $\phi_{\rm eq}(T)$, and $\phi_{\rm max}^{\rm other}$ stays below φ_c.

The per-trial alignment distributions show that the retrieval region has more structure than the thermodynamic theory predicts. Both panels of Figure 1 show the empirical CDF of the time-averaged alignment ϕ with the target pattern: the left panel varies α at fixed $T = 0.325$, and the right panel varies T at fixed $\alpha = 0.22$. In all cases, the distributions concentrate below unity, since thermal fluctuations at $T > 0$ reduce the time-averaged alignment, in accord with the Boltzmann equilibrium $\phi_{\rm eq}(T)$ (Petrova et al., 2026). In the left panel, at $\alpha = 0.10$ all trials concentrate near $\phi \approx 0.90$; as α increases toward 0.25, secondary steps appear at $\phi \approx 0.83$ and $\phi \approx 0.48$, indicating that a growing fraction of trials settles into states of lower alignment, until the primary peak has disappeared by $\alpha = 0.35$. The red curve ($\alpha = 0.35$), which lies in the theoretically non-retrieving zone, has a qualitatively different shape. The right panel shows an analogous trend with temperature: as T increases, the primary peak shifts to lower ϕ and the secondary steps grow.

These steps indicate the presence of metastable states at $\phi \approx 0.83$ and 0.48, which are not predicted by the thermodynamic theory of Petrova et al. (2026). This paper proposes a theory of these states, tests it against simulations, and resolves the discrepancy between theory and MCS.

The paper is organized as follows. Section 2 develops the theory of centroid hopping: the two-pattern geometry, the escape path and its entropic barrier, the thermal equilibrium at the centroid,

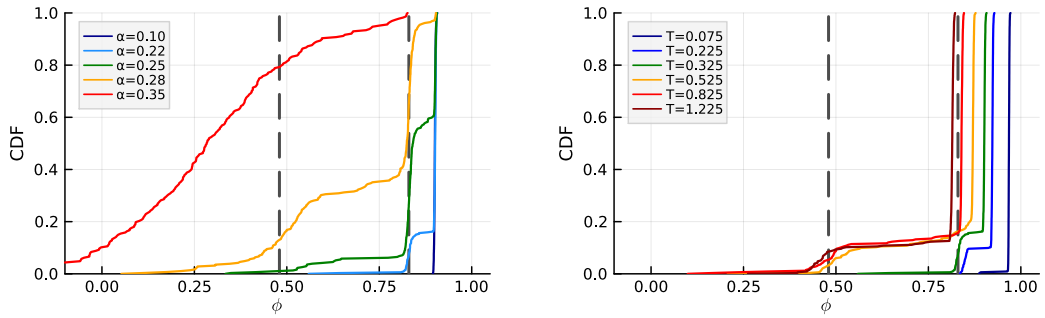

Figure 1: Empirical CDF of the per-trial time-averaged alignment ϕ with the target pattern (both replicas pooled). Dashed gray lines mark $\phi = 0.48$ and 0.83. **Left:** fixed $T = 0.325$, varying α. **Right:** fixed $\alpha = 0.22$, varying T. The primary peak shifts left with increasing T (following $\phi_{\rm eq}(T)$), while the secondary steps at $\phi \approx 0.83$ and 0.48 grow in weight.

and the hopping chain that explains the secondary peaks. Section 3 tests this mechanism with targeted MC simulations using a confirmed-barrier-crossing criterion, calibrates the multi-channel Kramers escape rate from the data, establishes its parameter-free large-N exponent, and constructs the three-zone phase diagram. Section 4 resolves the apparent contradiction between the corrected thermodynamic boundary and the MC observations, and discusses the implications for retrieval at large N. Appendix A derives the entropic barrier from the sphere geometry, and Appendix B develops the compound Poisson disorder average and its connection to the Kramers escape formula used in the main text.

2 Theory of centroid hopping

The Metropolis dynamics on the N-sphere is purely diffusive: the state performs a random walk with no momentum, accepting or rejecting proposals based on the energy change alone. The appropriate framework for escape from a metastable basin in such overdamped dynamics is the Kramers theory of thermally activated transitions (Kramers, 1940; Hänggi et al., 1990).

In the Kramers framework, the escape rate from a metastable basin scales as $\lambda \sim \exp(-\Delta F/T)$, where ΔF is the free-energy barrier along the escape coordinate (Hänggi et al., 1990). This barrier can be energetic, entropic, or both. In our system, the energy along the path at fixed target alignment is exactly constant ($\Delta E = 0$, as we show below), so the barrier is *entropic* in origin: the free-energy cost $\Delta F > 0$ arises because the state must align its perpendicular component with a specific direction on the sphere, reducing the available phase-space volume. At the saddle of the free energy the state trades a small amount of energy for volume, which lowers the barrier of distant patterns (Appendix A). The system must drift beyond its typical thermal fluctuation range to reach a region where a second pattern enters support, and the probability of this excursion is exponentially suppressed by the volume ratio (Appendix A).

We now analyse the geometry of this escape.

Consider the target pattern $\boldsymbol{\xi}^1$ and a spurious pattern $\boldsymbol{\xi}^\mu$ with inter-pattern alignment $\varphi_{1\mu} > 0$. We decompose the state in the orthonormal basis

$$\mathbf{e}_1 = \frac{\boldsymbol{\xi}^1}{\sqrt{N}}\,, \qquad \mathbf{e}_2 = \frac{\boldsymbol{\xi}^\mu/\sqrt{N} - \varphi_{1\mu}\,\mathbf{e}_1}{\sqrt{1-\varphi_{1\mu}^2}}\,, \tag{5}$$

writing $\boldsymbol{x}/\sqrt{N} = a\,\mathbf{e}_1 + v\,\mathbf{e}_2 + \mathbf{c}_\perp$, where $\mathbf{c}_\perp$ collects the remaining $N-2$ components perpendicular to both patterns, and $a^2 + v^2 + |\mathbf{c}_\perp|^2 = 1$. The alignments with the two patterns are

$$\phi_1 = a\,, \qquad \phi_\mu = a\,\varphi_{1\mu} + v\sqrt{1-\varphi_{1\mu}^2}\,. \tag{6}$$

At the retrieval equilibrium, $a = \phi_{\rm eq}(T)$ and $v = 0$ (star in Fig. 2); only pattern $\boldsymbol{\xi}^1$ is in support. The system can increase v (diffusing along $\mathbf{e}_2$) while keeping $\phi_1 = a$ fixed. Since only $\boldsymbol{\xi}^1$ is in support

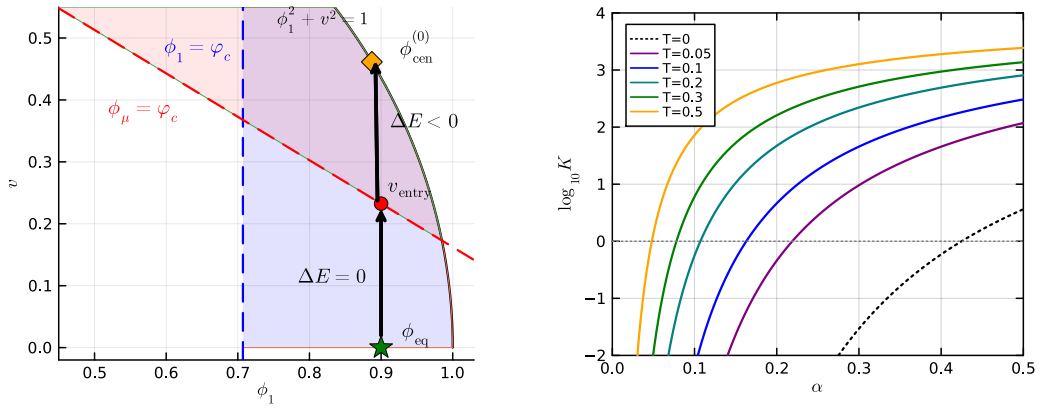

Figure 2: **Left:** the escape path in the (ϕ_1, v) plane. Blue region: $\boldsymbol{\xi}^1$ in support. Red region: $\boldsymbol{\xi}^\mu$ in support. Star: retrieval state $(\phi_{\rm eq}, 0)$. Red dot: entry point $v_{\rm entry}$, where $\boldsymbol{\xi}^\mu$ enters support. Diamond: centroid. Stage 1: entropic climb to the entry line, drawn at fixed ϕ_1 ($\Delta E = 0$); the saddle of the free energy lies on the same line at a nearby $\phi_1 = a^*$, the minimizer in (9). Stage 2: slide to the centroid ($\Delta E < 0$). **Right:** expected number of escape-enabling patterns K versus α for $M = 20{,}000$ at several temperatures. Dotted gray: $K = 1$.

and ϕ_1 does not change, the energy is exactly constant along this diffusion (Stage 1, $\Delta E = 0$). When v reaches

$$v_{\rm entry} = \frac{\varphi_c - \phi_{\rm eq}\,\varphi_{1\mu}}{\sqrt{1-\varphi_{1\mu}^2}}\,, \tag{7}$$

the spurious pattern enters support ($\phi_\mu = \varphi_c$). From this point the energy decreases and the system slides irreversibly to the centroid (Stage 2, $\Delta E < 0$). The condition $v \geq v_{\rm entry}$ is equivalent to $\phi_\mu \geq \varphi_c$, so the barrier crossing can be detected in MC simulations by monitoring all pattern alignments. More generally the support of $\boldsymbol{\xi}^\mu$ is entered on the line $a\,\varphi_{1\mu} + v\sqrt{1-\varphi_{1\mu}^2} = \varphi_c$ of the (a, v) plane, of which $(\phi_{\rm eq}, v_{\rm entry})$ is one point.

On the slice $a = \phi_{\rm eq}$ the entry point exists only if $\phi_{\rm eq}^2 + v_{\rm entry}^2 \leq 1$ (the state must remain on the sphere), that is for inter-pattern alignments above

$$\varphi_{\rm min}(\phi_{\rm eq}) = \varphi_c\,\phi_{\rm eq} - \sqrt{(1-\varphi_c^2)(1-\phi_{\rm eq}^2)}\,. \tag{8}$$

Patterns below $\varphi_{\rm min}$ are not unreachable: the state can lower a below $\phi_{\rm eq}$ and reach them at a finite cost, but that cost is a fixed fraction of N far above the barriers of the patterns considered here (Appendix A), so they contribute nothing to the escape rate at any accessible N. We therefore call the patterns above $\varphi_{\rm min}$ *escape-enabling* and write $K = M\,\Pr(\varphi_{1\mu} > \varphi_{\rm min})$ for their expected number, computed from the exact density (2); Fig. 2 (right) shows it for $M = 20{,}000$. K grows rapidly with α and depends on T through $\phi_{\rm eq}$: higher temperature lowers $\phi_{\rm eq}$, which lowers $\varphi_{\rm min}$ and admits more escape channels. At $T = 0$, $\phi_{\rm eq} = 1$ and $\varphi_{\rm min} = \varphi_c$, and lowering a has infinite cost, so the $K = 1$ crossing is then a true boundary of retrieval, which in the $N \to \infty$ limit occurs at $\alpha_{\rm th}$. At finite $N = \ln M/\alpha$, finite-size corrections shift all curves to the right relative to this limit (Fig. 2, right), so the $K = 1$ boundary depends on both T and N.

Since the energy is constant along the path at fixed a, no energy barrier opposes the escape there. The barrier is entropic: as v increases from 0 to $v_{\rm entry}$, the sphere constraint $|\mathbf{c}_\perp|^2 = 1 - a^2 - v^2$ forces the perpendicular volume to shrink, reducing the entropy $S_\perp \propto \frac{N-3}{2}\ln(R^2 - v^2)$ of the perpendicular subspace, $R^2 = 1 - \phi_{\rm eq}^2$. On the slice $a = \phi_{\rm eq}$ the free-energy cost is purely $-T\Delta S_\perp = T\,\frac{N-3}{2}\big[-\ln(1 - v_{\rm entry}^2/R^2)\big]$. But a is a fluctuating coordinate with a free energy of its own: with $\mathbf{c}_\perp$ integrated out, a state with only $\boldsymbol{\xi}^1$ in support has $\mathcal{F}(a, v) = u(a) - \frac{T}{2}\ln(1 - a^2 - v^2)$ per site, $u(a) = -b^{-1}\ln(1 - b + ba)$, the functional whose minimum at $v = 0$ defines $\phi_{\rm eq}(T)$ (Petrova et al., 2026). On the entry line the state may shift a away from $\phi_{\rm eq}$, trading energy against volume,

and the escape rate is set by the lowest point of $\mathcal{F}$ on that line (Appendix A):

$$\frac{\Delta F}{T} = \min_{a} \left[N \, \frac{u(a) - u(\phi_{\rm eq})}{T} + \frac{N-3}{2} \left(- \ln \frac{1 - \varphi_{1\mu}^2 - \varphi_c^2 - a^2 + 2a\varphi_{1\mu}\varphi_c}{(1-\varphi_{1\mu}^2)(1-\phi_{\rm eq}^2)} \right) \right], \tag{9}$$

where the minimizer a^* solves a quadratic (Appendix A); at $a = \phi_{\rm eq}$ the logarithm reduces to $-\ln(1 - v_{\rm entry}^2/R^2)$, the slice value, which is therefore an upper bound on (9). Through $\varphi_{1\mu}$ the barrier depends on the inter-pattern alignment: more distant patterns require a larger excursion and face a higher barrier, and for any fixed $\varphi_{1\mu}$ it grows linearly with N, making escape exponentially unlikely at large system sizes. For the nearest patterns the saddle lies close to the slice value; the difference matters only for distant patterns, where it turns the closed channels of the slice, $\varphi_{1\mu} < \varphi_{\rm min}$, into open ones whose barrier remains a large fraction of N (Appendix A). The barrier (9) is the free energy at the lowest point of an absorbing boundary, not at a stationary point of $\mathcal{F}$, so the multi-dimensional Langer theory (Langer, 1969; Hänggi et al., 1990) does not apply directly; the rate is the Kramers rate through the lowest point of that boundary.

After crossing the barrier, the system thermalizes, that is, reaches the Boltzmann equilibrium, in the basin where both $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^\mu$ are in support. In this basin the energy depends on $\phi_1 + \phi_\mu$, which from (6) is a function of (a, v) alone; the perpendicular component $\mathbf{c}_\perp$ contributes only through the sphere constraint $|\mathbf{c}_\perp|^2 = 1 - a^2 - v^2$. Integrating out $\mathbf{c}_\perp$ produces an entropy $s \propto \frac{1}{2}\ln(1 - a^2 - v^2)$, which penalizes large a and v. Balancing this entropy against the two-pattern energy $u_2 = -b^{-1}\ln[(1-b+b\phi_1)+(1-b+b\phi_\mu)]$ (the LSR energy (1) restricted to the two in-support patterns; see Petrova et al. (2026) for the general finite-temperature framework) and parametrizing the equilibrium by $\phi_1 = \phi_\mu \equiv \phi$ (the symmetric minimum) yields the self-consistency equation

$$\frac{1}{1 - b + b\,\phi} = \frac{2\,T\,\phi}{(1 + \varphi_{1\mu}) - 2\phi^2}\,. \tag{10}$$

At $T = 0$ this reduces to $\phi = \sqrt{(1+\varphi_{1\mu})/2}$, the geometric centroid. At finite T, thermal fluctuations shift ϕ to a lower value $\phi_{\rm cen}(\varphi_{1\mu}, T)$. Since $\varphi_{1\mu}$ varies slowly with α and $\phi_{\rm cen}$ is a weak function of T, the centroid position is approximately universal across the metastable regime; Section 3 compares it with the steps of Fig. 1.

After reaching the centroid $(\boldsymbol{\xi}^1, \boldsymbol{\xi}^\mu)$, the system can escape again by the same mechanism, recruiting a new pattern $\boldsymbol{\xi}^\nu$. The sphere constraint makes it difficult for three patterns to be in support simultaneously and the MC data confirm that transitions are pair-to-pair: one pattern enters support while another drops out. The result is a chain of pair-to-pair transitions:

$$\underbrace{\boldsymbol{\xi}^1}_{\phi_1 = \phi_{\rm eq}} \xrightarrow{\text{1st hop}} \underbrace{(\boldsymbol{\xi}^1,\, \boldsymbol{\xi}^\mu)}_{\phi_1 = \phi_{\rm cen}} \xrightarrow{\text{2nd hop}} \underbrace{(\boldsymbol{\xi}^\mu,\, \boldsymbol{\xi}^\nu)}_{\phi_1 \ll \phi_{\rm cen}} \to \cdots \tag{11}$$

After the second hop, the target alignment is $\phi_1 = (\varphi_{1\mu} + \varphi_{1\nu})/\sqrt{2(1+\varphi_{\mu\nu})}$. Since $\boldsymbol{\xi}^\nu$ was selected for high alignment with $\boldsymbol{\xi}^\mu$, $E[\varphi_{1\nu}|\varphi_{\mu\nu}] = \varphi_{1\mu}\varphi_{\mu\nu}$, which places the second step far below the centroid (Section 3). The system performs a random walk on the pattern overlap graph, always at some two-pattern centroid. In the LSR model the energy diverges when no pattern is in support, so the paramagnetic state of the Hopfield model, the phase with no macroscopic alignment with any stored pattern (Amit et al., 1987), has no analog.

3 Monte Carlo simulations

The bulk survey and the theory of Section 2 establish that the centroid states exist and explain their alignments: the centroid position from (10) is $\phi_{\rm cen} \approx 0.82$–$0.84$ across the metastable regime, matching $\phi_{\rm max}^{\rm other} \approx 0.82$ in every non-retrieval bin of the bulk data and the step at $\phi \approx 0.83$ in Fig. 1, and the residual overlap after the second hop, $\phi_1 \approx 0.51$ at $\alpha = 0.20$, matches the step at $\phi \approx 0.48$. We now test the escape *timescale* quantitatively, asking whether the Kramers mechanism accounts for the rate at which trials leave the retrieval basin.

The targeted simulations cover $\alpha \in [0.20,\, 0.34]$ and $T \in [0.15,\, 2.0]$, with $n_{\rm dis} = 2000$ disorder samples per point and run lengths from 10^4 to 5×10^5 MC steps. Each trial is initialized at $\phi_1 = 1$. Three successive measurement series cover 68 distinct (α, T) points. The statistical error of an

escaped fraction is binomial, below 0.011 at 2000 samples, and depends on the sample count, not on the system size.

Comparing runs that detect escape through any pattern ($\phi_\mu \geq \varphi_c$ for any $\mu \neq 1$) with runs monitoring only the nearest pattern ($\phi_{\mu_{\max}} \geq \varphi_c$), we find that the latter gives escape times 15–100$\times$ longer, confirming that the total rate is set by the sum over K channels. The escape is detected in two steps:

1. At each recording step, all M alignments ϕ_μ are checked. When any $\phi_\mu \geq \varphi_c$ for $\mu \neq 1$, a *candidate escape* is flagged, and the identity of pattern μ and its inter-pattern alignment $\varphi_{1\mu}$ are recorded.
2. The escape is *confirmed* by measuring the diffusion coefficient D_v in the perpendicular direction toward $\boldsymbol{\xi}^\mu$ over the next 50 recording intervals. A genuine escape initiates Stage 2 ($\Delta E < 0$), which accelerates the motion: $D_v^{\rm post}/D_v^{\rm ret} > 2$. If the ratio is below this threshold, the system has reached the support boundary and returned without sliding to the centroid, and the candidate is discarded.

This criterion uses the physical barrier-crossing condition $\phi_\mu \geq \varphi_c$ from Section 2. A simpler approach based on thresholding ϕ_1 gives escape times biased low by 5–20$\times$ at $T \geq 0.4$, because thermal fluctuations momentarily push ϕ_1 below the threshold without the system leaving the retrieval basin.

The confirmation threshold is the one arbitrary element of this criterion, and the conclusions do not depend on it. Repeating the classification from the stored per-trial diagnostics with thresholds 3 or 5 in place of 2 changes the median escape times by at most a factor 1.4 and moves the apparent boundary $\tau = T_{\rm MC}$ by about one grid step in α.

The multi-channel Kramers picture predicts that the total escape rate for a fixed disorder sample is proportional to K, the number of escape-enabling patterns, times the Boltzmann factor for the entropic barrier. The disorder average of this rate and the resulting escape-time formula are discussed in Appendix B. Assigning the lowest barrier to all K channels gives the closed form

$$\tau \;\approx\; \frac{\tau_{\rm rel}}{A\,K}\,\exp\!\left(\frac{\Delta F_{\min}}{T}\right), \tag{12}$$

where $\tau_{\rm rel} = R^2/D_v$ is the relaxation time of the perpendicular coordinate in the retrieval basin, $\Delta F_{\min} \equiv \Delta F(\varphi_{1,\max})$ is the barrier at the nearest-pattern alignment $\varphi_{1,\max} = (1 - e^{-2\alpha})^{1/2}$, $K = K(\alpha, T)$ is the number of escape-enabling patterns, and A is a dimensionless prefactor. The diffusion coefficient D_v entering $\tau_{\rm rel}$ is measured from separate MC simulations within the retrieval basin (distinct from the $D_v^{\rm post}/D_v^{\rm ret}$ ratio used for escape confirmation above). The LSR energy landscape makes the dynamics strongly anisotropic (D_v differs from the target-direction diffusion coefficient D_{ϕ_1} by a factor of 8–10), so the free-diffusion estimate $D_{\rm free} = \sigma^2/(2N)$ cannot be used. Equation (12) is a qualitative summary of the mechanism (Appendix B), and no quantitative result below rests on it. The calibration, the $\tau = T_{\rm MC}$ contour, and the lifetime estimates are all evaluated from the disorder-averaged rate (Appendix B)

$$\frac{1}{\tau} = \frac{A\,(M-1)}{\tau_{\rm rel}} \int_{\varphi_{\min}}^{1} e^{-\Delta F(\varphi)/T}\, f(\varphi)\, d\varphi\,, \tag{13}$$

which keeps the φ-dependent barrier under the integral and has a single free parameter, the prefactor A.

To calibrate A we compare escape probabilities, not fitted escape times, because the low-temperature runs are censored: at $\alpha = 0.20$, $T \leq 0.3$ only 5 to 18% of the 2000 samples escape before the run ends, and a time fitted to such data reflects the low-barrier tail of the disorder distribution rather than the typical barrier. For each measured (α, T) point we evaluate the survival (20) of Appendix B at the run length of that point, with the exact density (2), the barrier (9), and the measured D_v, and read off the predicted escape probability within the window. Calibrating A at the single point $(\alpha, T) = (0.20, 0.50)$ gives $A = 33$. Figure 3 (left) compares measured and predicted probabilities for all 50 points with $T \leq 0.8$; at higher T escape is not a rare event within the window and the waiting-time model does not apply. At $\alpha \leq 0.24$ the single A places every point on the diagonal at all measured temperatures, with median deviation 0.04. At $\alpha \geq 0.26$ the independent-channel sum over-predicts escape, consistent with inter-channel correlations slowing the dynamics (Appendix B).

Whether the entropic barrier survives at large N is decided by a competition of exponents: the number of escape channels grows exponentially with N, while each barrier grows linearly in N. The saddle-point evaluation of the disorder-averaged rate (13) resolves it (Appendix B): up to factors polynomial in N,

$$\frac{1}{\tau} \sim \frac{A}{\tau_{\rm rel}} e^{N(\alpha-\alpha_{\rm th})} \tag{14}$$

at every temperature: the entropy of the escape channels and their barriers compensate exactly, and the exponent changes sign at the thermodynamic threshold (3). For $\alpha < \alpha_{\rm th}$, the fraction of trials that escape within any window growing polynomially with N vanishes with the rate, so the metastable regime extends to $\alpha_{\rm th}$. A direct test varies M at fixed α (Appendix B): at $\alpha = 0.20$ the escaped fraction falls monotonically over six system sizes, $N = 38$ to 73, and at $\alpha = 0.28$ the median escape time grows with N as (14) predicts.

Figure 3 (right) shows the three-zone phase diagram. The background is the disorder-averaged alignment $\langle\phi\rangle(\alpha, T)$ from the bulk survey (see Introduction). The blue dashed curve is the isoline $\Delta F(\varphi_{1,\max}) = NT$ of the barrier (9) of the nearest stored pattern in the thermodynamic limit. To its left every stored pattern is separated from the retrieval state by a free-energy barrier above NT, so the escape time exceeds e^N up to a prefactor and retrieval is stable on any accessible timescale. The line marks a lifetime, not a transition: the barrier varies continuously across it (Appendix A), and the only boundary of retrieval in the thermodynamic limit is $\alpha_{\rm th}$. The black dotted curve, the vertical line $\alpha_{\rm th}$ joined at low temperature by the finite-temperature branch $\alpha_c(T) = -\frac{1}{2}\ln[1-(1-f_{\rm ret})^2]$ of the thermodynamic boundary of Petrova et al. (2026), with $f_{\rm ret}(T) = \mathcal{F}(\phi_{\rm eq}, 0)$ the retrieval free energy per site and the maximum spurious alignment taken from the exact density (2), is the corrected theoretical boundary (3) beyond which the state thermally wandering near the retrieval minimum already fills the support of a spurious pattern. There the escape rate (14) no longer decreases with N, the two basins merge into a centroid, and retrieval is lost. Below $T \approx 0.2$ the branch $\alpha_c(T)$ lies to the right of $\alpha_{\rm th}$ and marks where the retrieval state ceases to have the lowest free energy (Petrova et al., 2026); the exponent (14) gives the rate of leaving the retrieval basin, not which basin is the equilibrium one. This is the non-retrieval boundary in the thermodynamic limit, and its position restores the theoretical result of Petrova et al. (2026) once the exact density (2) is used in place of the Gaussian approximation.

Between this line and $\alpha_{\rm th}$ lies the metastable regime. Here escape channels with barriers below NT exist, but they are still separated from the retrieval basin by an entropic barrier. The escape time grows exponentially with N, at the parameter-free rate (14), so in the limit of large N this metastable zone becomes indistinguishable from thermodynamic stability, and the theoretical retrieval region is recovered. In Monte Carlo simulations at finite N, however, the metastable state has a finite lifetime, and the apparent retrieval boundary depends on the simulation time. The green dotted curve shows the contour $\tau = T_{\rm MC}$ of the disorder-averaged escape time (13) with the same constant A: to its left, the escape time exceeds $T_{\rm MC}$ and the state appears stable; to its right, the system escapes during the simulation and hops away from the retrieval basin.

The magenta dash-dotted curve, $\alpha^*(T) = I(\varphi_c\phi_{\rm eq})$ with $I(\varphi) = -\frac{1}{2}\ln(1-\varphi^2)$ the rate function of the density (2), so that $\alpha_{\rm th} = I(\varphi_c)$, separates two escape routes. The exponent (14) of the disorder-averaged rate is set by the patterns at $\varphi^* = \varphi_c\phi_{\rm eq}$, where the integrand of the rate (13) is maximal (Appendix B). A typical sample contains such patterns only if its nearest pattern is at least that close, $\varphi_{1,\max} \geq \varphi^*$, that is for $\alpha \geq \alpha^*(T)$. To the right of the curve a typical sample escapes through these patterns with the exponent $N(\alpha_{\rm th} - \alpha)$. To the left, its nearest pattern is farther than φ^*, the escape goes through this pattern alone, and the exponent $\Delta F(\varphi_{1,\max})/T$ exceeds $N(\alpha_{\rm th} - \alpha)$, so typical samples escape more slowly than the average. Both the exponent and its derivative in α are continuous at α^*, so the curve is a crossover between escape routes, not a boundary of retrieval.

4 Discussion and summary

The thermodynamic theory of Petrova et al. (2026) predicts a sharp retrieval boundary for the LSR kernel. Using the Gaussian approximation for the inter-pattern alignment distribution, the boundary is at $\alpha_{\rm th} = \varphi_c^2/2 = 0.25$; using the exact density (2), it shifts to $\alpha_{\rm th} = -\frac{1}{2}\ln(1-\varphi_c^2) \approx 0.347$. Monte Carlo simulations, however, place the apparent retrieval boundary much closer to $\alpha \approx 0.25$,

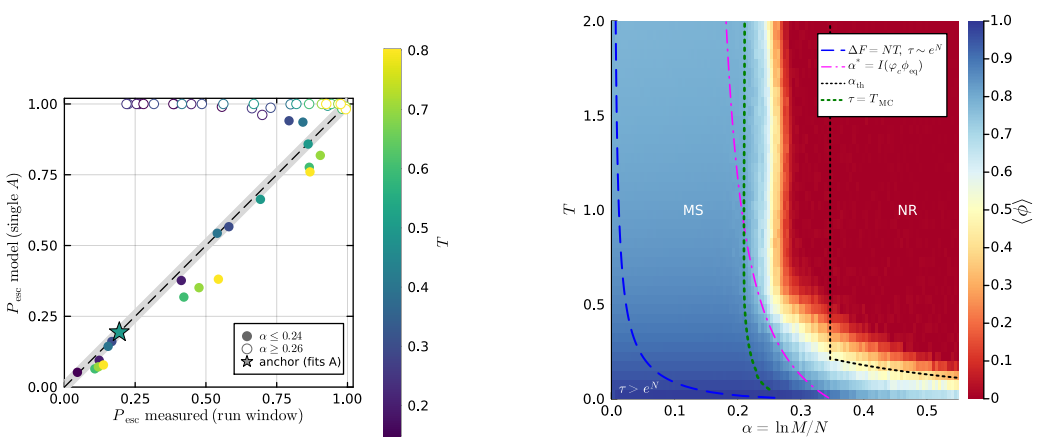

Figure 3: **Left:** measured versus predicted escape probability, within the run window of each point, for all 50 points with $T \le 0.8$. Model: survival (20) of Appendix B with the exact density (2), the barrier (9), the measured D_v, and the single constant $A = 33$ calibrated at the starred point. Filled: $\alpha \le 0.24$. Open: $\alpha \ge 0.26$, where independent channels over-predict escape. Gray band: twice the binomial error at 2000 samples. Color: temperature. **Right:** Three-zone phase diagram ($M =$ 20,000). Background: $\langle\phi\rangle(\alpha, T)$ from bulk survey. Blue dashed: isoline $\Delta F(\varphi_{1,\max}) = NT$ of the barrier of the nearest stored pattern, $N \to \infty$; to its left the escape time exceeds e^N. Black dotted: $\alpha_{\rm th}$ with $\alpha_c(T)$, thermodynamic limit. Green dotted: $\tau = T_{\rm MC}$ contour from (13) with $A = 33$. Magenta dash-dotted: $\alpha^*(T)$, where the nearest stored pattern reaches the saddle $\varphi^* = \varphi_c\phi_{\rm eq}$ of the disorder average. MS: metastable. NR: non-retrieval.

in seeming agreement with the Gaussian result and in apparent contradiction with the exact-density prediction.

The present analysis resolves the contradiction. The compact support of the LSR kernel creates an entropic barrier between the retrieval basin and the basins of spurious patterns (Section 2). The barrier height $\Delta F/T \propto N$ grows linearly with the dimension, so the escape time $\tau \propto \exp(\Delta F/T)$ is exponentially large in N. At the system sizes accessible to MC simulations ($N \approx 40$–50), this escape time is comparable to the simulation time $T_{\rm MC} \approx 4 \times 10^4$ near $\alpha \approx 0.21$. The coincidence with the Gaussian threshold is accidental: it depends on $T_{\rm MC}$ and N, not on the Gaussian approximation being correct. At larger N, the barrier grows and the kinetic boundary shifts rightward, eventually reaching the exact-density threshold $\alpha_{\rm th} \approx 0.347$, where the escape rate (14) stops decreasing with N and retrieval is lost. The shift is both derived and measured: the rate exponent $N(\alpha - \alpha_{\rm th})$ of (14) contains no fitted quantity, and the escaped fraction at fixed load falls monotonically with the system size (Section 3, Appendix B). The theoretical result of Petrova et al. (2026) is thus restored once the exact density is used.

The escape proceeds through K channels, each corresponding to a stored pattern whose support overlaps with the retrieval basin. We confirm this multi-channel mechanism by comparing single-channel and all-channel escape detection, whose ratio of escape times matches the predicted K (Section 3). After escape, the system hops between two-pattern centroids on the pattern overlap graph (11); a true paramagnetic state does not exist for the LSR energy because the energy diverges when no pattern is in support.

The disorder-averaged escape rate (13), with the single fitted constant A, reproduces the measured escape probabilities across the calibration range (Fig. 3, left). For practical retrieval at $N \gg 50$, the metastable state is effectively permanent: at $\alpha = 0.20$ and $T = 0.40$, the disorder-averaged escape time from (13) is $\tau \approx 2 \times 10^5$ MC steps at $N = 50$ and $\tau \approx 2 \times 10^8$ at $N = 100$, and typical samples are slower than this average (Appendix B).

One approximation limits the quantitative claims. The escape through a single pattern reduces exactly to a one-dimensional problem on the entry line, solved at its saddle (9); the approximation is to add the escapes through different patterns as independent routes. The routes are correlated, and the data

show the direction of the bias: at $\alpha \geq 0.26$ measured escape is slower than the model predicts (Fig. 3, left), so the model is conservative with respect to retrieval. The bias is a prefactor effect: it does not touch the exponential growth of the barrier with N, and the boundary $\alpha_{\rm th}$ is set by geometry alone ($\varphi_{1,\max} = \varphi_c$), with no rate model entering it.

Finite-temperature retrieval is the operating condition of neuromorphic hardware, where thermal fluctuations, device variability, and drift perturb the stored state continuously. For such systems the phase diagram gives the retrieval lifetime as a function of load and noise, and the finite-size analysis gives the growth of that lifetime with N.

The complete simulation code, the measured data, and the scripts that regenerate every figure of this paper are provided as supplementary material.

A question raised by the corrected threshold concerns the storage capacity of the LSE kernel: the Gaussian approximation used by Ramsauer et al. (2021) predicts a limiting load $\alpha = 0.5$. This bound originates from the tail of the Gaussian: at $\alpha = 0.5$ the predicted maximum spurious alignment $\sqrt{2\alpha}$ reaches 1, the largest value an alignment can take. For the exact density (2), however, the maximum inter-pattern alignment scales as $\varphi_{1,\max} = (1 - e^{-2\alpha})^{1/2}$, which remains strictly below 1 for all finite α. Whether retrieval persists beyond $\alpha = 0.5$ for the LSE kernel is an open question that we plan to address with dedicated Monte Carlo simulations.

Reproducibility statement

Every number in this paper comes from simulations whose code, parameters and outputs are released. The Monte Carlo protocol, the escape criterion and its confirmation threshold are stated in Section 3; the run windows, sample counts and grids are stated with them. All simulations were run on a single NVIDIA A6000 GPU with 50 GB of memory. The targeted simulations of Section 3 required approximately 60 GPU-hours, and the bulk survey of Section 1 a comparable amount on the same hardware. One (α, T) point with $M = 2 \times 10^4$ and 2000 disorder samples takes about 1.5 hours. The cost scales linearly in M times the number of samples, and at fixed α exponentially with N through $M = e^{\alpha N}$. The entropic barrier is derived in Appendix A and the disorder-averaged rate, together with its large-N limit, in Appendix B. The supplementary material contains the simulation scripts, the measured CSV files, the analysis scripts, and one plotting script per figure, with a README that maps each figure to its producing script and documents every data column.

AI use statement

Large language models (Claude, Anthropic) were used in this work in two roles. As a coding assistant: writing and debugging the simulation and analysis code, preparing the figures, and searching the literature. As a copy-editor: the text, written by the author, was revised sentence by sentence for readability, and the author checked each revision to confirm that the meaning was unchanged. The author takes full responsibility for the content.

References

Daniel J. Amit, Hanoch Gutfreund, and Haim Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173(1):30–67, 1987.

Mete Demircigil, J. Heusel, Matthias Löwe, S. Upgang, and F. Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017.

Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after Kramers. *Reviews of Modern Physics*, 62(2):251–341, 1990.

Benjamin Hoover, Zhaoyang Shi, Krishnakumar Balasubramanian, Dmitry Krotov, and Parikshit Ram. Dense associative memory with epanechnikov energy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554.

Hendrik A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

James S. Langer. Statistical theory of the decay of metastable states. *Annals of Physics*, 54(2): 258–275, 1969.

Tatiana Petrova. Thermal robustness of retrieval in dense associative memories: LSE vs LSR kernels. In *New Frontiers in Associative Memory Workshop at ICLR 2026*, 2026. arXiv:2603.13350.

Tatiana Petrova, Evgeny Polyachenko, and Radu State. Geometric entropy and retrieval phase transitions in continuous thermal dense associative memory. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research. PMLR, 2026. arXiv:2604.07401.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

A Origin of the entropic barrier

The Metropolis dynamics on $S^{N-1}(\sqrt{N})$ proposes isotropic displacements and projects them onto the sphere. Along the escape path (see Stage 1 in the left panel of Fig. 2), the energy is exactly constant ($\Delta E = 0$), so every proposal is accepted regardless of v. However, the volume of the perpendicular subspace at position v is $\mathrm{Vol}(v) \propto (R^2 - v^2)^{(N-3)/2}$, with $R^2 = 1 - \phi_{\mathrm{eq}}^2$ as in Section 2: it is maximal at $v = 0$ and shrinks as $v \to v_{\mathrm{entry}}$. The sphere projection after each MC step therefore creates a geometric drift toward $v = 0$, which arises not from energy rejection but from the greater available volume at low v. The equilibrium distribution is $P_{\mathrm{eq}}(v) \propto (R^2 - v^2)^{(N-3)/2}$, and escape requires the system to overcome this geometric bias from the maximum-volume point $v = 0$ to v_{entry}. Since $\Delta E = 0$, the barrier on this slice is purely entropic; the ratio of volumes gives

$$\exp\left(\frac{\Delta F}{T}\right) = \frac{\mathrm{Vol}(0)}{\mathrm{Vol}(v_{\mathrm{entry}})} = \left(\frac{R^2}{R^2 - v_{\mathrm{entry}}^2}\right)^{(N-3)/2}, \tag{15}$$

and the escape time scales as $\tau \sim \tau_{\mathrm{rel}} \exp(\Delta F/T)$, with the relaxation time $\tau_{\mathrm{rel}} = R^2/D_v$ of Section 3.

This volume ratio is the barrier on the slice $a = \phi_{\mathrm{eq}}$. The full free energy per site of a state with only $\boldsymbol{\xi}^1$ in support, after integrating out $\mathbf{c}_\perp$, is $\mathcal{F}(a, v) = u(a) - \frac{T}{2}\ln(1 - a^2 - v^2)$ with $u(a) = -b^{-1}\ln(1 - b + ba)$; its minimum at $v = 0$ is the equation of Petrova et al. (2026) for $\phi_{\mathrm{eq}}(T)$, $(bT+1)y^2 - (2+T+Tb)y + T = 0$ with $\phi_{\mathrm{eq}} = 1 - y$. Writing φ for $\varphi_{1\mu}$, the support of $\boldsymbol{\xi}^\mu$ is entered on the line $a\varphi + v\sqrt{1-\varphi^2} = \varphi_c$, where $(1-a^2-v^2)(1-\varphi^2) = 1 - \varphi^2 - \varphi_c^2 - a^2 + 2a\varphi\varphi_c$, so the free energy on the entry line, relative to the retrieval state and per site, is

$$g(a) = \frac{u(a) - u(\phi_{\mathrm{eq}})}{T} - \frac{1}{2}\ln\frac{1 - \varphi^2 - \varphi_c^2 - a^2 + 2a\varphi\varphi_c}{(1-\varphi^2)(1-\phi_{\mathrm{eq}}^2)}, \qquad \varphi_c < a < 1, \tag{16}$$

the target being in support only for $a > 1 - 1/b = \varphi_c$. The function diverges at both ends, $u \to \infty$ as $a \to \varphi_c$ and the logarithm where the bracket vanishes, so its minimum is interior, at the root $a^* \in (\varphi_c, 1)$ of

$$(1 + Tb)\, a^2 + \left[T(1-b) - Tb\,\varphi\varphi_c - 2\varphi\varphi_c\right] a + \left[\varphi^2 + \varphi_c^2 - 1 - T(1-b)\,\varphi\varphi_c\right] = 0, \tag{17}$$

and $\Delta F/T = N g(a^*)$, with $N \to N - 3$ in the entropic term, is (9). At $a = \phi_{\mathrm{eq}}$ the bracket equals $(1-\varphi^2)(1-\phi_{\mathrm{eq}}^2)(1 - v_{\mathrm{entry}}^2/R^2)$ and g is the volume ratio above: the value at one point of the line, hence an upper bound. The minimizer lies above ϕ_{eq} for $\varphi > \varphi_c\phi_{\mathrm{eq}}$ and below it for $\varphi < \varphi_c\phi_{\mathrm{eq}}$, with $a^* = \phi_{\mathrm{eq}}$ exactly at $\varphi = \varphi_c\phi_{\mathrm{eq}}$. The two barriers coincide where v_{entry} is small and separate as the pattern recedes. At $T = 0.325$, $\varphi = 0.574$ (the nearest pattern at $\alpha = 0.20$) gives 0.167 per site on the slice against 0.164 at the saddle, $\varphi = 0.426$ gives 0.555 against 0.446, and below $\varphi_{\min} = 0.326$, where the slice value is infinite, the saddle value stays finite: 0.74 at $\varphi_{\min}$ itself,

growing as $(\frac{1}{bT}+\frac{1}{2})\ln(1/\varphi)$ as $\varphi \to 0$. Over $T = 0.15$ to 2 the saddle barrier at $\varphi_{\min}(T)$ never falls below 0.67 per site, the bound used in Appendix B.

The difference between the two barriers is immaterial for the results. On the grid of Section 3 the saddle barrier changes the disorder-averaged rate (13) by a few per cent and the calibrated prefactor from $A = 36$ to $A = 33$. The figures are drawn with the saddle barrier. The large-N exponent is the same for both barriers: minimizing $g(a) + I(\varphi)$ jointly over (a, φ), with $I(\varphi)$ the rate function of (2) (Section 3), puts $\varphi = a\varphi_c$ and leaves $[\mathcal{F}(a, 0) - \mathcal{F}(\phi_{\rm eq}, 0)]/T + \alpha_{\rm th}$, minimal at $a = \phi_{\rm eq}$. The saddle of the rate integral therefore stays at $\varphi^* = \varphi_c\phi_{\rm eq}$ with $a^* = \phi_{\rm eq}$, and (24) of Appendix B holds with the saddle barrier exactly.

B Compound Poisson model and the Kramers formula

For a fixed disorder realization d, the per-channel Kramers rate is

$$\lambda_\mu = \frac{A}{\tau_{\rm rel}} \exp\left(-\frac{\Delta F(\varphi_{1\mu})}{T}\right) \tag{18}$$

and the total escape rate is $\lambda_d = \sum_{\mu\neq 1} \lambda_\mu$, to which only the escape-enabling patterns, $\varphi_{1\mu} > \varphi_{\min}$, contribute measurably. The survival for realization d is $S_d(t) = e^{-\lambda_d t}$. Assuming that the K channels contribute independently to the escape rate, and that the $M-1$ inter-pattern alignments $\varphi_{1\mu}$ are independent draws from $f(\varphi)$ (2), the disorder-averaged survival factorizes:

$$S(t) = \langle S_d(t)\rangle = \left[\int_{-1}^{1} e^{-\lambda(\varphi)\,t}\, f(\varphi)\, d\varphi\right]^{M-1}, \tag{19}$$

where $\lambda(\varphi) = (A/\tau_{\rm rel})\exp(-\Delta F(\varphi)/T)$; below $\varphi_{\min}$ the barrier exceeds $0.67N$ (Appendix A) and the rate is negligible, so the integrals below may be cut at $\varphi_{\min}$ (the figures integrate over all $\varphi > 0$, with negligible difference). Writing $S(t) = [1 - J(t)]^{M-1}$ with $J(t) = \int(1 - e^{-\lambda(\varphi)t})f(\varphi)\, d\varphi$, we note that $J(t) \le K/(M-1) \ll 1$ since only the $K \ll M$ escape-enabling patterns contribute. The approximation $(1-J)^{M-1} \approx e^{-(M-1)J}$ gives the compound Poisson form:

$$S(t) \approx \exp\left(-(M-1)\int_{\varphi_{\min}}^{1} \left(1 - e^{-\lambda(\varphi)\,t}\right) f(\varphi)\, d\varphi\right). \tag{20}$$

At early times, $1 - e^{-\lambda t} \approx \lambda t$ and this reduces to exponential decay $S(t) \approx e^{-t/\tau}$ with $1/\tau$ given by the rate (13) of the main text. Assuming that the barriers vary slowly across the K channels, we replace $\Delta F(\varphi)$ by $\Delta F_{\min} \equiv \Delta F(\varphi_{1,\max})$ and factor the integral as $e^{-\Delta F_{\min}/T} \times K/(M-1)$, giving

$$\lambda_d \;\approx\; \frac{A\,K}{\tau_{\rm rel}}\, \exp\left(-\frac{\Delta F_{\min}}{T}\right). \tag{21}$$

Inverting (21) recovers the closed form (12) of the main text.

Two assumptions enter the derivation of (21): (i) the barriers vary slowly across the K channels, and (ii) the channels contribute independently. Neither assumption holds quantitatively.

Regarding (i): the barrier $\Delta F(\varphi)/T$ ranges from 0 at the zero-barrier boundary $\varphi_c/\phi_{\rm eq}$ to $\gg 10$ at low alignments. At $(\alpha, T) = (0.20, 2.0)$, for instance, 64% of K comes from $\varphi < 0.20$ where $\Delta F/T > 38$, while $\Delta F_{\min}/T \approx 7$. Replacing $e^{-\Delta F/T}$ by $e^{-\Delta F_{\min}/T}$ in the exact disorder-averaged rate (13) overestimates the integral by a factor $\sim 5 \times 10^3$.

Regarding (ii): a thermal fluctuation that moves the state toward the support boundary simultaneously reduces the barrier for all patterns in the same angular sector, correlating the per-channel rates. Fitting the constant A of (13) to the measured escape times separately at each temperature yields values that vary by a factor of ≈ 5 between $T = 0.5$ and $T = 2.0$, likely reflecting these correlations. Additionally, transverse fluctuations in the $(N-2)$-dimensional perpendicular subspace, which are neglected by the one-dimensional projection onto v, may allow the system to explore directions with lower effective barriers, further modifying the escape rate relative to (13); the fluctuation of the target alignment a itself is included in (9) through the saddle. Correlations can only slow the escape relative

Table 1: Finite-size test at $\alpha = 0.20$: escaped fraction within a window of 2.6×10^5 steps, measured (meas.) and predicted by the survival (20) with $A = 33$ and $\tau_{\rm rel} \propto N^2$ (model).

		$T = 0.40$		$T = 0.25$	
N	M	meas.	model	meas.	model
38	2×10^3	0.43	0.65	0.27	0.37
44	6.3×10^3	0.30	0.43	0.17	0.21
50	2×10^4	0.21	0.24	0.10	0.11
58	10^5	0.12	0.10	0.044	0.042
66	5×10^5	0.075	0.034	0.017	0.015
73	2×10^6	0.035	0.013	0.010	0.006

to independent channels. The large-N statement below does not use the independence assumption: $\Pr(\text{escape by } t) \le t \sum_\mu \lambda_\mu$ is a union bound, valid for correlated channels as well.

Assumption (i) enters only the closed form (21), which no result uses; assumption (ii) enters the factorization (19), and the union bound above shows that the large-N conclusion survives without it.

LARGE-N LIMIT OF THE DISORDER-AVERAGED RATE

As a large-N law at fixed α the minimum-barrier replacement behind (21) fails, and the exponent must be taken from (13) itself; the two separate only when M varies at fixed α, the test performed below.

The rate (13) keeps the φ-dependent barrier under the integral. Inserting the density (2) and the slice form of the barrier, its value at $a = \phi_{\rm eq}$, which by Appendix A has the same large-N maximum as (9), with $v_{\rm entry}(\varphi)$ from (7) and $R^2 = 1 - \phi_{\rm eq}^2$, the integrand becomes a single N-th power,

$$e^{-\Delta F(\varphi)/T} f(\varphi) \;\sim\; \left[(1-\varphi^2)\left(1 - v_{\rm entry}^2/R^2\right)\right]^{(N-3)/2}, \tag{22}$$

and the bracket simplifies:

$$(1-\varphi^2)\left(1 - \frac{v_{\rm entry}^2}{R^2}\right) = (1-\varphi^2) - \frac{(\varphi_c - \phi_{\rm eq}\varphi)^2}{1-\phi_{\rm eq}^2} = \frac{1 - \phi_{\rm eq}^2 - \varphi_c^2 - \varphi^2 + 2\varphi_c\phi_{\rm eq}\varphi}{1-\phi_{\rm eq}^2}\,. \tag{23}$$

The numerator is a downward parabola in φ with maximum at $\varphi^* = \varphi_c\,\phi_{\rm eq}$, which lies inside the integration range ($\varphi_{\min} < \varphi^*$ by (8)). At φ^* the numerator equals $(1-\varphi_c^2)(1-\phi_{\rm eq}^2)$, the factor $(1-\phi_{\rm eq}^2)$ cancels, and

$$\max_\varphi (1-\varphi^2)\left(1 - v_{\rm entry}^2/R^2\right) = 1 - \varphi_c^2\,. \tag{24}$$

With $M - 1 = e^{\alpha N}$, the Laplace evaluation of the integral gives, up to factors polynomial in N, $1/\tau = (A/\tau_{\rm rel})\, e^{N[\alpha + \frac{1}{2}\ln(1-\varphi_c^2)]} = (A/\tau_{\rm rel})\, e^{N(\alpha - \alpha_{\rm th})}$ by the definition of $\alpha_{\rm th}$ (3), which is (14) of the main text. The temperature enters only through $\phi_{\rm eq}(T)$, and $\phi_{\rm eq}$ has dropped out of the exponent.

The rate is dominated by the interior saddle $\varphi^* = \varphi_c\,\phi_{\rm eq}$, to which neither the nearest pattern nor the bulk of the K channels contributes at leading order; this is why the minimum-barrier replacement, which lets the nearest pattern carry all of K, does not survive the limit.

One caveat remains. The exponent belongs to the disorder average, (13) being the average of the per-sample rate $\lambda_d = \sum_\mu \lambda_\mu$, while the simulations measure typical samples, whose largest alignment is $\varphi_{1,\max}$. For $\alpha < \alpha^*(T)$ (Section 3) the saddle lies above it, $\varphi^* = \varphi_c\,\phi_{\rm eq} > \varphi_{1,\max}$: the average is then carried by rare samples with a pattern near φ^*, and a typical sample escapes even more slowly than $e^{-N(\alpha_{\rm th} - \alpha)}$, so the conclusion is unchanged. At $\alpha = \alpha_{\rm th}$, $\varphi_{1,\max} = \varphi_c \ge \varphi^*$: the saddle is populated in typical samples and the exponent crosses zero exactly at $\alpha_{\rm th}$, at every temperature.

FINITE-SIZE TEST

The test varies M at fixed α, with the same run window (2.6×10^5 steps) and the same escape criterion at every size. At $\alpha = 0.20$ this gives six system sizes, $M = 2 \times 10^3$ to 2×10^6, i.e. $N = \ln M/\alpha = 38$

to 73. Table 1 lists the escaped fraction at two temperatures against the survival (20) with $A = 33$, the same window, and $\tau_{\rm rel} \propto N^2$, with no parameter adjusted. The measured fraction falls monotonically, at $T = 0.40$ as $\sim e^{-0.07N}$. The predicted one decays as $\sim e^{-0.12N}$ at both temperatures, and its slope lies below the asymptotic $\alpha_{\rm th} - \alpha = 0.147$ of (14) because at these sizes the polynomial prefactors of the Laplace evaluation and the growth of $\tau_{\rm rel}$ with N are not yet negligible; the asymptotic slope is approached only at larger N. At $T = 0.25$ the model is within twice the binomial error at $N \geq 50$; at $T = 0.40$ the measured decay is slower than predicted, by a factor 2.7 at $N = 73$, in the direction of the inter-channel correlations discussed above. At $N = 38$ the load $\alpha = 0.20$ lies well inside the apparent escape region, and at $N = 73$ almost every trial retrieves: the apparent boundary is kinetic, and it drifts toward $\alpha_{\rm th}$ as N grows.

At $\alpha = 0.28$, $T = 0.50$, escape is fast enough for its time to be measured directly. The median escape time at $M = 2 \times 10^4$, 2×10^5, 2×10^6, i.e. $N = 35$, 44, 52, is 4700, 11300, 21600 MC steps: escape slows with N by factors 2.4 and 1.9, against $e^{(\alpha_{\rm th}-\alpha)\Delta N} = 1.8$ and 1.7 from (14) (2.9 and 2.4 with the $\tau_{\rm rel} \propto N^2$ prefactor). Had the growth of K dominated, escape would have become faster at larger N. The data go the other way.