Geometric Entropy and Retrieval Phase Transitions in Continuous Dense Associative Memory

Anonymous Authors[1]

Abstract

We study the thermodynamic memory capacity of modern Hopfield networks (Dense Associative Memory models) with continuous states under geometric constraints, extending classical analyses of pairwise associative memory. We derive thermodynamic phase boundaries for Dense Associative Memory networks with exponential capacity $p = e^{\alpha N}$, comparing Gaussian (LSE) and Epanechnikov (LSR) kernels. For continuous neurons on an N-sphere, the geometric entropy depends solely on the spherical geometry, not the kernel. In the sharp-kernel regime, the maximum theoretical capacity $\alpha = 0.5$ is achieved at zero temperature; below this threshold, a critical line separates retrieval from a spin-glass phase. The two kernels differ qualitatively in their phase boundary structure: for LSE, the retrieval region extends to arbitrarily high temperatures as $\alpha \to 0$, but interference from spurious patterns is always present. For LSR, the finite support introduces a threshold α_{th} below which no spurious patterns contribute to the noise floor, producing a qualitatively different retrieval regime in this sub-threshold region. These results advance the theory of high-capacity associative memory and clarify fundamental limits of retrieval robustness in modern attention-like memory architectures.

1. Introduction

Associative memory networks store patterns such that the correct memory can be retrieved from a noisy or partial cue. The classic example is the Hopfield network (Hopfield, 1982), which can reliably retrieve a memory pattern up to a certain storage limit. In the original Hopfield model, each neuron is binary and the memory capacity scales linearly with the number of neurons N (on the order of $0.14N$ patterns in the symmetric binary case (Amit et al., 1985)). Numerous extensions have been proposed to achieve superlinear capacity, including models with p-body interactions yielding polynomial scaling (Gardner, 1988; Krotov & Hopfield, 2016a). More recently, Dense Associative Memory (DAM) or modern Hopfield networks have been shown to achieve exponential storage capacity (Demircigil et al., 2017; Ramsauer et al., 2021).

In modern Hopfield models, pairwise interactions are replaced by highly nonlinear energy functions, substantially improving storage capacity and robustness to interference in high-load regimes. When defined over continuous states constrained to an N-dimensional sphere, DAMs can store an exponential number of patterns ($p = e^{\alpha N}$) with perfect recall in the zero-temperature limit (Demircigil et al., 2017; Ramsauer et al., 2021; Lucibello & Mézard, 2024). These models have attracted renewed interest due to their formal equivalence to softmax attention mechanisms in Transformers, making DAMs a tractable theoretical framework for analyzing capacity limits, interference, and robustness in attention-style retrieval mechanisms (Ramsauer et al., 2021). However, existing theoretical results largely focus on zero-temperature behavior, leaving open how retrieval stability persists under noise and thermal fluctuations. In energy-based associative memories, such noise naturally corresponds to finite temperature, where retrieval is governed by a competition between energy and entropy.

Recent work has revisited associative memory from a thermodynamic perspective. Building on classical analyses of retrieval-to-spin-glass transitions (Amit et al., 1987), Rooke et al. (2026) introduced a stochastic thermodynamics framework for DAMs, while Lucibello & Mézard (2024) analyzed retrieval stability under exponential memory load, emphasizing entropic penalties in high-dimensional settings. Nevertheless, it remains unclear how such thermodynamic effects constrain retrieval robustness in continuous DAMs that are directly connected to modern neural architectures.

While both our analysis and that of Lucibello & Mézard (2024) address exponential memory load and thermodynamic limits, the latter primarily focuses on zero-temperature or energy-dominated regimes. In contrast, we

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

derive explicit finite-temperature phase boundaries for continuous DAMs on the N-sphere and show that retrieval stability is governed by a competition between kernel-dependent energy and a kernel-independent *geometric entropy* term induced solely by the spherical constraint. This separation enables a principled comparison of different kernels and reveals qualitative differences in thermal robustness, including a support threshold for compactly supported kernels below which retrieval is perfect at any temperature. In particular, it clarifies which aspects of retrieval robustness are determined by modeling choices, such as the similarity kernel, and which are imposed by high-dimensional geometry.

From this finite-temperature perspective, we obtain an explicit characterization of the thermodynamic retrieval phase in continuous DAMs under exponential memory load by deriving the phase boundary $\alpha_c(T)$ in the (α, T) plane separating retrieval from disordered behavior. We study DAMs on the N-sphere with two kernel-based energy functions: the Gaussian log-sum-exp (LSE) and the compactly supported log-sum-ReLU (LSR) (Hoover et al., 2025). Although both achieve exponential capacity as $T \to 0$, they differ qualitatively in how kernel sharpness mediates robustness to thermal noise. In high dimensions, sharper kernels incur a larger entropic cost to maintain high alignment, leading to distinct phase boundary structures.

Our main contributions are: (i) analytical characterization of the finite-temperature retrieval transition in continuous DAMs with exponential memory load; (ii) identification of a kernel-independent geometric entropy term on the N-sphere and its competition with kernel-dependent energy; and (iii) explicit phase boundaries for LSE and LSR, showing that LSR exhibits a support threshold below which retrieval is perfect at any temperature, while LSE maintains ever-present interference across all loads.

Paper outline. Section 2 introduces the DAM models and kernel-based energy functions studied in this work. Section 3 develops the thermodynamic framework for retrieval, including the replica formulation, geometric entropy induced by the spherical constraint, and the noise-floor comparison used to determine stability. Section 4 derives explicit phase boundaries $\alpha_c(T)$ for LSE and LSR and presents phase diagrams with Monte Carlo validation. Section 5 summarizes the main findings. Derivations are provided in Sections A and B.

2. Dense Associative Memory networks

In this section, we specify the class of energy-based associative memory models studied in this work and relate their energy functions to kernelized, attention-like retrieval.

In the study of Dense AM (Associative Memory), two types of Hamiltonians, i.e., energy functions, that define the internal energy of the system are distinguished. The first type is k-polynomial (Krotov & Hopfield, 2016b). Here, retrieval can be viewed as minimizing $H_k(\boldsymbol{x})$ over the state $\boldsymbol{x}$, so the form of H_k determines the geometry of basins of attraction and, ultimately, the storage capacity.

$$H_k(\boldsymbol{x}) = -\frac{1}{N^{k-1}} \sum_{\mu=1}^{M} (\boldsymbol{x} \cdot \boldsymbol{\xi}^\mu)^k \,, \qquad (1)$$

a special case of which is the classical Hopfield network (Hopfield, 1982) when $k = 2$. The vectors $\boldsymbol{x}$ and $\boldsymbol{\xi}^\mu$ belong to the same space and define, respectively, the current state of the system and the stored patterns (memories, $\mu = 1 \dots M$). They can either be simply elements of the binary neuron space $\{-1, +1\}^N$, or be continuous, for example, bounded on each coordinate by the interval $[-1, 1]$. This type is characterized by polynomial capacity $M_{\max} = \alpha N^{k-1}$ (Krotov & Hopfield, 2016b). In particular, for the Hopfield network with spins, the well known result is $M_{\max} \approx 0.14N$.

The second type of Hamiltonians is the exponential interaction type. The connection to polynomial interactions arises in the limiting case $k \to \infty$, where polynomial terms give way to exponential ones. This modification yields a maximum capacity of $M_{\max} \approx \exp(\alpha N)$, with $\alpha \approx 0.35$ for binary spin data (Demircigil et al., 2017).

A particularly relevant instance is the log-sum-exp energy, whose gradient-based dynamics induce a softmax-like weighting over stored patterns. Ramsauer et al. (2021) extended these exponential capacity results to continuous states using the Log-Sum-Exp (LSE) energy function,

$$H_{\mathrm{LSE}}(\boldsymbol{x}) \propto -\ln\left(\sum_{\mu=1}^{M} \exp(\beta_{\mathrm{net}}\, \boldsymbol{x} \cdot \boldsymbol{\xi}^\mu)\right), \qquad (2)$$

and showed that the update rule minimizing this energy is mathematically equivalent to the self-attention mechanism in Transformers, with stored patterns as keys and the state vector as query. To ensure bounded activations, vector magnitudes must be constrained. We adopt the N-sphere model, where continuous real-valued coordinates satisfy $\sum_{i=1}^{N} x_i^2 = N$. Under this constraint, Ramsauer et al. (2021) proved $M_{\max} \propto \exp(\alpha N)$ with $\alpha = 0.5$.

Krotov & Hopfield (2021) introduced an alternative formulation based on Euclidean distance $d^2(\boldsymbol{x}, \boldsymbol{\xi}^\mu) = \|\boldsymbol{x} - \boldsymbol{\xi}^\mu\|^2$, yielding the kernel-based energy function:

$$H_{\mathrm{LSE}}(\boldsymbol{x}) = -\frac{1}{\beta_{\mathrm{net}}} \ln p(\boldsymbol{x}) \,, \qquad (3)$$

with

$$p(\boldsymbol{x}) \equiv \sum_{\mu=1}^{M} \exp\left[-\frac{\beta_{\mathrm{net}}}{2} d^2(\boldsymbol{x}, \boldsymbol{\xi}^\mu)\right]. \qquad (4)$$

This form corresponds to the negative log-likelihood of a Gaussian kernel density estimate of the memory distribution (Hoover et al., 2025). Equivalently, the network assigns higher scores to states $\boldsymbol{x}$ that lie in high-density regions of the memory distribution under this kernel, making retrieval a form of mode-seeking in representation space. Under the N-sphere constraint, the dot product and distance formulations are equivalent:

$$d^2(\boldsymbol{x}, \boldsymbol{\xi}^\mu) = \|\boldsymbol{x}\|^2 + \|\boldsymbol{\xi}^\mu\|^2 - 2\boldsymbol{x} \cdot \boldsymbol{\xi}^\mu = 2N - 2\boldsymbol{x} \cdot \boldsymbol{\xi}^\mu . \quad (5)$$

The parameter β_{net}, termed the *inverse variance*, has nothing to do with the outer thermal bath. Instead, it controls the sharpness of the similarity kernel and thereby the shape of the energy landscape, with effective width σ given by:

$$\sigma^2 \equiv \beta_{\text{net}}^{-1} . \quad (6)$$

Recently, Hoover et al. (2025) proposed the log-sum-ReLU (LSR) energy function, inspired by optimal kernel density estimation using the Epanechnikov kernel. This formulation achieves exact memory retrieval with exponential capacity without relying on exponential kernels:

$$H_{\text{LSR}}(\boldsymbol{x}) = -\frac{1}{\beta_{\text{net}}} \ln\left(\sum_{\mu=1}^{M} \max\left[\epsilon, 1 - \frac{d^2(\boldsymbol{x}, \boldsymbol{\xi}^\mu)}{2\sigma^2}\right]\right), \quad (7)$$

where $\epsilon > 0$ is a small constant introduced to prevent singularities in the logarithm.

Memory retrieval occurs when patterns are sufficiently separated and the initial state lies within the basin of attraction of the target pattern. The relevant quantity is the *alignment*:

$$\phi^\mu(\boldsymbol{x}) = \frac{1}{N} \boldsymbol{x} \cdot \boldsymbol{\xi}^\mu , \quad (8)$$

with $\phi^\mu \approx 1$ when $\boldsymbol{x}$ is near $\boldsymbol{\xi}^\mu$ and $\phi^\mu \approx 0$ otherwise. The Euclidean distance (5) becomes $d^2(\boldsymbol{x}, \boldsymbol{\xi}^\mu) = 2N[1 - \phi^\mu(\boldsymbol{x})]$, scaling linearly with N. This scaling requires a corresponding rescaling of the LSR inverse variance in the thermodynamic limit (see Appendix B).

When the width σ is small, the minima are well separated. Because LSE kernels have global support, gradient descent converges to a local minimum regardless of initialization. For LSR, successful retrieval requires the initial state to lie within the basin of attraction of a stored pattern. As σ increases, the two models behave differently. For LSE, nearby patterns simply merge into a single basin. For LSR, there exists a narrow range of σ in which spurious local minima can emerge between patterns; further increasing σ eventually causes basins to merge as in the LSE case (Hoover et al., 2025).

3. Statistical Mechanics of Retrieval

In this section, we develop a thermodynamic description of memory retrieval that makes it possible to analyze robustness to noise and interference in the high-dimensional limit introduced above.

Thermal Dense AM involves two sources of randomness. The first is quenched disorder: the random distribution of patterns across state space, which together with β_{net} determines the energy landscape. The second is thermal noise from coupling to a heat bath at temperature T. We analyze phase transitions in the thermodynamic limit $N \to \infty$, averaging over the disorder in the patterns ξ^μ.

3.1. The Macroscopic Framework and the Replica Method

Our goal is to characterize typical retrieval behavior averaged over random memory sets, which requires computing disorder-averaged macroscopic quantities rather than properties of a single realization.

The Hamiltonians (3, 7) can be expressed in terms of the alignment vector $\boldsymbol{\phi} = (\phi^1, \ldots, \phi^M)^T$:

$$H_{\text{LSE}}(\boldsymbol{x}) \equiv U_{\text{LSE}}(\boldsymbol{\phi}) = $$
$$-\frac{1}{\beta_{\text{net}}} \ln\left[\sum_{\mu=1}^{M} \exp\left(-\frac{N(1 - \phi^\mu)}{\sigma^2}\right)\right] . \quad (9)$$

Thermodynamic behavior is governed by the partition function Z, with $\beta = 1/T$:

$$Z = \int d\boldsymbol{x} \exp\left(-\beta H(\boldsymbol{x})\right) = \int d\boldsymbol{x} \, [p(\boldsymbol{x})]^{\beta/\beta_{\text{net}}} . \quad (10)$$

The disorder-averaged free energy density $\langle f \rangle = -(N\beta)^{-1}\langle \ln Z \rangle$ is computed via the replica identity (Sherrington & Kirkpatrick, 1975):

$$\langle \ln Z \rangle = \lim_{n \to 0} \frac{\langle Z^n \rangle - 1}{n} . \quad (11)$$

This formalism provides a systematic way to compute typical macroscopic retrieval behavior in the presence of quenched disorder. Evaluating $\langle Z^n \rangle$ requires passing from the microscopic integral over N neurons to a macroscopic description in terms of alignments and overlaps (see Appendix A). The result takes the standard form:

$$\langle f \rangle = -\frac{1}{N\beta} \langle \ln Z \rangle \approx u(\phi) - Ts , \quad (12)$$

where s is the entropy density and $u(\phi)$ is the internal energy density of a single memory basin. This decomposition makes explicit the competition between energetic alignment with a target memory and the entropic cost of constraining

the state in high dimensions. For the LSE and LSR kernels (see Appendices A and B):

$$u_{\text{LSE}} = 1 - \phi\,, \tag{13}$$

and

$$u_{\text{LSR}} = -b^{-1} \ln\left[1 - b(1 - \phi)\right]. \tag{14}$$

where $b \equiv N\beta_{\text{net}}$ is the rescaled inverse variance.

3.2. Geometric Entropy on the N-sphere

Here we show that even in the absence of noise, geometric constraints alone induce an entropic pressure that limits retrieval in high-dimensional spaces.

To derive the entropy, we compute the volume Ω of configuration space satisfying the alignment constraints ϕ. We decompose x into orthogonal components: the projection $x_{\|}$ onto the subspace spanned by the M memories, and $x_{\perp}$ in the remaining $N - M$ dimensions. We assume $M < N$, though phase transition analysis requires only $M = 1$.

For linearly independent memories, the squared norm of the parallel component is:

$$\|\mathbf{x}_{\|}\|^2 = N\phi^T G^{-1}\phi \tag{15}$$

where $G_{\mu\nu} = \frac{1}{N}\xi^\mu \cdot \xi^\nu$ is the Gram matrix. Since the total norm is fixed to N:

$$\|\mathbf{x}_{\perp}\|^2 = N\left(1 - \phi^T G^{-1}\phi\right). \tag{16}$$

The volume $\Omega(\phi)$ is proportional to the surface area of an $(N - M)$-sphere with radius $R = \|\mathbf{x}_{\perp}\|$, giving:

$$\Omega(\phi) \propto \left[N\left(1 - \phi^T G^{-1}\phi\right)\right]^{(N-M-1)/2}. \tag{17}$$

In the large N limit, discarding ϕ-independent constants:

$$S(\phi) = \ln\Omega = \frac{N - M}{2}\ln\left(1 - \phi^T G^{-1}\phi\right). \tag{18}$$

For a single memory as $N \to \infty$, the entropy density becomes:

$$s(\phi) \equiv \frac{S(\phi)}{N} = \frac{1}{2}\ln(1 - \phi^2) \tag{19}$$

This expression holds for a single microstate with alignment ϕ. This dependence is purely geometric and does not involve any kernel-specific assumptions. In the thermal setting, under the replica symmetry ansatz, the system occupies a cloud of states centered at a point with alignment ϕ to the target pattern. We characterize this cloud by the self-overlap parameter:

$$q = \frac{1}{N}\langle x \rangle \cdot \langle x \rangle\,, \tag{20}$$

which measures the cloud's tightness on the N-sphere: as $T \to 0$, the cloud collapses to a point ($q \to 1$); as $T \to \infty$,

it spreads over the sphere ($q \to 0$). The quantity $1 - q$ gives the variance of fluctuations about the mean state.

Generalizing the geometric entropy to account for finite tightness q (see Appendix A):

$$s(\phi, q) = \frac{1}{2}\ln(1 - q) + \frac{q - \phi^2}{2(1 - q)}. \tag{21}$$

Along the equilibrium line $q = \phi^2$ (see Section 4), (21) reduces to (19).

3.3. Crosstalk and the Noise Floor

In classical associative memory with linear load $M = \alpha N$, interference between stored patterns manifests as Gaussian noise, captured by an additional term in the free energy density (Amit et al., 1987). This noise can trap the system in spurious states outside the true minima.

For exponential capacity $M = e^{\alpha N}$, the Gaussian approximation fails. Because M is exponential in N, random configurations can accidentally align with nearby patterns, creating a noise floor that competes with retrieval. This regime is best described by the Random Energy Model (REM) (Derrida, 1980). In this regime, retrieval fails not due to local noise, but due to the presence of exponentially many competing alignments.

We therefore adopt a different approach: we compute the retrieval free energy $f_{\text{ret}}(\phi, q) = u(\phi) - Ts(\phi, q)$ and compare it directly with the noise floor energy u_{noise}. Retrieval is thermodynamically stable if and only if:

$$\min_{\phi, q}\left[u(\phi) - Ts(\phi, q)\right] \leq u_{\text{noise}}. \tag{22}$$

As T increases, the entropy term grows, raising the retrieval free energy. As α increases, u_{noise} drops. The retrieval-to-disorder transition occurs at the critical temperature T_c where these values cross. Because the LSE and LSR energy wells are sharply peaked, this transition is first-order. We now derive u_{noise} for each kernel.

3.3.1. LSE Noise Floor u_{noise}

We first consider the noise floor induced by globally supported kernels. To derive the noise floor energy, we evaluate the partition function of the M non-retrieved patterns. For a random state uncorrelated with any memory, the alignment follows a Gaussian distribution $P(\phi) \sim \exp(-N\phi^2/2)$. The expected number of patterns at alignment ϕ is $n(\phi) = MP(\phi) = \exp(N[\alpha - \phi^2/2])$. Setting $n(\phi) \sim 1$ gives the maximum spurious alignment:

$$\phi_{\max} = \sqrt{2\alpha} \leq 1\,. \tag{23}$$

For the LSE kernel, the noise floor is determined by:

$$p_{\text{noise}} = \sum_{\mu=1}^{M} \exp\left(-\beta_{\text{net}} N(1 - \phi^{\mu})\right) \approx$$

$$\int d\phi \exp\left(N\left[\alpha - \beta_{\text{net}}(1 - \phi) - \frac{\phi^2}{2}\right]\right). \quad (24)$$

The internal energy density $u_{\text{noise}} = -[N\beta_{\text{net}}]^{-1} \ln p_{\text{noise}}$ is dominated by the maximum of the exponent:

$$g(\phi) = \alpha - \beta_{\text{net}} u(\phi) - \phi^2/2, \quad (25)$$

which yields a saddle point at $\phi^* = \beta_{\text{net}}$, constrained by (23). This leads to two regimes:

1. **Spin-glass regime** ($\beta_{\text{net}} > \sqrt{2\alpha}$): The saddle point lies beyond $\phi_{\max}$, so the integral is dominated by the maximum spurious alignment:

$$u_{\text{noise}} = 1 - \sqrt{2\alpha}. \quad (26)$$

2. **Paramagnetic regime** ($\beta_{\text{net}} < \sqrt{2\alpha}$): The saddle point $\phi^* = \beta_{\text{net}}$ lies within the available range, giving:

$$u_{\text{noise}} = 1 - \frac{\alpha}{\beta_{\text{net}}} - \frac{\beta_{\text{net}}}{2}. \quad (27)$$

3.3.2. LSR Noise Floor u_{noise}

For the Epanechnikov (LSR) kernel, the derivation must account for finite support: a pattern μ contributes to the noise only if $\phi^{\mu} > 1 - 1/b$.

The maximum spurious alignment is again $\phi_{\max} = \sqrt{2\alpha}$. However, a noise floor exists only if $\phi_{\max}$ exceeds the support boundary, defining a support threshold:

$$\alpha_{\text{th}} = \frac{1}{2}(1 - 1/b)^2. \quad (28)$$

For $\alpha < \alpha_{\text{th}}$, no random patterns fall within the support and no noise floor forms. For $\alpha > \alpha_{\text{th}}$, following the same REM logic as for LSE, the saddle point ϕ^* satisfies:

$$b = f(\phi^*) = \frac{\phi^*}{1 + \phi^* - (\phi^*)^2}. \quad (29)$$

This leads to two regimes:

1. **Spin-glass regime** ($b > f(\phi_{\max})$): The saddle point lies beyond $\phi_{\max}$, so the integral is dominated by the maximum spurious alignment:

$$u_{\text{noise}} = -b^{-1} \ln\left(1 - b(1 - \sqrt{2\alpha})\right). \quad (30)$$

Unlike LSE, this energy diverges as $\sqrt{2\alpha} \to 1 - 1/b$.

2. **Paramagnetic regime** ($b < f(\phi_{\max})$): The saddle point lies within the available range:

$$u_{\text{noise}} = -b^{-1} \ln\left(1 - b(1 - \phi^*)\right). \quad (31)$$

4. Retrieval–Spin-Glass Transition

This section turns the free-energy formulation from Section 3 into explicit phase boundaries that predict when retrieval remains stable under noise and interference. We focus only on the transition between retrieval and spin-glass phases. Before deriving phase boundaries, we clarify the optimization problem that determines the network's steady state.

In standard thermodynamics, equilibrium minimizes the free energy. However, for disordered systems analyzed via the replica method, the physical equilibrium corresponds to a saddle point of $f(\phi, q)$. This arises from the interplay between the thermodynamic limit ($N \to \infty$) and the replica limit ($n \to 0$).

The replicated partition function $\langle Z^n \rangle$ is evaluated as $N \to \infty$ using steepest descent. For fixed $n > 1$, the integral is dominated by the minimum of $f(\phi, q)$. However, obtaining the physical free energy requires analytic continuation to $n \to 0$. As n passes below 1, the number of independent off-diagonal elements in the overlap matrix, $n(n-1)/2$, becomes negative. This leads to a well-known inversion of stability criteria: the physical stationary point is a maximum with respect to q but a minimum with respect to ϕ (de Almeida & Thouless, 1978). Here ϕ sets the mean alignment with the target memory, while q controls the spread of thermal fluctuations around that mean on the sphere.

This saddle-point structure has geometric significance. The parameter q measures the tightness of the thermal cloud on the N-sphere, subject to the constraint $q \geq \phi^2$ (Section 3.2). Naively minimizing $f = u - Ts$ over q would drive $q \to 1$, implying that the state collapses to a single point regardless of temperature, which is inconsistent with finite-temperature behavior.

Instead, we seek the stationary point $\partial f/\partial q = 0$. For both LSE and LSR kernels, this occurs at

$$q = \phi^2, \quad (32)$$

where the geometric entropy $s(\phi, q)$ is maximized for given ϕ. Physically, this corresponds to isotropic thermal fluctuations in the subspace perpendicular to the target pattern. Evaluating the system at this saddle point correctly captures the entropic pressure leading to the first-order transition at T_c.

4.1. LSE Spin-Glass Transition

We first derive the equilibrium alignment and phase boundary for the globally supported LSE (Gaussian) kernel.

Minimizing $f_{\text{ret}}(\phi, q)$ using (13) and (21) over ϕ yields the

physical solution:

$$\phi_{\text{LSE}}(T) = \frac{1}{2}\Big[-T + \sqrt{T^2 + 4}\,\Big], \qquad (33)$$

satisfying $\phi \to 1$ as $T \to 0$, with corresponding free energy:

$$f_{\text{ret}}^{\text{LSE}}(T) = 1 - \phi_{\text{LSE}}(T) - \frac{T}{2}\ln\Big[1 - \phi_{\text{LSE}}^2(T)\Big]. \quad (34)$$

The phase boundary occurs when equality holds in (22). In the spin-glass regime:

$$\alpha_c^{\text{LSE}}(T) = \frac{1}{2}\Big[1 - f_{\text{ret}}^{\text{LSE}}(T)\Big]^2. \qquad (35)$$

As $T \to \infty$, $\alpha_c \to 0$, so LSE permits retrieval at arbitrarily high temperatures for sufficiently small α. In the limit $T \to 0$, one recovers the well-known result $\alpha_c = 0.5$ by Ramsauer et al. (2021).

4.2. LSR Spin-Glass Transition

Similarly, minimizing $f_{\text{ret}}(\phi, q)$ using (14) and (21) over ϕ yields:

$$-\frac{1}{1 - b(1 - \phi)} + \frac{T\phi}{1 - \phi^2} = 0. \qquad (36)$$

With $y = 1 - \phi$, this becomes a quadratic equation:

$$(bT + 1)y^2 - (2 + T + Tb)y + T = 0. \qquad (37)$$

For $b = 1$, the solution simplifies to $\phi_{\text{LSR}}(T) = 1/\sqrt{1 + T}$. For general $b > 1$, the solution must be obtained numerically from (37).

The finite support of the LSR kernel introduces a support boundary $\phi > 1 - 1/b$ and a corresponding threshold:

$$\alpha_{\text{th}}(b) = \frac{1}{2}\Big[1 - b^{-1}\Big]^2. \qquad (38)$$

For $\alpha < \alpha_{\text{th}}$, no spurious patterns fall within the kernel's support, providing enhanced retrieval stability. Above the threshold, the phase boundary with the spin-glass regime is:

$$\alpha_c^{\text{LSR}}(T) = \frac{1}{2}\Big[1 - f_{\text{ret}}^{\text{LSR}}(T)\Big]^2, \qquad (39)$$

as for LSE.

The key qualitative advantage of LSR with $b > 1$ emerges below the support threshold $\alpha_{\text{th}}(b) = \frac{1}{2}(1 - b^{-1})^2$. When $\alpha < \alpha_{\text{th}}$, no spurious patterns fall within the kernel's support region $\phi > 1 - 1/b$, and the REM-derived noise floor does not exist. The retrieval basin is completely isolated from interference, enabling perfect retrieval at any temperature. This sub-threshold regime has no analog in LSE, where interference from spurious patterns is always present regardless of load. While LSE permits retrieval at arbitrarily high temperatures for low loads, this robustness coexists with ever-present noise; LSR below threshold eliminates interference entirely.

4.3. Phase Diagrams and Monte Carlo Validation

Fig. 1 shows the phase diagrams in the (α, T) plane for LSE (left) and LSR (right) kernels. For LSR, we choose $b = 3.41$, which yields a support threshold $\alpha_{\text{th}} = 0.25$ below which retrieval is perfect at any temperature. This value of b illustrates the qualitative advantage of finite-support kernels while maintaining a substantial retrieval region above threshold. The retrieval region (shaded) corresponds to parameter values where the system maintains macroscopic alignment with a target pattern. Three distinct phases emerge:

1. **Retrieval phase** (low α, low T): The system converges to a state with high alignment $\phi \approx 1$ with the target memory. For LSE, this phase is bounded from above by the critical line $\alpha_c(T)$ at all temperatures. For LSR with $b > 1$, retrieval below the support threshold α_{th} is perfect at any temperature, as no spurious patterns fall within the kernel's support.

2. **Spin-glass phase** (high α, low T): When the memory load exceeds the critical value, interference from exponentially many stored patterns destabilizes retrieval. The system becomes trapped in disordered states with no macroscopic alignment to any single pattern.

3. **Paramagnetic phase** (high T): Thermal fluctuations dominate, and the system explores the sphere uniformly without preference for any memory.

The critical lines $\alpha_c(T)$ differ qualitatively between kernels. For LSE (solid curve), the retrieval region extends to arbitrarily high temperatures as $\alpha \to 0$. For LSR with $b = 3.41$ (right panel), the critical line terminates at the support threshold $\alpha_{\text{th}} = 0.25$. Below this threshold, the retrieval basin is completely isolated from spurious pattern interference, and retrieval is perfect at any temperature.

Figure 2 compares equilibrium properties as functions of temperature. The left panel shows the equilibrium alignment $\phi(T)$ obtained by minimizing the retrieval free energy, along with Monte Carlo validation (symbols, $N = 50$, $\alpha = 0.1$). For LSR with $b = 3.41$, the load $\alpha = 0.1$ lies below the support threshold $\alpha_{\text{th}} = 0.25$, and accordingly the system remains in the retrieval state across the full temperature range shown. For LSE, the alignment drops sharply around $T \approx 0.7$–0.8, reflecting the retrieval-to-disorder transition at this load. The right panel shows the corresponding retrieval free energies $f_{\text{ret}}(T)$.

Monte Carlo validation. These simulations provide a qualitative validation of the finite-temperature trends at low memory load, rather than a quantitative finite-size analysis. To verify the theoretical predictions, we performed Metropolis–Hastings simulations on networks of
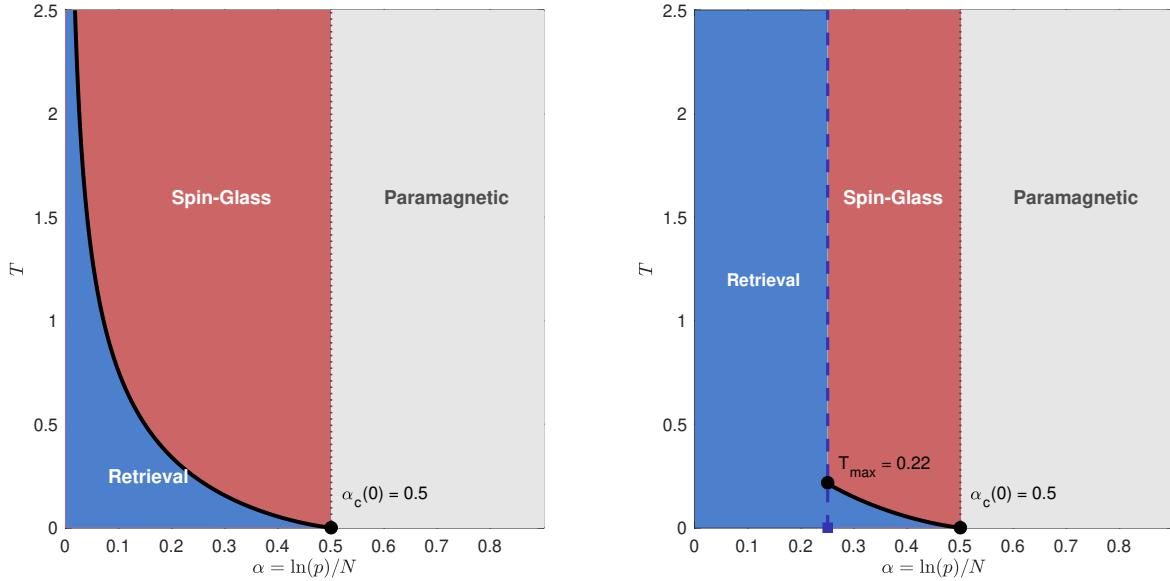
Figure 1. Phase diagrams for spherical DAM with exponential capacity $p = e^{\alpha N}$. **Left:** LSE kernel. The retrieval region (blue) permits retrieval at arbitrarily high temperatures for sufficiently low α. **Right:** LSR kernel with $b = 3.41$. The retrieval region exhibits a support threshold $\alpha_{\mathrm{th}} = 0.25$. Below this threshold, retrieval is perfect at any temperature. Above threshold, the critical line $\alpha_c(T)$ bounds the retrieval region, reaching $T_{\max} \approx 0.22$ at $\alpha = \alpha_{\mathrm{th}}$. Both kernels achieve $\alpha_c(0) = 0.5$ at zero temperature. The vertical boundary at $\alpha = 0.5$ separates the spin-glass phase from the paramagnetic phase.

size $N = 50$ with memory load $\alpha = 0.1$. For LSR with $b = 3.41$, this load lies below the support threshold $\alpha_{\mathrm{th}} = 0.25$, allowing us to test the predicted regime of perfect retrieval at any temperature. For LSE at the same load, the system should exhibit a retrieval-to-disorder transition at finite temperature. Patterns were sampled uniformly on the N-sphere, and the system was initialized near the target pattern. At each temperature, we equilibrated for 1500 steps followed by 1000 sampling steps, measuring the alignment $\phi = N^{-1}\mathbf{x} \cdot \boldsymbol{\xi}^1$ with the target.

The Monte Carlo results (symbols in Figure 2, left panel) show good agreement with theory. For LSE, the measured alignment follows the theoretical curve (33) and then drops sharply around $T \approx 0.7$–0.8, reflecting the retrieval-to-disorder transition where interference from spurious patterns destabilizes the retrieval basin.

For LSR with $b = 3.41$, the load $\alpha = 0.1$ lies below the support threshold $\alpha_{\mathrm{th}} = 0.25$. As predicted, the system remains in the retrieval state across the full temperature range, confirming that no spurious patterns fall within the kernel's support and retrieval is perfect at any temperature. For LSE at the same load, the alignment.

These simulations confirm the central qualitative prediction: LSR below threshold provides complete isolation from interference, while LSE exhibits ever-present noise that eventually destabilizes retrieval at sufficiently high temperature.

5. Conclusion

We have derived the thermodynamic phase boundaries for Dense Associative Memory networks operating in the exponential capacity regime $p = e^{\alpha N}$. By separating the free energy into kernel-dependent energy and geometry-dependent entropy, we obtained analytic expressions for the critical line $\alpha_c(T)$ that delineates the retrieval phase from disordered (spin-glass or paramagnetic) phases.

The two kernels exhibit qualitatively different phase boundary structures. For LSE, the retrieval region extends to arbitrarily high temperatures as $\alpha \to 0$, but interference from spurious patterns is always present. For LSR with $b > 1$, the finite support introduces a threshold $\alpha_{\mathrm{th}}(b)$ below which no spurious patterns contribute to the noise floor. In this sub-threshold regime, retrieval is perfect at any temperature, a qualitative advantage with no analog in LSE.

The practical implications are twofold. First, kernel choice involves a robustness-capacity trade-off: LSE provides thermal robustness across all loads with ever-present interference, while LSR offers complete isolation from interference below the support threshold, guaranteeing perfect retrieval regardless of temperature in this regime. Second, both kernels achieve the same zero-temperature capacity $\alpha_c(0) = 0.5$, confirming that exponential storage capacity is a geometric property of the spherical constraint rather than a kernel-specific feature.
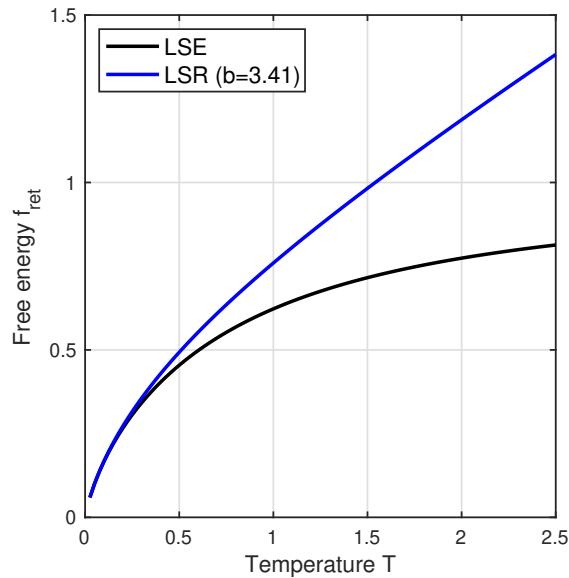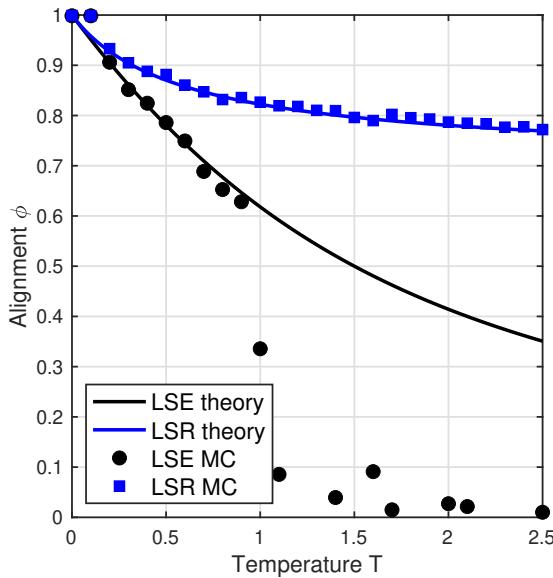
Figure 2. Equilibrium properties versus temperature. **Left:** Alignment $\phi(T)$ from theory (curves) and Monte Carlo simulation (symbols, $N = 50$, $\alpha = 0.1$). For LSR with $b = 3.41$, the system remains in the retrieval state across the temperature range shown, consistent with the phase boundary in Fig. 1 (right). For LSE, alignment drops sharply around $T \approx 0.7$–0.8, reflecting the phase transition at this load. **Right:** Retrieval free energy $f_{\text{ret}}(T)$ for LSE and LSR ($b = 3.41$).

Impact Statement

This paper analytically characterizes the thermodynamic stability of retrieval in continuous Dense Associative Memory networks on the N-sphere under finite temperature and exponential memory load. By separating kernel-dependent energy from geometry-induced entropy, it clarifies fundamental limits on retrieval robustness and explains how modeling choices, such as kernel support and sharpness, qualitatively affect stability under noise.

The work is purely theoretical: it introduces no new datasets, user-facing systems, or deployment procedures, and involves no personal data or human-subject experiments. Its broader impact is indirect. By providing a principled understanding of when and why high-capacity associative memories remain stable or fail, the results can inform the theoretical analysis and design of memory and attention-like mechanisms in machine learning. The societal implications are comparable to those of general advances in machine learning theory.

References

Amit, D. J., Gutfreund, H., and Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530–1533, 1985.

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

de Almeida, J. R. L. and Thouless, D. J. Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11 (5):983, may 1978. doi: 10.1088/0305-4470/11/5/028. URL https://doi.org/10.1088/0305-4470/11/5/028.

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

Derrida, B. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5): 2613, 1980.

Edwards, S. F. and Anderson, P. W. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

Gardner, E. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257, 1988.

Hoover, B., Shi, Z., Balasubramanian, K., Krotov, D., and Ram, P. Dense associative memory with epanechnikov energy. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ZbQ5Zq3zA3.

Hopfield, J. J. 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.

Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. *arXiv preprint arXiv:1606.01164*, 2016a. URL https://arxiv.org/abs/1606.01164.

Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems (NIPS)*, volume 29, pp. 1172–1180, 2016b.

Krotov, D. and Hopfield, J. J. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=X4y_10OX-hX.

Lucibello, C. and Mézard, M. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7):077301, 2024. doi: 10.1103/PhysRevLett.132.077301.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D., Kopp, M. K., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Rooke, S., Krotov, D., Balasubramanian, V., and Wolpert, D. Stochastic thermodynamics of associative memory. *arXiv preprint arXiv:2601.01253*, 2026.

Sherrington, D. and Kirkpatrick, S. Solvable model of a spin-glass. *Physical Review Letters*, 35(26):1792, 1975.

A. Derivation of the Free Energy Density

The disorder-averaged replicated partition function is:

$$\langle Z^n \rangle = \int \prod_{a=1}^{n} \mathrm{d}\boldsymbol{x}_a \, \delta(|\boldsymbol{x}_a|^2 - N) \left\langle \exp\left(-\beta \sum_{a=1}^{n} H(\boldsymbol{x}^a, \boldsymbol{\xi}^\mu)\right) \right\rangle . \tag{40}$$

Let $\phi_a = \boldsymbol{x}_a \cdot \boldsymbol{\xi}/N$ denote the alignment of replica a with a target memory $\boldsymbol{\xi}$, and introduce the overlap matrix:

$$q_{ab} = \frac{1}{N} \boldsymbol{x}_a \cdot \boldsymbol{x}_b . \tag{41}$$

On the N-sphere, diagonal elements are fixed: $q_{aa} = 1$. The off-diagonal elements q_{ab} ($a \neq b$) correspond to the Edwards–Anderson order parameter (Edwards & Anderson, 1975), measuring similarity between replica configurations.

Inserting the identity

$$1 = \int \prod_a \mathrm{d}\phi_a \delta\left(\phi_a - \boldsymbol{x}_a \cdot \boldsymbol{\xi}/N\right) \int \prod_{b<c} \mathrm{d}q_{bc} \delta\left(q_{bc} - \boldsymbol{x}_b \cdot \boldsymbol{x}_c/N\right) \tag{42}$$

into (40) and rearranging:

$$\langle Z^n \rangle = \int \prod_a \mathrm{d}\phi_a \int \prod_{b<c} \mathrm{d}q_{bc} \left\langle \left(\int \prod_d \mathrm{d}\boldsymbol{x}_d \, \delta(\|\boldsymbol{x}_d\|^2 - N)\delta(\phi_a - \boldsymbol{x}_a \cdot \boldsymbol{\xi}/N) \times \right.\right.$$

$$\left.\left. \delta(q_{bc} - \boldsymbol{x}_b \cdot \boldsymbol{x}_c/N) \exp\left(-\beta \sum_d H(\boldsymbol{x}_d, \boldsymbol{\xi}^\mu)\right) \right) \right\rangle . \tag{43}$$

In the retrieval basin of $\boldsymbol{\xi}$, the density $p(\boldsymbol{x})$ (4) is dominated by the target memory:

$$p(\boldsymbol{x}) \approx \exp\left[-\frac{\beta_{\mathrm{net}}}{2} d^2(\boldsymbol{x}, \boldsymbol{\xi})\right] . \tag{44}$$

The microscopic Hamiltonian can then be replaced by a macroscopic potential depending only on ϕ_a:

$$H(\boldsymbol{x}^a, \boldsymbol{\xi}^\mu) \to \mathcal{U}(\phi_a) = -\frac{1}{\beta_{\mathrm{net}}} \ln \exp\left(-\frac{N(1 - \phi_a)}{\sigma^2}\right) = N(1 - \phi_a) \tag{45}$$

for the LSE kernel. Since $\mathcal{U}(\phi_a)$ no longer depends on $\boldsymbol{x}_a$, it factors out of the microscopic integral:

$$\langle Z^n \rangle = \int \prod_a \mathrm{d}\phi_a \int \prod_{b<c} \mathrm{d}q_{bc} \exp\left(-\beta \sum_a \mathcal{U}(\phi_a)\right) \Omega(\phi_a, q_{bc}) , \tag{46}$$

where $\Omega(\phi_a, q_{bc})$ is the configuration space volume satisfying the macroscopic constraints:

$$\Omega(\phi_a, q_{bc}) = \left\langle \int \prod_d \mathrm{d}\boldsymbol{x}_d \, \delta(\|\boldsymbol{x}_d\|^2 - N) \prod_a \delta(\phi_a - \boldsymbol{x}_a \cdot \boldsymbol{\xi}/N) \prod_{b<c} \delta(q_{bc} - \boldsymbol{x}_b \cdot \boldsymbol{x}_c/N) \right\rangle . \tag{47}$$

This purely geometric quantity counts the arrangements of n vectors on an N-sphere with fixed norms $\sqrt{N}$, projections ϕ_a onto the memory, and mutual overlaps q_{bc}.

With $u(\phi) = \mathcal{U}(\phi)/N$ and defining the entropy density $s(\phi_a, q_{bc}) = \lim_{N \to \infty} N^{-1} \ln \Omega$:

$$\langle Z^n \rangle \approx \int \prod_a \mathrm{d}\phi_a \int \prod_{b<c} \mathrm{d}q_{bc} \exp\left(N\left[-\beta \sum_a u(\phi_a) + s(\phi_a, q_{bc})\right]\right) . \tag{48}$$

Under the replica symmetric ansatz (Amit et al., 1987), $\phi_a = \phi$ and $q_{ab} = q$ for all $a \neq b$, reducing the integral to two order parameters:

$$\langle Z^n \rangle \approx \int \mathrm{d}\phi \int \mathrm{d}q \exp\left(Nn\left[-\beta u(\phi) + s(\phi, q)\right]\right) . \tag{49}$$

To evaluate $s(\phi, q)$, we follow Gardner (1988). Decomposing each replica into components parallel and perpendicular to the target: $\boldsymbol{x}^a = \boldsymbol{x}_\parallel^a + \boldsymbol{x}_\perp^a$. The alignment constraint fixes $\boldsymbol{x}_\parallel^a = \sqrt{N}\phi\hat{\boldsymbol{\xi}}$. The overlap between replicas a and b is:

$$q = \frac{1}{N}\left(\boldsymbol{x}_\parallel^a \cdot \boldsymbol{x}_\parallel^b + \boldsymbol{x}_\perp^a \cdot \boldsymbol{x}_\perp^b\right) = \phi^2 + \frac{1}{N}\left(\boldsymbol{x}_\perp^a \cdot \boldsymbol{x}_\perp^b\right). \tag{50}$$

Since the perpendicular overlap is non-negative, this implies the geometric constraint $q \geq \phi^2$.

Gardner's calculation gives the log-volume for an isotropic cloud on the N-sphere as $\frac{1}{2}\ln(1-q) + q/[2(1-q)]$. With alignment ϕ constraining the longitudinal direction, the perpendicular overlap is $q_\perp = q - \phi^2$. Substituting into Gardner's result yields:

$$s(\phi, q) = \frac{1}{2}\left[\ln(1-q) + \frac{q - \phi^2}{1-q}\right]. \tag{51}$$

In the thermodynamic limit, the integral is dominated by the saddle point satisfying:

$$\beta\frac{\partial u}{\partial \phi} + \frac{\phi}{1-q} = 0, \qquad \frac{1}{1-q} - \frac{1-\phi^2}{(1-q)^2} = 0. \tag{52}$$

The second equation gives $q = \phi^2$: the thermal cloud tightness is determined by the alignment. Thus:

$$\langle Z^n \rangle \approx \exp\left(Nn[-\beta u(\phi) + s(\phi, q)]\right). \tag{53}$$

Applying the replica identity (11):

$$\langle f \rangle = -\frac{1}{N\beta}\langle \ln Z \rangle \approx u(\phi) - Ts(\phi, q). \tag{54}$$

B. Derivation of the Internal Energy Density for the LSR Kernel

To determine the internal energy density $u(\phi)$ for the LSR Hamiltonian (7), we must account for the scaling of Euclidean distances and kernel parameters in the limit $N \to \infty$. The distance is given by $d^2(\boldsymbol{x}, \boldsymbol{\xi}^\mu) = 2N(1 - \phi^\mu)$.

On the N-sphere, the volume of a basin with fixed Euclidean width σ^2 vanishes as $N \to \infty$. For the retrieval basin to span a macroscopic range of alignments $\phi \in [0, 1]$, the kernel width σ^2 must scale linearly with N. We therefore define the rescaled inverse variance:

$$b = N\beta_{\text{net}}. \tag{55}$$

This scaling ensures that the internal energy H is $O(N)$, so the energy density $u = H/N$ remains $O(1)$ and competes meaningfully with the geometric entropy s. This approach follows the scaling used in Hoover et al. (2025) and makes the LSR internal energy directly comparable to LSE, with both energy basins having identical slopes at perfect retrieval ($\phi = 1$).

Assuming the system is within the retrieval basin of pattern $\boldsymbol{\xi}$, the Hamiltonian sum is dominated by a single term. Substituting the rescaled parameters into (7):

$$H_{\text{LSR}}(\phi) \approx -\frac{N}{b}\ln\left[1 - b(1 - \phi)\right]. \tag{56}$$

Dividing by N gives the internal energy density:

$$u_{\text{LSR}}(\phi) = -b^{-1}\ln\left[1 - b(1 - \phi)\right]. \tag{57}$$

The Taylor expansion near perfect retrieval ($\phi \to 1$) gives:

$$u_{\text{LSR}}(\phi) \approx (1 - \phi) + \frac{b}{2}(1 - \phi)^2 + \cdots \tag{58}$$

Thus LSR matches the linear LSE behavior to first order, but introduces a quadratic correction proportional to the sharpness b.