CAN LOCAL ENERGY GEOMETRY PREDICT PER-PATTERN RETRIEVAL RELIABILITY IN DENSE ASSOCIATIVE MEMORIES?

Tatiana Petrova
Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg
tatiana.petrova@uni.lu

ABSTRACT

Capacity analyses of dense associative memories (DAMs) characterize global phase transitions but cannot predict which individual patterns will fail retrieval in a given finite-size system. We propose the basin isolation metric $I_\mu(\sigma)$, a Hessian-free diagnostic that measures the anharmonicity of the energy landscape around each stored pattern by probing radial energy profiles along random tangent directions. Evaluating on a spherical DAM with cubic interactions (n=3) across $N \in \{100, 200, 500, 1000\}$ in the near-transition regime, we find that at $N \leq 200$, I_μ outperforms pairwise overlap baselines (AUC-ROC up to 0.68), is reasonably robust to its scale parameter, and captures nonlinear geometric information not fully captured by simple overlap statistics. However, with a fixed number of probing directions K, the diagnostic degrades at $N \geq 500$, consistent with random tangent sampling becoming increasingly sparse relative to the growing tangent-space dimensionality. These results provide a geometric perspective on per-pattern retrieval variability and clarify the regime where local landscape probing remains informative.

1 INTRODUCTION

Associative memories store patterns as attractors of an energy-based dynamical system and retrieve them via energy minimization (Hopfield, 1982). Dense associative memories (DAMs) with higher-order polynomial interactions achieve storage capacity $M \sim N^{n-1}$ for interaction order n, vastly surpassing the classical $M \sim 0.14N$ limit (Krotov & Hopfield, 2016; Demircigil et al., 2017). These models underpin modern Hopfield networks (Ramsauer et al., 2020) and related memory-augmented architectures. As these systems scale, per-pattern reliability (knowing *which* stored patterns can be retrieved and which cannot) becomes a question of direct practical importance.

Existing theoretical analyses characterize *global* phase transitions (critical loads M^* above which typical retrieval fails) but treat all patterns as statistically equivalent (Amit et al., 1985; Lucibello & Mézard, 2023). Yet even below M^*, individual patterns vary substantially in retrieval robustness depending on their local energy landscape. Per-pattern reliability prediction would enable confidence estimation, selective consolidation of fragile memories (Fachechi et al., 2019), and targeted error correction.

Related work. Classical signal-to-noise analysis (Amit et al., 1985), scaling laws (Cabannes et al., 2023), and replica-theoretic approaches (Lucibello & Mézard, 2023) characterize average-case or asymptotic capacity without yielding finite-size per-pattern diagnostics. The Hessian spectrum at each minimum provides a direct local characterization but requires $O(N^2)$ storage and $O(N^3)$ eigendecomposition. Pairwise overlap statistics are inexpensive but capture only linear interference, missing the nonlinear basin deformations that polynomial energies create. To our knowledge, no prior work has proposed a Hessian-free, per-pattern geometric diagnostic for retrieval reliability in finite-size DAMs.

Contribution. We introduce the basin isolation metric $I_\mu(\sigma)$, which quantifies the *anharmonicity* of the energy landscape around each stored pattern by measuring how radial energy profiles deviate

Algorithm 1 Basin Isolation Metric $I_\mu(\sigma)$

Require: Pattern $\boldsymbol{\xi}^\mu$, all patterns $\boldsymbol{\Xi}$, radius σ, directions K, grid points n_{pts}
for $k = 1, \dots, K$ **do**
  Sample $\mathbf{v}_k$ uniformly on the unit sphere in the tangent space at $\boldsymbol{\xi}^\mu$
  $\{r_j\}_{j=0}^{n_{\text{pts}}-1} \leftarrow$ equispaced grid from 0 to σ ▷ includes r_0=0
  for $j = 0, \dots, n_{\text{pts}} - 1$ **do**
    $e_k(r_j) \leftarrow H\big(\text{Proj}_S(\boldsymbol{\xi}^\mu + r_j \mathbf{v}_k)\big)$ ▷ Eq. equation 1
  end for
  Fit $\hat{e}_k(r) = e_k(0) + \frac{1}{2}\hat{\kappa}_k r^2$ to first 10 points ▷ least squares
  Approximate $D_k = \sigma^{-1}\int_0^\sigma |e_k(r) - \hat{e}_k(r)|\, dr$ by the trapezoidal rule on $\{r_j\}$
end for
return $I_\mu(\sigma) = \frac{1}{K}\sum_{k=1}^K D_k$

from a locally-fitted quadratic. Through systematic experiments across four dimensions, we establish three findings: **(1)** I_μ outperforms overlap baselines at moderate N, capturing nonlinear basin deformation not fully reflected in pairwise overlap statistics; **(2)** at moderate N, the diagnostic is reasonably robust to its scale parameter, with the heuristic $\sigma \approx d_{\text{typ}}/3$ providing an effective default; and **(3)** all tested diagnostics degrade toward chance level at high N, suggesting a limitation of fixed-budget scalar diagnostics in this near-transition regime.

2 METHOD

2.1 MODEL

We consider a continuous DAM on the sphere $\|\mathbf{x}\|^2 = N$ with energy

$$H(\mathbf{x}) = -\frac{1}{n}\sum_{\mu=1}^{M}\left(\frac{\mathbf{x}\cdot\boldsymbol{\xi}^\mu}{N}\right)^n, \tag{1}$$

where $\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^M$ are i.i.d. uniform on $S^{N-1}(\sqrt{N})$ and n=3. Denoting the $N \times M$ pattern matrix $\boldsymbol{\Xi} = [\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^M]$ and the projection onto the sphere $\text{Proj}_S(\mathbf{z}) = \sqrt{N}\,\mathbf{z}/\|\mathbf{z}\|$, retrieval uses the iterative map $\mathbf{x}_{t+1} = \text{Proj}_S\big(\boldsymbol{\Xi}\, F'(\boldsymbol{\Xi}^\top \mathbf{x}_t/N)\big)$, with $F'(m) = m^{n-1}$ applied elementwise. For n=3, the theoretical capacity scales as $M^* \sim N^2$ (Krotov & Hopfield, 2016; Demircigil et al., 2017).

2.2 BASIN ISOLATION METRIC

Because retrieval is constrained to the sphere, the relevant local object is the tangent-space (Riemannian) Hessian. Around a constrained local minimum $\mathbf{x}^\star$, the projected energy admits a second-order expansion

$$H\big(\text{Proj}_S(\mathbf{x}^\star + \boldsymbol{\delta})\big) = H(\mathbf{x}^\star) + \tfrac{1}{2}\boldsymbol{\delta}^\top \mathcal{H}(\mathbf{x}^\star)\boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3),$$

for small $\boldsymbol{\delta} \in T_{\mathbf{x}^\star} S^{N-1}(\sqrt{N})$. Motivated by this local quadratic reference, we probe the energy around each stored pattern $\boldsymbol{\xi}^\mu$ without forming the Hessian: we sample K random tangent directions $\mathbf{v}_k$, evaluate the radial energy profile $e_k(r) = H\big(\text{Proj}_S(\boldsymbol{\xi}^\mu + r\mathbf{v}_k)\big)$ on n_{pts}=100 equispaced grid points from 0 to σ, fit a quadratic $\hat{e}_k(r) = e_k(0) + \frac{1}{2}\hat{\kappa}_k r^2$ to the first 10 points (where $\hat{\kappa}_k$ is an effective directional curvature coefficient, not an eigenvalue estimate), and compute the mean absolute deviation $D_k = \sigma^{-1}\int_0^\sigma |e_k(r) - \hat{e}_k(r)|\, dr$ approximated numerically over the grid. The metric averages over directions: $I_\mu(\sigma) = K^{-1}\sum_k D_k$. Larger $I_\mu(\sigma)$ indicates a more strongly deformed and hence less isolated local basin. The full procedure is given in Algorithm 1.

Tangent directions are sampled by drawing $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, projecting onto $T_{\boldsymbol{\xi}^\mu} S^{N-1}(\sqrt{N})$, and normalizing to unit length.

Choice of σ. For i.i.d. uniform patterns on $S^{N-1}(\sqrt{N})$, the typical inter-pattern distance is $d_{\text{typ}} = \sqrt{2N}$. We set $\sigma \approx d_{\text{typ}}/3 = \sqrt{2N}/3$, probing the "near field" of each basin, and validate this empirically in Section 3.3.

Table 1: AUC-ROC for predicting retrieval failure ($p_\mu < 0.5$). Bold: best per N; 95% bootstrap CI shown for I_μ only. J^*: Youden index for I_μ.

N	M	M/N^2	Fail%	AUC(I_μ)	AUC($\bar{O}_\mu$)	AUC($O_\mu^{\max}$)	J^*
100	300	.030	77.3	**.679** [.60,.75]	.653	.556	.301
200	1,200	.030	43.8	**.626** [.59,.66]	.587	.562	.195
500	11,000	.044	78.5	.559 [.55,.57]	**.562**	.463	.091
1000	48,000	.048	44.7	.523 [.52,.53]	**.535**	.475	.036

2.3 Baselines and evaluation

We compare I_μ against two overlap-based predictors: (i) maximal overlap $O_\mu^{\max} = \max_{\nu \neq \mu} \boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\nu / N$ (signed), and (ii) mean absolute overlap $\bar{O}_\mu = (M-1)^{-1} \sum_{\nu \neq \mu} |\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\nu / N|$.

Retrieval reliability p_μ is estimated from L=100 noisy trials. Each initialization is constructed as $\mathbf{x}_0 = \mathrm{Proj}_S(\boldsymbol{\xi}^\mu + \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is sampled isotropically in the tangent space with norm $0.3\sqrt{N}$. Iterations were terminated when $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_\infty < 10^{-7}$ or after 200 steps; success was declared when the terminal state achieved overlap >0.9 with the target pattern μ. Empirically, p_μ is near-bimodal in the regimes we study, so we evaluate all diagnostics as binary classifiers of retrieval failure ($p_\mu < 0.5$) using AUC-ROC with 1000-iteration bootstrap confidence intervals.

3 Experiments

We evaluate across $N \in \{100, 200, 500, 1000\}$ with K=100 tangent directions. For each N, M was calibrated to yield a nontrivial mixture of successes and failures (failure rates: 44–78%; Table 1). All experiments use fixed random seeds with separate RNG streams for pattern generation, metric computation, and retrieval testing.

3.1 Diagnostic performance

Table 1 summarizes the results. At $N \leq 200$, I_μ achieves the highest AUC (0.679 and 0.626), exceeding both overlap baselines. The gains are modest in absolute terms ($+0.026$ at N=100, $+0.039$ at N=200) but consistent with additional nonlinear geometric signal. The Youden index of 0.301 at N=100 indicates that I_μ should be viewed as a ranking statistic for potentially fragile memories rather than a deterministic oracle.

At $N \geq 500$, all diagnostics degrade. The mean overlap $\bar{O}_\mu$ is numerically slightly higher than I_μ (AUC 0.562 vs. 0.559 at N=500), and $O_\mu^{\max}$ falls slightly below chance (AUC < 0.5): for i.i.d. patterns in the near-transition regime with $M \sim N^2$, the per-pattern maximum overlap concentrates around $\sqrt{4 \log N / N} \to 0$ with vanishing variance, rendering it uninformative.

3.2 Classification analysis

Figure 2 shows that at N=200, I_μ dominates both baselines (AUC-PR $= 0.565$ vs. prevalence 0.438); at N=1,000, all curves collapse toward the diagonal. All AUCs move toward 0.5 at large N, while the Youden index of I_μ drops from 0.301 to 0.036, suggesting the degradation is not unique to I_μ among the scalar diagnostics considered here.

3.3 Scale sensitivity

We sweep σ over 15 values from $0.02\sqrt{2N}$ to $0.8\sqrt{2N}$ (Figure 3). At $N \leq 200$, AUC(I_μ) exhibits a broad plateau: I_μ exceeds both baselines across a wide range of σ, and the heuristic $\sigma = d_{\mathrm{typ}}/3$ falls within this plateau. Beyond $\sigma \approx d_{\mathrm{typ}}/3$, AUC degrades monotonically; at still larger scales, the Pearson correlation reverses sign, indicating a transition from local basin probing to global landscape averaging. At $N \geq 500$, I_μ never meaningfully exceeds $\bar{O}_\mu$ at any σ.

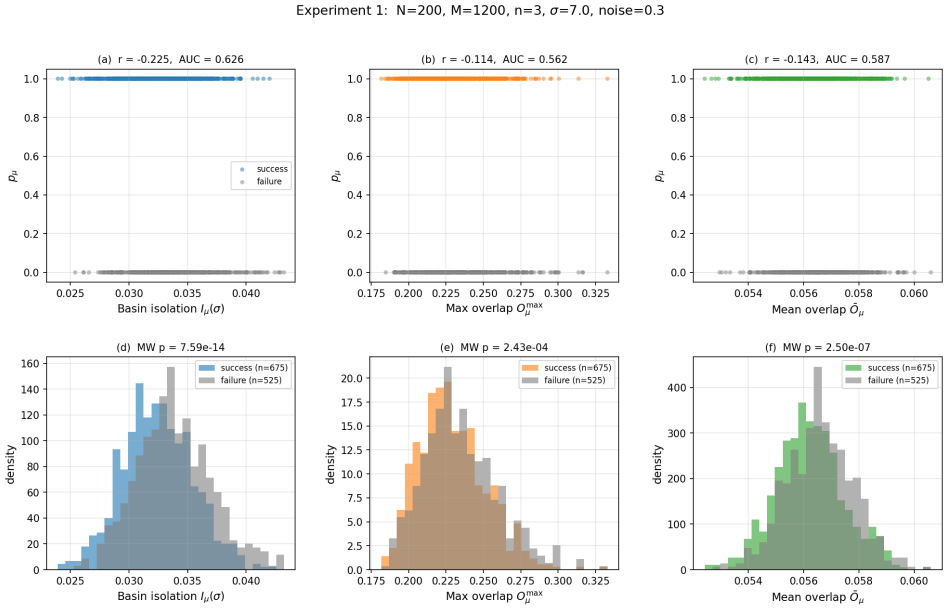

Figure 1: N=200, M=1,200. *Top:* p_μ vs. each predictor (left to right: I_μ, $O_\mu^{\max}$, $\bar{O}_\mu$). *Bottom:* predictor distributions split by retrieval success/failure. I_μ (left) shows distributional separation; $O_\mu^{\max}$ (center) shows near-complete overlap.

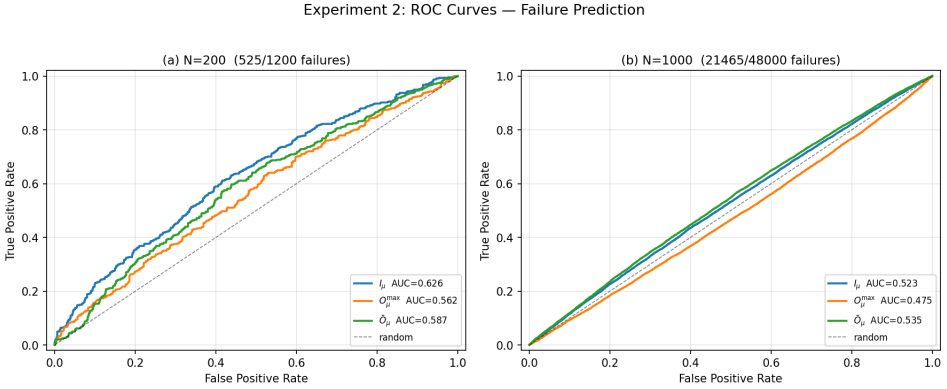

Figure 2: ROC curves at N=200 (left) and N=1,000 (right). I_μ (blue) vs. $O_\mu^{\max}$ (orange) vs. $\bar{O}_\mu$ (green). At N=200, I_μ dominates; at N=1,000, all curves approach the diagonal. Results at N=100 and N=500 are consistent (Table 1).

4 Discussion

What the metric reveals. The success of I_μ at moderate dimensions suggests that local energy anharmonicity carries predictive information beyond pairwise statistics. The polynomial energy equation 1 with n=3 generates higher-order terms encoding multi-pattern interference not fully captured by linear overlap. The transition scale $\sigma \approx d_{\text{typ}}/3$ provides an operational basin-radius proxy, otherwise difficult to define without full basin-of-attraction computation. At $N \geq 500$, performance degrades markedly. With K=100 fixed while the tangent-space dimension grows as $N-1$, directional sampling becomes progressively less informative: a fixed random set of probe directions is increasingly unlikely to intersect the narrow directions along which a basin is most strongly deformed. This degradation is not unique to I_μ: all AUCs move toward 0.5 at large N (Youden index of I_μ: 0.301 $\rightarrow$ 0.036). This suggests that near the transition in high dimensions, per-pattern variability may be governed by geometric features that these fixed-budget scalar diagnostics do not reliably capture. These results point to a limitation of fixed-budget local probing in the present near-transition setting, rather than a universal impossibility claim.

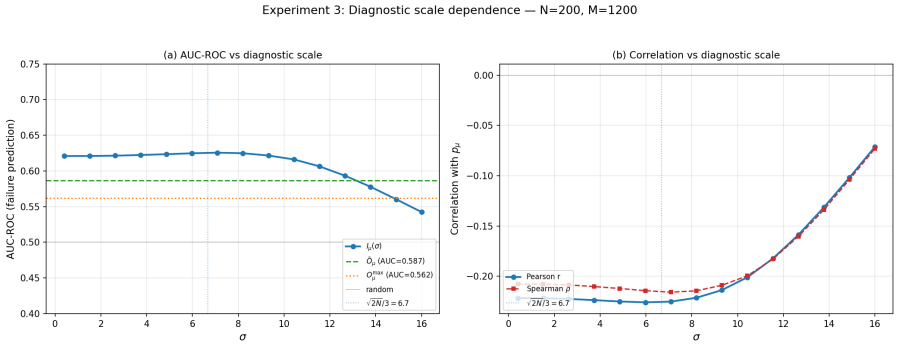

Figure 3: Scale sensitivity at N=200. *(a)* AUC vs. σ; overlap baselines as horizontal lines; dashed vertical: $\sigma = d_{\text{typ}}/3$. *(b)* Correlations weaken and eventually reverse sign at large σ.

Limitations and future work. Each N is a single disorder realization; bootstrap intervals quantify uncertainty across patterns within an instance only. Results concern the near-transition setting with one perturbation norm, one success threshold, and K=100. We test only n=3; behavior for other orders, including $n \to \infty$ (Ramsauer et al., 2020), remains open. Whether scaling K with N maintains diagnostic power and whether informed tangent directions (e.g., overlap gradients toward nearest competitors) could improve performance at large N are natural next steps.

Conclusion. We have introduced $I_\mu(\sigma)$ as a per-pattern diagnostic for retrieval reliability in DAMs, showing that it captures nonlinear geometric information not fully captured by simple overlap statistics. Our experiments establish both its utility at moderate dimensions and a practical limitation of fixed-budget local probing at high dimensions, offering a geometric perspective on retrieval variability that complements capacity-theoretic analyses.

REFERENCES

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

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. *arXiv preprint arXiv:2310.02984*, 2023.

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

Alberto Fachechi, Elena Agliari, and Adriano Barra. Dreaming neural networks: forgetting spurious memories and reinforcing pure ones. *Neural Networks*, 112:24–40, 2019.

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.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Carlo Lucibello and Marc Mézard. The exponential capacity of dense associative memories. *arXiv preprint arXiv:2304.14964*, 2023.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.