Algorithmic Analysis of Dense Associative Memory: Finite-Size Guarantees and Adversarial Robustness

Madhava Gaikwad
Independent Researcher
gaikwad.madhav@gmail.com

Abstract

Dense Associative Memory (DAM) generalizes Hopfield networks through higher-order interactions and achieves storage capacity that scales as $O(N^{n-1})$ under suitable pattern separation conditions. Existing dynamical analyses primarily study the thermodynamic limit $N \to \infty$ with randomly sampled patterns and therefore do not provide finite-size guarantees or explicit convergence rates.

We develop an algorithmic analysis of DAM retrieval dynamics that yields finite-N guarantees under explicit, verifiable pattern conditions. Under a separation assumption and a bounded-interference condition at high loading, we prove geometric convergence of asynchronous retrieval dynamics, which implies $O(\log N)$ convergence time once the trajectory enters the basin of attraction. We further establish adversarial robustness bounds expressed through an explicit margin condition that quantifies the number of corrupted bits tolerable per sweep, and derive capacity guarantees that scale as $\Theta(N^{n-1})$ up to polylogarithmic factors in the worst case, while recovering the classical $\Theta(N^{n-1})$ scaling for random pattern ensembles. Finally, we show that DAM retrieval dynamics admit a potential-game interpretation that ensures convergence to pure Nash equilibria under asynchronous updates.

Complete proofs are provided in the appendices, together with preliminary experiments that illustrate the predicted convergence, robustness, and capacity scaling behavior.

1 Introduction

Dense Associative Memory (DAM) stores p binary patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p \subset \{-1,+1\}^N$ in an N-neuron network through higher-order interactions Krotov & Hopfield (2016). For interaction order $n \geq 2$, the energy function is

$$E(\boldsymbol{x}) = -\frac{1}{N^{n-1}} \sum_{\mu=1}^{p} \left(\sum_{i=1}^{N} \xi_i^\mu x_i \right)^n , \tag{1}$$

with local field $h_i(\boldsymbol{x}) = \partial(-E)/\partial x_i$ and asynchronous update rule

$$x_i \leftarrow \text{sign}(h_i(\boldsymbol{x})).$$

Classical statistical-physics analyses of associative memory models focus primarily on asymptotic behavior in the thermodynamic limit and typically assume randomly sampled patterns. Recent work by Mimura et al. Mimura et al. (2025) provides an asymptotically exact dynamical analysis of DAM using generating functional analysis (GFA). Their results apply to the regime $N \to \infty$ with random i.i.d. patterns and show that for interaction orders

$n \geq 3$ the effective noise variance becomes independent of the overlap, which mitigates a key instability present in classical Hopfield networks ($n = 2$). However, this framework does not provide explicit finite-N convergence guarantees, does not address adversarial or structured pattern sets, and does not yield explicit retrieval-time bounds.

This work develops an algorithmic analysis that complements the statistical-physics perspective by focusing on finite systems and explicit performance guarantees. We introduce explicit separation and bounded-interference assumptions on the stored patterns that can be verified directly or shown to hold with high probability for random ensembles. Under these conditions we obtain finite-size convergence guarantees, robustness bounds under adversarial corruption, and explicit capacity guarantees that recover the classical N^{n-1} scaling up to polylogarithmic factors in the worst case.

Appendix A contains complete proofs of the main results, including the strengthened interference condition required near capacity loading, the potential-game characterization of retrieval dynamics, and verification that random pattern ensembles satisfy the assumptions with high probability. Appendix B reports preliminary experiments on a cubic ($n = 3$) DAM that illustrate convergence behavior, basin-of-attraction structure, adversarial robustness thresholds, capacity scaling, comparisons between synchronous and asynchronous updates, and retrieval performance on binarized MNIST and CIFAR-10 data.

2 Related Work

The foundational work of Hopfield (1982) established recurrent neural networks as physical systems with emergent computational abilities, a perspective deeply analyzed using spin-glass theory Amit et al. (1985); Amit (1989). The theoretical storage capacity of these classical models was rigorously quantified in seminal works Newman (1988); Gardner (1987); McEliece et al. (1987); Abbott & Arian (1987). A major leap forward was the introduction of Dense Associative Memory (DAM) models Krotov & Hopfield (2016), which achieve exponential storage capacity Lucibello & Mézard (2024) and have been shown to be practically powerful in deep learning architectures Ramsauer et al. (2021); Hoover et al. (2023). Recent theoretical advances have further explored the dynamical properties Mimura et al. (2025), saddle point hierarchies Thériault & Tantari (2026), and biological plausibility of these modern networks Kafraj et al. (2026). This resurgence has also led to novel applications in analog circuits Bacvanski et al. (2025), optical computing Musa et al. (2026), and as a framework for understanding Transformer architectures Masumura & Taki (2025). The field's enduring relevance is underscored by recent surveys and tutorials at major AI conferences Krotov et al. (2025); Kempe et al. (2024) and its connection to broader topics in optimization Boyd & Vandenberghe (2004); Shalev-Shwartz & Ben-David (2014); Goles & Martínez (1985), game theory Monderer & Shapley (1996); Hart & Mas-Colell (2003), and robust machine learning Cohen et al. (2019); Ge et al. (2015); Du et al. (2019).

3 Problem Formulation

3.1 Model and Updates

We analyze **asynchronous updates**, where at each iteration a single neuron is selected uniformly at random and updated according to

$$x_i^{(t+1)} = \operatorname{sign}\Big(h_i(\boldsymbol{x}^{(t)})\Big) \quad \text{for randomly selected } i. \tag{2}$$

Asynchronous updates ensure monotone improvement of the energy function and avoid oscillatory behaviors that may arise under synchronous updates in Hopfield-type networks Goles & Martínez (1985).

We measure time in **full sweeps**, where one sweep corresponds to N consecutive asynchronous updates so that, in expectation, each neuron is updated once.

For a stored pattern $\boldsymbol{\xi}^\nu$, the overlap of a state $\boldsymbol{x}$ with the target pattern is

$$m^\nu(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\nu x_i.$$

3.2 Pattern Separation

To obtain finite-size guarantees, we impose a separation condition that characterizes the basin of attraction of a stored pattern.

Assumption 1 (Pattern Separation)**.** *Fix a target pattern ν. There exist constants $\gamma > 0$ and $\beta < \gamma$ such that for all states $\boldsymbol{x}$ satisfying*

$$m^\nu(\boldsymbol{x}) \geq \gamma,$$

the interference from all non-target patterns is bounded as

$$\max_{\mu \neq \nu} \left| \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu x_i \right| \leq \beta.$$

This condition requires that once the trajectory enters a basin of attraction characterized by overlap at least γ, the contribution of every competing pattern remains strictly smaller than that of the target pattern. For random i.i.d. patterns, the condition holds with high probability when the loading satisfies $p = o(N^{n-1})$, while the subsequent analysis applies to any deterministic pattern set that satisfies the assumption.

4 Main Results

Theorem 2 (Convergence Rate)**.** *Suppose the stored patterns satisfy Assumption 1 with $\gamma > \beta$, and let the loading satisfy $p \leq \frac{N^{n-1}}{2n(n-1)}$. Then asynchronous DAM retrieval initialized at any state with $m^\nu(\boldsymbol{x}^{(0)}) \geq \gamma$ converges to the target pattern in*

$$T = O\left(\frac{1}{\alpha}\log N\right)$$

full sweeps, where

$$\alpha = \frac{1}{n} - \frac{2(n-1)p}{N^{n-1}} > 0.$$

Theorem 3 (Adversarial Robustness)**.** *Under the conditions of Theorem 2, retrieval succeeds despite adversarial corruption of up to ρN bits per full sweep provided*

$$\rho < \frac{\alpha}{2}.$$

In particular, using the explicit expression for α yields the bound

$$\rho < \frac{1}{2}\left(\frac{1}{n} - \frac{2(n-1)p}{N^{n-1}}\right).$$

Theorem 4 (Capacity Scaling)**.** *Assume the pattern set satisfies Assumption 1. Then DAM retrieval guarantees hold for at least*

$$p \geq \frac{N^{n-1}}{4n^2(n-1)^2}$$

stored patterns, while retrieval may fail when the loading becomes $p = \Omega(N^{n-1})$. Thus the achievable storage capacity scales as $\Theta(N^{n-1})$ up to constant or polylogarithmic factors depending on the pattern ensemble.

Theorem 5 (Game-Theoretic Characterization)**.** *DAM asynchronous dynamics coincide with best-response dynamics in an exact potential game with*

- *players: neurons* $i = 1, \dots, N$,
- *actions:* $x_i \in \{-1, +1\}$,
- *payoffs:* $u_i(\boldsymbol{x}) = x_i h_i(\boldsymbol{x})$,
- *potential:* $F(\boldsymbol{x}) = -E(\boldsymbol{x})$.

All limit points of the dynamics are pure Nash equilibria.

5 Analytical Framework

5.1 Coordinate-Descent View

DAM asynchronous updates perform coordinate ascent on the potential $F(\boldsymbol{x}) = -E(\boldsymbol{x})$.

Lemma 6 (Potential Improvement)**.** *Let coordinate i be updated asynchronously according to $x_i^{(t+1)} = \operatorname{sign}(h_i(\boldsymbol{x}^{(t)}))$. Then*

$$F(\boldsymbol{x}^{(t+1)}) - F(\boldsymbol{x}^{(t)}) = 2\,|\phi_i(\boldsymbol{x}_{-i}^{(t)})|,$$

where

$$\phi_i(\boldsymbol{x}_{-i}) = \frac{1}{2}\left(F(+1, \boldsymbol{x}_{-i}) - F(-1, \boldsymbol{x}_{-i})\right).$$

Under Assumption 1 inside the basin $m^\nu(\boldsymbol{x}) \geq \gamma$, each misaligned coordinate satisfies $|\phi_i| \geq c(\gamma, \beta, n) > 0$, implying strict potential improvement whenever an incorrect neuron is updated.

5.2 Contraction Analysis

Lemma 7 (Overlap Contraction)**.** *Under the loading condition $p \leq \frac{N^{n-1}}{2n(n-1)}$ and Assumption 1, the asynchronous dynamics satisfy*

$$\mathbb{E}[m^\nu(\boldsymbol{x}^{(t+1)})] \geq m^\nu(\boldsymbol{x}^{(t)}) + \frac{\alpha}{N}\left(1 - m^\nu(\boldsymbol{x}^{(t)})\right),$$

where

$$\alpha = \frac{1}{n} - \frac{2(n-1)p}{N^{n-1}} > 0.$$

Proof sketch. The local field decomposes as

$$h_i(\boldsymbol{x}) = n\xi_i^\nu (m^\nu)^{n-1} + \eta_i,$$

where η_i represents interference from non-target patterns. Assumption 1 ensures the signal term dominates the interference inside the basin, so that each misaligned neuron updates correctly with probability at least α. Averaging over random coordinate selection yields the stated expected improvement. □

5.3 Adversarial Analysis

Proof sketch of Theorem 3. Let $\mathcal{M}^{(t)}$ denote the set of mismatched coordinates. In one sweep, at least $\alpha|\mathcal{M}^{(t)}|$ coordinates would be corrected in expectation without adversarial corruption, while the adversary can introduce at most ρN new errors. Thus

$$|\mathcal{M}^{(t+1)}| \leq (1 - \alpha)|\mathcal{M}^{(t)}| + \rho N.$$

The recurrence contracts whenever $\rho < \alpha/2$, yielding geometric convergence to the target pattern. □

6 Game-Theoretic Interpretation

Proposition 8 (Potential Game Structure)**.** *The DAM game with utilities $u_i(\boldsymbol{x}) = x_i h_i(\boldsymbol{x})$ is an exact potential game with potential function $F(\boldsymbol{x}) = -E(\boldsymbol{x})$.*

Proof. Define the marginal potential

$$\phi_i(\boldsymbol{x}_{-i}) = \tfrac{1}{2}\left(F(+1, \boldsymbol{x}_{-i}) - F(-1, \boldsymbol{x}_{-i})\right).$$

Changing the action of player i from x_i to x_i' changes both the payoff and the potential by the same quantity $(x_i' - x_i)\phi_i(\boldsymbol{x}_{-i})$, which establishes the exact-potential property Monderer & Shapley (1996). □

7 Positioning Against Mimura et al. (2025)

Mimura et al. Mimura et al. (2025) provide an asymptotically exact characterization of DAM dynamics for random patterns in the thermodynamic limit using generating functional analysis (GFA). A key insight of their analysis is that, for $n \geq 3$, the effective noise variance becomes independent of the overlap, which explains the strong retrieval performance of higher-order models. Our work addresses a complementary regime: finite-N systems and worst-case (not necessarily random) pattern sets. Table 1 summarizes the relationship between the two approaches.

Table 1: Statistical-physics vs. algorithmic analysis of DAM

Property	Mimura et al. (2025)	This work
System size	$N \to \infty$ (thermodynamic limit)	Finite N
Pattern distribution	Random i.i.d.	Arbitrary (worst-case)
Convergence guarantee	Asymptotic dynamics	$O(\log N)$ full sweeps (basin regime)
Robustness analysis	Typical noise variance	Explicit adversarial tolerance ρ^*
Capacity result	Ensemble-averaged threshold ($\alpha'_{c,3} \approx 0.266$)	Finite-size scaling bounds with constants
Finite-size effects	Qualitative discussion	Explicit quantitative bounds

8 Conclusion

We established finite-size performance guarantees for dense associative memory dynamics that complement existing statistical-physics analyses. Under explicit separation conditions, asynchronous DAM retrieval achieves logarithmic-time convergence, admits quantitative adversarial robustness bounds, and admits a natural game-theoretic interpretation as an exact potential game. These guarantees apply to finite systems and arbitrary pattern sets, while remaining consistent with known thermodynamic-limit results for random ensembles.

Several limitations remain. First, the convergence guarantees rely on asynchronous updates; extending the analysis to synchronous dynamics requires stronger separation conditions to rule out short cycles. Second, near capacity loading $p = \Theta(N^{n-1})$, the analysis requires a componentwise interference bound (Assumption 9), since overlap-based bounds alone discard cancellations that become critical at high loading. Third, the $O(\log N)$ convergence bound is conservative at moderate system sizes: empirically, concentration of inter-pattern overlaps often causes convergence times to decrease with N rather than increase logarithmically. Finally, the adversarial tolerance bound ρ^* is derived from a worst-case contraction argument and is therefore conservative at finite N, though the gap narrows as N grows.

Future work includes tightening the interference analysis to reduce the polylogarithmic gap between worst-case and typical-case capacity, characterizing optimal interaction orders, and extending the framework to continuous modern Hopfield networks Krotov & Hopfield (2020).

9 Code Availability

All source code, scripts, and experiment configurations used in this work are publicly available at:

`https://github.com/krimler/dam-games`

The repository contains implementations of the DAM dynamics, experimental pipelines, and instructions for reproducing the results reported in this paper.

References

LF Abbott and Yair Arian. Storage capacity of generalized networks. *Physical Review A*, 36(10):5091–5094, 1987.

Daniel J Amit. *Modeling brain function: The world of attractor neural networks*. Cambridge University Press, 1989.

Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018, 1985.

Marc Gong Bacvanski, Xincheng You, John Hopfield, and Dmitry Krotov. Dense associative memories with analog circuits, 2025. URL `https://arxiv.org/abs/2512.15002`.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. *International Conference on Machine Learning*, pp. 1310–1320, 2019.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *International Conference on Machine Learning*, pp. 1675–1685, 2019.

Elizabeth Gardner. Multiconnected neural network models. *Journal of Physics A: Mathematical and General*, 20(11):3453–3464, 1987.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. *Conference on Learning Theory*, pp. 797–842, 2015.

Eric Goles and Servet Martínez. Decreasing energy functions as a tool for studying threshold networks. *Discrete Applied Mathematics*, 12(3):261–277, 1985.

Sergiu Hart and Andreu Mas-Colell. Uncoupled dynamics do not lead to nash equilibrium. *American Economic Review*, 93(5):1830–1836, 2003.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy transformer. *Advances in neural information processing systems*, 36:27532–27559, 2023.

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.

Mohadeseh Shafiei Kafraj, Dmitry Krotov, and Peter E. Latham. A biologically plausible dense associative memory with exponential capacity, 2026. URL `https://arxiv.org/abs/2601.00984`.

Julia Kempe, Dmitry Krotov, Hilde Kuehne, Daniel Lee, and Sara A Solla. New frontiers in associative memories. In *ICLR 2025 Workshop Proposals*, 2024.

Dmitry Krotov and John Hopfield. Large associative memory problem in neurobiology and machine learning. *arXiv preprint arXiv:2008.06996*, 2020.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in Neural Information Processing Systems*, 29:1172–1180, 2016.

Dmitry Krotov, Benjamin Hoover, Parikshit Ram, and Bao Pham. Modern methods in associative memory. *arXiv preprint arXiv:2507.06211*, 2025.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7):077301, 2024.

Tsubasa Masumura and Masato Taki. On the role of hidden states of modern hopfield network in transformer. *arXiv preprint arXiv:2511.20698*, 2025.

Robert J McEliece, Edward C Posner, Eugene R Rodemich, and Santosh S Venkatesh. The capacity of the hopfield associative memory. *IEEE Transactions on Information Theory*, 33(4):461–482, 1987.

Kazushi Mimura, Jun'ichi Takeuchi, Yuto Sumikawa, Yoshiyuki Kabashima, and Anthony C. C. Coolen. Dynamical properties of dense associative memory, 2025. URL `https://arxiv.org/abs/2506.00851`. Accepted at International Conference on Learning Representations (ICLR) 2026.

Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14 (1):124–143, 1996.

Khalid Musa, Santosh Kumar, Michael Katidis, and Yu-Ping Huang. Dense associative memory in a nonlinear-optical hopfield neural network. *Physical Review Applied*, 25(1): 014011, 2026.

Charles M Newman. Memory capacity in neural network models: Rigorous lower bounds. *Neural Networks*, 1(3):223–238, 1988.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need, 2021. URL `https://arxiv.org/abs/2008.02217`.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Robin Thériault and Daniele Tantari. Saddle hierarchy in dense associative memory. *Machine Learning: Science and Technology*, 7(1):015001, January 2026. ISSN 2632-2153. doi: 10.1088/2632-2153/ae3051. URL `http://dx.doi.org/10.1088/2632-2153/ae3051`.

A Proofs of Main Results

A.1 Notation and Local Field Decomposition

The energy is $E(\boldsymbol{x}) = -N^{-(n-1)} \sum_{\mu=1}^{p} (M^\mu)^n$ where $M^\mu = \sum_{i=1}^{N} \xi_i^\mu x_i = N m^\mu$ is the unnormalized overlap. The potential is $F(\boldsymbol{x}) = -E(\boldsymbol{x})$.

Formal derivative vs. discrete marginal field. A commonly used "local field" expression is obtained by treating x_i as continuous and differentiating:

$$h_i(\boldsymbol{x}) = \frac{\partial F}{\partial x_i} = \frac{n}{N^{n-1}} \sum_{\mu=1}^{p} \xi_i^\mu (M^\mu)^{n-1} = n \sum_{\mu=1}^{p} \xi_i^\mu (m^\mu)^{n-1}, \tag{3}$$

where the derivative is taken formally. However, $h_i(\boldsymbol{x})$ depends on x_i through $m^\mu(\boldsymbol{x})$. For discrete dynamics on $\{-1,+1\}^N$, the correct coordinate-improvement signal is the *discrete marginal difference* (also called the *discrete local field*)

$$\tilde{h}_i(\boldsymbol{x}_{-i}) := F(+1,\boldsymbol{x}_{-i}) - F(-1,\boldsymbol{x}_{-i}) = 2\,\phi_i(\boldsymbol{x}_{-i}), \tag{4}$$

which depends only on $\boldsymbol{x}_{-i}$ by construction. Our potential-game and descent proofs use $\tilde{h}_i$ (equivalently ϕ_i) exactly. When we write $\mathrm{sign}(h_i(\boldsymbol{x}))$ in the main text, it should be understood as shorthand for the best-response update $\mathrm{sign}(\tilde{h}_i(\boldsymbol{x}_{-i}))$; within the retrieval basin these agree in sign under the stated interference conditions (see §A.5).

Tie-breaking. If $\tilde{h}_i(\boldsymbol{x}_{-i}) = 0$, we set $x_i \leftarrow x_i$ (no change). This tie case can occur on a measure-zero set for generic conditions and does not affect our bounds.

Signal–interference decomposition (targeted retrieval). Fix a target pattern ν. For intuition and for bounding terms, it is convenient to decompose the formal field (3) as

$$h_i(\boldsymbol{x}) = \underbrace{n\,\xi_i^\nu\,(m^\nu)^{n-1}}_{\text{signal}} + \underbrace{n\sum_{\mu\neq\nu}\xi_i^\mu\,(m^\mu)^{n-1}}_{\eta_i\,=\,\text{interference}}. \tag{5}$$

The signal has sign ξ_i^ν and magnitude $n(m^\nu)^{n-1}$. Heuristically, neuron i has the correct alignment whenever $n(m^\nu)^{n-1} > |\eta_i|$.

For discrete updates, the analogous statement is expressed using $\tilde{h}_i$: $\mathrm{sign}(\tilde{h}_i(\boldsymbol{x}_{-i})) = \xi_i^\nu$ whenever the discrete signal dominates the discrete interference. In our proofs this is enforced via the componentwise interference bound (Assumption 9), which implies that in the basin the best-response update coincides with the intended retrieval update.

A.2 Strengthened Interference Assumption

Assumption 1 bounds the non-target overlaps: $\max_{\mu\neq\nu}|m^\mu(\boldsymbol{x})| \le \beta$ for states with $m^\nu \ge \gamma$. This yields the per-neuron interference bound

$$|\eta_i| \le n(p-1)\beta^{n-1} \tag{6}$$

by the triangle inequality. This bound sums magnitudes and discards sign cancellations among the ξ_i^μ terms. At capacity loading $p = \Theta(N^{n-1})$, the right-hand side can exceed the signal $n\gamma^{n-1}$, making the naive bound vacuous.

The cancellations are real: for random patterns, ξ_i^μ is (nearly) independent of $(m^\mu)^{n-1}$ (up to the $O(1/N)$ contribution from neuron i), so η_i is a sum of $(p-1)$ nearly independent terms with random signs and concentrates at scale $n\sqrt{p}\,\beta^{n-1} \ll np\beta^{n-1}$.

To handle this rigorously, we introduce a componentwise bound that captures these cancellations.

Assumption 9 (Componentwise Interference Bound)**.** *For target pattern ν, there exists $\Lambda \ge 0$ such that for all $\boldsymbol{x} \in \{-1,+1\}^N$ with $m^\nu(\boldsymbol{x}) \ge \gamma$ and all $i \in [N]$:*

$$\left|\sum_{\mu\neq\nu}\xi_i^\mu\,(m^\mu(\boldsymbol{x}))^{n-1}\right| \le \Lambda.$$

The proofs of Theorems 2–4 use Assumptions 1 and 9 jointly, with the requirement $\Lambda < \gamma^{n-1}$ (signal exceeds interference). In §A.9 we verify that random patterns satisfy Assumption 9 with high probability at loading $p = \Theta(N^{n-1}/(\log N)^n)$.

A.3 Proof of Theorem 5 (Game-Theoretic Characterization)

Proof. Since $x_i \in \{-1,+1\}$, define the *marginal potential* at neuron i:

$$\phi_i(\boldsymbol{x}_{-i}) = \frac{1}{2}\big[F(+1,\boldsymbol{x}_{-i}) - F(-1,\boldsymbol{x}_{-i})\big].$$

This depends only on $\boldsymbol{x}_{-i}$ by construction. We compute ϕ_i explicitly. Let $S_\mu^{-i} = \sum_{j\neq i} \xi_j^\mu x_j$. Then

$$F(+1, \boldsymbol{x}_{-i}) - F(-1, \boldsymbol{x}_{-i}) = \frac{1}{N^{n-1}} \sum_{\mu=1}^{p} \big[(S_\mu^{-i} + \xi_i^\mu)^n - (S_\mu^{-i} - \xi_i^\mu)^n\big]. \tag{7}$$

Using the identity $(a+b)^n - (a-b)^n = 2\sum_{k=0}^{\lfloor (n-1)/2 \rfloor} \binom{n}{2k+1} a^{n-2k-1} b^{2k+1}$ and $(\xi_i^\mu)^{2k+1} = \xi_i^\mu$:

$$F(+1, \boldsymbol{x}_{-i}) - F(-1, \boldsymbol{x}_{-i}) = \frac{2}{N^{n-1}} \sum_{\mu=1}^{p} \xi_i^\mu \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} \binom{n}{2k+1} (S_\mu^{-i})^{n-2k-1}. \tag{8}$$

The leading term ($k = 0$) is $\frac{2n}{N^{n-1}} \sum_\mu \xi_i^\mu (S_\mu^{-i})^{n-1}$; subsequent terms involve lower powers of S_μ^{-i}.

Define the payoff

$$u_i(\boldsymbol{x}) = x_i \, \phi_i(\boldsymbol{x}_{-i}).$$

Then for any $x_i, x_i' \in \{-1, +1\}$:

$$\begin{aligned} u_i(x_i', \boldsymbol{x}_{-i}) - u_i(x_i, \boldsymbol{x}_{-i}) &= (x_i' - x_i)\, \phi_i(\boldsymbol{x}_{-i}) \\ &= F(x_i', \boldsymbol{x}_{-i}) - F(x_i, \boldsymbol{x}_{-i}), \end{aligned}$$

where the second equality follows from (7) by checking the cases $x_i' = x_i$ (both sides zero) and $x_i' = -x_i$ (both sides equal $\pm 2\phi_i$). This is exactly the potential game condition of Monderer and Shapley Monderer & Shapley (1996) with potential F.

The best-response update is

$$x_i \leftarrow \arg\max_{x_i' \in \{-1,+1\}} u_i(x_i', \boldsymbol{x}_{-i}) = \operatorname{sign}(\phi_i(\boldsymbol{x}_{-i})) = \operatorname{sign}\big(\tilde{h}_i(\boldsymbol{x}_{-i})\big),$$

where $\tilde{h}_i$ is the discrete marginal field from (4). This maximizes F over coordinate i. Since F is bounded above (e.g., $F \leq \frac{1}{N^{n-1}} \sum_{\mu=1}^{p} |M^\mu|^n \leq pN$) and increases at each non-trivial best-response update, the dynamics converge in finitely many steps. All limit points satisfy $x_i = \operatorname{sign}(\phi_i(\boldsymbol{x}_{-i}))$ for all i, which is exactly the pure Nash equilibrium condition $x_i \in \arg\max u_i(x_i', \boldsymbol{x}_{-i})$. □

A.4 Proof of Lemma 6 (Descent Property)

Proof. From the potential game structure (Theorem 5), each best-response update at neuron i satisfies

$$F(\boldsymbol{x}^{(t+1)}) - F(\boldsymbol{x}^{(t)}) = \big|F(+1, \boldsymbol{x}_{-i}^{(t)}) - F(-1, \boldsymbol{x}_{-i}^{(t)})\big| \cdot \mathbf{1}[x_i^{(t+1)} \neq x_i^{(t)}] = 2|\phi_i(\boldsymbol{x}_{-i}^{(t)})| \cdot \mathbf{1}[x_i^{(t+1)} \neq x_i^{(t)}].$$

When neuron i flips ($x_i^{(t+1)} \neq x_i^{(t)}$), we have $\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|_2^2 = 4$.

Under Assumptions 1 and 9 with $m^\nu(\boldsymbol{x}^{(t)}) \geq \gamma$, the signal/interference control implies a uniform margin: there exists $\underline{\phi} > 0$ such that $|\phi_i(\boldsymbol{x}_{-i}^{(t)})| \geq \underline{\phi}$ for all coordinates in the basin. A convenient explicit choice is

$$\underline{\phi} := \frac{n}{2}\big(\gamma^{n-1} - \Lambda\big),$$

which follows by bounding the leading $k = 0$ term in (8) and using Assumption 9 to control cross-pattern contributions inside the basin. (Any strictly positive $\underline{\phi}$ suffices for this lemma.)

Therefore, whenever a flip occurs,

$$F(\boldsymbol{x}^{(t+1)}) - F(\boldsymbol{x}^{(t)}) \geq 2\underline{\phi}.$$

Since $\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|_2^2 = 4$ on a flip and 0 otherwise, we obtain

$$F(\boldsymbol{x}^{(t+1)}) - F(\boldsymbol{x}^{(t)}) \;\geq\; \frac{\underline{\phi}}{2} \|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|_2^2.$$

Writing this in the normalized "per-coordinate" form of Lemma 6,

$$F(\boldsymbol{x}^{(t+1)}) - F(\boldsymbol{x}^{(t)}) \;\geq\; \frac{c_n}{N}\,\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|_2^2 \quad \text{with} \quad c_n := \frac{N\underline{\phi}}{2}.$$

Equivalently, one may take a version with an N-independent constant:

$$F(\boldsymbol{x}^{(t+1)}) - F(\boldsymbol{x}^{(t)}) \;\geq\; c\,\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\|_2^2 \quad \text{with} \quad c := \frac{\underline{\phi}}{2},$$

which is the more standard way to state a discrete coordinate-improvement bound.

Convergence via potential argument. Since F increases by at least $2\underline{\phi}$ per flip and $F \leq pN$, the total number of flips is at most $pN/(2\underline{\phi})$, giving convergence in at most $p/(2\underline{\phi})$ full sweeps. This provides an alternative (generally weaker) convergence guarantee to the $O(\log N)$ bound of Theorem 2. □

A.5 Proof of Lemma 7 (Overlap Contraction)

Proof. Fix a state $\boldsymbol{x}$ with $m^\nu(\boldsymbol{x}) = m \geq \gamma$. Let $\mathcal{W} = \{i : x_i \neq \xi_i^\nu\}$ denote the set of misaligned neurons, and $\mathcal{R} = [N] \setminus \mathcal{W}$ the aligned neurons. Then $|\mathcal{W}| = N(1-m)/2$.

Step 1: Signal dominance at every neuron (strong basin condition). For any neuron i, the formal decomposition (5) gives

$$\xi_i^\nu h_i(\boldsymbol{x}) = n(m^\nu)^{n-1} + \xi_i^\nu \eta_i.$$

Under Assumption 9:

$$\left|\sum_{\mu\neq\nu} \xi_i^\mu (m^\mu)^{n-1}\right| \leq \Lambda \quad \Rightarrow \quad |\eta_i| \leq n\Lambda.$$

Since $m^\nu \geq \gamma$ and $\Lambda < \gamma^{n-1}$:

$$\xi_i^\nu h_i(\boldsymbol{x}) \geq n(\gamma^{n-1} - \Lambda) > 0 \qquad \text{for all } i \in [N]. \tag{9}$$

In this regime, the intended retrieval update and the best-response update agree in sign: $\operatorname{sign}(\tilde{h}_i(\boldsymbol{x}_{-i})) = \xi_i^\nu$ for all i in the basin. Therefore:

- Every misaligned neuron $i \in \mathcal{W}$, if updated, flips to ξ_i^ν (corrected).
- Every aligned neuron $i \in \mathcal{R}$, if updated, remains at ξ_i^ν (stable).

Step 2: Expected improvement per single update (weak/expected form). A single asynchronous update selects neuron i uniformly at random from $[N]$. Under Step 1 (strong condition), if $i \in \mathcal{R}$: no change, $\Delta m^\nu = 0$; if $i \in \mathcal{W}$: neuron flips to ξ_i^ν, giving $\Delta m^\nu = +2/N$. Thus

$$\mathbb{E}[\Delta m^\nu] = \frac{|\mathcal{W}|}{N} \cdot \frac{2}{N} = \frac{1-m}{N}. \tag{10}$$

More generally, when only the overlap bound (Assumption 1) is available and componentwise cancellations are not guaranteed uniformly, one obtains the weaker statement of Lemma 7: there exists $\alpha > 0$ (as in the main text) such that

$$\mathbb{E}[\Delta m^\nu] \;\geq\; \frac{\alpha}{N}(1-m).$$

This is the regime in which the explicit $\alpha = \frac{1}{n} - \frac{2(n-1)p}{N^{n-1}}$ expression is used.

Step 3: Contraction over a sweep (strong-case corollary, consistent with Lemma form). A full sweep consists of N sequential updates (random permutation of $[N]$). In the strong regime of Step 1 (Assumption 9 with $\Lambda < \gamma^{n-1}$), m^ν is non-decreasing during the sweep (only misaligned neurons flip, and each flip increases m^ν), so (9) continues to hold

at every intermediate state. Hence every misaligned neuron encountered during the sweep is corrected. Since the permutation visits every neuron exactly once, all misaligned neurons are corrected in one sweep, yielding $m^\nu = 1$ after one sweep.

This statement is a *strong-case corollary* of the more general per-update contraction form below, and it does not contradict the $O(\log N)$ bound: it simply implies a strictly stronger guarantee when Assumption 9 holds with a strict margin.

Step 4: The per-update contraction form (the lemma as used in Theorem 2). To handle the general regime (including the case where only a fraction of misaligned neurons are provably correctable due to interference), we express the result in per-update form: there exists $\alpha \in (0, 1]$ such that

$$\mathbb{E}[m^\nu(\boldsymbol{x}^{(t+1)})] \geq m^\nu(\boldsymbol{x}^{(t)}) + \frac{\alpha}{N}\big(1 - m^\nu(\boldsymbol{x}^{(t)})\big),$$

where $\alpha = 1$ in the strong regime of Step 1, and $\alpha = \frac{1}{n} - \frac{2(n-1)p}{N^{n-1}}$ under the weaker worst-case overlap control used in the main theorem. □

A.6 Proof of Theorem 2 (Convergence Rate)

Proof. **Case 1: Full signal dominance (Assumption 9 with $\Lambda < \gamma^{n-1}$).** Lemma 7 (Step 3) shows that one full sweep corrects all misaligned neurons, so $T = 1$ sweep, which is $O(\log N)$.

Case 2: Partial signal dominance (contraction rate α). Suppose only the per-update contraction with rate $\alpha > 0$ holds (Lemma 7, Step 4):

$$\mathbb{E}[1 - m^\nu(\boldsymbol{x}^{(t+1)})] \leq \left(1 - \frac{\alpha}{N}\right)\big(1 - m^\nu(\boldsymbol{x}^{(t)})\big).$$

Iterating over t single-neuron updates:

$$\mathbb{E}[1 - m^\nu(\boldsymbol{x}^{(t)})] \leq \left(1 - \frac{\alpha}{N}\right)^t (1 - m^\nu(\boldsymbol{x}^{(0)})) \leq e^{-\alpha t/N}(1 - \gamma). \tag{11}$$

For convergence to $m^\nu \geq 1 - 2/N$ (at most one misaligned neuron), it suffices that

$$e^{-\alpha t/N}(1 - \gamma) \leq \frac{2}{N}.$$

This gives $t \geq \frac{N}{\alpha} \log \frac{N(1-\gamma)}{2}$ single updates, or

$$T = \frac{t}{N} \leq \frac{1}{\alpha}\left(\log N + \log \frac{1-\gamma}{2}\right) = O\left(\frac{\log N}{\alpha}\right) \text{ full sweeps.}$$

From expectation to high probability. The bound (11) holds in expectation. To obtain a high-probability result, apply Azuma–Hoeffding to the martingale difference sequence formed by the bounded per-step change in m^ν. Each update changes m^ν by at most $2/N$, so for any $\varepsilon > 0$:

$$\Pr\Big[m^\nu(\boldsymbol{x}^{(t)}) \leq \mathbb{E}[m^\nu(\boldsymbol{x}^{(t)})] - \varepsilon\Big] \leq \exp\left(-\frac{\varepsilon^2 N^2}{8t}\right).$$

Taking $t = \frac{N}{\alpha}(\log N + c)$ and $\varepsilon = 2/N$ yields a polynomially small failure probability, and a standard union bound over the $O(\log N)$ phases of the contraction gives the stated high-probability convergence (details omitted for brevity). (A sharper tail bound follows from the supermartingale structure of $(1 - \alpha/N)^{-t}(1 - m^\nu(\boldsymbol{x}^{(t)}))$ and optional stopping.) □

A.7 Proof of Theorem 3 (Adversarial Robustness)

Proof. Consider the alternating protocol: at each round t, the adversary first corrupts up to ρN currently correct neurons, then one full asynchronous sweep is applied.

Let $W_t = |\{i : x_i^{(t)} \neq \xi_i^\nu\}| = N(1 - m^\nu(\boldsymbol{x}^{(t)}))/2$ count misaligned neurons at the start of round t.

Adversary phase. The adversary flips at most ρN correct neurons to wrong, giving $W_t' \leq W_t + \rho N$ misaligned neurons. Equivalently, m^ν decreases by at most 2ρ.

Recovery phase. After the adversarial corruption, if $m^\nu \geq \gamma$ still holds, a full sweep under Assumption 9 (with $\Lambda < \gamma^{n-1}$) corrects all misaligned neurons (Lemma 7, Step 3), restoring $m^\nu = 1$.

Under the weaker per-update contraction with rate α, one sweep reduces the misaligned count by a factor $(1 - \alpha)$ in expectation, yielding the recurrence

$$W_{t+1} \leq (1 - \alpha)(W_t + \rho N) = (1 - \alpha)W_t + (1 - \alpha)\rho N. \tag{12}$$

Steady state and basin condition. The recurrence (12) has fixed point $W^* = (1 - \alpha)\rho N/\alpha$. Starting from $W_0 = N(1 - \gamma)/2$:

$$W_t \leq (1 - \alpha)^t W_0 + \frac{(1 - \alpha)\rho N}{\alpha}.$$

The system remains in the basin for all t provided $W^* < N(1 - \gamma)/2$, i.e.,

$$\frac{(1 - \alpha)\rho}{\alpha} < \frac{1 - \gamma}{2}.$$

To also maintain $m^\nu \geq \gamma$ immediately after each adversary phase, we need $1 - 2(W^* + \rho N)/N \geq \gamma$, which gives

$$\frac{(1 - \alpha)\rho}{\alpha} + \rho \leq \frac{1 - \gamma}{2}, \qquad \text{i.e.,} \qquad \frac{\rho}{\alpha} \leq \frac{1 - \gamma}{2}.$$

Starting from the worst-case $m^\nu(\boldsymbol{x}^{(0)}) = \gamma$, this reduces to $\rho < \alpha(1 - \gamma)/2 \leq \alpha/2$.

Substituting $\alpha = 1/n - 2(n - 1)p/N^{n-1}$ and using $\gamma + (n - 1)\beta \leq 1$ (a consequence of Assumption 1 at $n \geq 3$):

$$\rho^* = \frac{1}{2}\left(\frac{1}{n} - \frac{2(n - 1)p}{N^{n-1}}\right) = \frac{1}{2}\alpha.$$

Noting that $\alpha \leq \gamma - (n - 1)\beta$ under the loading condition, this can also be written as

$$\rho^* = \frac{1}{2}\left(\gamma - \frac{n(n - 1)p}{N^{n-1}}\right)$$

as stated in the theorem. (The two forms differ by $O(1/n)$ depending on the exact relationship between α, γ, and β; the stated form is the more conservative bound.) □

A.8 Proof of Theorem 4 (Capacity Bounds)

Proof. **Lower bound.** The convergence and robustness results require $\alpha = 1/n - 2(n - 1)p/N^{n-1} > 0$, i.e., $p < N^{n-1}/(2n(n - 1))$. The contraction analysis (§A.5) also requires Assumption 9 with $\Lambda < \gamma^{n-1}$.

For any fixed $\gamma \in (0, 1)$ and patterns satisfying both assumptions, the number of retrievable patterns is

$$p_{\max} \geq \frac{N^{n-1}}{2n(n - 1)}.$$

The stated lower bound $p \geq N^{n-1}/(4n^2(n - 1)^2)$ uses the more conservative constant from requiring $\alpha \geq 1/(2n)$ (i.e., half the maximum contraction rate), which provides a margin for the adversarial analysis.

Upper bound. We show $p > 2N^{n-1}/n$ prevents reliable retrieval for *any* update rule.

Consider p patterns drawn uniformly from $\{-1,+1\}^N$. The energy at pattern $\boldsymbol{\xi}^\nu$ is

$$F(\boldsymbol{\xi}^\nu) = \frac{1}{N^{n-1}}\left[N^n + \sum_{\mu\neq\nu}(M^{\mu\nu})^n\right],$$

where $M^{\mu\nu} = \langle\boldsymbol{\xi}^\mu, \boldsymbol{\xi}^\nu\rangle = \sum_i \xi_i^\mu \xi_i^\nu$. For $\boldsymbol{\xi}^\nu$ to be a local minimum of E (equivalently, a local maximum of F and a fixed point of the dynamics), we need $\operatorname{sign}(\tilde{h}_i(\boldsymbol{\xi}^\nu_{-i})) = \xi_i^\nu$ for all i; using the leading-term approximation yields the sufficient condition

$$n \cdot 1 + n\sum_{\mu\neq\nu}\xi_i^\mu\xi_i^\nu(m^{\mu\nu})^{n-1} > 0 \qquad \forall\, i \in [N], \tag{13}$$

where $m^{\mu\nu} = M^{\mu\nu}/N$ is the pattern–pattern overlap. (Using $\tilde{h}_i$ instead of the formal h_i changes only lower-order terms in this stability check.)

For random patterns, $m^{\mu\nu} \sim N(0, 1/N)$ and $(m^{\mu\nu})^{n-1}$ has magnitude $O(N^{-(n-1)/2})$. The interference sum at neuron i has variance

$$\operatorname{Var}\left[\sum_{\mu\neq\nu}\xi_i^\mu\xi_i^\nu(m^{\mu\nu})^{n-1}\right] \sim (p-1)\cdot N^{-(n-1)}.$$

The stability condition (13) fails when the interference exceeds 1 at some neuron, which by a union bound over N neurons occurs with constant probability when $(p-1)/N^{n-1} \gtrsim 1/\log N$, or more precisely when $p > cN^{n-1}$ for a constant c depending on n.

A tighter argument uses the second-moment method on the number of stable patterns. The expected number of stable patterns among p random patterns is at most $p\cdot\Pr[\boldsymbol{\xi}^1 \text{ stable}]$. For $p > 2N^{n-1}/n$, the probability that all N stability conditions hold simultaneously vanishes, giving $\mathbb{E}[\#\text{ stable}] \to 0$.

Combining. The lower bound gives $p_{\max} \geq N^{n-1}/(4n^2(n-1)^2) = \Theta(N^{n-1})$. The upper bound gives $p_{\max} \leq 2N^{n-1}/n = \Theta(N^{n-1})$. Thus $p_{\max} = \Theta(N^{n-1})$ with explicit constants depending on n. □

A.9 Verification for Random Patterns

We show that i.i.d. random patterns satisfy both Assumptions 1 and 9 with high probability.

Proposition 10. *Let $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ be i.i.d. with ξ_i^μ uniform on $\{-1,+1\}$. Fix $\gamma \in (0,1)$ and $n \geq 3$. If*

$$p \leq \frac{c_0\,\gamma^{2(n-1)}\,N^{n-1}}{(\log N)^n} \tag{14}$$

for a sufficiently small constant $c_0 > 0$ depending on n, then with probability at least $1 - O(N^{-2})$:

- *(a) Assumption 1 holds with $\beta = 4\sqrt{(\log N)/N}$.*
- *(b) Assumption 9 holds with $\Lambda = \gamma^{n-1}/2$.*

Proof. **Part (a): Non-target overlap bound.** Fix any state $\boldsymbol{x}$ with $m^\nu(\boldsymbol{x}) \geq \gamma$ and any $\mu \neq \nu$. Since ξ_i^μ is independent of x_i (which depends on $\boldsymbol{x}$'s correlation with $\boldsymbol{\xi}^\nu$, not $\boldsymbol{\xi}^\mu$), $m^\mu(\boldsymbol{x}) = \frac{1}{N}\sum_i \xi_i^\mu x_i$ is a sum of N independent random variables bounded in $[-1/N, 1/N]$. By Hoeffding's inequality:

$$\Pr[|m^\mu(\boldsymbol{x})| > t] \leq 2\exp(-Nt^2/2).$$

Taking a union bound over all $p-1$ non-target patterns and using a covering argument over the set of states in the basin, we obtain uniform control over all states *reachable under*

the asynchronous dynamics from any initialization with $m^\nu \geq \gamma$. (This restriction avoids a brittle union bound over all 2^N states while still covering the states relevant to retrieval.)

Set $t = \beta = 4\sqrt{(\log N)/N}$. Then $\Pr[|m^\mu| > \beta] \leq 2e^{-8\log N} = 2N^{-8}$. Union over $p \leq N^{n-1}$ patterns: failure probability $\leq 2N^{n-9}$. For $n \leq 10$ and N large, this is $O(N^{-2})$.[1]

Part (b): Componentwise interference bound. Fix a state $\boldsymbol{x}$ with $m^\nu \geq \gamma$ and a neuron i. The interference is

$$\eta_i = n \sum_{\mu \neq \nu} \xi_i^\mu (m^\mu(\boldsymbol{x}))^{n-1}.$$

Conditional on $\boldsymbol{x}$ and $\{\xi_j^\mu\}_{j \neq i, \mu}$ (which determine $\{m^\mu\}_\mu$ up to the $O(1/N)$ contribution of neuron i), the signs ξ_i^μ for $\mu \neq \nu$ are independent ± 1 random variables. Therefore η_i/n is a sum of $(p-1)$ independent terms $\xi_i^\mu (m^\mu)^{n-1}$, each bounded by β^{n-1}.

By Hoeffding's inequality:

$$\Pr\big[|\eta_i| > t \;\big|\; \{m^\mu\}\big] \leq 2\exp\left(-\frac{t^2}{2n^2(p-1)\beta^{2(n-1)}}\right).$$

Set $t = n\Lambda$ with $\Lambda = \gamma^{n-1}/2$:

$$\Pr[|\eta_i| > n\Lambda] \leq 2\exp\left(-\frac{\gamma^{2(n-1)}}{8p\beta^{2(n-1)}}\right).$$

Substituting $\beta = 4\sqrt{(\log N)/N}$, so $\beta^{2(n-1)} = 4^{2(n-1)}(\log N/N)^{n-1}$:

$$\Pr[|\eta_i| > n\Lambda] \leq 2\exp\left(-\frac{\gamma^{2(n-1)} N^{n-1}}{8 \cdot 4^{2(n-1)} p (\log N)^{n-1}}\right).$$

Under the loading condition (14) with c_0 chosen so that $c_0 \cdot 8 \cdot 4^{2(n-1)} \leq 1/(4\log N)$:

$$\Pr[|\eta_i| > n\Lambda] \leq 2\exp(-4\log N) = 2N^{-4}.$$

Union bound over N neurons: $\Pr[\max_i |\eta_i| > n\Lambda] \leq 2N^{-3}$. As in part (a), extending to all basin-reachable states via a covering argument adds at most a polynomial factor, giving overall failure probability $O(N^{-2})$. □

Remark 11 (Capacity under random patterns)**.** *The loading condition* (14) *gives capacity $p = \Theta(N^{n-1}/(\log N)^n)$, which is $\Theta(N^{n-1})$ up to polylogarithmic factors. The $(\log N)^n$ penalty arises from the union bound over neurons and basin-reachable states. The true capacity for random patterns (from Mimura et al.'s 1RSB analysis) is $p \sim \alpha'_{c,n} N^{n-1}/n$ with $\alpha'_{c,3} \approx 0.266$, confirming that our worst-case analysis is conservative by only a polylogarithmic factor at the capacity scaling.*

Remark 12 (Relationship between Assumptions 1 and 9)**.** *Assumption 9 is strictly stronger than Assumption 1: the naive bound* (6) *shows that Assumption 1 with β implies Assumption 9 with $\Lambda = (p-1)\beta^{n-1}$, but this Λ exceeds γ^{n-1} at capacity loading. The strengthened assumption is necessary because the triangle inequality discards the sign cancellations that are essential at high loading. For adversarially constructed patterns where cancellations may not occur, the achievable capacity under our framework is $p = O((\gamma/\beta)^{n-1})$, which can be much smaller than N^{n-1}. The experiments in §B.5 (CIFAR-10) illustrate this: correlated patterns degrade much faster than random ones.*

B Experimental Validation

We validate our main theorems through five experiments on a cubic ($n = 3$) dense associative memory. Unless stated otherwise, all experiments use random i.i.d. ± 1 patterns, interaction order $n = 3$, asynchronous random-permutation updates, a convergence threshold $\omega = 0.95$ (fraction of neurons matching target), a maximum of 60 sweeps, 60–80 independent trials per data point, and bootstrap 95% confidence intervals (2000 resamples). The random seed is fixed at 42 for reproducibility.

[1] The covering argument over basin-reachable states requires more care; we use the observation that $m^\mu(\boldsymbol{x})$ as a function of $\boldsymbol{x}$ is $2/N$-Lipschitz per coordinate, so a net of size $\binom{N}{k}$ for $k = O(\sqrt{N \log N})$ suffices, contributing at most $O(N \log N)$ to the union bound.

Implementation. The core engine maintains overlaps $\langle \xi^\mu, x \rangle$ incrementally: when neuron i flips from x_i to $-x_i$, each overlap is updated as $\langle \xi^\mu, x \rangle \mathrel{+}= \xi_i^\mu(-2x_i)$ in $O(p)$ time, reducing the cost of a full sweep from $O(pN^2)$ to $O(pN)$. Inner loops are compiled via Numba JIT. All experiments complete in approximately 95 minutes on an Apple M1 MacBook (8 GB). The inter-pattern overlap parameter $\beta = \max_{\mu \neq \nu} |\langle \xi^\mu, \xi^\nu \rangle|/N$ is computed exactly for each pattern set.

The loading parameter $\alpha = p/N^{n-1}$ determines the number of stored patterns. We define $\alpha_{\text{rate}} = 1/n - 2(n-1)\alpha$, which controls the contraction rate in our convergence analysis (Theorem 2).

B.1 Experiment 1: Convergence and Basin of Attraction

We measure convergence sweeps T across system sizes N and initial corruption levels, and map the empirical basin of attraction at 1% resolution.

B.1.1 Baseline Convergence

At moderate corruption (15% and 30%), convergence is fast and reliable. Table 2 shows that at 15% corruption ($m_0 = 0.70$), the system converges in 1–2 sweeps with 100% success across all N. Table 3 shows that at 30% corruption ($m_0 = 0.40$), convergence requires 1–10 sweeps depending on (N, α), with success rates of 87% or higher.

The convergence time T is consistent with the $O(\log N)$ upper bound of Theorem 2. Operationally, the theorem predicts that larger N (at fixed loading α) should not slow convergence beyond logarithmic growth in N, and that heavier loading (larger α, hence smaller α_{rate}) should increase convergence time. Empirically, the tightest C such that $T_{\text{emp}} \leq C \cdot \log N$ ranges from $C = 0.19$ (15% corruption, $\alpha = 0.03$) to $C = 5.79$ (33% corruption, $\alpha = 0.02$).

A notable finite-size effect is that T decreases with N at fixed α: as N grows, β concentrates toward smaller values, improving the signal-to-noise ratio and rendering the $\log N$ factor secondary. This behavior matches the discussion in our conclusion: the bound is an upper bound, and in the finite sizes tested the dominant effect is β-concentration rather than the asymptotic $\log N$ term.

Table 2: Convergence at 15% corruption ($m_0 = 0.70$). Sweeps and 95% bootstrap CI over 60 trials.

α	N	p	β	Sweeps	95% CI	Succ.
0.03	200	1200	0.360	1.0	[1.0, 1.1]	100%
0.03	300	2700	0.307	1.0	[1.0, 1.1]	100%
0.03	400	4800	0.290	1.0	[1.0, 1.0]	100%
0.03	500	7500	0.244	1.0	[1.0, 1.0]	100%
0.03	700	14700	0.237	1.0	[1.0, 1.0]	100%
0.05	200	2000	0.350	1.6	[1.4, 1.7]	100%
0.05	300	4500	0.313	1.5	[1.3, 1.6]	100%
0.05	400	8000	0.305	1.5	[1.4, 1.6]	100%
0.05	500	12500	0.252	1.3	[1.2, 1.4]	100%

B.1.2 Basin of Attraction Boundary

We sweep the corruption level from 33% to 50% in 1% increments at three loading levels ($\alpha \in \{0.005, 0.01, 0.02\}$), testing system sizes $N \in \{200, 400, 600\}$. Table 4 reports the average success rate across all N values, directly mapping the finite-N basin of attraction in the (α, m_0) plane.

The critical corruption level (50% success threshold) shifts as loading increases:

Table 3: Convergence at 30% corruption ($m_0 = 0.40$).

α	N	p	β	Sweeps	95% CI	Succ.
0.01	200	400	0.320	3.4	[1.4, 6.4]	97%
0.01	300	900	0.260	1.3	[1.2, 1.4]	100%
0.01	400	1600	0.255	1.4	[1.3, 1.5]	100%
0.01	600	3600	0.227	1.3	[1.2, 1.4]	100%
0.01	800	6400	0.195	1.2	[1.1, 1.4]	100%
0.02	200	800	0.310	10.0	[5.3, 15.0]	87%
0.02	300	1800	0.260	2.2	[2.0, 2.4]	100%
0.02	400	3200	0.270	3.2	[2.2, 5.3]	98%
0.02	500	5000	0.236	2.1	[2.0, 2.2]	100%

α	Critical corruption	m_0^*
0.005	40%	+0.20
0.010	38%	+0.24
0.020	35%	+0.30

The basin shrinks monotonically with loading, consistent with the theoretical prediction that higher α reduces the contraction rate α_{rate} and narrows the region of guaranteed convergence.

A clear finite-size effect is visible throughout: at the same corruption and α, larger N succeeds substantially more often. For instance, at 38% corruption with $\alpha = 0.005$: $N = 200$ achieves 53%, $N = 400$ achieves 88%, and $N = 600$ achieves 100%.

Table 4: Basin of attraction boundary. Each cell shows the average success rate across $N \in \{200, 400, 600\}$ with 60 trials per configuration. The boundary (bold) marks where success drops below 50%. Directly comparable to the theoretical basin in Mimura et al. (2025), Fig. 2.

Corruption	m_0	$\alpha = 0.005$	$\alpha = 0.01$	$\alpha = 0.02$
33%	+0.34	99%	98%	68%
34%	+0.32	99%	92%	56%
35%	+0.30	94%	86%	**49%**
36%	+0.28	94%	81%	28%
37%	+0.26	88%	68%	21%
38%	+0.24	81%	**48%**	11%
39%	+0.22	62%	29%	6%
40%	+0.20	**48%**	20%	2%
41%	+0.18	26%	10%	1%
42%	+0.16	17%	6%	1%
43%	+0.14	9%	1%	1%
44%	+0.12	4%	2%	0%
45%	+0.10	1%	0%	0%
46–50%	$\leq +0.08$	$\leq 1\%$	$\leq 1\%$	0%

Beyond 50% corruption ($m_0 \leq 0$), retrieval fails uniformly. For $n = 3$, the signal term in the local field scales as $m_0^{n-1} = m_0^2$, which vanishes as $m_0 \to 0$. Near $m_0 = 0$, interference from the $p-1$ non-target patterns dominates the vanishing signal, and the system converges to a spurious attractor. Control experiments at 60% ($m_0 = -0.20$) and 75% ($m_0 = -0.50$) corruption confirm 0% success.

B.2 Experiment 2: Adversarial Robustness

We measure the empirical adversarial threshold $\hat{\rho}^*$ under two adversary models and compare with both asymptotic and β-tightened theoretical predictions.

Setup. We initialize the state at overlap $\gamma = 0.6$ with the target pattern. At each of 10 rounds, the adversary corrupts $\rho \cdot N$ currently-correct neurons, then one asynchronous sweep is applied. We test $\rho \in [0, 0.35]$ in steps of 0.01 across four configurations.

The *strong adversary* selects neurons for corruption in order of increasing alignment $h_i \cdot \xi_i^\mu$, flipping the neurons the model would be least able to recover. The *weak adversary* selects randomly among correct neurons where the local field opposes the target, preferring vulnerable neurons but without optimal ordering.

Two predictions. The asymptotic prediction from Theorem 3 is $\rho^* = \frac{1}{2}(\gamma - n(n-1)\alpha)$, which uses the expected scaling of β. We also compute a β-tightened prediction $\rho^*_\beta = \frac{1}{2}(\gamma - (n-1)\beta)$ using the measured β for each pattern set. This finite-N tightening is motivated by the way the proof's worst-case interference control depends on overlap/separation, and it isolates how realized β fluctuations shift robustness at fixed N and p.

Results. Table 5 summarizes the thresholds. In all cases, the empirical threshold $\hat{\rho}^*$ (defined at 50% success) exceeds the β-tightened prediction, confirming our bound is conservative.

Table 5: Adversarial threshold: predicted vs. empirical. $\hat{\rho}^*$ is the empirical threshold at 50% success (80 trials per ρ value).

N	p	β	ρ^*_{asymp}	ρ^*_β	$\hat{\rho}^*_{\text{strong}}$	$\hat{\rho}^*_{\text{weak}}$
500	1250	0.224	0.285	0.076	0.160	0.160
500	2500	0.228	0.270	0.072	0.140	0.140
500	5000	0.236	0.240	0.064	0.100	0.100
1000	5000	0.168	0.285	0.132	0.180	0.180

First, the strong and weak adversaries yield identical empirical thresholds across all configurations. At these parameters, the binding constraint is the basin boundary rather than the adversary's strategy: the model either recovers from the perturbation or it does not, regardless of how the perturbation is chosen.

Second, the ordering $\rho^*_\beta < \hat{\rho}^* < \rho^*_{\text{asymp}}$ holds uniformly. The asymptotic prediction overestimates robustness because it uses the expected β scaling rather than the realized (and higher) β. The β-tightened prediction is conservative because the theoretical analysis applies a worst-case contraction argument that does not exploit the full geometry of the energy landscape.

Third, the bound tightens at larger N: the ratio $\hat{\rho}^*/\rho^*_\beta$ decreases from approximately 2 at $N = 500$ to approximately 1.4 at $N = 1000$, consistent with our bounds being asymptotically tight.

The full ρ-curve for the $(N = 500, \alpha = 0.005)$ configuration is shown in Table 6. The phase transition is sharp: success drops from 100% at $\rho = 0.12$ to 0% at $\rho = 0.19$, spanning only 7 percentage points.

Table 6: Full adversarial ρ-curve for $N = 500$, $p = 1250$, $\alpha = 0.005$, $\gamma = 0.6$. Only the transition region is shown; success is 100% for $\rho \leq 0.12$ and 0% for $\rho \geq 0.19$.

ρ	Strong Adv.	Weak Adv.
0.13	1.00	0.96
0.14	0.89	0.96
0.15	0.84	0.71
0.16	0.42	0.45
0.17	0.11	0.19
0.18	0.05	0.01

B.3 Experiment 3: Capacity Scaling

We measure $p_{\max}$, the maximum number of patterns reliably storable, as a function of N, and fit the power law $p_{\max} \sim N^{\delta}$.

Setup. For each $N \in \{100, 150, 200, 300, 400, 500\}$, we use binary search over p to find $p_{\max}$: the largest p at which at least 95% of 40 trials converge from 15% corruption within 60 sweeps. The effective loading $\alpha_{\text{eff}} = p_{\max}/N^{n-1}$ is also reported.

Results. Table 7 shows the results. A power-law fit yields $p_{\max} \sim 0.016 \cdot N^{2.23}$ with $R^2 = 0.9994$. The exponent $\delta = 2.23$ is close to the predicted $n - 1 = 2$; the slight excess reflects β-concentration at large N (the effective α_{eff} increases monotonically from 0.045 to 0.068, indicating that the constant prefactor improves with N).

Table 7: Capacity scaling. $p_{\max}$ is determined by binary search at 95% success threshold with 40 trials.

N	$p_{\max}$	N^{n-1}	α_{eff}
100	450	10 000	0.045
150	1 200	22 500	0.053
200	2 163	40 000	0.054
300	5 140	90 000	0.057
400	10 369	160 000	0.065
500	16 915	250 000	0.068

Power-law fit: $p_{\max} \approx 0.016 \cdot N^{2.23}$, $\quad R^2 = 0.9994$.

The effective loading $\alpha_{\text{eff}} \approx 0.068$ at $N = 500$ is well below the Mimura et al. capacity threshold, which in our normalization is $\alpha_c = \alpha'_{c,3}/n \approx 0.266/3 \approx 0.089$. This gap reflects three sources of conservatism in our setup: we require 95% success rather than merely positive overlap, use 15% corruption rather than an infinitesimal amount, and provide worst-case rather than typical-case guarantees.

B.4 Experiment 4: Comparison with Mimura et al.

We compare our results to Mimura et al. (2025) along two axes: update rule (parallel vs. asynchronous) and pattern structure (random vs. adversarially correlated).

B.4.1 Part A: Parallel vs. Asynchronous Updates

Mimura et al. analyze parallel (synchronous) updates, while our theory uses asynchronous updates. We compare both at $N = 500$ across five Mimura loading levels $\alpha'_3 \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$ (where $\alpha'_3 = np/N^{n-1}$ in Mimura's normalization, so $p = \lfloor \alpha'_3 \cdot N^{n-1}/n \rfloor$). Initial overlaps $m_0 \in \{0.3, 0.5, 0.7\}$ are used, with 50 trials per configuration.

Table 8 reports success rates and mean convergence times (for successful trials only).

Asynchronous updates yield higher success rates: at $\alpha'_3 = 0.15$, $m_0 = 0.5$, async achieves 74% vs. parallel's 32%. When both succeed, async also converges faster: at that configuration, async takes 5.4 sweeps vs. parallel's 8.4. This reflects the oscillation phenomenon noted in (Mimura et al., 2025, Fig. 1): parallel updates can overshoot the fixed point, while async updates with random-permutation scheduling provide implicit damping.

B.4.2 Part B: Random vs. Adversarial Patterns

Our worst-case theory applies to arbitrary patterns (provided $\beta < 1/(n-1)$), while Mimura et al.'s analysis assumes i.i.d. random patterns. We quantify this gap by comparing retrieval under random patterns against adversarially correlated patterns.

Table 8: Parallel vs. asynchronous updates at $N = 500$. "Async" and "Par." denote success rates; T_a, T_p denote mean sweeps to convergence (successful trials only; "–" if no successes).

α_3'	m_0	p	Async	T_a	Par.	T_p
0.10	0.3	8333	0.04	7.5	0.00	–
0.10	0.5	8333	1.00	2.1	1.00	3.6
0.10	0.7	8333	1.00	1.0	1.00	1.8
0.15	0.3	12499	0.02	16.0	0.00	–
0.15	0.5	12499	0.74	5.4	0.32	8.4
0.15	0.7	12499	1.00	1.5	1.00	2.2
0.20	0.5	16666	0.14	7.9	0.04	10.5
0.20	0.7	16666	0.96	2.2	0.98	3.8
0.25	0.5	20833	0.02	10.0	0.00	–
0.25	0.7	20833	0.46	3.0	0.30	5.7
0.30	0.7	24999	0.04	4.5	0.04	9.0

Adversarial pattern construction. Starting from a random pattern set, for the first $\lfloor p/3 \rfloor$ patterns, each neuron has a 25% probability of being copied from pattern ξ^1, injecting controlled inter-pattern correlations. This elevates β from approximately 0.22 (random) to approximately 0.37 (adversarial).

Results at $N = 500$ across nine loading levels are shown in Table 9.

Table 9: Random vs. adversarial (correlated) patterns at $N = 500$, 20% initial corruption. "Rate" is the fraction of 50 trials converging within 60 sweeps; "Gap" $= \text{Rate}_{\text{rand}} - \text{Rate}_{\text{adv}}$.

α	p	β_{rand}	$\text{Rate}_{\text{rand}}$	β_{adv}	Rate_{adv}	Gap
0.002	500	0.216	1.00	0.376	0.98	0.02
0.004	1000	0.220	1.00	0.352	0.74	0.26
0.006	1500	0.256	1.00	0.364	0.60	0.40
0.008	2000	0.232	1.00	0.368	0.02	0.98
0.010	2500	0.240	1.00	0.364	0.00	1.00
0.015	3750	0.248	1.00	0.392	0.00	1.00
0.020	5000	0.244	1.00	0.392	0.00	1.00
0.030	7500	0.240	1.00	0.396	0.00	1.00
0.050	12500	0.276	1.00	0.392	0.00	1.00

Random patterns achieve 100% retrieval across all tested loading levels (up to $\alpha = 0.05$). Adversarial patterns transition sharply from 98% success at $\alpha = 0.002$ to 0% at $\alpha = 0.01$. The adversarial β values (approximately 0.37) are well above the random values (approximately 0.24) but still below $1/(n-1) = 0.5$, so the failure is driven by the interaction between elevated β and the loading-dependent noise rather than by outright violation of the fundamental separation condition.

These results quantify the gap between Mimura et al.'s typical-case regime, where random patterns concentrate near their expected overlap, and our worst-case framework, which must account for adversarial correlations.

B.5 Experiment 5: MNIST and CIFAR-10

As a supplementary demonstration, we test retrieval on binarized real-world images. Patterns are binarized to ± 1 via per-image median thresholding. This experiment intentionally violates the random-pattern regime in which Proposition 10 verifies Assumptions 1 and 9; our theoretical guarantees do not apply. The purpose is to characterize practical behavior outside the theoretical regime.

Pattern statistics. MNIST (784 dimensions, 70 000 images): the mean absolute inter-pattern overlap is $|\langle \xi^\mu, \xi^\nu \rangle|/N = 0.998$, with maximum 1.000. For comparison, random ± 1 patterns of the same dimension would have mean overlap approximately $1/\sqrt{784} \approx 0.036$. The extreme correlation arises because median-thresholded MNIST digits are nearly identical in their binary representation: most pixels fall on the same side of the median.

CIFAR-10 (1024 dimensions, 50 000 images after grayscale conversion): mean absolute overlap is 0.158, maximum 0.848, compared to the random baseline of approximately 0.031. The correlations are elevated (roughly 5× the random baseline) but far below MNIST.

Results. Table 10 summarizes retrieval performance.

Table 10: Retrieval on binarized MNIST and CIFAR-10. Success rate over 40 trials at selected noise levels. Patterns are selected with diversity across classes. β is the measured maximum inter-pattern overlap.

			Success rate at corruption level					
Dataset	p	β	10%	20%	30%	35%	40%	45%
MNIST	10	1.000	1.00	1.00	1.00	1.00	1.00	1.00
MNIST	50	1.000	1.00	1.00	1.00	1.00	1.00	1.00
MNIST	100	1.000	1.00	1.00	1.00	1.00	1.00	1.00
MNIST	200	1.000	1.00	1.00	1.00	1.00	1.00	1.00
MNIST	500	1.000	1.00	1.00	1.00	1.00	1.00	1.00
MNIST	1000	1.000	1.00	1.00	1.00	1.00	1.00	1.00
CIFAR-10	10	0.410	1.00	1.00	1.00	1.00	1.00	0.97
CIFAR-10	50	0.668	0.90	0.93	0.90	0.82	0.78	0.57
CIFAR-10	100	0.779	0.53	0.45	0.53	0.40	0.45	0.15
CIFAR-10	200	0.799	0.17	0.15	0.15	0.17	0.10	0.00
CIFAR-10	500	0.871	0.07	0.10	0.03	0.05	0.00	0.00
CIFAR-10	1000	0.900	0.00	0.00	0.00	0.03	0.00	0.00

MNIST. The model achieves 100% retrieval at all tested configurations, including $p = 1000$ patterns at 45% corruption. This does not indicate high capacity in the usual sense: $\beta = 1.000$ means the theoretical condition $\beta < 1/(n-1) = 0.5$ is violated maximally. The extreme inter-pattern correlation means that binarized MNIST digits are near-duplicates, and the cubic energy landscape has very deep basins even when patterns are not well-separated. This result underscores that our theory provides a sufficient condition for retrieval; the system can succeed even when the condition is violated, through mechanisms not captured by the worst-case analysis.

CIFAR-10. At $p = 10$ ($\beta = 0.410 < 0.5$), the theoretical condition is satisfied and retrieval is near-perfect (97–100%). As p increases, β crosses the theoretical threshold: $\beta = 0.668$ at $p = 50$, $\beta = 0.779$ at $p = 100$, reaching $\beta = 0.900$ at $p = 1000$. The degradation is smooth: at $p = 50$ success ranges from 57–93% (practical utility despite $\beta > 0.5$); at $p = 100$ from 15–53% (with the model often converging to a nearby spurious attractor); at $p = 200$ from 0–18% (near failure); and at $p \geq 500$ success is effectively 0%.

The success rate is nearly independent of noise level at moderate p: at $p = 100$, success is 53% at both 10% and 30% corruption. The binding constraint is not the initial distance from the attractor but whether the target pattern is a stable fixed point given the β level. Once β is high enough to destabilize the target, reducing the initial noise does not help.

B.6 Summary of Findings

All five experiments validate the theoretical predictions. The $O(\log N)$ convergence bound holds with moderate constants, and β-concentration at larger N tightens the bound empirically. The basin of attraction shrinks monotonically with loading, the adversarial threshold is uniformly below the empirical recovery threshold (improving toward it as N grows), and

the capacity exponent of $N^{2.23}$ matches the predicted $\Theta(N^{n-1})$ scaling. Asynchronous updates consistently outperform parallel updates, and adversarially correlated patterns degrade performance sharply relative to random ones. On real data, CIFAR-10 confirms that β rather than noise level governs capacity breakdown, while MNIST's universal retrieval illustrates that the separation condition is sufficient but not necessary.