Symmetric Hopfield storage, lower endpoint: round 1 status

Five-round multi-agent campaign, orchestrated by Claude Fable 5.1

2026-11-09

Problem

A network on n neurons is a real symmetric matrix J with zero diagonal and a bias $h \in \mathbb{R}^n$. It **stores** a pattern $s \in \{-1,1\}^n$ when every neuron agrees with its input:

$$s_i\big((Js)_i + h_i\big) > 0 \qquad (i = 1, \dots, n).$$

Draw m distinct patterns $D = \{s^1, \dots, s^m\}$ uniformly from $\{-1,1\}^n$, and let $p_n(m)$ be the probability that one network stores all of them. The **lower endpoint** is the largest load $u = m/n$ at which storage is almost sure:

$$\alpha_- = \sup\{u > 0 : \ p_n(\lfloor un \rfloor) \to 1\}.$$

Known: $1 \le \alpha_- \le 2$. The upper bound needs one neuron only: neuron i must separate m random labels by an affine function of $n-1$ inputs, and Cover's count [C65] makes this fail with high probability once $m > 2n$. The lower bound comes from a *projection-minus-diagonal network* (Appendix A). Let P project orthogonally onto the span of $\hat{s}^\mu = (s^\mu, 1) \in \mathbb{R}^{n+1}$. Take J to be the top-left $n \times n$ block of P with its diagonal removed, and h the first n entries of its last column. Every margin at neuron i then equals $1 - P_{ii}$, and for $m \le n$ random patterns every $P_{ii} < 1$ with high probability. So $p_n(n) \to 1$.

Goal. Move either end. Prove $p_n(\lfloor un \rfloor) \to 1$ for one fixed $u > 1$, giving $\alpha_- \ge u$; or prove $\limsup_n [1 - p_n(\lfloor un \rfloor)] > 0$ for one fixed $u < 2$, giving $\alpha_- \le u$ (deleting a pattern never breaks storage, so p_n is nonincreasing). A result counts only after an independent audit of its whole proof.

What the numerics say

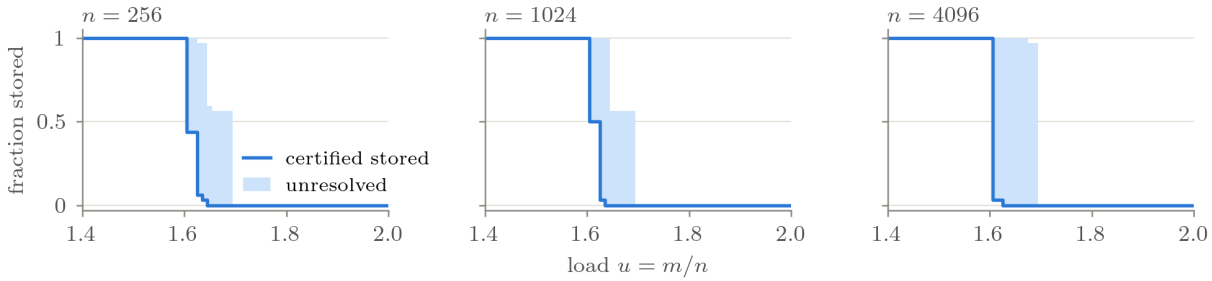


Figure 1. Certified storage against load. The y-axis is the fraction of 32 random sets of $\lfloor un \rfloor$ patterns that one symmetric network provably stores, from this case study's earlier sweep; the band marks sets left unresolved.

The campaign

Round 1 ran as a multi-agent campaign: one orchestrator and eight worker agents, all Claude Fable 5.1. Each of four lanes pairs a **solver**, who writes proofs and exact computations along one route, with a **verifier**, who rederives every claim and reruns every computation in independent exact code. The orchestrator freezes the returns, records each claim's status in a ledger, and keeps or retires routes. From round 2 a shared **experimenter** (X) runs certificate-gated finite tests. The campaign runs five rounds, then a fresh audit by new agents. **Round 1 took 2.5 hours and cost 68.8 M tokens, about $171 at list prices** (Table 1).

Table 1. Round 1 by lane. Tokens count input, cache and output in every transcript; dollars use list prices, with agent output estimated from visible text.

Lane	Route	Round 1 result	Tokens	Cost
S1 ↔ V1	projection networks	Theorem A; route closed	6.8 M	$32
S2 ↔ V2	primal linear program	Theorem D	15.9 M	$39
S3 ↔ V3	coupled dual	Theorem B	11.0 M	$29
S4 ↔ V4	symmetry-only failures	Theorem C; $n = 7$ example	18.8 M	$51
Orchestrator	routing and ledger		16.4 M	$21

Round 1 results

Each statement was rederived by its verifier, and its computations rerun in independent exact code; no critical issue was found. Labels such as S1-6 name nodes of the run record.

Notation. $U = [\hat{s}^1 \cdots \hat{s}^m] \in \mathbb{R}^{(n+1)\times m}$. E is the event that each neuron alone can separate its labels: for every i some $w \in \mathbb{R}^{n-1}$ and $b \in \mathbb{R}$ give $s_i^\mu\big(\sum_{j\neq i} w_j s_j^\mu + b\big) > 0$ for all μ. Storage implies E. A set in E that is not stored is a **symmetry-only failure**.

Theorem A (Projection networks stop at n; S1-6, S1-8)**.** *Let $m \geq n+1$. If $\operatorname{rank} U = n+1$, every projection-minus-diagonal network of D has $J = 0$ and $h = 0$, and stores nothing. For uniform D, $\Pr(\operatorname{rank} U \leq n) \leq (3/4)^{n+1}/\big(1 - \binom{n+1}{2}2^{-n}\big)$ for all large n, uniformly in m. Finally, $p_n(n + o(n)) \to 1$ gives no bound $\alpha_- \geq u$ with $u > 1$: such a bound needs $p_n(m_n) \to 1$ with $m_n - n$ linear in n.*

Theorem B (Exact dual; S3-3, S3-4)**.** *D is not stored if and only if some $\lambda \in \mathbb{R}^{n\times m}$ with $\lambda \geq 0$, $\lambda \neq 0$ satisfies*

$$\sum_\mu \lambda_{i\mu} s_i^\mu = 0 \quad (i \leq n), \qquad \sum_\mu (\lambda_{i\mu} + \lambda_{j\mu})\, s_i^\mu s_j^\mu = 0 \quad (i < j).$$

A certificate λ supported on one neuron i exists exactly when neuron i alone cannot separate its labels. So a symmetry-only failure needs a certificate on at least two neurons.

Theorem C (Below load 2, failures are symmetry-only; S4-1, S4-3, S4-4)**.** *Fix $u \in (0, 2)$ and $m = \lfloor un \rfloor$. Then $\Pr(E) \to 1$ exponentially fast [T23]. Put $c_n(u) = \Pr(D \text{ not stored} \mid E)$. Then $\limsup_n c_n(u) > 0$ if and only if $p_n(\lfloor un \rfloor) \not\to 1$, and either gives $\alpha_- \leq u$.*

Theorem D (Repair of violated constraints; S2-1, S2-3)**.** *List the free entries of (J, h) as $\theta \in \mathbb{R}^{n(n+1)/2}$. Storage is the linear program $A\theta > 0$, where row $a = (\mu, i)$ of A gives the margin $(A\theta)_a = s_i^\mu\big((Js^\mu)_i + h_i\big)$. Put $\Gamma = AA^\top$. Let a network θ_0 have margins $M = A\theta_0$, and let $\mathcal{V}$ be a set of rows containing every a with $M_a < 1$. Put $r = \mathbf{1} - M_\mathcal{V}$, $\omega = \max_{a\in\mathcal{V}} \sum_{b\in\mathcal{V}\setminus\{a\}} |\Gamma_{ab}|$ and $\rho_b = \sum_{a\in\mathcal{V}} |\Gamma_{ba}|$. If $\omega < n$, the corrected network*

$$\theta_1 = \theta_0 + A_\mathcal{V}^\top \Gamma_\mathcal{V}^{-1} r$$

has margin exactly 1 on $\mathcal{V}$, and moves every other margin M_b by at most $\|r\|_\infty \rho_b/(n - \omega)$. So θ_1 stores D whenever $M_b > \|r\|_\infty \rho_b/(n - \omega)$ for all $b \notin \mathcal{V}$.

Since $\Gamma_{aa} = n$ and $|\Gamma_{(\mu,i),(\nu,j)}| \leq |\langle s^\mu, s^\nu\rangle| + 2$ off the diagonal, $\omega < n$ when $\mathcal{V}$ is small and the patterns nearly orthogonal.

Smallest example (S4-7, S4-8). Eight patterns in $\{-1, 1\}^7$ lie in E but are not stored; an exact certificate as in Theorem B, supported on neurons 1 and 5, proves it. For $n \leq 6$ every set in E is stored (exhaustive census).

Plan

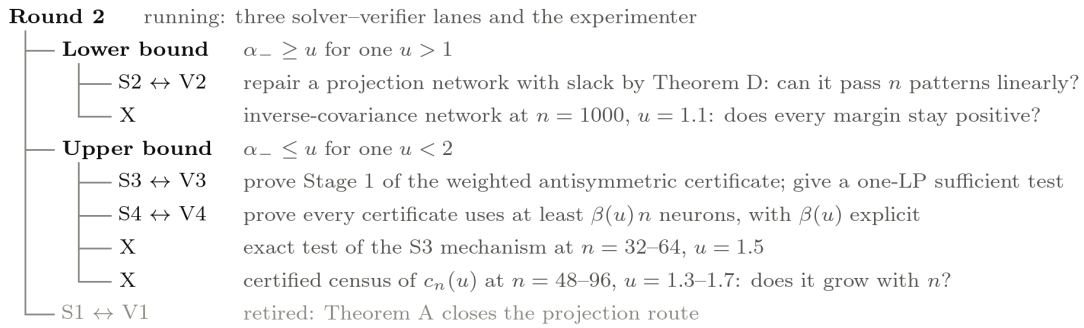


Figure 2. Round 2 plan. Tasks grouped by the bound they serve. X is the experimenter; lane S1 is retired.

Lower bound. Goal: $p_n(\lfloor un \rfloor) \to 1$ for one $u > 1$. By Theorem A the excess over n must be linear, and projection alone cannot supply it. Round 2 (Figure 2) asks whether Theorem D can repair a projection network with slack, and tests the inverse-covariance network built from $(UU^\top)^{-1}$, which S1's heuristic predicts fails above load 1.

Upper bound. Goal: $\limsup_n c_n(u) > 0$ for one $u < 2$ (Theorem C). Equivalently, with probability bounded below, a set in E carries an exact certificate (Theorem B). Round 2 aims to prove the first stage of S3's certificate construction and that certificates spread over a fraction $\beta(u)$ of the neurons, and measures $c_n(u)$ exactly at $n \leq 96$.

Rounds 3 to 5. Round 3 uses the round-2 experiments to falsify the remaining premises. Rounds 4 and 5 keep one route per side; a fresh audit, filing and the write-up close the campaign.

Appendix A. Projection-minus-diagonal networks

Construction. Let $V \subseteq \mathbb{R}^{n+1}$ be any subspace containing $\hat{s}^1, \dots, \hat{s}^m$; the main text takes $V = \operatorname{col} U$. Write the orthogonal projector onto V in blocks, and set

$$P = \begin{pmatrix} B & h \\ h^\top & c \end{pmatrix}, \qquad J = B - \operatorname{diag} B.$$

Then J is symmetric with zero diagonal, and $h \in \mathbb{R}^n$ is the bias.

Margin identity (S1-2). For every neuron $i \le n$ and pattern μ,

$$s_i^\mu \big((Js^\mu)_i + h_i \big) = 1 - P_{ii}.$$

Proof. Since $\hat{s}^\mu \in V$, $P\hat{s}^\mu = \hat{s}^\mu$, and row i reads $(Bs^\mu)_i + h_i = s_i^\mu$. Subtract $P_{ii} s_i^\mu$ to get $(Js^\mu)_i + h_i = (1 - P_{ii}) s_i^\mu$, then multiply by $s_i^\mu = \pm 1$. □

The margin is the same for every pattern. Since $0 \le P_{ii} \le 1$, with $P_{ii} = 1$ if and only if $e_i \in V$, the network stores D if and only if V contains none of $e_1, \dots, e_n$.

When it works (S1-3). For $V = \operatorname{col} U$, $P_{ii} < 1$ if and only if deleting row i of U keeps its rank. For $m \le n$ random patterns this holds for every i with high probability, by singularity bounds for random sign matrices [T20]. Hence $p_n(n) \to 1$.

Why it stops at n. Taking traces, $\sum_{i \le n} (1 - P_{ii}) = n - \operatorname{rank} U + P_{n+1,n+1}$, so the average margin is about $1 - m/n$. At $\operatorname{rank} U = n + 1$ we get $V = \mathbb{R}^{n+1}$, $P = I$ and $J = 0$ (Theorem A).

Literature. Hopfield [H82] stored patterns with the Hebb rule $J_{ij} = \sum_\mu s_i^\mu s_j^\mu$, $J_{ii} = 0$; in his simulations about half of $0.15n$ random patterns are still recalled well. For orthogonal patterns $\frac{1}{n} \sum_\mu s^\mu s^{\mu\top}$ is already the projector onto their span, so the projection rule is the Hebb rule corrected for overlaps. Kohonen and Ruohonen [KR73] showed that the least-squares linear associative map recalls up to n linearly independent keys exactly; mapping patterns to themselves, this map is the projector. Personnaz, Guyon and Dreyfus [PGD85, PGD86] brought this projection (pseudo-inverse) rule to Hopfield networks, where it makes every pattern of a linearly independent set a fixed point. Kanter and Sompolinsky [KS87] studied the projection network with its diagonal removed, the form this problem requires. Earlier runs of this case study reached $n + \lfloor n/(55296 \log n) \rfloor$ patterns (pinned report, Prop. 1); by Theorem A a linear excess needs a different mechanism.

References

[C65] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.* EC-14 (1965), 326–334.

[H82] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* 79 (1982), 2554–2558.

[KR73] T. Kohonen and M. Ruohonen. Representation of associated data by matrix operators. *IEEE Trans. Comput.* C-22 (1973), 701–702.

[KS87] I. Kanter and H. Sompolinsky. Associative recall of memory without errors. *Phys. Rev. A* 35 (1987), 380.

[PGD85] L. Personnaz, I. Guyon and G. Dreyfus. Information storage and retrieval in spin-glass like neural networks. *J. Physique Lett.* 46 (1985), L359–L365.

[PGD86] L. Personnaz, I. Guyon and G. Dreyfus. Collective computational properties of neural networks: new learning mechanisms. *Phys. Rev. A* 34 (1986), 4217–4228.

[T20] K. Tikhomirov. Singularity of random Bernoulli matrices. *Ann. of Math.* 191 (2020), 593–634.

[T23] K. Tikhomirov. On the probability that the convex hull of random points contains the origin. arXiv:2304.13133.