THE READOUT DECIDES WHICH HOPFIELD NETWORKS HAVE A CONSTANT-METRIC ENERGY

Anonymous authors
Paper under double-blind review

ABSTRACT

A modern Hopfield network retrieves a memory by iterating a map that scores the stored patterns against a query, weights them by a softmax, and forms the next state with a readout. Its iteration is backed by an energy that never increases from one step to the next. However, the frameworks that generalise its score and weighting keep the readout a weighted average of the stored patterns, which we call the *pattern readout*. We show that the readout decides which scores make each step a concave–convex step on an energy in a constant metric. We call a map with this property *admissible*. In particular, we consider scores whose gradient splits into a fixed per-pattern part and a part shared by all patterns. Among them, and for patterns in general position, the pattern readout is admissible in the Euclidean metric essentially only for the dot product. In contrast, the *gradient readout* takes the gradient of a potential of the scores. It is admissible for every score that is convex up to an added quadratic. Among 32 published retrieval maps, 8 are shown not to be admissible. For 5 of them, one reason is that their readout is not the gradient of their score's potential. One of them is the adaptive Hopfield network, which can enter cycles of period two as a multiple-instance pooling layer. We keep its score on squared distances, switch to the gradient readout, and constrain its learned weights to be non-negative and non-increasing. These changes give an exact energy that is bounded below throughout training. In our multiple-instance experiments, non-negative weights alone already stop the cycling under the pattern readout, so the energy adds a guarantee that no step raises it. Overall, our results identify the readout as the component that decides which scores give an admissible map, and they show how to test and design admissible retrieval maps.

1 INTRODUCTION

A modern Hopfield network stores memory patterns $\mathbf{\Xi} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M] \in \mathbb{R}^{d \times M}$ and answers a query $\mathbf{x} \in \mathbb{R}^d$ with a retrieval map $T(\mathbf{x})$, applied once or iterated. Most such models build this map in three steps (Millidge et al., 2022). A score rates each pattern against the query, and a separation such as the softmax turns the scores into weights. A readout then combines the patterns under these weights. The usual one is their weighted average, which we call the *pattern readout*. A Hopfield network differs from an arbitrary retrieval rule in having an energy $E(\mathbf{x})$ that never increases from one step of T to the next (Hopfield, 1982; 1984; Krotov & Hopfield, 2016; Ramsauer et al., 2021). This energy backs the iteration, and it also connects these models to attention (Ramsauer et al., 2021; Hoover et al., 2023).

Two frameworks generalise the modern Hopfield network, and both state their energy conditions on the score and the separation. Universal Hopfield networks (Millidge et al., 2022) combine several similarity scores with several separations. Adaptive Hopfield networks (Wang et al., 2026) sum the sorted per-coordinate similarities under learned weights, which we call the gate. Both keep the pattern readout for auto-associative retrieval. However, in both frameworks the energy is proved for a different map from the one that the model iterates (Section 4). Universal Hopfield networks prove descent for continuous-time dynamics, and the adaptive network's proof uses a score without the sort. A few works note that the readout matters, but each only for the model it studies. Santos et al. (2025) observe that their update "supports only the dot product as similarity function". Smart et al. (2025) point out that attention is a gradient descent only when its readout weights are tied to its scoring weights, and Dehmamy et al. (2026) locate the difference between attention and an energy

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

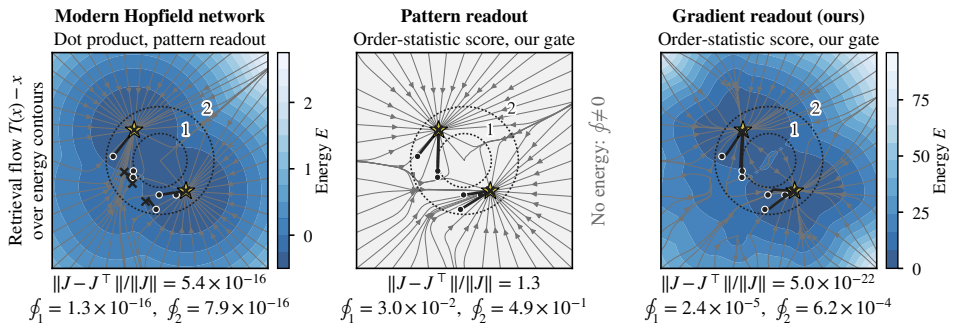

Figure 1: One learned score under two readouts. Each panel shows the retrieval flow and the energy contours on a two-dimensional slice through two stored patterns (stars). Left: the modern Hopfield network, with the dot product and the pattern readout. Middle: the adaptive network's order-statistic score, with our trained non-increasing gate, under the pattern readout. Right: the same score under the gradient readout. The bar beside each shaded panel gives its model's energy. The middle map has no constant-metric energy to contour, so its panel is left blank. Black paths show sixty steps of the map from half-masked training queries: each path starts at a dot and ends at a cross ($d = 32$, $M = 48$). Under each panel, the median Jacobian asymmetry $\|J - J^\top\|/\|J\|$ and the circulations $\oint_1$ and $\oint_2$ of T around the dotted circles 1 and 2 test the Euclidean metric. They vanish up to round-off and quadrature error in the left and right panels only. The middle map's Jacobian also has non-real eigenvalues, which exclude every constant metric.

gradient in the readout matrix. The condition behind these remarks is a symmetric Jacobian (Im et al., 2016; Cardin et al., 2025). These remarks leave two questions open. *Which scores give an energy under a given readout? Can a learned score keep an exact energy throughout training?*

To answer them, we study when each step of the map is a concave–convex step on an energy. We call this property *admissibility*. It asks for a constant metric G and a convex h in $E = \frac{1}{2}\mathbf{x}^\top G\mathbf{x} - h$ (Definition 1). Because the metric is constant, the definition excludes the Lagrangian dense associative memories (Krotov & Hopfield, 2021) and the Energy Transformer (Hoover et al., 2023), which reach an energy in a state-dependent metric. For a smooth map in the Euclidean metric, admissibility means a symmetric positive semidefinite Jacobian. Our main finding is that the readout decides which scores give an admissible map. Figure 1 shows this on one learned score. Under the pattern readout the map has no constant-metric energy, and under the *gradient readout* defined below it descends an energy.

Section 3 addresses the first question. We call a score *splitting* if its gradient is a fixed vector $\boldsymbol{\phi}_\mu$ for each pattern plus a vector $\boldsymbol{\psi}(\mathbf{x})$ shared by all patterns. In the Euclidean metric and under mild conditions on the patterns, the pattern readout $T(\mathbf{x}) = \boldsymbol{\Xi}\mathbf{p}$ meets this Jacobian condition for all softmax weights $\mathbf{p}$ exactly when $\boldsymbol{\phi}_\mu = \lambda\boldsymbol{\xi}_\mu + \mathbf{v}$ for one $\lambda \geq 0$ and one vector $\mathbf{v}$ (Theorem 1). These scores form a half-flat of dimension $d + 1$ that contains the dot product. In contrast, admissibility is nearly free under the gradient readout $T = G^{-1}\nabla h$, which takes the gradient of a potential h of the scores in a constant metric G. Under this readout, every score that is convex up to an added quadratic gives an admissible map (Proposition 1).

We apply these results to 32 published retrieval maps. Eight of them are shown not to be admissible. For five of the eight, one reason is that their readout is not the gradient of their score's potential. One of the eight is the adaptive Hopfield network, and its iteration can fail to converge. As a multiple-instance pooling layer, it can enter a cycle of period two, in which the state alternates between two points. By Proposition 3, two patterns that are each other's best match trap a saturated pattern readout between them, and on squared distances only a negative weight in the gate can create such a pair. We therefore keep the adaptive network's score on squared distances and change its gate and its readout. The resulting map has an exact energy throughout training and never enters a cycle in our multiple-instance experiments. A control that keeps the pattern readout shows that the non-negative gate alone already stops the cycling. In these runs, the energy therefore adds a guarantee that no

step raises the energy. Together, these results identify the readout as the part of a retrieval map that decides which scores give an admissible map.

Contributions.

1. We prove that, among splitting scores and for patterns in general position, the pattern readout is admissible in the Euclidean metric essentially only for the dot product (Theorem 1).
2. We show that under the gradient readout every score that is convex up to an added quadratic gives an admissible map (Proposition 1). A non-real or negative eigenvalue of a map's Jacobian on an open set of queries rules out admissibility and also Fenchel–Young energies with a twice-differentiable conjugate (Proposition 2).
3. We classify 32 published retrieval maps and show that 8 are not admissible, including one-step attention and the adaptive Hopfield network. We prove that a mutual best-match pair makes a saturated pattern readout cycle (Proposition 3).
4. From the adaptive network's score on squared distances, we build a map whose exact energy is bounded below throughout training (Theorem 2, Corollary 1). On five multiple-instance datasets, this map never enters a cycle, while the adaptive network trained the same way has a cycling bag in 250 of 500 runs (Sections 5 and 6).

2 Setting: retrieval maps and what "has an energy" means

Retrieval maps. Write $s_\mu(\mathbf{x}) := \mathrm{sim}(\boldsymbol{\xi}_\mu, \mathbf{x})$ for the score of pattern μ, $\beta > 0$ for the inverse temperature, and $\mathbf{p} := \mathrm{softmax}(\beta \mathbf{s}) \in \Delta^{M-1}$ for the weights. We compare two readouts,

$$T_{\boldsymbol{\Xi}}(\mathbf{x}) = \boldsymbol{\Xi}\,\mathbf{p} \qquad \text{and} \qquad T_\nabla(\mathbf{x}) = G^{-1}\nabla h(\mathbf{x}).$$

The gradient readout T_∇ takes the gradient of a potential h built from the scores, in a metric G. We call a score family *splitting* if its gradients have the form

$$\nabla_{\mathbf{x}} s_\mu(\mathbf{x}) = \boldsymbol{\phi}_\mu + \boldsymbol{\psi}(\mathbf{x}),$$

with a vector $\boldsymbol{\phi}_\mu$ that depends only on the pattern and a function $\boldsymbol{\psi}$ shared by all patterns. We collect the per-pattern parts in $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_M] \in \mathbb{R}^{d \times M}$. The dot product splits with $\boldsymbol{\Phi} = \boldsymbol{\Xi}$. The negative squared distance splits with $\boldsymbol{\Phi} = 2\boldsymbol{\Xi}$ and $\boldsymbol{\psi}(\mathbf{x}) = -2\mathbf{x}$.

What "energy" means here. We use the notion under which the results of Ramsauer et al. (2021) are proved. One step of T is a concave–convex step (Yuille & Rangarajan, 2003) on a quadratic in a fixed metric minus a convex potential h.

Definition 1: Admissible retrieval map

Let $\mathcal{D} \subseteq \mathbb{R}^d$ be open and convex, and let G be a constant symmetric positive-definite matrix. A map $T : \mathcal{D} \to \mathcal{D}$ is *G-admissible* if there is a convex $h : \mathcal{D} \to \mathbb{R}$ with

$$G\,T(\mathbf{x}) \in \partial h(\mathbf{x}) \quad \text{for every } \mathbf{x} \in \mathcal{D}.$$

Then the energy $E(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top G\mathbf{x} - h(\mathbf{x})$ satisfies

$$E(\mathbf{x}) - E(T(\mathbf{x})) \;\geq\; \tfrac{1}{2}\|T(\mathbf{x}) - \mathbf{x}\|_G^2 \;\geq\; 0 \quad \text{for every } \mathbf{x} \in \mathcal{D}.$$

The definition guarantees only descent, with no lower bound on E and no memories at the fixed points. So a check mark in Table 1 means only that no step of the map raises its energy. Appendix B explains why the metric must be constant.

By an integrability lemma (Lemmas 3 and 4, Appendix B), a locally Lipschitz T with Jacobian J_T is G-admissible exactly when GJ_T is symmetric positive semidefinite almost everywhere. In particular, if GJ_T fails this test on a set of positive measure, T is not G-admissible. For the pattern readout the Jacobian has a simple form. If $T = V\mathbf{p}$ with a matrix V that does not depend on $\mathbf{x}$, then the chain rule through the softmax gives (Lemma 5)

$$J_T = \beta\, V A\, \nabla\mathbf{s}, \qquad A := \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top, \qquad A\mathbf{1} = 0,$$

where $\nabla\mathbf{s} \in \mathbb{R}^{M \times d}$ stacks the score gradients as rows. For a splitting score, A annihilates the shared part, since $\nabla\mathbf{s} = \boldsymbol{\Phi}^\top + \mathbf{1}\boldsymbol{\psi}^\top$ and $A\mathbf{1} = 0$ give $A\,\nabla\mathbf{s} = A\boldsymbol{\Phi}^\top$.

3 WHAT AN ENERGY CONSTRAINS

3.1 WHAT THE PATTERN READOUT ADMITS

With a splitting score and $G = \mathbf{I}$, the pattern readout has $J_T = \beta\, \boldsymbol{\Xi} A \boldsymbol{\Phi}^\top$. Admissibility asks that this matrix be symmetric and positive semidefinite.

Theorem 1: The pattern readout generically admits only a half-flat of splitting scores

Let the scores split with per-pattern gradients $\boldsymbol{\Phi}$, and let $T(\mathbf{x}) = \boldsymbol{\Xi}\,\mathrm{softmax}(\beta \mathbf{s}(\mathbf{x}))$.

(i) If $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ for some $\lambda \geq 0$ and $\mathbf{v} \in \mathbb{R}^d$, then T is $\mathbf{I}$-admissible.

(ii) Let the patterns be pairwise distinct and, if $M \geq 3$, not all collinear. Then $\boldsymbol{\Xi} A \boldsymbol{\Phi}^\top$ is symmetric positive semidefinite for *every* $\mathbf{p}$ in the probability simplex if and only if $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ with $\lambda \geq 0$. For $M \geq 2$ the admissible set, uniformly in the weights, is a closed half-flat of dimension $d + 1$.

(iii) Let the $M(M-1)/2$ pair sums $\boldsymbol{\phi}_u + \boldsymbol{\phi}_v$ be pairwise distinct, which holds for generic $\boldsymbol{\Phi}$, and let the patterns be as in (ii). If $\boldsymbol{\Xi} A(\mathbf{x}) \boldsymbol{\Phi}^\top$ is symmetric for every $\mathbf{x}$ in an open set, then $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ for some real λ and some $\mathbf{v}$.

Proof sketch. The Jacobian expands over pairs of patterns,

$$\boldsymbol{\Xi} A \boldsymbol{\Phi}^\top = \sum_{u<v} p_u p_v\, (\boldsymbol{\xi}_u - \boldsymbol{\xi}_v)(\boldsymbol{\phi}_u - \boldsymbol{\phi}_v)^\top.$$

The products $p_u p_v$ are linearly independent functions on the simplex. Symmetry for every $\mathbf{p}$ therefore makes each pair term symmetric, which means $\boldsymbol{\phi}_u - \boldsymbol{\phi}_v = \lambda_{uv}(\boldsymbol{\xi}_u - \boldsymbol{\xi}_v)$. Non-collinear patterns force every λ_{uv} to take one common value λ. Positive semidefiniteness then gives $\lambda \geq 0$.

Each score on the half-flat gives a dot-product network at inverse temperature $\beta\lambda$ with a free offset for each pattern. This family includes nearest-neighbour retrieval and the negative squared distance, which is the case $\lambda = 2$ (Corollary 3). Part (i) extends the dot-product case of Santos et al. (2025). Part (ii) is its converse when a symmetric positive semidefinite Jacobian is required for all weights $\mathbf{p}$. A fixed model produces only the weights $\mathbf{p}(\mathbf{x})$ of its queries, so part (iii) asks for symmetry only on an open set of queries. For generic $\boldsymbol{\Phi}$ off the half-flat and patterns as in (ii), the pattern readout is therefore not admissible in the Euclidean metric (Appendix B.2). In a constant metric G the half-flat passes through $G\boldsymbol{\Xi}$ instead (Corollary 2). It contains the softmax form of the kernelised network of Wu et al. (2024a). For patterns as in (ii), a generic score off all these flats has no constant-metric energy under the pattern readout. We therefore change the readout.

3.2 THE GRADIENT READOUT: EVERY SCORE CONVEX UP TO A QUADRATIC

Proposition 1: Under the gradient readout every semi-convex score gives an admissible map

Call a score family $\{s_\mu\}$ semi-convex if there is a finite $\omega \geq 0$ such that $s_\mu + \omega\|\mathbf{x}\|^2$ is convex for every μ. Let the family be semi-convex. Then

$$h(\mathbf{x}) := \tfrac{1}{\beta}\,\mathrm{lse}\left(\beta(\mathbf{s}(\mathbf{x}) + \omega\|\mathbf{x}\|^2\mathbf{1})\right)$$

is convex, and every selection $T(\mathbf{x}) \in G^{-1}\partial h(\mathbf{x})$ with $G = (1 + 2\omega)\mathbf{I}$ is G-admissible on $\mathbb{R}^d$. The added quadratic leaves $\mathrm{argmax}_\mu\ s_\mu(\mathbf{x})$ unchanged at every state.

Here h is convex because lse is convex and non-decreasing in each input, and each input $s_\mu + \omega\|\mathbf{x}\|^2$ is convex. Moreover, the scores need not be differentiable. The gradient readout itself, however, is not new. Indeed, it is the update template of dense associative memory with a free similarity (Krotov et al., 2025, Eqs. 5.7 and 5.9), and Fanaskov & Oseledets (2025) use it in continuous time. What Proposition 1 adds is a condition under which each step is concave–convex in a constant metric.

3.3 TESTING A MAP THROUGH ITS JACOBIAN EIGENVALUES

Theorem 1 applies only to splitting scores. For any other map, we test admissibility on the map itself through the eigenvalues of its Jacobian.

Lemma 1: Complex eigenvalues exclude every constant metric

If $J_T(\mathbf{x})$ has a non-real eigenvalue, no constant symmetric positive-definite G makes $GJ_T(\mathbf{x})$ symmetric.

Indeed, if GJ_T is symmetric, then J_T is similar to the symmetric matrix $G^{-1/2}(GJ_T)G^{-1/2}$ and has real eigenvalues. Convexity of h makes GJ_T positive semidefinite and these eigenvalues non-negative. A non-real or negative eigenvalue therefore excludes every constant metric. The energy of Santos et al. (2025) goes further. It replaces $\frac{1}{2}\mathbf{x}^\top G\mathbf{x}$ by a convex $\mathbf{\Psi}$, which acts like a metric that changes with the state. Proposition 2 shows that the same test excludes these energies too.

Proposition 2: Non-real or negative eigenvalues also exclude a convex $\mathbf{\Psi}$

Let h be convex, and let $\mathbf{\Psi}$ be a closed proper convex function whose conjugate $\mathbf{\Psi}^*$ is differentiable on $\nabla h(D_h)$, where D_h is the set on which h is differentiable. Let $E(\mathbf{q}) = \mathbf{\Psi}(\mathbf{q}) - h(\mathbf{q})$, and let $T = \nabla\mathbf{\Psi}^*(\nabla h)$ be the concave–convex step of E. Where h and $\mathbf{\Psi}^*$ are twice differentiable, the Jacobian of T is

$$J_T(\mathbf{x}) = \nabla^2\mathbf{\Psi}^*\big(\nabla h(\mathbf{x})\big)\ \nabla^2 h(\mathbf{x}).$$

It is a product of two symmetric positive semidefinite matrices, so its eigenvalues are real and non-negative. Hence, if J_T exists and has a non-real or negative eigenvalue at every point of an open set U, then no such $\mathbf{\Psi}$ whose conjugate is twice differentiable on $\nabla h(U)$ makes T its concave–convex step. This holds even when $\nabla^2\mathbf{\Psi}^*$ is singular.

The energy of Santos et al. (2025) is the case $h = \Omega^* \circ \mathbf{\Xi}^\top$ with Ω convex. Proposition 2 asks for the eigenvalue on an open set, but we measure it at single points. For the piecewise smooth maps we probe, J_T is continuous at a generic point, so a non-real or negative eigenvalue found there also appears at every point of a small open set around it. The test has two limits: it does not cover a metric $G(\mathbf{x})$ that varies with the state, and it cannot see a jump in a discontinuous T (Appendix D).

4 Which published retrieval maps are admissible

We apply Theorem 1 and Propositions 1 and 2 to published retrieval maps that claim an energy or are used as a layer. These maps form a curated sample, not a complete survey. Appendix E states the selection rule and its exceptions, and it gives the evidence behind every verdict.

Of the 32 maps, 13 are admissible, 2 more are admissible only under a step-size condition, 6 have a proven energy outside Definition 1, 3 are undetermined and 8 fail. The 8 failures come from 6 papers, which make different claims about an energy for their maps (Table 7). For 4 of the failures, the paper claims no energy for the map. For the 3 universal-Hopfield failures, the paper's energy argument covers a related continuous-time dynamics instead. Only the adaptive network's paper states an energy for the map itself. The failures are of four kinds: a readout that is not the gradient of the score's potential, a score that no added quadratic makes convex, a separation that does not remove the part of the score shared by all patterns, and scores divided by their sum instead of passed through a softmax. These verdicts concern iteration and say nothing about retrieval in one step. For three of the six maps with an energy outside Definition 1, no positive-definite constant metric exists either, for generic patterns in one case and at the tested configurations in the other two. The count of failures leaves them out, because each has an energy of another kind.

Which failures can cycle. A map without an energy need not cycle. We iterated six failing maps as a pooling layer for multiple-instance classification (Section 6), including the adaptive network and modified forms of two universal-Hopfield maps. Two of the six cycle and four converge (Appendix H.6). To see which maps can cycle, write $S_{ab} := s_b(\boldsymbol{\xi}_a)$ for the score of pattern b at pattern a. We call b the *best match* of a when S_{ab} is the strict maximum of row a, and the *best-match graph* has an edge from each pattern to its best match.

Proposition 3: A saturated pattern readout jumps to best matches and can cycle

Let $T(\mathbf{x}) = \mathbf{\Xi}\ \mathrm{softmax}(\beta\mathbf{s}(\mathbf{x}))$ with distinct patterns and locally Lipschitz scores.

(i) If a is its own best match, then for every large enough β the map has a fixed point near $\boldsymbol{\xi}_a$ that attracts every orbit starting close enough to $\boldsymbol{\xi}_a$.

Table 1: Published retrieval maps, by verdict. A ✓ means that the row's map is G-admissible for a constant G (Definition 1). A × means that no constant metric and no convex potential make the map admissible. Each × comes from a non-real or negative Jacobian eigenvalue at a tested configuration (Proposition 2). An asterisk marks a model that we also read from its released code. Table 8 lists the other 14 rows, and Tables 9 and 10 give the measurements.

Model	Admissible	Why it holds or fails
Modern Hopfield network (Ramsauer et al., 2021)*	✓	The score and the readout use the same patterns, the base case of Theorem 1.
Sparse / entmax (Hu et al., 2023; Santos et al., 2024; Wu et al., 2024b; Hu et al., 2024a)*	✓	The modern Hopfield network with sparsemax or entmax as its separation (Corollary 3).
Kernelised / U-Hop (Wu et al., 2024a)*	✓	Its energy uses the constant metric $G = W^\top W$, positive definite for W of full column rank.
Squared distance, softmax sep (Millidge et al., 2022)	✓	The softmax removes the $-\|\mathbf{x}\|^2$ part shared by all patterns, which puts the score on the half-flat.
Dot product, non-decreasing sep (Millidge et al., 2022)	✓	Its Jacobian is symmetric, and positive semidefinite when the separation is non-decreasing on the scores that occur.
Dense assoc. memory, continuous relaxation (Krotov & Hopfield, 2016)	✓	Admissible when the power is convex, as for rectified or even powers. Odd powers can fail on signed patterns.
ClAM clustering (Saha et al., 2023)	✓	A gradient step of size at most 1 keeps the Jacobian positive semidefinite, and every published setting uses one.
Classical / tanh (Santos et al., 2025)	✓	Admissible in the metric $G = \Xi\Xi^\top$, positive definite when the patterns span $\mathbb{R}^d$.
DeepRC (Widrich et al., 2020), HELM (Paischer et al., 2022)	✓	Both iterate the modern Hopfield map, with the same patterns in the score and the readout.
Released Hopfield layer (Ramsauer et al., 2021)*	✓	Its loop uses the keys in both the score and the readout. The values are applied once, after the loop.
Adaptive Hopfield network (Wang et al., 2026)*	×	Its pattern readout is not the gradient of the sorted score's potential, even with a monotone gate.
One-step attention, value readout (Vaswani et al., 2017)*	×	The score uses the keys but the readout uses the values.
Nonparametric (Hu et al., 2025)	×	The score uses perturbed patterns but the readout uses clean ones.
Modern Hopfield attention (Masumura & Taki, 2025)	×	Its readout weights differ from its score weights, which its authors say violates the energy argument.
Kernel memory network (Iatropoulos et al., 2022)	×	Its pseudo-inverse readout is not the gradient of the kernel score's potential.
Universal Hopfield, Manhattan (Millidge et al., 2022)*	×	The ℓ_1 score's slope drops abruptly where a coordinate matches, and no added quadratic makes such a kink convex. The gradient readout does not repair it.
Universal Hopfield, Euclidean, non-shift-invariant sep (Millidge et al., 2022)	×	The separation does not remove the $-\|\mathbf{x}\|^2$ part shared by all patterns.
Universal Hopfield as released (Millidge et al., 2022)*	×	Reciprocal distances divided by their sum instead of passed through a softmax.

(ii) If $a \neq b$ are each other's best matches, then for every large enough β the map has an attracting orbit of period two, with one point near $\boldsymbol{\xi}_a$ and one near $\boldsymbol{\xi}_b$. Then no function E satisfies $E(T\mathbf{x}) < E(\mathbf{x})$ whenever $T\mathbf{x} \neq \mathbf{x}$, and no constant metric G makes T admissible.

(iii) If S is symmetric, the best-match graph has no cycle longer than two.

(iv) If $S_{aa} \geq S_{ab}$ for all a, b, or if S is a Gram matrix $S_{ab} = \boldsymbol{\xi}_a^\top H \boldsymbol{\xi}_b$ with $H \succeq 0$, then no two patterns are each other's best matches.

(v) Suppose the walk that starts at a and follows best matches meets only patterns with a strict best match, and let it end in a cycle of length k. Then there is an $r > 0$ such that, for every large enough β, every orbit that comes within distance r of $\boldsymbol{\xi}_a$ converges to an orbit of period exactly k, with one point near each pattern of that cycle. By (iii), $k \leq 2$ when S is symmetric.

When the softmax saturates, the pattern readout jumps from each pattern to its best match. Two patterns that are each other's best matches therefore trap it in a cycle of period two. A symmetric S leaves the best-match graph no cycle longer than two, the same bound that Goles-Chacc et al. (1985) and Bruck (1990) proved for symmetric networks under parallel update. Near a pattern whose walk meets only strict best matches, a saturated readout with a symmetric S therefore has period at most two (part (v)). Most saturating maps in Table 1 cannot contain a mutual best-match pair. Neither can the order-statistic score on squared distances when its gate is non-negative, because each pattern then scores highest at its own location. A negative gate entry removes this guarantee, and the adaptive network places no sign constraint on its gate. The other failing map that cycles uses the identity separation, which does not saturate, and the proposition says nothing about it.

5 AN ADAPTIVE SCORE THAT KEEPS ITS ENERGY

Among the failures of Table 1, only the adaptive network learns a score that is not linear in the query. We construct a variant of it that keeps an exact energy at every parameter value. By an exact energy we mean that Definition 1 holds at every query, ties included.

The order-statistic score. Wang et al. (2026) score a pattern by a learned weighted sum of its sorted per-coordinate similarities. We use the squared-Euclidean base $q_i(\boldsymbol{\xi}_i, x_i) = -(\xi_i - x_i)^2$ and write $\tilde{\mathbf{q}}_\mu$ for the decreasing rearrangement of $\mathbf{q}(\boldsymbol{\xi}_\mu, \mathbf{x})$. The score and the weights are

$$s_\mu(\mathbf{x}) = \sum_{l=1}^{d} c_l \, \tilde{q}_\mu(l), \qquad \mathbf{u} = \beta \mathbf{s}, \qquad \mathbf{p} = \mathrm{softmax}(\mathbf{u}). \tag{1}$$

The model learns the gate $\mathbf{c} \in \mathbb{R}^d$. Wang et al. (2026) parameterise it by an unconstrained cumulative sum and use the pattern readout $T = \boldsymbol{\Xi}\mathbf{p}$. We keep the score and change two things. The gate is constrained to be non-increasing with entries in $[0, 1]$, and we call such a gate monotone. The readout is the gradient of the score's potential.

Theorem 2: A monotone gate gives the gradient readout an exact energy

Let $\mathbf{c} \in [0, 1]^d$ be non-increasing, and let $\kappa > 0$ with $\kappa \geq \beta \|\mathbf{c}\|_\infty$. Then

$$T(\mathbf{x}) = \mathbf{x} + \frac{\beta}{\kappa} \sum_{\mu=1}^{M} p_\mu \, \mathbf{g}^*_\mu \odot (\boldsymbol{\xi}_\mu - \mathbf{x})$$

is the concave–convex step of $E(\mathbf{x}) = -\operatorname{lse}(\mathbf{u}(\mathbf{x}))$ in the metric $G = 2\kappa\mathbf{I}$, where $\mathbf{g}^*_\mu(\mathbf{x})$ is a permutation of $\mathbf{c}$ that attains $s_\mu(\mathbf{x})$. If $c_d > 0$, E is coercive. Hence, for every $\mathbf{x}$ and every trajectory $\mathbf{x}_{k+1} = T(\mathbf{x}_k)$,

$$E(\mathbf{x}) - E(T\mathbf{x}) \geq \kappa \|T\mathbf{x} - \mathbf{x}\|^2, \quad E(\mathbf{x}) \geq -\log M, \quad \sum_{k \geq 0} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq \frac{E(\mathbf{x}_0) + \log M}{\kappa}.$$

Proof sketch. For a non-increasing gate, the rearrangement inequality (Hardy et al., 1952, Theorem 368) turns the sorted sum into a maximum over the permutations σ of $\{1, \ldots, d\}$,

$$s_\mu(\mathbf{x}) = \max_{\sigma \in \mathfrak{S}_d} \sum_{i=1}^{d} c_{\sigma(i)} \, q_i(\boldsymbol{\xi}_\mu, \mathbf{x}).$$

Multiplied by β and added to $\kappa\|\mathbf{x}\|^2$, each member of this family is separable with second derivative $2(\kappa - \beta c_{\sigma(i)}) \geq 0$ in coordinate i. A maximum of convex functions is convex, and so is the log-sum-exp of convex functions. Hence $h := \operatorname{lse}(\mathbf{u} + \kappa\|\mathbf{x}\|^2 \mathbf{1})$ is convex on all of $\mathbb{R}^d$, and $E = \kappa\|\mathbf{x}\|^2 - h$. Where the maximising permutation is unique, differentiating h with it gives $\nabla h = 2\kappa T$. Where several permutations tie, each of them still gives a subgradient of h (Appendix C). In both cases $2\kappa T \in \partial h$, which is Definition 1 with $G = 2\kappa\mathbf{I}$.

Theorem 2 constrains the trained parameters, so the energy survives training only if the parameterisation keeps them inside its hypotheses.

Corollary 1: The energy holds throughout training

Let $c_l = f + (1 - f) \sum_{j \geq l} w_j$, with $\mathbf{w} = \mathrm{softmax}(\boldsymbol{\theta})$ and a floor $f \in (0, 1)$. For every $\boldsymbol{\theta} \in \mathbb{R}^d$ the gate is non-increasing, with entries in $[f, 1]$ and $c_1 = 1$. Setting $\kappa = \beta$ after every gradient step therefore meets the hypotheses of Theorem 2 at every $\boldsymbol{\theta}$ and every learned $\beta > 0$, so the energy exists at initialisation and after every gradient step.

The floor gives $c_d \geq f$ and hence coercivity, and $\kappa = \beta$ is the smallest κ that Theorem 2 allows. We use the squared-distance base because under $E = -\operatorname{lse}(\mathbf{u})$ the dot-product base is unbounded below for every non-zero gate that meets these hypotheses (Proposition 5). The construction carries over to sparsemax, α-entmax and a full-column-rank linear modulation (Proposition 4).

Table 2: Each prediction and its test on a map without an energy guarantee. In the first row that map is the same score and gate under the pattern readout, on synthetic regimes (Figure 4). In the other rows it is the adaptive network, over 500 training runs per model, with seeds and datasets that vary by row (Appendix H). A dash for our map marks a test that does not apply, since it never cycles. A paired sign-flip test with a Holm correction confirms the gaps in convergence and cycling on all five datasets, with every adjusted p at most 0.0035 (Table 18).

Prediction	From	Statistic	No guarantee	**Ours**
What the energy guarantees				
No step raises the energy	Theorem 2	synthetic regimes in which E rises at some step	24 of 27	**0 of 27**
An admissible map cannot cycle	Definition 1	training runs in which some held-out bag cycles	250 of 500	**0 of 500**
Why the adaptive network cycles				
A non-negative gate leaves no mutual best-match pair	Proposition 3(iv)	runs with a non-negative trained gate in which some bag cycles	0 of 142	—
A cycle of period two sits on a mutual best-match pair	Proposition 3(iii), (v)	cycling bags whose two orbit points lie at a mutual best-match pair	97–100%	—
What a cycle costs				
The implicit gradient is guaranteed exact only at a fixed point	implicit function theorem	held-out bags whose implicit and unrolled gradients differ by more than 10%	35–47%	**at most 0.1%**
A cycle can change the decision	—	held-out bags whose predicted label differs between steps 999 and 1000	9–27%	**0%**

Where the construction stops. The order of the gate in Theorem 2 cannot be dropped. If $\mathbf{c}$ is not non-increasing, then for patterns in general position no κ makes h convex (Proposition 6). Every base that is differentiable, non-positive and ρ-weakly convex keeps the energy with $\kappa \geq \beta\rho\|\mathbf{c}\|_\infty/2$ (Corollary 4). The Manhattan base $q_i = -|\xi_i - x_i|$ breaks the construction, because its slope drops abruptly at $x_i = \xi_i$ and no added quadratic makes such a kink convex. For any gate with $c_1 > 0$, no κ and no constant metric make its gradient readout admissible (Proposition 7).

6 Experiments

The experiments test the predictions of Sections 4 and 5, and Table 2 collects them with two further tests. All baselines are our implementations of the published equations, and three of them match the released code to machine precision (Appendix G).

Setup. We run two tests. The descent test iterates one monotone-gated order-statistic score under both readouts on synthetic patterns, in 27 regimes of $(d, M, \text{mask}, \beta)$ (Appendix G.2). The main test uses the retrieval map as the pooling layer in bag classification, the task on which earlier Hopfield papers evaluate their layers (Ramsauer et al., 2021; Hu et al., 2023; Santos et al., 2024; Wang et al., 2026). Its datasets are MUSK1 (Dietterich et al., 1997) and the four of the adaptive network's benchmark: Tiger, Fox and Elephant (Andrews et al., 2003) and UCSB (Gelasca et al., 2008). The models differ only in the retrieval map T, and the adaptive network uses only its squared-distance base, so its score is exactly ours. Statistics treat each seed as one sample.

All six predictions hold. For our map, the zeros in the first two rows follow from Theorem 2. In contrast, a map without an energy can cycle but need not (Section 4), so the counts of the maps without a guarantee are measurements, not predictions. The third and fourth rows, however, match Proposition 3 only empirically, and Figure 10 in Appendix H.3 shows the third row run by run on four of its five datasets. First, the proposition assumes a large β, but the trained β is finite. Second, these rows rely on the converse of part (ii): a cycle needs a mutual pair. Parts (iii) and (v) prove this converse only for large β, symmetric S and orbits that come close to a pattern whose walk has strict best matches (Appendix C.1). Finally, the gradient row shows that the adaptive network's implicit gradient is often inaccurate. This does not mean that the network fails to train (Appendix H.8).

The gate's sign alone stops the cycling. Our map differs from the adaptive network in both the sign of its gate and its readout. To tell which difference matters, we test three more maps that change

Table 3: A non-negative gate alone stops the cycling. Each map is a multiple-instance pooling layer, trained with the map unrolled $K = 5$ times at a learning rate of 0.1 on the Hopfield parameters and then iterated to $K' = 1000$. The first row is the adaptive network, the fifth is ours, and the last is the adaptive network's released configuration. Each entry is the range over five datasets of the twenty-seed mean. Converged, cycles and diverged are fractions of held-out bags, and a run that cycles has at least one cycling bag (Appendix H). Bold marks the maps that never cycle.

Gate	Readout	Converged	Cycles	Diverged	Runs that cycle	Iterations
signed	pattern	.483–.618	.381–.516	.000	250 of 500	2.0–3.1
signed	gradient	.041–.364	.000–.011	.540–.901	31 of 500	53.3–441.0
clipped at zero	pattern	.997–1.000	**.000**	.000	**0 of 500**	2.0–3.4
monotone	pattern	.999–1.000	**.000**	.000	**0 of 500**	4.8–13.3
monotone	gradient	.999–1.000	**.000**	.000	**0 of 500**	13.2–148.1
released	pattern	.445–.821	.178–.555	.000	189 of 500	4.0–11.1

one at a time (Table 3). Under the pattern readout, a non-negative gate removes the cycling, whether it is our monotone gate or the adaptive network's gate clipped at zero. The gaps in convergence and cycling are confirmed on all five datasets, consistent with Proposition 3(iv). The third map, the signed gate under the gradient readout, diverges in most held-out bags. A non-negative gate would prevent this, whatever the order of its entries: at its κ, every orbit would stay inside the box spanned by the patterns and the initial state. The negative entries of this gate remove that guarantee. In these runs, then, the energy adds a guarantee of descent but not convergence. Unlike the effect of the sign, this descent holds at every β and not only when the softmax saturates.

The cycling depends on the learning rate, not the base. Our setup departs from the adaptive network's released code in five ways, and none of them causes the cycling. Its released configuration restores four of them and still cycles (Table 3). With the dot-product base restored, the adaptive network at the same rate still cycles in 37 to 51% of held-out bags (Table 22). Unrolled once in training, as in its original code, it still cycles in 52 to 78% of held-out bags when iterated (Appendix H.2). The learning rate, however, does matter. We use a learning rate of 0.1 for every Hopfield parameter, as the adaptive network's paper does for its gate in retrieval. At 10^{-4} its squared-distance form cycles in no bag and loses no held-out AUC. Under the multiple-instance recipe of its own paper, taken from Hu et al. (2023), this form still cycles in 4.3 to 24.1% of held-out bags in our pipeline. Every one of the 443 runs that cycle under these recipes has the negative gate entry that Proposition 3(iv) requires for a mutual best-match pair (Appendix H.5).

Accuracy and drift are comparable. At the training depth, our map matches or exceeds the adaptive network's accuracy on every dataset. However, we tested accuracy for equivalence, not for superiority, so we claim no accuracy advantage (Appendix H.2). We also measure depth drift, the accuracy lost when the map runs past its training depth. The two maps drift about equally, except on Elephant, where a fifteen-seed test finds that ours drifts less. Finally, our map and the modern Hopfield network differ in none of the statistics we test.

7 Cost, scope and open questions

Our map pays for its energy in iterations. To measure this, we count applications of the map at inference on synthetic retrieval, where the two learned models also differ in their training details (Appendix G.4). At one step our map trails the adaptive network by up to 28.2 points. After 30 steps it is within one standard deviation at masks 0.5 and 0.6 and trails by 2.4 ± 1.4 points at mask 0.7 (Table 13), but at about 16 times the cost per query. We expect slow contraction to cause both the gap and the cost. In particular, at a stored pattern that wins by a margin, the continuous extension of our map's Jacobian within a region of fixed rank order has norm tending to $1 - c_d/\|\mathbf{c}\|_\infty$ as β grows (Theorem 3(iii)), which is close to one for the steep gates that training selects. Besides iterations, our map also has a slower implicit backward solve (Appendix H.10) and loses accuracy at mask 0.8 on MNIST (Appendix G.5). As for scope, the energy guarantees descent, not retrieval (Appendix G.3). Finally, two questions remain open: whether a monotone gate suffices under a state-dependent metric, and whether a preconditioned step can cut the iteration cost while keeping the descent inequality.

Reproducibility statement

Complete proofs of every result are in Appendices B and C, and Appendix D gives a numerical check of each. Experimental settings, hyperparameters and dataset details are in Appendices G and H, and Appendix H also lists every experimental protocol and every departure from it.

AI use statement

We used generative AI tools in this work. For the theory, they helped develop theoretical models, formulate and check mathematical claims, supply proof ingredients and write proofs. For the experiments, they helped refine hypotheses, design experiments and draft the pre-registered protocols, implement methods and verification code, including the code that generates the synthetic data, and interpret results. We also used them for literature search, for auditing our reimplementations against the original authors' released code, and for editing the manuscript. We did not use them to generate the research question. Translation, dataset cleaning and qualitative data analysis do not apply to this work. We have reviewed all AI-assisted work. Every proof was re-derived by the authors and is accompanied by a runnable numerical check with a negative control. All AI-assisted code was checked by independent tests, and all citations were verified against the cited source. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Susmit Agrawal, Krishn Vishwas Kher, Saksham Mittal, Swarnim Maheshwari, and Vineeth N. Balasubramanian. Memory-integrated reconfigurable adapters: A unified framework for settings with multiple tasks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2512.00940.

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, volume 15, 2003.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. arXiv:1909.01377.

Nicholas Barnfield, Juno Kim, Eshaan Nichani, Jason D. Lee, and Yue M. Lu. Sharp capacity thresholds in linear associative memory: From top-1 retrieval to tail-average learning. *arXiv preprint arXiv:2605.05189*, 2026.

Sumeet Batra and Gaurav S. Sukhatme. Zero shot generalization of vision-based RL without data augmentation. In *International Conference on Machine Learning*, volume 267 of *PMLR*, 2025. arXiv:2410.07441.

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès. SLOPE—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103–1140, 2015. doi: 10.1214/15-AOAS842. arXiv:1407.3824.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

Jehoshua Bruck. On the convergence properties of the Hopfield model. *Proceedings of the IEEE*, 78(10):1579–1585, 1990.

Thomas F. Burns and Tomoki Fukai. Simplicial Hopfield networks. In *International Conference on Learning Representations*, 2023. arXiv:2305.05179.

Piermarco Cannarsa and Carlo Sinestrari. *Semiconcave Functions, Hamilton–Jacobi Equations, and Optimal Control*, volume 58 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Boston, 2004. doi: 10.1007/b138356.

Franco Cardin, Alberto Lovison, Amos Maritan, and Aram Megighian. A new mathematical model for brain memory working. Optimal control behavior for Hopfield networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(10):103501, 2025.

Hamza Tahir Chaudhry, Jacob A. Zavatone-Veth, Dmitry Krotov, and Cengiz Pehlevan. Long sequence Hopfield memory. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2306.04532.

John M. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarathkrishna Swaminathan, Sihui Dai, Aurélie Lozano, Georgios Kollias, Vijil Chenthamarakshan, Jiří Navrátil, Soham Dan, and Pin-Yu Chen. Larimar: Large language models with episodic memory control. In *International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 10109–10126, 2024. arXiv:2403.11901.

Nima Dehmamy, Benjamin Hoover, Bishwajit Saha, Leo Kozachkov, Jean-Jacques Slotine, and Dmitry Krotov. NRGPT: An energy-based alternative for GPT. In *International Conference on Learning Representations*, 2026. arXiv:2512.16762.

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*, 168:288–299, 2017. arXiv:1702.01929.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997.

Vladimir Fanaskov and Ivan Oseledets. Associative memory and dead neurons. In *International Conference on Learning Representations*, 2025. arXiv:2410.13866.

Elisa Drelie Gelasca, Jiyun Byun, Boguslaw Obara, and B. S. Manjunath. Evaluation and benchmark for biological image segmentation. In *IEEE International Conference on Image Processing*, 2008.

Alexi Gladstone, Ganesh Nanduru, Md Mofijul Islam, Peixuan Han, Hyeonjeong Ha, Aman Chadha, Yilun Du, Heng Ji, Jundong Li, and Tariq Iqbal. Energy-based transformers are scalable learners and thinkers. In *International Conference on Learning Representations*, 2026. arXiv:2507.02092.

Eric Goles-Chacc, Françoise Fogelman-Soulié, and Didier Pellegrin. Decreasing energy functions as a tool for studying threshold networks. *Discrete Applied Mathematics*, 12(3):261–277, 1985.

Boliang Hao, Bailing Zhang, and Fangyu Wu. Hyperbolic associative memory networks. In *International Conference on Machine Learning*, 2026.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, second edition, 1952.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed J. Zaki, and Dmitry Krotov. Energy transformer. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2302.07253.

Benjamin Hoover, Duen Horng Chau, Hendrik Strobelt, Parikshit Ram, and Dmitry Krotov. Dense associative memory through the lens of random features. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2410.24153.

Benjamin Hoover, Zhaoyang Shi, Krishnakumar Balasubramanian, Dmitry Krotov, and Parikshit Ram. Dense associative memory with Epanechnikov energy. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2506.10801.

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.

John J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092, 1984.

Jerry Yao-Chieh Hu, Donglin Yang, Dennis Wu, Chenwei Xu, Bo-Yu Chen, and Han Liu. On sparse modern Hopfield model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2309.12673.

Jerry Yao-Chieh Hu, Pei-Hsuan Chang, Haozheng Luo, Hong-Yu Chen, Weijian Li, Wei-Po Wang, and Han Liu. Outlier-efficient Hopfield layers for large transformer-based models. In *International Conference on Machine Learning*, volume 235 of *PMLR*, 2024a. arXiv:2404.03828.

Jerry Yao-Chieh Hu, Dennis Wu, and Han Liu. Provably optimal memory capacity for modern Hopfield models: Transformer-compatible dense associative memories as spherical codes. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. arXiv:2410.23126.

Jerry Yao-Chieh Hu, Bo-Yu Chen, Dennis Wu, Feng Ruan, and Han Liu. Nonparametric modern Hopfield models. In *International Conference on Machine Learning*, volume 267 of *PMLR*, 2025. arXiv:2404.03900.

Yunzhe Hu, Difan Zou, and Dong Xu. Hyper-SET: Designing transformers via hyperspherical energy minimization. In *International Conference on Learning Representations*, 2026.

Georgios Iatropoulos, Johanni Brea, and Wulfram Gerstner. Kernel memory networks: A unifying framework for memory modeling. In *Advances in Neural Information Processing Systems*, volume 35, 2022. arXiv:2208.09416.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, volume 80 of *PMLR*, 2018.

Daniel Jiwoong Im, Mohamed Ishmael Belghazi, and Roland Memisevic. Conservativeness of untied auto-encoders. In *AAAI Conference on Artificial Intelligence*, 2016.

Arjun Karuvally, Terrence J. Sejnowski, and Hava T. Siegelmann. General sequential episodic memory model. In *International Conference on Machine Learning*, volume 202 of *PMLR*, 2023.

Arjun Karuvally, Pichsinee Lertsaroj, Terrence J. Sejnowski, and Hava T. Siegelmann. Exponential dynamic energy network for high capacity sequence memory. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2510.24965.

Dmitry Krotov. Hierarchical associative memory. *arXiv preprint arXiv:2107.06446*, 2021.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, volume 29, 2016. arXiv:1606.01164.

Dmitry Krotov and John J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021. arXiv:2008.06996.

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

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

André F. T. Martins and Ramón F. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, volume 48 of *PMLR*, 2016.

Tsubasa Masumura and Masato Taki. On the role of hidden states of modern Hopfield network in transformer. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2511.20698.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal Hopfield networks: A general framework for single-shot associative memory models. In *International Conference on Machine Learning*, volume 162 of *PMLR*, 2022. arXiv:2202.04557.

Fabian Paischer, Thomas Adler, Vihang Patil, Angela Bitto-Nemling, Markus Holzleitner, Sebastian Lehner, Hamid Eghbal-zadeh, and Sepp Hochreiter. History compression via language models in reinforcement learning. In *International Conference on Machine Learning*, volume 162 of *PMLR*, 2022.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

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. In *International Conference on Learning Representations*, 2021. arXiv:2008.02217.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1998.

Bishwajit Saha, Dmitry Krotov, Mohammed J. Zaki, and Parikshit Ram. End-to-end differentiable clustering with associative memories. In *International Conference on Machine Learning*, volume 202 of *PMLR*, 2023. arXiv:2306.03209.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *PMLR*, 2022.

Saul Santos, Vlad Niculae, Daniel McNamee, and André F. T. Martins. Sparse and structured Hopfield networks. In *International Conference on Machine Learning*, volume 235 of *PMLR*, 2024. arXiv:2402.13725.

Saul Santos, Vlad Niculae, Daniel McNamee, and André F. T. Martins. Hopfield-Fenchel-Young networks: A unified framework for associative memory retrieval. *Journal of Machine Learning Research*, 26(265):1–51, 2025. arXiv:2411.08590.

Matthew Smart, Alberto Bietti, and Anirvan M. Sengupta. In-context denoising with one-layer transformers: connections between attention and associative memory retrieval. In *International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 55950–55971, 2025. arXiv:2502.05164.

Akiyoshi Tomihari and Ryo Karakida. Recurrent self-attention dynamics: An energy-agnostic perspective from Jacobians. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2505.19458.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shurong Wang, Yuqi Pan, Zhuoyang Shen, Meng Zhang, Hongwei Wang, and Guoqi Li. Adaptive Hopfield network: Rethinking similarities in associative memory. In *International Conference on Learning Representations*, 2026. arXiv:2511.20609.

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, and Günter Klambauer. Modern Hopfield networks and attention for immune repertoire classification. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ezra Winston and J. Zico Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. arXiv:2006.08591.

Dennis Wu, Jerry Yao-Chieh Hu, Teng-Yun Hsiao, and Han Liu. Uniform memory retrieval with larger capacity for modern Hopfield models. In *International Conference on Machine Learning*, volume 235 of *PMLR*, 2024a. arXiv:2404.03827.

Dennis Wu, Jerry Yao-Chieh Hu, Weijian Li, Bo-Yu Chen, and Han Liu. STanHop: Sparse tandem Hopfield model for memory-enhanced time series prediction. In *International Conference on Learning Representations*, 2024b. arXiv:2312.17346.

Weimin Wu, Teng-Yun Hsiao, Jerry Yao-Chieh Hu, Wenxin Zhang, and Han Liu. In-context learning as conditioned associative memory retrieval. In *International Conference on Machine Learning*, volume 267 of *PMLR*, 2025.

Ronald R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190, 1988. doi: 10.1109/21.87068.

Alan L. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15 (4):915–936, 2003.

Xiangrong Zeng and Mário A. T. Figueiredo. Decreasing weighted sorted ℓ_1 regularization. *IEEE Signal Processing Letters*, 21(10):1240–1244, 2014. doi: 10.1109/LSP.2014.2331977. arXiv:1404.3184.

Jianyu Zhang, Niklas Nolte, Ranajoy Sadhukhan, Beidi Chen, and Léon Bottou. Memory mosaics. In *International Conference on Learning Representations*, 2025. arXiv:2405.06394.

APPENDIX CONTENTS

A Notation and the two readouts **17**

B Proofs for Sections 2 and 3 **17**

B.1 The descent inequality, the constant metric and the Jacobian condition 17

B.2 Proof of Theorem 1, and the members of the flat 19

B.3 Proof of Proposition 1 23

B.4 Proofs of Lemma 1 and Proposition 2 24

B.5 The escape families of Theorem 1(iii) 25

C Proofs for Sections 4 and 5 **25**

C.1 Proof of Proposition 3 (the saturated pattern readout) 25

C.2 Proof of Theorem 2 and Corollary 1 (the energy of the construction) 27

C.3 Theorem 3: the Jacobian of the gradient readout 28

C.4 Other separations, modulations and bases 30

C.5 The order of the gate cannot be dropped 32

D Additional numerical experiments **33**

E The published maps, row by row **36**

E.1 The population and how it was assembled 36

E.2 Admissible rows 41

E.3 Rows admissible only under a step-size condition 43

E.4 Rows with no energy guarantee 44

E.5 Rows with an energy outside Definition 1 46

E.6 Undetermined rows, and models covered by a row 47

E.7 One repair, four models, four classes 48

F Related work **49**

G Synthetic experiments **50**

G.1 Baselines 50

G.2 Descent and the two readouts 50

G.3 Basins, spurious points and capacity 51

G.4 Retrieval against the inference budget 54

G.5 MNIST 55

H Multiple-instance learning **56**

H.1 Protocol 56

H.2 Main results 59

H.3 Mechanism and label flips 63
H.4 The control: gate sign against readout 64
H.5 The training recipe and the learning rate 67
H.6 Other maps iterated 68
H.7 MUSK1 70
H.8 Implicit training 71
H.9 Forward solvers 72
H.10 Cost and scope in full 74

A NOTATION AND THE TWO READOUTS

Table 4 collects the notation of Section 2 and of the later sections, and Figure 2 draws the two readouts of Section 2 on one query.

Table 4: Notation.

Symbol	Meaning
$\mathbf{x} \in \mathbb{R}^d$	query, and the state of the iterated map
$\boldsymbol{\xi}_\mu, \boldsymbol{\Xi} = [\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M]$	stored patterns, M of them
$s_\mu(\mathbf{x}), S_{ab} := s_b(\boldsymbol{\xi}_a)$	score of pattern μ at $\mathbf{x}$, and the score matrix at the patterns
$\beta, \mathbf{u} = \beta\mathbf{s}$	inverse temperature, scaled scores
$\mathbf{p} = \mathrm{softmax}(\mathbf{u}), A = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$	separation weights and their covariance
$T_{\boldsymbol{\Xi}} = \boldsymbol{\Xi}\mathbf{p}, T_\nabla = G^{-1}\nabla h$	pattern readout, gradient readout
$G, h, E = \frac{1}{2}\mathbf{x}^\top G\mathbf{x} - h$	constant metric, convex potential, energy (Definition 1)
$\boldsymbol{\Phi}, \boldsymbol{\psi}$	per-pattern and shared parts of a splitting score's gradient
$J_T, \mathrm{asym}(J) = \|J - J^\top\|/\|J\|$	Jacobian of T, relative asymmetry
$\boldsymbol{\Psi}$	convex post-transform of a Fenchel–Young energy (Proposition 2)
$\mathbf{c}, \mathbf{g}^*_\mu$	gate, and the permutation of it that attains s_μ (equation (1))
q_i, ρ	per-coordinate base similarity, its weak-convexity modulus
κ	metric scale in $G = 2\kappa\mathbf{I}$, equal to $\beta\|\mathbf{c}\|_\infty$ in our experiments
$f, \boldsymbol{\theta}$	floor and parameters of the monotone gate (Corollary 1)
K, K'	applications of the map in training and at inference

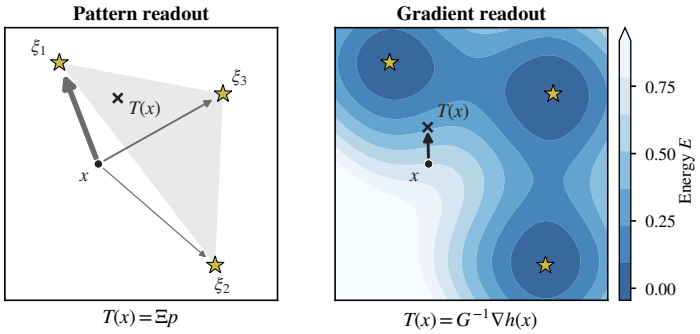

Figure 2: Both readouts weight three stored patterns (stars) by $\mathbf{p} = \mathrm{softmax}(\beta\mathbf{s}(\mathbf{x}))$ for a query $\mathbf{x}$, and here $s_\mu(\mathbf{x})$ is the negative of a smoothed Manhattan distance from $\mathbf{x}$ to $\boldsymbol{\xi}_\mu$ ($\varepsilon = 0.8$, $\beta = 2$). The pattern readout (left) averages the patterns under these weights, so its output $T(\mathbf{x})$ lies inside their triangle whatever the score. The gradient readout (right) moves $\mathbf{x}$ by $-G^{-1}\nabla E(\mathbf{x})$, one step down the energy $E(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top G\mathbf{x} - h(\mathbf{x})$. Arrow widths show the weights, and the blue shading is darker where E is lower. A search over constant metrics finds none that gives the left map an energy. Its panel is therefore not shaded.

B PROOFS FOR SECTIONS 2 AND 3

B.1 THE DESCENT INEQUALITY, THE CONSTANT METRIC AND THE JACOBIAN CONDITION

The descent inequality stated in Definition 1 is the following lemma. Theorem 2 uses it with $G = 2\kappa\mathbf{I}$.

Lemma 2: Each step lowers the energy by at least half its squared length

Let $h : \mathcal{D} \to \mathbb{R}$ be convex, let G be a constant symmetric positive-definite matrix, and let $E(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top G\mathbf{x} - h(\mathbf{x})$. If $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ and $G\mathbf{y} \in \partial h(\mathbf{x})$, then $E(\mathbf{x}) - E(\mathbf{y}) \geq \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2_G$.

Proof.

$$\begin{aligned} E(\mathbf{x}) - E(\mathbf{y}) &= \tfrac{1}{2}\mathbf{x}^\top G\mathbf{x} - \tfrac{1}{2}\mathbf{y}^\top G\mathbf{y} - h(\mathbf{x}) + h(\mathbf{y}) \\ &\overset{(a)}{\geq} \tfrac{1}{2}\mathbf{x}^\top G\mathbf{x} - \tfrac{1}{2}\mathbf{y}^\top G\mathbf{y} + \mathbf{y}^\top G(\mathbf{y} - \mathbf{x}) \\ &\overset{(b)}{=} \tfrac{1}{2}\|\mathbf{y} - \mathbf{x}\|_G^2, \end{aligned}$$

where (a) is the subgradient inequality $h(\mathbf{y}) \geq h(\mathbf{x}) + (G\mathbf{y})^\top(\mathbf{y} - \mathbf{x})$ and (b) completes the square. □

With $\mathbf{y} = T(\mathbf{x})$ this is the concave–convex argument of Yuille & Rangarajan (2003) with the rate kept. Theorem 2 is the case $G = 2\kappa\mathbf{I}$, where $\frac{1}{2}\|\cdot\|_G^2 = \kappa\|\cdot\|^2$.

Why the metric is constant. Two features of the definition are deliberate. First, the containment is a subgradient, so h need not be differentiable. The construction of Section 5 needs this, because its potential has a kink on a null set. Second, the metric is constant. The proven energies of the Lagrangian networks of Krotov & Hopfield (2021), the Energy Transformer (Hoover et al., 2023) and the ℓ_2-normalised post-transform of Santos et al. (2025) use a state-dependent metric or a convex post-transform. Definition 1 therefore does not certify them. For the ℓ_2 post-transform a numerical search over constant metrics finds none (Appendix D).

Definition 1 cannot simply let the metric vary with the state, because a convex potential stepped in a state-dependent metric need not descend. Take $h(\mathbf{x}) = \|\mathbf{x}\|^2$ and $G(\mathbf{x}) = \|\mathbf{x}\|^2\mathbf{I}$ on the half-space $x_1 > 0$. The step

$$G(\mathbf{x})^{-1}\nabla h(\mathbf{x}) = \frac{2\mathbf{x}}{\|\mathbf{x}\|^2}$$

is the inversion in the sphere $\|\mathbf{x}\|^2 = 2$. Every orbit off that sphere has period two, and no function decreases at every step of such an orbit. The three models above descend for reasons that a varying metric alone does not supply. The first two descend in continuous time. The third takes a concave–convex step with one fixed convex post-transform $\mathbf{\Psi}$ (Proposition 2) in place of the quadratic.

The phrase "no energy guarantee" means that no theorem here or in the map's source gives the map an energy. For a row of Table 1 it means more. No constant metric and no convex h make the map admissible, and we say which certificate shows it.

Lemma 3: A locally Lipschitz field is a gradient exactly when its Jacobian is symmetric

Let $\mathcal{D}$ be open and convex and $F : \mathcal{D} \to \mathbb{R}^d$ locally Lipschitz. Then $F = \nabla h$ for some $h \in C^{1,1}_{\mathrm{loc}}(\mathcal{D})$ if and only if the Jacobian $J_F(\mathbf{x})$ is symmetric for almost every $\mathbf{x} \in \mathcal{D}$. Moreover h is convex if and only if $J_F \succeq 0$ almost everywhere.

Proof. The converse direction mollifies F on shrinking convex subsets of $\mathcal{D}$, integrates the smooth field on each, and passes to the limit. By Rademacher's theorem F is differentiable almost everywhere.

(⇒) Where h is twice differentiable, $J_F = \nabla^2 h$ is symmetric. The reason is that for a Lipschitz ∇h the distributional Hessian is symmetric and coincides with the almost-everywhere derivative.

(⇐) Step 1: erosion and mollification. For $\varepsilon > 0$ let ρ_ε be a standard mollifier supported in the ball of radius ε, and let

$$\mathcal{D}_\varepsilon := \bigcap_{\|\mathbf{v}\| \leq \varepsilon} (\mathcal{D} - \mathbf{v}), \qquad F_\varepsilon := F * \rho_\varepsilon \ \text{ on } \mathcal{D}_\varepsilon.$$

The eroded set $\mathcal{D}_\varepsilon$ is open and convex, and $\mathcal{D} = \bigcup_{\varepsilon > 0} \mathcal{D}_\varepsilon$. The field F_ε is smooth, and its Jacobian

$$J_{F_\varepsilon} = J_F * \rho_\varepsilon$$

is symmetric. This step uses local Lipschitz continuity, since convolution commutes with the weak derivative and $F \in W^{1,\infty}_{\mathrm{loc}}$ makes the weak and the almost-everywhere derivatives coincide.

Step 2: a potential on each eroded set. The Poincaré lemma on the convex set $\mathcal{D}_\varepsilon$ gives h_ε with $\nabla h_\varepsilon = F_\varepsilon$, normalised at a fixed base point.

Step 3: the limit. As $\varepsilon \to 0$, $F_\varepsilon \to F$ locally uniformly. Hence h_ε converges locally uniformly to an h with $\nabla h = F$ on $\mathcal{D}$.

Convexity. If $J_F \succeq 0$ almost everywhere, the mollified potential $h * \rho_\varepsilon$ has Hessian $J_F * \rho_\varepsilon \succeq 0$ everywhere on $\mathcal{D}_\varepsilon$ and is convex there. It converges to h locally uniformly, and a locally uniform limit of convex functions is convex. Conversely, a convex h has a monotone gradient. Hence $\mathbf{v}^\top J_F(\mathbf{x})\mathbf{v} \geq 0$ for every $\mathbf{v}$ at every $\mathbf{x}$ where F is differentiable. □

Definition 1 is a subgradient containment, and Lemma 3 is about an equality. The next lemma joins them. It is the statement after Lemma 3 in Section 2.

Lemma 4: Admissibility is a Jacobian condition

Let $\mathcal{D}$ be open and convex, let G be a constant symmetric positive-definite matrix, and let $T : \mathcal{D} \to \mathcal{D}$ be locally Lipschitz. Then T is G-admissible if and only if GJ_T is symmetric positive semidefinite almost everywhere. In that case the potential h of Definition 1 is continuously differentiable with $\nabla h = GT$.

Proof. (⇒) Let h be the convex potential, and let $\mathbf{g} := GT$, a continuous selection from ∂h. Fix $\mathbf{x}$ and a direction $\mathbf{w}$. For $t > 0$ the subgradient inequalities at $\mathbf{x}$ and at $\mathbf{x} + t\mathbf{w}$ give

$$\mathbf{g}(\mathbf{x})^\top \mathbf{w} \;\leq\; \frac{h(\mathbf{x}+t\mathbf{w}) - h(\mathbf{x})}{t} \;\leq\; \mathbf{g}(\mathbf{x}+t\mathbf{w})^\top \mathbf{w}.$$

Letting $t \downarrow 0$ and using the continuity of $\mathbf{g}$ gives $h'(\mathbf{x}; \mathbf{w}) = \mathbf{g}(\mathbf{x})^\top \mathbf{w}$. The directional derivative is linear in $\mathbf{w}$, so the convex h is differentiable at $\mathbf{x}$ with $\nabla h(\mathbf{x}) = \mathbf{g}(\mathbf{x})$. Hence h is continuously differentiable, and $\nabla h = GT$ is locally Lipschitz. Lemma 3 with $F = GT$ gives that $J_F = GJ_T$ is symmetric almost everywhere. It is positive semidefinite almost everywhere because h is convex.

(⇐) The field $F := GT$ is locally Lipschitz, and $J_F = GJ_T$ is symmetric positive semidefinite almost everywhere. Lemma 3 gives a convex h with $\nabla h = GT$. Then $GT(\mathbf{x}) \in \partial h(\mathbf{x})$ at every $\mathbf{x}$. □

Every negative row of Section 4 uses the forward direction. The Lipschitz hypothesis excludes the discontinuous T of Section 5, which we prove admissible directly. The last lemma of this subsection is the factorisation that Section 2 states.

Lemma 5: The Jacobian of a fixed readout factors through the softmax

Let $\mathbf{p} = \mathrm{softmax}(\beta \mathbf{s}(\mathbf{x}))$, $A := \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$, and let $\nabla \mathbf{s} \in \mathbb{R}^{M \times d}$ have rows $\nabla_{\mathbf{x}} s_\mu(\mathbf{x})^\top$. For a readout $T(\mathbf{x}) = V\mathbf{p}$ with $V \in \mathbb{R}^{d \times M}$ constant in $\mathbf{x}$,

$$J_T = \beta\, V A\, \nabla \mathbf{s}, \qquad A\mathbf{1} = 0, \qquad \text{and if the scores split then } A\,\nabla \mathbf{s} = A\boldsymbol{\Phi}^\top.$$

Proof. The softmax has $\partial \mathbf{p}/\partial \mathbf{u} = A$, and $\partial \mathbf{u}/\partial \mathbf{x} = \beta \nabla \mathbf{s}$. The chain rule gives the product, since V is constant in $\mathbf{x}$. Next,

$$A\mathbf{1} = \mathbf{p} - \mathbf{p}\,(\mathbf{p}^\top \mathbf{1}) = \mathbf{p} - \mathbf{p} = 0.$$

If row μ of $\nabla \mathbf{s}$ is $\boldsymbol{\phi}_\mu^\top + \boldsymbol{\psi}(\mathbf{x})^\top$, then $\nabla \mathbf{s} = \boldsymbol{\Phi}^\top + \mathbf{1}\boldsymbol{\psi}^\top$ and

$$A\,\nabla \mathbf{s} = A\boldsymbol{\Phi}^\top + (A\mathbf{1})\,\boldsymbol{\psi}^\top = A\boldsymbol{\Phi}^\top.$$ □

B.2 Proof of Theorem 1, and the members of the flat

Theorem 1 (restated): The pattern readout generically admits only a half-flat of splitting scores, full statement

Let the scores split with per-pattern gradients $\boldsymbol{\Phi}$, and let $T(\mathbf{x}) = \boldsymbol{\Xi}\,\mathrm{softmax}(\beta \mathbf{s}(\mathbf{x}))$.

(i) If $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ for some $\lambda \geq 0$ and $\mathbf{v} \in \mathbb{R}^d$, then T is $\mathbf{I}$-admissible. For $\lambda > 0$ its decision rule is $\mathrm{argmax}_\mu(\lambda \boldsymbol{\xi}_\mu^\top \mathbf{x} + a_\mu)$ for the score's per-pattern offsets a_μ, which $\boldsymbol{\Phi}$ does not constrain: equal offsets give the dot product's rule, and $a_\mu = -\frac{\lambda}{2}\|\boldsymbol{\xi}_\mu\|^2$ gives nearest neighbour. For $\lambda = 0$ the map is constant.

(ii) Let the patterns be pairwise distinct and, if $M \geq 3$, not all collinear. Then $\boldsymbol{\Xi} A \boldsymbol{\Phi}^\top$ is symmetric positive semidefinite for *every* $\mathbf{p}$ in the probability simplex if and only if $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ with $\lambda \geq 0$. For $M \geq 2$ the admissible set, uniformly in the weights, is a closed half-flat of dimension $d + 1$.

(iii) Pointwise in $\mathbf{x}$, symmetry of $\boldsymbol{\Xi} A(\mathbf{x}) \boldsymbol{\Phi}^\top$ on an open set holds if and only if, for each distinct value $\mathbf{w}$ of the pair sums $\boldsymbol{\phi}_u + \boldsymbol{\phi}_v$, the antisymmetric parts of the corresponding pair terms, weighted by $e^{\beta(a_u + a_v)}$, cancel. If the $M(M-1)/2$ pair sums are pairwise distinct, which holds for generic $\boldsymbol{\Phi}$, this forces $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ for some real λ and some $\mathbf{v}$, under the same pattern hypotheses as (ii).

Proof. Throughout, the scores split. The difference $s_\mu - s_1$ then has the constant gradient $\boldsymbol{\phi}_\mu - \boldsymbol{\phi}_1$, and we write $s_\mu(\mathbf{x}) = \boldsymbol{\phi}_\mu^\top \mathbf{x} + a_\mu + \sigma(\mathbf{x})$ with offsets a_μ and a function σ shared by all patterns. The shared σ cancels in the softmax.

Part (i). With $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$, Lemma 5 gives

$$J_T = \beta \lambda \, \boldsymbol{\Xi} A \boldsymbol{\Xi}^\top,$$

which is symmetric positive semidefinite for $\lambda \geq 0$ because $A \succeq 0$. Lemma 3 gives a convex h with $\nabla h = T$. Explicitly, for $\lambda > 0$,

$$h(\mathbf{x}) = \frac{1}{\beta \lambda} \operatorname{lse}\big(\beta \lambda \boldsymbol{\Xi}^\top \mathbf{x} + \beta \mathbf{a}\big),$$

where $\mathbf{a}$ collects the offsets, and h is affine for $\lambda = 0$. Adding $\mathbf{v}\mathbf{1}^\top$ adds the same $\mathbf{v}^\top \mathbf{x}$ to every score and does not change the softmax. For $\lambda > 0$ the largest weight goes to the largest reduced score $\lambda \boldsymbol{\xi}_\mu^\top \mathbf{x} + a_\mu$, which is the stated decision rule.

Part (ii), Step 1: pair expansion. Using $A = \operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$ and $\sum_\mu p_\mu = 1$,

$$\boldsymbol{\Xi} A \boldsymbol{\Phi}^\top = \sum_{u<v} p_u p_v \, (\boldsymbol{\xi}_u - \boldsymbol{\xi}_v)(\boldsymbol{\phi}_u - \boldsymbol{\phi}_v)^\top. \tag{2}$$

The "if" direction of (ii) is part (i). The remaining steps prove the "only if" direction.

Step 2: each pair term is symmetric. A homogeneous quadratic that vanishes on the open simplex vanishes on the open orthant, and hence everywhere. The monomials $\{p_u p_v\}_{u<v}$ are therefore linearly independent on the open simplex. The antisymmetric part of equation (2) vanishes for every $\mathbf{p}$ if and only if every pair term is symmetric. The patterns are distinct, and the rank-one term $(\boldsymbol{\xi}_u - \boldsymbol{\xi}_v)(\boldsymbol{\phi}_u - \boldsymbol{\phi}_v)^\top$ is then symmetric exactly when

$$\boldsymbol{\phi}_u - \boldsymbol{\phi}_v = \lambda_{uv} \, (\boldsymbol{\xi}_u - \boldsymbol{\xi}_v) \qquad \text{for some real } \lambda_{uv}.$$

Step 3: one λ for all pairs. For three patterns a, b, c not on one line, the differences $\boldsymbol{\xi}_a - \boldsymbol{\xi}_b$ and $\boldsymbol{\xi}_b - \boldsymbol{\xi}_c$ are linearly independent. Step 2 for the pairs (a, c), (a, b) and (b, c) gives

$$\lambda_{ac}(\boldsymbol{\xi}_a - \boldsymbol{\xi}_b) + \lambda_{ac}(\boldsymbol{\xi}_b - \boldsymbol{\xi}_c) = \boldsymbol{\phi}_a - \boldsymbol{\phi}_c = \lambda_{ab}(\boldsymbol{\xi}_a - \boldsymbol{\xi}_b) + \lambda_{bc}(\boldsymbol{\xi}_b - \boldsymbol{\xi}_c),$$

and comparing coefficients gives $\lambda_{ab} = \lambda_{bc} = \lambda_{ac}$. Two pairs (a, b) and (b, c) with a, b, c on one line are joined through a pattern e off that line, which exists because the patterns are not all collinear. Neither a, b, e nor e, b, c lie on one line, and the argument above gives $\lambda_{ab} = \lambda_{be} = \lambda_{bc}$. Two disjoint pairs are joined through a pair that meets both. Every λ_{uv} therefore equals one λ, and $\boldsymbol{\phi}_u - \lambda \boldsymbol{\xi}_u$ is the same vector $\mathbf{v}$ for every u. That is, $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$. For $M = 2$ the single pair term already has this form.

Step 4: the sign. By Step 3 and equation (2), $\boldsymbol{\Xi} A \boldsymbol{\Phi}^\top = \lambda \, \boldsymbol{\Xi} A \boldsymbol{\Xi}^\top$. For $\mathbf{p}$ in the open simplex $\boldsymbol{\Xi} A \boldsymbol{\Xi}^\top$ is positive semidefinite and non-zero, because the patterns are distinct. Positive semidefiniteness of $\lambda \, \boldsymbol{\Xi} A \boldsymbol{\Xi}^\top$ then gives $\lambda \geq 0$.

Dimension. For $M \geq 2$ and distinct patterns the linear map $(\lambda, \mathbf{v}) \mapsto \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ is injective. Indeed $\lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top = 0$ gives $\lambda(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2) = 0$, hence $\lambda = 0$ and then $\mathbf{v} = 0$. The image of $\{\lambda \geq 0\} \times \mathbb{R}^d$ is therefore a closed half-flat of dimension $d + 1$.

Part (iii). Write $\boldsymbol{\delta}_{uv} := \boldsymbol{\xi}_u - \boldsymbol{\xi}_v$ and $\mathbf{m}_{uv} := \boldsymbol{\phi}_u - \boldsymbol{\phi}_v$. By equation (2), symmetry of $\boldsymbol{\Xi} A(\mathbf{x}) \boldsymbol{\Phi}^\top$ at $\mathbf{x}$ reads

$$\sum_{u<v} p_u p_v \, \Theta_{uv} = 0, \qquad \Theta_{uv} := \boldsymbol{\delta}_{uv} \mathbf{m}_{uv}^\top - \mathbf{m}_{uv} \boldsymbol{\delta}_{uv}^\top,$$

and each Θ_{uv} is constant in $\mathbf{x}$. Multiplying by the squared partition function of the reduced scores $\boldsymbol{\phi}_\mu^\top \mathbf{x} + a_\mu$ turns the condition into

$$\sum_{\mathbf{w}} e^{\beta \mathbf{w}^\top \mathbf{x}} \sum_{u<v \,:\, \boldsymbol{\phi}_u + \boldsymbol{\phi}_v = \mathbf{w}} e^{\beta(a_u + a_v)} \, \Theta_{uv} = 0 \qquad \text{on an open set,}$$

where $\mathbf{w}$ runs over the distinct values of the pair sums. Exponentials with distinct frequencies are linearly independent on every open set. The condition therefore holds if and only if every inner sum vanishes, which is the first claim of (iii). If the pair sums are pairwise distinct, each inner sum has one term. Then every Θ_{uv} vanishes, that is, $\mathbf{m}_{uv} \parallel \boldsymbol{\delta}_{uv}$ for every pair. Steps 2 and 3 of (ii) give $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ with a real λ. Step 4 does not apply, since only symmetry is assumed.

From admissibility to the half-flat. The weights depend on $\mathbf{x}$ only through the reduced scores $\boldsymbol{\phi}_\mu^\top \mathbf{x} + a_\mu$, so T is smooth. If T is $\mathbf{I}$-admissible, Lemma 4 makes $J_T = \beta \boldsymbol{\Xi} A \boldsymbol{\Phi}^\top$ symmetric positive semidefinite almost everywhere, and by continuity at every $\mathbf{x}$. The hypothesis of (iii) then holds on $\mathbb{R}^d$. If the pair sums are pairwise distinct, (iii) gives $\boldsymbol{\Phi} = \lambda \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$, and Step 4 at any single $\mathbf{x}$ gives $\lambda \geq 0$. For generic $\boldsymbol{\Phi}$ and patterns as in (ii), the pattern readout is therefore $\mathbf{I}$-admissible only on the half-flat. □

The non-collinearity hypothesis is needed. Take collinear patterns $\boldsymbol{\xi}_\mu = \boldsymbol{\xi}_0 + t_\mu \mathbf{e}$ and $\boldsymbol{\Phi} = \mathbf{v}\mathbf{1}^\top + \mathbf{e}\mathbf{f}^\top$ with f_μ non-decreasing in t_μ. Each pair term of equation (2) is then

$$p_u p_v \, (t_u - t_v)(f_u - f_v) \, \mathbf{e}\mathbf{e}^\top \succeq 0,$$

so $\boldsymbol{\Xi} A \boldsymbol{\Phi}^\top$ is symmetric positive semidefinite for every $\mathbf{p}$. This $\boldsymbol{\Phi}$ lies off the flat unless f_μ is affine in t_μ.

The same argument holds in every constant metric, with the half-flat through $G\boldsymbol{\Xi}$ in place of the one through $\boldsymbol{\Xi}$.

Corollary 2: In a constant metric G the half-flat passes through $G\boldsymbol{\Xi}$

Let G be symmetric positive definite, let the scores split with per-pattern gradients $\boldsymbol{\Phi}$, and let $T(\mathbf{x}) = \boldsymbol{\Xi} \operatorname{softmax}(\beta \mathbf{s}(\mathbf{x}))$. (i) If $\boldsymbol{\Phi} = \lambda G \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ for some $\lambda \geq 0$ and $\mathbf{v} \in \mathbb{R}^d$, then T is G-admissible. (ii) Under the pattern hypotheses of Theorem 1(ii), $G\boldsymbol{\Xi} A \boldsymbol{\Phi}^\top$ is symmetric positive semidefinite for every $\mathbf{p}$ in the probability simplex if and only if $\boldsymbol{\Phi} = \lambda G \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ with $\lambda \geq 0$. (iii) Under the pair-sum and pattern hypotheses of Theorem 1(iii), if $G\boldsymbol{\Xi} A(\mathbf{x}) \boldsymbol{\Phi}^\top$ is symmetric for every $\mathbf{x}$ in an open set, then $\boldsymbol{\Phi} = \lambda G \boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ for some real λ and some $\mathbf{v}$. If T is G-admissible, then moreover $\lambda \geq 0$.

Proof. By Lemma 5, $G J_T = \beta \, (G\boldsymbol{\Xi}) A \boldsymbol{\Phi}^\top$. The proof of Theorem 1(ii) and (iii) uses the patterns only through the left factor of $\boldsymbol{\Xi} A \boldsymbol{\Phi}^\top$, that is, through the differences $\boldsymbol{\xi}_u - \boldsymbol{\xi}_v$ in equation (2) and through the hypotheses that the patterns are distinct and not all collinear. The weights $\mathbf{p}(\mathbf{x})$ depend on $\boldsymbol{\Phi}$ and the offsets only. An invertible G keeps the vectors $G\boldsymbol{\xi}_\mu$ distinct and not all collinear. Replacing the left factor $\boldsymbol{\Xi}$ by $G\boldsymbol{\Xi}$ in that proof therefore gives (ii) and the first claim of (iii). The second claim of (iii) follows as in the last paragraph of that proof, since $G\boldsymbol{\Xi} A \boldsymbol{\Xi}^\top G$ is positive semidefinite and non-zero for $\mathbf{p}$ in the open simplex. For (i), the weights are $\mathbf{p} = \operatorname{softmax}(\beta\lambda \boldsymbol{\Xi}^\top G \mathbf{x} + \beta \mathbf{a})$, because $\mathbf{v}^\top \mathbf{x}$ and the shared part of the score cancel in the softmax. For $\lambda > 0$ the convex function

$$h(\mathbf{x}) = \frac{1}{\beta\lambda} \operatorname{lse}\big(\beta\lambda \boldsymbol{\Xi}^\top G \mathbf{x} + \beta \mathbf{a}\big)$$

has $\nabla h = G\boldsymbol{\Xi}\mathbf{p} = G\,T$. For $\lambda = 0$ the map is constant and $h(\mathbf{x}) = (G\,T)^\top \mathbf{x}$ is affine. Definition 1 holds in both cases. Theorem 1 is the case $G = \mathbf{I}$. The softmax form of the kernelised network of Wu et al. (2024a) has the score $\boldsymbol{\xi}_\mu^\top W^\top W \mathbf{x}$, so $\boldsymbol{\Phi} = G\boldsymbol{\Xi}$ with $G = W^\top W$, which is positive definite for W of full column rank. It is the case $\lambda = 1$ and $\mathbf{v} = 0$ of (i). □

The next corollary names the members of the flat that Table 1 cites.

Corollary 3: The squared distance and the dot product lie in the flat

Under the pattern readout, the following maps are $\mathbf{I}$-admissible. (a) The negative squared Euclidean distance under every locally Lipschitz separation whose Jacobian is symmetric positive semidefinite almost everywhere in $\mathbb{R}^M$ and annihilates $\mathbf{1}$, softmax among them. The score splits with $\mathbf{\Phi} = 2\mathbf{\Xi}$ and sits in the flat at $\lambda = 2$. (b) The dot product under two families of separations. The first is every Fenchel–Young map $\mathrm{sep} = \nabla\Omega^*$ for convex Ω, which covers softmax, sparsemax and α-entmax (Martins & Astudillo, 2016; Peters et al., 2019), none of which is elementwise. It also covers every locally Lipschitz separation whose Jacobian is symmetric positive semidefinite almost everywhere in $\mathbb{R}^M$. The second is every locally Lipschitz elementwise sep for which $\mathbf{\Xi}\,\mathrm{diag}(\mathrm{sep}'(\beta\mathbf{\Xi}^\top\mathbf{x}))\mathbf{\Xi}^\top$ is positive semidefinite for almost every $\mathbf{x}$, and every elementwise sep that is non-decreasing on the realised range, with no regularity assumed.

Proof. (a) The squared distance. Expanding the score gives

$$-\|\boldsymbol{\xi}_\mu - \mathbf{x}\|^2 = 2\boldsymbol{\xi}_\mu^\top\mathbf{x} - \|\boldsymbol{\xi}_\mu\|^2 - \|\mathbf{x}\|^2.$$

The last term is common to all μ and is annihilated by $A\mathbf{1} = 0$. The score therefore splits with $\mathbf{\Phi} = 2\mathbf{\Xi}$, and Theorem 1(i) applies with $\lambda = 2$ and offsets $-\|\boldsymbol{\xi}_\mu\|^2$. For softmax an explicit potential is

$$h(\mathbf{x}) = \frac{1}{2\beta}\,\mathrm{lse}\big(2\beta\mathbf{\Xi}^\top\mathbf{x} - \beta\mathbf{n}\big), \qquad n_\mu := \|\boldsymbol{\xi}_\mu\|^2.$$

For another separation, $J_{\mathrm{sep}}\mathbf{1} = 0$ almost everywhere makes sep constant along $\mathbf{1}$. The shared term $-\|\mathbf{x}\|^2$ then has no effect, and

$$T(\mathbf{x}) = \mathbf{\Xi}\ \mathrm{sep}\big(\beta(2\mathbf{\Xi}^\top\mathbf{x} - \mathbf{n})\big).$$

Lemma 3 applied in $\mathbb{R}^M$ gives $\mathrm{sep} = \nabla\Omega^*$ for a convex Ω^*. Hence $T = \nabla h$ for the convex potential

$$h(\mathbf{x}) = \frac{1}{2\beta}\Omega^*\big(\beta(2\mathbf{\Xi}^\top\mathbf{x} - \mathbf{n})\big).$$

(b) Fenchel–Young separations. For the dot product $T(\mathbf{x}) = \mathbf{\Xi}\ \mathrm{sep}(\beta\mathbf{\Xi}^\top\mathbf{x})$. For $\mathrm{sep} = \nabla\Omega^*$ the potential

$$h(\mathbf{x}) = \tfrac{1}{\beta}\Omega^*(\beta\mathbf{\Xi}^\top\mathbf{x})$$

is convex, and $\nabla h = \mathbf{\Xi}\nabla\Omega^*(\beta\mathbf{\Xi}^\top\mathbf{x}) = T$. Softmax, sparsemax and α-entmax are this case, and their Jacobians are not diagonal.

(b) Separations with a symmetric positive semidefinite Jacobian. Lemma 3 applied in $\mathbb{R}^M$ gives $\mathrm{sep} = \nabla\Omega^*$ with Ω^* convex, and the same h works.

(b) Elementwise separations. A pattern with $\boldsymbol{\xi}_\mu = 0$ contributes nothing to T, and we drop it. For a locally Lipschitz elementwise sep the Jacobian

$$J_T = \beta\,\mathbf{\Xi}\,\mathrm{diag}(\mathrm{sep}')\mathbf{\Xi}^\top$$

exists at almost every $\mathbf{x}$, since the preimage of a null set under $\mathbf{x} \mapsto \beta\boldsymbol{\xi}_\mu^\top\mathbf{x}$ is null for $\boldsymbol{\xi}_\mu \neq 0$. Lemma 3 then gives a convex potential where J_T is positive semidefinite, and

$$h(\mathbf{x}) = \tfrac{1}{\beta}\sum_\mu F(\beta\boldsymbol{\xi}_\mu^\top\mathbf{x}), \qquad F' = \mathrm{sep},$$

is one explicitly. A non-decreasing sep needs no regularity. Its integral F is convex, so h is convex, and $\mathrm{sep}(t) \in \partial F(t)$ gives $T(\mathbf{x}) \in \partial h(\mathbf{x})$ at every $\mathbf{x}$. □

The hypotheses of Corollary 3. Shift invariance alone is not enough in (a): $\mathrm{sep}(\mathbf{u}) = \mathrm{softmax}(-\mathbf{u})$ is shift-invariant and gives $J_T = -2\beta\,\mathbf{\Xi}A\mathbf{\Xi}^\top \preceq 0$. The hypothesis of (b) is needed on all of $\mathbb{R}^M$, because the realised scores can fill a null set: $d = 1$, $\mathbf{\Xi} = [1, -1]$ and $\mathrm{sep}(\mathbf{u}) = (-|u_1 + u_2| - u_1,\ 0)$ give $T(x) = -\beta x$. Continuity is not enough in place of Lipschitz continuity: $\mathrm{sep} = -C$ for the Cantor function C has $\mathrm{sep}' = 0$ almost everywhere, yet $T(x) = -C(\beta x)$ decreases. A non-decreasing separation is sufficient, not necessary (Appendix D).

Remarks on Theorem 1. Part (i) extends the dot-product case that Santos et al. (2025) establish, because their update is derived for the dot product. Part (ii) refines their remark that the update "supports only the dot product". When $M > d + 1$ the softmax weights of a trained model trace a set of dimension at most d inside the simplex. When $M \le d + 1$ and the differences $\boldsymbol{\phi}_\mu - \boldsymbol{\phi}_1$ are linearly independent, the weights fill the open simplex, and (ii) applies directly. Part (iii) says where the pointwise converse fails. On the set where two pair sums coincide there are admissible $\boldsymbol{\Phi}$ off the flat, and Appendix B.5 gives two families. Changing the separation on a linear score is not an escape from the flat. The Fenchel–Young maps of Santos et al. (2025) sit at its dot-product point, $\lambda = 1$ and $\mathbf{v} = 0$. Appendix D measures the half-flat along a random direction and along the flat itself.

B.3 Proof of Proposition 1

Proposition 1 (restated): Under the gradient readout every semi-convex score gives an admissible map

Call a score family $\{s_\mu\}$ semi-convex if there is a finite $\omega \ge 0$ such that $s_\mu + \omega\|\mathbf{x}\|^2$ is convex for every μ. For every semi-convex family, $h(\mathbf{x}) := \frac{1}{\beta} \operatorname{lse}\big(\beta(\mathbf{s}(\mathbf{x}) + \omega\|\mathbf{x}\|^2\mathbf{1})\big)$ is convex, so every selection $T(\mathbf{x}) \in G^{-1}\partial h(\mathbf{x})$ with $G = (1+2\omega)\mathbf{I}$ is G-admissible on $\mathbb{R}^d$, and at every state $\operatorname{argmax}_\mu\, s_\mu(\mathbf{x})$ is unchanged by the added quadratic.

Proof. Each $s_\mu + \omega\|\mathbf{x}\|^2$ is convex by hypothesis. The map $\mathbf{v} \mapsto \frac{1}{\beta}\operatorname{lse}(\beta\mathbf{v})$ is convex and non-decreasing in each argument, and a non-decreasing convex function of convex functions is convex. Hence h is convex. A convex h makes every selection from $G^{-1}\partial h$ G-admissible by Definition 1, for every constant positive-definite G and in particular for $G = (1 + 2\omega)\mathbf{I}$. A selection is needed in place of ∇h, because h has a kink wherever the score has one. The score of Section 5 has such kinks. The added term $\omega\|\mathbf{x}\|^2$ is common to all μ and does not change the score's argmax. □

Which metric. The scale of G is not free in practice. The identity

$$\tfrac{1}{\beta}\operatorname{lse}\big(\beta(\mathbf{v} + t\mathbf{1})\big) = \tfrac{1}{\beta}\operatorname{lse}(\beta\mathbf{v}) + t$$

gives $h = \frac{1}{\beta}\operatorname{lse}(\beta\mathbf{s}) + \omega\|\mathbf{x}\|^2$. The energy $E = \frac{1}{2}\mathbf{x}^\top G\mathbf{x} - h$ of Definition 1 is then

$$\begin{aligned} G = \mathbf{I}: &\quad E = \big(\tfrac{1}{2} - \omega\big)\|\mathbf{x}\|^2 - \tfrac{1}{\beta}\operatorname{lse}(\beta\mathbf{s}), \\ G = (1+2\omega)\mathbf{I}: &\quad E = \tfrac{1}{2}\|\mathbf{x}\|^2 - \tfrac{1}{\beta}\operatorname{lse}(\beta\mathbf{s}), \\ G = 2\omega\,\mathbf{I}: &\quad E = -\tfrac{1}{\beta}\operatorname{lse}(\beta\mathbf{s}). \end{aligned}$$

The first has a negative quadratic coefficient once $\omega > \frac{1}{2}$, and it is then unbounded below whenever the scores are bounded below. The second uses the metric of the statement. The third metric is positive definite for $\omega > 0$, and it cancels the quadratic. With $\omega = \kappa/\beta \ge \|\mathbf{c}\|_\infty$ it gives Theorem 2, whose h and $G = 2\kappa\mathbf{I}$ are these multiplied by β.

Where the fixed points sit. Every G with $G\,T \in \partial h$ gives an admissible T, and where the fixed points sit depends on G. For the squared distance $\omega = 1$ and $\nabla h = 2\boldsymbol{\Xi}\mathbf{p}$. Then $G = 3\mathbf{I}$ gives $T = \frac{2}{3}\boldsymbol{\Xi}\mathbf{p}$, and $G = 2\mathbf{I}$ gives $T = \boldsymbol{\Xi}\mathbf{p}$, the modern Hopfield update when the patterns have equal norms. The energy guarantees descent. It does not guarantee retrieval, as the repair experiment of Appendix E.7 shows.

Semi-convexity is not the almost-everywhere Hessian condition. Semi-convexity with a linear modulus (Cannarsa & Sinestrari, 2004, Section 1.1) is deliberately stronger than a lower bound on the almost-everywhere Hessian. A score with a first-derivative jump can satisfy the latter wherever it is twice differentiable and still fail to be semi-convex. The sorted score of equation (1) with an *unconstrained* gate is such a score. Its almost-everywhere Hessian is bounded below by $-2\|\mathbf{c}\|_\infty\mathbf{I}$, yet Proposition 6 shows that it has no energy guarantee. With the monotone gate of Section 5 the score is semi-convex and has an energy guarantee.

B.4 Proofs of Lemma 1 and Proposition 2

Lemma 1 (restated): Complex eigenvalues exclude every constant metric

If $J_T(\mathbf{x})$ has a non-real eigenvalue, no constant symmetric positive-definite G makes $GJ_T(\mathbf{x})$ symmetric.

Proof. Let $J = J_T(\mathbf{x})$ and suppose that GJ is symmetric. Then

$$J = G^{-1/2}\big(G^{-1/2}(GJ)\,G^{-1/2}\big)G^{1/2}$$

is similar to the symmetric matrix $G^{-1/2}(GJ)G^{-1/2}$. Its spectrum is therefore real. □

Negative eigenvalues. Symmetry of GJ alone does not exclude a negative eigenvalue. For example, $G = \mathbf{I}$ and $J = \mathrm{diag}(-2, 1, 3)$ give a symmetric GJ with spectrum $\{-2, 1, 3\}$. The other half of Definition 1 excludes one. The potential h is convex, so $\nabla h = G\,T$ is monotone and GJ_T is symmetric positive semidefinite almost everywhere (Lemma 4). The similarity in the proof then makes the spectrum of J_T non-negative almost everywhere. Proposition 2 contains this argument as its case $\mathbf{\Psi}(\mathbf{q}) = \frac{1}{2}\mathbf{q}^\top G\mathbf{q}$, and it also closes the route through a convex post-transform that Section 3.3 describes.

Proposition 2 (restated): Non-real or negative eigenvalues also exclude a convex $\mathbf{\Psi}$

Let h be convex, and let $\mathbf{\Psi}$ be a closed proper convex function whose conjugate $\mathbf{\Psi}^*$ is differentiable on $\nabla h(D_h)$, where D_h is the set on which h is differentiable. Let $E(\mathbf{q}) = \mathbf{\Psi}(\mathbf{q}) - h(\mathbf{q})$. The energy of Santos et al. (2025) is the case $h = \Omega^* \circ \mathbf{\Xi}^\top$ with Ω convex, and Definition 1 is the case $\mathbf{\Psi}(\mathbf{q}) = \frac{1}{2}\mathbf{q}^\top G\mathbf{q}$. Let $T = \nabla\mathbf{\Psi}^*(\nabla h)$ be its concave–convex step. Where h and $\mathbf{\Psi}^*$ are twice differentiable, the Jacobian of T is

$$J_T(\mathbf{x}) = \nabla^2\mathbf{\Psi}^*\big(\nabla h(\mathbf{x})\big)\;\nabla^2 h(\mathbf{x}).$$

It is a product of two symmetric positive semidefinite matrices, so its eigenvalues are real and non-negative. Hence, if J_T exists and has a non-real or negative eigenvalue at every point of an open set U, then no such $\mathbf{\Psi}$ whose conjugate is twice differentiable on $\nabla h(U)$ makes T its concave–convex step. This holds even when $\nabla^2\mathbf{\Psi}^*$ is singular, and it covers every Hopfield–Fenchel–Young energy whose post-transform meets this hypothesis.

Proof. Step 1: the step. The concave–convex step minimises $\mathbf{\Psi}(\mathbf{q}) - \nabla h(\mathbf{x})^\top\mathbf{q}$. Its minimiser satisfies $\nabla h(\mathbf{x}) \in \partial\mathbf{\Psi}(\mathbf{q})$, hence $\mathbf{q} \in \partial\mathbf{\Psi}^*(\nabla h(\mathbf{x}))$ (Rockafellar & Wets, 1998, Prop. 11.3), and $\mathbf{q} = \nabla\mathbf{\Psi}^*(\nabla h(\mathbf{x}))$ where $\mathbf{\Psi}^*$ is differentiable. Santos et al. (2025, Prop. 2 and Eq. 7) derive the same step for a differentiable $\mathbf{\Psi}$. The chain rule gives the product

$$J_T(\mathbf{x}) = PH, \qquad P := \nabla^2\mathbf{\Psi}^*\big(\nabla h(\mathbf{x})\big) \succeq 0, \qquad H := \nabla^2 h(\mathbf{x}) \succeq 0.$$

Step 2: the spectrum. For square matrices X and Y the products XY and YX share a characteristic polynomial χ. With $X = P^{1/2}$ and $Y = P^{1/2}H$ this gives

$$\chi_{PH} = \chi_{P^{1/2}(P^{1/2}H)} = \chi_{(P^{1/2}H)P^{1/2}} = \chi_{P^{1/2}HP^{1/2}}.$$

The matrix $P^{1/2}HP^{1/2}$ is symmetric positive semidefinite. The whole spectrum of PH, with multiplicity, is therefore real and non-negative. No invertibility is used.

Step 3: from points to the open set U. If $\mathbf{\Psi}^*$ is twice differentiable on $\nabla h(U)$, then both functions are twice differentiable at almost every point of U. The reason is that the convex h is twice differentiable almost everywhere by Alexandrov's theorem (Rockafellar & Wets, 1998, Thm. 13.51). At such a point where T is differentiable, J_T is the product, because $T = \nabla\mathbf{\Psi}^* \circ \nabla h$ on D_h, which has full measure. A non-real or negative eigenvalue at every point of U where J_T exists therefore contradicts the product formula. □

Why the dual form. The post-transforms used in practice are indicator functions, of a ball for ℓ_2-normalisation and of a set with empty interior for layer normalisation. The normalised state sits on the boundary of that set, where $\nabla^2\mathbf{\Psi}$ does not exist. The Hessian $\nabla^2\mathbf{\Psi}^*$ of the conjugate is available in closed form, and the dual form of the product therefore makes the certificate computable. Appendix D gives the positive control on the ℓ_2-normalised model.

B.5 THE ESCAPE FAMILIES OF THEOREM 1(III)

Theorem 1(iii) leaves open the non-generic set on which two pair sums coincide. Two families of splitting $\mathbf{\Phi}$ off the flat give a symmetric positive semidefinite $\mathbf{\Xi}A\mathbf{\Phi}^\top$ at every $\mathbf{x}$, and both live on that set. Appendix D measures both.

Equal columns. Take $M = 3$ patterns that do not lie on one line, and let

$$\boldsymbol{\phi}_2 = \boldsymbol{\phi}_3 = \mathbf{m}, \qquad \boldsymbol{\phi}_1 = \mathbf{m} + \tau\mathbf{g}, \qquad \mathbf{g} := r(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2) + (\boldsymbol{\xi}_1 - \boldsymbol{\xi}_3), \qquad r := e^{\beta(a_2 - a_3)} = p_2/p_3.$$

The ratio r is constant in $\mathbf{x}$ because $\boldsymbol{\phi}_2 = \boldsymbol{\phi}_3$, and the pair sums $\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_1 + \boldsymbol{\phi}_3$ coincide. In equation (2) the pair $(2, 3)$ contributes nothing, because $\boldsymbol{\phi}_2 = \boldsymbol{\phi}_3$, and the other two pairs give

$$\mathbf{\Xi}A\mathbf{\Phi}^\top = \tau\, p_1\big(p_2(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2) + p_3(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_3)\big)\mathbf{g}^\top = \tau\, p_1 p_3\, \mathbf{g}\mathbf{g}^\top,$$

which is symmetric positive semidefinite for $\tau > 0$. Its distance from the flat grows linearly in τ.

Parallelogram. Take $M = 4$ with $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3$ not on one line and $\boldsymbol{\xi}_4 = \boldsymbol{\xi}_2 + \boldsymbol{\xi}_3 - \boldsymbol{\xi}_1$, write $\boldsymbol{\delta}_{uv} := \boldsymbol{\xi}_u - \boldsymbol{\xi}_v$, and let

$$\boldsymbol{\phi}_2 = \boldsymbol{\phi}_1 - \alpha\boldsymbol{\delta}_{12}, \qquad \boldsymbol{\phi}_3 = \boldsymbol{\phi}_1 - \gamma\boldsymbol{\delta}_{13}, \qquad \boldsymbol{\phi}_4 = \boldsymbol{\phi}_2 + \boldsymbol{\phi}_3 - \boldsymbol{\phi}_1,$$

with $\alpha, \gamma > 0$, $\alpha \neq \gamma$, and score offsets satisfying $a_1 + a_4 = a_2 + a_3$ (zero offsets will do). All columns of $\mathbf{\Phi}$ are distinct, and $\boldsymbol{\phi}_1 + \boldsymbol{\phi}_4 = \boldsymbol{\phi}_2 + \boldsymbol{\phi}_3$. Hence $p_1 p_4 \equiv p_2 p_3$. Breaking the offset condition alone costs symmetry. With $p_1 p_4 = p_2 p_3$ the cross terms of equation (2) cancel, and

$$J_T = \beta\Big[(p_1 p_2 + p_3 p_4 + 2p_1 p_4)\,\alpha\,\boldsymbol{\delta}_{12}\boldsymbol{\delta}_{12}^\top + (p_1 p_3 + p_2 p_4 + 2p_1 p_4)\,\gamma\,\boldsymbol{\delta}_{13}\boldsymbol{\delta}_{13}^\top\Big],$$

which is symmetric positive semidefinite by construction.

The hypotheses of the parallelogram family. Setting $\alpha = \gamma$ gives $\mathbf{\Phi} = \alpha\mathbf{\Xi} + (\boldsymbol{\phi}_1 - \alpha\boldsymbol{\xi}_1)\mathbf{1}^\top$, which lies in the flat. With $\alpha < 0$ the display stays symmetric, and it loses positive semidefiniteness. Appendix D shows that the parallelogram condition on $\boldsymbol{\xi}_4$ is needed as well, and that a random $\mathbf{\Phi}$ with distinct pair sums is far from symmetric, as Theorem 1(iii) requires.

C PROOFS FOR SECTIONS 4 AND 5

C.1 PROOF OF PROPOSITION 3 (THE SATURATED PATTERN READOUT)

Proposition 3 (restated): A saturated pattern readout jumps to best matches and can cycle

Let $T(\mathbf{x}) = \mathbf{\Xi}\,\mathrm{softmax}(\beta\mathbf{s}(\mathbf{x}))$ with distinct patterns and locally Lipschitz scores.

(i) If a is its own best match, then for every large enough β the map has a fixed point near $\boldsymbol{\xi}_a$ that attracts every orbit starting close enough to $\boldsymbol{\xi}_a$.

(ii) If $a \neq b$ are each other's best matches, then for every large enough β the map has an attracting orbit of period two, with one point near $\boldsymbol{\xi}_a$ and one near $\boldsymbol{\xi}_b$. Then no function E satisfies $E(T\mathbf{x}) < E(\mathbf{x})$ whenever $T\mathbf{x} \neq \mathbf{x}$, and no constant metric G makes T admissible.

(iii) If S is symmetric, the best-match graph has no cycle longer than two.

(iv) If $S_{aa} \geq S_{ab}$ for all a, b, or if S is a Gram matrix $S_{ab} = \boldsymbol{\xi}_a^\top H \boldsymbol{\xi}_b$ with $H \succeq 0$, then no two patterns are each other's best matches.

(v) Suppose the walk that starts at a and follows best matches meets only patterns with a strict best match, and let it end in a cycle of length k. Then there is an $r > 0$ such that, for every large enough β, every orbit that comes within distance r of $\boldsymbol{\xi}_a$ converges to an orbit of period exactly k, with one point near each pattern of that cycle. By (iii), $k \leq 2$ when S is symmetric.

Proof. Step 0: saturation. Let $D = \max_{\mu,\nu} \|\boldsymbol{\xi}_\mu - \boldsymbol{\xi}_\nu\|$. Suppose that on a closed ball B the scores are L-Lipschitz and satisfy $s_b - s_\mu \geq \gamma > 0$ for every $\mu \neq b$. Then on B

$$\|T(\mathbf{x}) - \boldsymbol{\xi}_b\| \overset{(a)}{\leq} \delta_\beta := (M-1)\,D\,e^{-\beta\gamma},$$

$$\mathrm{Lip}_B(T) \overset{(b)}{\leq} \lambda_\beta := 2(M-1)\,D\,\beta L\,e^{-\beta\gamma},$$

where (a) uses $p_\mu \le e^{-\beta\gamma}$ for $\mu \ne b$ in $T(\mathbf{x}) - \boldsymbol{\xi}_b = \sum_{\mu\ne b} p_\mu(\boldsymbol{\xi}_\mu - \boldsymbol{\xi}_b)$. For (b), $\sum_\mu dp_\mu = 0$ gives $dT = \sum_{\mu\ne b} dp_\mu(\boldsymbol{\xi}_\mu - \boldsymbol{\xi}_b)$. The softmax has $\partial p_\mu/\partial u_\nu = p_\mu(\mathbb{1}_{\mu\nu} - p_\nu)$, which gives $|dp_\mu| \le 2p_\mu \max_\nu |du_\nu|$, and $|du_\nu| \le \beta L\|d\mathbf{x}\|$. Both δ_β and λ_β tend to zero as $\beta \to \infty$. Sparsemax and α-entmax with $\alpha > 1$ equal the indicator of the maximiser once $\beta\gamma$ exceeds a constant. For them $\delta_\beta = \lambda_\beta = 0$ from a finite β on.

Part (i): an attracting fixed point. Let $\gamma_a = S_{aa} - \max_{\mu\ne a} S_{a\mu} > 0$. By continuity there is $r > 0$ with $s_a - s_\mu \ge \gamma_a/2$ on $B = \bar{B}(\boldsymbol{\xi}_a, r)$ for every $\mu \ne a$. Once $\delta_\beta \le r$ and $\lambda_\beta < 1$ at $\gamma = \gamma_a/2$, the map sends B into $\bar{B}(\boldsymbol{\xi}_a, \delta_\beta) \subseteq B$ as a contraction. Banach's theorem gives a unique fixed point in B, within δ_β of $\boldsymbol{\xi}_a$, and every orbit that starts in B converges to it.

Part (ii): an orbit of period two. Let γ be the smaller of the two margins $S_{ab} - \max_{\mu\ne b} S_{a\mu}$ and $S_{ba} - \max_{\mu\ne a} S_{b\mu}$. Choose $r < \|\boldsymbol{\xi}_a - \boldsymbol{\xi}_b\|/2$ with $s_b - s_\mu \ge \gamma/2$ on $B_a = \bar{B}(\boldsymbol{\xi}_a, r)$ for $\mu \ne b$, and $s_a - s_\mu \ge \gamma/2$ on $B_b = \bar{B}(\boldsymbol{\xi}_b, r)$ for $\mu \ne a$. Once $\delta_\beta \le r$ and $\lambda_\beta < 1$, T maps B_a into B_b and B_b into B_a. Hence T^2 maps B_a into itself as a contraction. Its fixed point $\mathbf{x}_a$ gives $\mathbf{x}_b := T(\mathbf{x}_a) \in B_b$ with $T(\mathbf{x}_b) = \mathbf{x}_a$. The balls are disjoint, so $\mathbf{x}_a \ne \mathbf{x}_b$ and the orbit has period exactly two. A function E with $E(T\mathbf{x}) < E(\mathbf{x})$ whenever $T\mathbf{x} \ne \mathbf{x}$, applied at $\mathbf{x}_a$ and then at $\mathbf{x}_b$, would give

$$E(\mathbf{x}_a) > E(\mathbf{x}_b) > E(\mathbf{x}_a).$$

Definition 1 provides such an E, with $E(\mathbf{x}) - E(T\mathbf{x}) \ge \frac{1}{2}\|T\mathbf{x} - \mathbf{x}\|_G^2$. Therefore T is G-admissible for no G.

Part (iii): no long cycle for a symmetric S. Suppose $a_1 \to a_2 \to \cdots \to a_k \to a_1$ is a cycle of distinct patterns with $k \ge 3$. Then $a_{i+2} \ne a_i$. Since a_{i+2} is the best match of a_{i+1}, and by symmetry, $S_{a_{i+1}a_{i+2}} > S_{a_{i+1}a_i} = S_{a_ia_{i+1}}$ for every i, with indices taken around the cycle. Going once around the cycle gives the contradiction

$$S_{a_1a_2} < S_{a_2a_3} < \cdots < S_{a_ka_1} < S_{a_1a_2}.$$

Part (iv): no mutual pair. If a and b are each other's best matches, then $S_{ab} > S_{aa}$ and $S_{ba} > S_{bb}$. The first case of (iv) excludes this directly. In the second case

$$(S_{aa}S_{bb})^{1/2} \overset{(a)}{\le} \max(S_{aa}, S_{bb}) \overset{(b)}{<} S_{ab} \overset{(c)}{\le} (S_{aa}S_{bb})^{1/2},$$

where (a) uses $S_{aa}, S_{bb} \ge 0$, (b) is the mutual best match, and (c) is the Cauchy–Schwarz inequality for the semi-inner product H. The chain is a contradiction.

Part (v): the walk of best matches. Let $\mathcal{P}$ be the finite set of patterns on the walk from a, and write $b(e)$ for the best match of $e \in \mathcal{P}$. Let $\gamma > 0$ be the smallest of the margins $S_{e\,b(e)} - \max_{\mu\ne b(e)} S_{e\mu}$ over $e \in \mathcal{P}$. Choose $r < \min_{\mu\ne\nu} \|\boldsymbol{\xi}_\mu - \boldsymbol{\xi}_\nu\|/2$ with $s_{b(e)} - s_\mu \ge \gamma/2$ on $B_e = \bar{B}(\boldsymbol{\xi}_e, r)$ for every $e \in \mathcal{P}$ and every $\mu \ne b(e)$. Once $\delta_\beta \le r$ and $\lambda_\beta < 1$, both taken at the margin $\gamma/2$, Step 0 gives $T(B_e) \subseteq \bar{B}(\boldsymbol{\xi}_{b(e)}, \delta_\beta) \subseteq B_{b(e)}$ for every $e \in \mathcal{P}$. An orbit that enters B_a therefore visits the balls of the walk in order and reaches the ball B_{e_1} of a pattern on the final cycle $e_1 \to \cdots \to e_k \to e_1$. There T^k maps B_{e_1} into itself with Lipschitz constant $\lambda_\beta^k < 1$. The orbit converges to the orbit of the unique fixed point of T^k in B_{e_1}. The balls $B_{e_1}, \ldots, B_{e_k}$ are disjoint, and that fixed point has period exactly k under T. When S is symmetric, (iii) gives $k \le 2$. Parts (i) and (ii) are the cases $k = 1$ and $k = 2$ with a on the cycle. □

Reading Proposition 3. The two-cycle of part (ii) rules out every function that strictly decreases along the iteration and not only the energies of Definition 1. Step 0 uses only that, where one score leads by a fixed margin, the separation and its Jacobian tend to those of the indicator of the maximiser. The proposition therefore holds for sparsemax and α-entmax as well. The proof of part (iii) is the classical two-step energy argument of Goles-Chacc et al. (1985) and Bruck (1990), restricted to one-hot states. Part (iv) applies to most saturating rows of Table 1 for one of two reasons. A score that decreases with a distance has $S_{aa} \ge S_{ab}$, and so does the order-statistic score on the squared-distance base with a gate $\mathbf{c} \ge 0$, since $S_{aa} = 0$ and every sorted coordinate is non-positive. The dot-product rows, sparse and kernelised included, are Gram matrices. A negative gate entry can make $S_{ab} > 0 = S_{aa}$, which the adaptive network's released code allows and Corollary 1 forbids at every $\boldsymbol{\theta}$. Both of the adaptive network's bases give a symmetric S. By (iii) and (v), an orbit that

reaches the saturated neighbourhood of a pattern whose walk has strict best matches converges to a fixed point or to an orbit of period two.

The certificate of Proposition 2 and Proposition 3 work in complementary regimes. Suppose one score leads by a fixed margin and β grows. Then $J_T \to 0$, and every eigenvalue statistic with an absolute threshold passes. A mutual pair still forces a periodic orbit in this regime. The universal-Hopfield identity separation does not saturate, and the proposition says nothing about it. Section 6 tests the proposition on trained models.

C.2 Proof of Theorem 2 and Corollary 1 (the energy of the construction)

On the set where two permutations tie, the proof of Theorem 2 needs a subgradient of a log-sum-exp of convex functions, and the proof of Proposition 4 needs one of a conjugate Ω^*. The next lemma gives both at once.

Lemma 6: Weighted inner subgradients give a subgradient of the composition

Let $F : \mathbb{R}^M \to \mathbb{R}$ be convex, differentiable and non-decreasing in each argument, let $f_1, \dots, f_M : \mathbb{R}^d \to \mathbb{R}$ be convex, and let $h = F(f_1, \dots, f_M)$. Then h is convex. Write $\mathbf{p} = \nabla F\big(f_1(\mathbf{x}), \dots, f_M(\mathbf{x})\big)$. If $\boldsymbol{\eta}_\mu \in \partial f_\mu(\mathbf{x})$ for every μ, then $\sum_\mu p_\mu \boldsymbol{\eta}_\mu \in \partial h(\mathbf{x})$.

Proof. Convexity of h is the composition rule for a convex outer function that is non-decreasing in each argument (Boyd & Vandenberghe, 2004, Section 3.2.4). For every $\mathbf{y} \in \mathbb{R}^d$,

$$h(\mathbf{y}) \overset{(a)}{\geq} F\Big(\big(f_\mu(\mathbf{x}) + \boldsymbol{\eta}_\mu^\top(\mathbf{y}-\mathbf{x})\big)_{\mu=1}^M\Big) \overset{(b)}{\geq} h(\mathbf{x}) + \sum_\mu p_\mu\, \boldsymbol{\eta}_\mu^\top(\mathbf{y}-\mathbf{x}),$$

where (a) applies the subgradient inequality $f_\mu(\mathbf{y}) \geq f_\mu(\mathbf{x}) + \boldsymbol{\eta}_\mu^\top(\mathbf{y}-\mathbf{x})$ inside the non-decreasing F, and (b) is the gradient inequality of the convex F at $(f_\mu(\mathbf{x}))_\mu$. □

For $F = \mathrm{lse}$ the weights $\mathbf{p}$ are the softmax weights, and for $F = \Omega^*$ they are $\nabla\Omega^*$.

Theorem 2 (restated): A monotone gate gives the gradient readout an exact energy

Let $\mathbf{c} \in [0,1]^d$ be non-increasing, and let $\kappa > 0$ with $\kappa \geq \beta\|\mathbf{c}\|_\infty$. Then

$$T(\mathbf{x}) = \mathbf{x} + \frac{\beta}{\kappa} \sum_{\mu=1}^M p_\mu\, \mathbf{g}_\mu^* \odot (\boldsymbol{\xi}_\mu - \mathbf{x})$$

is the concave–convex step of $E(\mathbf{x}) = -\,\mathrm{lse}(\mathbf{u}(\mathbf{x}))$ in the metric $G = 2\kappa\mathbf{I}$, where $\mathbf{g}_\mu^*(\mathbf{x})$ is a permutation of $\mathbf{c}$ that attains $s_\mu(\mathbf{x})$. If $c_d > 0$, E is coercive. Hence, for every $\mathbf{x}$ and every trajectory $\mathbf{x}_{k+1} = T(\mathbf{x}_k)$,

$$E(\mathbf{x}) - E(T\mathbf{x}) \geq \kappa\|T\mathbf{x} - \mathbf{x}\|^2, \quad E(\mathbf{x}) \geq -\log M, \quad \sum_{k\geq 0} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq \frac{E(\mathbf{x}_0) + \log M}{\kappa}.$$

Proof. Step 1: rearrangement. Let $\tilde{\mathbf{q}}_\mu$ be the decreasing rearrangement of $\mathbf{q}_\mu := \mathbf{q}(\boldsymbol{\xi}_\mu, \mathbf{x})$. For a non-increasing $\mathbf{c}$ the rearrangement inequality (Hardy et al., 1952, Theorem 368) gives

$$s_\mu(\mathbf{x}) = \sum_{l=1}^d c_l\, \tilde{q}_\mu(l) = \max_{\sigma \in \mathfrak{S}_d} s_{\mu,\sigma}(\mathbf{x}), \qquad s_{\mu,\sigma}(\mathbf{x}) := -\sum_{i=1}^d c_{\sigma(i)}\, (\xi_{\mu,i} - x_i)^2.$$

Step 2: convexity. Each member $u_{\mu,\sigma} + \kappa\|\mathbf{x}\|^2 = \beta s_{\mu,\sigma} + \kappa\|\mathbf{x}\|^2$ is separable. Its second derivative in the coordinate x_i is

$$2\big(\kappa - \beta c_{\sigma(i)}\big) \;\geq\; 2\big(\kappa - \beta\|\mathbf{c}\|_\infty\big) \;\geq\; 0,$$

where the last inequality is the hypothesis on κ. Each member is therefore convex on $\mathbb{R}^d$. A maximum of convex functions is convex, and $u_\mu + \kappa\|\mathbf{x}\|^2$ is convex for every μ. Lemma 6 with $F = \mathrm{lse}$ then shows that $h := \mathrm{lse}(\mathbf{u} + \kappa\|\mathbf{x}\|^2\mathbf{1})$ is convex on all of $\mathbb{R}^d$.

Step 3: the step. The identity $\mathrm{lse}(\mathbf{v} + t\mathbf{1}) = \mathrm{lse}(\mathbf{v}) + t$ gives

$$E(\mathbf{x}) = -\mathrm{lse}(\mathbf{u}(\mathbf{x})) = \kappa\|\mathbf{x}\|^2 - h(\mathbf{x}),$$

the difference of two convex functions. The concave–convex step of Yuille & Rangarajan (2003), with a subgradient of h in place of its gradient, solves $2\kappa\,\mathbf{y} = \mathbf{v}$ for a subgradient $\mathbf{v} \in \partial h(\mathbf{x})$. Where h is differentiable, Danskin's theorem (Danskin, 1966) gives

$$\nabla_{\mathbf{x}} s_\mu = 2\,\mathbf{g}_\mu^* \odot (\boldsymbol{\xi}_\mu - \mathbf{x}),$$
$$\nabla h = 2\kappa\mathbf{x} + \beta \sum_\mu p_\mu \nabla s_\mu,$$

and $\nabla h/(2\kappa)$ is the displayed T. On the tie set, the member of each maximising permutation is convex, lies below the maximum and touches it at $\mathbf{x}$. Its gradient therefore lies in $\partial(u_\mu + \kappa\|\cdot\|^2)(\mathbf{x})$. Lemma 6 with $F = \mathrm{lse}$ turns these gradients into a subgradient of h, and that subgradient divided by 2κ is the displayed T with the chosen permutations. Hence every choice of maximising permutations gives a selection from $\frac{1}{2\kappa}\partial h$ at every $\mathbf{x}$, which is what Definition 1 asks for.

Step 4: descent with a rate. By Step 3, T is G-admissible with $G = 2\kappa\mathbf{I}$ and the convex h. The descent inequality of Definition 1, proved in Appendix B, then reads

$$E(\mathbf{x}) - E(T\mathbf{x}) \;\geq\; \tfrac{1}{2}\|T\mathbf{x} - \mathbf{x}\|^2_{2\kappa\mathbf{I}} \;=\; \kappa\|T\mathbf{x} - \mathbf{x}\|^2.$$

Summing it along $\mathbf{x}_{k+1} = T\mathbf{x}_k$ and using the lower bound of Step 5 gives, for every n,

$$\kappa \sum_{k=0}^{n-1} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \;\leq\; E(\mathbf{x}_0) - E(\mathbf{x}_n) \;\leq\; E(\mathbf{x}_0) + \log M,$$

and letting $n \to \infty$ gives the bound on the summed squared steps.

Step 5: lower bound and coercivity. Every u_μ is non-positive, since $\mathbf{c} \geq 0$ and $\mathbf{q} \leq 0$. The function lse is non-decreasing in each argument, and therefore $E(\mathbf{x}) = -\mathrm{lse}(\mathbf{u}(\mathbf{x})) \geq -\mathrm{lse}(\mathbf{0}) = -\log M$. Now let $c_d > 0$. Every sorted coordinate is non-positive and $c_l \geq c_d$ for every l. Hence

$$s_\mu(\mathbf{x}) \;\leq\; c_d \sum_{i=1}^{d} q_i(\boldsymbol{\xi}_\mu, \mathbf{x}) \;=\; -c_d\|\boldsymbol{\xi}_\mu - \mathbf{x}\|^2 \;\longrightarrow\; -\infty \quad \text{as } \|\mathbf{x}\| \to \infty,$$

and $E(\mathbf{x}) \geq -\max_\mu u_\mu(\mathbf{x}) - \log M$ tends to $+\infty$. □

Corollary 1 (restated): The energy holds throughout training

Let $c_l = f + (1-f)\sum_{j\geq l} w_j$, with $\mathbf{w} = \mathrm{softmax}(\boldsymbol{\theta})$ and a floor $f \in (0,1)$. For every $\boldsymbol{\theta} \in \mathbb{R}^d$ the gate is non-increasing, with entries in $[f,1]$ and $c_1 = 1$. Setting $\kappa = \beta$ after every gradient step therefore meets the hypotheses of Theorem 2 at every $\boldsymbol{\theta}$ and every learned $\beta > 0$, so the energy exists at initialisation and after every gradient step.

Proof. The tail sum $\sum_{j\geq l} w_j$ of the probability vector $\mathbf{w}$ is non-increasing in l, lies in $[0,1]$ and equals 1 at $l = 1$. Hence $\mathbf{c}$ is non-increasing, with $c_l \in [f,1]$, $c_1 = 1$ and $c_d \geq f > 0$, for every $\boldsymbol{\theta}$. Then $\|\mathbf{c}\|_\infty = c_1 = 1$, and $\kappa = \beta$ meets both $\kappa > 0$ and $\kappa \geq \beta\|\mathbf{c}\|_\infty$ at every $\beta > 0$. □

The floor. The floor gives $c_d \geq f$ for every $\boldsymbol{\theta}$, and with it the coercivity of Theorem 2. At $f = 0$ the gate is still positive at every finite $\boldsymbol{\theta}$, but c_d can underflow to zero in floating point.

C.3 Theorem 3: the Jacobian of the gradient readout

The gradient readout has a simple Jacobian. Under Theorem 2's hypotheses, inside every region of fixed rank order, J_T is a scaled Hessian of h and hence symmetric with a real non-negative spectrum (Theorem 3). The same formula holds inside every region of fixed rank order for every gate, monotone or not. A test of the Jacobian's symmetry or of its real non-negative spectrum therefore cannot see whether the gate is monotone.

Theorem 3: The gradient readout is a monotone operator

Let T be the map of Theorem 2 for a gate $\mathbf{c} \in \mathbb{R}^d$ and a $\kappa > 0$. Inside every sorting cell,

$$J_T = \mathbf{I} - \frac{\beta}{\kappa} \sum_{\mu} p_\mu \operatorname{diag}(\mathbf{g}_\mu^*) + \frac{\beta^2}{2\kappa} (\nabla \mathbf{s})^\top A \, (\nabla \mathbf{s}), \qquad A = \operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top \succeq 0.$$

(i) Under Theorem 2's hypotheses, J_T is symmetric with real non-negative spectrum and $\lambda_{\min}(J_T) \geq 1 - \beta \|\mathbf{c}\|_\infty / \kappa$, with equality in the limit of (iii). (ii) The display and the bound $\lambda_{\min}(J_T) \geq 1 - \beta \|\mathbf{c}\|_\infty / \kappa$ hold for every gate, monotone or not, and every $\kappa > 0$. Symmetry and a real spectrum therefore need no condition on κ or on the gate, and a non-negative spectrum needs only $\kappa \geq \beta \|\mathbf{c}\|_\infty$. Monotonicity of the gate is the separate hypothesis that makes h globally convex, and no test of the almost-everywhere Jacobian's symmetry or of its real non-negative spectrum can see it. (iii) Let the gate be as in Theorem 2. Then T is not unconditionally nonexpansive. If $\kappa \geq \beta \|\mathbf{c}\|_\infty$, then at every point where J_T exists and for any two patterns a and b, $\|J_T\| \geq \frac{\beta^2}{2\kappa} p_a p_b \|\nabla s_a - \nabla s_b\|^2$. So at $\kappa = \beta \|\mathbf{c}\|_\infty$, the supremum of $\|J_T\|$ over a bounded set that contains a tie between the two largest scores inside a sorting cell, with $\nabla s_a \neq \nabla s_b$ there, grows at least linearly in β. Let a stored pattern ν score strictly above every other pattern at $\boldsymbol{\xi}_\nu$, which holds for distinct patterns when $c_d > 0$. Within each sorting cell, J_T extends continuously to $\boldsymbol{\xi}_\nu$, and the extension tends to $\mathbf{I} - (\beta/\kappa) \operatorname{diag}(\mathbf{g}_\nu^*)$ as $\beta \to \infty$ for every fixed value of κ/β. At $\kappa = \beta \|\mathbf{c}\|_\infty$ its norm tends to $1 - c_d / \|\mathbf{c}\|_\infty$, which is below one when $c_d > 0$. Away from this limit $\|J_T\|$ can exceed one.

Proof. Step 1: the Jacobian. Inside a sorting cell the permutations $\mathbf{g}_\mu^*$ are constant, the potential $h = \kappa \|\mathbf{x}\|^2 + \operatorname{lse}(\mathbf{u})$ of Theorem 2 is smooth, and $T = \nabla h / (2\kappa)$. The walls between the cells lie in finitely many hyperplanes, so the display holds almost everywhere. Inside a cell

$$\nabla h = 2\kappa \mathbf{x} + \beta \sum_{\mu} p_\mu \nabla s_\mu, \qquad \frac{\partial \mathbf{p}}{\partial \mathbf{x}} = \beta A \, \nabla \mathbf{s}, \qquad \nabla^2 s_\mu = -2 \operatorname{diag}(\mathbf{g}_\mu^*).$$

Differentiating the first identity with the other two and dividing by 2κ gives the display of the theorem.

Step 2: part (i). J_T is a Hessian divided by 2κ, hence symmetric with real spectrum. The covariance term is positive semidefinite because $A \succeq 0$. Since $\mathbf{p}$ is a probability vector and $0 \leq \mathbf{g}_\mu^* \leq \|\mathbf{c}\|_\infty$ entrywise, the diagonal part satisfies

$$\mathbf{I} - \frac{\beta}{\kappa} \sum_{\mu} p_\mu \operatorname{diag}(\mathbf{g}_\mu^*) \;\succeq\; \Big(1 - \frac{\beta \|\mathbf{c}\|_\infty}{\kappa}\Big) \mathbf{I}.$$

Hence $\lambda_{\min}(J_T) \geq 1 - \beta \|\mathbf{c}\|_\infty / \kappa$, which is non-negative under the hypothesis $\kappa \geq \beta \|\mathbf{c}\|_\infty$. Step 5 gives the equality clause.

Step 3: part (ii). Step 1 used only that $\mathbf{x}$ lies inside one sorting cell. It used neither the order of $\mathbf{c}$ nor the size of κ, and the display holds for every gate and every $\kappa > 0$. The bound of Step 2 holds for every gate with entries of absolute value at most $\|\mathbf{c}\|_\infty$, and it gives a non-negative spectrum once $\kappa \geq \beta \|\mathbf{c}\|_\infty$. An almost-everywhere Jacobian is evaluated inside sorting cells only, and there it is symmetric with a non-negative spectrum for every gate once $\kappa \geq \beta \|\mathbf{c}\|_\infty$. Tests of symmetry and of a real non-negative spectrum therefore cannot detect a gate that is not monotone, although the Jacobian itself depends on the gate. Proposition 6 shows that the order matters on the walls between the cells.

Step 4: part (iii), the lower bound on $\|J_T\|$. Expanding $A = \operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$ with $\sum_\mu p_\mu = 1$ gives the pair expansion, and dropping every pair term but one gives, for any two patterns a and b,

$$(\nabla \mathbf{s})^\top A \, (\nabla \mathbf{s}) = \sum_{\mu < \nu} p_\mu p_\nu \, (\nabla s_\mu - \nabla s_\nu)(\nabla s_\mu - \nabla s_\nu)^\top$$

$$\succeq p_a p_b \, (\nabla s_a - \nabla s_b)(\nabla s_a - \nabla s_b)^\top.$$

The diagonal part is positive semidefinite once $\kappa \geq \beta \|\mathbf{c}\|_\infty$ (Step 2). Evaluating the quadratic form of J_T at the unit vector along $\nabla s_a - \nabla s_b$ gives

$$\|J_T\| \;\geq\; \frac{\beta^2}{2\kappa} \, p_a p_b \, \|\nabla s_a - \nabla s_b\|^2.$$

The tie set and the score gradients do not depend on β. At a point inside a sorting cell where s_a and s_b tie as the two largest scores, $p_a p_b \to \frac{1}{4}$ as β grows. At $\kappa = \beta\|\mathbf{c}\|_\infty$ the prefactor is $\beta^2/(2\kappa) = \beta/(2\|\mathbf{c}\|_\infty)$, and the bound grows linearly in β.

Step 5: part (iii), the limit at a stored pattern. Let $\gamma := s_\nu(\boldsymbol{\xi}_\nu) - \max_{\mu\neq\nu} s_\mu(\boldsymbol{\xi}_\nu)$. When $c_d > 0$ and the patterns are distinct, $\gamma > 0$. Every sorted coordinate is non-positive and $c_l \geq c_d$, and hence for $\mu \neq \nu$

$$s_\mu(\boldsymbol{\xi}_\nu) \;\leq\; -c_d\|\boldsymbol{\xi}_\mu - \boldsymbol{\xi}_\nu\|^2 \;<\; 0 \;=\; s_\nu(\boldsymbol{\xi}_\nu).$$

At $\boldsymbol{\xi}_\nu$ every coordinate of $\boldsymbol{\xi}_\nu - \mathbf{x}$ vanishes, so all the sorting cells of s_ν meet there, and J_T does not exist at $\boldsymbol{\xi}_\nu$ unless $\mathbf{c}$ is constant. Within one sorting cell the display extends continuously to $\boldsymbol{\xi}_\nu$, where $\nabla s_\nu = 0$ and $1 - p_\nu \leq (M-1)e^{-\beta\gamma}$. Gershgorin's theorem gives $\|A\| \leq 2(1-p_\nu)$. Hence, with κ/β fixed,

$$\Big\|\frac{\beta^2}{2\kappa}(\nabla\mathbf{s})^\top A\,\nabla\mathbf{s}\Big\| \;\leq\; \frac{\beta^2}{\kappa}\,(M-1)\,e^{-\beta\gamma}\,\|\nabla\mathbf{s}\|^2 \;\xrightarrow[\beta\to\infty]{}\; 0.$$

The diagonal part tends to $\mathbf{I} - (\beta/\kappa)\,\mathrm{diag}(\mathbf{g}_\nu^*)$. Every sorting cell permutes the same entries of $\mathbf{c}$. In every cell the eigenvalues of the limit are therefore $1 - (\beta/\kappa)c_l$ for $l = 1, \ldots, d$, the largest is $1-(\beta/\kappa)c_d$ and the smallest is $1-(\beta/\kappa)c_1$. At $\kappa = \beta\|\mathbf{c}\|_\infty = \beta c_1$ the eigenvalues lie in $[0, 1]$. The norm is $1-c_d/\|\mathbf{c}\|_\infty$, below one exactly when $c_d > 0$. The smallest eigenvalue is $0 = 1-\beta\|\mathbf{c}\|_\infty/\kappa$, which is the equality clause of (i). □

C.4 Other separations, modulations and bases

Section 5 states that the construction carries over to sparsemax, α-entmax and a full-column-rank linear modulation, that every base with a curvature bound keeps the energy, and that the dot-product base loses the lower bound. Proposition 4, Corollary 4 and Proposition 5 prove the three statements.

Proposition 4: The construction extends to Fenchel–Young separations and linear modulations
Let $\mathrm{sep} = \nabla\Omega^*$ be a Fenchel–Young separation, with Ω convex on the simplex and Ω^* differentiable. Let $W \in \mathbb{R}^{d'\times d}$ have full column rank, write $\mathbf{z} = W\mathbf{x}$ and $\boldsymbol{\zeta}_\mu = W\boldsymbol{\xi}_\mu$, and let $\mathbf{c} \in [0,1]^{d'}$ be non-increasing with $\kappa > 0$ and $\kappa \geq \beta\|\mathbf{c}\|_\infty$. Let $u_\mu(\mathbf{x}) = \beta\sum_r c_r\, q_{(r)}(\boldsymbol{\zeta}_\mu, \mathbf{z})$ be the gated order statistic of the squared distances in feature space, and let $\mathbf{g}_\mu^*$ be the permutation of $\mathbf{c}$ that attains it. Then

$$T(\mathbf{x}) = \mathbf{x} + \frac{\beta}{\kappa}\,W^+ \sum_{\mu=1}^{M} p_\mu\,\mathbf{g}_\mu^* \odot W(\boldsymbol{\xi}_\mu - \mathbf{x}), \qquad \mathbf{p} = \mathrm{sep}(\mathbf{u}(\mathbf{x})),$$

is the concave–convex step of $E(\mathbf{x}) = -\Omega^*(\mathbf{u}(\mathbf{x}))$ in the metric $G = 2\kappa W^\top W$. Hence $E(\mathbf{x}) - E(T\mathbf{x}) \geq \kappa\|W(T\mathbf{x} - \mathbf{x})\|^2$ and $E \geq -\max_{\mathbf{p}}(-\Omega(\mathbf{p}))$. With Ω the Shannon negentropy and $W = \mathbf{I}$ this is Theorem 2. With the Tsallis negentropies it covers sparsemax and α-entmax, where the bound is $-(1-M^{1-\alpha})/(\alpha(\alpha-1))$, and with a general W it covers a kernelised modulation.

Proof. The proof follows that of Theorem 2 in five steps, with Ω^* in place of lse and the feature $\mathbf{z} = W\mathbf{x}$ in place of $\mathbf{x}$.

Step 1: shift invariance. The function Ω^* is convex. The domain of Ω lies in the simplex, so that $\mathbf{p}^\top\mathbf{1} = 1$ on it and, for every real t,

$$\Omega^*(\mathbf{v} + t\mathbf{1}) \;=\; \sup_{\mathbf{p}}\big(\mathbf{p}^\top\mathbf{v} + t\,\mathbf{p}^\top\mathbf{1} - \Omega(\mathbf{p})\big) \;=\; \Omega^*(\mathbf{v}) + t.$$

Differentiating in $\mathbf{v}$ gives $\nabla\Omega^*(\mathbf{v} + t\mathbf{1}) = \nabla\Omega^*(\mathbf{v})$. The gradient of Ω^* is a probability vector, and Ω^* is therefore non-decreasing in each argument.

Step 2: a convex potential. Steps 1 and 2 of the proof of Theorem 2, applied in $\mathbb{R}^{d'}$ to the variable $\mathbf{z}$, show that each $u_\mu + \kappa\|\mathbf{z}\|^2$ is convex in $\mathbf{z}$. It is therefore convex in $\mathbf{x}$, because $\mathbf{z} = W\mathbf{x}$ is linear. Step 1 gives

$$h(\mathbf{x}) := \Omega^*\big(\mathbf{u}(\mathbf{x}) + \kappa\|W\mathbf{x}\|^2\mathbf{1}\big) = \Omega^*\big(\mathbf{u}(\mathbf{x})\big) + \kappa\|W\mathbf{x}\|^2,$$

and Lemma 6 with $F = \Omega^*$ shows that h is convex. The metric $G = 2\kappa W^\top W$ is positive definite because W has full column rank, and

$$E(\mathbf{x}) = -\Omega^*\big(\mathbf{u}(\mathbf{x})\big) = \tfrac{1}{2}\|\mathbf{x}\|_G^2 - h(\mathbf{x}).$$

Step 3: a subgradient. As in Step 3 of the proof of Theorem 2, each maximising permutation supplies

$$\boldsymbol{\eta}_\mu = W^\top\big(2\kappa\mathbf{z} + 2\beta\,\mathbf{g}_\mu^* \odot (\boldsymbol{\zeta}_\mu - \mathbf{z})\big) \in \partial\big(u_\mu + \kappa\|W\cdot\|^2\big)(\mathbf{x}).$$

By Step 1 the weights of Lemma 6 are $\nabla\Omega^*(\mathbf{u} + \kappa\|W\mathbf{x}\|^2\mathbf{1}) = \operatorname{sep}(\mathbf{u}) = \mathbf{p}$. The lemma with $F = \Omega^*$ gives

$$\mathbf{v}(\mathbf{x}) := \sum_\mu p_\mu \boldsymbol{\eta}_\mu = W^\top\Big(2\kappa\mathbf{z} + 2\beta\sum_\mu p_\mu\,\mathbf{g}_\mu^* \odot (\boldsymbol{\zeta}_\mu - \mathbf{z})\Big) \in \partial h(\mathbf{x}).$$

Step 4: the step and the descent inequality. The concave–convex step solves $G\,T(\mathbf{x}) = \mathbf{v}(\mathbf{x})$, that is

$$2\kappa W^\top W\,T(\mathbf{x}) = W^\top\Big(2\kappa W\mathbf{x} + 2\beta\sum_\mu p_\mu\,\mathbf{g}_\mu^* \odot W(\boldsymbol{\xi}_\mu - \mathbf{x})\Big).$$

Multiplying by $(2\kappa W^\top W)^{-1}$ and writing $W^+ = (W^\top W)^{-1}W^\top$, which has $W^+W = \mathbf{I}$, gives the displayed map. The descent inequality of Definition 1 with this G reads

$$E(\mathbf{x}) - E(T\mathbf{x}) \;\geq\; \tfrac{1}{2}\|T\mathbf{x} - \mathbf{x}\|_G^2 \;=\; \kappa\|W(T\mathbf{x} - \mathbf{x})\|^2.$$

Step 5: the lower bound. Every u_μ is non-positive because $\mathbf{c} \geq 0$ and each squared-distance similarity is non-positive. Hence

$$\Omega^*(\mathbf{u}) \;=\; \sup_{\mathbf{p}}\big(\mathbf{p}^\top\mathbf{u} - \Omega(\mathbf{p})\big) \;\leq\; \max_\mu u_\mu + \max_{\mathbf{p}}\big(-\Omega(\mathbf{p})\big) \;\leq\; \max_{\mathbf{p}}\big(-\Omega(\mathbf{p})\big),$$

and $E = -\Omega^*(\mathbf{u}) \geq -\max_{\mathbf{p}}(-\Omega(\mathbf{p}))$. □

Reading Proposition 4. The construction therefore carries over to the sparsemax and entmax members of the sparse row of Table 1, and to kernelised retrieval with a full-column-rank W. Two limits are unchanged. A base with no curvature bound is not covered (Section 5), and neither is the dot-product base under $E = -\operatorname{lse}(\mathbf{u})$ (Proposition 5). Appendix D measures the construction under three separations and on the multiple-instance datasets.

Corollary 4: Any base with a curvature bound gives an energy

Let each $q(\xi,\cdot)$ be differentiable, non-positive and ρ-weakly convex, meaning that $q(\xi,t) + \rho t^2/2$ is convex in t. With $\mathbf{c}$ as in Theorem 2, $\kappa > 0$ and $\kappa \geq \beta\rho\|\mathbf{c}\|_\infty/2$, the map

$$T(\mathbf{x}) = \mathbf{x} + \frac{\beta}{2\kappa}\sum_{\mu=1}^{M} p_\mu\,\mathbf{g}_\mu^* \odot \partial_{\mathbf{x}}\mathbf{q}(\boldsymbol{\xi}_\mu, \mathbf{x})$$

is the concave–convex step of $E = -\operatorname{lse}(\mathbf{u})$ in the metric $2\kappa\mathbf{I}$, with the descent inequality and the lower bound of Theorem 2. E is coercive if $c_d > 0$ and $q(\xi,t) \to -\infty$ as $|t| \to \infty$.

Proof. The proof of Theorem 2 uses the base in two places. The rearrangement step holds for any real vector $\mathbf{q}_\mu$. In the convexity step each member of the maximum is separable, with coordinate terms

$$\beta c_{\sigma(i)}\,q_i + \kappa x_i^2 \;=\; \beta c_{\sigma(i)}\Big(q_i + \frac{\rho}{2}x_i^2\Big) + \Big(\kappa - \frac{\beta\rho}{2}c_{\sigma(i)}\Big)x_i^2.$$

The first term is convex in x_i because $c_{\sigma(i)} \geq 0$ and q is ρ-weakly convex. The second is convex because $\kappa \geq \beta\rho\|\mathbf{c}\|_\infty/2$. Danskin's theorem gives $\nabla s_\mu = \mathbf{g}_\mu^* \odot \partial_{\mathbf{x}}\mathbf{q}(\boldsymbol{\xi}_\mu, \mathbf{x})$, and the step, the subgradient selection and the descent inequality follow as in Steps 3 and 4 of that proof. The lower bound uses $\mathbf{c} \geq 0$ and $q \leq 0$ only. For coercivity, $c_l \geq c_d$ and $\tilde{q}_\mu \leq 0$ give $s_\mu(\mathbf{x}) \leq c_d\sum_i q(\xi_{\mu,i}, x_i)$, which tends to $-\infty$ as $\|\mathbf{x}\| \to \infty$. □

Examples of Corollary 4. The squared distance has $\rho = 2$ and gives Theorem 2. The smoothed Manhattan base $q = \varepsilon - ((\xi - x)^2 + \varepsilon^2)^{1/2}$ has $\rho = 1/\varepsilon$. The Huber base $q = -H_\delta(\xi - x)$, with $H_\delta(e) = e^2/2$ for $|e| \le \delta$ and $\delta|e| - \delta^2/2$ otherwise, has $\rho = 1$ for every δ. The price of a small ε is a shorter step, since κ grows as $1/\varepsilon$. A smoothed Manhattan base defines a different model from the Manhattan model of Table 1, and it does not repair that model. The Manhattan base $q_i = -|\xi_i - x_i|$ has no finite ρ, and the corollary does not cover it. The dot-product base has $\rho = 0$ and fails only the sign condition $q \le 0$. Under $E = -\operatorname{lse}(\mathbf{u})$ it has no lower bound (Proposition 5).

Proposition 5: The dot-product base is unbounded below under $E = -\operatorname{lse}(\mathbf{u})$

Theorem 2's convexity argument also goes through for the dot-product base $q_i = \xi_i x_i$. But under $E = -\operatorname{lse}(\mathbf{u})$ the dot-product base is unbounded below, and so has no global minimum, for every non-zero gate that meets the hypotheses of Theorem 2, the uniform gate included, as soon as one pattern is non-zero. The squared-distance base is non-positive, so its energy is bounded below by $-\log M$, and its gradient $\nabla q(\boldsymbol{\xi}_i, x_i) = 2(\xi_i - x_i)$ vanishes at $\mathbf{x} = \boldsymbol{\xi}$ for every gate.

Proof. Convexity. The dot-product base has zero curvature, so $u_\mu + \kappa\|\mathbf{x}\|^2$ is convex for every $\kappa > 0$ by Step 2 of the proof of Theorem 2.

Unboundedness. For $t > 0$ the order of the coordinates $\xi_{\mu,i}(tx_i)$ does not depend on t, and therefore

$$s_\mu(t\mathbf{x}) = t\, s_\mu(\mathbf{x}).$$

Along a direction $\mathbf{x}$ with $s_\mu(\mathbf{x}) > 0$ for some μ,

$$E(t\mathbf{x}) = -\operatorname{lse}\big(\mathbf{u}(t\mathbf{x})\big) \;\le\; -\beta t\, s_\mu(\mathbf{x}) \;\longrightarrow\; -\infty \quad \text{as } t \to \infty.$$

Such a direction exists. Take a non-zero pattern $\boldsymbol{\xi}_\mu$ and $\mathbf{x} = \boldsymbol{\xi}_\mu$. Every $q_i = \xi_{\mu,i}^2$ is non-negative and the largest is positive. A non-zero gate that meets the hypotheses of Theorem 2 has $\mathbf{c} \ge 0$ and $c_1 > 0$, and hence $s_\mu(\boldsymbol{\xi}_\mu) \ge c_1 \max_i \xi_{\mu,i}^2 > 0$.

The hypothesis on the gate is needed. A signed gate can keep every score non-positive. In $d = 2$, $\mathbf{c} = (-1, 1)$ gives $s_\mu(\mathbf{x}) = -|\xi_{\mu,1}x_1 - \xi_{\mu,2}x_2| \le 0$, and then $E \ge -\log M$. □

C.5 The order of the gate cannot be dropped

Section 5 states that the first hypothesis of Theorem 2, a non-increasing gate, cannot be dropped. The next proposition makes this exact. For a function f and a direction $\mathbf{v}$, write $f'(\mathbf{x}; \mathbf{v})$ for the one-sided directional derivative, and call $f'(\mathbf{x}; \mathbf{v}) + f'(\mathbf{x}; -\mathbf{v})$ the jump of f at $\mathbf{x}$ along $\mathbf{v}$. A convex function has a non-negative jump at every point and along every direction.

Proposition 6: A gate that is not non-increasing breaks admissibility in general position

Let $\mathbf{c}$ be a gate that is not non-increasing, and let T be the map of Theorem 2 with this gate. (i) The rearrangement identity fails, and there are patterns and queries at which $u_\mu + \kappa\|\mathbf{x}\|^2$ has a first-derivative jump of negative sign at a rank swap. The jump does not depend on κ. (ii) Suppose that $M = 1$, or that the patterns are in general position, by which we mean that such a rank swap of one pattern occurs at a query where no other score has a crease. Then no κ makes h convex, and no constant symmetric positive-definite G makes T G-admissible.

Proof. Step 1: a crease with a negative jump. Let $c_l < c_{l+1}$ for some l. Consider a query at which the l-th and $(l+1)$-th largest per-coordinate similarities of one pattern μ cross as one coordinate x_i moves. Such a query exists because $\mathbf{x}$ is free. Choose $\boldsymbol{\xi}_\mu - \mathbf{x}$ with $|\xi_{\mu,i} - x_i| = |\xi_{\mu,j} - x_j|$ and exactly $l - 1$ coordinates of smaller absolute value. On either side of the crease $q_i = q_j$ the score s_μ is smooth, and the ranks l and $l + 1$ swap across it. Write $\mathbf{e} = \boldsymbol{\xi}_\mu - \mathbf{x}$. The jump of u_μ along a direction $\mathbf{v}$ is

$$u'_\mu(\mathbf{x}; \mathbf{v}) + u'_\mu(\mathbf{x}; -\mathbf{v}) \;=\; 2\beta\,(c_l - c_{l+1})\,|e_i v_i - e_j v_j|,$$

and its sign is that of $c_l - c_{l+1}$ and nothing else. With $c_l < c_{l+1}$ it is negative whenever $e_i v_i \ne e_j v_j$. It does not depend on κ, because $\kappa\|\mathbf{x}\|^2$ is smooth. For a non-increasing gate the same identity gives a non-negative jump. The rearrangement identity fails as well. For $\kappa \ge \beta\|\mathbf{c}\|_\infty$ it would make $u_\mu + \kappa\|\mathbf{x}\|^2$ a maximum of convex functions, whose jumps are non-negative.

Step 2: no κ makes h convex. The separation weights are continuous across the crease. At a query where only the score of pattern μ is creased, the other scores contribute nothing to the jump, and

$$h'(\mathbf{x};\mathbf{v}) + h'(\mathbf{x};-\mathbf{v}) \;=\; p_\mu\big(u'_\mu(\mathbf{x};\mathbf{v}) + u'_\mu(\mathbf{x};-\mathbf{v})\big) \;<\; 0$$

for every $\mathbf{v}$ with $e_i v_i \neq e_j v_j$, since $p_\mu > 0$. Such queries exist when $M = 1$ or when the patterns are in general position. A convex function has non-negative jumps. Hence h is not convex on any neighbourhood of the crease, whatever the value of κ.

Step 3: no constant metric. Step 2 shows more than the non-convexity of this particular h. Write the jump of ∇h across the crease, from the side where $q_j > q_i$ to the side where $q_i > q_j$, as $\lambda\mathbf{n}$, with

$$\lambda = p_\mu\,\beta\,(c_l - c_{l+1}) < 0, \qquad \mathbf{n} = \nabla(q_i - q_j),$$

where $\mathbf{n}$ is normal to the crease. Let $\mathbf{x}_0$ be the query of Step 2. For a constant symmetric positive-definite G take $\mathbf{x}_\pm = \mathbf{x}_0 \pm \varepsilon G^{-1}\mathbf{n}$, which straddle the crease since $\mathbf{n}^\top G^{-1}\mathbf{n} > 0$. The map $T = \nabla h/(2\kappa)$ is Lipschitz on each side near $\mathbf{x}_0$, and $T(\mathbf{x}_+) - T(\mathbf{x}_-) = \lambda\mathbf{n}/(2\kappa) + O(\varepsilon)$. Hence

$$\big(GT(\mathbf{x}_+) - GT(\mathbf{x}_-)\big)^\top(\mathbf{x}_+ - \mathbf{x}_-) \;=\; \frac{\lambda\varepsilon}{\kappa}\,\|\mathbf{n}\|^2 + O(\varepsilon^2),$$

which is negative for small ε. If T were G-admissible, GT would be a selection from the subdifferential of a convex function, and such a subdifferential is a monotone operator (Rockafellar & Wets, 1998, Thm. 12.17). No constant metric and no convex potential therefore make T admissible, which is the sense of Definition 1 that Table 1 uses. □

The same argument settles the second place where the construction stops, the downward kink of the Manhattan base.

Proposition 7: The Manhattan base breaks admissibility at every κ

Let $\mathbf{c}$ be a gate with $c_1 > 0$, let $\kappa > 0$, and let T be the map displayed in Corollary 4 for the Manhattan base $q_i = -|\xi_i - x_i|$, with any choice at the kinks. Then no constant symmetric positive-definite G and no convex function h' satisfy $G\,T(\mathbf{x}) \in \partial h'(\mathbf{x})$ at every $\mathbf{x}$.

Proof. Here $\partial_{x_i} q_i(\boldsymbol{\xi}_\nu, \mathbf{x}) = \operatorname{sign}(\xi_{\nu,i} - x_i)$. Fix a pattern μ and a coordinate i. The query $\mathbf{x}_0$ is taken on the hyperplane $x_i = \xi_{\mu,i}$, outside finitely many subsets of lower dimension, so that for every pattern ν the d similarities $q_j(\boldsymbol{\xi}_\nu, \mathbf{x}_0)$ are pairwise distinct and only the one with $j = i$ can vanish. Near $\mathbf{x}_0$ the permutations $\mathbf{g}^*_\nu$ are then constant on each side of the hyperplane, and T is Lipschitz on each side. The only terms of T that change across the hyperplane come from the patterns ν with $\xi_{\nu,i} = \xi_{\mu,i}$, which include μ. For each of them $q_i = 0$ is the largest similarity near $\mathbf{x}_0$, so its gate entry is c_1. Crossing the hyperplane in the direction $\mathbf{e}_i$ turns $\partial_{x_i} q_i$ from $+1$ into -1. The weights are continuous, so the jump of T across the hyperplane at $\mathbf{x}_0$ is $\lambda\mathbf{e}_i$ with

$$\lambda = -\frac{\beta c_1}{\kappa} \sum_{\nu\,:\,\xi_{\nu,i} = \xi_{\mu,i}} p_\nu(\mathbf{x}_0) < 0.$$

The jump does not depend on the metric, and it is negative at every κ. Take $\mathbf{x}_\pm = \mathbf{x}_0 \pm \varepsilon G^{-1}\mathbf{e}_i$, which straddle the hyperplane since $\mathbf{e}_i^\top G^{-1}\mathbf{e}_i > 0$. Then $T(\mathbf{x}_+) - T(\mathbf{x}_-) = \lambda\mathbf{e}_i + O(\varepsilon)$ and

$$\big(GT(\mathbf{x}_+) - GT(\mathbf{x}_-)\big)^\top(\mathbf{x}_+ - \mathbf{x}_-) \;=\; 2\lambda\varepsilon + O(\varepsilon^2),$$

which is negative for small ε. A selection from the subdifferential of a convex function is a monotone operator (Rockafellar & Wets, 1998, Thm. 12.17), so no such G and h' exist. □

D Additional numerical experiments

This section collects the measurements that accompany the formal results of Appendices B and C, one paragraph per result, and the probes behind Section 5. No number here is used in a proof.

The explicit potential of Corollary 3(a). For the negative squared distance under softmax, the proof of Corollary 3 gives the potential $h(\mathbf{x}) = \frac{1}{2\beta}\operatorname{lse}(2\beta\boldsymbol{\Xi}^\top\mathbf{x} - \beta\mathbf{n})$ with $n_\mu = \|\boldsymbol{\xi}_\mu\|^2$. Its gradient matches T to $3.6 \cdot 10^{-16}$.

A separation with a negative slope (Corollary 3(b)). A non-decreasing elementwise separation is sufficient for admissibility and not necessary. At $\beta = 2$, a separation with $\mathrm{sep}' = -0.7$ on part of its range and $\mathrm{sep}' = 1$ elsewhere is admissible on the antipodal pattern set $[\mathbf{e}_1, \mathbf{e}_2, -\mathbf{e}_1, -\mathbf{e}_2]$, where each negative slope is paired with a larger positive one. The smallest eigenvalue of its Jacobian over 4000 queries is $+0.6$. A convex hull that contains the origin is not enough. Random pattern sets with this property fail in 200 of 200 draws. On the antipodal set itself, $\mathrm{sep}' \equiv -1$ gives the smallest eigenvalue -4.0, the value -2β of $J_T = -\beta\boldsymbol{\Xi}\boldsymbol{\Xi}^\top = -2\beta\mathbf{I}$.

A positive control for the certificate (Proposition 2). The ℓ_2-normalised model of Santos et al. (2025) passes the certificate at every probe. Over 40 probes the largest imaginary part of an eigenvalue of J_T is $1.4 \cdot 10^{-15}$ and the smallest real part is $-9.6 \cdot 10^{-16}$, from a complex-step derivative of the map. That verdict is correct, because the model has a proven energy. A numerical search finds no constant G for this model.

The equal-columns family (Appendix B.5). Relative to $\|\boldsymbol{\Phi}\|$, the distance of the family from the flat is 0.02 to 0.81 over 40 configurations. Over the same configurations the asymmetry is at most $1.2 \cdot 10^{-13}$ and the smallest eigenvalue is at least $-1.2 \cdot 10^{-15}$. This family agrees with the dot product's decision in 30% to 81% of queries.

The parallelogram family (Appendix B.5). Measured over $d \in \{2, 3, 5\}$ and five seeds, the asymmetry is at most $4.4 \cdot 10^{-15}$, the smallest eigenvalue is at least $-6.6 \cdot 10^{-16}$, and the distance to the flat is at least 0.01. Parallelograms are abundant on the Hamming cube, and 15 of 15 bipolar configurations pass. At zero offsets this family agrees with the dot product's decision on every query. Two controls fail as they should. Perturbing $\boldsymbol{\xi}_4$ by 10^{-2} breaks symmetry, and a random $\boldsymbol{\Phi}$ with distinct pair sums has asymmetry of order one.

The half-flat, measured (Theorem 1). Figure 3 measures the admissible set of Theorem 1(ii). In a random direction, symmetry is lost at every distance from the flat. Along the flat, the spectrum changes sign exactly at $\lambda = 0$.

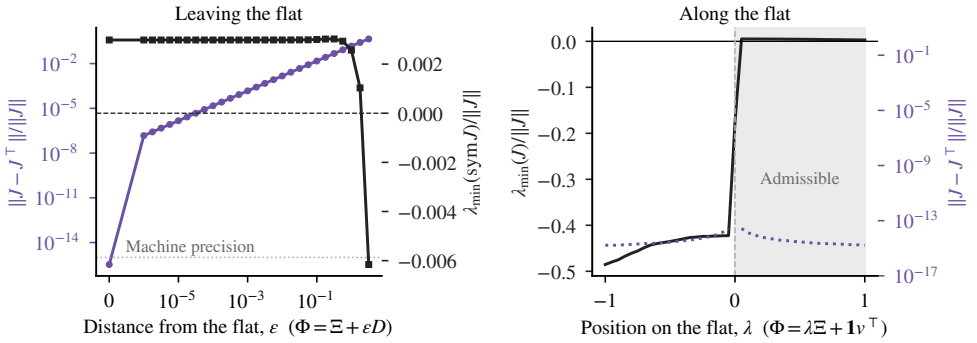

Figure 3: Theorem 1, measured. The Jacobian $J = \beta\,\boldsymbol{\Xi}A\boldsymbol{\Phi}^\top$ of the pattern readout for a splitting score with per-pattern gradients $\boldsymbol{\Phi}$, at $d = 32$, $M = 48$, $\beta = 4$, medians over 60 masked queries. Left: $\boldsymbol{\Phi} = \boldsymbol{\Xi} + \varepsilon D$ for a fixed random direction D of unit Frobenius norm. The relative asymmetry $\|J - J^\top\|/\|J\|$ (purple, left axis) rises linearly from machine precision at $\varepsilon = 0$, and the smallest eigenvalue of the symmetric part (black, right axis) stays positive until ε is of order one. Symmetry, not positivity, is the binding constraint. Right: $\boldsymbol{\Phi} = \lambda\boldsymbol{\Xi} + \mathbf{v}\mathbf{1}^\top$ for a fixed random $\mathbf{v}$. The asymmetry (purple, dotted) stays at machine precision for every $\lambda \neq 0$, because $A\mathbf{1} = 0$ removes the shared term. At $\lambda = 0$ the map is constant. The smallest eigenvalue of J (black) changes sign at $\lambda = 0$. The admissible set is the shaded half.

Theorem 3: an expanding fixed point without saturation. The limit in part (iii) of Theorem 3 assumes that one pattern wins by a margin. Without the margin T can expand at an exact fixed point. With the linear gate from 1 to 0.05 at $\kappa = \beta\|\mathbf{c}\|_\infty$, $\beta = 30$, $d = 24$ and $M = 65$ patterns placed as $\boldsymbol{\xi}_0$ plus 32 symmetric pairs along nearly one direction, $T(\boldsymbol{\xi}_0) = \boldsymbol{\xi}_0$ to $7 \cdot 10^{-18}$ with $p_0 = 0.31$, and every one-sided Jacobian at $\boldsymbol{\xi}_0$ has norm at least 2.72.

Proposition 4: the construction under other separations. The first check compares the two readouts at $\beta = 0.5$ and $W = \mathbf{I}$, under softmax, 1.5-entmax and sparsemax. Under each of the three separations the pattern readout with a signed gate has a Jacobian asymmetry of 1.4, against $2 \cdot 10^{-10}$ for the gradient readout. The pattern readout raises the energy $E = -\Omega^*(\mathbf{u})$ on 7 to 19% of steps. The monotone gate with the gradient readout raises it on none of 18 000 steps across three separations, three modulations and two temperatures. The second check uses the five multiple-instance datasets of Section 6. There the signed gate under the pattern readout cycles in 29 to 72 per cent of held-out bags under sparsemax, 1.5-entmax and a fixed kernel. The monotone gate with the gradient readout converges in at least 99.5 per cent of held-out bags under each (Table 26, Appendix H.6.2).

Proposition 5: the dot-product base diverges. We iterate the gradient readout on the dot-product base from a stored pattern, with the gate of Corollary 1 at $\boldsymbol{\theta} = 0$ and $f = 0.05$, $\kappa = \beta = 30$, $d = 24$ and $M = 40$ unit-norm patterns. In 30 steps $\|\mathbf{x}\|$ grows from 1.0 to 14.0 and E falls from -25.6 to -367. The gradient norm of E at a stored pattern grows with β for the uniform gate ($2.0 \cdot 10^2$ at $\beta = 200$), while it vanishes exponentially in β for the squared-distance base.

Proposition 6: every constant metric fails at a reversed crease. We evaluated the inner product of Step 3 of the proof of Proposition 6 under 20 random positive-definite G. The inner product is negative for all 20 of them, and it is positive for all 20 with a monotone gate at the same crease.

Section 5: the order of the gate. An unconstrained gate loses the energy guarantee, while sorting alone costs nothing. The loss has two parts, the certificate at every β and the behaviour at low β. The probe has $d = 16$ and $M = 12$ unit-norm patterns, the gate $\mathbf{c} = (1, \dots, 0.05)$ or its reverse, and five seeds of six starts, each a stored pattern plus Gaussian noise of standard deviation 0.45 per coordinate, run for 60 steps. Under the iterated map a reversed gate raises the energy on 842 of 1800 steps at $\beta = 1$ and on 729 at $\beta = 5$, against none for a monotone gate. At $\beta = 30$ the softmax saturates and the rises nearly vanish, at 24 of 1800.

Section 5: smoothed bases. Corollary 4 covers the smoothed Manhattan and the Huber bases. The probe has $d = 8$, $M = 6$, a random monotone gate per seed, $\beta = 1$ and 30, and five seeds of eight starts run for 400 steps. With a monotone gate both bases keep the exact energy, and neither violates the descent inequality in any of 16 000 steps per base and temperature.

Section 5: the Manhattan base. The Manhattan base $q_i = -|\xi_i - x_i|$ has no finite ρ. Its derivative jump $-2\beta c_1 p_\mu$ does not depend on κ. The first probe uses the uniform gate $\mathbf{c} = \mathbf{1}$ and steps at $\kappa = \beta = 1$ with $d = 8$ and $M = 6$, five seeds of eight starts run for 60 steps. Under the gradient readout the energy of the Manhattan model rises on 1087 of 2400 steps. The second probe uses $\kappa = 10$, and monotonicity fails on 3000 of 3000 pairs straddling a kink under a random positive-definite metric. This agrees with Proposition 7, which shows that the gradient readout does not repair the Manhattan model at any κ and in any constant metric. The repair of Appendix E.7 sets κ from the largest secant curvature it measures. Its steps are about 25 times shorter, and fewer of them raise the energy. Away from its kinks, every Jacobian statistic in this paper passes the Manhattan map.

What a Jacobian cannot see. Symmetry and the spectrum certify that T is a gradient. They do not certify that it is the gradient of the intended energy. Symmetry and a real spectrum are blind to a discontinuous T and to a step that is too long. The sign of the spectrum sees the long step. At half the required κ the bound $1 - \beta\|\mathbf{c}\|_\infty/\kappa$ of Theorem 3(ii) is -1, and the limit of its part (iii) attains it. Table 5 shows the two tests that have power without a Jacobian: the quantitative descent inequality of Definition 1 and monotonicity of GT over pairs of points (Rockafellar & Wets, 1998, Thm. 12.17).

Table 5: Which test detects which defect. Entries are the number of firings out of the transitions or pairs tested. The discontinuous readout is $T(\mathbf{x}) = \mathrm{diag}(\mathbf{m}(\mathbf{x}))\,\boldsymbol{\Xi}\mathbf{p}$ with scores $s_\mu = \mathbf{m}(\mathbf{x})^\top \mathbf{q}(\boldsymbol{\xi}_\mu, \mathbf{x})$, where $\mathbf{m}(\mathbf{x}) \in \{0,1\}^d$ marks the k coordinates with the largest summed similarity $\sum_\mu q_i(\boldsymbol{\xi}_\mu, \mathbf{x})$ and is shared by all patterns. It jumps at every wall where that set of coordinates changes.

Defect	Jacobian statistics	"did E rise"	Descent inequality	Monotonicity of GT
κ at half the required value	symmetric and real, $\lambda_{\min} < 0$ at 1800 of 1800 (minimum -1.00)	silent	9 of 1800 transitions	—
Discontinuous readout	pass	—	—	405–471 of 4000 pairs

E THE PUBLISHED MAPS, ROW BY ROW

Tables 6 and 7 give, for each row of Table 1, the condition behind its verdict and the energy its source claims. Table 8 gives the same for the other 14 rows. Tables 9 and 10 give, for every row with a numerical check, whether we read released code or the equations alone, the configuration, the measured statistic and a negative control. Every number that Table 1 leaves out is in these two tables.

An admissible verdict holds at every parameter value that meets the row's condition. A failure is shown at a probed configuration, by the certificate of Proposition 2 or by a search over metrics. For the adaptive network it is also shown at trained parameters, by the orbits of its cycling fits. A failure rules out an energy in the sense of Definition 1. A periodic orbit rules out more, namely every function that decreases at every step. Six failure rows were iterated on real data, and two of them enter period-two orbits (Section 6 and Appendix H.6).

E.1 THE POPULATION AND HOW IT WAS ASSEMBLED

Population rule. We classify published models that meet three conditions. (i) The model defines a retrieval map $T : \mathbb{R}^d \to \mathbb{R}^d$ on a continuous state space. (ii) It appeared at NeurIPS, ICML, ICLR or JMLR between 2016 and 2026. (iii) Its source claims an energy for it, or the model is deployed as a layer. The rule names its exclusions. Capacity results and a survey define no retrieval map beyond those already classified (Lucibello & Mézard, 2024; Barnfield et al., 2026; Krotov et al., 2025). Energy-based models with no stored patterns (Gladstone et al., 2026) are excluded as well. Binary-state models (Krotov & Hopfield, 2016; Demircigil et al., 2017; Chaudhry et al., 2023) are excluded because their $\mathcal{D}$ is not open and convex. Where such a model has a standard continuous relaxation on $\mathbb{R}^d$, its energy with $\frac{1}{2}\|\mathbf{x}\|^2$ added, the relaxation is in scope and its row says so. The hierarchical memory of Krotov (2021) is a preprint and does not meet (ii). A model that deploys one of the classified maps unchanged is covered by that map's row. An admissible row may carry a condition, and its table states the condition. A row that is admissible only for small enough steps, and whose condition a probed configuration fails, has a class of its own.

Where the rule bends. The rule is applied with judgement, and it bends for at least three rows. The released Hopfield layer has no venue paper of its own. The conservative Krotov–Hopfield flow runs in continuous time and defines no map. The kernel memory network meets rule (iii) only loosely, and we include it because it iterates its map by design.

How the population was assembled. The population was not drawn by a systematic search, and it is not complete. It began with the literature the adaptive network builds on and grew through our own reviews of that literature's neighbours. Our notes record a decision for 69 candidates, of which 33 are the sources of the 32 rows (32 papers and the released code of the Hopfield layer). A title screen made afterwards, for "Hopfield" or "associative memory" in the 8 of the rule's 44 venue-years whose title lists we fetched, finds 10 titles that we have not classified.

How the rows were measured. A row marked code in Tables 9 and 10 was read from released code as well as from its equations. A row marked eq. was read from its equations alone, and its statistic, where it has one, is measured on our reimplementation of those equations. Universal-Hopfield

Table 6: Provenance of Table 1, part 1 of 2: the ten admissible rows of Table 1. The condition is the hypothesis behind the verdict. The source's energy reads Same map when the cited paper states an energy for the map we classify, whatever its proof covers, Related map when its energy covers other dynamics, and None when it claims no energy for that map.

Model	Condition	Energy claimed by the source
Modern Hopfield network (Ramsauer et al., 2021)	None. Base case, $\mathbf{\Phi} = \mathbf{\Xi}$, $G = \mathbf{I}$.	Same map. Energy (2) and update (3), the concave–convex step of the proof of their Theorem 1.
Sparse / entmax (Hu et al., 2023; Santos et al., 2024; Wu et al., 2024b; Hu et al., 2024a)	None. No separation of the line moves $\mathbf{\Phi} = \mathbf{\Xi}$.	Same map. Each paper states an energy with its update, and three of the four prove descent.
Kernelised / U-Hop (Wu et al., 2024a)	At $G = W^\top W$, the paper's own metric, with W of full column rank.	Same map. Energy (2.2) and update (2.4).
Squared distance, softmax sep (Millidge et al., 2022)	Softmax separation, whose shift invariance removes the shared $-\|\mathbf{x}\|^2$ term.	Related map. Their Lyapunov argument is for the continuous-time dynamics of their eq. (2).
Dot product, non-decreasing sep (Millidge et al., 2022)	Where $\mathrm{sep}' \geq 0$ on the realised range.	Same map. Descent is proved in continuous time only.
Dense assoc. memory, continuous relaxation (Krotov & Hopfield, 2016)	Rectified polynomial and even powers everywhere. Odd powers can fail on signed patterns.	Related map. Their energy is for the binary original.
ClAM clustering (Saha et al., 2023)	Step $\alpha = dt/\tau \leq 1$, which every published setting meets.	Same map, energy (1) and update (4). Descent is proved in continuous time only.
Classical / tanh (Santos et al., 2025)	$G = \mathbf{\Xi}\mathbf{\Xi}^\top$, positive definite only when $\mathbf{\Xi}$ has rank d.	Same map. Energy (6) with the Ψ of their eq. (9).
DeepRC, HELM (Widrich et al., 2020; Paischer et al., 2022)	None. Base case.	Same map for DeepRC, whose Section 2 restates the energy of Ramsauer et al. (2021) with update (1). None for HELM.
Released Hopfield layer (key-space loop)	None. The loop runs in key space, so untied projections cannot break it.	Same map, eq. (3) of Ramsauer et al. (2021) with $X = K^\top$, in the rescaled state $\mathbf{z} = \beta\boldsymbol{\xi}$.

Table 7: Provenance of Table 1, part 2 of 2: the eight failures of Table 1, in its order and in the columns of Table 6.

Model	Condition	Energy claimed by the source
Adaptive Hopfield network (Wang et al., 2026)	Pattern readout. The certificate fires even with a monotone gate.	Same map. Their Table 1 and Theorem 2 state it, and the proof uses a score with no sort.
One-step attention, value readout (Vaswani et al., 2017)	None. Scores with K, reads out V.	None. Attention is applied once, with no energy.
Nonparametric (Hu et al., 2025)	A contamination off the flat. Excluded by a non-real eigenvalue and by a metric search.	None. The paper defines no energy.
Modern Hopfield attention (Masumura & Taki, 2025)	Untied $W_1 \neq W_2$. Tied, the spectrum is real.	None. The authors say untying violates the energy argument and leave an energy open.
Kernel memory network (Iatropoulos et al., 2022)	Outside the separation its properties require, and at some probes inside it.	None. Energy only as a picture.
Universal Hopfield, Manhattan (Millidge et al., 2022)	None. A kink in the score, and a max-cell witness for the limit $\beta \to \infty$.	Related map, the continuous-time dynamics of their eq. (2).
Universal Hopfield, Euclidean, non-shift-invariant sep (Millidge et al., 2022)	Non-shift-invariant separations, a combination the paper defines and does not run.	Related map, as for the Manhattan row.
Universal Hopfield as released (Millidge et al., 2022)	Reciprocal distances and sum normalisation.	Related map. Their proof covers at most three of the six similarities, and none in the form they run.

rows were differentiated in double precision at 12 probes with $d = 6$ and $M = 4$, on patterns in $(0, 1)$ and probes perturbed by Gaussian noise and clipped to $[0.02, 1]$. Sparse and Fenchel–Young rows were differentiated by complex step at 120 queries on each of five seeds with $d = 6$, $M = 9$ and $\beta = 1.5$. They were also checked by central differences at 40 queries with $d = 5$, $M = 8$ and $\beta = 1$, worst case over queries. A complex-step row leaves out the queries where the map is flat, with $\|J\|$ below 10^{-3} of its largest value, because there round-off is large relative to $\|J\|$. Every other row reports the worst case over all its queries.

Table 8: The other 14 rows, in the columns of Table 6: three further admissible models, the two rows admissible only under a step-size condition, the six rows with an energy outside Definition 1 and the three undetermined rows. Every failure is in Table 1. Three more models deploy the map of one-step attention with a value readout and are covered by that row, which Appendix E.6 names.

Model	Condition	Energy claimed by the source
Admissible		
In-context denoising (Smart et al., 2025)	One attention step at exact scaled-identity weights, an idealisation of its trained weights. The paper states the symmetry condition that general weights need.	Same map at those weights, energy (16) and step (17).
Larimar, mean read (Das et al., 2024)	The mean read only, an orthogonal projection with $G = \mathbf{I}$.	None.
ALDA scalar codebooks (Batra & Sukhatme, 2025)	One dimension per coordinate. There a Manhattan score gives a non-decreasing map, which is the gradient of a convex function.	None.
Admissible only under a step-size condition		
Random-feature DrDAM (Hoover et al., 2024)	An exact gradient step on the approximate energy $\hat{E}$, admissible where $\eta\nabla^2\hat{E} \preceq \mathbf{I}$. With the clip of its footnote 6, a small enough fixed step works on all of $\mathbb{R}^d$.	Same map, for the approximate energy $\hat{E}$ of their eq. (9). Nothing follows for the original energy.
Epanechnikov dense assoc. memory (Hoover et al., 2025)	The gradient dynamics of its energy, which its retrieval runs before its Algorithm 1. For $\epsilon > 0$ it is admissible at a support sphere only if the step is at most $\epsilon^2/(2+\epsilon)$.	Same map, energy (3). Their Theorem 1 uses a state-dependent step.
Energy by construction, outside Definition 1		
Lagrangian nets, Energy Transformer (Krotov & Hopfield, 2021; Hoover et al., 2023)	Descent in the state-dependent Hessian metric of the Lagrangian. Constant G not tested.	Same map. Both descend their energy by construction.
ℓ_2 / layer-norm post-transform (Santos et al., 2025)	A convex post-transform with a proven energy. No constant metric, since the search finds dimension 0, or 1 with a singular G.	Same map, energy (6) with the indicator Ψ of their Section 3.3.6. Their Proposition 2 derives the step for a differentiable Ψ.
Classical / tanh at $M < d$ (Santos et al., 2025)	The explicit h of the tanh row, with $\boldsymbol{\Xi}\boldsymbol{\Xi}^\top$ singular. No positive-definite G for generic patterns.	Same map, with the energy of the tanh row.
Kernelised retrieval at $D_\Phi < d$ (Hu et al., 2024b)	The U-Hop map with $G = WW^\top$ of rank $D_\Phi < d$, as in its MNIST run, so G is singular.	Same map, energy (1.3) with its sign corrected.
NRGPT, Hyper-SET (Dehmamy et al., 2026; Hu et al., 2026)	Gradient steps on an attention energy through a layer norm or a sphere constraint, a state-dependent metric. Constant G not tested.	Same map. Descent proved after a transient (NRGPT) or by construction.
Conservative Krotov–Hopfield flow (Fanaskov & Oseledets, 2025)	A continuous-time flow, made conservative by construction, and not a map.	Same map, energy by construction.
Undetermined		
Simplicial (Burns & Fukai, 2023)	Reading-dependent. On diluted networks the update is not the concave–convex step of the energy. No code is released.	Same map, energy (6) and update (7). Its results concern capacity and recall.
GSEMM, EDEN (Karuvally et al., 2023; 2025)	Asymmetric by design. Both descend only while the slow variable is frozen.	Related map, with the slow variable frozen.
Hyperbolic memory (Hao et al., 2026)	A Riemannian concave–convex step on the open Poincaré ball. Constant G not tested.	Same map. Descent proved locally, in the ball's metric.

Table 9: Evidence for the rows with an energy in a constant metric: the ten admissible rows of Table 1, then the three admissible rows and the two step-size rows of Table 8. Read is code for a model read from released code as well as its equations, and eq. for a model read from its equations alone (paragraph "How the rows were measured"). The statistic $\mathrm{asym}(J) := \|J - J^\top\|/\|J\|$ is the relative Jacobian asymmetry, maximised over queries unless the cell says otherwise. $\lambda_{\min}$ is the smallest eigenvalue, $|\mathrm{Im}|$ the largest imaginary part of an eigenvalue, and FD a central-difference check. The admissible rows are measured on a derivative of the implemented map, automatic or by complex step. The rows of DeepRC and HELM and of in-context denoising are the base case read from equations alone, with no numerical check, and are not listed.

Row	Read	Configuration	Statistic	Negative control	Control statistic
Modern Hopfield network	code	12 probes	asym $1.1 \cdot 10^{-15}$	scores divided by their own sum	asym 0.90
Sparse / entmax	code	complex step, 5 seeds, 120 queries, $d = 6$, $M = 9$, $\beta = 1.5$. FD at 40 queries, $d = 5$, $M = 8$, $\beta = 1$	asym $\leq 4.9 \cdot 10^{-14}$ over sparsemax and 1.5-entmax, and $8.2 \cdot 10^{-14}$ for softmax. $\lambda_{\min}/\|J\| \geq -1.2 \cdot 10^{-14}$. FD $8.7 \cdot 10^{-11}$–$7.2 \cdot 10^{-10}$	ℓ_2-norm or layer-norm post-transform, eq. (23) of Santos et al. (2025)	asym 1.06–1.16
Kernelised / U-Hop	code	40 queries, $d = 5$, $M = 8$, $\beta = 1$, complex step	asym $1.6 \cdot 10^{-13}$ at $G = W^\top W$. FD $4.4 \cdot 10^{-10}$ at $\beta = 0.05$	same map at $G = \mathbf{I}$	asym 1.03
Squared distance, softmax sep	eq.	12 probes, $d = 6$, $M = 4$	asym $9.4 \cdot 10^{-16}$, $\|\nabla h - T\|$ $3.6 \cdot 10^{-16}$, Hessian $\lambda_{\min}$ $-5.5 \cdot 10^{-17}$	identity separation	asym 1.43, $\|\mathrm{Im}\|$ 1.60
Dot product, non-decreasing sep	eq.	12 probes, $d = 6$, $M = 4$, with the identity, softmax, square, tenth power and log	asym $\leq 1.1 \cdot 10^{-15}$	square separation on signed patterns	asym $1.7 \cdot 10^{-16}$, scaled $\lambda_{\min}$ of the symmetric part -0.98
Dense assoc. memory, rectified polynomial	eq.	1000 draws, $d = 24$, $M = 40$, $\beta = 1$, signed Gaussian patterns, F of degree 3 and 11	asym 10^{-16}	unrectified $F = z^3/3$ and $z^{11}/11$	$\lambda_{\min}$ -4.1 and -5.1 in 1000 of 1000 draws
ClAM, $\alpha \in \{0.05, 0.5, 1\}$	eq.	20 probes, $d = 6$, $M = 4$, $\beta \in \{0.5, 5\}$	asym $\leq 8.7 \cdot 10^{-15}$, $\lambda_{\min} = 1 - \alpha$	$\alpha = 1.5$	$\lambda_{\min}$ -0.5, and the energy still never rises
Classical / tanh	eq.	40 queries, $d = 5$, $M = 8$, $\beta = 1$, complex step	asym $1.5 \cdot 10^{-16}$ at $G = \boldsymbol{\Xi}\boldsymbol{\Xi}^\top$, $\lambda_{\min}/\|GJ\|$ $-7.0 \cdot 10^{-17}$. A metric search finds a G with $\lambda_{\min}$ $+0.0975$	Euclidean metric	asym 0.97
Released Hopfield layer (key-space loop)	code	5 seeds $\times$ 400 draws, $d = 24$, $M = 40$, $\beta = 1/\sqrt{24}$, untied projections, LayerNorm folded in, complex step	asym $1.57 \cdot 10^{-16}$, $\lambda_{\min}$ $-1.6 \cdot 10^{-16}$, $\|\mathrm{Im}\|$ 0.0	update moved into value space	asym 1.53, $\lambda_{\min}$ -0.14, $\|\mathrm{Im}\|$ $7.4 \cdot 10^{-2}$
Larimar, mean read	eq.	5 slots, $d = 9$, 20 queries	asym $6.2 \cdot 10^{-16}$, eigenvalues 0 and 1 to $1.3 \cdot 10^{-15}$	none	none
ALDA	eq.	50 codebooks of 12 scalars, $\beta \in [0.1, 100]$, 20 001 grid points	smallest increment $-1.1 \cdot 10^{-15}$	the same score at $d = 6$, the Manhattan row	$\|\mathrm{Im}\|$ 0.13
DrDAM, $\eta = 0.1$	eq.	20 probes, $d = 6$, $M = 4$, $Y \in \{16, 4096\}$	asym $\leq 1.9 \cdot 10^{-16}$, $\lambda_{\min}$ from $+0.89$ to -1.23, the latter at $\beta = 20$, $Y = 16$	the map of their eq. (32)	asym 1.30 at $Y = 16$
Epanechnikov (log-sum-ReLU)	eq.	$d = 4$, one support, $\beta = 1$, $\epsilon \in \{0.1, 0.01\}$, 20 probes at the support sphere	$\lambda_{\max}(\nabla^2 E)$ 210 and $2.01 \cdot 10^4$, as $1/\epsilon + 2/\epsilon^2$	a fixed step of 0.01, the largest they report, for their log-sum-exp runs	not positive semidefinite at the sphere

Table 10: Evidence for the eight failures of Table 1, then for the rows of Table 8 with an energy outside Definition 1 in the notation of Table 9. The rows read from equations alone with no measurement, the Lagrangian networks, the Energy Transformer, NRGPT, Hyper-SET, the conservative flow and the three undetermined rows, are not listed.

Row	Read	Configuration	Statistic	Negative control	Control statistic
Adaptive Hopfield network	code	$d = 32$, $M = 64$, $\beta = 30$, mask 0.6, 40 queries per seed, 5 seeds, a monotone gate under the pattern readout	median asym 0.71–0.94. In the configuration of Figure 1 ($M = 48$, mask 0.5, 120 queries) the median is 1.3, and a non-real eigenvalue appears in 70.8% of Jacobians and 31.5% of eigenvalues	gradient readout of the same score	asym $\leq 3.3 \cdot 10^{-16}$ by complex step, and over 12 000 transitions 0 energy increases and 0 violations of the descent inequality
One-step attention, value readout	code	5 seeds $\times$ 400 draws, $d = 24$, $M = 40$, $\beta = 1/\sqrt{24}$, untied projections. Also $d = 8$, $M = 6$, 5 seeds, 8 starts, 60 steps	asym 1.53, $\lambda_{\min} - 0.14$, $\lvert\mathrm{Im}\rvert$ $7.4 \cdot 10^{-2}$. At $d = 8$, asym 1.54 at $\beta = 1$ and 1.53 at $\beta = 30$	key-space loop of the same layer	asym $1.57 \cdot 10^{-16}$
Nonparametric	eq.	40 queries, $d = 5$, $M = 8$, $\beta = 1$. The published score and a squared-distance analogue at $d = 8$, $M = 6$, 5 seeds, 8 starts, 60 steps	asym $2.0 \cdot 10^{-16}$, 0.058, 0.331 at contamination 0, 0.05, 0.3. At 0.3 one probe has a non-real eigenvalue, $\lvert \mathrm{Im}\, \lambda\rvert = 4.6 \cdot 10^{-3}$. At $d = 8$ and contamination 0.1, 0.3, the published score gives 0.38, 1.00 ($\beta = 1$) and 0.56, 1.31 ($\beta = 30$), and the analogue gives 0.51, 1.13 ($\beta = 1$) and 0.56, 1.32 ($\beta = 30$). A metric search finds no constant G	contamination zero	asym $2.0 \cdot 10^{-16}$, analogue 0.00 at both β
Modern Hopfield attention	eq.	joint state $(\mathbf{x}, \mathbf{h})$, $d = 8$, $M = 6$, $\alpha = \alpha' = 0.5$, 20 probes	$\lvert\mathrm{Im}\rvert$ 0.19, non-real at 20 of 20 probes, asym 0.55	tied, $W_1 = W_2$	$\lvert\mathrm{Im}\rvert$ $2.3 \cdot 10^{-9}$, real at 20 of 20
Kernel memory network	eq.	$N = 16$, $M = 8$, $\beta \in \{2, 4\}$, 20 noisy queries, $r = d_{\mathrm{median}}$ (outside the separation of its Properties 5.1 to 6) and $r = 0.4\, d_{\min}$ (inside)	Outside, $\lvert\mathrm{Im}\rvert$ 0.21 and 0.18, non-real at 12 and 8 of 20. Inside, asym 1.41, and non-real at 3 of 20 at $\beta = 2$	modern Hopfield update on the same patterns	asym $4.8 \cdot 10^{-16}$
Universal Hopfield, Manhattan	code	softmax separation, $d = 6$, $M = 4$, 12 probes at $\beta = 1$, 2000 probes at $\beta = 10$ and 100. Max separation, 4000 and 20 000 draws	asym 1.26 and $\lvert\mathrm{Im}\rvert$ 0.13 at $\beta = 1$ over the 12 probes. $\lvert\mathrm{Im}\rvert$ 1.22 in 76 of 2000 probes at $\beta = 10$, and 2.72 in 3 of 2000 at $\beta = 100$. Max separation: non-convex ℓ_1 cells, 5 of 4000 and 32 of 20 000 witnesses	Euclidean distance under the max separation	0 of 4000 and 0 of 20 000 witnesses
Universal Hopfield, Euclidean, non-shift-invariant sep	eq.	12 probes, $d = 6$, $M = 4$	$\lvert\mathrm{Im}\rvert$ 1.60, 2.43, 1.30, 2.02 for identity, square, tenth power, log. asym 1.38–1.43	the same distance under softmax	asym $9.4 \cdot 10^{-16}$
Universal Hopfield as released	code	12 probes, $d = 6$, $M = 4$	Manhattan under softmax asym 1.40, $\lvert\mathrm{Im}\rvert$ $1.3 \cdot 10^{-2}$. Dot product asym 0.90	normalisation switched off	dot product asym $1.1 \cdot 10^{-15}$, Manhattan stays at 1.40
ℓ_2 / layer-norm post-transform	code	$d = 5$, $M = 8$, $\beta = 1$, FD. 40 queries for asym, 25 for the metric search	asym 1.16 (ℓ_2) and 1.06 (layer norm) at $G = \mathbf{I}$. Metric solution space of dimension 0 and 1, the latter singular	softmax readout, no post-transform	a positive-definite G from the same search
Classical / tanh at $M < d$	eq.	$d = 5$, $M = 3$, 40 queries, FD	every symmetric H with HJ_T symmetric is a multiple of $\Xi\Xi^\top$, of rank 3	$\boldsymbol{\xi} = (1, 1)$, $d = 2$, $M = 1$, one query	asym $1.0 \cdot 10^{-12}$ at $G = \mathbf{I}$
Kernelised retrieval at $D_\Phi < d$	eq.	20 queries, $d = 6$, $M = 12$, $\beta = 1$, $D_\Phi = 3$	every symmetric H with HJ symmetric is a multiple of one singular matrix, smallest eigenvalue $-2.1 \cdot 10^{-16}$	$D_\Phi = d$	a positive-definite H

E.2 ADMISSIBLE ROWS

Each row of this subsection has an energy in a constant metric wherever the condition in Table 6 or Table 8 holds. The ten rows of Table 1 come first, in its order, and the three admissible rows of Table 8 follow.

Modern Hopfield network. The energy is equation (2) of Ramsauer et al. (2021),

$$E = -\operatorname{lse}(\beta, X^\top \boldsymbol{\xi}) + \tfrac{1}{2}\boldsymbol{\xi}^\top \boldsymbol{\xi} + \beta^{-1}\log N + \tfrac{1}{2}M^2,$$

and the update is their equation (3),

$$\boldsymbol{\xi}^{\text{new}} = X \operatorname{softmax}(\beta X^\top \boldsymbol{\xi}).$$

Similarity and readout share one matrix. Hence $\boldsymbol{\Phi} = \boldsymbol{\Xi}$ and $G = \mathbf{I}$, the base case of Theorem 1(i). The proof of their Theorem 1 (their Theorem A1) identifies equation (3) as the concave–convex step. The negative control divides the scores by their own sum, as the universal-Hopfield reimplementation does.

Sparse and entmax networks. Every update in this line has the shape

$$\mathbf{x} \leftarrow \boldsymbol{\Xi}\ \operatorname{sep}(\beta \boldsymbol{\Xi}^\top \mathbf{x}).$$

Each paper states the energy and the update, as equations (8) and (9) of Santos et al. (2024), (20) and (21) of Santos et al. (2025), (2.1) and (2.5) of Hu et al. (2023), (2.4) and (2.6) of Hu et al. (2024a), and (3.2) and (3.4) of Wu et al. (2024b). Replacing softmax by sparsemax, α-entmax, γ-normmax, SparseMAP or softmax$_1$ does not move $\boldsymbol{\Phi}$. The row therefore sits at the flat's dot-product point, $\lambda = 1$ and $\mathbf{v} = 0$, and Corollary 3(b) applies. The statistic is the worst complex-step asymmetry over sparsemax and 1.5-entmax. Softmax, measured in the same way, reaches $8.2 \cdot 10^{-14}$. The central-difference check sits at the finite-difference floor. The negative control is equation (23) of Santos et al. (2025). Its ℓ_2-normalised or layer-norm post-transform admits no constant G (Appendix E.5), although its concave–convex step still descends its energy.

Kernelised retrieval (U-Hop). The energy is equation (2.2) of Wu et al. (2024a),

$$E_K(\mathbf{x}) = \tfrac{1}{2}K(\mathbf{x}, \mathbf{x}) - \Psi^*_\alpha\big(\beta, K(\boldsymbol{\Xi}^\top \mathbf{x})\big),$$

with the feature map $\Phi(\mathbf{u}) = W\mathbf{u}$ of their equation (2.1) and the update of their equation (2.4). Their Assumption 2.1 requires W to be of full rank, and the energy carries the term $\frac{1}{2}\mathbf{x}^\top W^\top W \mathbf{x}$. Hence $G = W^\top W$ is the paper's own metric. The analytic Jacobian is $J_T = \beta \boldsymbol{\Xi} A \boldsymbol{\Xi}^\top G$. The finite-difference run is taken at $\beta = 0.05$, where this G no longer saturates the softmax. The negative control is the same map under $G = \mathbf{I}$.

Squared distance under the softmax separation. The framework is equation (1) of Millidge et al. (2022),

$$\mathbf{z} = P\ \operatorname{sep}\big(\operatorname{sim}(M, \mathbf{q})\big),$$

here with the Euclidean row of its Table 2 and the softmax row of its Table 3. As in its main text and its Appendix E, scores increase with similarity. The distance therefore enters negated, and the softmax has a positive exponent. The $-\|\mathbf{x}\|^2$ term of the expanded score is common to every pattern, and $A\mathbf{1} = 0$ removes it. The score therefore splits with $\boldsymbol{\Phi} = 2\boldsymbol{\Xi}$ at $\lambda = 2$ (Corollary 3(a)), and its explicit potential is

$$h(\mathbf{x}) = \tfrac{1}{2\beta}\operatorname{lse}_\mu\big(2\beta\, \boldsymbol{\xi}_\mu^\top \mathbf{x} - \beta\|\boldsymbol{\xi}_\mu\|^2\big).$$

The negative control is the identity separation, which is not shift-invariant.

Dot product under a non-decreasing separation. The same equation (1), with the dot-product row of Table 2, each separation of Table 3 and the same sign convention, gives

$$T(\mathbf{x}) = \boldsymbol{\Xi}\ \operatorname{sep}(\beta \boldsymbol{\Xi}^\top \mathbf{x}).$$

Its Jacobian $\beta \boldsymbol{\Xi} \operatorname{diag}(\operatorname{sep}')\boldsymbol{\Xi}^\top$ is symmetric for every elementwise sep and positive semidefinite wherever $\operatorname{sep}' \geq 0$ (Corollary 3(b)). The square and the tenth power need non-negative scores, and the logarithm needs positive ones. The negative control is the square separation on signed patterns. There symmetry survives and positive semidefiniteness fails.

Dense associative memory, continuous relaxation. Krotov & Hopfield (2016) define the model on binary states. The original is therefore a named exclusion of the population rule. The row is its continuous relaxation, which adds $\frac{1}{2}\|\mathbf{x}\|^2$ to their energy $-\sum_\mu F(\boldsymbol{\xi}_\mu^\top \mathbf{x})$, as Ramsauer et al. (2021) add it to theirs, and takes the concave–convex step in the Euclidean metric. The step reads out through F', as their one-step update does (their eqs. (10) and (11)), so

$$T(\mathbf{x}) = \boldsymbol{\Xi} F'(\boldsymbol{\Xi}^\top \mathbf{x}), \qquad J_T(\mathbf{x}) = \boldsymbol{\Xi}\,\mathrm{diag}\left(F''(\mathbf{u})\right)\boldsymbol{\Xi}^\top.$$

The Jacobian is always symmetric. Their rectified $F(z) = \max(z, 0)^n$ with $n \geq 2$ is convex. Hence $F'' \geq 0$ for every real argument, and the relaxation is positive semidefinite and admissible everywhere. So is every even power z^n, and $n = 2$ is their standard model. The energy

$$E(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|^2 - \sum_\mu \max(\boldsymbol{\xi}_\mu^\top \mathbf{x}, 0)^n$$

is unbounded below along $t\boldsymbol{\xi}_\mu$ as $t \to \infty$ for $n \geq 3$, and for $n = 2$ once $\|\boldsymbol{\xi}_\mu\|^2 > 1/2$. This is the case that Proposition 5 describes for the dot-product base. The binary original is bounded on its cube. An *odd* power z^n with $n \geq 3$ has $F'' < 0$ wherever a score is negative. On signed patterns the energies $z^3/3$ and $z^{11}/11$ then have a negative Jacobian eigenvalue in every draw, while the asymmetry stays at machine precision. A test of symmetry alone cannot detect this failure. Not every signed set fails, since antipodal patterns cancel the odd terms and make $T = 0$. Non-negative scores make the Jacobian positive semidefinite. They do not make the map admissible, since the map can leave the region where they hold.

ClAM. The energy is equation (1) of Saha et al. (2023),

$$E(\mathbf{v}) = -\beta^{-1} \log \sum_\mu \exp\left(-\beta\|\boldsymbol{\rho}_\mu - \mathbf{v}\|^2\right),$$

and the update is their equation (4). With $\alpha = dt/\tau$ the update is

$$T(\mathbf{x}) = (1 - \alpha)\mathbf{x} + \alpha\boldsymbol{\Xi}\mathbf{p},$$

a combination of the identity and the map of the squared-distance row. Its Jacobian $(1 - \alpha)\mathbf{I} + 2\alpha\beta\boldsymbol{\Xi}A\boldsymbol{\Xi}^\top$ is symmetric. It is positive semidefinite for every $\mathbf{x}$ exactly when $\alpha \leq 1$, a bound that involves neither β nor the patterns. Table 5 of that paper sets $\alpha = 1/T \in [0.05, 0.5]$, so every published configuration is admissible. That paper proves descent in continuous time only. The negative controls are $\alpha = 1.5$, where $\lambda_{\min} = -0.5$ and the energy still never increases, and $\alpha = 3$, where it increases at every step. The first control shows that a gradient step can descend and still fail Definition 1, which asks for a convex potential and not only for descent.

Classical network with tanh. The model is the classical instance of Santos et al. (2025), Section 3.3.1, with energy (6), $\Omega = \frac{1}{2}\|\mathbf{y}\|^2$ and the binary-entropy Ψ of their equation (9). In our $d \times M$ convention it updates

$$\mathbf{x} \leftarrow \tanh(\beta\boldsymbol{\Xi}\boldsymbol{\Xi}^\top \mathbf{x}).$$

The Jacobian is $D(\mathbf{x})\,\Gamma$, with $\Gamma = \boldsymbol{\Xi}\boldsymbol{\Xi}^\top$ symmetric and $D(\mathbf{x}) = \mathrm{diag}(\beta\,\mathrm{sech}^2(\cdot))$ positive. Hence

$$\Gamma J_T = \Gamma D(\mathbf{x}) \Gamma \succeq 0,$$

a symmetric matrix. Definition 1 then holds with the constant $G = \Gamma$, which is positive definite when the pattern matrix has rank d. A search over metrics recovers a positive-definite G independently. The negative control is the Euclidean metric. Appendix E.5 treats the case $M < d$, where Γ is singular.

DeepRC and HELM. Section 2 of Widrich et al. (2020) restates the update of Ramsauer et al. (2021) as its equation (1),

$$\boldsymbol{\xi}^{\mathrm{new}} = X\,\mathrm{softmax}(\beta X^\top \boldsymbol{\xi}),$$

and its equation (2) applies W_V outside the update. The row is therefore the base case. Its attention pooling, equation (4), scores with one network, reads out another and claims no energy. Equation (1) of Paischer et al. (2022),

$$\mathbf{x}_t^\top = \sigma(\beta\,\mathbf{o}_t^\top P^\top E^\top)E,$$

scores with the frozen embedding E and reads out E. The random projection P sets only the initial condition. Neither row carries a statistic, because both are the base case.

Released Hopfield layer (key-space loop). The layer released with the code of Ramsauer et al. (2021) iterates in head space. Its loop is

$$\mathbf{z} \mapsto \beta K^\top \operatorname{softmax}(K\mathbf{z}),$$

which is equation (3) of Ramsauer et al. (2021) with $X = K^\top$ in the rescaled state $\mathbf{z} = \beta\boldsymbol{\xi}$. It is admissible with $G = \mathbf{I}$ and $h(\mathbf{z}) = \beta \operatorname{lse}(K\mathbf{z})$. The value readout runs once after the loop, as equation (10) of that paper writes it, and the released layer applies its output projection after that. Untying W_Q, W_K and W_V therefore cannot break the energy. The measurement uses fully untied random projections with the LayerNorm folded in. The negative control moves the update into value space, which gives the one-step attention row of Appendix E.4.

In-context denoising. Smart et al. (2025) study the attention step of their equation (12),

$$\hat{\mathbf{x}} = W_{PV} X \operatorname{softmax}(X^\top W_{KQ} \tilde{\mathbf{x}}).$$

Trained, its weights approach the scaled identities $W_{PV} \approx \alpha\mathbf{I}$ and $W_{KQ} \approx \beta\mathbf{I}$ (their p. 7). At exact scaled-identity weights the step is $\alpha X \operatorname{softmax}(\beta X^\top \mathbf{x})$, the gradient of the convex function $(\alpha/\beta) \operatorname{lse}(\beta X^\top \mathbf{x})$, which is the base case. Their energy (16) and gradient step (17) say the same. The row certifies this idealisation. It does not certify the approximate trained weights, since a small rotation of W_{PV} can give the Jacobian a non-real eigenvalue. Their Appendix I.2 states the symmetry condition that general weights would need.

Larimar, mean read. The read of Das et al. (2024) (their Algorithm 1) samples an addressing matrix around ZM^+ and returns its product with the memory M. The mean read

$$\mathbf{z} \mapsto \mathbf{z}M^+M$$

is the orthogonal projection onto the row space of M. It is the gradient of the convex function $\frac{1}{2}\mathbf{z}M^+M\mathbf{z}^\top$, so $G = \mathbf{I}$. The sampled read is random and is not a map. The model claims no energy and is deployed as a layer.

ALDA scalar codebooks. Batra & Sukhatme (2025) read out each latent coordinate from a scalar codebook $\mathbf{v}$ (their equation 6 and Algorithm 2) as

$$x \mapsto \sum_k \operatorname{softmax}(-\beta|x - v_k|)_k \, v_k,$$

a Manhattan softmax in one dimension. A map of one variable is the gradient of a convex function exactly when it is non-decreasing. This map is non-decreasing, because its derivative $\beta \operatorname{Cov}_{\mathbf{p}}(v, \operatorname{sign}(v - x))$ is a covariance of v with a non-decreasing function of v. The map acts coordinate by coordinate. Its Jacobian is therefore diagonal and non-negative. The same score fails in six dimensions, as the Manhattan row of Appendix E.4 shows. The kink that excludes that row therefore needs more than one dimension.

E.3 Rows admissible only under a step-size condition

Both rows of this subsection take gradient steps on an energy. Each is admissible where the step is short relative to the curvature of that energy. For DrDAM a small enough fixed step works on all of $\mathbb{R}^d$, because its implementation clips the argument of the logarithm. For Epanechnikov memory the admissible step tends to zero as $\epsilon \to 0$.

Random-feature DrDAM. Hoover et al. (2024) replace the energy of their equation (11) by that of their equation (9), clipped as in their footnote 6,

$$\hat{E}(\mathbf{x}) = -\beta^{-1} \log \max\big(\langle \varphi(\sqrt{\beta}\mathbf{x}), \mathbf{T}\rangle, \, \epsilon\big),$$

and descend it by their equation (5) with the gradient of their equation (10), computed by automatic differentiation. The map is an exact gradient step, so its Jacobian is symmetric. It is $\mathbf{I}$-admissible for $\hat{E}$ where $\eta\nabla^2\hat{E} \preceq \mathbf{I}$. The inner product is a trigonometric sum that reaches zero, and their footnote 6 clips it at a small $\epsilon > 0$. Without the clip, no step size would work on all of $\mathbb{R}^d$. With it, the curvature

of $\hat{E}$ is bounded where the clip is inactive, and the clip adds only convex kinks to $\frac{1}{2}\|\mathbf{x}\|^2 - \eta\hat{E}$, so a small enough fixed step, which depends on ϵ, β and the features, works everywhere. At the step size $\eta = 0.1$ of their Appendix D the condition holds at $\beta = 2$. It fails at $\beta = 20$ with $Y = 16$ features, a deliberately small Y, where the exact energy rises on 82 of 200 steps while $\hat{E}$ never rises. Nothing follows for the original energy. The negative control is the map of their equations (29) to (32), which their Theorem 2 analyses. It is a pattern readout with random-feature weights, and it is not a gradient at finite Y (asym 1.30 at $Y = 16$ and 0.17 at $Y = 4096$, $\beta = 20$).

Epanechnikov memory. Hoover et al. (2025) define the log-sum-ReLU energy of their equation (3),

$$E(\mathbf{x}) = -\beta^{-1} \log\Big(\epsilon + \sum_{\mu} \mathrm{ReLU}\big(1 - \tfrac{\beta}{2}\|\mathbf{x} - \boldsymbol{\xi}_\mu\|^2\big)\Big),$$

with a small $\epsilon \geq 0$ whose value their descent runs do not give. Their retrieval takes gradient steps, their equation (6) with the gradient of their equation (7). It then runs their Algorithm 1, which moves the state to the mean of the patterns whose support contains it. The row classifies the gradient dynamics. The step sizes they report, 0.01 to 0.0001 and 0.002, are for their log-sum-exp runs, and their log-sum-ReLU runs state none. As for DrDAM, a step $\mathbf{x} - \alpha\nabla E$ is $\mathbf{I}$-admissible where $\alpha\nabla^2 E \preceq \mathbf{I}$. Inside one support,

$$\nabla^2 E = \frac{\mathbf{I}}{s} + \frac{\beta(\mathbf{x} - \boldsymbol{\xi}_\mu)(\mathbf{x} - \boldsymbol{\xi}_\mu)^\top}{s^2}, \qquad s = \epsilon + 1 - \tfrac{\beta}{2}\|\mathbf{x} - \boldsymbol{\xi}_\mu\|^2.$$

Its largest eigenvalue reaches $1/\epsilon + 2/\epsilon^2$ at the support sphere. The step is therefore admissible there only if $\alpha \leq \epsilon^2/(2+\epsilon)$, and no fixed step works as $\epsilon \to 0$. The energy has a concave kink at the sphere. That kink is a convex kink of $\frac{1}{2}\|\mathbf{x}\|^2 - \alpha E$ and costs nothing. The proof of their Theorem 1 (their p. 19) uses the state-dependent step $\epsilon + \mathrm{ReLU}(1 - \frac{\beta}{2}\|\mathbf{x} - \boldsymbol{\xi}_\mu\|^2)$, which returns $\boldsymbol{\xi}_\mu$ in one step inside a basin. Algorithm 1 is piecewise constant, so every Jacobian statistic passes it, and we do not classify it.

E.4 Rows with no energy guarantee

These are the eight failures of Table 1, in its order. A non-real Jacobian eigenvalue shows all eight. For the nonparametric row it appears at contamination 0.3, and a search over metrics agrees.

Adaptive Hopfield network. The score is equation (4) of Wang et al. (2026), computed from a squared-distance base and a dot-product base. Each base is sorted per pattern and gated per rank by

$$c_i = \sum_{j \geq i} w_j,$$

where $\mathbf{w}$ is a raw parameter with no constraint, normalisation or sign restriction. The released code uses a block-constant version of this gate, $c_i = \sum_{j \geq \lceil i/4 \rceil} w_j$, with stride 4. Their Theorem 2 reads "Energy $E(x)$ will be monotonically decreasing, convergent, bounded from below, for isotropic noisy, and biased variants, if the following energy is used: $E(x) = -\mathrm{lse}(s(\Xi, x))$" (their p. 8). Its proof, in their Appendix A.2.2, uses the score $-\|\mathbf{x} - \boldsymbol{\xi}_\mu\|^2 + \mathbf{b}^\top(\mathbf{x} - \boldsymbol{\xi}_\mu)$, which has no footprint and no sort. The readout is $\boldsymbol{\Xi}\mathbf{p}$ and the score does not split. Theorem 1 therefore does not apply, and the certificate decides. The statistic of Table 10 is measured with a monotone gate under the pattern readout, so the asymmetry comes from the readout alone. A signed gate under the same readout gives an asymmetry of 1.4 (Appendix D). The 70.8% of Table 10 is the fraction of Jacobians, computed in closed form, with an eigenvalue whose imaginary part exceeds 10^{-8} in modulus. By Lemma 1 each such Jacobian excludes every constant metric.

Training does not keep the gate monotone. In the MNIST run of Appendix G.5 the trained gate has a negative entry in 11 of 15 configurations, 1 of 5 at mask 0.7 and 5 of 5 at 0.8 and 0.9, and its entries range over $[-4.19, 408.27]$. The row therefore records the absence of a guarantee at trained parameters as well. For a trained signed gate whose map cycles, as in 197 of the 400 fits of Section 6, the period-two orbit is the evidence. It excludes every function that decreases at every step, since such a function would need $E(\mathbf{x}_a) > E(\mathbf{x}_b) > E(\mathbf{x}_a)$. The negative control is the gradient readout of the same score. Its 12 000 transitions in Table 10 are five seeds of 40 queries and 60 steps, with no energy increase and no violation of the descent inequality. The 1800 transitions of a separate check find no violation of the descent inequality either (Appendix G.2).

One-step attention with a value readout. The map of Vaswani et al. (2017) is

$$\mathrm{softmax}(\beta QK^\top)V.$$

It scores with the keys and reads out the values. Hence $\mathbf{\Phi} = K \neq V$, and the Jacobian $\beta VAK^\top$ of Lemma 5 is not symmetric. The map is the value readout that the released Hopfield layer runs once after its loop (Appendix E.2), and it serves as the negative control of that row. Its own negative control is the key-space loop of the same layer. The Energy Transformer paper already names the missing gradient term. Attention as Vaswani et al. (2017) define it is applied once, and that paper claims no energy for it.

Nonparametric model. The update is equations (3.5) and (3.7) of Hu et al. (2025), the linear-kernel member of its Proposition E.1,

$$T(\mathbf{x}) = \mathbf{\Xi}\,\mathrm{softmax}(\beta\mathbf{\Xi}_\delta^\top\mathbf{x}), \qquad \mathbf{\Xi}_\delta = \mathbf{\Xi} + \Delta.$$

Here $\Delta = [\delta\boldsymbol{\xi}_1, \ldots, \delta\boldsymbol{\xi}_M]$ is a per-pattern perturbation and not of the form $(\lambda - 1)\mathbf{\Xi} + \mathbf{v}\mathbf{1}^\top$, since such a perturbation would keep the score in the flat of Theorem 1. The score uses the contaminated patterns, and the readout uses the clean ones. Hence $\mathbf{\Phi} = \mathbf{\Xi}_\delta \neq \mathbf{\Xi}$, and the Jacobian $\beta\mathbf{\Xi}A\mathbf{\Xi}_\delta^\top$ of Lemma 5 is not symmetric. At contamination 0.3 one of the 40 probes has a non-real eigenvalue, with imaginary part $4.6 \cdot 10^{-3}$. Lemma 1 then excludes every constant metric. A search over metrics also finds no constant symmetric G for equation (3.5). A squared-distance analogue with the same contaminated score and clean readout grows in asymmetry with the contamination in the same way. The negative control is contamination zero, where the analogue sits on the flat with $\mathbf{\Phi} = 2\mathbf{\Xi}$. The paper defines no energy of its own. Its row therefore records a mismatch between score and readout.

Modern Hopfield attention. Masumura & Taki (2025) discretise Model B of Krotov & Hopfield (2021) into the update of their equation (9),

$$\mathbf{x}_{n+1} = \alpha\mathbf{x}_n + (1 - \alpha)\,\mathrm{softmax}(\mathbf{h}_n)W_1^\top, \qquad \mathbf{h}_{n+1} = \alpha'\mathbf{h}_n + (1 - \alpha')\mathbf{x}_{n+1}W_2,$$

with $\alpha = \alpha' = 0.5$, one of their settings (p. 5). Their footnote 1 (p. 4) says that they do not tie W_1 and W_2, that untying them violates the energy argument of Krotov & Hopfield (2021), and that an energy for the untied model is open. On the joint state $(\mathbf{x}, \mathbf{h})$ the untied map has a non-real eigenvalue at every probe. By Lemma 1 and Proposition 2, no constant metric and no Fenchel–Young energy with a twice-differentiable conjugate make it admissible. The negative control ties $W_1 = W_2$, where the spectrum is real at every probe. The layer is applied once per block, with new weights in each block.

Kernel memory network. Iatropoulos et al. (2022) iterate the synchronous update of their equation (20) with the kernel of their equation (21),

$$\mathbf{s} \leftarrow XK^+K(X, \mathbf{s}), \qquad K(\mathbf{x}, \mathbf{y}) = \exp\big(-(\|\mathbf{x} - \mathbf{y}\|/r)^\beta\big).$$

Their Property 4 makes every pattern a fixed point. They use an energy landscape as a picture (p. 8) and write no energy for this network. The model therefore meets rule (iii) of the population only loosely, and we include it because it iterates its map by design. Its Jacobian is not symmetric. Their Properties 5.1 to 6 require a separation between the patterns (p. 8). Outside it, at r equal to the median distance between patterns, the Jacobian has non-real eigenvalues up to 0.21 at $\beta = 2$ and 0.18 at $\beta = 4$. No constant metric makes the map admissible there. Inside it, at $r = 0.4$ times the smallest distance, the Jacobian is close to rank one and its asymmetry at $G = \mathbf{I}$ is 1.41, which by itself excludes the Euclidean metric only. At $\beta = 2$, 3 of the 20 probes there also have a non-real eigenvalue, which excludes every constant metric. The negative control is the modern Hopfield update on the same patterns.

Universal Hopfield network, Manhattan similarity. The model is again equation (1) of Millidge et al. (2022), with the Manhattan row of Table 2 and the softmax and max rows of Table 3. It is the paper's own recommended model, used in its Figures 2, 4, 7 and 8. The Lyapunov argument of that paper (its Appendix A) is for the continuous-time dynamics of its equation (2), which already carry the gradient of the similarity. It concludes that the energy "always decreases over time, as long as the Hessian of the activation functions f and g are positive definite" (their Section 4). It does not cover

the single-step update of equation (1) that the released code runs and that this row iterates. The score gradient $\operatorname{sign}(\boldsymbol{\xi}_\mu - \mathbf{x})$ depends on $\mathbf{x}$. The score is therefore not affine, and it does not split. Under the softmax separation a non-real eigenvalue excludes every constant metric by Proposition 2. The fraction of probes with a non-real eigenvalue falls as β grows, because the softmax saturates and $\beta A \to 0$.

The max separation, the limit $\beta \to \infty$ of the softmax map, has a certificate that needs no Jacobian, because its ℓ_1 nearest-neighbour cells are not convex. One two-dimensional witness was verified by hand. The argument has three steps. *Step 1.* Suppose $G\,T \in \partial h$ with h convex, and take two points $\mathbf{a}, \mathbf{b}$ of the nearest-neighbour cell of $\boldsymbol{\xi}_\mu$. Then $G\boldsymbol{\xi}_\mu$ lies in $\partial h(\mathbf{a})$ and in $\partial h(\mathbf{b})$. *Step 2.* Both points therefore lie in the convex set $\partial h^*(G\boldsymbol{\xi}_\mu)$, and so does their midpoint. *Step 3.* Suppose the midpoint lies strictly inside the cell of another pattern $\boldsymbol{\xi}_\nu$. Then T is constant near it, and h is differentiable there with gradient $G\boldsymbol{\xi}_\nu$. Hence $G\boldsymbol{\xi}_\mu = G\boldsymbol{\xi}_\nu$, which distinct patterns rule out. The ℓ_1 witnesses are such pairs, 5 of 4000 random draws. The negative control is the Euclidean distance under the same max separation, whose Voronoi cells are convex. The gradient readout does not repair the row at any κ, because its score has a downward first-derivative jump (Proposition 7).

Universal Hopfield network, Euclidean distance under a non-shift-invariant separation. The same equation (1), with the Euclidean row of Table 2, is taken with the identity, square, tenth-power and logarithm rows of Table 3. The logarithm is applied to the distance itself, because the negated distance is negative. The cancellation that saves the softmax combination needs $A\mathbf{1} = 0$, which only a shift-invariant separation provides. Under the identity the $-\|\mathbf{x}\|^2$ term survives, and the Jacobian

$$J_T = 2\boldsymbol{\Xi}\boldsymbol{\Xi}^\top - 2(\boldsymbol{\Xi}\mathbf{1})\mathbf{x}^\top$$

has a rank-one term that breaks symmetry unless $\boldsymbol{\Xi}\mathbf{1} = 0$. The negative control is the softmax combination of the same distance.

Universal Hopfield network as released. Appendix E of Millidge et al. (2022) states that the reported results use the normalised reciprocal of each distance and normalise the separation outputs to sum to one. The dot product of their equation (13) is normalised internally. Every score vector is therefore divided by its own sum before the separation and again after it. Division by a query-dependent scalar does not cancel in a softmax. We ran the released code on a grid of its eight similarity functions by five of its separations (identity, softmax, square, tenth power and log). Every combination of that grid loses symmetry in the Euclidean metric, the reimplementation of the dot-product network included. In 18 of these 40 combinations a non-real eigenvalue also rules out every constant metric. Two defects must be kept apart. Switching the normalisation off restores the dot product only, and the Manhattan combination stays broken, because the reciprocal is a separate defect. Under Proposition 1 the row is repairable only on a bounded domain, because the reciprocal's curvature is unbounded at a stored pattern (Appendix E.7). The framework's Lyapunov proof (their Appendix A) needs a similarity that sums over coordinates, which covers at most three of the six similarities, the Euclidean and Manhattan distances and the dot product of their Table 2, and none in the normalised form that their Appendix E runs.

E.5 Rows with an energy outside Definition 1

Each row of this subsection has an energy that Definition 1 does not reach. The energy lives in a state-dependent metric, behind a convex post-transform, in a singular metric or on a continuous-time flow.

Lagrangian networks and the Energy Transformer. Equations (1) to (4) of Krotov & Hopfield (2021) give the dynamics, the energy, the Lagrangians and the descent. The descent holds when the Hessians of the Lagrangians are positive semidefinite. Since $\partial E/\partial g_i$ from their equation (2) is the right-hand side of their equation (1), the flow is gradient descent in the outputs. By the chain rule it is gradient flow in $\mathbf{v}$ under the Hessian of L_v, which is constant only when L_v is quadratic. In Hoover et al. (2023), equations (2), (3), (5) and (6) give the LayerNorm Lagrangian, the attention energy, the Hopfield energy and the update. The descent of their equations (7) and (8) again has a state-dependent metric, $M^A = \partial^2 L/\partial x\,\partial x$. Both models descend an energy by construction. Both sit outside the constant-metric requirement of Definition 1.

ℓ_2-normalised and layer-norm post-transforms. Equation (23) of Santos et al. (2025) applies an ℓ_2 normalisation or a layer norm after the update. Its energy is their energy (6) with the indicator Ψ of their Section 3.3.6, and their Proposition 2 derives the step for a differentiable Ψ. The post-transform is convex and the energy is proven, but Ψ plays the role of a state-dependent metric. A search over constant metrics at $d = 5$, $M = 8$ and $\beta = 1$ finds a solution space of dimension 0 for the ℓ_2 post-transform. For the layer norm it finds a solution space of dimension 1, spanned by a singular matrix. Neither admits a positive-definite constant G. The ℓ_2 model passes the certificate of Proposition 2 at every probe (Appendix D), which is the correct verdict for a model with a proven energy. The negative control is the softmax readout with no post-transform, for which the same search finds a positive-definite G.

Classical network with tanh at $M < d$. At $M < d$ the matrix $\Gamma = \Xi\Xi^\top$ of the tanh row is singular. Suppose that a symmetric H makes HJ_T symmetric at every $\mathbf{x}$. Then

$$HD(\mathbf{x})\Gamma = \Gamma D(\mathbf{x})H \quad \text{for every realised } D(\mathbf{x}).$$

For generic patterns the realised $D(\mathbf{x})$ span the diagonal matrices. That forces $H\mathbf{e}_i \parallel \Gamma\mathbf{e}_i$ for every i. Hence $H = \Gamma\,\mathrm{diag}(\boldsymbol{\alpha})$ has rank at most M, and it is not positive definite. A metric search at $d = 5$ and $M = 3$ finds only multiples of Γ. Special patterns can escape. At $\beta = 1$ the pattern $\boldsymbol{\xi} = (1, 1)$ gives $T = \nabla \log\cosh(x_1 + x_2)$ and $G = \mathbf{I}$.

Kernelised retrieval with fewer features than dimensions. Hu et al. (2024b) take the feature map $\Phi(\mathbf{x}) = W^\top\mathbf{x}$ with $W \in \mathbb{R}^{d\times D_\Phi}$ of full column rank (their Definition 2.6), so $D_\Phi \le d$. Their MNIST run has $D_\Phi = 200$ and $d = 784$ (their Table 3). The update of their equation (1.3) is the U-Hop map with $G = WW^\top$, which is singular when $D_\Phi < d$. A search at 20 queries over every symmetric H with HJ_T symmetric finds a one-dimensional solution space spanned by a singular matrix. No positive-definite metric exists. The negative control is $D_\Phi = d$, where the same search finds a positive-definite metric. Their equation (1.3) prints $+\,\mathrm{lse}$. With that sign corrected, the energy descends in the singular metric $WW^\top$, as in the tanh row at $M < d$.

NRGPT and Hyper-SET. Dehmamy et al. (2026) update

$$\mathbf{x} \leftarrow \mathbf{x} - \eta\,\partial E/\partial\mathbf{g},$$

where $\mathbf{g}$ is the layer-normalised state (their equation 6). They prove that the energy decreases after a transient for $\eta = c\,\mathrm{diag}(\gamma)$ (their Proposition 2.1). The gradient is taken in $\mathbf{g}$ and not in $\mathbf{x}$. The step therefore descends in the metric of the normalisation's Jacobian, which depends on the state. Their equation (13) notes that standard attention and the gradient of their attention energy share the form $M\mathbf{g}\,\mathrm{SM}(\mathbf{g}^\top J\mathbf{g})$ and differ only in the matrix M that reads out. Hu et al. (2026) descend an attention energy with one query-key-value matrix per head, through a projection onto spheres (their equations 3 to 9). This is again a state-dependent geometry. Both models sit outside Definition 1 with the Energy Transformer.

Conservative Krotov–Hopfield flow. Fanaskov & Oseledets (2025) note that the energy of the Krotov–Hopfield system is non-increasing on its trajectories only if $W = W^\top$ and a Hessian is positive semidefinite (their p. 3). They make the flow conservative with the right-hand side of their equation 5 and Proposition 5,

$$\dot{\mathbf{u}} = R(\mathbf{u})\big(g(W\mathbf{u} + \mathbf{b}) - \mathbf{u}\big).$$

One choice in their Example 6 is a Euclidean gradient flow for $W \neq W^\top$. The model runs in continuous time and defines no map. It has its energy by construction.

E.6 Undetermined rows, and models covered by a row

The first three rows of this subsection stay undetermined. The last paragraph names the models that a row covers.

Simplicial network. The continuous section of Burns & Fukai (2023) gives equations (5) to (7). Equation (6) is the convex-plus-concave energy of Ramsauer et al. (2021), and equation (7) carries the pattern readout Ξ. Whether (7) is the concave–convex step of (6) turns on what S_σ means

there. The paper defines it as a product of spins, and equation (7) multiplies it by a $P \times |\sigma|$ block, which that definition does not allow. Under the product-of-spins reading the score is a polynomial of degree $\dim(K) + 1$, and the step differs from (7) whenever a setwise simplex exists. Under the sub-vector reading, with D the diagonal matrix of vertex degrees, the step is $D\boldsymbol{\Xi}\mathbf{p}$ rather than $\boldsymbol{\Xi}\mathbf{p}$, and (7) descends a different energy with $G = D$. The paper's headline results use randomly diluted networks, where degrees vary. For them (7) is not the step of (6) under either reading. The complete pairwise skeleton with $D \propto \mathbf{I}$ is the positive control. No code is released, and admissibility stays undetermined. The paper's own empirical claims concern capacity and recall, not energy descent.

GSEMM and EDEN. GSEMM (Karuvally et al., 2023) and EDEN (Karuvally et al., 2025) add a slow variable to a Krotov–Hopfield network so that the fast state moves through a sequence. Both analyse their energy with the slow variable frozen, GSEMM as a control parameter (its p. 4) and EDEN under a slow timescale much longer than the fast one (its p. 6). Both are asymmetric by design.

Hyperbolic memory. Hao et al. (2026) define an energy on the Poincaré ball through the Minkowski inner product (their equation 5) and derive a Riemannian concave–convex step (their equations 6 to 10). The update of their Algorithm 1 is a self-map of the open ball, an open convex subset of $\mathbb{R}^d$, so Definition 1 applies to it. The authors prove descent in the state-dependent metric of the ball, and only locally (their p. 5). We have not tested a constant metric, and the row stays undetermined.

Models covered by an existing row. Memory Mosaics (Zhang et al., 2025) predict with a Gaussian kernel regression over pairs of keys and values (their equation 2). The in-context retrieval of Wu et al. (2025) computes (their equation 5)

$$V_t \operatorname{softmax}(K_t^\top \mathbf{q}).$$

MIRA (Agrawal et al., 2025) reads out adapter weights through the universal-Hopfield template (their equation 1), with a separation that is mainly affine and not softmax (their p. 4). Each scores with keys, reads out values and runs once. Each is therefore covered by the row of one-step attention with a value readout. For MIRA the row's mechanism carries over, and its softmax statistic does not.

E.7 One repair, four models, four classes

The proposition is a recipe as well as an existence statement, and it applies to every model of Table 1 whose score has a curvature bound. We ran it in the normalisation of Theorem 2, $E = -\operatorname{lse}(\mathbf{u})$ in the metric $G = 2\kappa\mathbf{I}$, with

$$\kappa = \max\left(\beta\rho/2,\ 10^{-3}\right).$$

Here ρ is the curvature modulus of the base where one exists. For a row without one, the Manhattan and reciprocal rows, ρ is the largest secant curvature measured on that row. Beyond this choice of ρ, nothing is tuned per model. We ran the recipe on four of the failure rows and on a squared-distance analogue of one of them, at $\beta = 1$ and $\beta = 30$ over five seeds. Table 11 gives the outcome, four classes rather than the two that the proposition alone suggests. Only one class is a repair, and only the analogue reaches it.

The squared-distance analogue of the nonparametric model has a global curvature bound, and it is the only row that the recipe repairs outright. The published nonparametric score is a dot product, and it lands in the class of the attention row. The released universal-Hopfield code replaces each distance by its reciprocal, whose curvature is unbounded at a stored pattern. That score is semi-convex at every fixed radius, with no global constant. A fixed κ therefore buys a bounded domain, and the map's own trajectory leaves that domain as β grows. At $\beta = 30$ its energy rises in 12% of steps and the descent inequality fails in 15%. Manhattan has no bound at any scale and is the named exclusion of Section 5. Two failure rows are not run. The adaptive network has its own construction in Section 5. The Euclidean distance under a non-shift-invariant separation fails in its separation. The recipe's softmax replaces that separation and returns the admissible softmax row of Table 1.

The attention row shows the main limit of the recipe. Attention that scores with the keys and reads out the values has a linear score, so every hypothesis of Proposition 1 holds and the repaired map

Table 11: One repair recipe (Proposition 1) on four failure rows of Table 1 and on a squared-distance analogue of the nonparametric row, $\beta \in \{1, 30\}$, five seeds. Energy rises and descent violated are fractions of steps under the repaired map. Manhattan's two fractions are at $\beta = 30$, and at $\beta = 1$ they are 13% and 48%. Not decided marks a probe that did not decide the row. The class column reads each row in the recipe's normalisation, $E = -\operatorname{lse}(\mathbf{u})$ in the metric $G = 2\kappa\mathbf{I}$.

Model	Curvature bound	$\operatorname{asym}(J)$ after repair	Energy rises	Descent violated	Iterate	Class
Nonparametric, squared-distance analogue	global, $\rho = 2$	10^{-16} (from 0.5–1.3)	0	0	bounded at both β	repaired
Nonparametric as published (Hu et al., 2025)	global, $\rho = 0$	symmetric, with a real non-negative spectrum	0	0	grows $1.4 \cdot 10^6$-fold in 60 steps, $E \to -8.0 \cdot 10^7$	admissible, nothing to retrieve
Universal Hopfield, reciprocal with an offset	at every fixed radius, none global	$2 \cdot 10^{-19}$ at $\beta = 1$, $4 \cdot 10^{-18}$ at $\beta = 30$	0 at $\beta = 1$, 12% at $\beta = 30$	0 at $\beta = 1$, 15% at $\beta = 30$	leaves the bounded domain as β grows	repaired on a bounded domain only
Universal Hopfield, Manhattan	none at any scale	$4 \cdot 10^{-18}$, $\lambda_{\min} = 1.0$, no imaginary part	22%	73%	not decided	excluded (Section 5)
One-step attention, value readout (Vaswani et al., 2017)	global, $\rho = 0$	symmetric, with a real non-negative spectrum	0	0	grows $9 \cdot 10^5$-fold in 60 steps, $E \to -2.7 \cdot 10^7$	admissible, nothing to retrieve

is admissible. The retrieved index agrees with the score's argmax in every probe at $\beta = 30$ and in 81% of probes at $\beta = 1$. The iterate still diverges, because $E = -\operatorname{lse}(\mathbf{u})$ is unbounded below for a linear score, which is the dot-product base under the uniform gate (Proposition 5). The divergence occurs at every κ. The rate in the table is that of the floor $\kappa = 10^{-3}$ that $\rho = 0$ selects. The divergence belongs to this energy. Proposition 1 does not cause it, since at $\omega = 0$ and $G = \mathbf{I}$ the same score gives the modern Hopfield update on the keys, which retrieves. Admissibility does not imply retrieval. Every statistic in this row passes a map that retrieves nothing, and only the norm of the iterate tells the two maps apart.

Separating the four classes needs three probes, and none of them is a Hessian. Shrinking a secant at a fixed location detects a first-derivative jump. Moving the probe toward a pattern at a fixed secant length detects a singularity. Tracking $\|\mathbf{x}\|$ detects an unbounded energy. Manhattan's repaired map passes every almost-everywhere Jacobian statistic in this paper. The one score that the theory excludes is therefore caught only by measured energy ascent and by the descent inequality.

F Related work

Integrability criteria. The integrability criterion has several antecedents. Santos et al. (2025) state the constraint and the repair in one sentence, and Theorem 1 gives its exact form. Im et al. (2016) prove the same symmetry condition for untied auto-encoders and measure it by a Jacobian-symmetry ratio and by the curl. Cardin et al. (2025), Smart et al. (2025) and Dehmamy et al. (2026) state it for associative memory, for a softmax attention layer and for an attention energy, the last locating the difference in the matrix that reads out. Fanaskov & Oseledets (2025) make the Krotov–Hopfield flow conservative by replacing its right-hand side with the gradient of an energy, the move of Proposition 1 for that model. The gradient of an energy with a free similarity is the update template of Krotov et al. (2025, Eqs. 5.7 and 5.9). Lemma 1 is standard, and Tomihari & Karakida (2025) read complex Jacobian eigenvalues of attention dynamics as oscillation. What is new here is a certificate that survives the Fenchel–Young post-transform.

Sorted sums and the rearrangement inequality. Our energy rests on a classical fact. A sorted sum with non-increasing weights is a maximum over permutations (Hardy et al., 1952, Theorem 368), and it is therefore convex (Boyd & Vandenberghe, 2004, Section 3.2.3). The same fact underlies ordered weighted averaging with non-increasing weights (Yager, 1988), the sorted-ℓ_1 penalty (Bogdan et al., 2015) and the decreasing weighted sorted ℓ_1 norm (Zeng & Figueiredo,

2014). We claim only the use of this fact for an energy. What is new is an exact energy for a learned order-statistic similarity at every parameter value.

Equilibrium and energy-based models. Deep equilibrium models (Bai et al., 2019) train a layer at its fixed point. Monotone operator equilibrium networks (Winston & Kolter, 2020) parameterise that layer so that the fixed point exists and is unique at every parameter value. Our monotone gate plays that role for an energy. It guarantees descent, and it does not guarantee a unique fixed point. Sander et al. (2022) obtain a Wasserstein gradient flow for attention by making it doubly stochastic under a symmetric kernel.

G SYNTHETIC EXPERIMENTS

G.1 BASELINES

Section 6 states that every baseline is our own implementation of the published equations. Three of them are checked against the released code to machine precision: the modern Hopfield network, the sparse separations and the adaptive network's model class. The adaptive network's released experiment scripts import two modules that neither its public repository nor its anonymous mirror contains. We therefore could not run those scripts, and we could not reproduce its published numbers. Its rows in our tables are our implementation of its equations. We compare architectures on our tasks. We do not compare against the numbers that its paper reports.

G.2 DESCENT AND THE TWO READOUTS

Theorems 2 and 3 predict a symmetric Jacobian with a real, non-negative spectrum for the gradient readout. No theorem covers this score under the pattern readout, and we expect neither property there.

Energy along trajectories. Figure 4 shows the 27 regimes that Section 6 counts, and it adds a second gate. With the adaptive network's own trained signed gate, the energy rises under the pattern readout in 13 of the 27 regimes, against 24 of 27 with our monotone gate. Over five seeds the gradient readout raises the energy in none of 12 000 transitions and violates the descent inequality in none of 1800, exhaustive counts along trajectories and not independent samples. The same map at half the required κ violates the descent inequality (Table 5).

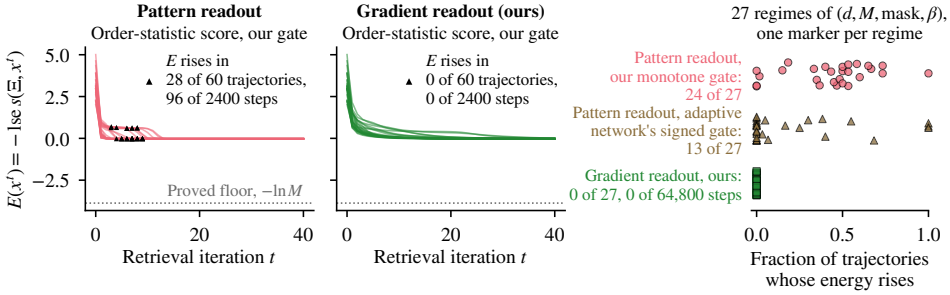

Figure 4: The energy $E = -\operatorname{lse}(\mathbf{u})$ stated by Wang et al. (2026), measured along each map. The left two panels trace one regime near the median of the 27, the adaptive network's order-statistic score with our trained monotone gate under both readouts. Each draws 26 of its 60 trajectories, and its legend counts all 60. The right panel summarises 27 regimes over $(d, M, \text{mask}, \beta)$, one seed each. Under the pattern readout the energy rises in 13 of 27 regimes with the adaptive network's own trained signed gate and in 24 of 27 with our monotone gate. The signed gate here is the unconstrained parameterisation trained with Adam on the squared-distance base at a fixed β, a simpler trainer than the one of Table 13. Under the gradient readout the energy rises in 0 of 27 regimes and in 0 of 64 800 steps.

Spectrum. Figure 5 draws the spectrum of J_T for the three maps of Figure 1. Eigenvalues off the real axis appear under the pattern readout of the order-statistic score. The gradient readout's spectrum is real and non-negative, as Theorem 3(i) states.

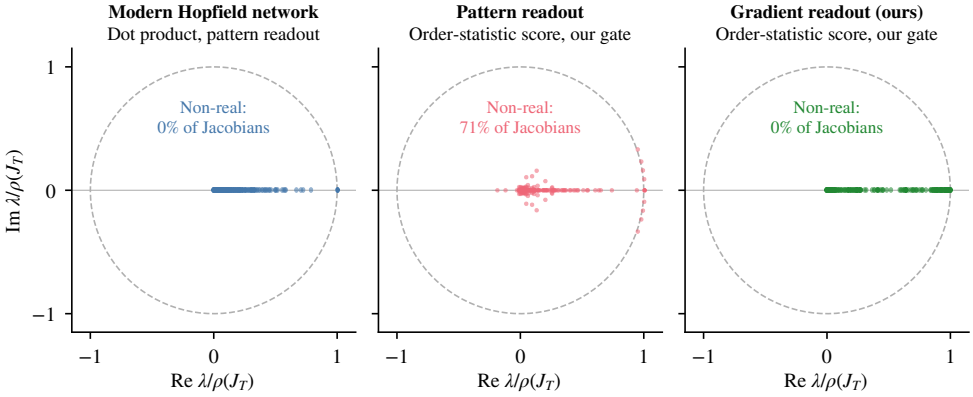

Figure 5: Spectrum of the retrieval Jacobian J_T in the complex plane, normalised by the spectral radius, for 120 masked queries in the configuration of Figure 1. The dashed circle is the unit circle. The middle and right panels share the score exactly and differ only in the readout. Off-axis mass appears in one panel. By Lemma 1 and Proposition 2 it excludes every constant metric, and every Fenchel–Young energy whose conjugate is twice differentiable on the image of ∇h near that point. The percentages are counts in this configuration. The five-seed range of Table 10 is a separate run at $M = 64$. The gradient readout's spectrum is real and non-negative, as Theorem 3(i) states.

Circulation. The closed line integral $\oint T \cdot d\ell$ needs no Jacobian. On the larger loop of Figure 6, from $1\times$ to $256\times$ quadrature, its magnitude falls from $6.0 \cdot 10^{-4}$ to $1.4 \cdot 10^{-6}$ for the gradient readout. For the pattern readout it stays at 0.49, and it is 0.4869 at every level. The fall is roughly first order and not monotone, because the gradient readout jumps where two coordinates swap rank.

Basins. The basins of attraction do not show the difference between the readouts. On a two-dimensional slice through six stored patterns, the two readouts send 99.6% of the slice to the same memory (Figure 7). Over the same slice the pattern readout's Jacobian asymmetry is of order one wherever the map still moves the query. Figure 6 maps this asymmetry for the three panels of Figure 1.

G.3 Basins, spurious points and capacity

At its initial gate and $\beta = 10$, no query's endpoint under the constrained map lands within 0.25 of its own pattern on random patterns in $d = 64$. Theorem 2 gives descent and a lower bound. It does not say where the minimisers sit.

Protocol. A plan fixed before its code was written measures where the minimisers sit, on random unit patterns in $d = 64$, with $M \in \{32, 64, 128, 256, 512\}$, $\beta \in \{1, 3, 10, 30, 100\}$ and five pattern draws. The draw is the unit. The study is exploratory and tests nothing. Retrieval rates below are means over the five draws. A query is a stored pattern with Gaussian noise of standard deviation 0.04 and with 10, 30 or 50% of its coordinates replaced by $N(0, 1/d)$ noise, as in the synthetic experiment of Appendix G.4. A second set of 200 queries starts uniformly on the sphere. A query is retrieved when its final state lies within 0.25 of its own pattern, about a third of the typical half-distance between two patterns. Our map runs at the gate of Corollary 1 at $\boldsymbol{\theta} = 0$. Two controls separate the gate from the readout (Figure 8). The first is our map at the uniform gate, which is exactly the modern Hopfield network at 2β. We added this control after most of the grid had been read. The second is the same gate under the pattern readout, which differs from our map only in the readout.

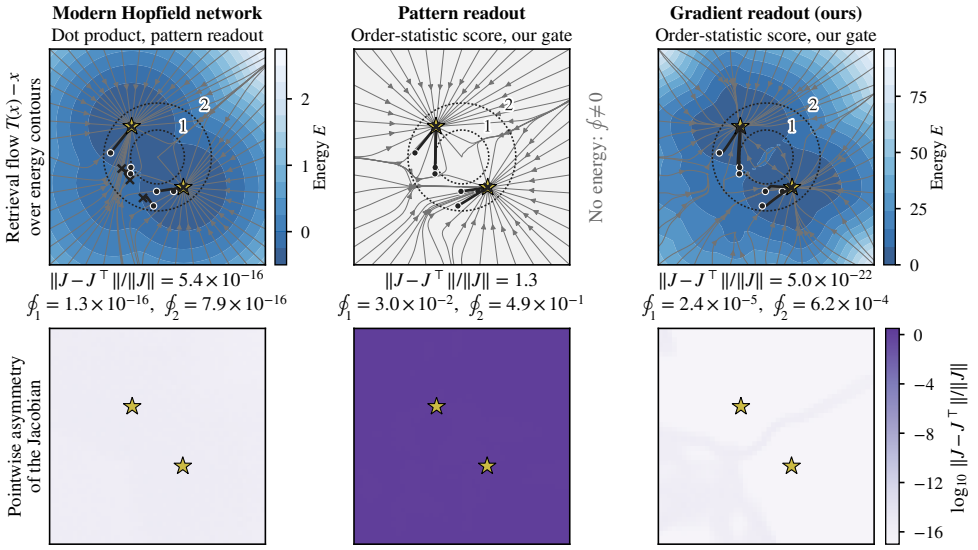

Figure 6: Figure 1 with its second row. The top row is Figure 1 itself. The bottom row maps the asymmetry pointwise on a $\log_{10}$ scale, from machine precision (pale) to order one (dark). It is the local obstruction to the flow being a gradient, and it is dark across the whole slice for the pattern readout alone. The circles of $\oint_1$ and $\oint_2$ are centred midway between the two stars, with radii 0.34 and 0.68 times the distance between them. The larger one is the loop whose circulation Appendix G.2 refines. Three of the modern Hopfield network's six queries converge to other stored patterns, which lie off this slice, so their end points (crosses) project inside it.

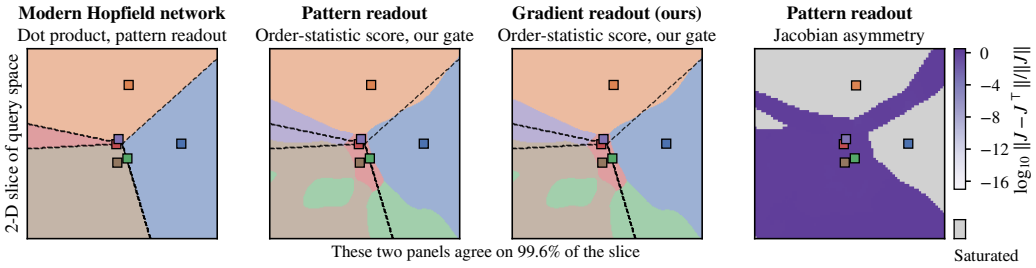

Figure 7: The two readouts reach the same memories, and only the Jacobian asymmetry shows that one of them has no energy. The slice is two-dimensional and passes through six stored patterns at $d = 32$. The first three panels colour each query by the memory it reaches after 400 applications of the map, one flat colour per memory and white for a metastable end point, which no query reaches. Dashed lines are the Voronoi boundaries and squares the patterns. The middle and right panels share the score, with our trained monotone gate, and differ only in the readout. They agree on 99.6% of the slice. The last panel is $\log_{10} \|J - J^\top\|/\|J\|$ of the pattern readout over the same slice. Grey marks the 42% of the slice where the softmax has saturated and $\|J\| < 10^{-10}$. On the rest, where the map still moves the query, the asymmetry is at least 0.83 (median 1.19).

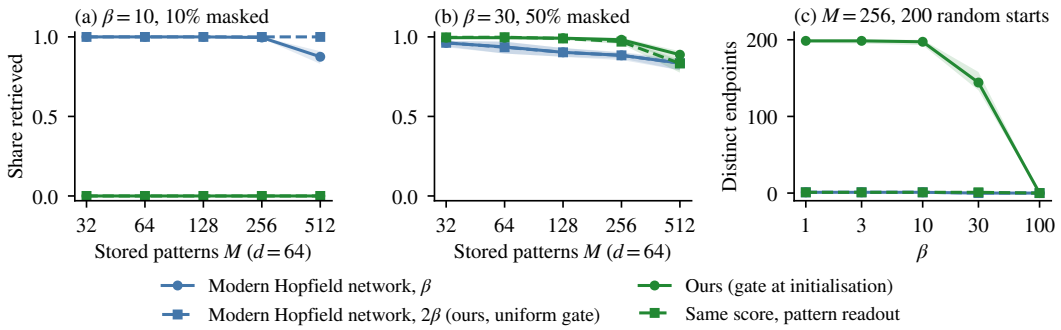

Figure 8: Retrieval against load, and the spurious endpoints, on random unit patterns in $d = 64$. Lines are means over five pattern draws and bands their range. (a) $\beta = 10$ with 10% of the coordinates masked. The gate at initialisation retrieves no query under either readout. The uniform gate, the modern Hopfield network at 2β, retrieves every query, and the modern Hopfield network at β falls below 0.95 only at $M = 512$. (b) $\beta = 30$ with half of the coordinates masked. Our map retrieves more queries than the modern Hopfield network at β and at 2β at every load, and its pattern-readout twin does so at every load but the largest. (c) Distinct endpoints away from every pattern, out of 200 random starts at $M = 256$, against β. The modern Hopfield network at either temperature and the twin end at one such point or none, and our map at up to 200.

Retrieval. Every map has a threshold in β. At $\beta \leq 3$ no map retrieves more than 0.015 of the queries. The gate moves the threshold from between 3 and 10 to between 10 and 30. At $\beta = 10$ our map and its pattern-readout twin retrieve no query at any load. At the same temperature and at every load, the uniform gate retrieves every query with a tenth masked, and on average over the draws at least 99.5% of the queries with 30% masked. At $\beta = 30$ and 100 our map retrieves at least 95% of the queries with a tenth masked at every load up to $M = 512$. At these temperatures the gate helps under heavy masking. With half of the coordinates masked, our map retrieves more queries than the uniform gate at every load, by 0.032 to 0.155. It retrieves at least 95% of them up to $M = 256$ at $\beta = 30$ and up to $M = 512$ at $\beta = 100$. With half masked the modern Hopfield network stops at $M = 32$, at β and at 2β. On this grid the readout changes none of these capacities. The classical network stores 4 of its own ± 1 patterns in the median draw, by exact recovery and the same 95% rule. Its queries replace the same shares of coordinates by random signs, and its loads are $M \in \{2, 4, 8, 16, 32, 64\}$.

Spurious points. Below its threshold our map's 200 random starts end at 134 to 200 distinct points, the median over draws. Every such point lies within 0.17 of the patterns' mean. The modern Hopfield network and the pattern-readout twin end at one such point or none. A probe added after the grid had been read examines 100 of our points at each of three settings, $(M, \beta) = (128, 3)$, $(128, 10)$ and $(256, 30)$. There the almost-everywhere Jacobian has spectral radius 0.53 to 0.66 in the median and at most 0.68. After a perturbation of norm 10^{-4} or 10^{-2} the iteration ends within $4.2 \cdot 10^{-3}$ of the point, at an energy within $1.5 \cdot 10^{-5}$ of its own. The random starts therefore end in a dense cluster of points at nearly equal energy. We do not claim that each point attracts within its sorting cell, because one further step can cross into a neighbouring cell. The stability test that the plan fixed in advance asks a perturbed run to return within 10^{-6}, and none of our spurious points passes it. At $\beta = 30$ the cluster appears in two of the five draws at $M = 2d$ and in every draw from $M = 4d$ on. It takes 0.08, 0.78 and 0.98 of the random starts at $M = 128$, 256 and 512. Its points lie 0.6 to 1.9 above the lowest energy among the fixed points reached from the patterns, the median over the points at each load. Every query with at most 30% masked is still retrieved there. At $\beta = 100$ no random start ends away from a pattern. These statements concern the gate at initialisation on this grid. A gate trained with the recipe of Table 13 at mask 0.6 has 27 to 29 of its 64 entries below 0.9. Its temperature rises to between 125.5 and 155.4. At that temperature no random start ends away from a pattern at any M up to 512. The modern Hopfield network at the same temperature does the same. At $\beta = 30$ the trained gate has the cluster at every load. Its absence therefore depends on the trained temperature.

G.4 RETRIEVAL AGAINST THE INFERENCE BUDGET

Protocol. Patterns are Gaussian, normalised to unit norm, in $d = 64$, $M = 128$. A query starts from a stored pattern drawn uniformly at random and adds Gaussian noise of standard deviation 0.04 to every coordinate. It then replaces $\mathrm{round}(d \cdot \mathrm{mask})$ coordinates, drawn without replacement, by $N(0, 1/d)$ noise. A replaced coordinate therefore carries neither the pattern nor the additive noise. Both learned models are trained with AdamW at learning rate 0.1 for 150 steps. At every step each model receives 256 fresh queries from an identically seeded generator, so the two models see the same query sequence. Table 12 lists what differs between them. The modern Hopfield network receives the best $\beta \in \{10, 30, 100, 300, 1000\}$ on a separate validation draw of 400 queries from the same patterns, each β run for 30 applications. Accuracy is the fraction of 400 test queries whose retrieval lands nearest the pattern it was generated from, after the stated number of applications of the map. Standard deviations over the five seeds are sample standard deviations, with denominator $n - 1$.

Table 12: Model configuration in the synthetic retrieval experiment. The adaptive network follows its released code.

	Adaptive network	Ours
Bases	squared distance and dot product	squared distance
Gate	block-constant, stride 4, signed weights	full d-dimensional, Corollary 1, floor $f = 0.05$
Parameters	34	65
β	initialised at 1, learned	$\log \beta$ learned from $\beta = 30$
Loss	Brier	negative log-likelihood
Weight decay	AdamW default	none
κ	—	$\beta \|\mathbf{c}\|_\infty$
Readout	$\boldsymbol{\Xi}\mathbf{p}$	gradient readout, Theorem 2

Table 13: Retrieval accuracy on synthetic patterns, $d = 64$, $M = 128$, masked queries, five seeds (mean $\pm$ one standard deviation, sd, with $n = 5$ seeds and 400 test queries per seed). Both learned models are trained with AdamW on the same query stream. They also differ in bases, gate, temperature, loss and weight decay (Table 12), so the table compares the two models as a whole. The modern Hopfield network (MHN) receives the best β of a validation sweep and is applied 30 times (Appendix G.4, paragraph "Protocol"). "@k" is the number of applications of the map at inference. A query counts as retrieved when the stored pattern nearest to the output is its own, and "no retrieval" applies that rule to the query itself. The last row keeps our trained gate and temperature and reads out with $\boldsymbol{\Xi}\mathbf{p}$, without retraining. It is exploratory, not registered, and has no energy guarantee.

	Mask 0.5	Mask 0.6	Mask 0.7
No retrieval	0.912 ± 0.009	0.741 ± 0.028	0.443 ± 0.017
MHN (best β) @30	0.911 ± 0.008	0.741 ± 0.028	0.444 ± 0.016
Adaptive @1	1.000 ± 0.001	0.992 ± 0.004	0.874 ± 0.021
Adaptive @30	1.000 ± 0.001	0.992 ± 0.004	0.873 ± 0.023
Ours @1	0.971 ± 0.008	0.874 ± 0.015	0.592 ± 0.030
Ours @30	1.000 ± 0.000	0.992 ± 0.004	0.850 ± 0.027
Ours @300	1.000 ± 0.000	0.992 ± 0.004	0.850 ± 0.027
Ours, pattern readout @1	1.000 ± 0.000	0.993 ± 0.004	0.857 ± 0.029

Results. Section 7 states the one-step and thirty-step gaps of Table 13. The table shows three more things. On every seed the constrained model's accuracy after 30 steps equals its accuracy after 300. Iterating does not change the adaptive network's accuracy, since its one-step and thirty-step rows agree to within 0.001. The modern Hopfield network scores within 0.001 of no retrieval at every mask, because on unit-norm patterns its dot product ranks the patterns as the distance to the query does. The learned scores gain by ranking the patterns better. Appendix G.5 repeats the comparison on MNIST.

The readout swap. The mechanism we expect is the limit in Theorem 3(iii). Near a stored pattern the trained gradient readout contracts by approximately the factor $1 - c_d/\|\mathbf{c}\|_\infty$ per step. The factor holds to within one per cent on every map we measured, the maps at the floor $f = 0.05$ that we use

included (paragraph "Floor" below). The contraction is slow for the steep gates that training selects, while $\boldsymbol{\Xi}\mathbf{p}$ can jump to a pattern in one application. At a uniform gate the two readouts coincide. The cost therefore needs a steep gate. The last row of Table 13 tests this mechanism by reading out our trained score with $\boldsymbol{\Xi}\mathbf{p}$. In an exploratory swap without retraining, our trained gate under the pattern readout reaches at one step at least the accuracy that the gradient readout reaches after 300, at every mask and on every seed. At mask 0.7 that row still trails the adaptive network at one step by 1.7 ± 1.5 points. This is a descriptive difference at five seeds between models that also differ in bases, gate, temperature, loss and weight decay (Table 12). That row and the adaptive network both converge on every test query within 1000 applications, so the cycling of Section 6 does not arise on this task.

Gate initialisation. The adaptive network's gate starts at $\mathbf{w} = 0$, as in its released code, which makes $\mathbf{c} \equiv 0$ at step zero. Ours starts at the ramp that $\mathrm{softmax}(\boldsymbol{\theta} = 0)$ gives. On MNIST and in the multiple-instance experiments the adaptive network starts at the same ramp instead, which departs from its released code. No run isolates this departure. The released configuration of Table 3 restores it together with three other departures.

Losses and weight decay. The adaptive network is trained on the Brier loss, the default of its released code, and ours on the negative log-likelihood. Both losses are strictly proper and have the same Bayes decision. The adaptive network's weight decay is the AdamW default on every parameter, as in its released code, and ours has none. Retraining our gate at that decay moves one-step accuracy by at most $+0.0015$, and removing the decay from the adaptive network's temperatures leaves its one-step accuracy unchanged to four decimals. Its gate was not retrained undecayed on this task; on the multiple-instance tasks that control is worth about one point to it.

Floor. We fixed the floor $f = 0.05$ before any run. It is a round value well clear of the double-precision underflow that $f = 0$ permits, and it keeps the coercivity clause of Theorem 2 valid in floating point. We did not choose it by accuracy. The floor also sets the cost, since a smaller c_d slows the last coordinate. Every iteration count we report holds at $f = 0.05$. Table 14 sweeps the floor at mask 0.6 and 300 steps. Accuracy falls slowly with f. The completion distance to the true pattern is an order of magnitude worse at $f = 0$ than at any $f \leq 0.10$, and about four times worse than at $f = 0.25$. At $f = 0$ the trained gate's last entry c_d is small but not zero, between $1.3 \cdot 10^{-3}$ and $2.7 \cdot 10^{-3}$ over the five seeds. The parameterisation of Corollary 1 gives $c_1 = 1$ at every f. Near a stored pattern the trained map then shrinks an error on one coordinate by the factor $1 - c_d$ per step, the limit of Theorem 3(iii). We measured this on the ten trained maps at $f = 0$ and $f = 0.02$ over 300 steps, and on the five at $f = 0.05$ over 30 steps. It holds to within one per cent on each. At $f = 0.05$, 300 steps would shrink the error below the offset of the fixed point from the pattern, where the ratio no longer measures the contraction. At $f = 0$, 300 steps leave 44 to 68 per cent of such an error, and at $f = 0.02$ they leave about 0.2 per cent. The energy is constant along a coordinate only at $c_d = 0$ exactly, which double precision can reach by underflow when $f = 0$.

Table 14: Floor sweep, $d = 64$, $M = 128$, mask 0.6, 300 steps, five seeds. "Distance" is the completion distance to the true pattern. Bold marks the floor we use, $f = 0.05$. The sweep draws its training stream from a different generator than Table 13, so it is a separate run of the same model.

Floor f	0	0.02	**0.05**	0.10	0.15	0.25
Accuracy	0.995 ± 0.003	0.995 ± 0.003	0.992 ± 0.005	0.980 ± 0.006	0.969 ± 0.003	0.943 ± 0.006
Distance	0.301	0.008	0.011	0.027	0.042	0.077

G.5 MNIST

Table 15 gives the setting and the results.

At one step the constrained model is within one standard deviation of the adaptive network at masks 0.7 and 0.9, with 0.955 ± 0.021 against 0.945 ± 0.049 at mask 0.7, and it trails at mask 0.8. After 150 steps, against the adaptive network at one step, the constrained model trails by 3.4 points at mask 0.8. At mask 0.9 every model is below 0.23, and every gap between ours and the adaptive network is inside one standard deviation. The adaptive network's mask-0.7 accuracy has the largest

Table 15: MNIST retrieval, $d = 784$, $M = 256$, five seeds, matched AdamW (learning rate 0.1, 120 steps, fresh draws), 192 test queries with additive noise of standard deviation 0.02. Standard deviations are sample standard deviations, with denominator $n - 1$. Here the adaptive network receives a full 784-dimensional gate and the squared-distance base only. Both learned models are trained on the negative log-likelihood and learn $\log \beta$ from $\beta = 300$. Weight decay 0.01 is applied to the adaptive network's gate only. The modern Hopfield network is applied 30 times. The constrained model trails at masks 0.8 and 0.9 at its own budget, at 0.9 within one standard deviation.

Mask	MHN (best β)	Adaptive @1	Adaptive @30	Ours @1	Ours @150
0.7	0.915 ± 0.015	0.945 ± 0.049	0.945 ± 0.049	0.955 ± 0.021	0.972 ± 0.014
0.8	0.660 ± 0.047	0.798 ± 0.034	0.798 ± 0.034	0.750 ± 0.029	0.764 ± 0.033
0.9	0.180 ± 0.017	0.217 ± 0.031	0.215 ± 0.030	0.214 ± 0.023	0.202 ± 0.023

dispersion in the table, from 0.885 to 1.000 across seeds. Its trained gate has a negative entry in 11 of 15 configurations (Appendix E.4). Our gate lies in $[0.05, 1]$ by construction, and its measured range is $[0.051, 1.000]$. The completion distance to the true pattern, for the constrained model at 150 steps against the adaptive network at one step, is 0.018 against 0.046 at mask 0.7, 0.199 against 0.167 at mask 0.8 and 0.759 against 0.672 at mask 0.9. At one step the constrained model's distance is above 1.1 at every mask, because one bounded step does not reach a pattern.

H Multiple-instance learning

H.1 Protocol

Design. We use the retrieval map as the pooling layer in bag classification, the setting in which this literature deploys Hopfield layers (Ramsauer et al., 2021; Hu et al., 2023; Santos et al., 2024; Wang et al., 2026). The datasets are the four of the adaptive network's benchmark, which Ramsauer et al. (2021) and Hu et al. (2023) share. They are Tiger, Fox and Elephant (Andrews et al., 2003) and UCSB (Gelasca et al., 2008). MUSK1 (Dietterich et al., 1997), the probe of our first registration, is in none of these benchmarks. Appendix H.7 reports it, and our costs are largest there. Every experiment of this section follows a written protocol, which we call a registration. The first registration specifies the statistics and their decision rules.

In Table 17 the adaptive network uses its squared-distance base only, which shares our score exactly. The first registration declares this deviation from the adaptive network's released code. A later registration restores the dot-product base (Table 22), and another runs the whole released configuration (Table 3). The released code runs the map for one step. Theorem 2 of Wang et al. (2026) states that its energy decreases under iteration for isotropic noisy and biased variants, and its proof uses a score without the footprint (Appendix E). The footprint is their name for the sums of the k largest per-coordinate similarities, for $k = 1, \ldots, d$. The protocol uses stratified five-fold cross-validation and matched AdamW, with everything identical except T. The two maps still differ in the sign of the gate as well as in the readout. Every fit that cycles has a negative gate entry (Appendix H.3), and for a non-negative gate Proposition 3(iv) rules out a mutual pair, a property of the score alone. Table 17 therefore compares the two models as a whole, and a later registration separates the gate from the readout (Appendix H.4). The registrations behind these tables run five seeds, and the twenty-seed registration adds fifteen (Appendix H.2).

Datasets. Table 16 lists the five datasets. Fox, Tiger and Elephant are in the form released by Santos et al. (2024). UCSB Breast Cancer is the instance matrix released with Hu et al. (2023).

Model and training. Every pooling layer is the bare retrieval map, with one head and no instance encoder, projection or layer normalisation, unlike the Hopfield pooling layer of Ramsauer et al. (2021). The pooled vector feeds a linear classifier trained with binary cross-entropy and AdamW for 200 epochs. The learning rate is 0.1 for the gate and the temperature and 0.01 with no weight decay for the classifier and the query, the same for every model, and the map is unrolled $K = 5$ steps at training. The frequency of the cycling depends on this rate (Appendix H.5). Every model learns $\log \beta$ from $\beta = 30$, where the adaptive network's released code starts a raw β at 1. The adaptive

Table 16: The multiple-instance datasets. "Instances" counts all instances over all bags, and d is the instance dimension.

Dataset	Bags	Positive bags	Instances	d
Fox	200	100	1320	230
Tiger	200	100	1220	230
Elephant	200	100	1391	230
MUSK1	92	47	476	166
UCSB Breast Cancer	58	26	2002	708

network's gate has one entry per coordinate here, where its released code uses a block-constant gate of stride 4.

Weight decay. Weight decay is 0.01 on the gate and the temperature of every model, ours and the modern Hopfield network included. We call this rule the *release protocol*. It applies the adaptive network's released code symmetrically, since that code passes every parameter to AdamW at the default decay and contains no multiple-instance code of its own. The release protocol departs from the first registration, which decayed the adaptive network's gate alone (the *registered protocol*). The departure was decided after the adaptive network had been run under both rules and before our model or the modern Hopfield network had been trained under the release protocol. The registered protocol was also run in full. Every verdict is the same under it, and the adaptive network's depth drift is smaller, 0.0134 against 0.0248 (five-seed means over MUSK1, Fox, Tiger and Elephant).

The rule is symmetric in form but not in effect. Decay pulls the gate parameter towards zero. Zero is a working decreasing ramp under our parameterisation and a uniform map under the adaptive network's, so the shared rule costs the adaptive network more than it costs us. A third protocol removes the decay from every Hopfield parameter of every model, which removes the asymmetry, and every rule fixed in advance returns the same verdict on every dataset. Without the decay the adaptive network gains about one point of one-step accuracy, a mean of $+0.012$ against -0.001 for ours. It cycles in 39 to 63 per cent of bags rather than 43 to 56 (both at five seeds), while ours converges in at least 99.8 per cent of them with no cycles. Its depth drift is worse without the decay, 0.0320 against 0.0248. On Tiger its trained gate reaches -107 rather than -80 (a mean over fits of the smallest entry), because the decay was the only pressure holding an unconstrained gate near zero. In the mean over these 20 seed-dataset pairs, the adaptive network drifts most and the modern Hopfield network least under each of the three protocols.

Evaluation and the dynamical statistics. Evaluation runs the map to $K' = 1000$. Write $\mathbf{x}_t$ for the state of a held-out bag after t applications of T. The bag has *converged*, or is in a *period-2 cycle*, when

$$\text{converged:} \quad \min_{0 \le t < 1000} \|\mathbf{x}_{t+1} - \mathbf{x}_t\| < 10^{-8},$$

$$\text{period-2 cycle:} \quad \text{not converged}, \quad \|\mathbf{x}_{1000} - \mathbf{x}_{998}\| < 10^{-8} \quad \text{and} \quad \|\mathbf{x}_{1000} - \mathbf{x}_{999}\| > 10^{-6}.$$

The remaining bags are in neither class. On every dataset they average at most 0.35 per cent for the adaptive network and 0.05 per cent for ours. "Converged" and "cycles" are the fractions of a seed's held-out bags in each class. A bag has *diverged* when its state stops being finite or leaves the ball of radius 10^6, and a diverged bag counts as wrong at every later depth. "Fits that cycle" counts the fits with at least one cycling bag, out of 100 per dataset. "Iterations" is the number of applications needed to reach a residual $\|T(\mathbf{x}) - \mathbf{x}\|$ of 10^{-6}, a median over the bags of a seed that reach it. The condition number of $\mathbf{I} - J_T$ is taken at the final iterate, a median over all bags of a seed. Both medians are averaged over seeds. Accuracy at depth K', written $\mathrm{acc}(K')$, is the held-out accuracy of a seed after K' applications, with the five folds pooled by bag count. The *depth drift* of a seed is

$$\mathrm{acc}(5) - \min_{K' \in \{5,\,10,\,30,\,100,\,300,\,1000\}} \mathrm{acc}(K'),$$

the accuracy lost by iterating past the training depth, and tables give its mean over seeds. Every K' after 5 in this grid is even. For a map in a period-2 orbit the statistic is therefore the drop, if any, from the orbit's odd phase to its even phase.

The gradient statistics. Both gradients are taken with respect to the Hopfield parameters $\boldsymbol{\theta}$ of the bag's binary cross-entropy L, meaning the gate parameters and $\log\beta$, or $\log\beta$ alone for the modern Hopfield network. The unrolled gradient g_{unr} backpropagates through all 1000 applications of the evaluation. The implicit gradient is formed at the 1000th iterate $\mathbf{x}^*$,

$$g_{\mathrm{imp}} = \Big(\frac{\partial T}{\partial \boldsymbol{\theta}}(\mathbf{x}^*)\Big)^{\top}\mathbf{v}, \qquad \big(\mathbf{I} - J_T(\mathbf{x}^*)\big)^{\top}\mathbf{v} = \frac{\partial L}{\partial \mathbf{x}^*}.$$

A bag has a *bad gradient* when $\|g_{\mathrm{imp}} - g_{\mathrm{unr}}\| > 0.1\,\|g_{\mathrm{unr}}\|$, and a bag whose implicit gradient is not finite counts as a disagreement. The error is relative, so a bag whose unrolled gradient is small can inflate it. The registered statistic therefore keeps only the bags with $\|g_{\mathrm{unr}}\| > 10^{-12}$. "Bad grad." in Table 17 is the fraction of kept bags with a bad gradient, pooled over the twenty seeds, and "bags kept" counts the kept bags. The tables of Appendix H.6 report the unrestricted fraction over all bags, averaged over seeds, under the header "Bad grad. (all bags)". The *uphill fraction* is the share of kept bags whose implicit gradient has a negative cosine with the unrolled one.

On UCSB the restriction keeps 6 per cent of bags (70 of 1160 over the twenty seeds, and 64 of 870 over the fifteen new ones), so we do not quote it.

Statistical analysis. Every inferential statement in this paper takes the *seed* as its independent unit. In the multiple-instance experiments a seed fixes the stratified five-fold partition and every initialisation. Each bag is held out exactly once per seed, the folds of a seed are combined with bag weights, and the result is one number per seed, model and dataset. Bags, folds and iterations are never counted as replicates. A "fraction of bags" is therefore a quantity measured within a seed, and it is not a sample size. The bags themselves are fixed, so the intervals below describe the variation due to partition and initialisation on these datasets. They do not describe the sampling of new bags or new datasets. Fox, Tiger and Elephant share a 230-dimensional feature vector in which the same 87 coordinates are identically zero in all three, and they are not independent tasks. In the synthetic and MNIST experiments a seed redraws the patterns, the queries and the initialisation. Unless stated otherwise, entries are means $\pm$ one standard deviation (sd) over seeds with the denominator $n-1$, the multiple-instance tables omit the leading zero, and "median over bags" means the median within a seed averaged over seeds. One decision rule fixed in advance, the depth-drift rule of the first registration, uses the population standard deviation (denominator n), and its verdict is the same under either convention.

Five seeds cannot carry a test, because the smallest attainable two-sided p of an exact paired test is $2/2^5 = 0.0625$. Every five-seed comparison is therefore descriptive. Where a registration uses one standard deviation, in the drift rule and in the iterated-rows registration's accuracy clause, it is a descriptive rule. It is not a test of equivalence. For the synthetic table the paired differences against the adaptive network at one step have 95% t intervals (4 degrees of freedom) of $[-0.0009, +0.0019]$, $[-0.0031, +0.0021]$ and $[-0.041, -0.007]$ for ours at 300 steps at masks 0.5, 0.6 and 0.7, and of $[-0.038, -0.020]$, $[-0.141, -0.096]$ and $[-0.293, -0.270]$ for ours at one step.

The confirmatory tests. The confirmatory analysis uses the fifteen seeds of the twenty-seed registration alone, which were run after the five-seed runs behind Tables 17 and 27. For each statistic and dataset we take the per-seed paired difference and test its mean with the exact two-sided sign-flip test over all 2^{15} sign patterns (smallest attainable $p = 6.1 \cdot 10^{-5}$). We correct with Holm's method over the registered family of five statistics by five datasets. The family has 24 realised comparisons, because the uphill statistic is undefined on UCSB. MNIST-bags has its own later registration, with the same analysis on its own fifteen new seeds and a separate Holm family of its five statistics. There the gaps in convergence and in cycling are confirmed (adjusted $p = 3 \cdot 10^{-4}$) and depth drift is not. Its two gradient statistics are reported and not quoted, because the restriction keeps 10 per cent of the adaptive network's bags. Intervals are 95% percentile bootstrap intervals over seeds (10 000 resamples). A t interval with 14 degrees of freedom is 11 to 18 per cent wider in every comparison and changes no verdict. The converged and cycle fractions are nearly complementary, so the family counts one effect twice, which makes the correction conservative. The same five statistics against the modern Hopfield network are the negative control, corrected as a family of their own, and none is confirmed.

Accuracy is tested for equivalence, not for difference. The margin 0.03 was declared before the run. The verdict is read from the 90% interval, the interval form of two one-sided tests at level

0.05, with no correction across the ten comparisons. An interval that excludes zero and straddles the margin is therefore inconclusive by design. Table 18 gives every comparison, with the standardised paired effect d_z, the mean difference over its standard deviation, as a descriptive addition that was not registered. The zero counts of the descent checks (0 of 64 800, 0 of 18 000, 0 of 12 000 and 0 of 1800 steps) are exhaustive checks of a deterministic inequality along trajectories. Their steps are not independent samples, and we attach no interval to them. Timings are medians of five repetitions with one thread requested. The software is Python 3.9.6, NumPy 2.0.2 and PyTorch 2.8.0 in double precision.

H.2 Main results

Table 17: Multiple-instance learning in full, with the retrieval map as the pooling layer. Each model is trained with the map unrolled $K = 5$ times and evaluated at $K' = 1000$, over twenty seeds, with a learning rate of 0.1 on the Hopfield parameters. "Converged" and "cycles" are fractions of held-out bags, given as mean $\pm$ sd over seeds of the per-seed fraction. "Bad grad." is the registered fraction of kept bags whose implicit and unrolled gradients differ by more than 10%, pooled over the twenty seeds. A bag is kept when its unrolled gradient has norm above 10^{-12}, and "bags kept" counts these bags out of 4000 per dataset and 1160 on UCSB. †UCSB keeps 6 per cent of the adaptive network's bags, too few to quote. "Depth drift" is the accuracy lost by iterating past the training depth, a mean over seeds. Iterations to a residual of 10^{-6} are medians over the bags that reach it, which for the adaptive network is 49 to 62 per cent of them. The condition number of $\mathbf{I} - J_T$ is a median over all bags at the final iterate. Both are per-seed medians averaged over seeds. Bold marks a gap that the confirmatory tests on fifteen new seeds confirm (Table 18) and that the text claims, and nothing else is bold. Elephant's depth drift is confirmed there, under the release protocol, and is not bold, because we claim no drift advantage. The modern Hopfield network converges in at least 99.9% of bags on every dataset. Datasets and their order follow the adaptive network's benchmark, which reports AUC where we report accuracy. Appendix H.1 defines every statistic, and Table 27 gives MUSK1.

	Tiger		Fox		Elephant		UCSB	
Metric	Ada.	ours	Ada.	ours	Ada.	ours	Ada.	ours
converged	.534 ± .220	**1.000 ± .001**	.618 ± .219	**1.000**	.540 ± .201	**1.000 ± .002**	.488 ± .220	**1.000**
cycles	.465 ± .220	**0**	.381 ± .218	**0**	.457 ± .201	**0**	.512 ± .220	**0**
bad grad.	.431	**.000**	.349	**.000**	.413	**.001**	.957†	.000
bags kept	1666	3734	1548	2553	2119	3698	70	1053
depth drift	.0263	.0025	.0115	.0020	.0260	.0037	.0164	.0138
iterations	2.7	13.9	2.6	14.4	3.1	13.2	2.0	51.7
$\mathrm{cond}(\mathbf{I} - J_T)$	1.00	1.48	1.00	1.55	1.00	1.38	1.00	5.07

Cycling and the fixed point. The costs that the limit in Theorem 3(iii) leads us to expect are in the same rows of Table 17 as the differences in convergence, and Figure 9 shows the dynamics behind them. At a learning rate of 0.1, about half of the adaptive network's trained fits enter period-2 cycles when iterated, nearly always in every held-out bag (Appendix H.3). Its median residual after 1000 steps is 0.38 to 0.96. Over all twenty seeds it cycles in 38 to 52% of bags. Ours converges in at least 99.9%. Every bag has a fixed point, because the map is continuous and sends the convex hull of the bag's instances into itself. In the cycling bags plain iteration does not reach it. An implicit layer whose forward pass iterates the map then differentiates at a point that is not a fixed point.

Root-finders do reach these points. Broyden's method reaches a fixed point in 50 to 86% of the cycling bags within 1000 evaluations, on average over seeds, and Anderson acceleration in 0 to 52% (Appendix H.9). In an exploratory run on five of the seeds, Broyden's method with a backtracking line search reaches 91 to 97% of them on four datasets and 22% on UCSB. At 99.2% or more of the points found, the almost-everywhere Jacobian of the map has spectral radius above one, so plain iteration is unstable there. Where a root-finder converges, a layer whose forward pass is that root-finder would nearly always return such a point. The models here were trained through the iteration, not through a root-finder.

Training at a smaller depth does not remove the cycles. Trained at $K = 1$, the depth that its released code uses, the adaptive network's map still cycles over five seeds in 57%, 76%, 78% and 56% of

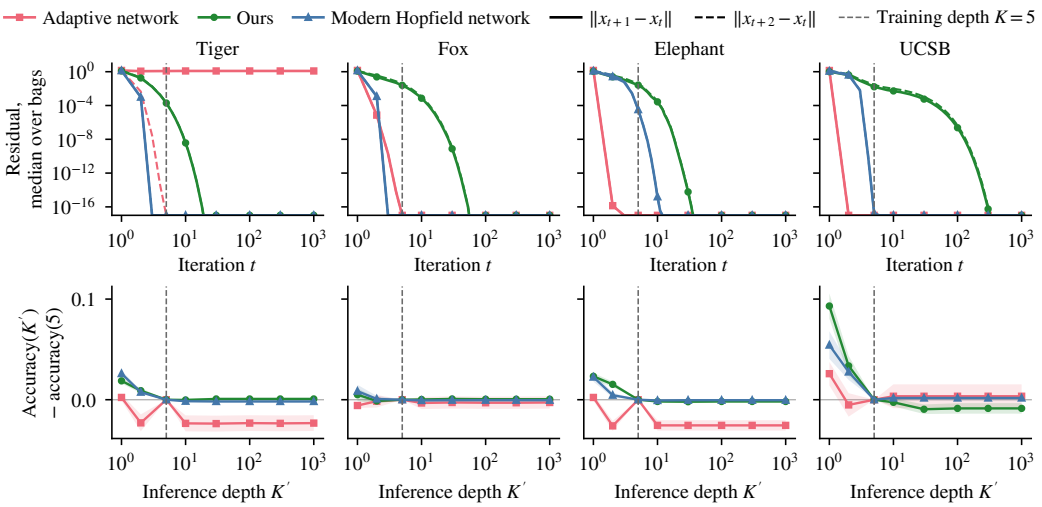

Figure 9: The pooling map iterated, one column per dataset. Top: residual along the iteration for one held-out fold (one fold of one seed), median over bags, solid $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|$ and dashed $\|\mathbf{x}_{t+2} - \mathbf{x}_t\|$. A converging map sends both to zero. A period-2 orbit sends only the dashed one to zero while the solid one stays at order one, which is the adaptive network on Tiger. Its fits on Fox, Elephant and UCSB converge, as about half of its fits do (Table 17). The gradient readout converges in nearly every bag (Table 17) and takes more steps than the modern Hopfield network, as the limit of Theorem 3(iii) leads us to expect. Bottom: held-out accuracy at depth K' minus accuracy at the training depth $K = 5$ (dashed), per seed with the five folds pooled by bag count, mean and standard error over the twenty seeds of Table 17. A map with a fixed point returns the same answer at every depth once it has converged. Ours on UCSB converges slowest of the four (Table 17), and its accuracy still moves at $K' = 30$. MUSK1 is Figure 11.

held-out bags on Tiger, Fox, Elephant and UCSB, and in 52% on MUSK1. With the dot-product base of its released code restored, the map cycles in 37 to 51% of held-out bags over twenty seeds. The gaps in convergence and in cycling are confirmed against it on all five datasets, with a largest adjusted p of 0.0017 (Table 22).

The constrained map converges. Theorem 2 gives descent and step lengths that tend to zero. Convergence is consistent with that and is not implied by it. The map pays in iterations and conditioning, as the limit in Theorem 3(iii) leads us to expect. Among the four datasets of Table 17 it pays most on UCSB, at 52 iterations and a condition number of 5.1. It pays more on MUSK1 (Appendix H.7), at 148 iterations and 14.7. There 65 of 100 fits end with their smallest gate entry within 0.01 of the floor, the regime in which $\mathbf{I} - J_T$ is worst conditioned.

Gradient fidelity. An equilibrium layer is trained by differentiating a converged map through the implicit function theorem, which costs $O(1)$ memory against $O(K)$ for unrolling. That derivative is correct only at a fixed point. Over the twenty seeds the adaptive network's implicit and unrolled gradients differ by more than a tenth in 37 to 48 per cent of held-out bags on the four datasets of Table 17. In more than half of those bags the unrolled gradient has norm at most 10^{-12}, where a relative error carries no information. Among the kept bags, the registered statistic, the fraction is 35 to 43 per cent on Fox, Tiger and Elephant, and UCSB keeps too few such bags to quote. The constrained map stays at or below 0.1 per cent on every dataset under both counts. So does the modern Hopfield network under the registered count. Unrestricted, its fraction runs from 0 to 0.6 per cent over the four datasets, a mean of 0.3.

The disagreement is not an artefact of conditioning. The smallest singular value of $\mathbf{I} - J_T$ is above 10^{-6} in every bag of every model. What fails for the adaptive network is therefore convergence to a fixed point, and the linear solve does not fail.

Two views of one failure. Admissibility guarantees descent, and its absence does not by itself prevent convergence, as the iterated rows of Appendix H.6 show. For the adaptive network, convergence and gradient fidelity are two views of the same failure to converge, so they are not independent evidence. Among its kept bags the two gradients differ by more than a tenth in 0 to 0.2 per cent of the bags that converge and in 99.9 to 100 per cent of those that do not, over the five datasets.

Depth drift. Depth drift is the primary pre-registered statistic, and it is the weakest of the five. Over twenty seeds on the four datasets of Table 17, the means are 0.0200 for the adaptive network, 0.0055 for ours and 0.0021 for the modern Hopfield network, which beats us. Ours drifts less than the adaptive network in 46 of 80 seed-dataset pairs. On UCSB ours drifts nearly as much as the adaptive network (0.0138 against 0.0164), as it does on MUSK1 (0.0168 against 0.0179), and the modern Hopfield network alone sits at the floor on both. The pre-registered rule compares ours with the adaptive network only. It needs a gap of at least 0.05 that also exceeds one seed standard deviation, and it calls every dataset a tie. The confirmatory test nonetheless confirms one drift gap, on Elephant (Appendix H.2).

Twenty seeds. The twenty-seed registration, recorded before its code, adds fifteen seeds on all five datasets and analyses them alone, since five seeds cannot carry a test (Appendix H.1). The gaps in convergence and in cycling are confirmed on all five datasets, three of which, Fox, Tiger and Elephant, share one feature construction. The gap in gradient error is confirmed on Fox, Tiger and Elephant against both forms of the adaptive network (Table 22). It is confirmed on MUSK1 against the distance form only, because the gradient comparisons against the model with both bases are undefined there. The test is an exact paired sign-flip test with a Holm correction over the registered family, and for these gaps every adjusted p is at most 0.0035 (Table 18). The mean difference in the converged fraction is 0.33 to 0.52, and its smallest 95% lower bound is 0.23, on Fox. The uphill fraction is confirmed on the same four datasets as the gradient error, against the same forms. The standardised paired effects of the confirmed gaps in these four statistics are 1.2 to 2.2. This range leaves out the UCSB gradient cell, which is reported and not quoted.

Depth drift is confirmed on Elephant only, under the release protocol that the fifteen seeds use. Against the model with both bases it is confirmed on Fox and not on Elephant. Under the registered protocol the adaptive network drifts less (Appendix H.1), so we claim no drift advantage. None of the five statistics separates ours from the modern Hopfield network, which drifts less than ours on MUSK1 and UCSB. The seed varies the partition and the initialisation on fixed bags. Every interval here therefore covers training randomness. It does not cover the sampling of new data.

Accuracy at the trained depth. Accuracy at the trained depth was registered as an equivalence test with a margin of 0.03, declared before the run. Against the adaptive network, accuracy is equivalent on Fox, where every model is near chance. It differs by $+0.031$, $+0.023$ and $+0.030$ on MUSK1, Tiger and Elephant, with 90% intervals that exclude zero and do not clear the margin. It is above the margin on UCSB ($+0.086$, interval $[0.051, 0.122]$). The registered test is one of equivalence, so an interval that excludes zero and straddles the margin is inconclusive by design, and we claim no accuracy advantage. Table 19 gives the accuracies themselves, for all three models trained at $K = 5$. Ours and the modern Hopfield network are both more accurate after one step than at the training depth on Tiger, Elephant and UCSB. Iterating past one step costs them accuracy too.

Table 18: The confirmatory analysis on fifteen new seeds ($n = 15$ seeds per comparison, the adaptive network (Ada.) against ours). The gradient-error and uphill rows carry the registered restriction to bags whose unrolled gradient exceeds 10^{-12}, as does the bad-gradient row of Table 17. Mean paired difference, 95% percentile bootstrap interval over seeds, standardised paired effect d_z (exploratory), the number of seeds with a positive difference, the exact two-sided sign-flip p and its Holm adjustment over the 24 realised comparisons. Bold marks a confirmed gap, meaning adjusted $p < 0.05$ with the predicted sign. †The registered restriction to bags whose unrolled gradient exceeds 10^{-12} keeps 64 of 870 UCSB bags over these fifteen seeds, so this comparison is reported and not quoted.

Dataset	Statistic	Mean	95% interval	d_z	Seeds > 0	p	Holm p
MUSK1	converged, ours − Ada.	$+.\mathbf{503}$	$[+.347, +.648]$	1.66	13/15	$2.4 \cdot 10^{-4}$	.002
	cycles, Ada. − ours	$+.\mathbf{503}$	$[+.347, +.647]$	1.66	13/15	$2.4 \cdot 10^{-4}$	.002
	grad. error > 0.1, Ada. − ours	$+.\mathbf{528}$	$[+.347, +.695]$	1.51	12/15	$4.9 \cdot 10^{-4}$	.003
	uphill, Ada. − ours	$+.\mathbf{296}$	$[+.191, +.397]$	1.42	12/15	$4.9 \cdot 10^{-4}$	.003
	depth drift, Ada. − ours	$-.005$	$[-.012, +.003]$	-0.32	4/15	.307	.613
Fox	converged, ours − Ada.	$+.\mathbf{325}$	$[+.226, +.427]$	1.57	14/15	$1.2 \cdot 10^{-4}$	.002
	cycles, Ada. − ours	$+.\mathbf{324}$	$[+.225, +.426]$	1.56	14/15	$1.2 \cdot 10^{-4}$	.002
	grad. error > 0.1, Ada. − ours	$+.\mathbf{305}$	$[+.196, +.419]$	1.33	14/15	$1.2 \cdot 10^{-4}$	.002
	uphill, Ada. − ours	$+.\mathbf{164}$	$[+.106, +.224]$	1.35	14/15	$1.2 \cdot 10^{-4}$	.002
	depth drift, Ada. − ours	$+.011$	$[+.003, +.021]$	0.63	10/15	.017	.068
Tiger	converged, ours − Ada.	$+.\mathbf{450}$	$[+.339, +.572]$	1.90	15/15	$6.1 \cdot 10^{-5}$	.001
	cycles, Ada. − ours	$+.\mathbf{449}$	$[+.339, +.570]$	1.90	15/15	$6.1 \cdot 10^{-5}$	.001
	grad. error > 0.1, Ada. − ours	$+.\mathbf{424}$	$[+.289, +.572]$	1.47	15/15	$6.1 \cdot 10^{-5}$	.001
	uphill, Ada. − ours	$+.\mathbf{216}$	$[+.131, +.308]$	1.21	14/15	$1.2 \cdot 10^{-4}$	.002
	depth drift, Ada. − ours	$+.017$	$[+.002, +.036]$	0.50	8/15	.080	.239
Elephant	converged, ours − Ada.	$+.\mathbf{469}$	$[+.355, +.577]$	2.05	15/15	$6.1 \cdot 10^{-5}$	.001
	cycles, Ada. − ours	$+.\mathbf{466}$	$[+.352, +.574]$	2.03	15/15	$6.1 \cdot 10^{-5}$	.001
	grad. error > 0.1, Ada. − ours	$+.\mathbf{447}$	$[+.319, +.568]$	1.77	15/15	$6.1 \cdot 10^{-5}$	.001
	uphill, Ada. − ours	$+.\mathbf{233}$	$[+.168, +.295]$	1.80	15/15	$6.1 \cdot 10^{-5}$	.001
	depth drift, Ada. − ours	$+.\mathbf{025}$	$[+.012, +.038]$	0.96	12/15	.003	.015
UCSB	converged, ours − Ada.	$+.\mathbf{518}$	$[+.397, +.633]$	2.19	15/15	$6.1 \cdot 10^{-5}$	.001
	cycles, Ada. − ours	$+.\mathbf{518}$	$[+.397, +.633]$	2.19	15/15	$6.1 \cdot 10^{-5}$	.001
	grad. error > 0.1, Ada. − ours†	$+.973$	$[+.926, +1.000]$	12.41	15/15	$6.1 \cdot 10^{-5}$	.001
	uphill, Ada. − ours	undefined on at least one seed					
	depth drift, Ada. − ours	$+.000$	$[-.015, +.017]$	0.00	4/15	1	1

Table 19: Accuracy against inference depth on the pooling task of Table 17. All three models are trained with the map unrolled $K = 5$ times and evaluated after K' applications, mean ± sd over the twenty seeds, so $K' = 5$ is the training depth. Appendix H.2 gives the registered test at the training depth.

Model	K'	Tiger	Fox	Elephant	UCSB	MUSK1
Adaptive	1	$.741 \pm .025$	$.568 \pm .031$	$.763 \pm .026$	$.592 \pm .044$	$.792 \pm .025$
	5	$.739 \pm .027$	$.574 \pm .032$	$.761 \pm .028$	$.566 \pm .051$	$.786 \pm .031$
	1000	$.715 \pm .023$	$.571 \pm .030$	$.736 \pm .022$	$.570 \pm .049$	$.781 \pm .031$
Ours	1	$.779 \pm .016$	$.577 \pm .035$	$.817 \pm .023$	$.739 \pm .047$	$.808 \pm .028$
	5	$.761 \pm .018$	$.572 \pm .031$	$.794 \pm .019$	$.646 \pm .064$	$.811 \pm .029$
	1000	$.761 \pm .020$	$.572 \pm .032$	$.792 \pm .021$	$.637 \pm .059$	$.799 \pm .033$
MHN	1	$.774 \pm .026$	$.582 \pm .038$	$.793 \pm .025$	$.700 \pm .047$	$.810 \pm .037$
	5	$.748 \pm .029$	$.574 \pm .033$	$.771 \pm .025$	$.646 \pm .066$	$.795 \pm .034$
	1000	$.746 \pm .028$	$.573 \pm .033$	$.770 \pm .026$	$.647 \pm .066$	$.795 \pm .034$

More bags. The MNIST-bags registration asks whether the cycling is an artefact of 58 to 200 bags, and a later one adds fifteen seeds to it. On MNIST-bags (Ilse et al., 2018) with 500 bags and the setup unchanged, over twenty seeds the adaptive network converges in 40% of held-out bags and cycles in 60%, against 100% convergence for ours and for the modern Hopfield network. With both of its bases it cycles in 38%. On the fifteen new seeds alone the gap in convergence is 0.59 (95% interval $[0.51, 0.67]$, Holm-adjusted $p = 3 \cdot 10^{-4}$), with the predicted sign in every seed. Instances there are pooled raw pixels with no encoder. Every model's accuracy is therefore within 0.15 of chance, and we report none.

H.3 MECHANISM AND LABEL FLIPS

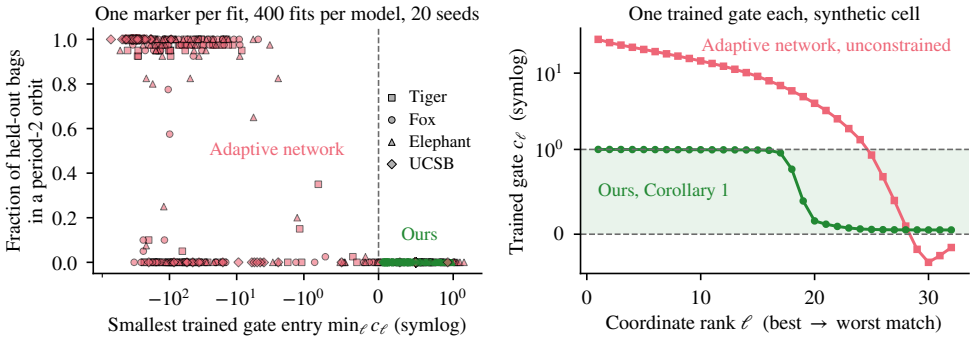

Figure 10: Cycling appears only with a negative gate entry. Left: the adaptive network's 400 runs on its four benchmark datasets (red). The horizontal axis is a run's smallest trained gate entry, and the vertical axis is its fraction of held-out bags that cycle with period two. Every run that cycles has a negative entry, and no run with a non-negative gate cycles. Our map (green) keeps every entry at least the floor $f = 0.05$ of Corollary 1 by construction and never cycles. Right: one trained gate of each model against coordinate rank, in the synthetic setting of Figure 1, both trained with AdamW for 120 steps on the same queries. The shaded band marks entries in $[0, 1]$.

The mechanism. The mechanism works in one direction, and the sign of the gate separates the fits more sharply than non-monotonicity does (Figure 10). The adaptive network has 400 fits at a learning rate of 0.1, from twenty seeds and five folds on each of the four datasets of Table 17, and we split them by the sign of the trained gate. Every one of the 197 fits that cycled had a negative gate entry, and none of the 108 fits with a non-negative gate cycled in a single bag. MUSK1 adds 53 cycling fits of its 100, each with a negative entry, and none of its 34 fits with a non-negative gate cycles. A fit that cycles does so almost everywhere, in a median of 100 per cent of its held-out bags, and 177 of the 197 cycle in at least 90 per cent. The converse fails, since 95 of the 292 fits with a negative entry never cycled. Non-monotonicity does not separate the fits at all, because 391 of the 400 are non-monotone.

Proposition 3 gives the mechanism. The trained β is finite, and the walk below starts outside the hypothesis of part (v). The agreement is therefore empirical. With the squared-distance base and $\mathbf{c} \geq 0$ no two instances are each other's best matches. A negative c_l multiplies a non-positive sorted coordinate and can make $s_b(\boldsymbol{\xi}_a) > 0 = s_a(\boldsymbol{\xi}_a)$. Two instances that are each other's best matches carry an orbit of period two once the softmax saturates. The label-flip registration tests this bag by bag. Every one of the 10 390 cycling bags of either form of the adaptive network contains such a pair, and the two ends of the orbit sit on one in 97 to 100% of them. The saturated walk from the trained initial state predicts whether a bag cycles in 98 to 100% of held-out bags. Corollary 1 forbids $c_l < 0$ at every $\boldsymbol{\theta}$. The gates of Corollary 1 also lie above the floor f and are non-increasing.

Does the cycling change the decision? It does, in 9 to 27% of the adaptive network's held-out bags and in none of ours. Two points of an orbit can fall on the same side of the linear head, so a cycle need not change the prediction. A registration recorded before its code measures the change bag by bag, on fifteen further seeds of all five datasets, disjoint from the twenty of Table 17. The *parity flip* is the fraction of held-out bags whose predicted label after 999 applications of the map differs from the label after 1000, and it is zero for a map that converges. For the adaptive network it is 9 to 26% with its distance base and 9 to 27% with both of its bases, against zero in every bag for ours and for the modern Hopfield network. The gap is confirmed on all five datasets against both forms, with a Holm-adjusted p of $6.1 \cdot 10^{-4}$ in every comparison, the smallest the test can return (Table 20). Among the adaptive network's cycling bags the label alternates in 21 to 61%, pooled over the cycling bags of the fifteen seeds. Its label after 1000 steps also differs from the label at the trained depth of five in 9 to 27% of bags, a mean over those seeds. Both ranges take their ends from the model with both bases (Table 20). The same depth flip is 1 to 5% for ours and 0.2 to 0.8% for

the modern Hopfield network. On these seeds the adaptive network cycles in 29 to 53% of held-out bags and ours in none.

Label flips and the best-match walk. The label-flip registration fits the three models of Table 17 and the model with both bases on these fifteen seeds, five folds of each of the five datasets. The setup of Table 17 is unchanged, except that the implicit-gradient pass is not run. For each held-out bag the best-match graph is computed over the bag's distinct instances with the trained gate or gates, and the walk starts at the instance of largest score at the trained initial state. Both forms of the adaptive network give a symmetric S in every bag, and no best-match graph has a cycle longer than two, as Proposition 3(iii) requires. All three registered predictions hold. The distance-base flip is at least 0.05 on all five datasets, every cycling bag contains a mutual pair, and the walk agrees with the cycle classification in at least 98% of bags on every dataset. The largest separation weight at the last step has median 1.000 in cycling and in settled bags, so the trained softmax is saturated.

Table 20: Label flips on fifteen further seeds, disjoint from the twenty of Table 17, under a registration recorded before its code. "Flip" is the fraction of held-out bags whose predicted label differs between 999 and 1000 applications of the map, and "depth flip" the fraction whose label after 1000 applications differs from the label at the trained depth of five. Means over seeds, bag-weighted over folds. Every flip gap against ours is confirmed (exact paired sign-flip test, Holm over ten comparisons, adjusted $p = 6.1 \cdot 10^{-4}$ in every comparison). Ours and the modern Hopfield network flip in no bag, and the latter's depth flip is 0.002 to 0.008. "Walk" is the fraction of held-out bags whose cycle classification the saturated walk of Proposition 3 predicts.

	Adaptive, distance base				Adaptive, both bases				Ours
Dataset	cycles	flip	depth flip	walk	cycles	flip	depth flip	walk	depth flip
MUSK1	.453	.113	.113	1.000	.444	.092	.091	.999	.051
Fox	.524	.204	.204	.988	.466	.180	.180	.995	.009
Tiger	.525	.172	.173	.981	.524	.161	.161	.993	.007
Elephant	.286	.095	.096	.983	.463	.130	.131	.996	.008
UCSB	.457	.255	.255	1.000	.448	.274	.275	1.000	.028

H.4 THE CONTROL: GATE SIGN AGAINST READOUT

Design and columns. Table 3 crosses the gate with the readout and adds the adaptive network's released configuration. Both registrations use the setup of Table 17 on all five datasets over twenty seeds. Each cell of Table 3 is the range over the five datasets of the twenty-seed mean, and Table 21 gives every dataset. The columns are defined in Appendix H.1. The signed gate under the gradient readout converges in only 4 to 36 per cent of bags, so its iteration cell averages only the seeds with such a bag, 4 of 20 on UCSB, 6 on MUSK1, 17 on Fox, 18 on Tiger and 20 on Elephant. "Released" has both bases, a block gate of stride four and a raw temperature. The first and fifth rows are the models of Table 17, with MUSK1 added. Only the fifth has an energy guarantee (Theorem 2).

Gate sign against readout. Is it the energy that removes the cycling, or the sign of the gate? Under the pattern readout the sign of the gate removes it in these runs. A registration recorded before its code fills in the design of gate against readout with three new maps, on all five datasets over twenty seeds (Table 3). Two keep the pattern readout and make the gate non-negative. One is our monotone gate, trained through the pattern readout, unlike the last row of Table 13, which swaps the readout after training. The other is the adaptive network's signed gate clamped at zero, which is not constrained to be monotone and has no energy guarantee. The third map keeps the signed gate and reads out by the gradient. Both non-negative gates remove the cycling. Each converges in at least 99.7% of held-out bags on every dataset and cycles in none. In 159 of the 500 fits of the clamped gate, training drives every entry to zero, and the map then returns the bag's mean, a constant map that is trivially admissible. None of the other 341 fits cycles either, and 317 of them end with a gate that is not monotone. Their gaps in convergence and in cycling against the adaptive network are confirmed on all five datasets, on the fifteen confirmatory seeds, with a Holm-adjusted p of $1.2 \cdot 10^{-3}$ in each of the 20 registered comparisons, where convergence and cycling count one effect twice. Under the pattern readout, then, the sign of the gate stops the cycling in these runs, consistent with Proposition 3(iv), which rules out a mutual best-match pair. No energy is needed for it.

One registered prediction failed. We predicted that our gate under the pattern readout would need at most five iterations. It needs 7.7 to 13.3 on four of the five datasets and 4.8 on MUSK1, against 2.0 to 3.4 for the clamped gate and 13.2 to 148 for our gradient readout. On Tiger, Fox and Elephant most of our map's iterations therefore come with the monotone gate under either readout. The gradient readout adds most of them only on UCSB and MUSK1. The registration allowed the conclusion of the previous paragraph only if this prediction held. For a failed prediction it prescribed the reading that the gradient readout adds convergence, and these data contradict that reading. We draw the conclusion from the clamped gate and from the prediction's two convergence clauses, which held, and we drop its claim about the rate.

Under the gradient readout the signed gate diverges in 54 to 90% of held-out bags. The registration's rule reads this divergence as evidence that the gradient readout needs the monotone gate (Proposition 6). We read it differently. The map changes the gate's sign and its order together, and the sign alone explains the divergence. At the $\kappa = \beta\|\mathbf{c}\|_\infty$ that this map uses, a gate with $\mathbf{c} \geq 0$ makes each coordinate of $T(\mathbf{x})$ a convex combination of x_i and the patterns' i-th coordinates, so no orbit leaves the box they span, whatever the gate's order. The order is what convexity needs, and no map here varies it at a fixed sign. In these runs the energy therefore adds the descent guarantee. It does not add convergence. Apart from the constant maps, neither non-negative map under the pattern readout has an energy guarantee.

At the training depth, the 90% interval of our gate's accuracy under the pattern readout minus the gradient readout lies inside ± 0.03 on Tiger, Fox and Elephant, which was the registered question. On UCSB and MUSK1 it crosses the margin, and the question stays open there. On MUSK1 the pattern readout is lower by 0.021 ($[-0.038, -0.003]$). Table 21 gives every dataset.

Table 21: Every dataset of Table 3 for the three maps of the gate-sign-against-readout registration and the released configuration, over twenty seeds. Converged, cycles and accuracy at the training depth $K' = 5$ are means $\pm$ sd over seeds of the per-seed value, diverged and depth drift are means over seeds, and iterations are as in Table 3. The signed gate under the gradient readout counts every diverged bag as wrong at every later depth, so its depth drift measures divergence and is not comparable with the other rows. Against the adaptive network of Table 17, the gaps in convergence and in cycling of the clipped gate and of the monotone gate under the pattern readout are confirmed on every dataset, with a Holm-adjusted p of $1.2 \cdot 10^{-3}$ in each of the 20 registered comparisons. The registered accuracy question compares our gate under the two readouts at $K' = 5$ over the fifteen confirmatory seeds. The 90% interval of the difference, pattern readout minus gradient readout, is $[-0.009, +0.007]$ on Tiger, $[-0.014, +0.020]$ on Fox, $[-0.011, +0.011]$ on Elephant, $[-0.032, +0.032]$ on UCSB and $[-0.038, -0.003]$ on MUSK1, against a margin of ± 0.03.

Dataset	Map	Converged	Cycles	Diverged	Iterations	Depth drift	Acc. @5
Tiger	signed, gradient	.315 ± .215	.008 ± .010	.540	101.2	.3910	.732 ± .023
	clipped, pattern	.997 ± .005	.000 ± .000	.000	3.4	.0013	.752 ± .029
	monotone, pattern	.999 ± .002	.000 ± .000	.000	12.6	.0018	.756 ± .015
	released	.821 ± .186	.178 ± .186	.000	8.4	.0080	.764 ± .022
Fox	signed, gradient	.364 ± .245	.001 ± .002	.610	59.9	.3438	.570 ± .032
	clipped, pattern	.999 ± .003	.000 ± .000	.000	3.0	.0007	.564 ± .024
	monotone, pattern	1.000 ± .001	.000 ± .000	.000	11.7	.0030	.571 ± .031
	released	.725 ± .215	.272 ± .215	.000	8.8	.0177	.572 ± .037
Elephant	signed, gradient	.321 ± .151	.011 ± .019	.570	53.3	.4155	.745 ± .032
	clipped, pattern	.999 ± .005	.000 ± .000	.000	3.4	.0010	.783 ± .018
	monotone, pattern	.999 ± .002	.000 ± .000	.000	13.3	.0045	.798 ± .025
	released	.784 ± .147	.213 ± .148	.000	11.1	.0175	.771 ± .028
UCSB	signed, gradient	.041 ± .105	.000 ± .000	.889	441.0	.5267	.583 ± .049
	clipped, pattern	1.000 ± .000	.000 ± .000	.000	2.0	.0000	.677 ± .064
	monotone, pattern	1.000 ± .000	.000 ± .000	.000	7.7	.0078	.653 ± .050
	released	.667 ± .165	.333 ± .165	.000	4.0	.0328	.612 ± .052
MUSK1	signed, gradient	.077 ± .133	.000 ± .000	.901	100.2	.7272	.802 ± .036
	clipped, pattern	.999 ± .002	.000 ± .000	.000	2.1	.0016	.807 ± .028
	monotone, pattern	.999 ± .002	.000 ± .000	.000	4.8	.0033	.795 ± .034
	released	.445 ± .177	.555 ± .177	.000	6.7	.0511	.801 ± .028

The released configuration. Table 17 departs from the adaptive network's released code in five ways: its temperature, its gate's stride, the gate's initial value, its bases and its training depth. A second registration restores the first four at once, with a raw temperature from one, a block gate of

stride four started at zero, and both bases (last row of Table 3). It keeps the training depth $K = 5$, where the released code runs the map for one step, and Appendix H.2 reports the adaptive network trained at $K = 1$. A check first matches the map to the released code to $1.6 \cdot 10^{-16}$. At the learning rate of 0.1 of Table 17, the released configuration cycles in 18 to 56% of held-out bags on the five datasets (Appendix H.5 varies the rate). Of its 500 fits, 189 cycle, against 250 of 500 in the configuration of Table 17 with MUSK1 added. Its trained temperatures can change sign, and a negative one flips the sign of that base's whole gate. The sign that matters is therefore that of each base's gate times its temperature. As in Table 17, none of the 80 fits in which both of these are non-negative cycles. For the dot-product base this is an observation, since Proposition 3(iv) covers the squared-distance base only. This registration predicts cycling and makes no confirmatory comparison with our map. The released code has no multiple-instance code, so this run shows its model class and initialisation in our setup and does not reproduce the published runs.

Both bases. The setup of Table 17 gives the adaptive network its squared-distance base alone, which shares our score exactly. Its equation (4) also has a dot-product base, sorted per pattern and gated per rank with a temperature of its own, and its released code trains both. A registration recorded before its code restores that base, so the pooling map is the model class of the released code with both bases at stride 1. The map matches the released implementation to $3.1 \cdot 10^{-16}$, and at a zero dot gate it equals the squared-distance map exactly. The dot gate starts at the ramp of the distance gate. Only the new map is trained. We predicted that the map still cycles in more than a tenth of held-out bags on at least three of the five datasets. It does on all five, in 37 to 51% over twenty seeds (Table 22).

The confirmatory analysis is that of the twenty-seed registration, on its fifteen new seeds. Its family has 21 defined comparisons rather than 25, because the gradient and uphill statistics are undefined on MUSK1 and UCSB. On one MUSK1 seed no held-out bag of the full model has an unrolled gradient above 10^{-12}. A smaller family is less conservative, but at 25 comparisons every verdict is the same. The gaps in convergence and in cycling are confirmed on all five datasets, the gradient and uphill gaps on Fox, Tiger and Elephant, and depth drift on Fox alone. On those three datasets the full model's registered bad-gradient fraction is 38 to 55%. Accuracy at the trained depth is equivalent on Fox and inconclusive on MUSK1, Tiger and Elephant. On UCSB ours is above the margin ($+0.074$, 90% interval $[0.032, 0.115]$), as it is against the squared-distance model, and we report it without a claim.

With the zero initialisation of the dot gate in the released code, the map trained at a learning rate of 0.1 cycles in 36 to 63% of bags over five seeds, and trained at $K = 1$ in 41 to 83%. On MNIST-bags it cycles in 38%, against 60% for the squared-distance base. Of its 500 fits on the five datasets, 243 cycle in at least one bag. Of those, 236 have a negative distance-gate entry, and the other 7 have a non-negative distance gate and a negative weight on the dot base. None of the 82 fits with both gates non-negative cycles.

Table 22: The adaptive network with both of its bases, the model class of its released code, as the pooling map of Table 17, over twenty seeds. Converged and cycles are means ± sd over seeds of the per-seed bag fraction. Ours converges in at least 99.9% of bags with no cycles on every dataset. "Bad grad." is the registered count, over the bags whose unrolled gradient exceeds 10^{-12} pooled over the seeds, and a dash marks a dataset where that restriction keeps under a tenth of the bags. Depth drift is a mean over seeds. The last column gives the Holm-adjusted p of the five comparisons against ours on the fifteen new seeds, in the order converged, cycles, bad gradient, uphill and depth drift, with Holm's correction over the 21 defined comparisons. A dash marks an undefined comparison, and bold a confirmed gap.

Dataset	Converged	Cycles	Bad grad.	Depth drift	Holm p against ours
MUSK1	.578 ± .206	.422 ± .206	—	.0255	**.002**, **.002**, —, —, .760
Fox	.491 ± .191	.506 ± .192	.546	.0160	**.001**, **.001**, **.001**, **.001**, **.020**
Tiger	.502 ± .191	.497 ± .191	.493	.0123	**.001**, **.001**, **.001**, **.001**, .125
Elephant	.502 ± .241	.498 ± .241	.375	.0160	**.001**, **.001**, **.001**, **.001**, .160
UCSB	.627 ± .214	.373 ± .214	—	.0414	**.001**, **.001**, —, —, .760
MNIST-bags	.619 ± .265	.381 ± .265	—	.0165	not in the family

H.5 THE TRAINING RECIPE AND THE LEARNING RATE

Every run of Table 17 trains the Hopfield parameters at a learning rate of 0.1 for 200 full-batch epochs, with the head and the query at 0.01. The rate 0.1 is the gate's rate in the adaptive network's retrieval experiments. For multiple-instance learning its paper follows another recipe, that of Hu et al. (2023), and its Table 8 tunes the number of epochs over $\{10, 20, 40\}$, the learning rate over $\{10^{-3}, 10^{-4}, 10^{-5}\}$ and a learning-rate decay over $\{0.9, 0.75\}$, which the script of Hu et al. (2023) applies after every epoch. A protocol fixed before its code was written runs two arms in our pipeline, on seeds 0 to 4 and the five folds, for the adaptive network's squared-distance form and for its released configuration. Every number below is a mean over the five seeds, and the analysis is descriptive.

The learning rate alone. The first arm changes only the learning rate η_h of the Hopfield parameters, over $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. Table 23 gives the fraction of held-out bags that cycle. It falls with the rate on every dataset for both forms. At 10^{-4} the squared-distance form cycles in no held-out bag, and its gate stays non-negative in 124 of 125 fits. The released configuration cycles there only on MUSK1, in 0.007 of the bags. The held-out AUC at the training depth is not monotone in the rate, but at 10^{-4} it is higher than at 0.1 on every dataset. At 10^{-4} the squared-distance form scores 0.901 against 0.878 at 0.1 on MUSK1, 0.600 against 0.590 on Fox, 0.850 against 0.826 on Tiger, 0.889 against 0.836 on Elephant and 0.714 against 0.616 on UCSB. The released configuration scores higher at 10^{-4} on all five datasets as well. In this pipeline, then, the squared-distance form's gate trained at 0.1 both cycles and scores lower than its gate at 10^{-4}, which stays non-negative in 124 of 125 fits. We make no accuracy claim from these values.

Table 23: The fraction of held-out bags that cycle, against the learning rate η_h of the Hopfield parameters, with every other part of the protocol of Table 17. Means over seeds 0 to 4.

Model	η_h	MUSK1	Fox	Tiger	Elephant	UCSB
Adaptive, distance base	10^{-1}	.557	.553	.513	.429	.493
	10^{-2}	.350	.333	.295	.319	.452
	10^{-3}	.052	.000	.057	.032	.110
	10^{-4}	.000	.000	.000	.000	.000
Released configuration	10^{-1}	.593	.408	.220	.158	.372
	10^{-2}	.465	.123	.000	.000	.079
	10^{-3}	.370	.000	.000	.000	.000
	10^{-4}	.007	.000	.000	.000	.000

The recipe of the adaptive network's paper. The second arm follows the recipe of the released script of Hu et al. (2023). It uses AdamW on every parameter at weight decay 10^{-4}, mini-batches of four bags, gradient clipping at norm one, the learning rate multiplied by the decay after every epoch, a validation split of each training fold, and the checkpoint of the best validation AUC. It runs the eighteen configurations of the grid above, and the configuration is selected in each outer fold by that fold's own validation AUC at the last epoch, the value that the script reports. Our pipeline keeps its bare map, one query and a linear head, where their network has embedding layers, several heads and bag dropout. This is therefore their recipe in our pipeline, and it is not a reproduction of their results or of the adaptive network's reported numbers. At the selected configuration the squared-distance form cycles in 0.043 of held-out bags on MUSK1, 0.088 on Fox, 0.169 on Tiger, 0.154 on Elephant and 0.241 on UCSB, which is at least 0.10 on Tiger, Elephant and UCSB. It cycles only at the grid's largest rate, 10^{-3}, which validation selects in most fits, except on UCSB at 10^{-4} with a decay of 0.9 and at least 20 epochs, where it cycles in 0.034 of the bags. Selection by the best epoch, selection on the first fold as in the released script, and the variant without gradient clipping each give at least 0.10 on three datasets as well. The released configuration cycles in 0.054 of held-out bags on MUSK1 and in none on the other four datasets. Across 7780 distinct trained models, 443 have a cycling bag, and every one of them has a negative gate entry, as the mechanism of Section 6 predicts.

The recipe gives the squared-distance form a lower held-out AUC on every dataset than the protocol of Table 17 does in this pipeline. For example, it reaches a held-out AUC of 0.721 on MUSK1 at the selected configuration, against 0.878 at $\eta_h = 0.1$. We report every test AUC below 0.6 in the same way for every model. Such values occur on Fox for every model under the recipe, and for both

forms of the adaptive network at $\eta_h = 10^{-1}$ and 10^{-2} (and for the released configuration at 10^{-3}). They also occur on UCSB for the squared-distance form under the recipe.

Our map under the recipe. Our map cycles in no bag at any configuration of the recipe. We predicted that it would also converge in at least 0.99 of held-out bags, and this prediction failed. Its lowest converged fraction, a mean over seeds for one dataset and configuration, is 0.962, on UCSB, and it is reached at every learning rate of the grid there. The bags in which it does not converge neither cycle nor diverge. They have not reached a residual of 10^{-8} within 1000 steps.

H.6 Other maps iterated

H.6.1 More rows iterated

Summary. Is the failure to converge a property of inadmissibility, or of this model? Inadmissibility alone does not cause it, because four of the six failure rows that we iterated converge. The iterated-rows registration iterates six more rows of Table 1 in the same setup, with nothing learned but β. Three are admissible: sparsemax retrieval, 1.5-entmax retrieval, and kernelised retrieval read at its own metric with a fixed W. Three are not: the Manhattan model, one-step attention with a value readout, and the nonparametric model at contamination 0.3 (Table 24). The three admissible rows converge in at least 99.8% of bags with no cycles, which is consistent with the descent their admissibility gives and is not implied by it. The three inadmissible rows also converge, in every bag on every dataset. Over the same four datasets and the same five seeds, the adaptive network converges in .443 to .568 of bags and cycles in .429 to .557. Its unrestricted bad-gradient fraction there is .363 to .512, and Table 17 gives its twenty-seed values under the restricted statistic. The registration's conclusion for this outcome was that admissibility is sufficient but not necessary for a fixed point. Only the second half holds. Admissibility guarantees descent, not convergence.

The inadmissible rows converge for the reason that the mechanism of Appendix H.3 gives, read in reverse. For a distance-based score every instance is a fixed point up to softmax leakage, so a saturated separation lands on one instance. The Manhattan model's fixed points sit on an instance in 86–99% of bags. For attention we measured the final state on one fold of one seed of MUSK1 and of Fox only. There it lies on one value row in the median bag. Among these softmax rows the distance-based scores have $s_a(\boldsymbol{\xi}_a) \geq s_b(\boldsymbol{\xi}_a)$, and the kernelised score is a Gram matrix. Neither of them can therefore have a mutual pair. Attention and the nonparametric score have no such guarantee.

A later registration iterates the two remaining universal-Hopfield failure rows, each in a modified form described below, and the nonparametric model at contamination 1 (Table 25). Two of these maps converge in every bag. The Euclidean distance under the identity separation, a combination that the universal-Hopfield paper defines and does not run, cycles in 38–83% of bags on four datasets, against our registered prediction, because its weights grow with distance. Of the six failure rows iterated, the adaptive network included and two universal-Hopfield rows in modified forms, two cycle and four converge. Every saturating row agrees with Proposition 3. In the registered runs none of their held-out bags has a best-match cycle, except one of Elephant's 1000 under the nonparametric model at contamination 1, and that bag converges. The identity separation does not saturate, and the proposition says nothing about it.

The six maps. Table 24 iterates the maps

$$\begin{aligned}
\text{sparsemax:}\quad & T = \boldsymbol{\Xi}\,\mathrm{sparsemax}(\beta\boldsymbol{\Xi}^\top\mathbf{x}), & \text{1.5-entmax:}\quad & T = \boldsymbol{\Xi}\,\mathrm{entmax}_{1.5}(\beta\boldsymbol{\Xi}^\top\mathbf{x}),\\
\text{kernelised:}\quad & T = \boldsymbol{\Xi}\,\mathrm{softmax}(\beta\boldsymbol{\Xi}^\top G\mathbf{x}), & \text{Manhattan:}\quad & T = \boldsymbol{\Xi}\,\mathrm{softmax}(-\beta\|\boldsymbol{\xi}_\mu - \mathbf{x}\|_1),\\
\text{attention:}\quad & T = V^\top\,\mathrm{softmax}(\beta K\mathbf{x}), & \text{nonparametric:}\quad & T = \boldsymbol{\Xi}\,\mathrm{softmax}(\beta\boldsymbol{\Xi}_\delta^\top\mathbf{x}).
\end{aligned}$$

Here $G = W^\top W$ with $W = \mathbf{I} + 0.3\,N(0,1)/\sqrt{d}$ fixed per fold, and $K = \boldsymbol{\Xi}W_k$ and $V = \boldsymbol{\Xi}W_v$ with fixed $W_k \neq W_v$. The contaminated patterns are $\boldsymbol{\Xi}_\delta = \boldsymbol{\Xi} + 0.3\,N(0,1)/\sqrt{d}$, where the noise is a fixed function of the instance. Everything else is the protocol of Table 17, with β the only learned Hopfield parameter. The predictions, the symmetric decision rule and the fixed matrices were fixed in the registration, which was recorded before its code. The table gives the per-dataset values.

UCSB in the iterated-rows registration. That registration also adds UCSB Breast Cancer, the dataset of this literature that the earlier registrations had not run, with all nine maps. At a learning

Table 24: Six rows of Table 1 iterated as the pooling layer, per dataset, five seeds, same protocol as Table 17. The converged, cycle, on-an-instance and bad-gradient columns are fractions of bags. "On an instance" is the fraction of all bags whose final iterate has converged and lies within 10^{-3} of a stored instance. "Bad grad. (all bags)" is the unrestricted fraction of all bags whose implicit and unrolled gradients differ by more than 0.1, not the restricted statistic of Table 17. Accuracy is at one step, mean $\pm$ sd over five seeds. Attention reads out value rows, so the on-an-instance column is undefined for it.

Dataset	Row	Converged	Cycles	On an instance	Bad grad. (all bags)	Accuracy @1
MUSK1	Sparsemax	1.000	0	.97	.009	.813 ± .024
	1.5-entmax	1.000	0	.96	.020	.793 ± .025
	Kernelised, fixed W	1.000	0	.73	.000	.811 ± .020
	Manhattan	1.000	0	.99	.002	.800 ± .028
	Attention, value readout	1.000	0	—	.000	.791 ± .026
	Nonparametric, $\delta = 0.3$	1.000	0	.74	.000	.791 ± .036
Fox	Sparsemax	1.000	0	.95	.026	.562 ± .037
	1.5-entmax	.999	0	.89	.033	.573 ± .031
	Kernelised, fixed W	1.000	0	.59	.001	.579 ± .018
	Manhattan	1.000	0	.98	.001	.552 ± .045
	Attention, value readout	1.000	0	—	.003	.567 ± .023
	Nonparametric, $\delta = 0.3$	1.000	0	.59	.000	.587 ± .033
Tiger	Sparsemax	.999	0	.93	.036	.752 ± .026
	1.5-entmax	.998	0	.91	.026	.769 ± .032
	Kernelised, fixed W	1.000	0	.58	.000	.776 ± .023
	Manhattan	1.000	0	.97	.004	.728 ± .031
	Attention, value readout	1.000	0	—	.002	.776 ± .026
	Nonparametric, $\delta = 0.3$	1.000	0	.61	.000	.763 ± .043
Elephant	Sparsemax	.999	0	.92	.048	.770 ± .026
	1.5-entmax	.999	0	.85	.044	.765 ± .046
	Kernelised, fixed W	1.000	0	.63	.002	.798 ± .036
	Manhattan	1.000	0	.86	.008	.764 ± .019
	Attention, value readout	1.000	0	—	.004	.798 ± .033
	Nonparametric, $\delta = 0.3$	1.000	0	.60	.000	.789 ± .040
UCSB	Sparsemax	1.000	0	.99	.003	.662 ± .034
	1.5-entmax	1.000	0	1.00	.000	.676 ± .056
	Kernelised, fixed W	1.000	0	.88	.000	.672 ± .097
	Manhattan	1.000	0	1.00	.000	.652 ± .054
	Attention, value readout	1.000	0	—	.000	.710 ± .060
	Nonparametric, $\delta = 0.3$	1.000	0	.91	.000	.669 ± .064

rate of 0.1 the adaptive network cycles in 49% of its bags and every other map converges in all of them, which replicates the pattern on a fifth dataset. One registered clause fails there and is reported as such. At one step the adaptive network trained at $K = 5$ scores $.607 \pm .041$, against $.714 \pm .043$ for the modern Hopfield network and $.724 \pm .063$ for ours. The gap is outside one seed standard deviation, so on UCSB accuracy is not a tie. Trained at $K = 1$ the adaptive network scores $.597 \pm .109$ and cycles in 56% of bags.

The remaining failure rows iterated. The registration of the remaining failure rows was recorded together with its first results. It iterates three more maps in the protocol of Table 24,

$$T = \boldsymbol{\Xi}\,\mathrm{softmax}\big(\beta\,\mathbf{r}/\mathbf{1}^\top\mathbf{r}\big), \qquad T = \boldsymbol{\Xi}\,\mathbf{a}/\mathbf{1}^\top\mathbf{a}, \qquad T = \boldsymbol{\Xi}\,\mathrm{softmax}(\beta\boldsymbol{\Xi}_\delta^\top\mathbf{x}) \ \text{ at } \delta = 1,$$

with $r_\mu = 1/(\|\boldsymbol{\xi}_\mu - \mathbf{x}\| + 10^{-2})$ and $a_\mu = \|\boldsymbol{\xi}_\mu - \mathbf{x}\|^2$. The first is a normalised universal Hopfield network. Its released code scores with the reciprocal of the summed absolute or squared differences, with no offset. The second is the Euclidean distance under the identity separation, with the normalisation of the released code. Its paper defines this combination and does not run it. The unnormalised form of the row in Table 1 is unbounded under iteration and is not run. The third is the nonparametric model at contamination 1, where the noise has about the norm of an instance. We predicted that all three converge. Two do, in every bag (Table 25). The identity row cycles on four datasets and converges on UCSB, whose bags are the largest.

The identity row's cycles are period-2 orbits. On one MUSK1 fold of one seed, 17 of the 19 held-out bags meet the period-2 test of the protocol after 1000 steps. None of them reaches residual 10^{-8}, and each has $\|\mathbf{x}_{1000} - \mathbf{x}_{998}\| < 10^{-8}$. In those 17 bags, $\|\mathbf{x}_{1000} - \mathbf{x}_{999}\|$ is between 0.24 and 1.63. The weights grow with distance, so the map sends a query toward the farthest instances and back. No temperature enters this map. Both implicit gradients are therefore zero, and the bad-gradient

column is not reported. The recipe of Proposition 1 replaces the identity separation by softmax, which is the admissible squared-distance row of Table 1, and we claim no repair from this run.

Table 25: The remaining failure rows iterated, per dataset, five seeds. Both universal-Hopfield maps are modified forms of their rows in Table 1, as the text describes. Columns as in Table 24.

Dataset	Row	Converged	Cycles	On an instance	Accuracy @1
MUSK1	Reciprocal, normalised	1.000	0	.96	$.824 \pm .025$
	Euclidean, identity sep	.157	.835	0	$.811 \pm .025$
	Nonparametric, $\delta = 1$	1.000	0	.82	$.809 \pm .031$
Fox	Reciprocal, normalised	1.000	0	.98	$.596 \pm .013$
	Euclidean, identity sep	.567	.393	0	$.579 \pm .021$
	Nonparametric, $\delta = 1$	1.000	0	.66	$.591 \pm .022$
Tiger	Reciprocal, normalised	1.000	0	.99	$.792 \pm .021$
	Euclidean, identity sep	.530	.434	0	$.803 \pm .018$
	Nonparametric, $\delta = 1$	1.000	0	.70	$.768 \pm .029$
Elephant	Reciprocal, normalised	1.000	0	.99	$.821 \pm .022$
	Euclidean, identity sep	.590	.383	.01	$.825 \pm .009$
	Nonparametric, $\delta = 1$	1.000	0	.54	$.787 \pm .024$
UCSB	Reciprocal, normalised	1.000	0	.90	$.714 \pm .029$
	Euclidean, identity sep	1.000	0	0	$.745 \pm .031$
	Nonparametric, $\delta = 1$	1.000	0	.85	$.707 \pm .024$

H.6.2 The gate against the separation

A registration fixed after the remaining failure rows had been run asks whether the failure to converge belongs to softmax or to the gate. It places the adaptive network's signed gate under the pattern readout with sparsemax, with 1.5-entmax and with a fixed kernel $W = \mathbf{I} + 0.3\, N/\sqrt{d}$ drawn once per fold. It places the monotone gate with the gradient readout of Proposition 4 under the same three, in the protocol of Table 24. The only learned Hopfield parameters are the gate and β. We predicted that every signed-gate map cycles in more than a tenth of bags on at least three datasets. That prediction held on all five (Table 26). We also predicted that every monotone-gate map converges in at least 99% of bags and cycles in at most 1% on at least four datasets, with a bad-gradient fraction below .05 on every dataset. Its convergence and cycle clauses held on all five. The mechanism of Appendix H.3 held without exception on 375 fits. Every one of the 188 fits that cycled had a negative gate entry, and none of the 107 fits with a non-negative gate cycled.

The bad-gradient clause failed as written for the monotone gate under the kernel. Its registered column reads .03 to .41, above .05 on four of the five datasets. In 833 of the 869 bags it counts, the unrolled gradient has norm below 10^{-15}, and in 866 it is at most 10^{-12}. A relative error carries no information there. On the bags whose unrolled gradient exceeds 10^{-12}, the restriction of the twenty-seed analysis, the fraction is at most .004. That restriction was added to this run's summary after its first fold had run, and it is not part of the registered rule. Under the same restriction the signed gate under the kernel has .33 to .60 on four datasets. On UCSB the restriction keeps 20 of 290 bags, so we do not quote that value.

H.7 MUSK1

MUSK1 (Dietterich et al., 1997) was the probe of our first registration. It is not in the benchmark of the adaptive network, of Ramsauer et al. (2021) or of Hu et al. (2023), so Table 17 reports their four datasets and this subsection reports MUSK1. The multiple-instance ranges of Tables 2 and 3 in the body include MUSK1. Table 27 is the column that Table 17 would carry for MUSK1, produced by the same code over the same twenty seeds. Figure 11 is the MUSK1 column of Figure 9. The gap in the dynamics is the one the other four datasets show. On MUSK1 the fifteen confirmatory seeds confirm the gaps in convergence and in cycling, and the gaps in gradient error and in the uphill fraction against the distance form (Table 18). The bold cells of Table 27 mark the first three of these gaps. The costs are the largest we measured. Ours needs 148 iterations against 13 to 52 on the other four, the condition number is 14.7, and ours drifts as much as the adaptive network.

Table 26: The signed gate and the monotone gate under three separations, five datasets, five seeds. Columns as in Table 25, and "Bad grad. (all bags)" is the unrestricted fraction of all bags with implicit-gradient error above 0.1, as in Table 24. For the monotone gate under the kernel that column counts bags whose unrolled gradient vanishes. Under sparsemax and 1.5-entmax most bags have both gradients exactly zero, and the column counts them as agreeing. For the signed gate these include 1747 of the 1921 bags that do not converge under sparsemax and 1467 of the 1665 under 1.5-entmax, so under those two separations the column does not see the cycles.

Dataset	Gate	Separation	Converged	Cycles	On an instance	Bad grad. (all bags)	Accuracy @1
MUSK1	signed	sparsemax	.600	.400	.59	.017	.802 ± .039
	signed	1.5-entmax	.580	.420	.57	.041	.796 ± .038
	signed	softmax, fixed W	.520	.478	.50	.361	.776 ± .018
	monotone	sparsemax	1.000	0	.98	0	.815 ± .013
	monotone	1.5-entmax	1.000	0	.94	0	.800 ± .018
	monotone	softmax, fixed W	1.000	0	.70	.343	.815 ± .030
Fox	signed	sparsemax	.481	.519	.47	.027	.576 ± .034
	signed	1.5-entmax	.640	.358	.60	.025	.571 ± .015
	signed	softmax, fixed W	.482	.518	.45	.497	.576 ± .036
	monotone	sparsemax	.999	0	.96	.001	.581 ± .033
	monotone	1.5-entmax	1.000	0	.86	0	.589 ± .032
	monotone	softmax, fixed W	1.000	0	.81	.408	.581 ± .025
Tiger	signed	sparsemax	.452	.548	.43	.087	.730 ± .011
	signed	1.5-entmax	.487	.511	.47	.060	.748 ± .028
	signed	softmax, fixed W	.454	.545	.40	.515	.731 ± .030
	monotone	sparsemax	.997	0	.74	.003	.763 ± .016
	monotone	1.5-entmax	1.000	0	.62	.001	.775 ± .022
	monotone	softmax, fixed W	1.000	0	.42	.194	.783 ± .032
Elephant	signed	sparsemax	.413	.587	.38	.079	.779 ± .017
	signed	1.5-entmax	.540	.455	.49	.124	.765 ± .019
	signed	softmax, fixed W	.604	.396	.47	.355	.763 ± .027
	monotone	sparsemax	.996	0	.76	.004	.811 ± .032
	monotone	1.5-entmax	.995	0	.64	.006	.802 ± .028
	monotone	softmax, fixed W	1.000	0	.53	.100	.798 ± .017
UCSB	signed	sparsemax	.714	.286	.71	.003	.593 ± .055
	signed	1.5-entmax	.521	.479	.52	.024	.624 ± .054
	signed	softmax, fixed W	.276	.724	.28	.697	.597 ± .040
	monotone	sparsemax	1.000	0	1.00	0	.645 ± .057
	monotone	1.5-entmax	1.000	0	1.00	0	.638 ± .052
	monotone	softmax, fixed W	1.000	0	.76	.031	.752 ± .070

Table 27: MUSK1, the column that Table 17 would carry. Same statistics, same code, same twenty seeds and the same bolding rule, with 1840 bags.

Metric	Ada.	ours
converged	.483 ± .280	**.999 ± .002**
cycles	.516 ± .280	**0**
bad grad.	.475	**.000**
bags kept	432	1164
depth drift	.0179	.0168
iterations	2.1	148.1
$\mathrm{cond}(\mathbf{I} - J_T)$	1.00	14.7

H.8 Implicit training

Does the disagreement stop training? It does not stop the layer from training. The implicit-training experiment uses the setup of Table 17 for 150 epochs on two of the five folds, with an exact dense solve of $(\mathbf{I} - J_T)^\top \mathbf{v} = \partial L / \partial \mathbf{x}^*$ per bag in the backward pass. It trains all three models on Fox, Tiger, Elephant and MUSK1 over five seeds. The registered rule counts a mean accuracy loss of at least 0.05, a non-finite loss in a fifth of the fits or a rising loss in half of them as a failure. Under that rule every model trains on every dataset. The adaptive network and ours reach their own unrolled-trained accuracy to within one seed standard deviation on every dataset, with no divergence and a falling loss in every fit. This describes the data and is not a test of equivalence. The modern Hopfield network does the same on three datasets and loses 4.5 points on Fox, 1.6 seed standard deviations, inside the pre-registered tolerance.

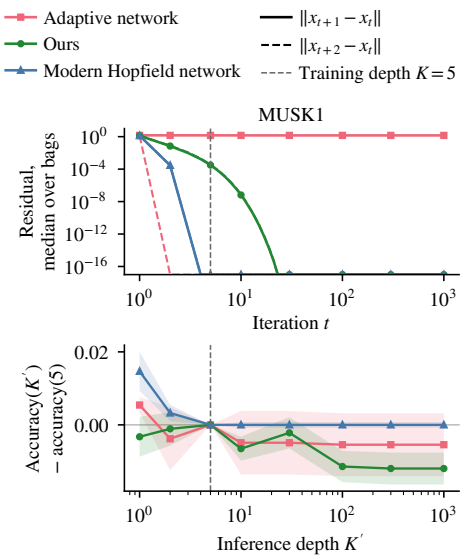

Figure 11: The MUSK1 column of Figure 9, drawn in the same way. Top: the adaptive network is in a period-2 orbit on this fold. Bottom: mean and standard error over twenty seeds. Ours still moves at $K' = 100$, the drop that the depth-drift row of Table 27 measures per seed.

The adaptive network's forward pass still fails to reach a fixed point on Fox and Tiger while that happens, at median residuals of 1.7 and 1.5, against $2 \cdot 10^{-5}$ to $8 \cdot 10^{-4}$ for the modern Hopfield network and $4 \cdot 10^{-3}$ to $8 \cdot 10^{-3}$ for ours. On MUSK1 and Elephant its median residual is $2 \cdot 10^{-6}$ and $1 \cdot 10^{-2}$. A later registration adds UCSB, the fourth dataset of Table 17, under the same protocol. All three models train there by the same rule. The mean change in accuracy is $+0.03$ for the adaptive network, $+0.06$ for ours and $+0.01$ for the modern Hopfield network. Each is inside one seed standard deviation, which is 0.12 to 0.16 because a seed holds 25 test bags and one bag is 0.04 of accuracy. The adaptive network's median forward residual on UCSB is 1.8, against 10^{-8} for the modern Hopfield network and $4 \cdot 10^{-3}$ for ours. The implicit function theorem is applied where its hypothesis fails, and momentum-based optimisation absorbs the resulting gradient error. Our claim therefore concerns the fidelity of the gradient when the forward pass iterates the map. It is not a claim that the adaptive network cannot be trained.

H.9 Forward solvers

Setup. The forward-solver registration asks whether the root-finders of deep-equilibrium practice reach a fixed point in the bags where plain iteration cycles. It refits the adaptive network, ours and the modern Hopfield network as trained for Table 17, and the released configuration of Table 3, on twenty seeds and five folds. Each solver starts from the layer's initial state and works in double precision, with a budget of 1000 map evaluations per bag, as in Table 17. Anderson acceleration uses memory 5, full mixing and a Tikhonov term of 10^{-10} relative to the residuals. Broyden's method starts from the plain step and takes full steps. A bag counts as reached when the residual $\|T(\mathbf{x}) - \mathbf{x}\|$ at the returned point is below 10^{-8}. A reach rate below is the mean, over the seeds that have a cycling bag, of each seed's share of its cycling bags. Those are 18 to 20 seeds for the adaptive network and 14 to 20 for its released configuration.

Reaching the fixed points. In the adaptive network's cycling bags, Broyden's method reaches a fixed point in 0.81, 0.81, 0.86, 0.51 and 0.50 of bags on Tiger, Fox, Elephant, UCSB and MUSK1 (Figure 12a). Anderson acceleration reaches one in 0.43, 0.36, 0.52, 0 and 0.49. Its registered prediction of at least half on every dataset therefore fails on four of the five. In the released configuration Broyden's method reaches 0.88 to 1.00 and Anderson acceleration 0.32 to 1.00. There the prediction fails on UCSB alone. We added three more settings after the registration. They ran on the first five seeds and are exploratory. In the adaptive network's cycling bags, a Tikhonov term 10^4

times larger adds at most 0.093 to Anderson's rate, and a mixing of one half at most 0.035. With a backtracking line search on the residual norm, Broyden's method reaches 0.91 to 0.97 of these bags on Tiger, Fox, Elephant and MUSK1, and 0.22 on UCSB. At every point these settings find there, the spectral radius exceeds one.

The points found. Over the adaptive network's cycling bags of all twenty seeds, the points that Anderson and Broyden find have an almost-everywhere Jacobian of spectral radius 30 to 684 in the median. It exceeds one at 99.2% or more of them (Figure 12b), so plain iteration is unstable at those points. The eigenvalue of largest modulus is real and negative at 94.5% or more of them. The spectral radius is below one at 6, 2 and 17 points on Tiger, Fox and Elephant, and at none on UCSB and MUSK1. By the Jacobian test such a point would attract within its sorting cell. No iteration was run from these points to confirm it. If one does attract, a cycle and an attracting fixed point coexist in its bag. Where plain iteration ends there then depends on its start. The matrix $\mathbf{I} - J_T$ that an implicit backward pass inverts has median condition number 161 to 1980 at the points found, and none above 10^6. The label at a point found equals the label at the last plain iterate $\mathbf{x}_{1000}$ in 0.72 to 0.89 of cases, and the label at its image $T(\mathbf{x}_{1000})$ in 0.58 to 0.88. The two phases of the orbit carry different labels in 0.22 to 0.61 of those cases.

An energy safeguard. For the two models with an energy, ours and the modern Hopfield network, the registration also runs Anderson acceleration with an energy safeguard. It keeps an accelerated candidate only when the energy falls by at least the amount that a plain step guarantees, and otherwise it takes the plain step. The energy therefore cannot rise in exact arithmetic. Along every trajectory its rise relative to $\max(1, |E|)$ was at most $4.5 \cdot 10^{-16}$, which is rounding error. Without the safeguard it rose in 23 to 64% of our bags. For our map the safeguarded solver needs fewer evaluations to 10^{-6} than plain iteration, by a factor of 1.37, 1.56, 1.28, 2.25 and 5.13 on the five datasets, the geometric mean over seeds (Figure 12c). The gain is confirmed on all five datasets over the fifteen confirmatory seeds by an exact sign-flip test, with every Holm-adjusted p equal to 0.0003. This count measures map evaluations per bag, not time. The cost paragraph (Appendix H.10) therefore also times the solves. For the modern Hopfield network, which plain iteration solves in four to seven evaluations, the factor is 0.71 to 0.98.

Where both reach 10^{-8}, the safeguarded solver ends with the label of plain iteration in at least 99.78% of our bags. It ends more than 10^{-6} from the end point of plain iteration in 16 to 43% of them. Its registered prediction of the same point in at least 95% of bags therefore fails on all five datasets. The distance between the two end points has median $7 \cdot 10^{-9}$ to $3 \cdot 10^{-7}$ and 90th percentile $3 \cdot 10^{-5}$ to $6 \cdot 10^{-3}$, and its largest value is 1.2. The safeguard makes the residuals tend to zero. It does not select the fixed point, if any, that the iterates approach.

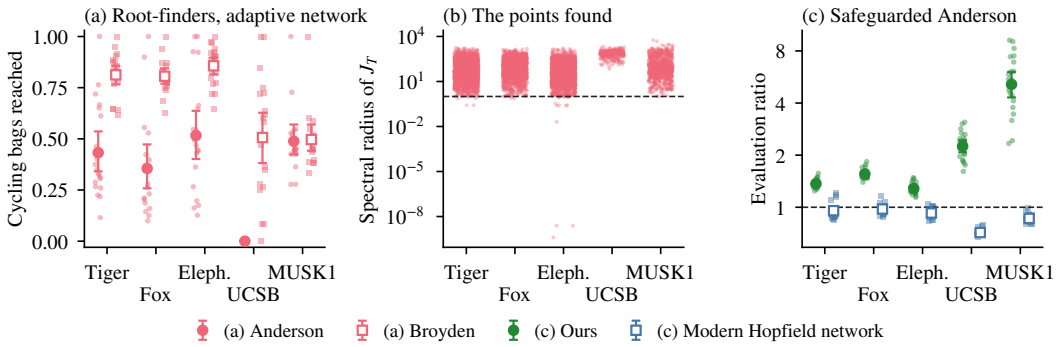

Figure 12: Forward solvers on the multiple-instance maps, twenty seeds. (a) The share of the adaptive network's cycling held-out bags in which Anderson acceleration (memory 5) and Broyden's method reach a residual below 10^{-8} within 1000 map evaluations from the layer's initial state. Faint dots are seeds, and markers the mean over seeds with its 95% percentile bootstrap interval. (b) The spectral radius of the almost-everywhere Jacobian of the adaptive network's map at every point that Anderson or Broyden finds in a cycling bag, on a log scale. Above the dashed line plain iteration is unstable at the point. (c) For our map and the modern Hopfield network, the evaluations that plain iteration needs to reach 10^{-6} divided by those of energy-safeguarded Anderson acceleration, the geometric mean over a seed's bags. Faint dots are seeds, and markers the geometric mean over seeds with its 95% percentile bootstrap interval.

H.10 Cost and scope in full

Where the cost comes from. The constrained model pays in iterations (Table 13), in iteration count and conditioning at the final iterate (Table 17), and in accuracy at mask 0.8 on MNIST (Appendix G.5), where the two models also differ in the gate's parameterisation and weight decay. In wall-clock terms one application of our map costs about as much as one of the adaptive network's on MUSK1, and about half on the synthetic task. The medians of five timings, with one thread requested and untrained parameters, are 0.088 against 0.086 ms per bag-step on MUSK1, with all 92 bags padded to 40 instances, and 0.19 against 0.36 ms per query on 400 synthetic queries at mask 0.6. The cost is therefore almost entirely the iteration count.

Per query and per bag. At the budget where the two tie, 30 steps for ours against one for the adaptive network (Table 13), ours costs about 16 times more per query. It is about 70 times more per bag on MUSK1, our worst case, at the twenty-seed iteration counts of Table 27, where the adaptive network's count is a median over the bags that reach the tolerance. The modern Hopfield network, which needs no sort, takes 0.0029 ms per bag-step and 0.013 ms per query at 30 steps. Against it the ratio is about 430 per query at 30 steps each, and about 1100 per bag at the twenty-seed iteration counts on MUSK1 of 148 for ours and 4.2 for it.

Accelerated solves. Anderson acceleration of our forward pass with an energy safeguard cuts its evaluations by a factor of 1.3 to 5.1 (Appendix H.9). The energy rises along it by no more than rounding error. We timed the forward solves of ours and of the adaptive network on the held-out bags of one fold of five fits per dataset, each bag alone on one thread, to a residual of 10^{-6}. On MUSK1 the safeguarded solver makes our solve 4.9 times faster in total than plain iteration, and on UCSB 2.1 times. On Tiger, Fox and Elephant, where plain iteration needs a median of 12 to 15 evaluations, it runs at 0.6 to 0.7 times the speed of plain iteration. On MUSK1 our median bag costs 90 times the adaptive network's median converged bag by plain iteration, against about 70 at the twenty-seed counts. With the safeguarded solver it costs 9 times. Our score under the pattern readout reaches in one step at least the accuracy of 30 gradient steps (Table 13), at 0.18 ms per query, but it has no energy guarantee.

The backward solve. Implicitly trained fits pay in the backward solve. We measured the Jacobian spectrum of these fits on one fold of MUSK1, Fox, Tiger and Elephant. There the spectral radius

of J_T has median 0.95 for our map. A fixed-point solve of the backward system therefore contracts slowly in the typical bag of those fits, unlike the unrolled-trained fits of Table 17. An Anderson solve of the backward system, with memory 5, reaches a relative error below 10^{-2} within 1000 iterations in 35 per cent of our fits, against 80 and 90 per cent for the other two models ($n = 20$ fits per model, four datasets by five seeds, descriptive).

Scope. Theorem 1(iii) is a statement for generic $\mathbf{\Phi}$. Theorem 2 is stated for the squared-Euclidean base and gives descent and a lower bound, not that the minimisers are the memories at every β (Appendix G.3).