On the Structural Impossibility of Attractor-Agnostic Incremental Concept Learning in Dense Weight Architectures

A Formal Analysis via Spectral Perturbation Theory

Andreas Bean
Independent Researcher
andreas.bean@beanbox.at

April 2026 (v2)

Abstract

We prove that safe incremental learning of structurally new concepts in any dense associative memory is structurally impossible—not for algorithmic reasons, but for geometric ones rooted in spectral perturbation theory.

The core result (Theorem 2) shows that in any associative memory with symmetric weight matrix W, no perturbation ΔW—of any rank or produced by any optimisation method—can introduce a new attractor a^* without necessarily changing the eigenstructure of W: either existing eigenvectors are rotated or existing eigenvalues shift. In either case, existing stored attractors are at risk. The result is purely structural: it follows from the symmetry of ΔW and the completeness of the eigenbasis, independent of any specific learning algorithm or architecture beyond the requirement that W is a dense real matrix.

Beyond attractor survival, we establish a deeper consequence: the introduction of a new concept necessarily disrupts the pairwise semantic relations between all existing concepts in truncated spectral space (Proposition 7). Knowing the existing attractor positions does not resolve the difficulty—it merely makes the geometric obstruction explicit (Section 5.2).

Catastrophic forgetting, in its deepest form, is not an engineering deficiency. It is a structural consequence of implicit topology. The biological connectome avoids this limitation through *explicit sparse topology*: a new synapse (r, s) perturbs only four entries of the Laplacian L, leaving concepts whose representations are distant from (r, s) structurally unaffected (Section 6).

Contents

1 Introduction

The problem of catastrophic forgetting has been studied for over three decades [5, 6]. The standard framing treats it as an engineering problem: the network overwrites old weights when exposed to new data, and the remedy is better regularisation or rehearsal.

We argue that this framing is incomplete for an important class of learning problems. There is a categorical distinction between:

1. *New facts about existing concepts*: associating a known concept with new information. This requires only weight modification within the existing eigenspace and is, in principle, tractable.
2. *Structurally new concepts*: concepts whose attractor structure does not yet exist in the eigenspace of the weight matrix. Introducing such an attractor provably requires global eigenstructure reorganisation—no local operation suffices.

The second case is not an engineering problem. It is a structural one. For structurally new concepts in dense weight architectures, any weight change that introduces the new attractor necessarily modifies the eigenbasis of W, thereby displacing the spectral representations of *all* existing concepts and disrupting *all* pairwise semantic relations between them.

This paper provides:

(i) A formal proof (Theorem 2) that any symmetric ΔW introducing a new concept a^* must change the eigenstructure of W.

(ii) A spectral survival condition (Proposition 5) quantifying when existing attractors survive.

(iii) A new result (Proposition 7) showing that even when attractors survive as fixed points, their pairwise semantic relations in truncated spectral space are disrupted.

(iv) An analysis of the semantic placement problem (Section 5.2): why knowing the existing attractor positions does not make safe concept addition tractable.

(v) A comparison with the biological connectome, where explicit sparse topology enables local insertability (Section 6).

The broader wave-dynamic framework from which this analysis emerged is developed in [1].

2 Background and Related Work

2.1 Hopfield Networks and Associative Memory

The classical Hopfield network [3] stores patterns $\xi^1, \ldots, \xi^p \in \{-1, +1\}^n$ as attractors of the energy function $E(s) = -\frac{1}{2} s^\top W s$, where the weight matrix is constructed by Hebbian learning: $W = \frac{1}{n} \sum_k \xi^k (\xi^k)^\top$.

The eigenstructure of this Hebbian matrix plays a central role: stored patterns are approximate eigenvectors of W (Lemma 1), and the stability of each pattern as an attractor is determined by the spectral gap around its corresponding eigenvalue (Proposition 5).

2.2 Catastrophic Forgetting

McCloskey and Cohen [5] identified catastrophic interference as a fundamental problem in connectionist models. Subsequent work has proposed many mitigation strategies: Elastic Weight Consolidation (EWC) [4] and Progressive Neural Networks [7], among others.

None of these approaches address the structural question: *can any weight change—local or global—introduce a qualitatively new attractor without disturbing existing ones, even in principle?* We show the answer is no.

The present work differs from the continual learning literature in that we do not propose a new algorithm or regularisation scheme. We prove a structural lower bound: no algorithm can guarantee safe concept addition in a dense weight architecture. This reframes catastrophic forgetting as a geometric problem, not an optimisation one.

2.3 Spectral Perturbation Theory

The Davis-Kahan theorem [2] provides quantitative bounds on eigenvector perturbations: for symmetric matrices W and perturbation ΔW,

$$\sin \angle(e_i, \tilde{e}_i) \leq \frac{\|\Delta W\|_2}{\delta_i},$$

where $\delta_i = \min_{j \neq i} |\lambda_i - \lambda_j|$ is the spectral gap. Our contribution goes further: we show that for introducing a new attractor, the shift is not merely possible but structurally necessary—it cannot be avoided by any choice of ΔW.

3 Formal Framework

Definition 1 (Associative Memory)**.** *Let $W \in \mathbb{R}^{n \times n}$ be symmetric with $W_{ii} = 0$. The energy function is $E(s) = -\frac{1}{2} s^\top W s$ for $s \in \{-1, +1\}^n$. The asynchronous update rule is $s_i \leftarrow \operatorname{sign}\Big(\sum_j W_{ij} s_j\Big)$.*

Definition 2 (Attractor)**.** *A pattern $\xi \in \{-1, +1\}^n$ is an attractor of W if and only if $\operatorname{sign}(W\xi) = \xi$, i.e., ξ is a fixed point of the update rule.*

Definition 3 (Structurally New Concept)**.** *A vector $a^* \in \mathbb{R}^n$ is a* structurally new concept *with respect to trained network W if $\langle a^*, \xi^k \rangle \approx 0$ for all stored attractors $\xi^1, \ldots, \xi^p$. That is, a^* lies outside the span of the stored patterns.*

Definition 4 (Spectral Semantic Representation)**.** *Let $W = \sum_i \lambda_i e_i e_i^\top$ be the eigendecomposition of W, with $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, and let $\Phi_K = [e_1, \ldots, e_K]$ denote the K lowest eigenvectors (low-frequency modes).*

The spectral semantic representation *of pattern ξ^m relative to W is:*

$$c_m(W) \;:=\; \Phi_K^\top \xi^m \;\in\; \mathbb{R}^K.$$

The spectral semantic relation *between concepts m and k is:*

$$\rho_{mk}(W) \;:=\; \cos\big(c_m(W),\, c_k(W)\big) \;=\; \frac{(\xi^m)^\top \Phi_K \Phi_K^\top \xi^k}{\|\Phi_K^\top \xi^m\|\, \|\Phi_K^\top \xi^k\|}.$$

This quantity depends on W through $\Phi_K = \Phi_K(W)$.

Remark 1. *The full-eigenspace case $K = n$ gives $\Phi_n \Phi_n^\top = I_n$, so $\rho_{mk}(W) = \cos(\xi^m, \xi^k)$ independent of W—a trivially constant quantity. The semantically meaningful case is the* truncated *representation with $K \ll n$, which captures the low-frequency (global, abstract) component of each concept's pattern on the eigenbasis of W.*

In the wave-dynamic framework of [1], the FieldReader operates exactly as $c_m(W) = \Phi_K^\top \psi^*(T)$: the K-dimensional spectral summary of the steady-state field is the computational representation of meaning.

4 Main Results

4.1 Attractor Approximation

Lemma 1 (Attractors Approximate Eigenvectors)**.** *Let $W = \frac{1}{n}\sum_{k=1}^{p} \xi^k(\xi^k)^\top$ be the Hebbian weight matrix for patterns $\xi^1, \ldots, \xi^p$. Then for each stored pattern ξ^m:*

$$W\xi^m = \xi^m + \frac{1}{n}\sum_{k\neq m} \langle \xi^k, \xi^m \rangle \xi^k.$$

When $p \ll n$ and patterns are approximately orthogonal, the cross-talk term is small, and ξ^m is an approximate eigenvector of W with eigenvalue ≈ 1.

Proof. Direct computation: $W\xi^m = \frac{1}{n}\sum_k \xi^k(\xi^k)^\top \xi^m = \frac{1}{n}\sum_k \langle \xi^k, \xi^m\rangle \xi^k = \xi^m + \frac{1}{n}\sum_{k\neq m}\langle \xi^k, \xi^m\rangle \xi^k$, using $\langle \xi^m, \xi^m\rangle = n$. When $p \ll n$, patterns are approximately orthogonal. Concretely, for $\xi^k \in \{-1, +1\}^n$ chosen independently and uniformly, each cross-term $\langle \xi^k, \xi^m\rangle$ has mean zero and standard deviation $\sqrt{n}$. The residual vector $\frac{1}{n}\sum_{k\neq m}\langle \xi^k, \xi^m\rangle \xi^k$ has $p-1$ terms each of expected squared norm $n/n^2 = 1/n$, so the residual norm is $O(\sqrt{p/n})$, vanishing as $n \to \infty$ at fixed p [3]. □

4.2 The Impossibility Theorem

Theorem 2 (New Concept Requires Global Eigenstructure Change)**.** *Let $W \in \mathbb{R}^{n\times n}$ be symmetric with eigendecomposition $W = \sum_i \lambda_i e_i e_i^\top$, where $\{e_i\}$ is an orthonormal basis of $\mathbb{R}^n$. Let $a^* \in \mathbb{R}^n$ be a structurally new concept with $\langle a^*, e_i\rangle \approx 0$ for $i = 1, \ldots, p$ (orthogonal to all stored attractors), and assume a^* is not already an eigenvector of W (i.e., $Wa^* \neq \mu a^*$ for any μ).*

Then for any symmetric perturbation ΔW that makes a^* an attractor of $W' = W + \Delta W$ (i.e. $W'a^* = \mu a^*$ for some $\mu > 0$), the eigenstructure of W' necessarily changes relative to W:

- *(a) at least one existing eigenvector e_j is no longer an eigenvector of W' (eigenvector rotation), or*
- *(b) at least one eigenvalue shifts ($\lambda'_j \neq \lambda_j$ for some j).*

The approximate case $W'a^* \approx \mu a^*$ follows by continuity of eigenvalues and eigenvectors under small perturbations (Weyl's inequality).

Proof. We work in the exact regime $W'a^* = \mu a^*$; the approximate case follows by continuity (Weyl's inequality). The proof splits according to whether the eigenvectors of W are preserved as eigenvectors of W'.

Case 1: Some eigenvector e_j is no longer an eigenvector of W'.

If $W'e_j \neq \lambda' e_j$ for any λ', then e_j has been rotated in W'; case (a) holds.

Case 2: All $e_1, \ldots, e_n$ remain eigenvectors of W' (eigenvalues must shift).

Suppose $W'e_j = (\lambda_j + c_j)e_j$ for all j, so $\Delta W \cdot e_j = c_j e_j$.

Step A (Residual vector). Since $W'a^* = \mu a^*$:

$$\Delta W \cdot a^* = \mu a^* - Wa^* = \sum_i (\mu - \lambda_i)\langle e_i, a^*\rangle e_i =: r.$$

Since a^* is not an eigenvector of W, at least one summand $(\mu - \lambda_j)\langle e_j, a^*\rangle \neq 0$, so $r \neq 0$.

Step B (Compatibility via symmetry). For any j, compute $\langle e_j, \Delta W \cdot a^* \rangle$ two ways. From Step A it equals $(\mu - \lambda_j)\langle e_j, a^* \rangle$. From the Case 2 assumption and symmetry of ΔW: $\langle e_j, \Delta W \cdot a^* \rangle = \langle \Delta W e_j, a^* \rangle = c_j \langle e_j, a^* \rangle$. Equating:

$$c_j \langle e_j, a^* \rangle = (\mu - \lambda_j)\langle e_j, a^* \rangle \quad \text{for all } j. \tag{$*$}$$

Step C (Eigenvalue shift). Since $a^* \neq 0$ and $\{e_j\}$ is a complete basis, there is at least one j with $\langle e_j, a^* \rangle \neq 0$. Equation $(*)$ gives $c_j = \mu - \lambda_j$ for *every* such j. Suppose $c_j = 0$ for every j with $\langle e_j, a^* \rangle \neq 0$; then $\lambda_j = \mu$ for all those j, and therefore

$$W a^* = \sum_j \lambda_j \langle e_j, a^* \rangle e_j = \mu \sum_j \langle e_j, a^* \rangle e_j = \mu a^*,$$

contradicting the hypothesis that a^* is not an eigenvector of W. Hence $c_j \neq 0$ for at least one j: at least one eigenvalue shifts.

In both cases the eigenstructure of W' necessarily differs from that of W. This holds for any symmetric ΔW, regardless of rank or optimisation method. □ □

Remark 2 (Formal Verification in Lean 4 and Isabelle/HOL)**.** *The full Case 2 argument has been independently machine-verified using two proof assistants:*

- ***Lean 4 (version 4.29.0)**, without external libraries.*
- ***Isabelle 2025-2 / HOL**, via the Isabelle build system (`HOL-Analysis` session heap);* zero `sorry` *or unverified axiom in either theory file.*

Two complementary Isabelle theory files cover the complete argument:

`EigenvalueShift.thy` **(algebraic core).** *Five fully machine-checked results for Theorem 2:*

- `eigenvalue_shift_core`*: if $\lambda_i \neq \lambda_j$ and both are forced to μ under W', then at least one eigenvalue shifts.*
- `eigenvalue_shift_necessary`*: the existential wrapper of the above.*
- `cancel_active_component`*: the integral-domain cancellation $a \cdot b = c \cdot a \wedge a \neq 0 \Rightarrow b = c$ (Step C).*
- `diagonal_eigenvector_equation`*: component-wise cancellation $\lambda'_j \alpha_j = \mu \alpha_j \wedge \alpha_j \neq 0 \Rightarrow \lambda'_j = \mu$.*
- `full_theorem2_case2`*: the complete Case 2 statement with* `h_eigvec` *taken as hypothesis.*

`FullTheorem2.thy` **(end-to-end proof including Steps A+B).** *Built on the* `HOL-Analysis` *session heap; proves the full theorem from first principles without any hypothesis gap:*

- `step_AB`*: from an orthonormal eigenbasis $\{e_j\}$, the expansion $a^* = \sum_j \alpha_j e_j$, and $W' a^* = \mu a^*$ (Steps A+B), it follows that $\lambda'_j \alpha_j = \mu \alpha_j$ for all j—formally derived using* `linear_sum`*,* `linear_scale`*,* `inner_sum_right`*,* `inner_scaleR_right`*, and* `sum.delta'`*.*
- `full_theorem2_end_to_end`*: the complete Case 2 conclusion $\exists k.\ \lambda'_k \neq \lambda_k$ derived directly from the geometric hypotheses, with no assumed lemmas.*

Remark 3 (From Eigenstructure Change to Attractor Disruption)**.** *Theorem 2 establishes a* necessary *structural change in W'. This does not by itself guarantee that a stored attractor is destroyed—a small perturbation may rotate an eigenvector by very little. The quantitative threshold is given by Proposition 5: attractor ξ^m is at risk whenever the required eigenvalue change $|\mu - \lambda_m|$ meets or exceeds the spectral gap δ_m. The full impossibility argument combines*

(i) any introduction of a^* forces an eigenstructure change (Theorem 2) with (ii) that change disrupts ξ^m whenever $|\mu - \lambda_m| \geq \delta_m$ (Proposition 5).

Corollary 3 (No Algorithm Circumvents the Result)**.** *Theorem 2 applies to rank-1, rank-k, and full-rank perturbations. The result is structural, not algorithmic: no choice of learning rate, regulariser, or optimiser can avoid the eigenstructure change when introducing a genuinely new attractor.*

Corollary 4 (Two Qualitatively Different Learning Cases)**.**

1. ***New facts about existing concepts**: If the new information lies within the current eigenspace (a^* already spanned by $\{e_i\}_{i \leq p}$), learning is tractable in principle. Catastrophic forgetting remains a risk due to the global nature of ΔW, but is not structurally necessary.*
2. ***Structurally new concepts**: If a^* is orthogonal to all existing eigenvectors, its introduction necessarily disrupts the eigenstructure of W. By Proposition 5, at least one existing attractor is at risk. This case is structurally unavoidable without full retraining.*

4.3 Quantitative Bounds

Proposition 5 (Survival Condition for Existing Attractors)**.** *Under the conditions of Theorem 2, let $\delta_m = \min_{j \neq m} |\lambda_m - \lambda_j|$ be the spectral gap of attractor ξ^m. A necessary condition for ξ^m to survive the introduction of new concept a^* with target eigenvalue μ is:*

$$|\mu - \lambda_m| < \delta_m.$$

If this condition is violated, any perturbation ΔW introducing a^* has $\|\Delta W\|_2 \geq |\mu - \lambda_m| \geq \delta_m$, beyond the Davis-Kahan safe radius for e_m.

Proof. We establish a case-independent lower bound on $\|\Delta W\|_2$, then apply the Davis-Kahan theorem.

Lower bound (both cases). For any matrix M and vector $v \neq 0$, $\|M\|_2 \geq \|Mv\|/\|v\|$. Applying this with $v = a^*$ and using Step A of Theorem 2:

$$\|\Delta W\|_2 \;\geq\; \frac{\|\Delta W a^*\|}{\|a^*\|} \;=\; \frac{\|\mu a^* - W a^*\|}{\|a^*\|} \;=\; \frac{1}{\|a^*\|}\Big(\sum_i (\mu - \lambda_i)^2 \langle e_i, a^* \rangle^2\Big)^{1/2}.$$

The term $i = m$ alone contributes $|\mu - \lambda_m|\,|\langle e_m, a^* \rangle|/\|a^*\|$, so $\|\Delta W\|_2 \geq |\mu - \lambda_m|\,|\langle e_m, a^* \rangle|/\|a^*\|$. This bound holds regardless of whether case (a) or case (b) of Theorem 2 applies.

Survival condition via Davis-Kahan. For $\xi^m \approx e_m$ to remain a stable attractor of W', it suffices that the m-th eigenspace of W' stays close to e_m. By the Davis-Kahan theorem [2], this holds when $\|\Delta W\|_2 < \delta_m$. Combining with the lower bound:

$$|\mu - \lambda_m| \, \frac{|\langle e_m, a^* \rangle|}{\|a^*\|} < \delta_m.$$

Since a^* is a structurally new concept with non-trivial semantic content in direction e_m, we have $|\langle e_m, a^* \rangle|/\|a^*\| > 0$, and the stated necessary condition $|\mu - \lambda_m| < \delta_m$ is the tightest form of this bound in the case $|\langle e_m, a^* \rangle| = \|a^*\|$ (i.e. $a^* \parallel e_m$, the maximally sensitive configuration). □

Remark 4. *Proposition 5 predicts that concepts with dense semantic neighbourhoods (small δ_m) are most vulnerable: any new concept whose target eigenvalue is farther than δ_m from λ_m risks destroying ξ^m. Concepts in sparse semantic regions (large δ_m) are robust. This is consistent with empirical observations that training disrupts common, densely-represented concepts more than rare ones.*

Proposition 6 (Geometric Threshold: Case 1 vs. Case 2)**.** *Let $P_p = \sum_{i=1}^{p} e_i e_i^\top$ be the orthogonal projection onto the eigenspace of the p stored concepts, and define the* novelty angle*:*

$$\cos\theta(a^*) = \frac{\|P_p\, a^*\|}{\|a^*\|} \in [0, 1].$$

(Heuristic; formal proof of sharpness is an open problem.) *The natural operational threshold is $\cos\theta(a^*) = 1/\sqrt{p}$:*

- *$\cos\theta > 1/\sqrt{p}$: the projection of a^* onto the existing eigenspace exceeds the expected cross-talk of a random unit vector. The new concept is substantially representable within the current eigenspace:* ***Case 1*** *(new fact).*
- *$\cos\theta < 1/\sqrt{p}$: a^* lies effectively outside the current eigenspace. By Corollary 4, its introduction requires a global eigenstructure change:* ***Case 2*** *(structurally new concept).*

5 Relational Disruption and the Placement Problem

Theorem 2 establishes that introducing a new concept requires changing the eigenstructure of W. We now show that this change has a further consequence beyond attractor survival: it disrupts the pairwise semantic relations between *all* existing concepts in truncated spectral space.

5.1 Pairwise Semantic Relations under Dense Perturbation

Proposition 7 (Relational Disruption under Eigenvector Rotation)**.** *Let W store concepts $\xi^1, \ldots, \xi^p$ and let ΔW introduce new concept a^* as in Theorem 2, and suppose case (a) of Theorem 2 holds: at least one eigenvector e_j ($j \le K$) is rotated, i.e. $e'_j \neq e_j$.*

For any pair (m, k) with $\langle e_j, \xi^m\rangle \neq 0$ or $\langle e_j, \xi^k\rangle \neq 0$:

$$\rho_{mk}(W') \;\neq\; \rho_{mk}(W).$$

In a dense weight matrix W with generic stored patterns, all eigenvectors have non-zero overlap with all patterns ($\langle e_j, \xi^m\rangle \neq 0$ for all j, m almost surely). Therefore the semantic relation ρ_{mk} changes for all *pairs (m, k), not only those directly involving the new concept.*

Remark 5 (Case (b): Relational Disruption via Mode Reordering)**.** *In case (b) of Theorem 2, eigenvectors are unchanged ($\Phi'_K = \Phi_K$), so $\rho_{mk}(W') = \rho_{mk}(W)$ for all existing pairs—no direct relational disruption occurs as long as the set of K lowest eigenmodes is unchanged.*

Relational disruption in case (b) arises if the eigenvalue shift $c_j = \mu - \lambda_j$ causes the j-th mode to leave the top-K set: specifically, if $\lambda'_j = \mu$ crosses λ_{K+1}, the mode formerly ranked j is replaced by e_{K+1} in Φ'_K. This mode reordering *produces a discontinuous change in Π_K and therefore in all relations ρ_{mk} with non-zero overlap on the affected mode. The condition for reordering is $\mu \ge \lambda_{K+1}$, i.e. the new concept is placed in the high-frequency part of the spectrum. For concepts embedded in the low-frequency semantic regime ($\mu \le \lambda_K$), no reordering occurs and case (b) produces no relational disruption.*

Proof. The spectral projection matrix $\Pi_K = \Phi_K \Phi_K^\top$ changes to $\Pi'_K = \Phi'_K (\Phi'_K)^\top$ under the perturbation. Write $\delta\Pi = \Pi'_K - \Pi_K$. If $e'_j = e_j + \delta e_j$ for some $j \le K$ (with $\|\delta e_j\| = O(\|\Delta W\|_2/\delta_j)$ by Davis-Kahan), then to first order:

$$\delta\Pi \;=\; (\delta e_j) e_j^\top + e_j (\delta e_j)^\top + O(\|\delta e_j\|^2).$$

Write $u_m := \Pi_K \xi^m$ and $u_k := \Pi_K \xi^k$, so that $\rho_{mk}(W) = (u_m^\top u_k)/(\|u_m\|\,\|u_k\|)$. To first order in $\delta\Pi$:

$$u'_m = (\Pi_K + \delta\Pi)\xi^m = u_m + \delta\Pi\,\xi^m =: u_m + \delta u_m,$$
$$u'_k = u_k + \delta u_k,$$

with $\delta u_m = \delta\Pi\,\xi^m$ and $\delta u_k = \delta\Pi\,\xi^k$. The cosine changes to first order as:

$$\rho_{mk}(W') - \rho_{mk}(W) \;=\; \frac{u_m^\top \delta u_k + \delta u_m^\top u_k}{\|u_m\|\,\|u_k\|} - \frac{(u_m^\top u_k)(u_m^\top \delta u_m + u_k^\top \delta u_k)}{\|u_m\|^2\,\|u_k\|^2} \cdot \frac{\|u_m\|\,\|u_k\|}{1} + O(\|\delta\Pi\|^2).$$

The numerator of the leading term is:

$$u_m^\top \delta u_k + \delta u_m^\top u_k \;=\; (\xi^m)^\top \delta\Pi^\top u_k + u_m^\top \delta\Pi\,\xi^k \;=\; (\langle \delta e_j, \xi^m\rangle\langle e_j, \xi^k\rangle + \langle e_j, \xi^m\rangle\langle \delta e_j, \xi^k\rangle)\cdot\|u_m\|\,\|u_k\| + \cdots$$

using $\delta\Pi = (\delta e_j)e_j^\top + e_j(\delta e_j)^\top$ (symmetry of Π_K). This is non-zero whenever $\langle e_j, \xi^m\rangle \neq 0$ or $\langle e_j, \xi^k\rangle \neq 0$, establishing $\rho_{mk}(W') \neq \rho_{mk}(W)$. For generic dense patterns $\xi^m \in \{-1, +1\}^n$, all inner products $\langle e_j, \xi^m\rangle$ are non-zero almost surely (Haar-random eigenvectors have full-support projections), so the relation changes for all pairs (m, k). □ □

Remark 6. *The full-eigenspace relation $\rho_{mk}^{(n)}(W) = \frac{(\xi^m)^\top \xi^k}{\|\xi^m\|\|\xi^k\|}$ is constant in W because $\Phi_n \Phi_n^\top = I_n$. The relational disruption therefore does not affect the raw input-space overlap between stored patterns—it affects the* way those patterns are decoded through the current eigenstructure. *In the wave-dynamic framework, this is the critical quantity: the FieldReader computes meaning as a spectral projection, so a change in Φ_K changes what every concept means, not merely where it sits in state space.*

5.2 The Semantic Placement Problem

Proposition 7 reveals a deeper obstacle than attractor survival alone. Even if one knows the full set of existing attractor positions $\{c_m\}_{m=1}^p$ in spectral space, adding a new concept a^* with desired semantic relations $\rho_{m,a^*}^{\text{target}}$ to all existing concepts is generically impossible in a dense architecture.

Remark 7 (The Placement Problem)**.** *Proposition 7 establishes that any perturbation ΔW introducing a new concept a^* produces a non-zero change $\Delta\rho_{mk}$ for all existing concept pairs (m, k) with overlap on the perturbed eigenvector e_j. This change is governed by the global coupling term*

$$\Delta\rho_{mk} \;\propto\; \langle e_j, \xi^m\rangle\langle e_j, \xi^k\rangle \cdot |\mu - \lambda_j|,$$

which cannot be selectively suppressed for individual pairs: the factor $|\mu - \lambda_j|$ is determined by the target eigenvalue of a^* and is shared across all pairs simultaneously.

This has a precise consequence for the role of attractor knowledge. Even if the full set of existing attractor positions $\{c_m\}_{m=1}^p$ is known—for example via sparse autoencoder extraction—this information does not provide a mechanism to embed a^* without disturbing existing relational geometry. Knowing where the concepts are does not resolve how to place a new one among them: every existing pairwise relation ρ_{mk} imposes a constraint on the position of a^*, and all constraints share the same coupling through the perturbed eigenvector. The geometric obstruction is not informational but structural—a direct consequence of the dense substrate in which all concepts are simultaneously coupled through the shared eigenbasis. Sparse topology resolves this by breaking the global coupling: a new synapse (r, s) perturbs only those eigenmodes with significant amplitude at nodes r and s, leaving the relational geometry among distant concepts structurally intact (Section 6).

6 Sparse Topology as the Structural Solution

The analysis above implies that truly local concept addition requires an architecture in which the weight structure is *sparse* and concepts have *explicit local addresses*. The biological connectome is exactly such a structure.

Let $L = D - A$ be the graph Laplacian of the connectome, with eigenbasis $\Phi = [\varphi_1, \varphi_2, \ldots, \varphi_n]$ and eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. A new Hebbian synapse between neurons r and s with weight Δw introduces the perturbation:

$$\Delta L \;=\; \Delta w\,(e_r - e_s)(e_r - e_s)^\top,$$

which modifies only four entries of L in the nodal basis. First-order eigenvector perturbation theory gives:

$$\Delta\varphi_j \;=\; \Delta w \sum_{k\neq j} \frac{(\varphi_k(r) - \varphi_k(s))(\varphi_j(r) - \varphi_j(s))}{\lambda_j - \lambda_k}\,\varphi_k.$$

Two structural protections emerge immediately.

Low-frequency mode protection. The low-eigenvalue modes φ_j (large spatial wavelength, encoding global abstract semantic structure) are slowly varying: $|\varphi_j(r) - \varphi_j(s)| \approx 0$ for any spatially adjacent pair (r, s). These modes are therefore essentially undisturbed by any single local synapse, regardless of where it is. The global semantic structure of the network—the arrangement of abstract categorical attractors—is topologically robust.

Distant-concept protection (sketch). For a stored concept ξ^m whose neural activation pattern has negligible amplitude at nodes r and s (i.e. $|\xi^m_r|, |\xi^m_s| \approx 0$), the first-order change in the spectral representation $c_m = \Phi_K^\top \xi^m$ follows from the eigenvector perturbation formula above. The ℓ-th component of Δc_m is $\Delta\varphi_\ell^\top \xi^m$, which by the perturbation formula equals:

$$\Delta\varphi_\ell^\top \xi^m \;=\; \Delta w \sum_{k\neq \ell} \frac{(\varphi_k(r) - \varphi_k(s))(\varphi_\ell(r) - \varphi_\ell(s))}{\lambda_\ell - \lambda_k}\,(\varphi_k^\top \xi^m).$$

When $\xi^m_r \approx \xi^m_s \approx 0$, the inner products $\varphi_k^\top \xi^m$ receive negligible contribution from nodes r and s; the dominant factor $(\varphi_k^\top \xi^m)$ is therefore determined by the rest of the network and is not directly suppressed by the localisation condition. *Note:* The argument that $\Delta c_m \approx 0$ is rigorous only for low-frequency modes φ_ℓ (where $|\varphi_\ell(r) - \varphi_\ell(s)| \approx 0$ by the smoothness of low-frequency eigenvectors) combined with $\xi^m_r \approx \xi^m_s \approx 0$. For high-frequency modes both factors can be large; this case is not covered by the present argument and is left as an open problem. The claim of full distant-concept protection therefore holds under the additional assumption that the concept's representation is dominated by low-frequency modes, consistent with the UToM framework but not proved in full generality here.

Remark 8 (Coordinate Identity in Sparse vs. Dense Architectures)**.** *The local insertability of the connectome rests on a property that dense weight matrices structurally lack:* coordinate identity*.*

In the biological connectome, every synapse (i, j) has an unambiguous address: neuron i is a physical object with a location, morphology, and connectivity pattern that exists independently of what the network has learned. A Hebbian update on (i, j) modifies exactly that physical connection without implying any change elsewhere. The address of a synapse is prior to *the content it encodes.*

In a dense weight matrix $W \in \mathbb{R}^{d\times d}$, no entry (i, j) has identity outside its relation to all other entries simultaneously. A stored concept is not localised to any region of W; it is distributed across all d^2 entries as a global eigenvector superposition. No single entry $(W)_{ij}$ can be identified

as "the synapse for concept A" without knowing the full eigenbasis—i.e., without knowing all other concepts. The address of a concept in W presupposes knowledge of all other concepts in W: the circularity is exact.

This is not a hardware limitation. It is a consequence of the permutation symmetry *of dense hidden units: any permutation π of the hidden dimensions produces a weight matrix $W' = P_\pi W P_\pi^\top$ that is computationally identical to W, so individual dimensions carry no intrinsic meaning. The biological connectome breaks this symmetry explicitly through physical instantiation: neurons are not interchangeable. The address of a synapse is independent of what the network knows.*

7 Consequences

7.1 Catastrophic Forgetting as Structural Necessity

The standard framing of catastrophic forgetting is algorithmic: the network "forgets" because gradient descent overwrites weights. Theorem 2 and Proposition 7 reframe this at a more fundamental level.

For structurally new concepts, forgetting is not caused by a poorly chosen optimiser. It is caused by the impossibility of introducing a new eigenspace direction without globally reorganising the existing eigenstructure—and thus simultaneously disrupting all pairwise semantic relations in spectral space. There is no submatrix of W that corresponds to a single concept; a new concept requires a new eigenvector, which shifts all others.

This explains why methods such as EWC and rehearsal succeed for new facts about existing concepts (Case 1 of Corollary 4) but fail for structurally new concepts (Case 2): they address the algorithmic symptom, not the structural cause.

7.2 The Irreducible Cost of Continuation Training

The only escape from the impossibility is *global retraining*: presenting all existing training data $\mathcal{D}_{\text{old}}$ jointly with the new concept's data $\mathcal{D}_{\text{new}}$. This works because the combined dataset implicitly encodes all attractor positions, allowing the optimiser to find a new globally consistent eigenstructure.

The cost is irreducible: every training epoch must present $N_{\text{old}} + N_{\text{new}}$ examples. As the knowledge base grows, this cost grows without bound. Whether initialisation from existing weights reduces the total number of required epochs is an open empirical question; the global eigenstructure reorganisation required by Theorem 2 suggests that the benefit may be limited.

7.3 Implications for Architecture Design

Theorem 2 and the analysis of Section 6 jointly suggest that truly incremental concept learning requires architectures with *explicit, locally modifiable topology*: structures where adding a new concept corresponds to adding a local subgraph with identifiable synapse addresses, not reorganising a global weight matrix.

The biological connectome is one such architecture. Designing artificial analogues—systems where concepts have explicit local addresses and new connections can be added without disturbing the global eigenbasis—is an open engineering problem whose solution cannot come from within the dense-weight paradigm.

8 Open Problems

1. **Quantitative bound on relational disruption.** Proposition 7 establishes that $\rho_{mk}(W') \neq \rho_{mk}(W)$ for generic dense ΔW. A quantitative bound on $|\rho_{mk}(W') - \rho_{mk}(W)|$ in terms of $\|\Delta W\|_2$, the spectral gap δ_j, and the pattern overlaps $\langle e_j, \xi^m \rangle$ would strengthen the result. The first-order expression from the proof of Proposition 7 provides a starting point.

2. **Formal proof of the Case 1/Case 2 threshold.** Proposition 6 provides a heuristic threshold $\cos\theta(a^*) = 1/\sqrt{p}$; a formal proof of its sharpness is open.

3. **Constructive architecture for local insertability.** Section 6 shows that sparse topology enables local insertability. A constructive characterisation of the sparsity pattern and locality conditions sufficient for safe concept addition at per-epoch cost $O(N_{\text{new}})$ would provide a blueprint for practical architectures.

4. **Interaction of relational disruption with concept similarity.** Proposition 7 implies that concepts with large overlap on the perturbed mode e_j (semantically close to a^*) experience the largest relational disruption. A formal analysis of this similarity-disruption coupling—and its interaction with the survival condition (Proposition 5)—would yield a complete characterisation of which concepts are most at risk when a given new concept is added.

9 Conclusion

We have proved that in any dense associative memory, no weight perturbation can introduce a structurally new concept without shifting at least one existing eigenvalue or rotating at least one existing eigenvector. This result is structural, not algorithmic: it holds for any symmetric ΔW of any rank, and no optimisation method can circumvent it (Corollary 3).

Beyond attractor survival, the introduction of a new concept disrupts the pairwise semantic relations between *all* existing concepts in truncated spectral space (Proposition 7). This holds even when the existing attractor positions are fully known: knowledge of where the concepts are does not provide a mechanism to embed a new one without disturbing the relational geometry among existing ones. The placement problem—simultaneously satisfying proximity to related concepts, distance from unrelated ones, and preservation of all pairwise semantic relations—is not merely underdetermined in practice but geometrically obstructed in principle: the global coupling term

$$\Delta\rho_{mk} \propto \langle e_j, \xi^m \rangle \langle e_j, \xi^k \rangle \cdot |\mu - \lambda_j|$$

cannot be selectively suppressed for individual pairs, regardless of what information is available at update time (Remark 7).

The biological solution is not algorithmic but structural. In a dense weight matrix, the representation of any concept is distributed across all d^2 parameters simultaneously, with no locality or addressability. The brain, by contrast, uses a sparse graph—the connectome—where every synaptic weight has an explicit address (i, j) independent of learned content. A Hebbian update on (i, j) perturbs only those eigenmodes with significant amplitude at nodes i and j, leaving the existing semantic map intact for all concepts whose representations are localised elsewhere. Adding a new Hebbian edge introduces a new attractor without globally displacing existing ones.

This local insertability is categorically unavailable to dense-weight architectures: there is no (i, j)-address in W because every entry couples every concept to every other concept simultaneously (Remark 8). Catastrophic forgetting, in its most fundamental form, is not a limitation of current optimisation methods—it is a consequence of the implicit, dense substrate in which meanings are stored.

References

[1] A. Bean. A theory of meaning, working memory, and the structure of conscious experience. Zenodo, 2026. https://doi.org/10.5281/zenodo.19414276.

[2] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[3] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[4] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[5] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

[6] R. Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2):285–308, 1990.

[7] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.