

Memorisation and forgetting in a learning Hopfield neural network: bifurcation mechanisms, attractors and basins

Adam E. Essex,^{1, a)} Natalia B. Janson,^{1, b)} Rachel A. Norris,¹ and Alexander G. Balanov^{2, c)}

¹⁾*Department of Mathematical Sciences, Loughborough University, Loughborough LE11 3TU, UK*

²⁾*Department of Physics, Loughborough University, Loughborough LE11 3TU, UK*

Despite explosive expansion of artificial intelligence based on artificial neural networks (ANNs), these are employed as “black boxes”, as it is unclear how, during learning, they form memories or develop unwanted features, including spurious memories and catastrophic forgetting. Much research is available on isolated aspects of learning ANNs, but due to their high dimensionality and non-linearity, their comprehensive analysis remains a challenge. In ANNs, knowledge is thought to reside in connection weights or in attractor basins, but these two paradigms are not linked explicitly. Here we comprehensively analyse mechanisms of memory formation in an 81-neuron Hopfield network undergoing Hebbian learning by revealing bifurcations leading to formation and destruction of attractors and their basin boundaries. We show that, by affecting evolution of connection weights, the applied stimuli induce a pitchfork and then a cascade of saddle-node bifurcations creating new attractors with their basins that can code true or spurious memories, and an abrupt disappearance of old memories (catastrophic forgetting). With successful learning, new categories are represented by the basins of newly born point attractors, and their boundaries by the stable manifolds of new saddles. With this, memorisation and forgetting represent two manifestations of the same mechanism. Our strategy to analyse high-dimensional learning ANNs is universal and applicable to recurrent ANNs of any form. The demonstrated mechanisms of memory formation and of catastrophic forgetting shed light on the operation of a wider class of recurrent ANNs and could aid the development of approaches to mitigate their flaws.

Keywords: Hopfield neural network, learning, dynamical system, memory formation, catastrophic forgetting, bifurcation, attractor, basin

© Copyright 2025 A.E. Essex, N.B. Janson, R.A. Norris, A.G. Balanov. This article is distributed under a Creative Commons Attribution (CC BY) License.

Artificial neural networks (ANNs) are now abundant in a broad range of pivotal technologies, including communication, security, health and finance. Being rough imitations of the biological brain, they provide a powerful paradigm for pattern recognition, categorisation, data processing and optimisation tasks, which are conventionally attributed to artificial intelligence (AI). However, one of the major drawbacks of such AI is the lack of transparency in the mechanisms related to memory formation, reasoning and decision-making, known as the explainability problem. Here, by analysing the dynamics and, more specifically, the phase space of an archetypal high-dimensional Hopfield neural network, we study mechanisms of its memory formation as it learns from external stimuli. Our analysis reveals typical bifurcations, which occur in the course of training and lead to the formation of attractors and of their basins associated with memorised categories. We show that bifurcations play the dual role in learning: besides creating memories, they are responsible for the well-known deficiencies of ANNs, namely, spurious memories and catastrophic forgetting. Our results constitute an uncommon instance of analyses and visualisations conducted on genuinely high-dimensional dynamical systems, and of a comprehensive study of the full range of

dynamical phenomena involved in learning. They shed light on the dynamical mechanisms of memory formation, spurious memories and catastrophic forgetting in a wide class of recurrent ANNs, and thus provide a promising ground for further understanding of reasoning and decision-making, as well as for designing ANNs with better explainability.

I. INTRODUCTION

Artificial neural networks (ANNs) are embedded in many aspects of modern technologies, including sound and image recognition, optimisation, data analysis, dynamics prediction and medical diagnostics^{1,2}. Inspired by a biological intelligent system (the brain), such networks are formed by many interconnected functional non-linear elements, called neurons, in which the inter-neuron coupling strengths (weights) are adjusted in the course of learning.

Despite the extensive development of ANNs and the growing area of their application, they exhibit critical flaws. These include spurious memories³ when the emerging attractors do not represent any real categories, or catastrophic forgetting^{4–7} when previously learnt categories disappear abruptly as learning continues. The latter presents a major obstacle to the realisation of lifelong learning in ANNs^{8,9}. Also, the inner operation of ANNs is notoriously difficult to explain¹⁰.

Here we focus exclusively on recurrent ANNs, i.e. those

^{a)}E-mail: A.Essex@lboro.ac.uk

^{b)}E-mail: N.B.Janson@lboro.ac.uk

^{c)}E-mail: A.Balanov@lboro.ac.uk

that can be modelled as dynamical systems (DSs), and in what follows we often omit the descriptor ‘‘recurrent’’. In recurrent ANNs, learning modifies the vector field, giving rise to new structures in the phase space, such as attractors or saddle limit sets. On the one hand, in machine learning literature it is overwhelmingly assumed that in ANNs the knowledge is contained in their weights. On the other hand, it is widely accepted that memories or categories are represented as attractors^{11–13}, or more appropriately, as attractor basins^{14,15}. Whereas both paradigms are well justified and linked, an explicit connection between the weights and attractors coding individual memories is missing. Moreover, the mechanisms underlying the formation of memories and categories during learning are not transparent, and ANNs are employed as black boxes¹⁶. Because of this, common methods for mitigating catastrophic forgetting are empirical and not directly tied to its underlying causes^{6,8,17–20}.

It is often appreciated that in learning ANNs bifurcations can occur, however, their influence is deemed mostly detrimental^{21–23}. With this, in 1988 it was theoretically suggested that bifurcations are *required* for memory formation¹⁴, which implies that they should be crucial for successful learning. These two viewpoints seem to present a paradox needing resolution. However, due to the challenges of bifurcation analysis in learning NNs (discussed in Sections IV and VC), the respective studies are currently missing.

Meanwhile, significant efforts are being directed towards developing techniques that provide insights into the internal workings of NNs and/or improve their interpretability^{24,25}. However, understanding the mechanisms of memory formation and decision-making remains one of the major challenges of ANN research.

In this paper we address the paradox above and reveal the dual role of bifurcations in learning recurrent ANNs, while simultaneously clarifying their inner workings. We do so by considering a paradigmatic recurrent ANN – continuous-time Hopfield NN^{12,13} with Hebbian learning^{26,27}. Specifically, in an 81-neuron network, we uncover bifurcation mechanisms of memory formation and of catastrophic forgetting, and provide evidence that these are two manifestations of the same process. We also reveal the basins of attraction formed by the end of learning.

The paper has the following structure. In Sec. II we describe the continuous-time Hopfield NN undergoing Hebbian learning used in this study. Section III reports observations of bifurcations occurring in the course of learning and leading to the formation of memories. In Sec. IV a systematic bifurcation analysis is performed along a one-dimensional trajectory followed by the NN in the space of weights as it learns from stimuli. In Sec. V we clarify and verify a hypothesis about the bifurcation mechanisms of catastrophic forgetting in ANNs using the Hopfield NN as a case study. Section VI discusses the structure of the boundaries of attraction basins and their relation to memory of the network. The conclusive remarks and the discussion of the broader significance of our findings are given in Sec. VII.

II. HOPFIELD NEURAL NETWORK

Hopfield NNs^{12,13} are a type of ANN capable of pattern recognition, which feature the so-called associative memory. Despite the recent proliferation of deep (multi-layer) NNs, a single-layer Hopfield NN is an important paradigm of a NN and remains highly relevant to date^{28,29}. It consists of a set of coupled neurons (functional units), which are connected to each other with certain strengths, usually referred to as weights.

Here we use the NN model in the same form as in Ref.³⁰:

$$\frac{dx_i}{dt} = -x_i + g \sum_{j=1}^N F(\omega_{ij}) F(x_j) + A I_i(t), \quad (1)$$

$$\frac{d\omega_{ij}}{dt} = \frac{1}{B_{ij}} (-\omega_{ij} + F(x_i) F(x_j)), \quad (2)$$

where $x_i(t)$ ($i = 1, \dots, N$) is the state of i^{th} neuron at time t , $\omega_{ij}(t)$ is the weight of connection between i^{th} and j^{th} neurons, and $I_i(t)$ is external input to i^{th} neuron, which is used as the training stimulus applied in the course of NN learning. Following Ref.³⁰, we consider the case without self-connections $\omega_{ii}=0$ and with symmetric couplings $\omega_{ij} = \omega_{ji}$ for all $i, j=1, \dots, N$. The activation function $F(x)$, which here takes the form³⁰

$$F(x) = \left(\frac{2}{\pi} \right) \arctan \left(\frac{\lambda \pi x}{2} \right), \quad (3)$$

determines the response of the i^{th} neuron to the collective input from all other neurons, g is the coupling gain parameter, and A is the strength of the external input.

The network learns from the input signal $\mathbf{I}(t) = (I_1(t), \dots, I_N(t))$ containing a sequence of N -dimensional vectors, each coding some pattern, by adjusting its weights ω_{ij} according to some built-in rule. Often this is the Hebbian learning rule governed by (2). The rate of learning is regulated by parameter B_{ij} .

Once the patterns are learned, the weights ω_{ij} are fixed at their latest values, and $\mathbf{I}(t)$ is set to zero. After that the network (1) can recognise a new pattern as belonging to some stored category, or ‘‘recall a memory’’. This is achieved by setting initial conditions of (1) to a vector representing this new pattern and observing how the state $\mathbf{x} = (x_1, \dots, x_N)$ converges to an attractor to whose basin the initial conditions belong. The respective attractor is assumed to represent the most typical element of the given category, or the undistorted memory.

Hopfield NNs have some important properties. They are robust to noise and can recall memories from cues, which are distorted or incomplete versions of the memorised patterns. They also have the property of a content-addressable memory³¹, meaning that the network can recall a stored pattern based on its content, rather than its physical address as in random-access memory (RAM) of conventional computers. This is achieved by *identifying* the content with location: any pattern is coded by a state vector \mathbf{x} , but instead of the physical location within a real device, any state vector has its unique

location in the phase space of the NN. When recalling a memory from a cue, the NN automatically goes through the states in its phase space towards the required attractor.

Operation of the Hopfield NN is often explained in terms of an energy function³⁰, to whose local minimum the state of the NN converges with time. Namely, each state of the network is associated with an energy level, and the network evolves from the states with higher energy to those with lower energy. The state with the locally lowest energy represents an attractor, which in the Hopfield NN is the stable fixed point.

Remarkably, if in (3) $\lambda \rightarrow \infty$, thus making $F(x)$ a Heaviside step function, for $x_j \neq 0$ with $j=1, \dots, N$, Eq. (1) can be rewritten as

$$\frac{dx_i}{dt} = -\frac{\partial V}{\partial x_i}, \quad (4)$$

where

$$V = \sum_{i=1}^N \left[\frac{(x_i - A I_i(t))^2}{2} - g x_i \sum_{j=1}^N F(\omega_{ij}) F(x_j) \right]. \quad (5)$$

In this form, Eq. (1) has a mechanical interpretation as a model describing the motion of an overdamped particle with location $\mathbf{x}(t)$ in the potential energy landscape $V=V(\mathbf{x}, t)$. Thus, the shape of V changes with time, as determined by the dynamics of the weights ω_{ij} and stimulus I_i . In that case, the model (1)–(2) can be regarded as a particular case of a dynamical system with plastic self-organising vector field^{32,33}.

The local minima of V occur at the stable fixed points, which either by themselves^{11–13}, or together with their attraction basins¹⁵, represent stored patterns, or memories that the network has learned. For (1)–(2) with $N=81$, this is illustrated in Fig. S1(a)–(b) of Supplementary Notes, showing two different cross-sections (coloured surfaces) of the same post-training V by two different three-dimensional flat surfaces in $(81+1)$ -dimensional space, each going through a different attractor. The respective attractors (black circles) are visible at the bottoms of cross-sections of V . Notably, V could in some sense be regarded as an energy function discussed in Refs.^{12,13}.

We study (1)–(2) with the following parameter values: $N=81$, $A=30$, $B_{ij}=B=300$ ($\forall i, j$), $g=0.3$ and $\lambda=1.4$.

By analogy with Ref.³⁰, we design the input signal $\mathbf{I}(t)$ from a training set of six N -dimensional vectors, each assumed to code some non-specified pattern. The components I_i^k ($k=1, \dots, 6$) of these vectors are chosen at random from the set $\{1, -1\}$, so that 1 or -1 have equal probability of occurrence, and the values of the components are statistically independent of each other. Two different training sets, Sets 1 and 2, used in our study for the network of $N=81$ neurons are given in Tabs. S1–S2 of Supplementary Notes.

Starting from $k=1$, the signal $\mathbf{I}(t)$ is equal to a vector $\mathbf{I}^k=(I_1^k, \dots, I_N^k)$ from the training set during the initial $t_s=12$ time units. For the subsequent t_s time units, $\mathbf{I}(t)$ becomes equal to \mathbf{I}^{k+1} . In the same manner $\mathbf{I}(t)$ goes through all vectors \mathbf{I}^k until k reaches 6. After that, the signal known as the “training epoch” in ANN literature³⁴, and here lasting $6t_s$ time units, is repeated periodically throughout the maximal training time

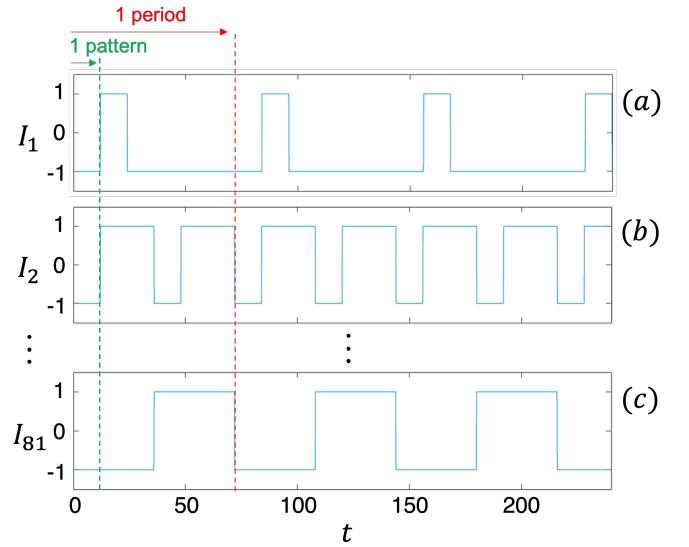


FIG. 1. Illustration of construction of input signals $I_i(t)$ to the NN (1)–(2) with $N=81$ from the training Set 1 of vectors \mathbf{I}^k ($k=1, \dots, 6$) given in Tabs. S1–S2 of Supplementary Note. Vertical green dashed line indicates the time t_s during which a single vector \mathbf{I}^k is applied to the NN. Red dashed line indicates one full period of stimulus $\mathbf{I}(t)$ (duration of “training epoch”), which is equal to $6t_s$. Panels (a), (b) and (c) show inputs $I_i(t)$ to the 1st, 2nd and 81st neurons, respectively.

of 6000 time units. This way, inputs to individual neurons take the form of telegraph signals shown in Fig. 1.

At the start of learning, the initial conditions (ICs) of (1)–(2) were randomly and uniformly distributed in $[-1, 1]$ for all x_i , and in $[-0.01, 0.01]$ for all ω_{ij} , the latter ensuring the absence of pre-existing memories.

The training duration of 6000 time units was chosen experimentally by ensuring that the weights ω_{ij} converge to small-amplitude oscillations about some fixed values (see Fig. 2) for a sufficiently long time before the end of learning. Small oscillations of individual ω_{ij} , which occur during the learning phase, are due to disturbance of (1)–(2) by a periodic signal $\mathbf{I}(t)$, and their period is equal to $6t_s=72$.

III. MEMORY FORMATION: OBSERVATION OF ATTRACTORS

After learning, the NN becomes an autonomous dynamical system (DS):

$$\frac{dx_i}{dt'} = -x_i + g \sum_{j=1}^N F(\omega_{ij}) F(x_j) = u_i(\mathbf{x}). \quad (6)$$

Equation (6) is obtained from (1) by setting $A=0$ and fixing the values of ω_{ij} , and t' is the time in the memory retrieval phase, which we distinguish from time t of learning used in (1)–(2).

Note, that during learning, the model (1)–(2) represents a non-autonomous DS. In such systems it can be possible to define and detect non-autonomous attractors³⁵. However, doing

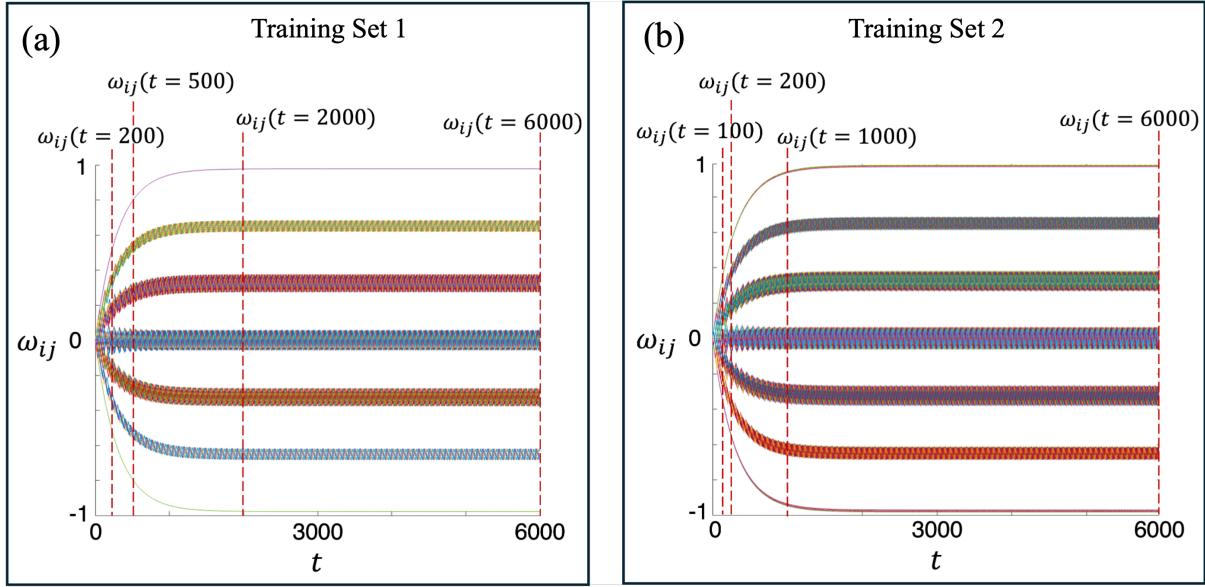


FIG. 2. Evolution of connection weights $\omega_{ij}(t)$ (solid lines in various colours) during learning by the NN (1)–(2) with $N=81$, $A=30$, $B_{ij}=B=300$ ($\forall i, j$), $g=0.3$ and $\lambda=1.4$. Panels show learning from two different training sets (see Tabs. S1–S2): (a) Set 1 and (b) Set 2. Vertical dashed lines mark four different stages of learning corresponding to times t given in brackets, at which instantaneous weights ω_{ij} are collected to reveal memories formed, as illustrated in Fig. 3.

so is not practical here, since we wish to study memories existing in the *autonomous* system (6) at various stages of learning, i.e. when (6) is considered with various sets of *fixed* ω_{ij} occurring at various times t during the learning phase described by (1)–(2). Such memories will be associated with conventional attractors and their basins.

In (6), $\mathbf{u}=(u_1, \dots, u_N)$ is the phase velocity vector field of the NN formed at the given stage of learning, which is parametrised by the set of ω_{ij} . For the NN of the size $N=81$, the time evolution of all ω_{ij} during the learning phase is illustrated in Fig. 2. Namely, Figs. 2(a) and (b) illustrate training with two different training sets, Set 1 and Set 2 (Tabs. S1–S2), respectively. Both graphs are qualitatively the same, and are in agreement with a similar graph in Ref.³⁰.

Despite the similarity of graphs in Figs. 2(a) and (b), for different training sets, both the processes of learning, and the post-learning memories formed, are very different. This is illustrated by Fig. 3, where for two training sets, Set 1 (a)–(d) and Set 2 (e)–(h), the projections of phase portraits of (6) onto the (x_1, x_2) -plane are given at various stages of learning. The respective stages are marked in Fig. 2(a) and (b), see values of t in brackets next to vertical dashed lines.

It is desirable that, as a result of learning, attractors representing vectors \mathbf{I}^k either coincide with them, or are found at predictable locations, which could be linearly linked to \mathbf{I}^k . However, for continuous-time and continuous-space NNs this does not happen because of their non-linearity. This is clearly seen in Figs. 3(d) and (h): among various attractors formed by the end of learning (black diamonds and red crosses), none have components equal to ± 1 , $\pm A$ or $\pm C$, where C is some constant. Thus, to associate attractors with vectors \mathbf{I}^k , a different approach is needed, as described below.

In Fig. 3 attractors associated with input vectors \mathbf{I}^k (“true memories”) are shown as black diamonds, and attractors representing spurious memories as red crosses. To reveal attractors and to distinguish between them, we considered three types of ICs specified below. Descriptions in brackets refer to trajectories launched from these ICs.

1. Type 1: Identical to six vectors \mathbf{I}^k from the training set (dashed coloured lines).
2. Type 2: Within small vicinities of six vectors \mathbf{I}^k from the training set. Namely, to each \mathbf{I}^k , small vectors were added, whose components were randomly and uniformly distributed in $[-0.5, 0.5]$ (solid lines coloured as in Type 1 for respective \mathbf{I}^k).
3. Type 3: 1000 randomly and uniformly distributed within an N -dimensional hyper-cube of side length 10 and the centre at the origin (grey dashed lines).

If (6) converges to the given attractor from Type 1 and 2 ICs corresponding to a single \mathbf{I}^k , we deem this attractor and its basin a true memory of the respective \mathbf{I}^k . If the system converges to the same attractor from multiple input patterns \mathbf{I}^k and their respective neighborhoods, we interpret this attractor and its basin as representing a blended memory of the corresponding \mathbf{I}^k s. Any attractor whose basin does not encompass any \mathbf{I}^k or its vicinity is considered a spurious memory.

Type 2 ICs are important for identification of true memories, since to ensure that their recall is robust to noise, the respective attractor basins should not be negligibly small.

Let us observe stages of memory formation as the NN (1)–(2) is processing the repeatedly applied six vectors from Set 1. By $t=200$ (Fig. 3(a)) the NN develops only two attractors

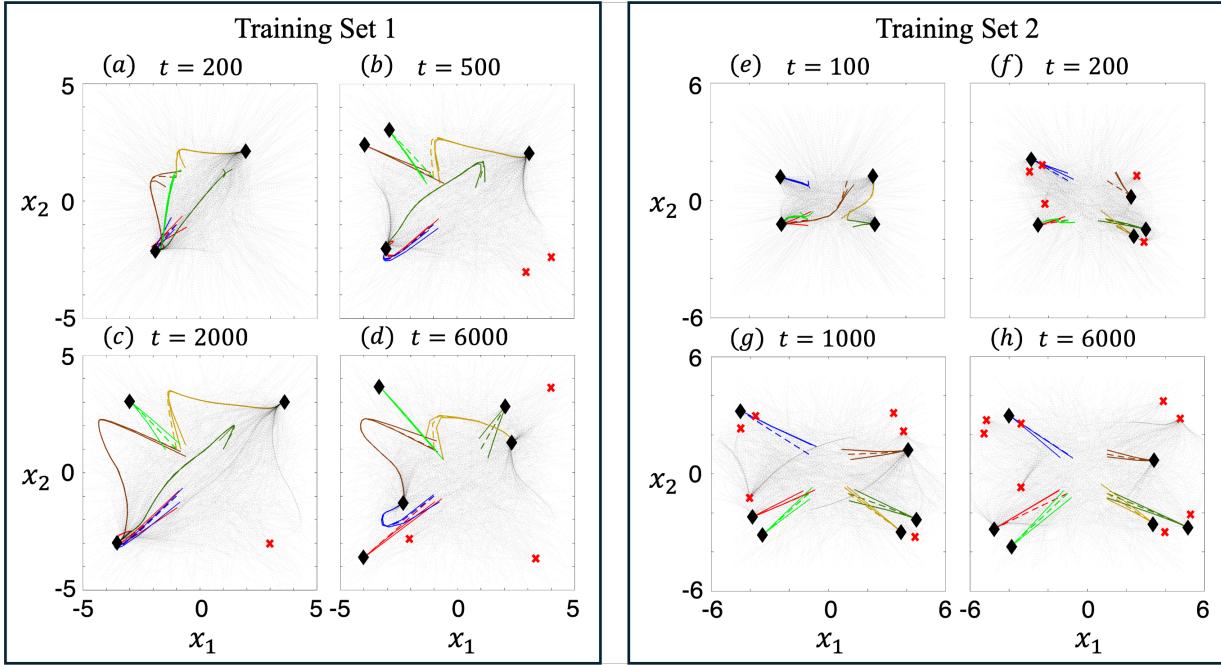


FIG. 3. Memories formed at various stages of learning by the NN (1)–(2) with $N=81$, $A=30$, $B_{ij}=B=300$ ($\forall i, j$), $g=0.3$ and $\lambda=1.4$. Panels illustrate stages of learning from two different training sets (see Tabs. S1–S2): (a)–(d) Set 1 and (e)–(h) Set 2. Panels display projections onto the (x_1, x_2) -plane of phase portraits of the NN (6) with sets of fixed values of ω_{ij} taken at various stages of learning corresponding to times t of (1)–(2) indicated above each panel, also marked in Fig. 2. Each panel shows attractors associated with true memories of input vectors \mathbf{I}^k (black diamonds), attractors representing spurious memories (red crosses) not associated with any \mathbf{I}^k , phase trajectories originating from six vectors \mathbf{I}^k and their vicinities (Type 1 and 2 ICs, dashed and solid lines with one colour corresponding to one \mathbf{I}^k), and phase trajectories launched from random ICs (Type 3 ICs, grey dashed lines).

(black diamonds). The one on the right is associated with the true memory of \mathbf{I}^3 , since it attracts phase trajectories from the respective Type 1 and 2 ICs (yellow lines). The one on the left is associated with a blended memory of five patterns, since it attracts all trajectories from ICs of Types 1 and 2 for $k \neq 3$ (other coloured lines).

By $t=500$ (Fig. 3(b)), in addition to attractors present at $t=200$ (see Fig. 3(a)), four new attractors are formed. With this, the old attractors have moved away from each other (compare black diamonds along the main diagonal in (b) and (a)), and one of them is still associated with \mathbf{I}^3 . Among the four new attractors, two are spurious memories (red crosses). However, two of the new attractors with basins become true memories of \mathbf{I}^5 and \mathbf{I}^6 , as evidenced by the trajectories from the respective ICs (brown and light-green lines). Thus, one of the old attractors is now associated with a blend of \mathbf{I}^1 , \mathbf{I}^2 and \mathbf{I}^4 (see red, dark green and blue lines, respectively).

By $t=2000$ (Fig. 3(c)), a separate memory of \mathbf{I}^5 disappears, and the old attractor on the left is now associated with a blended memory of \mathbf{I}^1 , \mathbf{I}^2 , \mathbf{I}^4 (red, dark green and blue lines, respectively) and of \mathbf{I}^5 (brown lines). Thus, \mathbf{I}^5 was forgotten.

By the end of learning at $t=6000$ (Fig. 3(d)), the NN has formed five memories. Four of them are separate memories of \mathbf{I}^1 (red lines), \mathbf{I}^2 (dark green lines), \mathbf{I}^3 (yellow lines) and \mathbf{I}^6 (light green lines). One attractor is associated with a blend of \mathbf{I}^1 and \mathbf{I}^5 (red and brown lines, respectively). Thus, during learning, attractors were being formed gradually to be-

come associated with some of the individual applied patterns (proper memories) and those associated with more than one applied pattern (incorrectly formed memories).

To explore the sensitivity of learning to input data, the NN (1)–(2) was subjected to the same learning procedure, but with a different training signal $\mathbf{I}(t)$ – the one formed from the training Set 2 (see Tabs. S1–S2). The stages of memory formation are illustrated in Fig. 3(e)–(h). Note, that the statistical properties of Sets 1 and 2 are the same, but the actual values of their vectors \mathbf{I}^k are different.

Despite the statistical similarity of both sets, they result in very different processes of learning, as is evident from comparing the two boxes in Fig. 3. Attractors appear at very different locations and become associated with true or spurious memories in a seemingly random manner. These observations are in line with a recent attempt to explain learning by the Hopfield NN using a plastic self-organising vector field, and with the conclusion about its inherent non-explainability³⁶. Unlike with Set 1, with Set 2 the NN develops six proper memories, each associated with an individual vector \mathbf{I}^k .

Our simulations suggest that, although training may lead to creation of attractors whose number equals, or even exceeds, the number of patterns the NN needs to memorise, not all patterns could be represented by a unique attractor. Thus, the development of the full memory critically depends on the content of the training set.

IV. MEMORY FORMATION: BIFURCATION MECHANISMS

After observing stages of memory formation in the NN in the course of its learning, we need to reveal the underlying *mechanisms*. Below we quote some results from the area of machine learning that are relevant to recurrent NNs, where those may differ substantially from the continuous-time Hopfield NN considered here. Namely, they could have a discrete time and/or a different evolution rule for the neuron states. Moreover, modern machine learning mostly uses non-Hebbian learning algorithms, since they are more efficient.

Despite these differences, in all recurrent NNs the *conceptual* principles behind learning are fundamentally the same. Namely, during learning the weights need to be adjusted in such a way, that at the end a certain structure of the vector field in the state space of the NN is formed, which ensures an appropriate number and configuration of attractors and their basins – often metaphorically called “attractor landscape”²³. Therefore, better understanding of Hopfield NNs with Hebbian learning will aid understanding of learning in other recurrent NNs.

In 1988, a theoretical idea was put forward that if a NN initially had only one attractor, then to associate different training vectors with distinct attractors, during learning this NN needs to go through bifurcations (or “catastrophes”)¹⁴. Recently, evidence was obtained that in non-small (up to 200-neuron) discrete-time NNs, bifurcations do occur during learning, since the numbers or the types of attractors could change between the consecutive training epochs^{23,37}. In these studies, the nature of bifurcations or of attractors involved were not clearly specified, and bifurcation parameters were not introduced.

Despite the importance of bifurcations for learning, there have been no rigorous and systematic studies explicitly revealing bifurcations in a learning NN and demonstrating how they lead to memory formation. The absence of such studies likely arises from the conceptual and technical challenges involved in analysing bifurcations in learning NNs, as explained below.

Bifurcations of co-dimension one can be detected by fixing all parameters of a DS except one, and allowing this single parameter to monotonously increase or decrease, while monitoring a certain limit set and its stability. A local bifurcation³⁸ is the instant when this stability changes, thus leading to a dramatic change in the system behaviour³⁹. A more sophisticated well-established method of bifurcation analysis is continuation^{40–42}, which can detect bifurcations of co-dimensions one and higher, when theoretically any number of parameters can be changed simultaneously. However, these approaches work well when the number of control parameters in a DS is relatively small and is comparable with the number of state variables. Such conventional bifurcation analysis was successfully done for relatively small non-learning NNs with two, three, or four neurons^{43–49}.

However, in NNs of a non-small size N , the number of control parameters ω_{ij} is *much* larger than the number N of state variables. For example, in the Hopfield NN (6) with $\omega_{ij}=\omega_{ji}$, $\omega_{ii}=0$ and $N=81$, the number M of distinct weights calculated

as $M=\frac{N^2-N}{2}$ is 3240, which is much more than 81. In this situation, conventional methods of bifurcation analysis involving all significant parameters are not practical, since all weights are equally important and none are more significant than others.

To reveal bifurcation mechanisms of memory formation in the learning NN (1)–(2), here we adopt the following strategy. We note that as the NN learns, *all* of its weights ω_{ij} change simultaneously and continuously. Thus, during learning the state of the subsystem (2) follows a one-dimensional path in the M -dimensional space of all ω_{ij} , which we refer to as the “weight trajectory” and visualise in Sec. V. Such a path can be parametrised by a single parameter, such as arc-length⁵⁰.

However, in our case for the given training signal $\mathbf{I}(t)$ and with the given initial conditions, the state of (2) at any time t is fully determined by t , and therefore the path it follows can be parametrised by t . Therefore, for the DS (6) we can assume that $\omega_{ij}=\omega_{ij}(t) \forall i, j$, so that t could be treated as a *bifurcation parameter*.

Importantly, bifurcations detected with the increase of a single parameter t will most likely have co-dimension one. The probability to come across a bifurcation with higher co-dimension is negligibly small.

We search for bifurcations while being aware that in (6), at any combination of ω_{ij} s, stable fixed points can be the only attractors, and any fixed point (stable or unstable) can have only real eigenvalues and be of a node type. These facts imply that oscillations in (6) are impossible, which is supported by the proof involving the concept of an energy function⁵¹.

A. Early stages of learning

Early stages of learning with the training Set 1 are illustrated in Figs. 4 and 5. Namely, Fig. 4 presents a one-parameter bifurcation diagram showing components x_1 of all fixed points of (6) as functions of t . Here, t grows from 0 to 38, during which the system fully “processes” the first three vectors \mathbf{I}^k applied to it ($t \in [0, 36]$) and starts to process the fourth one ($t \in [36, 38]$). At the start of learning ($t \in [0, 6.5]$), the NN (6) has a single attractor at the origin and no other fixed points (Fig. 5(a)).

Figure 4 demonstrates the development of new potential memories in the form of stable fixed points (black lines) with their basins. Besides attractors, saddle points with a single positive eigenvalue (blue lines) are also critical to memory formation for the following reasons. Firstly, their $(N-1)$ -dimensional stable manifolds can serve as boundaries of attractor basins representing memories, as explained in detail in Sec. VI. Secondly, they can participate in saddle-node bifurcations with attractors causing their deaths, thus potentially destroying memories. We will refer to such saddles as “useful”. Saddle points with more than one positive eigenvalue (red lines) are not immediately relevant to memory formation.

Consider a sequence of events induced by the growth of t . As t reaches 6.5 (Fig. 4(a)), a pitchfork bifurcation occurs (pink circle) that destabilises the fixed point at $\mathbf{0}$ (blue line starting at $t=6.5$) and gives rise to two new attractors (black

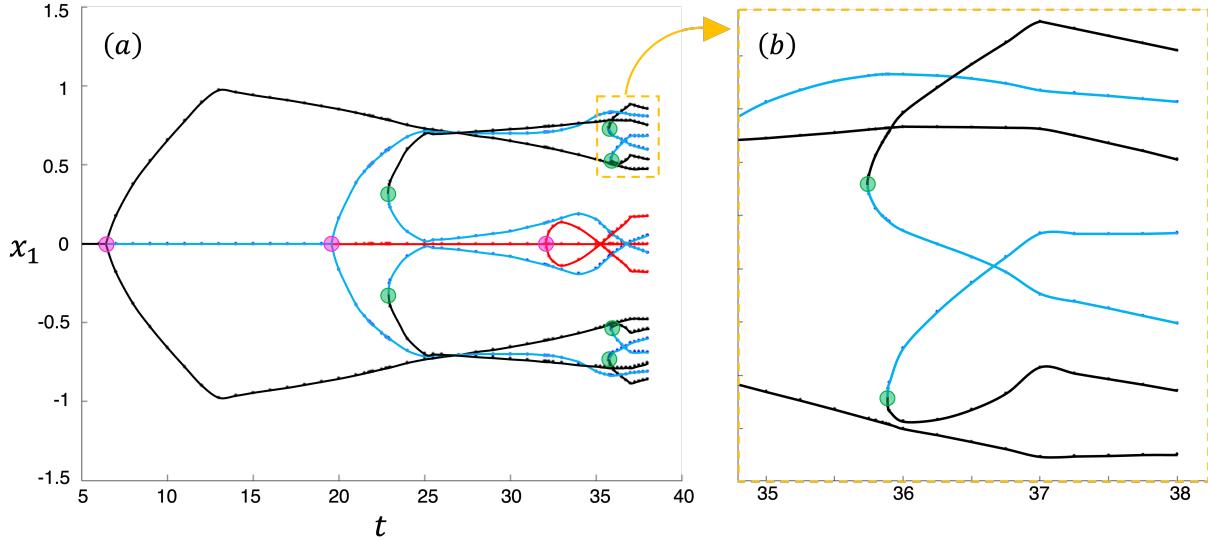


FIG. 4. One-parameter bifurcation diagram illustrating memory formation in *early* stages of learning by the NN (1)–(2) from the training Set 1 (Tabs. S1–S2) with $N=81$. Panels show x_1 -coordinates of fixed points of (6) (solid lines) as functions of control parameter t , which coincides with time t in (1)–(2). Stability of fixed points is indicated by the line colour: stable point, i.e. attractor potentially associative with memory (black), “useful” saddle point with a single positive eigenvalue (blue), and other saddle points (red). Dots along each branch indicate values of t at which the respective fixed point was numerically found and analysed; the dots are connected by interpolating cubic splines. Translucent circles mark bifurcation points: pitchfork at $t=6.5$ (pink), saddle-node at $t=22.8783$ and at $t\approx35.9$ (green). All saddle-node bifurcations occur in pairs due to symmetry in (6). Panel (a) shows the bifurcation diagram for $t \in [5, 38]$; (b) is a close-up of the rectangular selection in (a). All bifurcations involving attractors are additionally illustrated by phase portraits in Fig. 5.

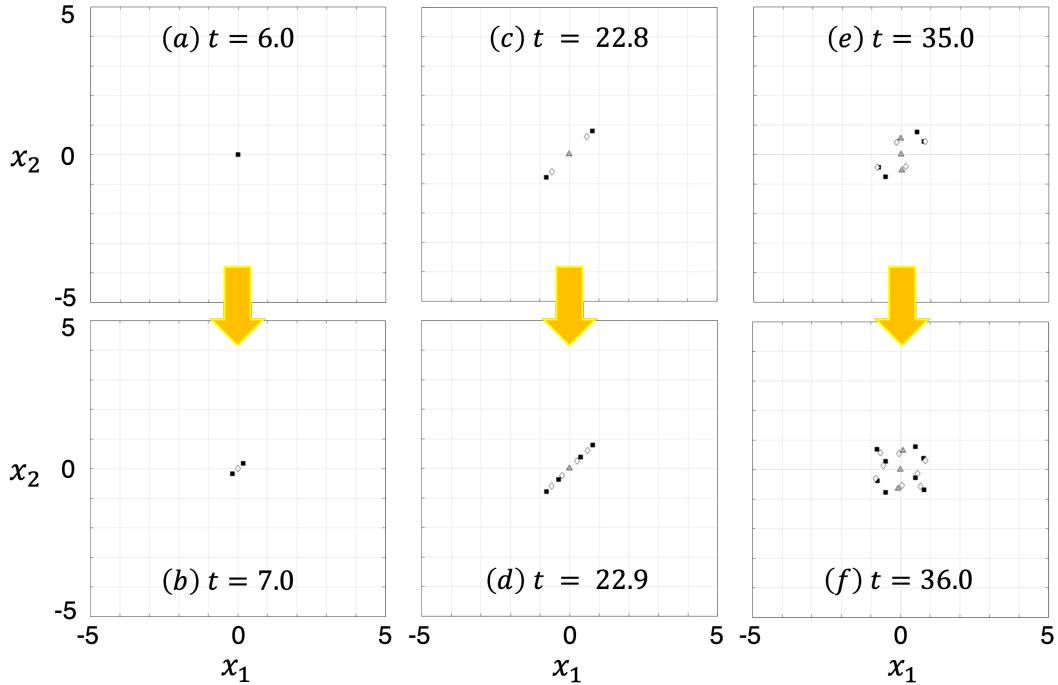


FIG. 5. Illustration of the first four bifurcations producing new potential memories in the NN (1)–(2) with $N=81$ learning from training Set 1, compare with Fig. 4. Panels show projections of fixed points on the (x_1, x_2) -plane of the NN (6) before (first row) and after (second row) the respective bifurcations, and provide the value of parameter t . Fixed points are marked as: attractors potentially associative with memories (black boxes), “useful” saddle points with one positive eigenvalue (white diamonds), and all other saddle or unstable points (grey triangles). Yellow arrows indicate the flow of time t in (1)–(2). Bifurcations illustrated are: (a)–(b) pitchfork bifurcation at $t=6.5$, which destabilises the point $\mathbf{0}$ and produces two new attractors; (c)–(d) a pair of saddle-node bifurcations at $t=22.8783$ creating two new attractors (compare with Fig. 4(a)), and (e)–(f) two pairs of saddle-node bifurcations at $t\approx35.9$ creating four new attractors (compare with Fig. 4(b)).

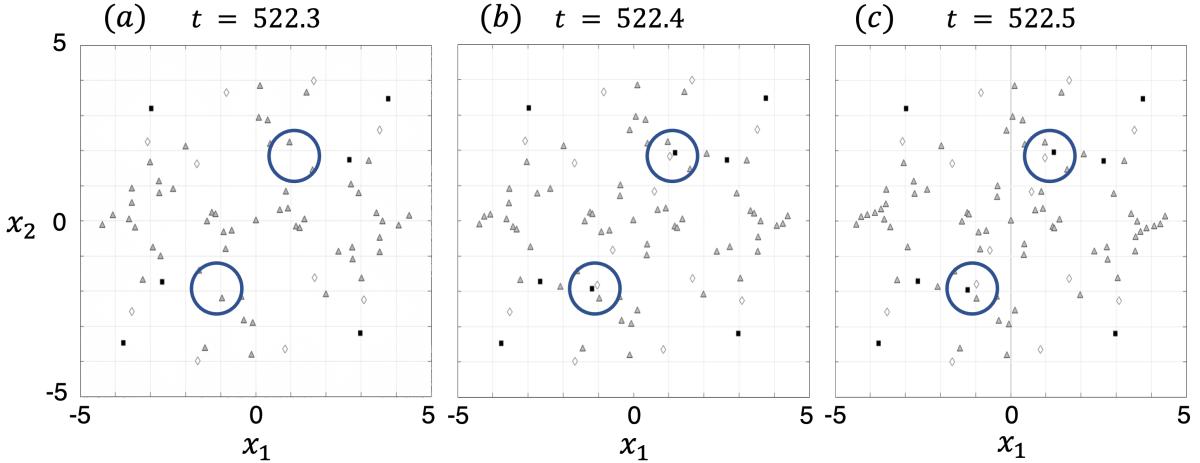


FIG. 6. Illustration of memory-creating bifurcations at later stages of learning ($t > 38$), specifically, of a pair of saddle-node bifurcations taking place in the NN (1)–(2) with $N=81$ that learns from training Set 1 (compare with Fig. 3(a)–(d)). Panels show projections onto the (x_1, x_2) -plane of fixed points of the NN (6) (a) before the bifurcation at $t=522.3$, (b) right after the bifurcation at $t=522.4$, and (c) further away from the bifurcation at $t=522.5$, with t being the control parameter of (6) equal to time t of (1)–(2). Blue circles highlight areas of the phase space where bifurcations take place. Fixed points are marked as: attractors potentially associative with memories (black boxes), “useful” saddle points with one positive eigenvalue (white diamonds), and all other saddle or unstable points (grey triangles).

lines extending for $t \in [6.5, 38]$). The same effect is illustrated by phase portraits in Fig. 5 as one goes from (a) a single attractor at $\mathbf{0}$ (black box) at $t=6$ to (b) a pair of attractors (black boxes) and a saddle at $\mathbf{0}$ with one positive eigenvalue (white diamond) at $t=7$.

After that, a cascade of saddle-node bifurcations (green circles in Fig. 4) produces more attractors and “useful” saddles. Namely, at $t=22.8783$ the first pair of saddle-node bifurcations give rise to two symmetric pairs of stable and saddle points. This is illustrated in Fig. 5 as one goes from (c) two pairs of attractors (black boxes) and “useful” saddles (white diamonds) at $t=22.8$ to (d) four such pairs at $t=22.9$ meaning that the NN acquires four potential memories.

The subsequent two pairs of saddle-node bifurcations occurs almost simultaneously at $t \approx 35.9$, as more clearly seen in Fig. 4(b). The transition of the phase portrait from before to after bifurcations is illustrated in Fig. 5(e)–(f). As a result, four more pairs of attractors and of “useful” saddles appear. Note, that saddle-node bifurcations always occur in pairs at symmetric locations in the phase space due to the symmetry of u_i in (6) resulting, in turn, from the symmetry of $F(x)$.

Interestingly, bifurcations giving rise to new attractors occur at the ends of the time intervals of length $t_s=12$, during which individual vectors \mathbf{I}^k are applied to (1)–(2). For smaller t_s , such bifurcations might not have happened. This is consistent with the idea from psychology that memories need time to “sink in”⁵².

A cascade of pitchfork bifurcations of the origin (pink circles in Fig. 4(a)), each endowing the fixed point at $\mathbf{0}$ with one more positive eigenvalue, and also producing new saddle fixed points with *more* than one positive eigenvalue, do not affect memory formation and are indicated only for completeness.

Note, that in (6) the fixed point at $\mathbf{0}$ is stable only if all ω_{ij} are close to zero. In that case, the point at $\mathbf{0}$ is the only attrac-

tor in the system, which is in agreement with early studies of similar NNs⁵³. This is true in the beginning of learning illustrated here because of the chosen ICs, i.e. $|\omega_{ij}(0)| \leq 0.01$, see Fig. 2(a) at small t . However, if in (1)–(2) the ICs for ω_{ij} are not small, then already at $t=0$ the NN (6) can have an unstable fixed point at $\mathbf{0}$ and multiple stable fixed points away from $\mathbf{0}$. In that case, the pitchfork bifurcation at relatively small t does not occur, and only saddle-node bifurcations take place.

B. Later stages of learning

At later stages of learning, saddle-node bifurcations remain the prime mechanism of memory formation, as demonstrated in Fig. 6. Namely, as t grows from 522.3 (Fig. 6(a)) to 522.4 (Fig. 6(b)), two pairs of stable and saddle fixed points appear in two symmetrically located areas of the phase space highlighted by blue circles. As t grows to 522.5, within each newly born pair of fixed points, these points move away from each other (Fig. 6(c)). The given behaviour of the fixed points, together with the analysis of their eigenvalues, clearly points to saddle-node bifurcations.

C. End of learning

As a result of the sequence of bifurcations discussed above, the NN (1)–(2) with $N=81$ trained from Set 1, by the time $t=6000$ develops fixed points shown in Fig. 7 (compare with Fig. 3(d)). Namely, Fig. 7 shows projections onto the (x_1, x_2) -plane of *all* fixed points of the NN (6) considered with the values ω_{ij} of (2) at $t=6000$. Attractors are depicted with coloured boxes, whose colours are matched with those of their basins shown in Figs. 13 and 14 and discussed in Sec. VI. “Use-

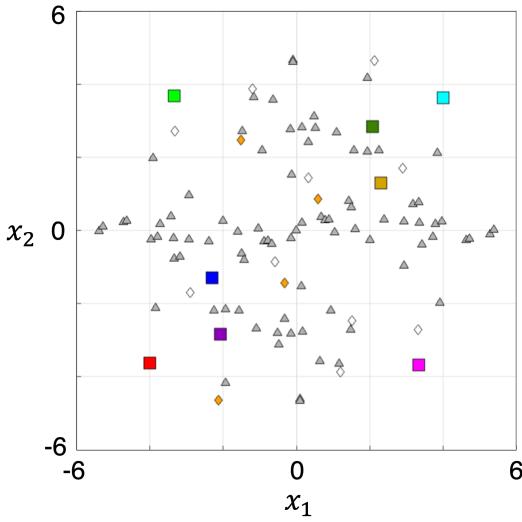


FIG. 7. All fixed points developed in the 81-neuron network (1)–(2), described in caption to Fig. 3 and trained with Set 1, at the end of learning at $t=6000$. Red, dark-green, yellow, blue, and light green boxes represent attractors identical to those shown in Fig. 3(d) by black diamonds and associated with true memories, their colours matching those of the trajectories in Fig. 3(d) launched from respective \mathbf{I}^k 's and their vicinities. Purple, pink and cyan boxes show spurious memories identical to red crosses in Fig. 3(d). Diamonds depict “useful” saddles with a single positive eigenvalue. Of these, four orange diamonds mark saddles whose manifolds are shown in Fig. 14 as basin boundaries. Grey triangles indicate all other saddle or unstable fixed points.

ful” saddle fixed points with a single positive eigenvalue are depicted by diamonds. Of these, orange diamonds indicate four saddles whose manifolds forming boundaries of attraction basins are illustrated in Fig. 14. Grey triangles indicate all other saddle or unstable fixed points.

V. CATASTROPHIC FORGETTING: BIFURCATION MECHANISMS

Here we address bifurcation mechanisms of catastrophic forgetting, which remains one of the major deficiencies of modern ANNs^{4–6}. The existing literature suggests that a full elimination, rather than the reduction, of catastrophic forgetting is only possible if ANNs expand their sizes to accommodate new knowledge^{8,19,20}. With this, the existing research suggests that in ANNs of *fixed* size catastrophic forgetting can only be mitigated, but not eradicated⁷.

It is conventionally assumed that the function of ANNs is encoded in connection weights, and memories are represented by attractors and their basins. However, an explicit connection between the weights’ values and representations of particular memories or categories in ANNs has not been established⁵⁴. Consequently, conventional strategies for mitigating catastrophic forgetting have been predominantly empirical, since the nature of the underlying mechanisms have remained largely unknown^{6,18}.

While in 1988 it was suggested that during learning a NN could undergo bifurcations¹⁴, since 1993 bifurcations have largely been viewed as detrimental to learning²¹. This is because they can cause gradient “explosions” in loss functions, which underlie popular algorithms governing evolution of weights in the course of learning (which are different from those of the Hebbian learning considered here)²³.

In an explicit mathematical language, a “bifurcation occurring during learning” can be interpreted as an event when a bifurcation manifold in the weight space of a non-learning NN is crossed by the weight trajectory of its counterpart learning NN. It was proposed that for a good learning, the initial weights of the learning NN should be set to values corresponding to a sufficient number of attractors, and the weight trajectory should steer clear of all “bifurcation boundaries”^{21,22}, i.e. of bifurcation manifolds. This is difficult to achieve because in non-linear high-dimensional multi-parameter DSSs, the locations of bifurcation manifolds are not known in advance.

Notably, for a long time the damages caused by the bifurcations during learning were not explicitly linked to catastrophic forgetting. Interestingly, Ref.⁵⁵ proposes methods to learn without forgetting by confining the weight trajectories to appropriate parts of the weight space, whose boundaries it does not identify with bifurcation manifolds. Only in 2023 an explicit link was proposed and proven between catastrophic forgetting and bifurcations (as well as the gradient explosion of a loss function) for a special kind of a NN – a piecewise-linear NN with discrete time⁵⁶ – which is different from the smooth continuous-time NNs studied here. Note, that bifurcations in non-smooth DSSs may not coincide with those in the smooth ones⁵⁷. With this, NNs modelled by smooth DSSs are mathematically closer to both the biological NNs, and the neuromorphic devices. Thus, it is particularly important to reveal the general principles behind catastrophic forgetting in smooth NNs.

In this Section we propose and test a hypothesis that catastrophic forgetting in smooth NNs occurs when, during learning, the weight trajectory crosses some bifurcation manifold in the weight space. To gather evidence supporting or opposing this hypothesis, we perform two case studies of (1)–(2) with a small (Sec. V B) and a large (Sec. V C) number N of neurons.

A. Catastrophic forgetting: observation of attractors

Firstly, we observe that during the learning phase, as the weights evolve continuously, attractors can not only be born, but can also disappear abruptly. One example of disappearance of a pair of attractors in the NN (1)–(2) with $N=81$ learning from the training Set 1 is illustrated in the phase portraits of Fig. 8. Namely, at $t=507.4$ (Fig. 8(a)), inside the blue circles there are two pairs of stable (black boxes) and saddle (white diamonds) fixed points. At $t=508.9$ (Fig. 8(b)), these pairs of points come very close to each other, and at $t=509.0$ (Fig. 8(c)) they no longer exist. Before the fixed points disappear, their largest eigenvalues approach zero: for the stable fixed points from the negative values, and for the saddle points

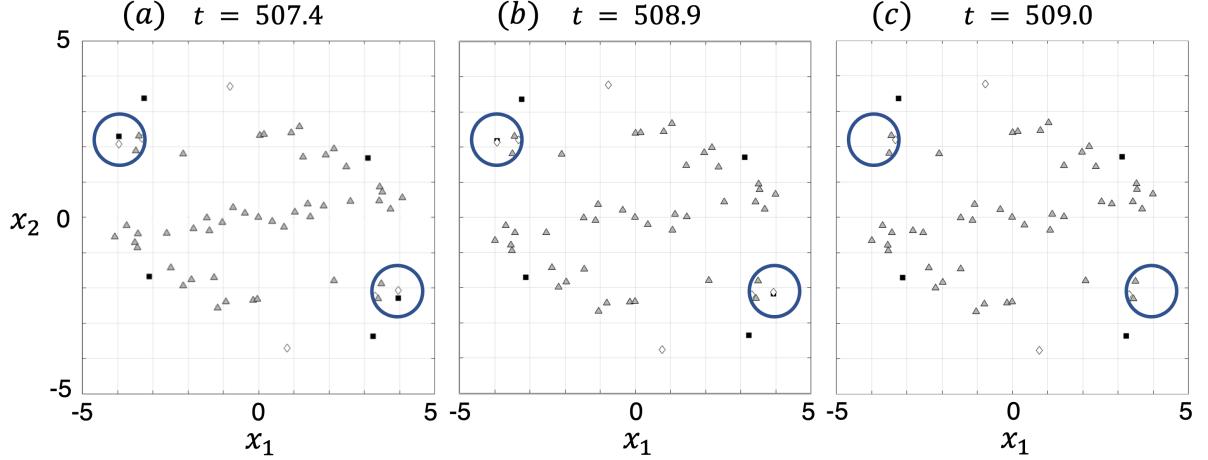


FIG. 8. Illustration of memory-destroying bifurcations, i.e. those underlying catastrophic forgetting. A pair of saddle-node bifurcations taking place in the NN (1)–(2) with $N=81$ learning from training Set 1, compare with Fig. 3(a)–(d). Panels show projections onto the (x_1, x_2) -plane of fixed points of the NN (6) (a) before the bifurcation at $t=507.4$, (b) immediately before the bifurcation at $t=508.9$, and (c) after the bifurcation at $t=509.0$, with t being the control parameter of (6) equal to time t of (1)–(2). Blue circles highlight areas of the phase space where bifurcations take place. Fixed points are marked as: attractors potentially associative with memories (black boxes), “useful” saddle points with one positive eigenvalue (white diamonds), and all other saddle or unstable points (grey triangles).

from the positive values. The phase portraits, together with eigenvalues, suggest that a pair of sub-critical saddle-node bifurcations took place.

If any of the attractors discussed above coded a valid category, their sudden death would constitute catastrophic forgetting.

B. Catastrophic forgetting in a small network: full evidence

It is well appreciated that a small NN cannot learn efficiently because of the highly limited number of attractors it could develop. However, before analysing the 81-neuron network with 3240 control parameters (weights), in this Subsection we consider a three-neuron network with only three distinct weights, because it permits bifurcation analysis in the space of *all* of its control parameters. This analysis will prepare the ground for the bifurcation analysis of the NN of size 81, which is done in Sec. V C.

The idea of our analysis is as follows. *First*, we will obtain a three-parameter bifurcation diagram of an autonomous system (6) with $N=3$ in the space $(\omega_{12}, \omega_{13}, \omega_{23})$, which will contain some two-dimensional bifurcation manifolds. Here, we will employ the assumptions of a standard bifurcation analysis, that the weights ω_{ij} are *not* parametrised by t and can take any values independently of each other.

Second, we will follow learning by the NN (1)–(2) with $N=3$, and will record the weight trajectory $\boldsymbol{\omega}(t)=(\omega_{12}(t), \omega_{13}(t), \omega_{23}(t))$ of subsystem (2). *Thirdly* and finally, in the parameter space $(\omega_{12}, \omega_{13}, \omega_{23})$ we will superimpose the bifurcation manifolds with the weight trajectory.

For the NN with $N=3$, the input signal $\mathbf{I}(t)$ was constructed from the training set of three-dimensional vectors \mathbf{I}^k ($k=1, 2, 3$) whose coordinates are ± 1 chosen at random with

equal probabilities (see Tab. S3 in Supplementary Notes). The procedures to form $\mathbf{I}(t)$ from the set of \mathbf{I}^k , and to generate initial conditions for x_i and ω_{ij} , are equivalent to those described in Sec. II for $N=81$ with suitable adjustments accounting for the different value of N and the number of vectors \mathbf{I}^k . The values of A and B are also the same as for $N=81$. We only change the value of g to $g=5$, because with $g=0.3$ the small NN produces no bifurcations while processing $\mathbf{I}(t)$.

The only bifurcation occurring in (6) with $N=3$ is a pitchfork bifurcation of the fixed point at $\mathbf{0}$, which destabilises $\mathbf{0}$ and creates a symmetric pair of attractors. To obtain a three-parameter bifurcation diagram, we fix ω_{23} and perform a standard two-parameter bifurcation analysis on the plane $(\omega_{12}, \omega_{13})$ using software XPPAUT⁵⁸. After revealing the curve of pitchfork bifurcation for the given ω_{23} , we change the value of ω_{23} and repeat the two-parameter bifurcation analysis. We do so for a set of values of ω_{23} sampled from the interval $[-0.1, 0.1]$ with the step 0.01. As a result, we obtain a collection of one-dimensional curves of pitchfork bifurcation. We then place these curves in $(\omega_{12}, \omega_{13}, \omega_{23})$ -space and use as a frame to render the manifold (the surface) of pitchfork bifurcation visualised in Fig. 9 as the blue surface marked S_{PF} .

Figure 9(a)–(b) shows how $\boldsymbol{\omega}(t)$ (red line) behaves relative to the bifurcation manifold S_{PF} (blue surface) during learning. At $t=0$ the trajectory starts inside the region bounded by S_{PF} , for which there is a single fixed point in the NN – a stable point at $\mathbf{0}$ signifying the absence of memories. Each “zag” (an approximately straight line segment) of the trajectory lasts approximately $t_s=12$ time units, which is equal to the time of application of a single vector \mathbf{I}^k ($k=1, 2, 3$) from the training set given in Tab. S3 of Supplementary Notes. The exception is the very first zag, which lasts approximately 13 time units. This is due to a delay of about 1 time unit between the end of the previous \mathbf{I}^k and the abrupt turn of the weight trajectory

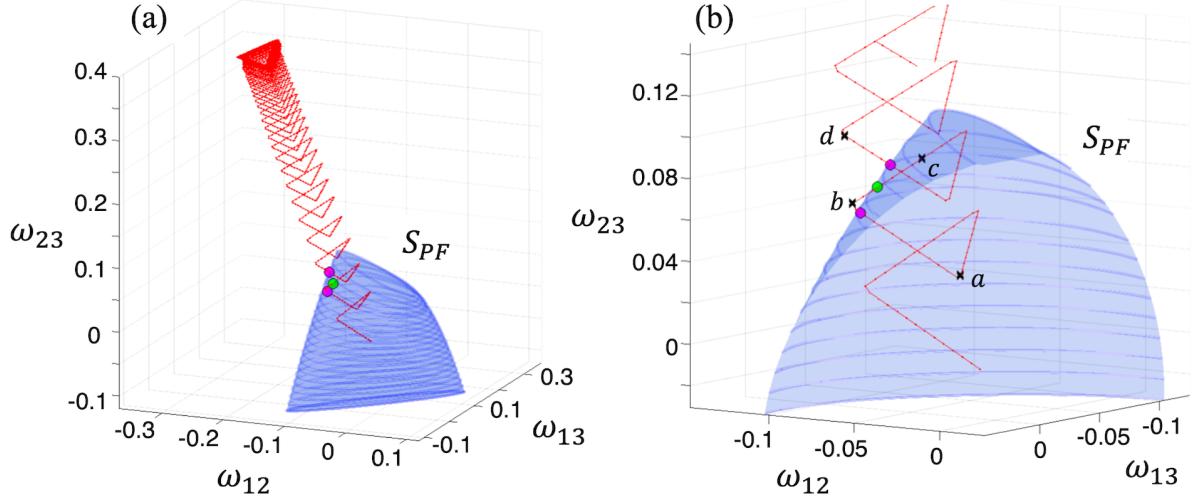


FIG. 9. Illustration of bifurcation mechanisms of memory formation and forgetting in a *small* NN (1)–(2) with $N=3$, $A=30$, $B=300$, $g=5$ and $\lambda=1.4$, trying to memorise vectors \mathbf{I}^k ($k=1, 2, 3$, see Tab. S3 of Supplementary Notes). Panels show a superposition in the space of weights $(\omega_{12}, \omega_{13}, \omega_{23})$ of a three-parameter bifurcation diagram of the NN (6) with $N=3$ and $g=5$ with the weight trajectory $\boldsymbol{\omega}(t)=(\omega_{12}(t), \omega_{13}(t), \omega_{23}(t))$ (red line) of subsystem (2) as the NN (1)–(2) learns. Panel (b) is a zoomed version of (a). S_{PF} (blue surface) is the manifold of pitchfork bifurcation of the fixed point at $\mathbf{0}$. Within the volume bounded by S_{PF} , the point $\mathbf{0}$ is stable; outside, the point $\mathbf{0}$ is saddle and there are additionally two stable fixed points. The trajectory $\boldsymbol{\omega}(t)$ (red line) starts near the origin inside the volume bounded by S_{PF} . Each zag of $\boldsymbol{\omega}(t)$ lasts approximately $t_s=12$ time units, but see text for a more accurate description. Within 48 time units, $\boldsymbol{\omega}(t)$ crosses S_{PF} three times by moving out (pink circle near point b in (b)), in (green circle in (a) and (b)) and out again (pink circle near point d in (b)). Panel (b) shows time stamps (black crosses) illustrated by phase portraits in Fig. 10.

in response to application of the next \mathbf{I}^k . Each m^{th} zag starts at approximately $(1+t_s(m-1))$ and ends at $(1+t_sm)$ time units. Note, that the period of $\mathbf{I}(t)$ is $3t_s=36$.

In Fig. 9, small dots on $\boldsymbol{\omega}(t)$ mark the values of t sampled each time unit, and panel (b) is a zoomed version of (a) showing the details of $\boldsymbol{\omega}(t)$ crossing S_{PF} .

By $t=36$, after making three zags, $\boldsymbol{\omega}(t)$ reaches point a (Fig. 9(b)). It is still inside the region bounded by S_{PF} , where the stable origin $\mathbf{0}$ is the only fixed point, as illustrated by the phase portrait in Fig. 10(a). During the next zag, at $t\approx48$, $\boldsymbol{\omega}(t)$ crosses S_{PF} (pink circle near point b in Fig. 9(b)), thus destabilising $\mathbf{0}$ and creating a pair of stable fixed points nearby. The phase portrait shortly after the pitchfork bifurcation, corresponding to $t=49$ and point b in Fig. 9(b), is given in Fig. 10(b) and shows two attractors (black boxes) potentially associative with memories.

During the next zag, at $t\approx52$, $\boldsymbol{\omega}(t)$ crosses S_{PF} again (green circle in Fig. 9(a)–(b)) and returns to the region bounded by S_{PF} , where it travels for a while. A phase portrait typical of this travel, corresponding to $t=56$ and marked by point c in Fig. 9(b), is given in Fig. 10(c). The fixed point $\mathbf{0}$ (signifying no memories) becomes stable again, and all other attractors (potential memories) disappear, which signifies catastrophic forgetting.

During the seventh zag that occurs for $t\in[73, 85]$, at $t\approx79$, $\boldsymbol{\omega}(t)$ crosses S_{PF} (pink circle near point c in Fig. 9(b)) and brings back two symmetric attractors. The phase portrait after the bifurcation, at $t=84$ (point d in Fig. 9(b)), is shown in Fig. 10(d). The attractor locations here are different from those at point b because the parameters ω_{ij} have different re-

spective values. After that, the trajectory $\boldsymbol{\omega}(t)$ moves away from S_{PF} in an oscillatory manner (Fig. 9(a)), never crosses S_{PF} again, and converges to a closed set resembling a limit cycle.

Note, that this small three-neuron network develops only a maximum of two attractors while trying – and failing – to memorise three patterns. This illustrates the well-known fact that a small NN cannot learn efficiently. Moreover, in the course of learning, due to a bifurcation, it catastrophically (albeit temporarily) loses all memories it manages to form. The above analysis is an explicit demonstration, in the case when the *full* bifurcation diagram in the space of weights can be obtained, that catastrophic forgetting is caused by the crossings between the weight trajectory and the bifurcation manifold.

C. Catastrophic forgetting in a large network: method and some evidence

To verify a hypothesis, that the causes of catastrophic forgetting are the crossings by the weight trajectory of the bifurcation manifolds, ideally, one needs a bifurcation diagram in the space of *all* weights of a NN. However, while it is theoretically possible to perform a bifurcation analysis of an 81-neuron network in the 3240-dimensional space of weights, doing so would not only be computationally demanding, but the results would be impossible to visualise using standard tools.

In Sec. V C 1 we present a possible form of an explicit evidence that could support the above hypothesis when a complete bifurcation diagram is unavailable. Given that one can

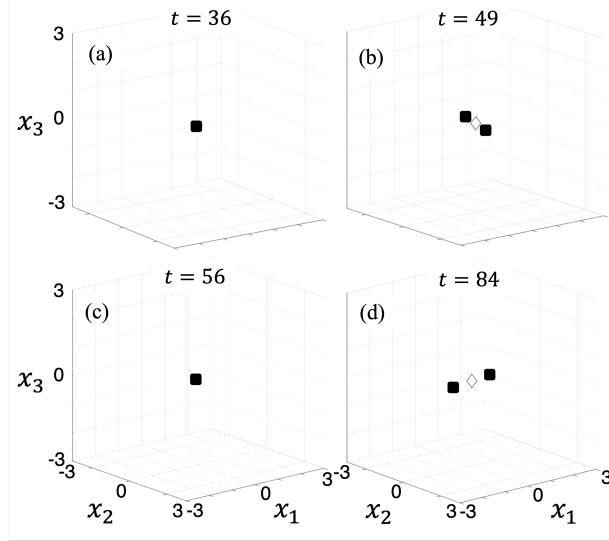


FIG. 10. Phase portraits illustrating stages of memory formation and forgetting in the small learning NN (1)–(2) described in caption to Fig. 9. Panels (a)–(d) show fixed points of (6) at various values of parameter t corresponding to various stages of learning by the NN (1)–(2), marked by black crosses on the weight trajectory in Fig. 9(b) next to matching letters a – d . Fixed points are: attractors (black boxes) and “useful” saddles with a single positive eigenvalue (white diamonds). Between (a) $t=36$ and (b) $t=49$ the NN increases the number of potential memories from one to two (black boxes). By (c) $t=56$ their number reverts to one, which could constitute forgetting. However, by (d) $t=84$ the NN has two attractors again (black boxes), which are placed at different locations compared to (b) because their birth occurs at a different combination of weights, as evidenced by different locations of two pink circles in Fig. 9(b).

feasibly visualise in a three-dimensional space *cross-sections* of the weight space and of the bifurcation manifolds, consider what these cross-sections would show when a bifurcation takes place during learning.

1. Crossing a bifurcation manifold: a detection method

Assume that there exists some bifurcation manifold S in the space of *all* weights of (6). As its counterpart NN (1)–(2) learns and t evolves, S remains fixed. For any value of t , at which the respective bifurcation does *not* occur, the current state $\omega(t)$ is at a finite distance from S . Only at the instant of bifurcation, $\omega(t)$ crosses S . For a small network, this situation is visualised in Fig. 9.

Now, consider *cross-sections* of S in a non-small NN (6) as its counterpart (1)–(2) undergoes various stages of learning. In order to visualise how $\omega(t)$ intersects with S , one needs to choose the cross-sections of S meaningfully. Figure 11 schematically illustrates the method we propose to achieve this. Assume that at some time $t=t_n$ during the learning phase, we wish to construct a cross-section $\widehat{S}(t)|_{t=t_n}=\widehat{S}(t_n)$ of S in a subspace of weights $(\omega_{i_1j_1}, \omega_{i_2j_2}, \omega_{i_3j_3})$. We propose to consider an intersection between S and a three-dimensional flat

surface in the M -dimensional weight space defined by the set of equations $\omega_{ij}=\omega_{ij}(t_n)$ for all ω_{ij} except the three selected weights $\omega_{i_1j_1}$, $\omega_{i_2j_2}$ and $\omega_{i_3j_3}$. Alternatively, $\widehat{S}(t_n)$ can be understood as the result of slicing S by $(M-3)$ mutually orthogonal hyper-planes of dimension $(M-1)$, each defined by $\omega_{ij}=\omega_{ij}(t_n)$ for appropriate ω_{ij} .

Thus, as learning progresses, for every new value $t=t_n$, there is a new secant surface and therefore a *new* cross-section $\widehat{S}(t_n)$ of the same S . This is illustrated in Fig. 11, where different cross-sections $\widehat{S}(t_n)$ (blue and grey surfaces) are schematically shown for three consecutive times t_n : (a) t_1 , (b) t_2 (bifurcation), and (c) t_3 , with $t_1 < t_2 < t_3$.

Now, consider a projection $\widehat{\omega}(t)$ (red line in Fig. 11, the same in (a)–(c)) of the weight trajectory $\omega(t)$ onto the $(\omega_{i_1j_1}, \omega_{i_2j_2}, \omega_{i_3j_3})$ -space relative to $\widehat{S}(t_n)$. At each $t=t_n$ when there is no bifurcation (Fig. 11(a) and (c)), the current state $\widehat{\omega}(t_n)$ (yellow circle) is at a finite distance from the respective $\widehat{S}(t_n)$. With this, there can exist intersections between the *full* $\widehat{\omega}(t)$ and the instantaneous cross-section $\widehat{S}(t_n)$ (green circles in Fig. 11(a) and (c)). However, these intersections do not signify the actual crossings in the space of *all* weights between $\omega(t)$ and S , and hence do not mark bifurcations. This is because for any point $\widehat{\omega}(t^*)$ on $\widehat{\omega}(t)$, at which the latter crosses $\widehat{S}(t_n)$ with $t^*\neq t_n$, its counterpart cross-section $\widehat{S}(t^*)$ is generally *different* from $\widehat{S}(t_n)$.

Only at an instant of the bifurcation, the *current* state $\widehat{\omega}(t_n)$ lands on the respective $\widehat{S}(t_n)$ (red circle in Fig. 11(b) corresponding to $t=t_2$). Other possible intersections between $\widehat{\omega}(t)$ and $\widehat{S}(t_2)$ (green circle in Fig. 11(b)) would not indicate bifurcations, as explained above.

2. Crossing a bifurcation manifold in a large NN: numerical evidence

In Sec. IV we observed that in the NN (1)–(2) with $N=81$, the primary mechanism of memory formation is the saddle-node bifurcation. Thus, for the NN learning from Set 1, as parameter t increases, we focus on the first pair of saddle-node bifurcations marked by the first pair of green circles in Fig. 4, illustrated by phase portraits in Fig. 5(c)–(d), and occurring at $t=22.8783$.

Following the method described in Sec. V C 1, below we visualise the instant when the weight trajectory $\omega(t)$ of the large NN crosses the manifold S_{SN} of the saddle-node bifurcation. In other words, we obtain a graph conveying the message of Fig. 11(b) for the NN under study. To do this, in (6) we fix all ω_{ij} , except ω_{12} , ω_{13} and ω_{23} , at the values they take in (1)–(2) at $t=22.8783=t_2$. We then perform a three-parameter bifurcation analysis of (6) as ω_{12} , ω_{13} and ω_{23} are allowed to vary freely and independently of each other.

This way, in the space $(\omega_{12}, \omega_{13}, \omega_{23})$ we obtain the cross-section $\widehat{S}_{SN}(t_2)$ (Fig. 12, blue surface) between S_{SN} and a three-dimensional flat surface in the M -dimensional space (with $M=3240$) defined by the set of equations $\omega_{ij}=\omega_{ij}(t_2)$ for all ω_{ij} excluding the three selected weights.

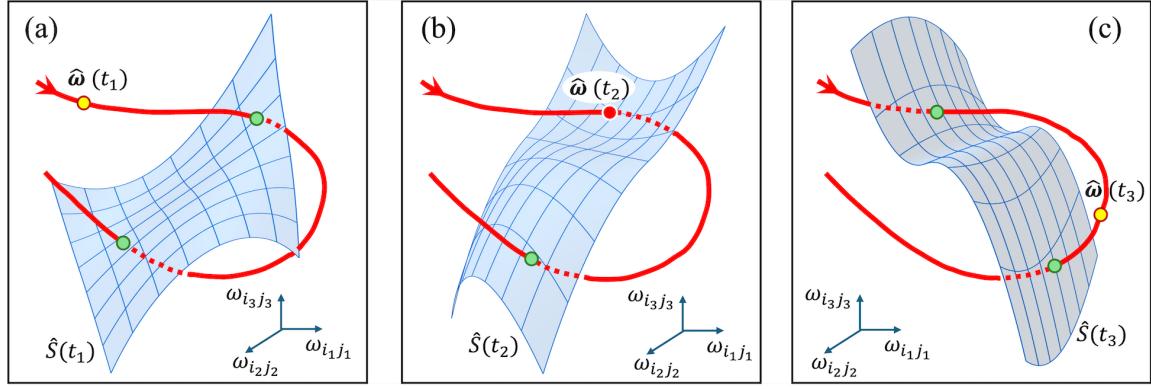


FIG. 11. A sketch illustrating a method to visualise bifurcation mechanisms of memory formation and forgetting in a non-small learning NN. Panels show cross-sections $\hat{S}(t_n)$ (blue surfaces) of the same bifurcation manifold S , which exists in the M -dimensional weight space of (6) with $N > 3$, with different three-dimensional flat surfaces defined by $\omega_{ij} = \omega_{ij}(t_n)$ for all ω_{ij} (except $\omega_{i_1 j_1}$, $\omega_{i_2 j_2}$ and $\omega_{i_3 j_3}$) at the values they take in (1)–(2) at time t_n . In the three-dimensional space of selected weights, for each $t=t_n$, there is a different $\hat{S}(t_n)$. Also shown is a projection $\hat{\omega}(t)$ (red line) of the weight trajectory $\omega(t)$ of (2). Panels illustrate the states (a) before the bifurcation at $t=t_1$ (yellow circle), (b) at the bifurcation at $t=t_2$ (red circle), and (c) after the bifurcation at $t=t_3$ (yellow circle), with $t_1 < t_2 < t_3$. The instant of bifurcation at $t=t_2$ is visualised in (b) as the crossing by $\hat{\omega}(t)$ (red circle) of $\hat{S}(t_2)$. In (a)–(c) other visible crossings between $\hat{\omega}(t)$ and $\hat{S}(t_n)$ (green circles) do not represent real crossings between $\omega(t)$ and S , since each green circle corresponds to t different from t_n , $n=1, 2, 3$, for which $\hat{S}(t_n)$ is constructed.

Specifically, to construct the cross-section $\hat{S}_{SN}(t_2)$, we sample the values of ω_{23} from $[-1, 1]$ with a step 0.1. For each fixed ω_{23} , a one-dimensional bifurcation curve was obtained in $(\omega_{12}, \omega_{13})$ -plane using XPPAUT⁵⁸. A collection of these curves was placed in $(\omega_{12}, \omega_{13}, \omega_{23})$ -space and used as a frame to render the surface $\hat{S}_{SN}(t_2)$.

Next, we register the weight trajectory of (2), i.e. $\omega(t) = (\omega_{12}(t), \dots, \omega_{(N-1)N}(t))$, such that $\omega_{ii}=0$, $\omega_{ij}=\omega_{ji}$, $\forall i, j=1, \dots, N$. The components of $\omega(t)$ are given in Fig. 2(a), and this $\omega(t)$ was used to produce Fig. 4. Finally, we superimpose the projection $\hat{\omega}(t)$ of $\omega(t)$ onto the $(\omega_{12}, \omega_{13}, \omega_{23})$ -space (red line in Fig. 12(a)–(b)) with $\hat{S}_{SN}(t_2)$.

Along the portion of $\hat{\omega}(t)$ visible slightly to the left of $\hat{S}_{SN}(t_2)$ in Fig. 12(b), i.e. at $6.5 < t < t_2$, the NN (6) has two attractors (see Fig. 5(c)). At $t=22.8783=t_2$, $\hat{\omega}(t)$ crosses $\hat{S}_{SN}(t_2)$ (red circle in Fig. 12), thus signifying the crossing between $\omega(t)$ and S_{SN} and the occurrence of the pair of saddle-node bifurcations giving rise to two more attractors. At $t_2 < t < 35.9$, the NN (6) has four attractors altogether (see Fig. 5(d)).

At $t=35.9$ two pairs of saddle-node bifurcations occur – to some other fixed points. Note, that the location of $\hat{\omega}(35.9)$ in Fig. 12(b) provides a rough idea of where the cross-section of the manifold of the latter bifurcation is located at $t=35.9$.

The additional crossings between $\hat{\omega}(t)$ and $\hat{S}_{SN}(t_2)$ (green circles in Fig. 12(b)) do not signify bifurcations, as explained in Sec. VC 1.

As t grows to very large values, $\hat{\omega}(t)$ converges to a closed curve, see condensation of red lines in the lower part of $\hat{\omega}(t)$ in Fig. 12(a). This is qualitatively similar to the behaviour of $\omega(t)$ in the small NN illustrated by Fig. 9(a). Thus, towards the end of learning, ω_{23} oscillates around $-\frac{1}{3}$, whereas ω_{12} and ω_{13} oscillate around zero (compare with Fig. 2(a)).

3. Catastrophic forgetting: evidence-based inference

When a complete bifurcation diagram of a NN in the space of its weights is unavailable, an explicit numerical proof, that catastrophic forgetting occurs when the weight trajectory crosses a bifurcation surface, can in principle be obtained by constructing sequences of appropriate cross-sections of the bifurcation manifolds. The method to achieve this is described in Sec. VC 1 and illustrated for an 81-neuron network in Sec. VC 2.

Recall that the Hopfield NN considered here starts from zero memories and acquires new memories as its weight trajectory goes through the space of weights and crosses various bifurcation manifolds. One example of such a crossing is explicitly demonstrated in Fig. 12. Catastrophic forgetting means that attractors, already formed in (1)–(2) by a certain stage of learning, should suddenly disappear as t keeps increasing continuously. This implies that as t grows, the weight trajectory must cross the same bifurcation manifolds it crossed previously, only in the opposite direction, as demonstrated for the small NN in Fig. 9.

To explicitly demonstrate this effect in a large NN, even for a single instance of catastrophic forgetting, one needs to construct a series of cross-sections of a certain bifurcation manifold S at several consecutive values of t . These values of t should include two instants of the respective bifurcation (when the same attractors are first born and then die), and in addition instants before the first bifurcation, in between the two bifurcations, and after the second bifurcation.

Numerical calculations of such cross-sections for non-small NNs are technically challenging, laborious and computationally demanding. Here we calculate only one of the required cross-sections in Fig. 12, whereas obtaining the complete and explicit numerical proof of the suggested mechanism of cata-

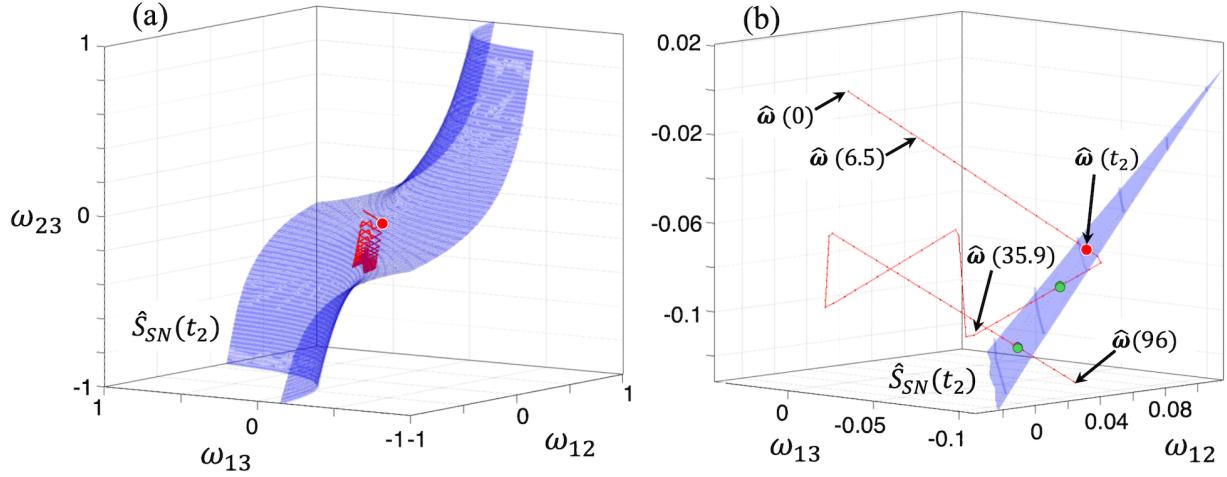


FIG. 12. Compare with Fig. 11(b). Numerical visualisation of a bifurcation responsible for the formation of two potential memories (attractors with basins) in a large NN (1)–(2) with $N=81$, $g=0.3$, $B=300$, $A=30$ learning from the training Set 1 (Tabs. S1–S2 of Supplementary Notes). $\hat{S}_{SN}(t_2)$ (blue surface) is a cross-section of a manifold S_{SN} of a pair of saddle-node bifurcations in the NN (6) with $N=81$ and $g=0.3$, visualised in the sub-space of weights $(\omega_{12}, \omega_{13}, \omega_{23})$ at $t=t_2=22.8783$. To obtain $\hat{S}_{SN}(t_2)$, all other ω_{ij} were fixed at values they take at $t=t_2$ as (1)–(2) undergoes the first pair of saddle-node bifurcations marked by the first pair of green circles in Fig. 4. Also in (a)–(b) there is a projection $\hat{\omega}(t)$ of the “weight trajectory” $\omega(t)$, $\hat{\omega}(t)=(\omega_{12}(t), \omega_{13}(t), \omega_{23}(t))$ (red line), of (2). At $6.5 < t < t_2$, the NN (1)–(2) has two attractors (Fig. 5(c)). At $t_2 < t < 35.9$, it has four attractors (Fig. 5(d)), two of which are born from the respective pair of saddle-node bifurcations. Panel (b) shows a portion of $\hat{\omega}(t)$ for $t \in [0, 96]$, during which eight input vectors I^k are applied consecutively with $k=1, 2, 3, 4, 5, 6, 1, 2$. In (a)–(b) red circle at intersection between $\hat{\omega}(t)$ and $\hat{S}_{SN}(t_2)$ signifies a real intersection between $\omega(t)$ and S_{SN} , i.e. the instant of bifurcation. In (b) green circles at other visible intersections between $\hat{\omega}(t)$ and $\hat{S}_{SN}(t_2)$ do not signify real bifurcations, see Sec. V C 1 and Fig. 11.

trophic forgetting will be the subject of our future work.

However, the numerical evidence collected here allows us to make an inference regarding the validity of our hypothesis. Firstly, we discover that throughout learning, the weight trajectory oscillates (zig-zags) with a considerable amplitude due to the applied periodic stimulus $I(t)$. Secondly, we observe that in the space of weights, the bifurcation manifolds are probably located quite close to each other. The latter is evidenced by the rather small distance in the subspace of weights between cross-sections of two different bifurcation manifolds: $\hat{S}_{SN}(t_2)$ and of another saddle-node bifurcation, which must go through the point $\hat{\omega}(35.9)$ in Fig. 12(b). Thirdly, Fig. 2 illustrates the well-known fact that, if the initial weights are inside the box $|\omega_{ij}| < 1$, the whole weight trajectory is confined to the same box. Fourthly, by the end of learning, the weight trajectory almost reaches some corners of the box $|\omega_{ij}| < 1$, since some weights tend to ± 1 (see Fig. 2). Fifthly, bifurcation manifolds can stretch beyond the boundaries of the box $|\omega_{ij}| < 1$ confining the weight trajectory, as seen in Fig. 12(a).

The five observations above suggest that there is quite a high probability that a zig-zagging weight trajectory repeatedly crosses the same bifurcation manifolds, rather tightly packed in the bounded volume of the space of weights, in opposite directions. If that happens, catastrophic forgetting occurs. Thus, the numerical facts already available provide a good evidence in favour of the hypothesis that catastrophic forgetting is caused by intersections between the weight trajectory and the bifurcation manifolds in large Hopfield NNs

with Hebbian learning.

VI. BASINS OF ATTRACTION

The question about representation of memories in NNs is fundamental, but remains largely unresolved. To date the most popular idea about the way an ANN represents a *single* memory is that it is represented by an attractor in the phase space of the NN^{11–13}. However, when the ANN is used to recognise a previously unseen input as belonging to a certain *category*, it is more appropriate to associate this category with the whole basin of attraction^{14,15}.

Therefore, the question about the size and the shape of attractor basins formed in a NN during learning becomes very important. Some investigations on the boundaries of attraction basins in small two- or three-neuron networks have been reported in Refs.^{59,60}. For large discrete-time and discrete-state NNs, it can be possible to evaluate some statistical characteristics of the radiiuses of attraction basins^{61–65}. However, detailed analysis of the sizes and shapes of attraction basins in continuous-state NNs with more than three neurons has not been done to the best of our knowledge.

Here we present an insight into the structure of the attraction basins of all attractors (Fig. 7) formed by the end of learning in the NN (1)–(2) with $N=81$ and the training Set 1. We visualise the attractor basins in several *two-dimensional* cross-

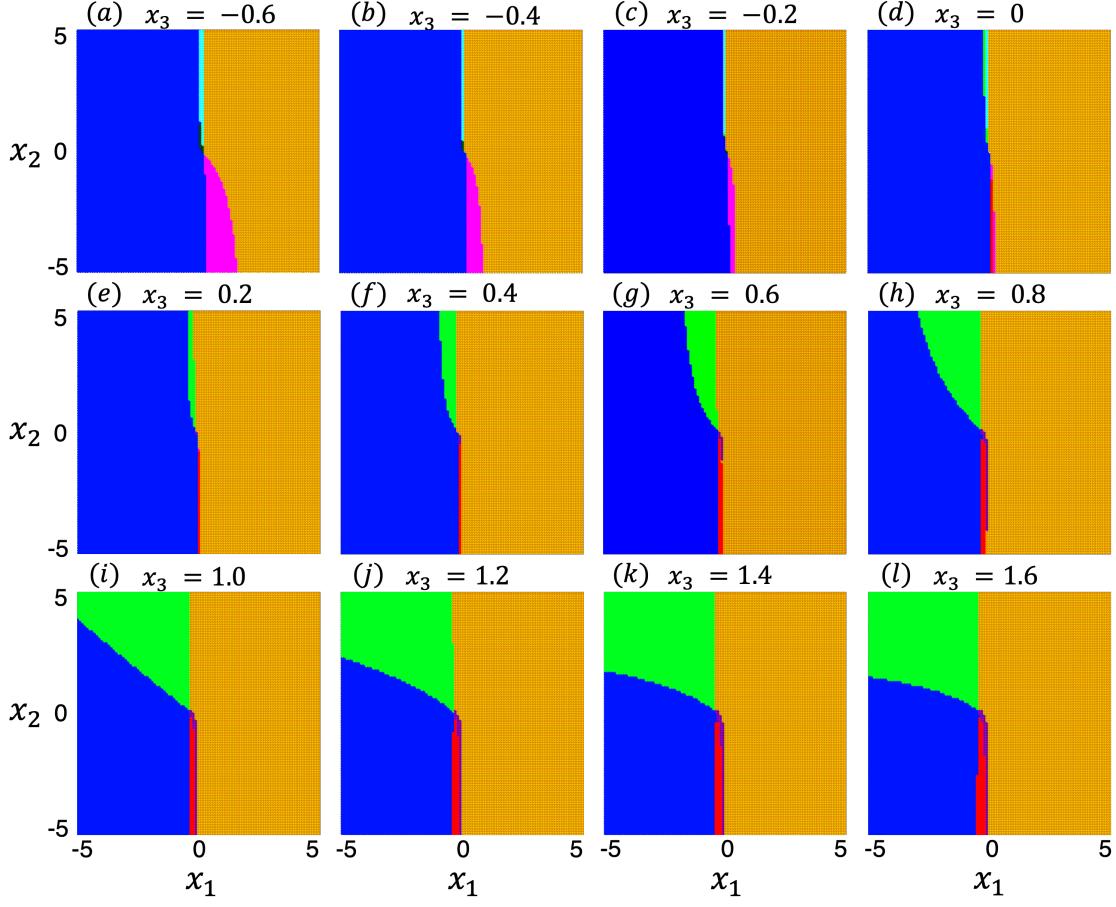


FIG. 13. Attractor basins (shaded regions in various colours) developed in the NN (1)–(2) with $N=81$, $g=0.3$, $B=300$ and $A=30$, by the end of learning from Set 1 at $t=6000$. Twelve different two-dimensional cross-sections (a)–(l) of the same set of basins are shown as the phase space intersects with different planes defined by (7), where x_3^* is displayed above each panel. Colours of attractor basins match those of respective attractors in Fig. 7. Attractors and their basins are found for the autonomous version (6) of the NN, where ω_{ij} are set to values achieved by (1)–(2) at the end of learning.

sections of the phase space of (6) with the planes defined by

$$x_3 = x_3^*, \quad x_4 = 0, \quad \dots, \quad x_{81} = 0, \quad (7)$$

where x_3^* is a value in $[-0.6, 1.6]$ sampled with step 0.2.

In practice, to numerically calculate the basin cross-sections, we prepare a grid of initial conditions (ICs) on the secant plane representing the plane (7). Namely, on (x_1, x_2) -plane we split a region $x_{1,2} \in [-5, 5]$ into square boxes with the side 0.1 to obtain 101×101 nodes. The vectors of ICs are then formed by taking (x_1, x_2) -coordinates of these nodes and augmenting them by the values of x_3, \dots, x_{81} defined by (7).

For each IC, we register an attractor to which the phase trajectory eventually converges. Then on the secant plane represented by (x_1, x_2) -coordinates (since all other coordinates are fixed as in (7)) we mark this IC by the colour matching the colour of the respective attractor in Fig. 7.

The cross-sections of the same set of attractor basins are shown in Fig. 13, where different panels correspond to secant planes (7) that differ only in the values of x_3 , as indicated above each panel. Note, that attractors themselves cannot be seen in the cross-sections considered, since they do

not belong to the respective secant planes. The gradual increase of x_3 from (a) to (l) allows one to reveal the structure of the basins in a *three-dimensional* cross-section of the 81-dimensional phase space. The incremental reshaping of the two-dimensional cross-sections of the basin boundaries with the gradual increase of x_3 from (a) to (l) is an evidence of their continuity in the full space. One can also observe the complex shapes of these boundaries, and notice different basin sizes.

Interestingly, the basins with the largest sizes in the chosen cross-sections (blue and yellow shades) belong to the symmetric pair of attractors born first in the course of training from the pitchfork bifurcation marked by the first pink circle in Fig. 4(a). Note the symmetry of their basins visible in Fig. 13(d), which corresponds to $x_3=0$, implying that the secant plane goes through the origin. The largest visible sizes of these attractor basins could be explained by the fact that our secant planes are quite close to the origin, in whose vicinity the first attractors emerged. Other attractors are born later and further away from the origin, so their basins occupy much smaller areas in the given cross-sections.

It is important to understand how the *boundaries* of attrac-

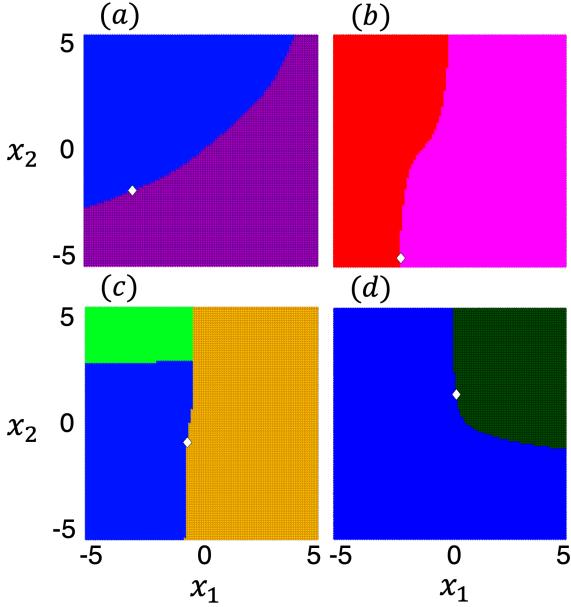


FIG. 14. Evidence that the boundaries of attractor basins are formed by the stable manifolds of the saddle fixed points with one positive eigenvalue (white diamonds). Here, more cross-sections are shown of the same set of attractor basins as those illustrated in Fig. 13. Namely, every panel shows a cross-section with a different plane, each going through a different saddle point (white diamond), see details in text. Note, that the saddles lie exactly on the basin boundaries. Colours of attractor basins (shaded regions) match those of respective attractors in Fig. 7 and of their basins in Fig. 13.

tion basins (or separatrices) are formed. They should be represented by some $(N-1)$ -dimensional manifolds. Given that the only limit sets in the NN (6) are fixed points, suitable manifolds should belong to the saddle fixed points with a single positive eigenvalue marked by diamonds in Fig. 7.

To verify this, we calculate a few more cross-sections of the same basins of attraction as those illustrated in Fig. 13, but this time obtained with the secant planes going directly through (as close as the numerical error allows) various saddle points with exactly one positive eigenvalue. Namely, we consider cross-sections of the 81-dimensional space with four different planes defined by $x_i = s_i^m$, $i=3,\dots,81$, $m=1,\dots,4$, where s_i^m is the i^{th} coordinate of the m^{th} saddle point chosen from points shown in Fig. 7 as orange diamonds. If our assumption is true, in the given cross-sections, the respective saddle points should belong to the basin boundaries.

In Fig. 14(a)–(d) the resultant four cross-sections of attractor basins are shown together with the respective saddle points s^m (white diamonds). The colour-coding of the basins is the same as in Fig. 13. One can see that, within the numerical error, the saddles lie on the basin boundaries, which therefore must coincide with the stable manifolds of these saddles.

VII. SUMMARY AND DISCUSSION

a. Summary of results. We comprehensively studied bifurcation mechanisms of memory formation and catastrophic forgetting in an archetypal relatively *large* recurrent ANN – the continuous-time Hopfield NN with Hebbian learning. We proposed a rigorous approach to perform this analysis in a NN of any size. Our analysis demonstrates that, when the NN starts learning with no attractors potentially associative with memories, the *same* bifurcation mechanism underlies two opposite effects: memory formation and destruction.

We discovered how *spurious* memories are formed in the NN with identical neurons and symmetric connections. Namely, when bifurcations occur in the course of learning, attractors are born in symmetric pairs, quads, octets, etc. Normally, only one of the simultaneously born attractors is associated with the new stimulus, whereas the rest become spurious memories at least for some part of the training phase.

We also revealed the structure of attractor *basins* formed as a result of learning, and their link to the saddle fixed points. This research highlights the importance for memory formation not only of the attractors, but of the less visible saddle objects. Our work is a rare example of studies and visualisations of truly high-dimensional and multi-parameter nonlinear DSs. The findings of this paper agree well with the mechanisms of real and spurious memory formation considered from the view point of dynamical systems with plastic self-organising vector fields (termed “plastic dynamical systems” for brevity)³⁶.

b. Significance of the method for bifurcation analysis in learning NNs. Although memory formation in ANNs has long been associated with bifurcations theoretically^{14,21}, to date this connection has not been rigorously verified. The primary reason is the dimensionality of the parameter (weight) space of NNs, which is abnormally and intractably high for the standard tools of bifurcation analysis. The existing numerical evidence for bifurcations occurring during learning^{23,37} is indirect, non-rigorous and incomplete due to the absence of a well-defined continuously changing bifurcation parameter linkable to the stage of learning.

Our proposal of a rigorous method to analyse bifurcations in a learning recurrent NN is based on formally introducing its two versions. The first version is a *learning* NN (such as (1)–(2)) whose state and weights evolve with continuous time t , where t is the stage of learning. The second version is a *non-learning* NN with fixed weights (such as (6)), which evolves with a different time t' , and for which all weights are formally functions of a single parameter t that can serve as the bifurcation parameter required. In this setting, the bifurcation analysis of a learning NN is formally reduced to the standard bifurcation analysis of a non-learning NN depending on a single parameter t .

This reduction is applicable to any recurrent NN of any dimension employing any learning algorithm. Bifurcation analysis can be a powerful tool to explain operations of learning NNs, even in cases where energy functions are not defined, such as Hopfield-type NNs with asymmetric weights.

c. Broader significance of results. Our study takes a significant step towards demystifying the “black box” of recurrent learning ANNs by integrating isolated aspects of their operation as fragments of a larger unified picture. Indeed, although the Hopfield NN with Hebbian learning is different from recurrent ANNs that are more popular in machine learning, they share the same learning principle: guiding the weight trajectory to a point in the weight space at which the ANN has the desired configuration of attractors and basins.

There are many ways to choose the start, the end, and the interim steps of this journey. With Hebbian learning, the weight trajectory is formed as part of self-organisation of the whole NN being a non-autonomous non-linear dissipative DS.

In machine learning, the choice of steps is justified by various criteria often based on the loss function. The prevailing preference is to initialise the weights at values ensuring a good number of attractors, and to avoid bifurcations during learning in order to prevent gradient explosions of the loss^{21–23}. However, the appropriateness of the initial weights vector depends on its location relative to bifurcation manifolds, which for non-small NNs cannot realistically be found. Also, the way a new stimulus is associated with one of the existing attractors is unclear. Moreover, such learning pre-conditions the existence of attractors not associated with stimuli (spurious memories) during much or all of the training phase.

With this, the mechanisms of memory formation similar to those uncovered in the sample recurrent ANN can be used to prepare a network for conventional learning in a relatively predictable and controlled way. One can guide the NN through the necessary number of attractor-creating bifurcations. If there are not enough attractors at any stage of learning, one could increase their number by forcing the weight trajectory to cross bifurcation manifolds in appropriate directions. One could also avoid catastrophic forgetting by preventing unwanted crossings of bifurcation manifolds.

Overall, by exposing the relationship between memory formation, catastrophic forgetting, and bifurcations in learning recurrent ANNs, our findings can help create learning algorithms targeting the *causes* of unwanted or desirable effects. Specifically, these algorithms could utilise methods from the theory of non-linear DSs for the control of attractors^{66–69} and of bifurcations^{70–72}.

d. Future work. The hypothesis about the causes of catastrophic forgetting cannot be verified by simply detecting crossings between the weight trajectory and bifurcation manifolds, since these manifolds cannot be calculated (except for small NNs, see Sec. V B). Our approach to do this using a series of cross-sections of these manifolds is still technically demanding and laborious, but is feasible with the computational tools available and could be used in the future for a more thorough verification of this hypothesis.

Also, bifurcation analysis may help reveal the learning mechanisms in deep ANNs – the most powerful yet poorly understood AI available to date¹⁶.

VIII. AUTHOR CONTRIBUTIONS

AEE performed all numerical calculations reported in paper, wrote the first draft and edited the final draft. NBJ conceived this research. RN produced the first preliminary results, not illustrated in paper, under the supervision of NBJ. AEE, NBJ and AGB analysed the data and interpreted the results. NBJ and AGB supervised and conceptualised the work, and co-wrote the final version of the paper. AGB coordinated the paper preparation.

IX. DATA AVAILABILITY STATEMENT

No data was generated during this research.

X. ACKNOWLEDGEMENTS

AEE received funding from EPSRC (UK) to support his PhD studies under grant EP/W523987/1.

- ¹Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature* **521**, 436–444 (2015).
- ²L. Alzubaidi *et al.*, “Review of deep learning: concepts, cnn architectures, challenges, applications, future directions,” *J. Big Data* **8**, 53 (2021).
- ³D. Amit, *Modeling brain function: The world of attractor neural networks* (Cambridge University Press, 1989).
- ⁴M. McCloskey and N. J. Cohen, “Catastrophic interference in connectionist networks: The sequential learning problem,” *Psychol. Learn. Motiv.* **24**, 109–165 (1989).
- ⁵R. French, “Catastrophic forgetting in connectionist networks,” *Trends Cogn. Sci.* **3**, 128–135 (1999).
- ⁶J. Kirkpatrick *et al.*, “Overcoming catastrophic forgetting in neural networks,” *PNAS* **114**, 3521–3526 (2017).
- ⁷R. Kemker, M. Mcclure, A. Abitino, T. Hayes, and C. Kanan, “Measuring catastrophic forgetting in neural networks,” *Proceedings of the AAAI Conference on Artificial Intelligence* **32**, 3390–3398 (2018).
- ⁸J. Yoon, E. Yang, J. Lee, and S. J. Hwang, “Lifelong learning with dynamically expandable networks,” (2018), arXiv:1708.01547 [cs.LG].
- ⁹G. Parisi, “Continual lifelong learning with neural networks: A review,” *Neural Netw.* **113**, 54–71 (2019).
- ¹⁰G. Marcus, “Deep learning: a critical appraisal,” (2018).
- ¹¹S.-I. Amari, “Learning patterns and pattern sequences by self-organizing nets of threshold elements,” *IEEE Transactions on Computers* **C-21**, 1197–1206 (1972).
- ¹²J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554–2558 (1982).
- ¹³J. Hopfield, “Neurons with graded response have collective computational properties like those of two-state neurons,” *Proc. Natl. Acad. Sci. U.S.A.* **81**, 3088–3092 (1984).
- ¹⁴F. J. Pineda, “Dynamics and architecture for neural computation,” *J. Complex.* **4**, 216–245 (1988).
- ¹⁵D. Barack and J. Krakauer, “Two views on the cognitive brain,” *Nat. Rev. Neurosci.* **22**, 359–371 (2021).
- ¹⁶S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, “Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence,” *Information Fusion* **99**, 101805 (2023).
- ¹⁷R. Ratcliff, “Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions,” *Psychological Review* **97**, 285–308 (1990).

- ¹⁸A. Robins and S. McCallum, "Catastrophic forgetting and the pseudorehearsal solution in Hopfield-type networks," *Connection Science* **10**, 121–135 (1998).
- ¹⁹A. Rusu, N. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," arXiv:1606.04671 (2016).
- ²⁰H. Fayek, L. Cavedon, and H. Wu, "Progressive learning: A deep learning framework for continual learning," *Neural Networks* **128**, 345–357 (2020).
- ²¹K. Doya, "Bifurcations of recurrent neural networks in gradient descent learning," *IEEE Transactions on neural networks* **1**, 218 (1993).
- ²²R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 28(3), edited by S. Dasgupta and D. McAllester (PMLR, Atlanta, Georgia, USA, 2013) pp. 1310–1318.
- ²³U. Haputhanthri, L. Storan, Y. Jiang, A. Shai, H. A. Orhun, M. Schnitzer, F. Dinc, and H. Tanaka, "Why do recurrent neural networks suddenly learn? Bifurcation mechanisms in neuro-inspired short-term memory tasks," in *ICML 2024 Workshop on Mechanistic Interpretability* (2024).
- ²⁴M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlöterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," *ACM Comput. Surv.* (2023).
- ²⁵R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu, "Explaining deep neural networks: A survey on the global interpretation methods," *Neurocomputing* **513**, 165–180 (2022).
- ²⁶D. Hebb, *The organization of behavior; a neuropsychological theory* (Wiley, Oxford, England, 1949).
- ²⁷W. Gerstner and W. Kistler, "Mathematical formulations of Hebbian learning," *Biol. Cybern.* **87**, 404–415 (2002).
- ²⁸Z. Yu, A. Abdulghani, A. Zahid, H. Heidari, M. Imran, and O. Abbasi, "An overview of neuromorphic computing for artificial intelligence enabled hardware-based hopfield neural network," *IEEE Access* **8**, 67085–67099 (2020).
- ²⁹D. Krotov, "A new frontier for Hopfield networks," *Nat. Rev. Phys.* **5**, 366–367 (2023).
- ³⁰D. Dong and J. Hopfield, "Dynamic properties of neural networks with adapting synapses," *Network: Computation in Neural Systems* **3**, 267–283 (1992).
- ³¹R. Körberle, "Neural networks as content addressable memories and learning machines," *Comput. Phys. Commun.* **56**, 43–50 (1989).
- ³²N. Janson and C. Marsden, "Dynamical system with plastic self-organized velocity field as an alternative conceptual model of a cognitive system," *Sci. Rep.* **7**, 17007 (2017).
- ³³N. Janson and P. Kloeden, "Robustness of a dynamical systems model with a plastic self-organising vector field to noisy input signals," *Eur. Phys. J. Plus* **136**, 720 (2021).
- ³⁴L. Alzubaidi, J. Zhang, A. Humaidi, et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data* **8**, 53 (2021).
- ³⁵P. Kloeden and M. Yang, *An Introduction to Nonautonomous Dynamical Systems and their Attractors* (World Scientific, 2020) <https://www.worldscientific.com/doi/pdf/10.1142/12053>.
- ³⁶N. Janson, A. Essex, and A. Balanov, "Designing explainable cognitive systems and explaining neural networks with plastic dynamical systems," <http://dx.doi.org/10.2139/ssrn.5123206> (2024), SSRN:ssrn.5123206.
- ³⁷A. Ribeiro, K. Tiels, L. Aguirre, and T. Schön, "Beyond exploding and vanishing gradients: analysing rnns training using attractors and smoothness," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 108, edited by S. Chiappa and R. Calandra (PMLR, 2020) pp. 2370–2380.
- ³⁸Global bifurcations are irrelevant to this study.
- ³⁹I. Kuznetsov, *Elements of applied bifurcation theory*, Vol. 112 (Springer Verlag, 1998).
- ⁴⁰E. Allgower and K. Georg, *Numerical Continuation Methods*, Vol. 13 (Springer (Berlin, Heidelberg), 1990).
- ⁴¹E. Doedel, A. Champneys, T. Fairgrieve, Y. Kuznetsov, B. Sandstede, and X. Wang, "AUTO97: Continuation and bifurcation software for ordinary differential equations (with HomCont)," *Tech. Rep. (Technical Report, Concordia University, 1997)*.
- ⁴²B. Ermentrout, "Simulating, analyzing, and animating dynamical systems: A guide to xppaut for researchers and students," *SERBIULA (sistema Libre 2.0)* (2002), 10.1137/1.9780898718195.
- ⁴³P. Das and W. Schieve, "A bifurcation analysis of the four dimensional generalized hopfield neural network," *Phys. D: Nonlinear Phenom.* **88**, 14–28 (1995).
- ⁴⁴R. Beer, "On the dynamics of small continuous-time recurrent neural networks," *Adaptive Behavior* **3**, 469–509 (1995).
- ⁴⁵R. Haschke and J. J. Steil, "Input space bifurcation manifolds of recurrent neural networks," *Neurocomputing* **64**, 25–38 (2005), trends in Neurocomputing: 12th European Symposium on Artificial Neural Networks 2004.
- ⁴⁶W.-Z. Huang and Y. Huang, "Chaos of a new class of hopfield neural networks," *Applied Mathematics and Computation* **206**, 1–11 (2008).
- ⁴⁷J. Cervantes-Ojeda, M. Gómez-Fuentes, and R. Bernal-Jaquez, "Empirical analysis of bifurcations in the full weights space of a two-neuron DTRNN," *Neurocomputing* **237**, 362–374 (2017).
- ⁴⁸Z. Njitacke, J. Kengne, T. Fozin, B. Leutcha, and H. Fotsin, "Dynamical analysis of a novel 4-neurons based hopfield neural network: emergences of antimonotonicity and coexistence of multiple stable states," *Int. J. Dyn. Contr.* **7**, 823–841 (2019).
- ⁴⁹Z. Hu and C. Wang, "Hopfield neural network with multi-scroll attractors and application in image encryption," *Multimedia Tools and Applications* **83**, 97–117 (2023).
- ⁵⁰R. Sharpe and R. Thorne, "Numerical method for extracting an arc length parameterization from parametric curves," *Computer-Aided Design* **14**, 79–81 (1982).
- ⁵¹J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," *PNAS* **81**, 3088–3092 (1984).
- ⁵²L. Smith and D. Liles, "The effects of exposure time and retention time on location memory in visual information processing," *Proc. of the Human Factors Soc. Annual Meeting* **26**, 812–815 (1982).
- ⁵³K. Doya, "Bifurcations in the learning of recurrent neural networks," in *[Proceedings] 1992 IEEE International Symposium on Circuits and Systems*, Vol. 6 (1992) pp. 2777–2780.
- ⁵⁴It is worth noting that the knowledge of the weights alone could be insufficient to establish their connection to the categories they supposedly represent, see discussion and references in Refs.^{32,36}.
- ⁵⁵G. Raghavan, B. Tharwat, S. N. Hari, D. Satani, R. Liu, and M. Thomson, "Engineering flexible machine learning systems by traversing functionally invariant paths," *Nat. Mac. Intell.* **6**, 1179–1196 (2024).
- ⁵⁶L. Eisenmann, Z. Monfared, N. Göring, and D. Durstewitz, "Bifurcations and loss jumps in rnns training," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23 (Curran Associates Inc., Red Hook, NY, USA, 2023).
- ⁵⁷O. Makarenkov and J. S. Lamb, "Dynamics and bifurcations of nonsmooth systems: A survey," *Physica D: Nonlinear Phenomena* **241**, 1826–1844 (2012), dynamics and Bifurcations of Nonsmooth Systems.
- ⁵⁸B. Ermentrout, "XPPAUT," *Scholarpedia* **2**, 1399 (2007).
- ⁵⁹J. Mira and J. Álvarez, eds., *Computational Methods in Neural Modeling* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2003).
- ⁶⁰W. Krauth, M. Mezard, and J.-P. Nadal, "Basins of attraction in a perceptron-like neural network," *Complex Systems* **2**, 387–408 (1988).
- ⁶¹A. Storkey and R. Valabregue, "The basins of attraction of a new hopfield learning rule," *Neural Networks* **12**, 869–876 (1999).
- ⁶²N. Davey and S. P. Hunt, "The capacity and attractor basins of associative memory models," in *Foundations and Tools for Neural Modeling*, edited by J. Mira and J. V. Sánchez-Andrés (Springer Berlin Heidelberg, Berlin, Heidelberg, 1999) pp. 330–339.
- ⁶³F. Zhang and X. Zhang, "The average radius of attraction basin of hopfield neural networks," in *Advances in Neural Networks - ISNN 2008*, edited by F. Sun, J. Zhang, Y. Tan, J. Cao, and W. Yu (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008) pp. 253–258.
- ⁶⁴C. Lin, T. H. Yeap, and I. Kiringa, "On the basin of attraction and capacity of restricted hopfield network as an auto-associative memory," in *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)* (2023) pp. 146–154.
- ⁶⁵S. Sampath and V. Srivastava, "On stability and associative recall of memories in attractor neural networks," *PLoS ONE* **15** (2020).

- ⁶⁶P. Parmananda and M. Eiswirth, "Stabilizing unstable fixed points using derivative control," *J. Phys. Chem.* **100**, 16568–16570 (1996).
- ⁶⁷J. Claussen, T. Mausbach, A. Piel, and H. Schuster, "Memory difference control of unknown unstable fixed points: Drifting parameter conditions and delayed measurement," *Phys. Rev. E* **58**, 7256–7260 (1998).
- ⁶⁸B. Epureanu and E. Dowell, "Optimal multi-dimensional OGY controller," *Physica D* **139**, 87–96 (2000).
- ⁶⁹A. Ahlborn and U. Parlitz, "Stabilizing unstable steady states using multiple delay feedback control," *Phys. Rev. Lett.* **93** (2004).
- ⁷⁰E. Abed and J. Fu, "Local feedback stabilization and bifurcations control, I. Hopf bifurcation," *Syst. Control Lett.* **7**, 11–17 (1986).
- ⁷¹E. Abed and J. Fu, "Local feedback stabilization and bifurcations control, II. Stationary bifurcation," *Syst. Control Lett.* **8**, 467–473 (1987).
- ⁷²G. Chen, J. Moiola, and H. Wang, "Bifurcation control: Theories, methods, and applications," *Int. J. Bifurcat. Chaos* **10**, 511–548 (2000).