RNNs with gracefully degrading continuous attractors

Ábel. Ságodi, Piotr Sokół, Il Memming Park

Champalimaud Centre for the Unknown
Chmampalimaud Foundation, Lisbon, Portugal
{abel.sagodi,piotr.sokol,memming.park}@research.fchampalimaud.org

Abstract

Attractor networks are essential theoretical components in recurrent networks for memory, learning, and computation. However, the continuous attractors that are essential for continuous-valued memory suffer from structural instability—infinitesimal changes in parameter can destroy the continuous attractor. Moreover, the perturbed system can exhibit divergent behavior with associated exploding gradients. This poses a question about the utility of continuous attractors for systems that learn using gradient signals. To address this issue, we use Fenichel's persistence theorem from dynamical systems theory to show that bounded attractors are stable in the sense that all perturbations maintain the stability. This ensures that if there is a restorative learning signal, there will be no exploding gradients for any length of time for backpropagation. In contrast, unbounded attractors may devolve into divergent systems under certain perturbations, leading to exploding gradients. This insight also suggests that there can exist homeostatic mechanisms for certain implementations of continuous attractors that maintain the structure of the attractor sufficiently for the neural computation it is used in. Finally, we verify in a simple continuous attractor that all perturbations preserve the invariant manifold and demonstrate the principle numerically in ring attractor systems.

1 Introduction

Recurrent neural/neuronal networks (RNNs) can process sequential observations and model temporal dependencies of arbitrary length. At the same time, they are fundamentally limited by their finite-sized hidden states which form the only channel between the past and the future. To store information over a long period of time, as many difficult tasks demand, RNNs can learn or be designed to have "persistent memory". When the information of interest is continuous-valued, a natural solution is to use continuous attractors. Continuous attractors are prevalent in theoretical neuroscience as tools to model neural representation and computation ranging from internal representations of head directions and eye positions to perceptual decision-making and working memory [21]. Continuous attractors are also at the core of long short-term memory (LSTM) [17] units and the neural Turning machine (NTM) [16] to provide digital computer memory like properties not natural to recurrent networks. In fact, the critical weakness of continuous attractors is their inherent brittleness as they are rare in the parameter space, i.e., infinitesimal changes in parameters destroys the continuous attractor structures implemented in RNNs [30, 34]. However, not all RNN implementations of continuous attractors behave similarly in their brittleness.

We found that in the space of RNNs, some have neighbourhoods with highly undesirable exploding gradients. We will describe some continuous attractors which have such neighbourhoods. Consider an RNN (without input or output for now) expressed in continuous time as an ordinary differential equation:

$$\dot{\mathbf{x}} = -\mathbf{x} + [\mathbf{W}\mathbf{x} + \mathbf{b}]_+ \tag{1}$$

Preprint. Under review.

where $\mathbf{x} \in \mathbb{R}^d$ is the hidden state of the network, $\mathbf{b} > 0$ is the bias, and $[\cdot]_+ = \max(0, \cdot)$ is the threshold nonlinearity per unit. In discrete time, this corresponds to a ReLU RNN (see Sec. 3.2). The non-trivial activity of this network is limited to the (non-negative) first quadrant, outside of which the hidden state decays to the origin.

When $d = 2$, we can build two kinds of continuous attractors. First, through positive feedback, $\mathbf{W} = [0, 1; 1, 0]$ and no bias $\mathbf{b} = \mathbf{0}$, we can create a continuous attractor, i.e., $\dot{\mathbf{x}} = 0$ on the $x_1 = x_2 \geq 0$ half-line, and surrounding attractive flow (Fig. 1A left). We refer to it as an **unbounded line attractor (UBLA)**. For any point on the line attractor, linearization results in eigenvalues 0 and -2, corresponding to the zero flow and attractive flow respectively. When $\mathbf{W}$ is perturbed, the null eigenvalue can easily become non-zero and the continuous line attractor disappears. If it becomes negative, the system bifurcates to a stable fixed point at the origin (Fig. 1A bottom). However, if it becomes positive (Fig. 1A top), *the resulting flow diverges to infinity along the diagonal*. Corresponding to the divergent flow, the backpropagating gradient over time exponentially grows in magnitude, thus rendering gradient descent impractical without truncation in time.

The second kind of continuous attractor is created through negative feedback. By choosing $\mathbf{W} = [0, -1; -1, 0]$ and $\mathbf{b} = [1; 1]$, we get $\dot{\mathbf{x}} = 0$ on the $x_1 = -x_2 + 1$ line segment in the first quadrant as the continuous attractor. We refer to it as the **bounded line attractor (BLA)**. Again, linearization on the attractor shows two eigenvalues, 0 and -2, and perturbations again cause the null eigenvalue to be non-zero and the line attractor disappears. However, surprisingly, the bifurcations are qualitatively different. It either bifurcates into a single stable fixed point (Fig. 1B top) or two stable fixed points separated with a saddle node in between (Fig. 1B bottom). Neither of these two cases show a divergent flow, but rather consists of one or two basins of attraction. It implies only vanishing gradients for this system and **exploding gradients will not be present for an arbitrarily long time**.

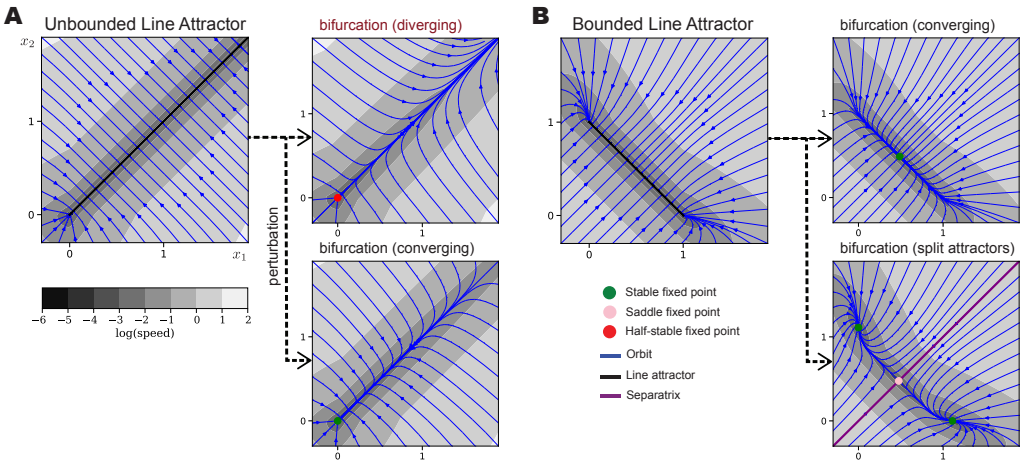

Figure 1: Motivating case study of the two systems implementing the same computation but one near exploding gradients. Phase portraits for unbounded and bounded linear attractors (1). Under perturbation of parameters, each of them can bifurcate to one of the two potential systems without the continuous line attractor. Note that the parameters for the UBLA are near a diverging system associated with exploding gradient behavior.

Avoiding exploding gradients is of paramount importance in the context of long-range temporal learning (Sec. 2.2). Learning generally induces stochasticity in parameters, while spontaneous synaptic fluctuations are present in biological neuronal networks (Sec. 2.3). Given these observations, we predict that BLA would be a more stable motif for computation than UBLA in the presence of noise and continuous learning. Since BLA, but not UBLAs, avoids exploding gradients, if the desired computation requires a line attractor of finite range, BLA would be both easier to maintain and learn. Is this only true for ReLU parameterized RNNs, or does it generalize?

In this paper, we lay out a new theory of general continuous-valued memory in the context of learning to answer the following questions:

1. Can we avoid exploding gradients under parameter perturbation?

2. Do we need to worry about the brittleness of the continuous attractor solutions in practice?

Our theory provides answers to both questions under mild assumptions in an architecture agnostic manner. Using Fenichel's invariant manifold theorem, we derive a sufficient condition for RNNs implementing continuous attractors to remain free of exploding gradients. Moreover, even after a bifurcation, these RNNs still approximately behave like the original continuous attractor for a while. Together these theoretical results practically nullify the concern of the fine tuning problem in theoretical neuroscience and suggest general principles for evaluating and designing new architectures and initialization strategies for RNNs in machine learning.

2 Theory of gracefully degrading continuous attractors

In this section, we apply Fenichel's work on invariant manifolds to RNNs and translate the results for the machine learning and theoretical neuroscience audience.

2.1 Invariant Continuous Attractor Manifold Theory

We start by formulating RNNs implementing a continuous attractor in continuous time: $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$. Let l be the intrinsic dimension of the manifold of equilibria that defines the continuous attractor. We will reparameterize the dynamics around the manifold with coordinates $\mathbf{y} \in \mathbb{R}^l$ and the remaining ambient space with $\mathbf{z} \in \mathbb{R}^{d-l}$. To describe an arbitrary bifurcation of interest, we introduce a sufficiently smooth function g and a bifurcation parameter $\epsilon \geq 0$, such that the following system is equivalent to the original ODE:

$$\dot{\mathbf{y}} = \epsilon \mathbf{g}(\mathbf{y}, \mathbf{z}, \epsilon) \qquad \text{(tangent)} \tag{2}$$

$$\dot{\mathbf{z}} = \mathbf{h}(\mathbf{y}, \mathbf{z}, \epsilon) \qquad \text{(normal)} \tag{3}$$

where $\epsilon = 0$ gives the condition for the continuous attractor $\dot{\mathbf{y}} = \mathbf{0}$. We denote the corresponding manifold of l dimensions $\mathcal{M}_0 = \{(\mathbf{y}, \mathbf{z}) \mid \mathbf{h}(\mathbf{y}, \mathbf{z}, 0) = 0\}$.

We need the flow normal to the manifold to be hyperbolic, that is *normally hyperbolic*, meaning that the Jacobians $\nabla_{\mathbf{z}}\mathbf{h}$ evaluated on any point on the $\mathcal{M}_0$ has $d - l$ eigenvalues with non-zero real parts, and $\nabla_{\mathbf{y}}\mathbf{g}$ has l eigenvalues with zero real parts. More specifically, for continuous attractors, the real part of the eigenvalues of $\nabla_{\mathbf{z}}\mathbf{h}$ will be negative, representing sufficiently strong attractive flow toward the manifold. Equivalently, for the ODE, $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, the variational system is of constant rank, and has exactly l eigenvalues with negative real parts and $(n - l)$ eigenvalues with zero real parts everywhere along the continuous attractor.

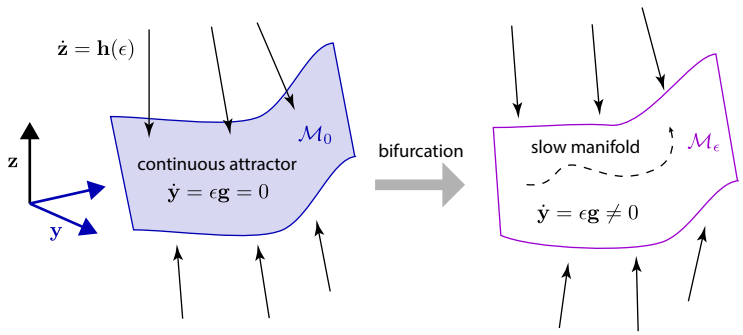

Figure 2: Fenichel's invariant manifold theorem applied to compact continuous attractor guarantees the flow on the slow manifold is locally invariant and continues to be attractive. The dashed line is a trajectory "trapped" in the slow manifold (locally invariant).

When $\epsilon > 0$, the continuous attractor bifurcates away. What can we say about the fate of the perturbed system? The continuous dependence theorem [9] says that the trajectories will change continuously as a function of ϵ without a guarantee on how quickly they change. Moreover, the topological structure and the asymptotic behavior of trajectories change discontinuously due to the

bifurcation. Surprisingly there is a strong connection in the geometry due to Fenichel's theorem [14]. We informally present a special case due to [19]:

Theorem 1 (Fenichel's Invariant Manifold Theorem)**.** *Let $\mathcal{M}_0$ be a connected, compact, normally hyperbolic manifold of equilibria originating from a sufficiently smooth ODE. For a sufficiently small perturbation $\epsilon > 0$, there exists a manifold $\mathcal{M}_\epsilon$ diffeomorphic to $\mathcal{M}_0$ and locally invariant under the flow of* (2)*. Moreover, $\mathcal{M}_\epsilon$ has corresponding smoothness.*

The manifold $\mathcal{M}_\epsilon$ is called the *slow manifold* which is no longer necessarily a continuum of equilibria. However, the local invariance implies that trajectories remain within the manifold except potentially at the boundary. Furthermore, the non-zero flow on the slow manifold is slow and given in the $\epsilon \to 0$ limit as $\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\tau} = \mathbf{g}(c^\epsilon(\mathbf{y}), \mathbf{y}, 0)$ where $\tau = \epsilon t$ is a rescaled time and $c^\epsilon(\cdot)$ parameterizes the l dimensional slow manifold. In addition, the stable manifold(s) of $\mathcal{M}_0$ are similarly approximately maintained [19], allowing the continuous attractor to remain attractive.

These conditions are met (up to numerical precision) for the BLA example in Fig. 1B. As the theory predicts, BLA bifurcates into a 1-dimensional slow manifold (dark colored regions) that contains fixed points, and overall still attractive. On the contrary, the UBLA does not satisfy the compactness condition, hence the theory does not predict its persistence. Importantly, the "slow" flow on the perturbed system is not bounded.

In practice, the sufficient conditions for RNNs implementing continuous attractors to have this graceful breakdown is for the continuous attractor manifold to be of finite dimension throughout, connected, and bounded. However, in systems with an invariant manifold with dimension at least three, it is possible that a slow manifold with chaotic dynamics is created through a perturbation. This would have as consequence that the perturbed system acquires positive Lyapunov exponents (corresponding to the chaotic orbit), which then can still lead to exploding gradients.

2.2 Implications on Machine Learning

Extending the memory time constant of RNNs have long been an important area of research with much focus on random weights [3, 8, 15, 18, 20, 24, 32, 42]. Various initializations for the recurrent weights have been proposed to help learning: initialization with the identity matrix [**?**], with a random orthogonal matrix [18, 33], with a unitary matrix [3] and with a block diagonal weight matrix that creates a quasi-periodic system with limit cycles [39]. However, despite the capacity to maintain representation of continuous quantities for arbitrary duration of time, continuous attractor mechanism has not been pursued in machine learning research because of its brittleness. The stochasticity in gradients inherited from the training data, regularization strategy, and multi-task learning objectives act as a perturbation on the recurrent dynamics, and continuous attractors break down even if it could be learned. Remedies emerged in machine learning to hard-code continuous-valued memory structures within the RNNs—e.g., the cell state in vanilla LSTM. However, our theory shows that the geometric structure of the manifold and the flow around the manifold play a critical role in enabling gradient descent learning of continuous attractors using standard methods such as backpropagation through time (BPTT) [41].

It is well known that asymptotic exploding gradients comes from positive Lyapunov exponents [13, 27, 43]. It has also been pointed out that bifurcations can cause arbitrarily large gradients [12] as well as discontinuity in the Lyapunov spectrum [2]. These gradient propagation theories suggested that bifurcations should be avoided, including the continuous attractors.

As far as we know, there is no architecture agnostic theory describing the loss landscape around RNN solutions. We remark that due to the singular nature of the center manifold that supports the continuous attractor, the usual analysis approach of linearization fails. *Our theory successfully connects the invariant manifold theory and the gradient signal propagation theory in RNNs to describe two types of loss landscape around continuous attractor solutions.* In one case, when the theorem holds, the landscape is shallow in all directions due to (asymptotically) vanishing gradients induced by the attractor structure—we have the gracefully degrading continuous attractor. On the other case, we can find examples where the theorem does not hold, and the continuous attractor solution is at the boundary of network configurations with exploding gradients, meaning the loss landscape is very steep in some directions.

While exploding gradients would prevent gradient descent to correct for deviations from the optima, for gracefully degrading ones, one can apply restorative forces via gradient descent to be in the vicinity of the brittle continuous attractor solution (see Sec. 3.5).

2.3 Implications on Neuroscience

Continuous attractors are biologically plausible, theoretically elegant, consistent with neural recordings, and avoids the asymptotic exploding and vanishing gradient problem [2]. As a conceptual tool in computational and theoretical neuroscience, continuous attractors are widely used when working memory of continuous values is needed [6, 11, 21]. When used to accumulate stimulus, continuous attractors are also called neural integrators that are hypothesized to be the underlying computation for the maintenance of eye positions, heading direction, self-location, target location, sensory evidence, working memory, decision variables, to name a few [31, 34, 35]. Neural representation of continuous values have been observed as persistent activity in the prefrontal cortex of primates, ellipsoid body of the fly, and hypothalamus [28, 29, 31]. A typical computational implementation of a continuous attractor are bump attractor network models which require a mean-field limit [7, 30, 38] and finite sized networks with threshold linear units [29, 40], see also Sec. 3.6.1.

However, the so-called "fine-tuning problem" describing the theoretical and practical brittleness of continuous attractors was recognized in [36]. Since biological neural systems have constantly fluctuating synaptic weights [37], this has been a big puzzle in the field. There have been efforts and remedies to lessen the degradation for particular implementations, often focusing on keeping the short-term behavior close to the continuous attractor case [5, 23, 25, 26, 30].

Our theory shows that not all continuous attractors are born equal, and there are gracefully degrading continuous attractors. In finite time, trajectories are well-behaved, contrary to the asymptotic behavior captured by the Lyapunov exponents. Animal behavior is finite time in nature and the longer the temporal distance the harder it is to learn in general. The conditions are favorable in the recurrent neuronal networks: (1) mutual inhibition is widely present and evidence points to inhibition dominated dynamics, (2) the neural state space is bounded due to physiological constraints– by a non-negative firing rate below and a maximum firing rate above.

3 Experiments

3.1 Continuous-valued click integration task

Although there are several standard benchmark tasks for temporal dependence learning with memory components, due to the discrete nature of the relevant stimulus space, the solutions do not necessarily require a continuously-valued memory. We designed a simple integration task where one of the optimal solutions is the continuous attractor.

We extend the Poisson clicks task which has been used for studying perceptual decision making [**?**]. Neural integrators are considered to be crucial for the task for accumulating evidence over time and maintaining a continuous representation of the accumulated evidence. For our experiments, we used discrete time representations over T time bins and the real-valued[1] stimulus encoded as difference of two non-negative values:

$$I_{t,i} = m_{t,i} \cdot u_{t,i} \qquad \text{(continuous clicks)} \tag{4}$$

$$O_t^* = \sum_{s=0}^{t} (I_{s,1} - I_{s,2}) \qquad \text{(desired output)} \tag{5}$$

where $m_{t,i}$ are independent Bernoulli random variables with probability 0.2 and $u_{t,i}$ are independent random variables with uniform distribution on the unit interval. We used mean squared error (MSE) of the 1-dimensional output over time as the loss function over all time bins. We used $T = 500$ time bins per trial unless specified otherwise. The gradients were computed in batch mode with 1000 randomly generated trials. The main challenges of this task are (1) addition and subtraction and (2) maintain the accumulated clicks for extended period of time.

[1] Approximated as single precision IEEE 754 floating point.

3.2 RNN solutions to the integration task

We use vanilla RNN implementations with the standard parameterization:

$$\begin{aligned} \mathbf{x}_t &= \sigma(\mathbf{W}_{\text{in}}\mathbf{I}_t + \mathbf{W}\mathbf{x}_{t-1} + \mathbf{b}) \\ O_t &= \mathbf{W}_{\text{out}}\mathbf{x}_t + \mathbf{b}_{\text{out}} \end{aligned} \tag{6}$$

where $\mathbf{x}_t \in \mathbb{R}^d$ is the hidden state, $\mathbf{I}_t \in \mathbb{R}^K$ is the input, $\sigma : \mathbb{R} \to \mathbb{R}$ an activation function which acts on each of the hidden dimension, and $\mathbf{W}, \mathbf{b}, \mathbf{W}_{\text{in}}, \mathbf{W}_{\text{out}}, \mathbf{b}_{\text{out}}$ are parameters. Assuming an Euler integration with unit time step, the discrete-time RNN of (6) corresponds to the ODE:

$$\dot{\mathbf{x}} = -\mathbf{x} + \sigma(\mathbf{W}_{\text{in}}\mathbf{I} + \mathbf{W}\mathbf{x} + \mathbf{b}) \tag{7}$$

For tractable analysis, we consider 2 dimensional systems with ReLU activation. We study three different RNN implementations of an integrator, the Identity RNN (iRNN), UBLA and BLA. On the original clicks task the UBLA and BLA networks count the click differences directly, while iRNN counts the clicks separately and then subtracts these representations through the output mapping. The behaviors of UBLA and BLA in the absence of stimulus are shown in Fig. 1, while the behavior of the iRNN is trivial since there is no flow. These networks are defined as follows.

Identity RNN

$$\mathbf{W}_{\text{in}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \mathbf{W} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \mathbf{W}_{\text{out}} = \begin{pmatrix} -1 \\ 1 \end{pmatrix},\ \mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \mathbf{b}_{\text{out}} = 0. \tag{8}$$

Unbounded line attractor

$$\mathbf{W}_{\text{in}} = \alpha \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix},\ \mathbf{W} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ \mathbf{W}_{\text{out}} = \frac{1}{2\alpha} \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ \mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \mathbf{b}_{\text{out}} = -\frac{\beta}{\alpha}. \tag{9}$$

While the line attractor is unbounded from above, it only extends to the center from below. The step size along line attractor α determines the maximum number of clicks as the difference between the two channels; the capacity is β/α number of clicks.

Bounded line attractor We formulate this implementation of a bounded integrator with a parameter that determines step size along line attractor α. Analogously as for UBLA, these parameters determine the capacity of the network. The inputs push the input along the line attractor in two opposite directions, see below. UBLA and BLA need to be initialized at $\beta(1, 1)$ and $\frac{\beta}{2}(1, 1)$, respectively, for correct decoding, i.e., output projection.

$$\mathbf{W}_{\text{in}} = \alpha \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},\ \mathbf{W} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix},\ \mathbf{W}_{\text{out}} = \frac{1}{2\alpha} \begin{pmatrix} 1 \\ -1 \end{pmatrix},\ \mathbf{b} = \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ \mathbf{b}_{\text{out}} = 0. \tag{10}$$

3.3 Asymmetric loss landscape due to exploding gradients

To illustrate the effect of bifurcations from continuous attractor solution, we take a 1-dimensional slice of the loss surface. Specifically, we take the self-recurrent connection and systematically vary:

$$\mathbf{W}_{1,1} \leftarrow \mathbf{W}_{1,1} + \Delta. \tag{11}$$

The result is shown in Fig. 3A. For UBLA and iRNN, positive perturbations shows exponentially increasing loss due to the bifurcation to an exploding gradient dynamical system. In all other cases, including all perturbations of BLA, leads to vanishing gradient, hence the loss is bounded.

3.4 Bifurcation probability and random perturbations

We consider all parametrized perturbations of the form $\mathbf{W} \leftarrow \mathbf{W} + \mathbf{V}$ for a random matrix $\mathbf{V} \in \mathbb{R}^{2\times 2}$. The BLA can bifurcate in the following systems, characterized by their invariant sets: a system with single stable fixed point, a system with three fixed points (one unstable and two stable) and a system with two fixed points (one stable and the other a half-stable node) and a system with a (rotated) line

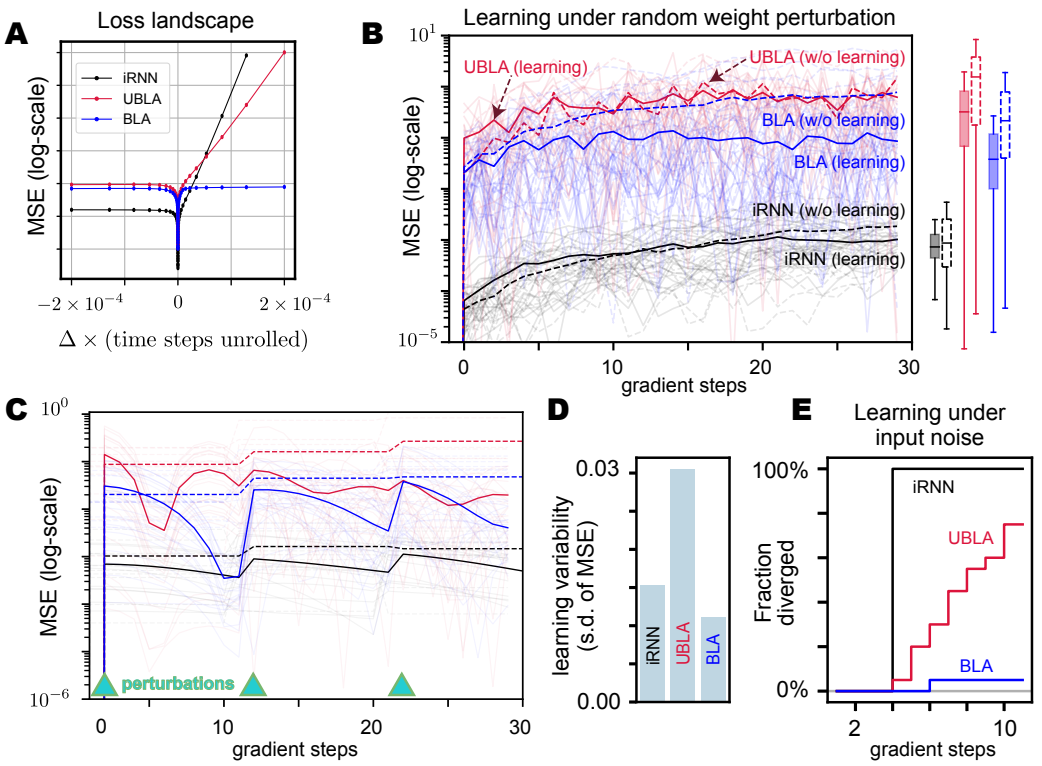

Figure 3: Comparing three continuous attractor solutions to the clink integration task. (A) Single parameter perturbation showing exploding gradients for iRNN and UBLA. (B) Effect of gradient descent in repairing the continuous attractor. RNNs without gradient descent (dashed line) are shown for reference. Box plots show distribution of the loss for the last 10 steps. (C) Interleaved weight perturbations showing gradual recovery for BLA and iRNN. (D) High variability across time of UBLA and iRNN due to exploding gradients. (E) Noise injected to objective induces gradient descent to drive solutions to diverge away.

attractor. Only the first two bifurcations (Fig. 1) can happen with nonzero chance for arbitrary the type of perturbations we consider. The perturbations that leave the line attractor intact and or two a system with two fixed points have measure zero in parameter space. The types of perturbation with measure zero are codimension 2 bifurcations. The perturbation that results in one fixed point happen with probability $\frac{3}{4}$, while perturbations lead to a system with three fixed points with probability $\frac{1}{4}$.

3.5 Maintaining a neural integrator

To investigate the differential effect of gradient descent on the three neural integrator models, we performed three learning experiments using the continuous-valued click integration task.

In the first experiment, Gaussian random noise is injected to all parameters inducing a constant diffusion of the parameters. This emulates the biological synaptic variability. Fig. 3B shows that for BLA, gradient descent (with constant learning rate, chosen from a grid search) was able to counter the diffusion. Note that the magnitude of perturbation were identical for all networks, but iRNN is more robust in terms of MSE. The distribution of MSE for the task can be used as proxy for the misadjustment from the optimal solution. BLA with learning has superior misadjustment to UBLA and BLA without learning.

To dissociate the effect of misadjustment from gradient descent and external perturbation, we interleaved a larger perturbation with 10 gradient steps. Fig. 3C shows consistent learning in the average MSE for BLA and iRNN, but not in UBLA. The individual runs indicate larger variability induced by learning over time, which we quantify in Fig. 3D. iRNN and UBLA both have larger

Table 1: Summary of parameters for experiments.

Parameter	Input noise	Paramter noise
Trial length (time steps)	200	500
Input sequence length (T_I)	100	50
Learning rate[2]	10^{-4}	10^{-9}
Perturbation parameter (σ)	10^{-6}	10^{-5}

learning noise than BLA, consistent with the absence of exploding gradient configuration around the optimum.

Finally, we investigated the effect of learning when perturbations are induced by the backpropagated gradient which is structured by the recurrent dynamics. To inject noisy gradients naturally, we added noise to the input to the first 10 time steps during the trial that were not integrated. Then we looked for networks that diverged due to exploding gradients which we identify as having a loss of 1 or higher. Fig. 3E shows that UBLA and iRNN quickly diverges under gradient descent than the BLA.

The parameters that were used for the effects of input- and parameter-type noise are summarized in Table 1. The bias parameters of the UBLA and BLA networks were set to $\beta = 100$, while $\alpha = 1$. The hidden state at the beginning of a trial is a learnable parameter. The optimal learning rates for the input-type noise experiments were chosed from a set of three values ($\{10^{-8}, 10^{-9}, 10^{-10}\}$) based on best performance of the task of 20 runs.

A sequence of inputs is generated of maximum length T_I generated as described above. For the experiments with D-type noise, we generate noise in the input for a sequence length of 10 time steps and distribution width sampled from the uniform distribution $(-\sigma, \sigma)$ with $\sigma = 10^{-6}$. For the experiments with S-type noise, we perturb the weights with a matrix with entries sampled from $\mathcal{N}(0, \sigma)$. We inject this noise into the system either after each learning epoch (Fig. 3B) or in pulses only at epochs 1,11 and 21 (Fig. 3C). The first experiment should demonstrate the effects of the restorative force of learning.

3.6 Perturbation of continuous attractors

We will first of all look at a simple (non-biological) system that has a ring attractor to demonstrate the consequences of the Persistence Theorem. The system we will analyse is defined by the following ODE: $\dot{r} = r(1 - r),\ \dot{\theta} = 0$. This system has as fixed points the ring with radius one centered around zero, i.e., $(0, 0) \cup \{(1, \theta) \mid \theta \in [0, 2\pi)\}$.

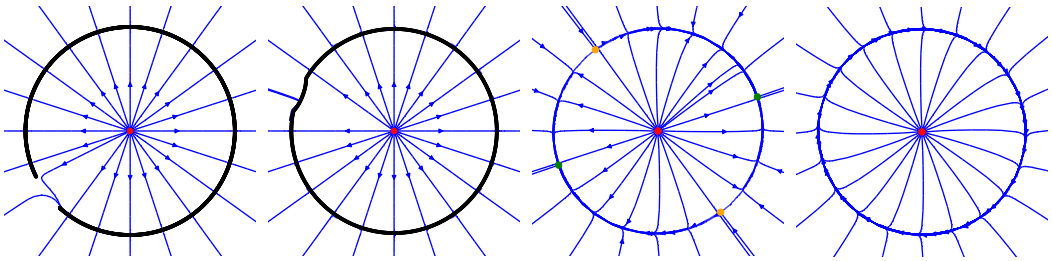

Figure 4: Perturbations to a simple implementation of a ring attractor all leave the invariant manifold intact. (Left) Examples of a local perturbation to the vector field through the addition of a bump to the vector field along the ring attractor. (Leftmost) An example of a bump perturbation that results in the ring breaking up and becoming diffeomorphic to a line. (Left, middle) An example of a bump perturbation that maintains the ring structure, but deforms it locally. (Right) Examples of a global perturbation to the vector field through the addition of a small term to the connectivity matrix. (Right, middle) A global perturbation that results in a system with four fixed points along the persistent invariant manifold. The two saddle nodes (yellow dots) are connected to the stable fixed points (green dots) through connecting orbits. (Rightmost) A global perturbation that results in a limit cycle.

3.6.1 Finite-sized head direction network model

Continuous attractor models of the head direction representation suggest that the representation emerges from the interactions of recurrent connections among neurons that form a circular or ring-like structure [1, 29, 44]. Since continuous attractors are susceptible to noise and perturbations the precise representation of the head direction can in principle easily be disrupted. We will the implications of the Persistence theorem in the case of the implementation of a continuous ring attractor presented in [29]. As this continuous ring attractor is bounded its invariant manifold persists and, hence, no divergent orbits are created under small perturbations to this system. This model is composed of N heading-tuned neurons with preferred headings $\theta_j \in \{\frac{2\pi i}{N}\}_{i=1\ldots N}$ radians. For sufficiently strong local excitation (given by the parameter J_E) and broad inhibition, this network will generate a stable bump of activity, one corresponding to each head direction. This continuum of fixed points for a one dimensional manifold homeomorphic to the circle.

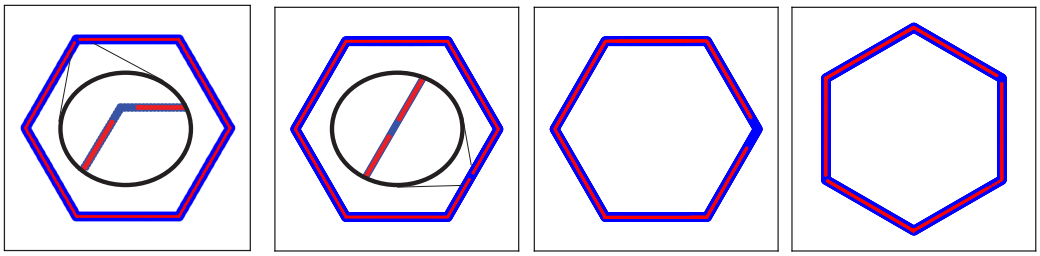

Figure 5: Perturbations to the ring attractor introduced in [29]. (Left) Examples of a local perturbation to the vector field through the addition of a bump to the vector field along the ring attractor. (Leftmost) An example of a bump perturbation which cuts out a piece of the ring attractor, the stable fixed points of the system are shown in black. A heteroclinic orbit is created on the bounday of the hole with a slow flow. (Left, middle) An example of a bump perturbation that deforms the ring attractor, but maintains its topology. (Right) Examples of a global perturbation to the vector field through the addition of a small term to the connectivity matrix.

We evaluate the effect of perturbations on a network of size $N = 6$ with the optimal value $J_E^* = 2.4$ through local and parametrized perturbations, similarly as for the toy ring attractor model. The space of possible perturbations in Sec. 2.1 is very large. To be able to form an image of what happens in the theorem we will work out the effect of some examples of perturbations. Hereby, we will focus on two kinds of perturbations, local and parametric. We define a local perturbation (i.e., a change to the ODE with compact support) through the bump function $\Psi(x) = \exp\left(\frac{1}{\|x\|^2-1}\right)$ for $\|x\| < 1$ and zero outisde, by multiplying it with a uniform, unidirectional vector field. We further investigate the effect of parameteric perturbations of the form if a term $\mathbf{W}x$ to the original ODE with the entries of $\mathbf{W}$ sampled from a normal distribution with mean zero and variance $\frac{1}{100}$. All such perturbations leave at least a part of the continuous attractor intact and preserve the invariant manifold, i.e. the parts where the fixed points disappear a slow flow appears.

4 Discussion

We have derived a general theory that describes the loss landscape of RNNs around continuous attractor solutions. This theory implies backpropagating gradient does not explode for systems near compact continuous attractors because of divergent dynamics. This insight also suggests that there can exist homeostatic mechanisms for certain implementations of continuous attractors that maintain the structure of the attractor sufficiently for the neural computation it is used in, which we demonstrate in simple networks.

It is well-known that activation functions play a critical role in propagating these gradients effectively through the network [**? ? ?**]. The activation function can have many forms. The ones that are C^r for $r \geq 1$ are the ones to which the Persistence Theorem applies. The Persistence Theorem further specifies how the smoothness of the activation can have implications on the smoothness of the persistence invariant manifold. For situations where smoothness of the persistent invariant manifold

is of importance, smoother activation functions might be preferrable, such as the Exponential Linear Unit (ELU)[10] or the Continuously Differentiable Exponential Linear Units (CELU) [4].

References

[1] Z. Ajabi, A. T. Keinath, X.-X. Wei, and M. P. Brandon. Population dynamics of head-direction neurons during drift and reorientation. *Nature*, 615(7954):892–899, 2023.

[2] (anonymized). Persistent learning signals without continuous attractors. 2023.

[3] M. Arjovsky, A. Shah, and Y. Bengio. Unitary evolution recurrent neural networks. In *International conference on machine learning*, pages 1120–1128. PMLR, 2016.

[4] J. T. Barron. Continuously differentiable exponential linear units. *arXiv preprint arXiv:1704.07483*, 2017.

[5] M. Boerlin, C. K. Machens, and S. Denève. Predictive coding of dynamical variables in balanced spiking networks. *PLoS computational biology*, 9(11):e1003258, Nov. 2013. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1003258. URL `http://dx.doi.org/10.1371/journal.pcbi.1003258`.

[6] Y. Burak and I. R. Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS Computational Biology*, 5(2), 2009. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000291.

[7] M. Camperi and X. J. Wang. A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *Journal of computational neuroscience*, 5(4):383–405, Dec. 1998. ISSN 0929-5313. doi: 10.1023/a:1008837311948. URL `http://dx.doi.org/10.1023/a:1008837311948`.

[8] M. Chen, J. Pennington, and S. S. Schoenholz. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 873–882. PMLR, June 2018. URL `https://proceedings.mlr.press/v80/chen18i.html`.

[9] C. Chicone. *Ordinary Differential Equations with Applications*. Springer Science & Business Media, Sept. 2006. ISBN 9780387357942. URL `https://market.android.com/details?id=book-Vn9EQ5onbvUC`.

[10] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), 2015.

[11] P. Dayan and L. F. Abbott. Theoretical neuroscience: Computational and mathematical modeling of neural systems. 2001.

[12] K. Doya. Bifurcations of recurrent neural networks in gradient descent learning. *IEEE Transactions on neural networks*, 1(75):218, 1993.

[13] R. Engelken, F. Wolf, and L. F. Abbott. Lyapunov spectra of chaotic recurrent neural networks. *Physical Review Research*, 2023. URL `http://arxiv.org/abs/2006.02427`.

[14] N. Fenichel and J. Moser. Persistence and smoothness of invariant manifolds for flows. *Indiana University Mathematics Journal*, 21(3):193–226, 1971.

[15] M. S. Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634, Feb. 2009. ISSN 0896-6273, 1097-4199. doi: 10.1016/j.neuron.2008.12.012. URL `http://dx.doi.org/10.1016/j.neuron.2008.12.012`.

[16] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014. URL `http://arxiv.org/abs/1410.5401`.

[17] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, Oct. 2017. ISSN 2162-2388, 2162-237X. doi: 10.1109/TNNLS.2016.2582924. URL `http://dx.doi.org/10.1109/TNNLS.2016.2582924`.

[18] M. Henaff, A. Szlam, and Y. LeCun. Recurrent orthogonal networks and long-memory tasks. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2034–2042, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/henaff16.html`.

[19] C. K. R. T. Jones. Geometric singular perturbation theory. In L. Arnold, C. K. R. T. Jones, K. Mischaikow, G. Raugel, and R. Johnson, editors, *Dynamical Systems: Lectures Given at the 2nd Session of the Centro Internazionale Matematico Estivo (C.I.M.E.) held in Montecatini Terme, Italy, June 13–22, 1994*, pages 44–118. Springer Berlin Heidelberg, Berlin, Heidelberg, 1995. ISBN 9783540494157. doi: 10.1007/BFb0095239. URL `https://doi.org/10.1007/BFb0095239`.

[20] G. Kerg, K. Goyette, M. Puelma Touzel, G. Gidel, E. Vorontsov, Y. Bengio, and G. Lajoie. Non-normal recurrent neural network (nnRNN): learning long time dependencies while improving expressivity with transient dynamics. In *Advances in Neural Information Processing Systems 32*, pages 13613–13623. 2019. URL http://papers.nips.cc/paper/9513-non-normal-recurrent-neural-network-nnrnn-learning-long-time-dependencies-while-improving-expressspdf.

[21] M. Khona and I. R. Fiete. Attractor and integrator networks in the brain. *Nature reviews. Neuroscience*, 23(12):744–766, Dec. 2022. ISSN 1471-003X, 1471-0048. doi: 10.1038/s41583-022-00642-0. URL http://dx.doi.org/10.1038/s41583-022-00642-0.

[22] J. F. Kolen and S. C. Kremer. *Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies*, pages 237–243. 2001. doi: 10.1109/9780470544037.ch14.

[23] A. A. Koulakov, S. Raghavachari, A. Kepecs, and J. E. Lisman. Model for a robust neural integrator. *Nature neuroscience*, 5(8):775–782, Aug. 2002. ISSN 1097-6256. doi: 10.1038/nn893. URL http://dx.doi.org/10.1038/nn893.

[24] R. Legenstein and W. Maass. What makes a dynamical system computationally powerful? In S. S. Haykin, J. C. Príncipe, T. J. Sejnowski, and J. McWhirter, editors, *New Directions in Statistical Signal Processing: From Systems to Brain*, pages 127–154. MIT Press, 2007. ISBN 9780262083485. URL http://www.igi.tu-graz.ac.at/maass/psfiles/165.pdf.

[25] S. Lim and M. S. Goldman. Noise tolerance of attractor and feedforward memory models. *Neural computation*, 24(2):332–390, Feb. 2012. ISSN 0899-7667, 1530-888X. doi: 10.1162/NECO_a_00234. URL http://dx.doi.org/10.1162/NECO_a_00234.

[26] S. Lim and M. S. Goldman. Balanced cortical microcircuitry for maintaining information in working memory. *Nature neuroscience*, 16(9):1306–1314, Sept. 2013. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.3492. URL http://dx.doi.org/10.1038/nn.3492.

[27] J. M. Mikhaeil, Z. Monfared, and D. Durstewitz. On the difficulty of learning chaotic dynamics with RNNs. In *Advances in Neural Information Processing*, 2022. URL http://arxiv.org/abs/2110.07238.

[28] A. Nair, T. Karigo, B. Yang, S. Ganguli, M. J. Schnitzer, S. W. Linderman, D. J. Anderson, and A. Kennedy. An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell*, 186(1):178–193.e15, Jan. 2023. ISSN 0092-8674. doi: 10.1016/j.cell.2022.11.027. URL https://www.sciencedirect.com/science/article/pii/S0092867422014714.

[29] M. Noorman, B. K. Hulse, V. Jayaraman, S. Romani, and A. M. Hermundstad. Accurate angular integration with only a handful of neurons. *bioRxiv*, 2022. doi: 10.1101/2022.05.23.493052. URL https://www.biorxiv.org/content/early/2022/05/25/2022.05.23.493052.

[30] A. Renart, P. Song, and X.-J. Wang. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron*, 38(3):473–485, May 2003. ISSN 0896-6273. doi: 10.1016/s0896-6273(03)00255-1. URL http://dx.doi.org/10.1016/s0896-6273(03)00255-1.

[31] R. Romo, C. D. Brody, A. Hernández, and L. Lemus. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399(6735):470–473, June 1999. ISSN 0028-0836. doi: 10.1038/20939. URL http://dx.doi.org/10.1038/20939.

[32] T. K. Rusch and S. Mishra. UnICORNN: A recurrent model for learning very long time dependencies. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9168–9178. PMLR, 2021. URL https://proceedings.mlr.press/v139/rusch21a.html.

[33] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Represenatations*, 2014. URL http://arxiv.org/abs/1312.6120.

[34] H. S. Seung. How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23):13339–13344, . ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.93.23.13339. URL https://pnas.org/doi/full/10.1073/pnas.93.23.13339.

[35] H. S. Seung, D. D. Lee, B. Y. Reis, and D. W. Tank. Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron*, 26(1):259–271, Apr. 2000. ISSN 0896-6273. doi: 10.1016/s0896-6273(00)81155-1. URL http://dx.doi.org/10.1016/s0896-6273(00)81155-1.

[36] S. H. Seung. Continuous attractors and oculomotor control. *Neural Networks*, 11(7-8):1253–1258, . ISSN 08936080. doi: 10.1016/S0893-6080(98)00064-1. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608098000641.

[37] G. Shimizu, K. Yoshida, H. Kasai, and T. Toyoizumi. Computational roles of intrinsic synaptic dynamics. 70:34–42. ISSN 09594388. doi: 10.1016/j.conb.2021.06.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0959438821000672.

[38] W. E. Skaggs, J. J. Knierim, H. S. Kudrimoti, and B. L. McNaughton. A model of the neural basis of the rat's sense of direction. In *Advances in neural information processing systems*, volume 7, pages 173–180, 1995. URL https://www.ncbi.nlm.nih.gov/pubmed/11539168.

[39] P. Sokol, I. Jordan, E. Kadile, and I. M. Park. Adjoint dynamics of stable limit cycle neural networks. In *53rd Asilomar Conference on Signals, Systems and Computers*, 2019. doi: 10.1109/IEEECONF44664.2019.9049080.

[40] D. Spalla, I. M. Cornacchia, and A. Treves. Continuous attractors for dynamic memories. *eLife*, 10, Sept. 2021. ISSN 2050-084X. doi: 10.7554/eLife.69499. URL http://dx.doi.org/10.7554/eLife.69499.

[41] N. Toomarian and J. Barhen. Adjoint-functions and temporal learning algorithms in neural networks. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 113–120. Morgan-Kaufmann, 1991. URL http://papers.nips.cc/paper/425-adjoint-functions-and-temporal-learning-algorithms-in-neural-networks.pdf.

[42] T. Toyoizumi and L. F. Abbott. Beyond the edge of chaos: amplification and temporal integration by recurrent networks in the chaotic regime. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 84(5 Pt 1):051908, Nov. 2011. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.84.051908. URL http://dx.doi.org/10.1103/PhysRevE.84.051908.

[43] R. Vogt, M. Puelma Touzel, E. Shlizerman, and G. Lajoie. On Lyapunov exponents for RNNs: Understanding information propagation using dynamical systems tools. *Frontiers in Applied Mathematics and Statistics*, 8, 2022. ISSN 2297-4687. doi: 10.3389/fams.2022.818799. URL https://www.frontiersin.org/articles/10.3389/fams.2022.818799.

[44] K. Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

5 Supplementary Material

RNNs are also capable of learning complex patterns and relationships in the data, which makes them in particular useful as models for neural computations. This is achieved through the use of non-linear activation functions and the training of the network using backpropagation through time (BPTT). BPTT allows the network to adjust the weights of the connections between neurons based on the error signal that is propagated backwards through time.

However, training RNNs by using back-propagation through time to compute error-derivatives can be difficult. Early attempts suffered from vanishing and exploding gradients [22] and this meant that they had great difficulty learning long-term dependencies. Many different methods have been proposed for overcoming this difficulty.