ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Stability of State and Costate Dynamics in Continuous Time Recurrent Neural Networks

Alessandro Betti[1], Marco Gori[2], Stefano Melacci[2] *

1 – IMT School for Advanced Studies, Lucca, Italy

2 – DIISM, University of Siena, Siena, Italy

Abstract. The notion of stability plays a crucial role in ensuring the safe development of a model in a lifelong learning context. This paper investigates the fundamental aspects of stability in a class of continuous-time recurrent neural networks which include both state and costate variables. The latter are directly inherited from optimal control theory, and they act as adjoint variables closely related to gradient terms. Stability is investigated both in terms of state and of costate dynamics, showing the key conditions that must be satisfied to produce bounded dynamics in the forward and learning stages.

1 Introduction

In the last few years, the machine learning community renewed the interest in recurrent neural networks and related topics, focusing on novel instances of state-space models and analyzing their connections with Transformers-like neural architectures [1]. This interest is paired to the one of learning from non-i.i.d. data provided over time, such as in the case of agents interacting in a dynamic environment [2]. This is at the core of continual/lifelong learning [3], time series processing [4], and streaming machine learning [5], even if with sometimes different terminology, approaches, and metrics. From a bare machine learning standpoint, while sequence processing is very common when learning offline from datasets of nicely segmented sequences, it is less common when learning from a single, possibly lifelong, sequence provided over time [6]. It becomes even more challenging when learning is instantiated in an online manner, without focusing on storing data to replicate offline-learning dynamics (Collectionless AI [7][1]). This paper focuses on continuous-time recurrent models [8, 9] that not only develop an internal *state* to summarize the data observed so far, but also learn online in a forward manner (i.e., without backpropagating gradients over the temporal dimension) from a stream of data. Recent work has promoted the idea of Neural ODEs/CDEs and related approaches [10, 11], still relying on boundary-value problems, and introducing adjoint variables to model the learning process when going backward in time, which can be related to the *costate*

*This work was supported by the University of Siena (Piano per lo Sviluppo della Ricerca - PSR 2024, F-NEW FRONTIERS 2024), under the project "TIme-driveN StatEful Lifelong Learning" (TINSELL). It was also supported by the project "CONSTR: a COllectionless-based Neuro-Symbolic Theory for learning and Reasoning", PARTENARIATO ESTESO "Future Artificial Intelligence Research - FAIR", SPOKE 1 "Human-Centered AI" Università di Pisa, "NextGenerationEU", CUP I53C22001380006.

[1]See also `https://cai.diism.unisi.it`

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

variables in optimal control theory. Differently, the novel framework of Hamiltonian Learning [12] focuses on solutions that work forward-in-time, thus specifically considering a continuous stream of data. Hamiltonian Learning is rooted on a control-theory-based formulation of the learning problem, where *costate* variables have their own forward dynamics. In this paper, the stability of *state* and *costate* dynamics are studied, both inheriting and adapting results from existing literature which might be known to a smaller audience, and specifically analyzing the stability of the temporal evolution of the costate variables.

2 State Space Dynamics

We consider a neural network whose topology is defined by means of the directed graph $G = (V, A)$, being V and A the set of vertices (neurons) and edges (weighed connections), respectively. The input processed by a neural network with n neurons consists of the information coming from the environment in which the model "lives", and it is mathematically represented by a trajectory $u\colon [0, +\infty) \to \mathbb{R}^d$. We indicate with $x_i(t)$ the output of the i-th neuron at time t, and we assume that the first d neurons of the topology are the input ones, thus $x_i(t) = u_i(t)$ for $i = 1, \dots d$ and $\forall t \in [0, T]$. Furthermore we will indicate with x without subscripts the $n - d$ dimensional vector $(x_{d+1}, \dots, x_n)$, also referred to as *state*. The notation $\mathrm{pa}(i)$ is used to indicate the set of parents of the i-th neuron. Finally, without any lack of generality, we assume the output of the network to consist of the last m values of x, i.e., $x_{n-m+1}, x_{n-m+2}, \dots x_n$.

Let us consider a dynamical system defined for almost every $t \in [0, T]$,

$$\gamma \dot{x}_i(t) = -\varepsilon x_i(t) + \sigma\Big(\sum_{j \in \mathrm{pa}(i)} w_{ij}(t) x_j(t) \Big) \qquad \text{for } i = d+1, \dots, n, \tag{1}$$

where $\gamma > 0$ is a time constant of the dynamics of the neurons, $\varepsilon > 0$ is a regularization parameter (discussed in the following), and w_{ij} are real valued *weights* of the model, $(j, i) \in A \mapsto w_{ij}$. The function $\sigma \in C^1(\mathbb{R}, S)$ with $S \subset \mathbb{R}$, is an activation function, that we will often assume to be bounded; for definiteness

$$S \subset [-1, 1], \tag{2}$$

and monotonically nondecreasing ($\sigma' \geq 0$). Notice that in Eq. (1) we put ourselves in the general setting where dynamics are driven by weights that have an explicit temporal dependence. This is an important feature, as we will mention in the next section, as soon as we regard the learning process of the network as part of the dynamic of the system. Eq. (1) is also paired with initial conditions

$$x_i(0) = x_i^0 \in \mathbb{R} \qquad \text{for } i = d+1, \dots, n. \tag{3}$$

Let $a_i = \sum_{j \in \mathrm{pa}(i)} w_{ij}(t) x_j(t)$ be a shorthand for the activations without explicitly indicating their dependence on x or w or time. Once the trajectories of the weights are fixed and the initial conditions are given, we can give the following.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Definition 1. We denote with $x_\varepsilon(\cdot; x^0)$ any solution of Eq. (1) that satisfies also the initial condition (3), with $x_{\varepsilon i}(t) = u_i(t)$ for $i = 1, \dots, d$ and $\forall t \in [0, T]$.

Regularization parameter ε. The first stability aspect of the "forward"[2] dynamic that we analyze is related to the role of ε, which also motivates why we refer to it as *regularization parameter*. The key result that clarifies its importance can be summed up in the following proposition.

Proposition 1 (A priori estimate on x)**.** *Suppose σ satisfies* (2)*, then*[3]

$$\|x_\varepsilon(t; x^0)\|_2 \leq \|x^0\|_2 + \frac{\sqrt{n}}{\varepsilon} \qquad \forall t \geq 0.$$

Proof. Let us study how $\|x\|^2/2$ change over time. Assuming that x solves (1):

$$\frac{d}{dt}\frac{1}{2}\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i \dot{x}_i = \sum_{i=1}^n \left(-\frac{\varepsilon}{\gamma}x_i^2 + \frac{1}{\gamma}x_i\sigma(a_i)\right)$$
$$\leq -\frac{\varepsilon}{\gamma}\sum_{i=1}^n x_i^2 + \frac{1}{\gamma}\left|\sum_{i=1}^n x_i\sigma(a_i)\right| \leq -\frac{\varepsilon}{\gamma}\sum_{i=1}^n x_i^2 + \frac{1}{\gamma}\sum_{i=1}^n |x_i\sigma(a_i)|$$

Using hypothesis (2) on the activation function we therefore have $(\|x\|_2^2/2)' \leq -(\varepsilon\|x\|_2^2 - \|x\|_1)/\gamma$. Now we want to study the sign of the function $f(x) := \varepsilon\|x\|_2^2 - \|x\|_1$ since we have just shown that $\|x\|_2$ decreases when $f \geq 0$. Notice that it is sufficient to study the behaviour of such function for $x_i > 0$, $i = 1, \dots, n$ since f is invariant under any transformation that maps $(x_1, \dots, x_n) \mapsto (s_1x_1, \dots, s_nx_n)$ with any choice of $s_i \in \{+1, -1\}$ for $i = 1, \dots, n$. We begin to look for the values of x with positive coordinates such that $f(x) = 0$. From the definition of f we have $0 = \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i/\varepsilon = \sum_{i=1}^n \big(x_i - 1/(2\varepsilon)\big)^2 - n/(4\varepsilon^2)$. This is precisely the equation of a sphere with center $(1/2\varepsilon, 1/2\varepsilon, \dots, 1/2\varepsilon)$ and radius $\sqrt{n}/2\varepsilon$. Moreover, if we switch to hypershperical coordinates, and we restrict ourselves in the region $x_i > 0$, $i = 1, \dots, n$, the function f along the radial variable assumes the form $\varepsilon r^2 - Dr$ (where $D > 0$ depends only on the values of the various angular variables). This means that for $r \in (0, D/\varepsilon)$ the function is negative, then for $r \geq D/\varepsilon$ it becomes positive. In any other sectors where at least one coordinate is negative, say the j-th one, $x_j < 0$,[4] the same reasoning applies. In this case, the sphere on which the zeros of the function lie has its center at $(s_1/2\varepsilon, s_2/2\varepsilon, \dots, s_n/2\varepsilon)$ with $s_j = -1$ and same radius $\sqrt{n}/2\varepsilon$. However, since all these spheres are contained in the bigger sphere with

[2]Throughout the paper we will sometimes refer to the dynamics of the neuron states x described by Eq. (1) as *forward dynamics* to distinguish it from the additional equations, that we will present in Section 3, that instead describes how the weights changes over time. Classically this other set of equations, because of the language related to the backprop algorithm are called *backward* dynamic/step.

[3]Here we use the notation $\|\cdot\|_2$ to stand for the Euclidean norm in $\mathbb{R}^n$, instead we will indicate with $\|\cdot\|_1$ the 1-norm in $\mathbb{R}^n$.

[4]Technically these sectors are called orthant, and they are the n-dimensional generalization of quadrants in 2 dimensions.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

center 0 and double the radius, we can eventually say that $f(x) \geq 0$ whenever $\|x\|_2 \geq \sqrt{n}/\varepsilon$. This means that if $\|x^0\|_2 < \sqrt{n}/\varepsilon$, then the norm of the solution can grow up until it reaches the value $\sqrt{n}/\varepsilon$ but it cannot go above. On the other hand if $\|x^0\|_2 > \sqrt{n}/\varepsilon$ then it will decrease from the beginning until it reaches $\sqrt{n}/\varepsilon$. As a result, for sure $\|x_\varepsilon(t; x^0)\|_2 \leq \|x^0\|_2 + \sqrt{n}/\varepsilon$ for all $t > 0$. □

Notice that this proof relies in a crucial way to the presence of the regularization parameter ε and as $\varepsilon \to 0$ the result expressed by Prop. 1 is empty.

Antisymmetric weight matrix. While the regularization term ε enforces the stability of the forward dynamics, it also favors exponentially suppressing modes, a fact that can be problematic in tasks where memory is important. Therefore another effective approach to ensure the stability of Eq. (1) is to enforce specific spectral properties of the weight matrix W, that is the matrix collecting all the w_{ij}'s. Specifically it has been shown that [13] a simple policy to ensure a stable forward propagation is to constraint the weight matrix to be antisymmetric.

3 Hamiltonian Learning

We can now study the stability of the overall learning process, using a unified dynamical framework where also the "backward" phase can be described in terms of an ODE system for a set of adjoint variables. In particular we focus on the learning scheme that has been recently proposed in [12], based on results of [14], where a class of learning problems over time is formulated using optimal control. Following this approach, the way in which the weights of the model change due to interaction with the environment are described through the dynamics of so called *costate* variables p and q. A "sign flip" strategy allows the model to learn by only going forward in time [12]. The costate p has the same the dimension of the state variable x in Eq. (1) and can be thought as a surrogate of the δ-error in backprop [14, Corollary 1]. The costate q is associated directly with the temporal variations of the weights, so for each weight w_{ij} there is a costate q_{ij}.

In the present work we study the stability of the *free dynamics* of the overall system, i.e., the behavior when no input or supervision is provided,[5] with a weight decay term with strength $\alpha > 0$. For convenience in the notation, we rewrite the main equations using a generic weight matrix W, instead of the graph-based formalism we followed in Section 2. In particular, we indicate with Q the costate matrix associated to W, where $Q_{ij} = q_{ij}$ and $W_{ij} = w_{ij}$. With these assumptions, the set of equations that defines the Hamiltonian Learning apporoach (cfr. [12, Eq. $\mathcal{E} \star 1$–$\mathcal{E} \star 4$] and [14, Eq. (18)]) are:

$$\begin{cases} \gamma \dot{x}_i = -\varepsilon x_i + \sigma\Big(\sum_{j=1}^n W_{ij} x_j\Big) \\ \dot{W}_{ij} = -Q_{ij} \\ \dot{p}_i = -(\theta + \varepsilon/\gamma) p_i + (1/\gamma) \sum_{k=1}^n (W')_{ik} \sigma'\Big(\sum_{j=1}^n W_{kj} x_j\Big) p_k \\ \dot{Q}_{ij} = -\theta Q_{ij} + (1/\gamma) \sigma'\Big(\sum_{m=1}^n W_{im} x_m\Big) p_i x_j + \alpha W_{ij} \end{cases} \tag{4}$$

[5] We broadly mean any kind of signal to which we compare the output of the model.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

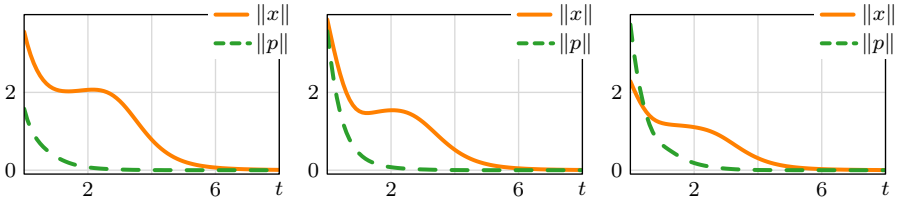

Fig. 1: Evolution of the norm of state x and costate p for $n = 10$, $\varepsilon = \gamma = \alpha = \theta = 1$ for three different random initialization of Eq. (4).

where $\theta > 0$ is a dissipation term.

Conjecture 1. *The system in Eq.* (4) *is stable.*

Notice that a proof of this conjecture in the case where the weight matrix W is antisymmetric is easily conceivable: as we already remarked in Section 2, antisymmetry alone guarantees [13] stability of x, since the Jacobian associated to the right hand side of the dynamical system is proportional (through a positive constant) to the weight matrix W itself. Looking at Eq. (4) we then realize that the same thing applies also the equation of the costate p. In fact, in this case, the Jacobian will be proportional to W' that, if W is antisymmetric, is an antisymmetric matrix itself. Moreover, regardless of the structure of W we experimentally evaluated that when $\varepsilon > 0$, the system in Eq. (4) appears to be stable, as shown in Fig. 1.

Local stability of a single neuron. We formally explore the stability in the case of a single neural unit. Let us set $\varepsilon = \gamma = 1$, then for $n = 1$ the system in Eq. (4) is in the normal form $\dot{u} = F(u)$ with

$$F(u) = \begin{pmatrix} -u_1 + \sigma(u_1 u_2) \\ -u_4 \\ -(\theta+1)u_3 + \sigma'(u_1 u_2)u_3 u_2 \\ -\theta u_4 + \sigma'(u_1 u_2)u_3 u_1 + u_2 \end{pmatrix}.$$

Hence a fixed point of the dynamics, i.e., a solution of $F(u) = 0$, is $u = 0$. Moreover the Jacobian matrix $(\partial F/\partial u)(u)$ is

$$\begin{pmatrix} -1+\sigma'(u_1u_2)u_2 & \sigma'(u_1u_2)u_1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ \sigma''(u_1u_2)u_2^2u_3 & [\sigma''(u_1u_2)u_1u_2+\sigma'(u_1u_2)]u_3 & -(\theta+1)+\sigma'(u_1u_2)u_2 & 0 \\ [\sigma''(u_1u_2)u_1u_2+\sigma'(u_1u_2)]u_3 & 1+\sigma''(u_1u_2)u_1^2u_3 & \sigma'(u_1u_2)u_1 & -\theta \end{pmatrix}.$$

The Jacobian evaluated at $u = 0$, i.e. $(\partial F/\partial u)(0)$, has the following eigenvalues: $\lambda_1 = -1$, $\lambda_2 = -1 - \theta$, $\lambda_3 = (-\theta - \sqrt{\theta^2 - 4})/2$, $\lambda_4 = (-\theta + \sqrt{\theta^2 - 4})/2$ which are all real and negative giving an asymptotically locally stable behaviour [15].

4 Conclusions and Future Work

The stability of the state and costate dynamics were analyzed for a class of continuous-time recurrent neural model, learning over time in a possibly lifelong

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

manner. Existing results were re-framed in the context of Hamiltonian Learning [12], highlighting the importance of the type of neuron model and of the structure of the weight matrices to yield stable dynamics both in the computation of *state* and *costate*. These results promote further investigations in the field of learning over time, especially in a collectionless framework [7], which represents the research direction of our future work.

References

[1] Matteo Tiezzi, Michele Casoni, Alessandro Betti, Marco Gori, and Stefano Melacci. State-space modeling in long sequence processing: A survey on recurrence in the transformer era. *arXiv 2406.09062*, 2024.

[2] Alessandro Betti, Marco Gori, and Stefano Melacci. Learning visual features under motion invariance. *Neural Networks*, 126:275–299, 2020.

[3] Eli Verwimp, Rahaf Aljundi, Shai Ben-David, Matthias Bethge, Andrea Cossu, et al. Continual learning: Applications and the road forward. *Transactions on Machine Learning Research*, 2024.

[4] Zonglei Chen, Minbo Ma, Tianrui Li, Hongjun Wang, and Chongshou Li. Long sequence time-series forecasting with deep learning: A survey. *Information Fusion*, 97:101819, 2023.

[5] Albert Bifet, Ricard Gavalda, Geoffrey Holmes, and Bernhard Pfahringer. *Machine learning for data streams: with practical examples in MOA*. MIT press, 2023.

[6] Michele Casoni, Tommaso Guidi, Matteo Tiezzi, Alessandro Betti, Marco Gori, and Stefano Melacci. Pitfalls in processing infinite-length sequences with popular approaches for sequential data. In *Artificial Neural Networks in Pattern Recognition*, pages 37–48. Springer Nature Switzerland, 2024.

[7] Marco Gori and Stefano Melacci. Collectionless artificial intelligence. *arXiv 2309.06938*, 2023.

[8] Pearlmutter. Learning state space trajectories in recurrent neural networks. In *International 1989 Joint Conference on Neural Networks*, pages 365–372. IEEE, 1989.

[9] Huaguang Zhang, Zhanshan Wang, and Derong Liu. A comprehensive review of stability analysis of continuous-time recurrent neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 25(7):1229–1262, 2014.

[10] James Morrill, Cristopher Salvi, Patrick Kidger, and James Foster. Neural rough differential equations for long time series. In *International Conference on Machine Learning*, pages 7829–7838. PMLR, 2021.

[11] Kazuki Irie, Francesco Faccio, and Jürgen Schmidhuber. Neural differential equations for learning to program neural nets through continuous learning rules. In *Advances in Neural Information Processing Systems*, volume 35, pages 38614–38628. Curran Associates, Inc., 2022.

[12] Stefano Melacci, Alessandro Betti, Michele Casoni, Tommaso Guidi, Matteo Tiezzi, and Marco Gori. A unified framework for neural computation and learning over time. *arXiv 2409.12038*, 2024.

[13] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1):014004, 2017.

[14] Alessandro Betti and Marco Gori. Nature-inspired local propagation. In *Advances in Neural Information Processing Systems*, 2024.

[15] Clark Robinson. *Dynamical systems: stability, symbolic dynamics, and chaos*. CRC press, 1998.