**Abstract**

Recurrent neural networks (RNNs) are wide-spread machine learning tools for modeling sequential and time series data. They are notoriously hard to train because their loss gradients backpropagated in time tend to saturate or diverge during training. This is known as the exploding and vanishing gradient problem. Previous solutions to this issue either built on rather complicated, purpose-engineered architectures with gated memory buffers, or - more recently - imposed constraints that ensure convergence to a fixed point or restrict (the eigenspectrum of) the recurrence matrix. Such constraints, however, convey severe limitations on the expressivity of the RNN. Essential intrinsic dynamics such as multistability or chaos are disabled. This is inherently at disaccord with the chaotic nature of many, if not most, time series encountered in nature and society. Here we offer a comprehensive theoretical treatment of this problem by relating the loss gradients during RNN training to the Lyapunov spectrum of RNN-generated orbits. We mathematically prove that RNNs producing stable equilibrium or cyclic behavior have bounded gradients, whereas the gradients of RNNs with chaotic dynamics always diverge. Based on these analyses and insights, we offer an effective yet simple training technique for chaotic data and guidance on how to choose relevant hyperparameters according to the Lyapunov spectrum.

**Introduction**

Recurrent neural networks (RNNs) are widely used across various fields in engineering and science for learning sequential tasks or modeling and predicting time series (Lipton et al., 2015). Yet, they struggle when long-term temporal dependencies, very slow, or hugely varying time scales are involved (Hochreiter, 1991; Bengio et al., 1994; Schmidt et al., 2021; Li et al., 2018; Rusch & Mishra, 2021a). Time series or sequential data with such properties are, however, very common in fields like climate physics (Thomson, 1990), neuroscience (Fusi et al., 2007; Russo & Durstewitz, 2017), ecology (Turchin & Taylor, 1992), or language processing (Cho et al., 2014b). Training RNNs on such data is hard because the loss gradients backpropagated in time easily saturate or diverge in this process. This is commonly referred to as the exploding and vanishing gradient problem (EVGP) (Hochreiter, 1991; Bengio et al., 1994; Pascanu et al., 2013).

One solution to the EVGP is based on specifically designed RNN architectures with gating mechanisms, such as long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) or gated recurrent units (GRU) (Cho et al., 2014a). These architectures allow states at earlier time steps to more easily influence activity much later through a kind of protected memory buffer, thus alleviating the EVGP by structural design. In practice, such models need to be backed up by further techniques like gradient clipping to keep the gradients in check (Pascanu et al., 2013). The relatively complex architectural design of these networks impedes their mathematical analysis and requires reverse engineering after training (Maheswaranathan et al., 2019; Monfared & Durstewitz, 2020a;b; Schmidt
et al., 2021). Partly to forego these complications, a variety of other solutions has been proposed recently, imposing restrictions on the recurrence matrix to bound the gradients (Arjovsky et al., 2016; Chang et al., 2019), or enforcing global stability by design or regularization (Erichson et al., 2021; Kolter & Manek, 2019). Often these procedures dramatically curtail the expressivity of the RNN (Kerg et al., 2019; Orhan & Pitkow, 2020; Schmidt et al., 2021); in particular, they rule out chaotic dynamics.

This is at odds with the plethora of chaotic phenomena in nature, engineering, and society. Chaotic dynamics are commonplace, almost default in any complex physical or biological system. This includes scientific areas as diverse as neuroscience (Durstewitz & Gabriel, 2007; van Vreeswijk & Sompolinsky, 1996), physiology (Kesma et al., 2020), geophysics (Sivakumar, 2004), climate systems (Tziperman et al., 1997), astrophysics (Laskar & Robutel, 1993), ecology (Duarte et al., 2010), chemical reactions (Field & Györgyi, 1993), cell (Olsen & Degn, 1977) or population (May, 1987) biology. Chaotic phenomena are also crucial for the understanding of societal and epidemiological processes, such as the spread of diseases (Mangiarotti et al., 2020), or in economics (Faggini, 2014). They are further relevant in purely technical contexts such as electrical engineering (Tehtiga et al., 2019; Kamdjue Kengne et al., 2021) or laser optics (Kantz et al., 1993). They have even been suggested to play an up-to-now largely neglected, but potentially very significant role in speech recognition (Sabanal & Nakagawa, 1996) and natural language processing (Inoue et al., 2021). Hence, in almost any practical setting, chaotic phenomena abound. They cannot, in general, be ignored when devising RNN training algorithms.

Here we offer a comprehensive theoretical treatment of the relation between RNN dynamics and the behavior of the loss gradients during training. We find a close connection between an RNN’s loss gradients and the largest Lyapunov exponent of its freely generated orbits. We mathematically prove that RNNs producing stable fixed point or cyclic behavior have bounded gradients. Crucially, however, the loss gradients of RNNs producing chaotic dynamics always diverge. Hence, the chaotic nature of many time series data induces a principle problem, and, despite significant efforts in the past to solve the EVGP, training RNNs on such data remains an open issue. While our main contribution in this work is of theoretical nature, we propose an easy, yet effective training technique, sparsely forced BPTT, for chaotic data which is able to learn the underlying dynamics despite exploding gradients.

2 RELATED WORKS

Exploding and vanishing gradients. While ‘classical’ remedies of the EVGP (Hochreiter, 1991; Bengio et al., 1994; Pascanu et al., 2013) rest on purpose-tailored architectures with gating mechanisms, which safeguard information flow across longer temporal distances (Hochreiter & Schmidhuber, 1997; Cho et al., 2014a), the focus has recently shifted to simpler RNNs that address the EVGP by restricting the recurrence matrix to be orthogonal (Henaff et al., 2016; Helfrich et al., 2018; Jing et al., 2019), unitary (Arjovsky et al., 2016), or antisymmetric (Chang et al., 2019), or by ensuring globally stable fixed point solutions through co-trained Lyapunov functions (Kolter & Manek, 2019). However, all these approaches impose strong limitations on the dynamical repertoire of the RNN, enforcing global convergence to fixed points or simple cycles. In doing so, they drastically reduce the expressiveness of these models (Kerg et al., 2019; Orhan & Pitkow, 2020). To address this problem, Erichson et al. (2021) somewhat relaxed the constraints on the recurrence matrix by introducing a skew-symmetric decomposition combined with a Lipschitz condition on the activation function. Another recent approach discretizes oscillator ODEs to arrive at a stable system of coupled (Rusch & Mishra, 2021a) or independent (Rusch & Mishra, 2021b) oscillators which increase the RNN’s expressiveness while bounding its gradients. By design, neither of these architectures is capable of producing chaotic dynamics, however, as the underlying ODEs do not allow for exponential divergence of close-by trajectories (a prerequisite for chaos). Hence, as it currently stands, only gated architectures (LSTM, GRU) or regularized PLRNNs (Schmidt et al., 2021) are in principle able to produce chaotic dynamics, but do not fully alleviate the exploding gradient problem yet (as will be shown below).

Learning dynamical systems. Surprisingly disconnected from the work on the EVGP and learning long-term dependencies, a huge and long-standing literature deals with training RNNs on nonlinear

---

1 We make this point more formal in Proposition 2 in Appx. A.1.1.
we can recursively rewrite eq. (1) as

Assuming we start at some initial value of the dynamical variables (namely, hidden and cell states; see Appx. A.1.3). We have

where \( F_\theta \) is an element-wise activation function. For instance, for standard RNNs (\( z_t \) is an element-wise activation function like tanh or a rectified linear unit (ReLU)). For LSTMs, \( F_\theta(z_{t-1}, s_t) \) looks considerably more complicated, and the latent states \( z_t = (h_t, c_t) \) are given by a concatenation of two different types of dynamical variables (namely, hidden and cell states; see Appx. A.1.3).

Assuming we start at some initial value \( z_1 \in \mathbb{R}^M \), and given a sequence of external inputs \( S = \{ s_t \} \), we can recursively rewrite eq. (1) as

\[
z_T = F_\theta(F_\theta(\ldots F_\theta(z_1, s_2)\ldots))) =: F_\theta^{T-1}(z_1, s_2).
\]
In DS theory, we characterize the long-term behavior of such sequences by its spectrum of Lyapunov exponents. The Lyapunov exponents estimate the exponential growth rates in different local directions of the system’s state space, and the largest Lyapunov exponent gives the dominant exponential behavior. Let us denote the system’s Jacobian at time $t$ by

$$J_t := \frac{\partial F_\theta(z_{t-1}, s_t)}{\partial z_{t-1}} = \frac{\partial z_t}{\partial z_{t-1}}.$$  

(3)

For instance, for standard RNNs we would have $J_t = W \text{diag}(f'(W z_{t-1} + B s_t + h))$, where $\text{diag}$ denotes a diagonal matrix for which the $i$-th diagonal entry is the derivative of $f$ w.r.t. $z_{t-1}$. Then, the maximal Lyapunov exponent along an RNN trajectory $\{z_1, z_2, \cdots, z_T, \cdots\}$ is defined as

$$\lambda_{max} := \lim_{T \to \infty} \frac{1}{T} \log \left| \prod_{r=0}^{T-2} J_{T-r} \right|,$$

(4)

where $\| \cdot \|$ denotes the spectral norm (or any subordinate norm) of a matrix. If $\lambda_{max} < 0$ nearby trajectories will ultimately converge to a fixed point or cycle, while for $\lambda_{max} > 0$ (a necessary condition for chaos) initially nearby trajectories will exponentially separate, i.e. we will have divergence along one (or more) directions in state space. This accounts for the sensitive dependence on initial conditions in chaotic systems.

Now let $\mathcal{L}(W, B, h)$ be some loss function employed for RNN training that decomposes in time as $\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t$. Suppose we fancy BPTT as our training algorithm (similar derivations could be performed for RTRL), we recursively develop the loss gradients w.r.t. some RNN parameter $\theta$ in time (i.e., across layers of the RNN unrolled in time) as

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t}{\partial \theta} \quad \text{with} \quad \frac{\partial \mathcal{L}_t}{\partial \theta} = \sum_{r=1}^{t} \frac{\partial \mathcal{L}_t}{\partial z_t} \frac{\partial z_t}{\partial z_r} \frac{\partial^+ z_r}{\partial \theta},$$

(5)

and

$$\frac{\partial z_t}{\partial z_r} = \frac{\partial z_t}{\partial z_{t-1}} \frac{\partial z_{t-1}}{\partial z_{t-2}} \cdots \frac{\partial z_{t-r+1}}{\partial z_r} = \prod_{k=0}^{t-r-1} \frac{\partial z_{t-k}}{\partial z_{t-k-1}} = \prod_{k=0}^{t-r-1} J_{t-k},$$

(6)

where $\partial^+$ denotes the immediate derivative. Now observe that the behavior of the loss gradients crucially depends on the product series of Jacobians in eqn. (6): If the maximum absolute eigenvalues of the Jacobians $J_t$ will, in the geometric mean, be larger than 1 (i.e., $\left\| \prod_{r=0}^{T-2} J_{T-r} \right\|^{1/T} > 1$), gradients will explode as $T \to \infty$, while they will saturate if $\left\| \prod_{r=0}^{T-2} J_{T-r} \right\|^{1/T} < 1$. Thus, the key point to note is that the same terms that occur in the definition of the Lyapunov spectrum, eqn. (4), resurface in the loss gradients, eqn. (5) & (6). This accounts for the tight links between system dynamics and gradients. The detailed mathematical picture is both more complicated and more interesting, however, as we will work out in the following sections.

### 3.2 Fixed points and cyclic dynamics

Let us start by considering the simplest types of autonomous dynamics that can occur in RNNs (or any discrete-time DS): fixed points and cycles. In fact, by far most of the literature on global stability in RNNs and on loss gradients focused on just fixed points ([Chang et al. 2019](https://example.com), [Kolter & Manek 2019](https://example.com), [Erichson et al. 2021](https://example.com)), with only few authors who recently started to also connect cyclic behavior to loss gradients ([Schmidt et al. 2021](https://example.com), [Rusch & Mishra 2021a](https://example.com)). Recall that a fixed point of a recursive map $z_t = F(z_{t-1})$ is defined as a point $z^*$ for which we have $z^* = F(z^*)$. Likewise, a $k$-cycle ($k > 1$) is a set of temporally consecutive periodic points $P_k := \{z_t, z_{t+k}, \ldots, z_{t+k-1}\}$ = $\{z_t, F(z_t), \ldots, F^{k-1}(z_t)\}$ that we obtain from recursive application of the map such that each of the cyclic points $z_{t+k} \in P_k$ is a fixed point of the $k$ times iterated map $F^k$ (with $k$ being the smallest positive integer for which this holds). To simplify the subsequent treatment, we will collectively refer to fixed points and cycles as $k$-cycles ($k \geq 1$). Further recall that a fixed point or $k$-cycle is called **stable** if the maximum absolute eigenvalue of the Jacobian evaluated at that point is smaller than 1, **neutrally stable** if exactly 1, and **unstable** otherwise. Although the results we develop in this
and the following sections will hold more widely, we will restrict our attention to recursive maps \( F_\theta \) from the class of RNNs \( \mathcal{R} = \{ \text{standardRNN}, \text{LSTM}, \text{GRU}, \text{PLRNN} \} \) (see Appx. A.1 for details).

Based on the observations made in the previous sections we can state the following theorem that links RNN dynamics and loss gradients:

**Theorem 1.** Consider an RNN \( F_\theta \in \mathcal{R} \) parameterized by \( \theta \), and assume that it converges to a stable fixed point or \( k \)-cycle \( \Gamma_k \) \((k \geq 1)\) with \( B_{\Gamma_k} \) as its basin of attraction. Then for every \( z_1 \in B_{\Gamma_k} \) (i) the Jacobian \( \frac{\partial z_T}{\partial z_1} \) exponentially vanishes as \( T \to \infty \); (ii) for \( \Gamma_k \) the tangent vectors \( \frac{\partial z_T}{\partial \theta} \), and thus the gradient of the loss function, \( \frac{\partial L}{\partial \theta} \), will be bounded from above, i.e. will not diverge for \( T \to \infty \); and (iii) for the PLRNN \([23] \) both \( \| \frac{\partial z_T}{\partial z_1} \| \) and \( \| \frac{\partial L}{\partial \theta} \| \) will remain bounded for every \( z_1 \in B_{\Gamma_k} \) as \( T \to \infty \).

**Proof.** (i) Assume that \( \Gamma_k \) is a stable \( k \)-cycle \((k \geq 1)\) denoted by

\[
\Gamma_k = \{ z_1, z_2, \ldots, z_T, \ldots \} = \{ z_{t+1}, z_{t-k+1}, \ldots, z_{t+1-(k-1)}, z_{t+1}, z_{t+1-k}, \ldots, z_{t+1-(k-1)}, \ldots \}.
\]  

(7) Then, the largest Lyapunov exponent of \( \Gamma_k \) is given by

\[
\lambda_{\Gamma_k} = \lim_{T \to \infty} \frac{1}{T} \ln \left\| J_T^* J_T^{*-1} \cdots J_2^* \right\| = \lim_{j \to \infty} \frac{1}{j} \ln \left\| \left( \prod_{s=0}^{k-1} J_{t+1-s}^* \right)^j \right\|.
\]  

(8) By assumption of stability of \( \Gamma_k \) we have \( \lambda_{\Gamma_k} < 0 \) and also \( \rho(\prod_{s=0}^{k-1} J_{t+1-s}) < 1 \), which implies

\[
\lim_{T \to \infty} \left\| J_T^* J_T^{*-1} \cdots J_2^* \right\| = \lim_{j \to \infty} \left( \prod_{s=0}^{k-1} J_{t+1-s}^* \right)^j = 0.
\]  

(9) Now suppose that \( O_{z_1} \) is an orbit of \( \{1\} \) converging to \( \Gamma_k \), i.e. \( z_1 \in B_{\Gamma_k} \). Since \( O_{z_1} \) and \( \Gamma_k \) have the same largest Lyapunov exponent, we have

\[
\lambda_{O_{z_1}} = \lim_{T \to \infty} \frac{1}{T} \ln \left\| J_T J_T^{-1} \cdots J_2 \right\| = \lambda_{\Gamma_k} < 0,
\]  

(10) and hence for \( z_1 \in B_{\Gamma_k} \)

\[
\lim_{T \to \infty} \left\| \frac{\partial z_T}{\partial z_1} \right\| = \lim_{T \to \infty} \left\| J_T J_T^{-1} \cdots J_2 \right\| = 0.
\]  

(11) (ii) & (iii) See Appx. A.2.1. \( \square \)

**Remark 1.** The result of Theorem 1 part (i) will be generally true for any first-order-Markovian recursive map \([7]\), but the conclusions in part (ii) may hinge on its specific definition.

The results above ensure that loss gradients will not diverge (explode) as \( T \to \infty \) in RNNs that are “well-behaved” in the sense that they converge to a fixed point or cycle. This is a generalization of the results given in Theorem 1 in [Schmidt et al., 2021], where this was shown only a) for the specific class of PLRNNs and b) for specific constraints imposed on the eigenvalue spectrum of the RNN’s Jacobians which were relaxed in our theorem above.

While our treatment above is centered on the “exploding-gradients” case, various architectural modifications or regularization techniques can ensure that gradients do not vanish either, i.e. remain bounded from below as well. This was established, for instance, in [Schmidt et al., 2021] for PLRNNs using ‘manifold attractor regularization’. In Appx. A.2.1 we show that the results from Theorem 2 from [Schmidt et al., 2021] on doubly bounded (from below and above) loss gradients can indeed be extended to the more general case covered by Theorem 1 above.

### 3.3 Chaotic dynamics

We will now consider the all-important chaotic case. Let \( F \) be a recursive map and \( O_{z_1} = \{ z_1, z_2, z_3, \cdots \} \) be an orbit of \( F \). The orbit is chaotic if (i) it is not asymptotically periodic and (ii) has at least one positive Lyapunov exponent ([Glendinning & Simpson, 2021](#) [Meiss, 2007](#). If
the system’s invariant set is bounded, condition (ii) is considered a standard signature of chaos, as in this case two nearby orbits separate exponentially fast, but at the same time their mutual separation cannot go to infinity so that there are also folds. The following theorem states the sufficient condition for exploding gradients:

**Theorem 2.** Suppose that an RNN \( F_\theta \in \mathcal{R} \) (parameterized by \( \theta \)) has a chaotic attractor \( \Gamma^* \) with \( \mathcal{B}_{\Gamma^*} \) as its basin of attraction. Then, for every orbit with \( z_1 \in \mathcal{B}_{\Gamma^*} \), (i) the Jacobians connecting temporally distal states \( z_T \) and \( z_t \) (\( T > t \)), \( \frac{\partial z_T}{\partial z_t} \), will exponentially explode for \( T \to \infty \), and (ii) the tangent vector \( \frac{\partial z_T}{\partial z_t} \) and so the gradients of the loss function, \( \frac{\partial c_T}{\partial \theta} \), will diverge as \( T \to \infty \).

**Proof.** Let the RNN \( F_\theta \in \mathcal{R} \) have a chaotic orbit denoted by \( \Gamma^* = \{ z_1^*, z_2^*, \ldots, z_T^*, \ldots \} \). Then, denoting by \( J^*_T \) the Jacobian of \( (1) \) at \( z_T^* \in \Gamma^* \), the largest Lyapunov exponent of \( \Gamma^* \) is given by

\[
\lambda = \lim_{T \to \infty} \frac{1}{T} \ln \| J^*_T J^*_{T-1} \cdots J^*_2 \|.
\]  

(12)

Since \( \Gamma^* \) is chaotic, so \( \lambda > 0 \). Hence, from (12), it is concluded that

\[
\lim_{T \to \infty} \| J^*_T J^*_{T-1} \cdots J^*_2 \| = \lim_{T \to \infty} \| \frac{\partial z_T^*}{\partial z_1^*} \| = \infty, \quad T \gg t.
\]  

(13)

Now, according to Oseledec’s multiplicative ergodic Theorem, nearly all the points in the basin of attraction of \( \Gamma^* \) have the same largest Lyapunov exponent \( \lambda \). Thus (13) holds for every \( z_1 \in \mathcal{B}_{\Gamma^*} \).

(ii) See Appx. A.2.2

**Remark 2.** The first part of Theorem 2 holds for all first-order-Markovian recursive maps (7). Note that for LSTMs, \( \frac{\partial z_T}{\partial z_1} \) denotes the full Jacobian of both hidden and cell states.

We collect some further mathematical results and remarks related to Theorem 2 in Appx. A.3.1

Hence, the essential result is that for all popular RNNs \( \mathcal{R} \) and activation functions, loss gradients will inevitably diverge if the RNN latent states converge to a chaotic attractor.

### 3.4 Quasi-periodicity

Quasi-periodicity is a long-term behavior which occurs on a torus and, superficially, bears some similarity to chaos in the sense that, strictly speaking, orbits are also aperiodic. That is, as \( T \to \infty \), trajectories will never close up with themselves, i.e. the time-domain solution will never repeat itself. Moreover, every trajectory becomes arbitrarily close to any point on the torus, that is, it is dense. One important difference between quasi-periodic and chaotic systems is, however, that in a quasi-periodic system, as time passes, two close initial conditions are linearly diverging, while in a chaotic system the divergence is exponential.

**Theorem 3.** Assume that an RNN \( F_\theta \in \mathcal{R} \) (parameterized by \( \theta \)) has a quasi-periodic attractor \( \Gamma \) with \( \mathcal{B}_\Gamma \) as its basin of attraction. Then, for every \( z_1 \in \mathcal{B}_\Gamma \)

\[
\forall 0 < \epsilon < 1 \exists T_0 > 1 \text{ s.t. } \forall T \geq T_0 \implies (1 - \epsilon)^T < \left\| \frac{\partial z_T}{\partial z_1} \right\| < (1 + \epsilon)^T - 1.
\]  

(14)

**Proof.** See Appx. A.2.3

According to Theorem 3, for every orbit converging to a quasi-periodic attractor, the Jacobians \( \frac{\partial z_T}{\partial z_1} \) may diverge or vanish as \( T \to \infty \), but this will not occur exponentially fast as \( T \to \infty \). Thus, even for bounded non-chaotic RNNs we may sometimes stumble into the problem of diverging gradients. Although this may be a less common scenario, we point out it may occur if we train RNNs on real data from oscillatory systems with incommensurate frequencies, as for instance encountered in electronic engineering.
3.5 Other connections between dynamics and gradients

As the previous sections elucidated, there is a direct link between the norms of the Jacobians of the RNN along trajectories and the EVGP. By observing this link, we can formulate some general conditions that will have implications for the behavior of the gradients regardless of the limiting behavior of the RNN, as collected in the following theorem (see also Appx. A.3.2):

**Theorem 4.** Let \( O_{z_t} = \{ z_1, z_2, \ldots, z_T, \ldots \} \) be a sequence (orbit) generated by an RNN \( F_\theta \in \mathcal{R} \) parameterized by \( \theta \), and \( P_T := J_T - I \), \( T = 2, 3, \ldots \).

(i) Assume that \( O_{z_t} \) is an orbit for which \( \| \frac{\partial z_t}{\partial \theta} \| \leq \xi \) \( \forall t \). If \( \sum_{T=2}^{\infty} \| J_T \| < \infty \), then the Jacobian \( \frac{\partial z_T}{\partial z_1} \), the tangent vector \( \frac{\partial z_T}{\partial \theta} \) and thus the gradient of the loss function, \( \frac{\partial \mathcal{L}}{\partial \theta} \), will be bounded for \( T \to \infty \).

(ii) If \( \sum_{T=2}^{\infty} \| P_T \| < \infty \), then the Jacobian \( \frac{\partial z_T}{\partial z_1} \) will neither vanish nor explode as \( T \to \infty \).

**Proof.** See Appx. A.2.4.

Part (i) of Theorem 4 relaxes some of the conditions required in Theorem 1 for bounded gradients by imposing a Lipschitz condition on the immediate derivatives. Part (ii) generalizes conditions satisfied, for instance, in orthogonal (unitary) RNNs (Arjovsky et al., 2016; Henaff et al., 2016) or fully regularized PLRNNs (Schmidt et al., 2021).

4 Empirical and simulation results

Our theoretical results imply that chaotic time series pose a principle challenge for RNN training that cannot easily be circumvented through specifically designed architectures, constraints, or regularization criteria. If the underlying DS we aim to capture is chaotic, loss gradients propagated back in time will inevitably explode. Hence we need to curtail gradients in an ideal way, as illustrated in the next sections.

4.1 Training on systems with exploding gradients by sparse teacher forcing

Here we revive the old idea of TF (Williams & Zipser, 1989) as a mechanism for effectively truncating error gradients while training. However, we would like to do this such that important information about the system dynamics does not get lost, for which Lyapunov theory offers some guidance. For this we suggest sparsely forced BPTT, which - in contrast to conventional TF - does not force the system back onto the true trajectory all or most of the time, but effectively “re-calibrates” it only at certain time points chosen wisely according to the system’s local divergence rates.

Assume we want to train an RNN with hidden states \( z_t \in \mathbb{R}^M \) and linear (or affine) output layer on a time-series \( \{ x_1, x_2, \ldots, x_T \} \) generated by a chaotic system. The linear output layer \( \hat{x}_t = B z_t \), \( B \in \mathbb{R}^{N \times M} \), maps the RNN hidden states into the observation space. This allows us to modify the original TF procedure by constructing a control series \( \{ \tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_T \} \) from the observations by “inverting” the linear output mapping

\[
\tilde{z}_t = (B^T B)^{-1} B^T x_t.
\]

The idea is to supply this control signal only sparsely, separated by the learning interval \( \tau \) between consecutive forcings. Hence, defining \( T = \{ n\tau + 1 \}_{n \in \mathbb{N}_0} \) as the set of all time points at which we force the RNN onto the ‘true’ values, the RNN updates can be written as

\[
z_{t+1} = \begin{cases} RNN(\tilde{z}_t) & \text{if } t \in T \\ RNN(z_t) & \text{else} \end{cases}.
\]

This forcing is applied after calculation of the loss, such that \( \mathcal{L}_t = \| x_t - B z_t \|^2 \) irrespective of whether \( t \) is in \( T \) or not (and of course it is applied only during training, not at test time!). Replacing hidden states \( z_t \) with their teacher-forced signals \( \tilde{z}_t \) simply breaks divergence between true and predicted trajectories at time points \( t \in T \), and also cuts off the Jacobians by breaking the temporal contingency (for details see Appx. A.7). The learning interval \( \tau \) thus controls how many time steps
are included in the gradient calculation and has to be chosen with care such as to balance the effects of exploding gradients vs. those of loosing relevant time scales and long-term dependencies. Here we suggest to choose $\tau$ in accordance with the system’s Lyapunov spectrum or, more specifically, the predictability time $\tau_{\text{pred}}$ (Bezruchko & Smirnov, 2010)

$$\tau_{\text{pred}} = \frac{\ln 2}{\lambda_{\text{max}}}.$$  \hspace{1cm} (17)

We emphasize that this simple recipe for addressing the exploding gradient problem is based on modifying the training routine, and is thus in principle applicable to any model architecture.  

4.2 Example 1: Lorenz and Rössler systems in chaotic regime

Let us illustrate these ideas on two classical textbook examples of chaotic DS, the chaotic Lorenz attractor and the Rössler system (see Appx. A.4 for details). Trajectories of length $T = 1000$ were repeatedly drawn from these systems, on which we trained a PLRNN, a vanilla RNN with tanh activation function, and a LSTM by stochastic gradient descent (SGD) to minimize the MSE loss between predicted and actual observations. As optimizer we used Adam (Kingma & Ba, 2015) from PyTorch (Paszke et al., 2017) with a learning rate of 0.001. For all models, training proceeded solely by sparsely forced BPTT and did not employ gradient clipping or any other technique that may interfere with optimal loss truncation.

In nonlinear DS reconstruction, we are mainly interested in reproducing invariant properties of the underlying system such as the attractor geometry (or topology; Takens (1981); Sauer et al. (1991)) or the frequency composition (i.e., time-averaged properties), while measures like ahead-prediction errors are less meaningful especially on chaotic time series (Wood, 2010; Koppe et al., 2019). Thus, in evaluating training performance, here we follow Koppe et al. (2019) in using a Kullback-Leibler divergence $D_{\text{stsp}}$ to quantify the agreement between observed and generated probability distributions across state-space to assess the overlap in attractor geometry (Appx. A.5). Moreover, we employ a dimension-wise frequency correlation measure (PSC) to quantify the agreement of power-spectra of the observed and generated time-series (Appx. A.5). Fig. 1 shows the dependence of the reconstruction quality on the learning interval $\tau$ for all RNN architectures on (a) the Lorenz and (b) the Rössler system. Continuous lines = sparsely forced BPTT. Dashed lines = classical BPTT with gradient clipping. Prediction time indicated vertically in black.

2All code produced here is available at [placeholder].
the optimal learning interval that agrees well with the predictability time defined in eqn. \([17]\), where estimates for the maximal Lyapunov exponent were taken from the literature \([\text{Rosenstein et al., 1993}]\). As a reference, dashed lines represent the reconstruction performance for all architectures when trained with classical BPTT and gradient clipping. The training procedure was the same as for sparsely forced BPTT, except that instead of supplying a control-signal, gradients were normalized to 1 prior to each parameter update. As evidenced by the much worse performance, gradient clipping does not effectively address the EVGP, even for LSTMs. This suggests that mere normalization may wipe out essential information about the dynamics.

Figure 2: Lorenz attractor (blue) and example reconstructions by a LSTM (orange) trained with a learning interval (a) chosen too small \((\tau = 5)\), (b) chosen optimally \((\tau = 30)\), and (c) chosen too large \((\tau = 200)\).

4.3 EXAMPLE 2: CHAOTIC WEATHER DATA

As for an empirical example, we trained all RNNs (vanilla RNN, PLRNN, LSTM) on a temperature time series recorded at the Weather Station at the Max Planck Institute for Biogeochemistry in Jena, Germany, spanning the time period between 2009 and 2016. To expose the chaotic behavior and obtain a robust estimate of the maximal Lyapunov exponent, trends and yearly cycles were removed, and nonlinear noise-reduction was performed \([\text{Kantz et al., 1993}; \text{Appx. A.4}]\). The maximal Lyapunov exponent was determined with the TISEAN package \([\text{Hegger et al., 1999}]\), as shown in Figure 3 (a). The value obtained is in close agreement with the literature \([\text{Millán et al., 2010}]\).

Figure 3 shows that also for these empirical data the optimal training interval \(\tau\) agrees well with the predictability time, eqn. \([17]\), for all trained RNNs. Furthermore, as was the case for the DS benchmarks, gradient clipping was not able to satisfactorily tackle the EVGP, even when paired with architectures like LSTMs explicitly designed for alleviating this problem.

Figure 3: (a) The maximal Lyapunov exponent was determined as the slope of the average log-divergence of nearest neighbors in embedding space \((m = \text{embedding dimension})\). (b) Reconstruction quality assessed by attractor overlap (lower = better) and power-spectrum correlation (higher = better). Black vertical lines = \(\tau_{\text{pred}}\).

5 DISCUSSION AND CONCLUSIONS

In this paper we proved that RNN dynamics and loss gradients are intimately related for all major types of RNNs and activation functions. If the RNN is “well behaved” in the sense that its autonomous dynamics converges to a fixed point or cycle, loss gradients will remain bounded, and established remedies \([\text{Hochreiter & Schmidhuber, 1997}; \text{Schmidt et al., 2021}]\) can be used to refrain them from vanishing. However, if the dynamics are chaotic, gradients will always explode. This constitutes a principle problem in RNN training that cannot easily be mastered through architectural design or gradient clipping. It is furthermore a practically highly relevant one, as most time series we encounter in nature, and many from man-made systems as well, are inherently chaotic. While we
do not offer a full solution to this problem here, we suggest it might be tackled in training by taking
a system’s local divergence rates as measured through the Lyapunov spectrum into account. Hence,
rather than conquering the EVGP by structural design or specific constraints or regularization terms,
we recommend to put the focus more on the training process itself. As a step into this direction, we
proposed sparsely forced BPTT, a training technique that pulls trajectories back on track at times
determined by the maximal Lyapunov exponent. We empirically demonstrated that this leads to
optimal reconstruction results for two prominent DS benchmarks and one empirical time series,
regardless of the specific RNN architecture employed in training. We stress that our goal above all
was to provide a mathematically grounded perspective on the problem, with the empirical section
focused on elucidating the practical implications of the theoretical results. For instance, Lyapunov
exponents are empirically often hard to determine, in case of which $\tau$ may be regarded more as a
hyper-parameter that needs tuning. Situations where the RNN is strongly guided by external inputs
(unlike the examples in Sect. [4]) also need further consideration, as external forcing can profoundly
change the dynamics.

AUTHOR CONTRIBUTIONS

ZM, JM, DD conceived & designed the study, ZM contributed the theoretical sect. with inputs from
JM & DD, JM contributed the empirical sect. with inputs from ZM & DD, and all authors wrote the
manuscript.

ACKNOWLEDGMENTS

Funded by the German Research Foundation (DFG) under Germany’s Excellence Strategy – EXC-
2181 – 390900948 (STRUCTURES), and through grant Du 354/10-1 to DD. ———

REFERENCES

Henry D. I. Abarbanel. Predicting the Future: Completing Models of Observed Complex Systems.
Understanding Complex Systems. Springer-Verlag, New York, 2013. ISBN 978-1-4614-7217-
9. doi: 10.1007/978-1-4614-7218-6. URL https://www.springer.com/gp/book/9781461472179

Henry D. I. Abarbanel, Paul J. Rozdeba, and Sasha Shirman. Machine Learning: Deepest Learning
as Statistical Data Assimilation Problems. Neural Computation, 30(8):2025–2055, August 2018.
ISSN 1530-888X. doi: 10.1162/neco_a_01094.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks.
In Proceedings of the 33rd International Conference on International Conference on Machine
Learning - Volume 48, ICML’16, pp. 1120–1128. JMLR.org, 2016.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence
prediction with recurrent Neural networks. In Proceedings of the 28th International Conference
on Neural Information Processing Systems - Volume 1, NIPS’15, pp. 1171–1179, Cambridge,
MA, USA, December 2015. MIT Press.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient
descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166, 1994. doi: 10.1109/
72.279181.

Boris P. Bezruchko and Dmitry A. Smirnov. Extracting Knowledge From Time Series: An
Introduction to Nonlinear Empirical Modeling. Springer Series in Synergetics. Springer-Verlag,
Berlin Heidelberg, 2010. ISBN 978-3-642-12600-0. doi: 10.1007/978-3-642-12601-7. URL
https://www.springer.com/gp/book/9783642126000

Bo Chang, Minmin Chen, Eldad Haber, and Ed H. Chi. AntisymmetricRNN: A dynamical system
view on recurrent neural networks. In International Conference on Learning Representations,
2019. URL https://openreview.net/forum?id=ryxepo0cFX

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties
of neural machine translation: Encoder–decoder approaches. In Proceedings of SSST-8, Eighth
Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 2014.

Kenji Doya. Bifurcations in the learning of recurrent neural networks. In Proceedings 1992 IEEE International Symposium on Circuits and Systems, volume 6, pp. 2777–2780 vol.6, May 1992. doi: 10.1109/ISCAS.1992.230622.

N. Benjamin Erichson, Omri Azencot, Alejandro Queiruga, Liam Hodgkinson, and Michael W. Mahoney. Lipschitz recurrent neural networks. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=-N7PBXqOUJZ

Marisa Faggini. Chaotic time series analysis in economics: Balance and perspectives. Chaos: An Interdisciplinary Journal of Nonlinear Science, 24(4):042101, December 2014. ISSN 1054-1500. doi: 10.1063/1.4903797. URL https://aip.scitation.org/doi/full/10.1063/1.4903797 Publisher: American Institute of Physics.

Stefano Fusi, Wael F. Asaad, Earl K. Miller, and Xiao-Jing Wang. A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. Neuron, 54(2):319–333, April 2007. ISSN 0896-6273. doi: 10.1016/j.neuron.2007.03.017.

Zoubin Ghahramani and Sam Roweis. Learning nonlinear dynamical systems using an em algorithm. In M. Kearns, S. Solla, and D. Cohn (eds.), Advances in Neural Information Processing Systems, volume 11. MIT Press, 1999. URL https://proceedings.neurips.cc/paper/1998/file/0ebcc77dc72360d0eb8e9504c78d38bd-Paper.pdf

Paul A. Glendinning and David J. W. Simpson. A constructive approach to robust chaos using invariant manifolds and expanding cones. Discrete & Continuous Dynamical Systems, 41(7): 3367–3387, 2021.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org

Alex Graves, Greg Wayne, Malcolm Reynolds, and et al. Hybrid computing using a neural network with dynamic external memory. Nature, 538:471—476, 2016. doi: 10.1038/nature20101.
Rainer Hegger, Holger Kantz, and Thomas Schreiber. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9(2):413–435, June 1999. ISSN 1054-1500. doi: 10.1063/1.166424. URL https://aip.scitation.org/doi/citedby/10.1063/1.166424 Publisher: American Institute of Physics.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal Recurrent Neural Networks with Scaled Cayley Transform. In *International Conference on Machine Learning*, pp. 1969–1978. PMLR, July 2018. URL http://proceedings.mlr.press/v80/helfrich18a.html ISSN: 2640-3498.

Mikael Henaff, Arthur Szlam, and Yann LeCun. Recurrent Orthogonal Networks and Long-Memory Tasks. In *International Conference on Machine Learning*, pp. 2034–2042. PMLR, June 2016. URL http://proceedings.mlr.press/v48/henaff16.html ISSN: 1938-7228.

Daniel Hernandez, Antonio Khalil Moretti, Ziqiang Wei, Shreya Saxena, John Cunningham, and Liam Paninski. Nonlinear Evolution via Spatially-Dependent Linear Dynamics for Electrophysiology and Calcium Data. *arXiv preprint arXiv:1811.02459*, 2020. URL http://arxiv.org/abs/1811.02459.

Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen, 1991.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735

Katsuma Inoue, Soh Ohara, Yasuo Kuniyoshi, and Kohei Nakajima. Transient chaos in bert. *arXiv:2106.03181 [nlin]*, June 2021. URL http://arxiv.org/abs/2106.03181 arXiv: 2106.03181.

Li Jing, Caglar Gulcehre, John Peurifoy, Yichen Shen, Max Tegmark, Marin Soljacic, and Yoshua Bengio. Gated Orthogonal Recurrent Units: On Learning to Forget. *Neural Computation*, 31(4):765–783, April 2019. doi: 10.1162/neco_a_01174.

Léandre Kamdjeu Kengne, Justin R. Mboupda Pone, and Hilaire B. Fotsin. On the dynamics of chaotic circuits based on memristive diode-bridge with variable symmetry: A case study. *Chaos, Solitons & Fractals*, 145:110795, April 2021. ISSN 0960-0779. doi: 10.1016/j.chaos.2021.110795. URL https://www.sciencedirect.com/science/article/pii/S0960077921001478

Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2 edition, 2003. doi: 10.1017/CBO9780511755798.

Holger Kantz, Thomas Schreiber, Ingo Hoffmann, Thorsten Buzug, Gerd Pfister, Leci G. Flepp, Josef Simonet, Remo Badii, and Ernst Brun. Nonlinear noise reduction: A case study on experimental data. *Physical Review E*, 48(2):1529–1538, August 1993. doi: 10.1103/PhysRevE.48.1529. URL https://link.aps.org/doi/10.1103/PhysRevE.48.1529 Publisher: American Physical Society.

Matthew B. Kennel, Reggie Brown, and Henry D. I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, 45:3403–3411, Mar 1992. doi: 10.1103/PhysRevA.45.3403. URL https://link.aps.org/doi/10.1103/PhysRevA.45.3403

Giancarlo Kerg, Kyle Goyette, Maximilian Puelma Touzel, Gauthier Gidel, Eugene Vorontsov, Yoshua Bengio, and Guillaume Lajoie. Non-normal recurrent neural network (nrmn): learning long time dependencies while improving expressivity with transient dynamics. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/9d7099d87947faa8d07a272dd6954b80-Paper.pdf
Mounira Kesmia, Soraya Boughaba, and Sabir Jacquir. Control of continuous dynamical systems modeling physiological states. *Chaos, Solitons & Fractals*, 136:109805, July 2020. ISSN 0960-0779. doi: 10.1016/j.chaos.2020.109805. URL [https://www.sciencedirect.com/science/article/pii/S096007792030206X](https://www.sciencedirect.com/science/article/pii/S096007792030206X)

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL [http://arxiv.org/abs/1412.6980](http://arxiv.org/abs/1412.6980)

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014. URL [http://arxiv.org/abs/1312.6114](http://arxiv.org/abs/1312.6114)

J. Zico Kolter and Gaurav Manek. Learning Stable Deep Dynamics Models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper/2019/file/0a4bbceda17a6253386bc9eb45240e25-Paper.pdf](https://proceedings.neurips.cc/paper/2019/file/0a4bbceda17a6253386bc9eb45240e25-Paper.pdf)

Georgia Koppe, Hazem Toutounji, Peter Kirsch, Stefanie Lis, and Daniel Durstewitz. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. *PLOS Computational Biology*, 15(8):1–35, 08 2019. doi: 10.1371/journal.pcbi.1007263. URL [https://doi.org/10.1371/journal.pcbi.1007263](https://doi.org/10.1371/journal.pcbi.1007263)

Jacques Laskar and Philippe Robutel. The chaotic obliquity of the planets. *Nature*, 361(6413): 608–612, February 1993. ISSN 1476-4687. doi: 10.1038/361608a0. URL [https://www.nature.com/articles/361608a0](https://www.nature.com/articles/361608a0)

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5457–5466, 2018. doi: 10.1109/CVPR.2018.00572.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhat-tacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL [https://openreview.net/forum?id=c8P9NQVtmnO](https://openreview.net/forum?id=c8P9NQVtmnO)

Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv:1506.00019 [cs]*, October 2015. URL [http://arxiv.org/abs/1506.00019](http://arxiv.org/abs/1506.00019) arXiv: 1506.00019.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2): 130–141, March 1963. ISSN 0022-4928, 1520-0469. doi: 10.1175/1520-0469(1963)020⟨ 0130:DNF⟩ 2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml](https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml) Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences.

Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper/2019/file/d921c3c762b1522c475ac8fc0811bb0f-Paper.pdf](https://proceedings.neurips.cc/paper/2019/file/d921c3c762b1522c475ac8fc0811bb0f-Paper.pdf)

Sylvain Mangiarotti, Matthieu Peyre, Yixiao Zhang, Maciej Huc, Friederike Roger, and Yvonne Kerr. Chaos theory applied to the outbreak of COVID-19: an ancillary approach to decision making in pandemic context. *Epidemiology and Infection*, 148:e95, May 2020. ISSN 0950-2688. doi: 10.1017/S0950268820000990. URL [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7231667/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7231667/)
Pantelis R. Vlachas, Georgios Arampatzis, Caroline Uhler, and Petros Koumoutsakos. Learning the Effective Dynamics of Complex Multiscale Systems. *arXiv:2006.13431 [nlin, physics:physics]*, July 2020. URL [http://arxiv.org/abs/2006.13431](http://arxiv.org/abs/2006.13431) arXiv: 2006.13431.

Henning U. Voss, Jens Timmer, and Jürgen Kurths. Nonlinear dynamical system identification from uncertain and indirect measurements. *International Journal of Bifurcation and Chaos*, 14(06):1905–1933, June 2004. ISSN 0218-1274, 1793-6551. doi: 10.1142/S0218127404010345. URL [https://www.worldscientific.com/doi/abs/10.1142/S0218127404010345](https://www.worldscientific.com/doi/abs/10.1142/S0218127404010345).

Joseph H. M. Wedderburn. *Lectures on Matrices*. New York: American mathematical society, New York : Dover Publications, 1964.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, June 1989. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1989.1.2.270. URL [https://direct.mit.edu/neco/article/1/2/270-280/5490](https://direct.mit.edu/neco/article/1/2/270-280/5490).

Simon N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310), August 2010. ISSN 1476-4687. doi: 10.1038/nature09319. URL [https://www.nature.com/articles/nature09319](https://www.nature.com/articles/nature09319).
A Appendix

A.1 Theorems: Preliminaries

A.1.1 RNN derivatives

Considering the loss function $\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t$ of an RNN $F_\theta \in \mathcal{R}$ parameterized by $\theta$, we have

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t}{\partial \theta},$$

(18)

where

$$\frac{\partial \mathcal{L}_t}{\partial \theta} = \frac{\partial \mathcal{L}_t}{\partial z_t} \frac{\partial z_t}{\partial \theta}.$$  \hspace{1cm} (19)

The tangent vector $\frac{\partial z_T}{\partial \theta}$ has the form

$$\frac{\partial z_T}{\partial \theta} = \frac{\partial^+ z_T}{\partial \theta} + (T-2) \sum_{t=1}^{T-2} \left( \prod_{r=0}^{t-1} J_{T-r} \right) \frac{\partial^+ z_{T-t}}{\partial \theta},$$

(20)

where $\partial^+$ denotes the immediate partial derivative. Since for an RNN $F_\theta \in \mathcal{R}$ the activation function is element-wise, with $\theta$ the $m$-th element of a parameter vector $\theta$ (or belonging to the $m$-th row of a parameter matrix $\theta$), we have

$$\frac{\partial^+ z_T}{\partial \theta} = \begin{pmatrix} 0 & \cdots & 0 & \frac{\partial^+ z_{m,T}}{\partial \theta} & 0 & \cdots & 0 \end{pmatrix}^T.$$  \hspace{1cm} (21)

For instance, let $\theta = W$ be a weight matrix, then

$$\frac{\partial \mathcal{L}}{\partial W} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial w_{11}} & \frac{\partial \mathcal{L}}{\partial w_{12}} & \cdots & \frac{\partial \mathcal{L}}{\partial w_{1M}} \\ \frac{\partial \mathcal{L}}{\partial w_{21}} & \frac{\partial \mathcal{L}}{\partial w_{22}} & \cdots & \frac{\partial \mathcal{L}}{\partial w_{2M}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{L}}{\partial w_{M1}} & \frac{\partial \mathcal{L}}{\partial w_{M2}} & \cdots & \frac{\partial \mathcal{L}}{\partial w_{MM}} \end{pmatrix}.$$  \hspace{1cm} (22)

In this case, for the standard RNN we have

$$\frac{\partial^+ z_T}{\partial w_{mk}} = \begin{pmatrix} 0 & \cdots & 0 & z_{k,T-1} \xi_{mk}(z_{T-1}) & 0 & \cdots & 0 \end{pmatrix}^T = 1_{(m,k)} \xi_{mk}(z_{T-1}) z_{T-1},$$

(23)

where $\xi_{mk}(z_{T-1}) = f'_{w_{mk}} \left( \sum_{j=1}^{M} w_{mj} z_{j,T-1} + \sum_{j=1}^{M} b_{mj} s_{j,T} + h_m \right)$, and $f'_{w_{mk}}$ stands for the derivative of $f$ with respect to $w_{mk}$.

Therefore, for standard RNNs, (20) becomes

$$\frac{\partial z_T}{\partial w_{mk}} = 1_{(m,k)} \xi_{mk}(z_{T-1}) z_{T-1} + \sum_{t=1}^{T-2} \left( \prod_{r=0}^{t-1} J_{T-r} \right) 1_{(m,k)} \xi_{mk}(z_{T-t-1}) z_{T-t-1}.$$  \hspace{1cm} (24)

A.1.2 Piecewise-linear RNN (PLRNN)

The PLRNN has the generic form [Koppe et al., 2019; Schmidt et al., 2021]

$$z_t = F(z_{t-1}) = A z_{t-1} + W \phi(z_{t-1}) + C s_t + h + \epsilon_t,$$

(25)

where $\phi(z_{t-1}) = \max(z_{t-1}, 0)$ is the element-wise rectified linear unit (ReLU) function, $z_t \in \mathbb{R}^M$. 

17
The term $\partial$ where $s$strate the input, forget, cell, and output gates, $h$ is the bias vector, $s_t \in \mathbb{R}^K$ the external input weighted by $C \in \mathbb{R}^{M \times K}$, and $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$ a Gaussian noise term with diagonal covariance matrix $\Sigma$.

Equation (25) can be rewritten as

$$z_t = (A + WD_{\Omega(t-1)})z_{t-1} + Cs_t + h + \varepsilon_t =: W_{\Omega(t-1)}z_{t-1} + Cs_t + h + \varepsilon_t,$$

(26)

where $D_{\Omega(t)} := \text{diag}(d_{\Omega(t)})$ is a diagonal matrix with $d_{\Omega(t)} := (d_1, d_2, \cdots, d_M)$ an indicator vector such that $d_m(z_{m,t}) := d_m = 1$ whenever $z_{m,t} > 0$, and zeros otherwise.

For the PLRNN (26) we have

$$J_t = \frac{\partial z_t}{\partial z_{t-1}} = W_{\Omega(t-1)},$$

(27)

and $\|W_{\Omega(t-1)}\| \leq \|A\| + \|W\|.$

Furthermore, the derivatives (20) for the PLRNN (26) are

$$\frac{\partial z_T}{\partial w_{mk}} = 1_{(m,k)}D_{\Omega(T-1)}z_{T-1} + \sum_{j=2}^{T-1} \left( \prod_{i=1}^{j-1} W_{\Omega(T-i)} \right) 1_{(m,k)}D_{\Omega(T-j)}z_{T-j}.\tag{28}$$

A.1.3 Long Short-Term Memory (LSTM)

The LSTM is defined by the equations

$$\begin{align*}
i_t &= \sigma(W_{i_t}s_t + W_{hi}h_{t-1} + bid) \\
f_t &= \sigma(W_{f_t}s_t + W_{hf}h_{t-1} + bid) \\
g_t &= \tanh(W_{tg}s_t + W_{hg}h_{t-1} + b_g) \\
o_t &= \sigma(W_{o_t}s_t + W_{ho}h_{t-1} + b_o) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t) \tag{29}
\end{align*}$$

where $\{s_t\}$ is the input sequence, $W$ denotes weight matrices, $b$ bias terms, $i_t, f_t, g_t, o_t$ demonstrate the input, forget, cell, and output gates, $h_t$ and $c_t$ are the hidden and cell states at time $t$ respectively, $\sigma$ is the sigmoid activation function, and $\odot$ represents the element-wise (Hadamard) product (see Hochreiter & Schmidhuber [1997], Graves et al. [2016], Vlachas et al. [2018] for further information on LSTMs).

Defining $z_t := (h_t, c_t)^T$, the LSTM (29) can be represented as the first-order recursive map

$$z_t = F_\theta(z_{t-1}) = \left( o_t \odot \tanh(f_t \odot c_{t-1} + i_t \odot g_t) \\
f_t \odot c_{t-1} + i_t \odot g_t \right). \tag{30}$$

The term $\frac{\partial L_t}{\partial \theta}$ in (18) for some LSTM parameter $\theta$ can be written as

$$\frac{\partial L_t}{\partial \theta} = \sum_{r=1}^{t} \frac{\partial L_t}{\partial h_t} \frac{\partial h_t}{\partial z_t} \frac{\partial z_t}{\partial z_r} \frac{\partial z_r}{\partial \theta}.\tag{31}$$

A necessary condition for LSTMs to have a chaotic orbit is given by:
Proposition 1. Let the LSTM given by (29) have a chaotic attractor $\Gamma^*$ with $B_{\Gamma^*}$ as its basin of attraction. Then for every $z_1 = (h_1, c_1)^T \in B_{\Gamma^*}$,

$$\gamma := \lim_{T \to \infty} \frac{1}{T} \ln \left\| J_T J^*_T \cdots J_2 \right\| > 1.$$  \hspace{1cm} (32)

Proof. The Jacobian matrix of (30) for $t > 1$ can be written in the block form

$$\frac{\partial z_t}{\partial z_{t-1}} = J_t = \begin{pmatrix} \frac{\partial h_t}{\partial h_{t-1}} & \frac{\partial h_t}{\partial c_{t-1}} \\ \frac{\partial c_t}{\partial h_{t-1}} & \frac{\partial c_t}{\partial c_{t-1}} \end{pmatrix}.$$  \hspace{1cm} (33)

Further, due to the chain rule, we have

$$J_t J_{t-1} = \begin{pmatrix} \frac{\partial h_t}{\partial h_{t-1}} & \frac{\partial h_t}{\partial c_{t-1}} \\ \frac{\partial c_t}{\partial h_{t-1}} & \frac{\partial c_t}{\partial c_{t-1}} \end{pmatrix} \begin{pmatrix} \frac{\partial h_{t-1}}{\partial h_{t-2}} & \frac{\partial h_{t-1}}{\partial c_{t-2}} \\ \frac{\partial c_{t-1}}{\partial h_{t-2}} & \frac{\partial c_{t-1}}{\partial c_{t-2}} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial h_t}{\partial h_{t-2}} & \frac{\partial h_t}{\partial c_{t-2}} \\ \frac{\partial c_t}{\partial h_{t-2}} & \frac{\partial c_t}{\partial c_{t-2}} \end{pmatrix},$$  \hspace{1cm} (34)

and by induction we obtain

$$\frac{\partial z_t}{\partial z_1} = J_t J_{t-1} J_{t-2} \cdots J_2 = \begin{pmatrix} \frac{\partial h_t}{\partial h_1} & \frac{\partial h_t}{\partial c_1} \\ \frac{\partial c_t}{\partial h_1} & \frac{\partial c_t}{\partial c_1} \end{pmatrix}.$$  \hspace{1cm} (35)

Now assume that (30) has a chaotic orbit given by

$$\Gamma^* = \{ z_1^*, z_2^*, \ldots, z_T^*, \ldots \}.$$  \hspace{1cm} (36)

According to (35), the largest Lyapunov exponent of $\Gamma^*$ is given by

$$\lambda_{\Gamma^*} = \lim_{T \to \infty} \frac{1}{T} \ln \left\| J_T J^*_T \cdots J_2 \right\| = \lim_{T \to \infty} \frac{1}{T} \ln \left\| \begin{pmatrix} \frac{\partial h_t}{\partial h_1} & \frac{\partial h_t}{\partial c_1} \\ \frac{\partial c_t}{\partial h_1} & \frac{\partial c_t}{\partial c_1} \end{pmatrix} \right\|.$$  \hspace{1cm} (37)

Since $\Gamma^*$ is chaotic, so $\lambda_{\Gamma^*} > 0$, which gives

$$\lim_{T \to \infty} \frac{1}{T} \ln \left\| \begin{pmatrix} \frac{\partial h_t}{\partial h_1} & \frac{\partial h_t}{\partial c_1} \\ \frac{\partial c_t}{\partial h_1} & \frac{\partial c_t}{\partial c_1} \end{pmatrix} \right\| > 1.$$  \hspace{1cm} (38)

Based on Oseledec’s multiplicative ergodic Theorem, (37) holds for every $z_1 \in B_{\Gamma^*}$. This completes the proof. \hfill \Box

A.1.4 Gated Recurrent Unit (GRU)

A GRU network is defined by the equations

$$z_t = \sigma (W_z s_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma (W_r s_t + U_r h_{t-1} + b_r)$$

$$h_t = (1 - z_t) \odot \tanh (W_h s_t + U_h (r_t \odot h_{t-1}) + b_h) + z_t \odot h_{t-1},$$  \hspace{1cm} (38)

where $r_t$ represents the reset gate, $z_t$ the update gate, $s_t$ and $h_t$ denote the inputs and the hidden state respectively, $W_z, W_r, W_h \in \mathbb{R}^{M \times N}$ and $U_z, U_r, U_h \in \mathbb{R}^{M \times M}$ are weight matrices, $b_z, b_r, b_h \in \mathbb{R}^M$ are bias vectors, and $\sigma$ is the element-wise logistic sigmoid function (for more details about GRUs see [Cho et al. 2014a]).
A.1.5 Unitary evolution RNN (uRNN)

The uRNN, proposed in (Arjovsky et al., 2016), is defined as the nonlinear DS

\[ z_t = \sigma_b(W z_{t-1} + V s_t), \]  

(39)

for which \( W \in U(M) \) is an unitary matrix, \( V \in \mathbb{C}^{M \times N} \), \( b \in \mathbb{R}^M \) is the bias parameter, \( s_t \) is the real- or complex-valued input of dimension \( N \), and

\[
[\sigma_b(z)]_i = \begin{cases} 
(z_i + b_i) \frac{z_i}{|z_i|} & \text{if } |z_i| + b_i \geq 0 \\
0 & \text{if } |z_i| + b_i < 0
\end{cases}.
\]

(40)

**Proposition 2.** The uRNN given by (39) cannot have any chaotic orbit.

**Proof.** For any arbitrary orbit \( O_{z_1} \) of (39) we have

\[
\|J_T J_{T-1} \cdots J_2\| = \prod_{k=0}^{T-2} \|D_{T-k} W^T\|, \quad (41)
\]

where \( D_t = \text{diag} (\sigma'_b(W z_{t-1} + V s_t)) \). Since \( W \) is unitary and so a norm preserving matrix, it is concluded that

\[
\prod_{k=0}^{T-2} \|D_{T-k} W^T\| \leq \prod_{k=0}^{T-2} \|D_{T-k} W^T\| = \prod_{k=0}^{T-2} \|D_{T-k}\| = 1, \quad (42)
\]

which implies

\[
\lambda_{\text{max}} = \lim_{T \to \infty} \frac{1}{T} \ln \|J_T J_{T-1} \cdots J_2\| \leq 0. \quad (43)
\]

This rules out the existence of chaos (since \( \lambda_{\text{max}} > 0 \) is a necessary condition for \( O_{z_1} \) to be chaotic).

A.2 Theorems: Proofs

A.2.1 Proof of Theorem I, parts (ii) & (iii)

**Proof.** (ii) If \( J \) is the Jordan normal form of \( \prod_{s=0}^{k-1} J_{t^k-s} \), then \( \prod_{s=0}^{k-1} J_{t^k-s} = P J P^{-1} \), where

\[
J = \begin{pmatrix}
J_{m_1}(\lambda_1) & 0 & \cdots & 0 \\
0 & J_{m_2}(\lambda_2) & \cdots & 0 \\
0 & 0 & \ddots & \cdots \\
0 & \cdots & 0 & J_{m_p}(\lambda_p)
\end{pmatrix}, \quad (44)
\]

and \( m_i \) is the algebraic multiplicity of each eigenvalue \( \lambda_i \). Since \( \rho(\prod_{s=0}^{k-1} J_{t^k-s}) < 1 \), so the eigenvalue \( \lambda_i \) associated with each Jordan block satisfies \( |\lambda_i| < 1 \) (\( i = 1, \cdots, p \)). Moreover, every \( m_i \times m_i \) Jordan block has the form

\[
J_{m_i}(\lambda_i) = \begin{pmatrix}
\lambda_i & 0 & \cdots & 0 \\
0 & \lambda_i & \cdots & 0 \\
0 & 0 & \ddots & \cdots \\
0 & 0 & \cdots & \lambda_i
\end{pmatrix}.
\]

(45)
Accordingly
\[
\left\| \left( \prod_{s=0}^{k-1} J_{t^s-s} \right)^j \right\| = \| P \, J^j \, P^{-1} \| \leq p \| J^j \| , \tag{46}
\]
in which \( p = \| P \| \| P^{-1} \| \). Furthermore, for \( j \in \mathbb{N} \), \( J^j \) is a block diagonal matrix of the form
\[
J^j = \begin{pmatrix}
J_{m_1}^j (\lambda_1) & 0 & 0 & \cdots & 0 \\
0 & J_{m_2}^j (\lambda_2) & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \cdots & \vdots \\
0 & \cdots & 0 & J_{m_{p-1}}^j (\lambda_{p-1}) & 0 \\
0 & \cdots & \cdots & 0 & J_{m_p}^j (\lambda_p)
\end{pmatrix}, \tag{47}
\]
in which every \( m_i \times m_i \) Jordan block has the form
\[
J_{m_i}^j (\lambda_i) = \begin{pmatrix}
\lambda_i^j & (\frac{j}{1}) \lambda_i^{j-1} & (\frac{j}{2}) \lambda_i^{j-2} & \cdots & (\frac{j}{m_i-1}) \lambda_i^{j-m_i+1} \\
0 & \lambda_i^j & (\frac{j}{1}) \lambda_i^{j-1} & \cdots & (\frac{j}{m_i-2}) \lambda_i^{j-m_i+2} \\
\vdots & \vdots & \ddots & \cdots & \vdots \\
0 & 0 & \cdots & \lambda_i^j & (\frac{j}{1}) \lambda_i^{j-1} \\
0 & 0 & \cdots & 0 & \lambda_i^j
\end{pmatrix}. \tag{48}
\]

In addition, for every block \( J_{m_i}^j (\lambda_i) \), we have
\[
\| J_{m_i}^j (\lambda_i) \| \leq \sqrt{m_i} \| J_{m_i}^j (\lambda_i) \|_\infty = \sqrt{m_i} \sum_{q=1}^{m_i} \left| (J_{m_i}^j (\lambda_i))_{1q} \right|
\]
\[
= \sqrt{m_i} \sum_{q=1}^{m_i} \left( \frac{j}{q-1} \right) |\lambda_i|^{j-q+1} = |\lambda_i|^j \sqrt{m_i} \left( |\lambda_i|^{1-m_i} \sum_{q=1}^{m_i} \left( \frac{j}{q-1} \right) |\lambda_i|^{m_i-q} \right)
\]
\[
\leq |\lambda_i|^j \sqrt{m_i} \left( |\lambda_i|^{1-m_i} \sum_{q=1}^{m_i} |\lambda_i|^{m_i-q} \right) =: |\lambda_i|^j m_i N_{\lambda_i}. \tag{49}
\]

Moreover, for any \( 1 < \tilde{r}_i < \frac{1}{|\lambda_i|} \), there exists some \( l_i \) such that \( j m_i < \tilde{r}_i^j \) for \( j \geq l_i \). This means for \( j \geq l_i \)
\[
\| J_{m_i}^j (\lambda_i) \| \leq N_{\lambda_i} |\tilde{r}_i| \lambda_i|^{j}, \tag{50}
\]
such that \( |\tilde{r}_i \lambda_i| = \tilde{r}_i |\lambda_i| < 1 \).

Besides, for \( J^j = J_{m_1}^j (\lambda_1) \oplus J_{m_2}^j (\lambda_2) \oplus \cdots \oplus J_{m_p}^j (\lambda_p) \)
\[
\| J^j \| = \max_{1 \leq i \leq p} \| J_{m_i}^j (\lambda_i) \| =: \| J_{m}^j (\lambda) \|. \tag{51}
\]
Hence, from (46), (50) and (51), it is deduced that for \( j \geq l \)
\[
\left\| \left( \prod_{s=0}^{k-1} J_{t^s-s} \right)^j \right\| \leq p N_{\lambda} |\tilde{r} \lambda|^{j} =: \bar{p} r^j, \tag{52}
\]
in which \( r = |\tilde{r} \lambda| < 1 \).
Furthermore, let for $\Gamma_k$

\[
\max_{T \geq 1} \left\{ \| J_T^* \| \right\} = \max_{0 \leq s \leq k-1} \left\{ \| J_{t^*k-s} \| \right\} = \tilde{m},
\]

\[
\max_{T \geq 1} \left\{ \left\| \frac{\partial z_T}{\partial \theta} \right\| \right\} = \max_{0 \leq s \leq k-1} \left\{ \left\| \frac{\partial z_{t^*k-s}}{\partial \theta} \right\| \right\} = \xi,
\]

\[
\max_{T \geq 1} \left\{ \| z_T \| \right\} = \max_{0 \leq s \leq k-1} \left\{ \| z_{t^*k-s} \| \right\} = \bar{q}.
\]  \hspace{1cm} (53)

Hence, defining $z_0 = 0$, for this $k$-cycle

\[
\left\| \frac{\partial z_T}{\partial \theta} \right\| = \left\| \frac{\partial z_T}{\partial \theta} \right\| + \sum_{t=1}^{T-2} \left( \prod_{r=0}^{t-1} J_T^{*} \right) \frac{\partial z_{T-t}}{\partial \theta} \\
\leq \tilde{q} \xi \left( 1 + \sum_{t=1}^{T-1} \left\| \prod_{r=0}^{t-1} J_{T-r}^{*} \right\| \right).
\]  \hspace{1cm} (54)

On the other hand, for $T = kj$, from \ref{eq:52} and \ref{eq:53} we have

\[
\sum_{t=1}^{T-1} \left\| \prod_{r=0}^{t-1} J_{T-r}^{*} \right\| = \sum_{t=1}^{k-1} \left\| \prod_{r=0}^{t-1} J_{kt-r}^{*} \right\| = \sum_{t=1}^{k-1} \left\| \prod_{r=0}^{t-1} J_{kt-r}^{*} \right\| + \sum_{t=k}^{2k-1} \left\| \prod_{r=0}^{t-1} J_{kt-r}^{*} \right\| \\
+ \sum_{t=2k}^{3k-1} \left\| \prod_{r=0}^{t-1} J_{kt-r}^{*} \right\| + \cdots + \sum_{t=(j-2)k}^{(j-1)k-1} \left\| \prod_{r=0}^{t-1} J_{kt-r}^{*} \right\| \\
+ \sum_{t=(j-1)k}^{jk-1} \left\| \prod_{r=0}^{t-1} J_{kt-r}^{*} \right\| \\
= \sum_{t=1}^{k-1} \left\| \prod_{r=0}^{t-1} J_{kt-r}^{*} \right\| + \sum_{t=2}^{j} \sum_{t=(t-1)k}^{tk-1} \left\| \prod_{r=0}^{t-1} J_{kt-r}^{*} \right\| \\
\leq (\tilde{m} + \tilde{m}^2 + \cdots + \tilde{m}^{k-1}) + \sum_{t=2}^{j} p \left( 1 + \tilde{m} + \tilde{m}^2 + \cdots + \tilde{m}^{k-1} \right) r^{t-1}.
\]  \hspace{1cm} (55)

Thus, considering $(\tilde{m} + \tilde{m}^2 + \cdots + \tilde{m}^{k-1}) = \mathcal{M}$, it is deduced that

\[
\lim_{T \rightarrow \infty} \left\| \frac{\partial z_T}{\partial \theta} \right\| = \lim_{j \rightarrow \infty} \left\| \frac{\partial z_j}{\partial \theta} \right\| \leq \tilde{q} \xi (1 + \mathcal{M} + \frac{\bar{p} r (1 + \mathcal{M})}{1 - r}) = \tilde{M} < \infty,
\]  \hspace{1cm} (56)

which, by \ref{eq:49}, implies $\frac{\partial z_T}{\partial \theta}$ will be bounded for $T \rightarrow \infty$.

(iii) Consider the PLRNN given by \ref{eq:25}, where for simplicity we ignore the external inputs and noise terms. Let $\{ z_{t_1}, z_{t_2}, z_{t_3}, \ldots \}$ be an orbit which converges to $\Gamma_k$. Hence

\[
\lim_{n \rightarrow \infty} d(z_{t_n}, \Gamma_k) = 0,
\]  \hspace{1cm} (57)

which implies there exists a neighborhood $U$ of $\Gamma_k$ and $k$ sub-sequences $\{ z_{tk_m} \}_{m=1}^{\infty}, \{ z_{tk_{m+1}} \}_{m=1}^{\infty}, \ldots, \{ z_{tk_{k+1}} \}_{m=1}^{\infty}$ of the sequence $\{ z_{tn} \}_{n=1}^{\infty}$ such that all these sub-sequences belong to $U$ and

a) $z_{tk_{m+s}} = F^k(z_{tk_{(m-1)+s}}), s = 0, 1, 2, \ldots, k - 1,$

b) $z_{tk_{m+1}} = F^{k+1}(z_{tk_{m}}), m = 0, 1, 2, \ldots, k - 2,

\frac{\partial z}{\partial \theta}$ will be bounded for $T \rightarrow \infty$.
b) \( \lim_{n \to \infty} z_{tk^{m+s}} = z_{t^s k^{-s}}, s = 0, 1, 2, \cdots, k - 1, \)

c) for every \( z_{t_n} \in U \) there is some \( s \in \{0, 1, 2, \cdots, k - 1\} \) such that \( z_{t_n} \in \{z_{tk^{m+s}}\}_{m=1}^\infty. \)

In this case, for every \( z_{t_n} \in U \) with \( z_{t_n} \in \{z_{tk^{m+s}}\}_{m=1}^\infty \), there exists some \( \tilde{n} \in \mathbb{N} \) such that \( z_{t_n} = z_{tk^{\tilde{n}+s}} \) and \( \lim_{\tilde{n} \to \infty} z_{tk^{\tilde{n}+s}} = z_{t^s k^{-s}}. \) Therefore, continuity of \( F \) results in

\[
\lim_{\tilde{n} \to \infty} F(z_{tk^{\tilde{n}+s}}) = F(z_{t^s k^{-s}}),
\]

and so by (58)

\[
\lim_{n \to \infty} W_{\Omega(t_{\tilde{k}n+s})} z_{tk^{\tilde{n}+s}} = W_{\Omega(t^s k^{-s})} z_{t^s k^{-s}}.
\]

Assuming \( \lim_{n \to \infty} W_{\Omega(t_{\tilde{k}n+s})} = L, \) since (60) holds for every \( z_{t^s k^{-s}} \), substituting \( z_{t^s k^{-s}} = e_1^T = (1, 0, \cdots, 0)^T \) in (60), we can prove that the first column of \( L \) equals the first column of \( W_{\Omega(t^s k^{-s})}. \)

Performing the same procedure for \( z_{t^s k^{-s}} = e_i^T, i = 2, 3, \cdots, M, \) yields

\[
\lim_{n \to \infty} W_{\Omega(t_{\tilde{k}n+s})} = W_{\Omega(t^s k^{-s})}.
\]

According to (57), \( U \) contains an infinite number of terms of the sequence \( \{z_{t_n}\}_{n=1}^\infty, \) i.e.

\[
\exists N \in \mathbb{N} \text{ s.t. } n \geq N \implies z_{t_n} \in U.
\]

Suppose that \( z_{t_n} \in U \) for some \( n \geq N. \) Thus, there exists some \( s \in \{0, 1, 2, \cdots, k - 1\} \) such that \( z_{t_n} \in \{z_{tk^{m+s}}\}_{m=1}^\infty. \) Without loss of generality let \( s = 0. \) Hence, there is some \( \tilde{n} \in \mathbb{N} \) such that \( z_{t_n} = z_{tk^{\tilde{n}}} \) and \( \lim_{\tilde{n} \to \infty} z_{tk^{\tilde{n}}} = z_{t^s k^{-s}}. \) In this case, moving forward in time gives

\[
\begin{align*}
z_{t_n} &= z_{tk^{\tilde{n}}} \left( z_{t_{\tilde{n}}} \in \{z_{tk^{m}}\}_{m=1}^\infty \right), \\
z_{t_{n+1}} &= z_{tk^{\tilde{n}+1}} \left( z_{t_{n+1}} \in \{z_{tk^{m+1}}\}_{m=1}^\infty \right), \\
z_{t_{n+2}} &= z_{tk^{\tilde{n}+2}} \left( z_{t_{n+2}} \in \{z_{tk^{m+2}}\}_{m=1}^\infty \right), \\
&\quad \vdots \\
z_{t_{n+k-1}} &= z_{tk^{\tilde{n}+k-1}} \left( z_{t_{n+k-1}} \in \{z_{tk^{m+k-1}}\}_{m=1}^\infty \right), \\
z_{t_{n+k}} &= z_{tk^{\tilde{n}+k}} \left( z_{t_{n+k}} \in \{z_{tk^{m+k}}\}_{m=1}^\infty \right), \\
z_{t_{n+k+1}} &= z_{tk^{\tilde{n}+k+1}} \left( z_{t_{n+k+1}} \in \{z_{tk^{m+k+1}}\}_{m=1}^\infty \right), \\
&\quad \vdots \\
z_{t_{n+2k-1}} &= z_{tk^{\tilde{n}+2k-1}} \left( z_{t_{n+2k-1}} \in \{z_{tk^{m+2k-1}}\}_{m=1}^\infty \right), \\
z_{t_{n+2k}} &= z_{tk^{\tilde{n}+2k}} \left( z_{t_{n+2k}} \in \{z_{tk^{m+2k}}\}_{m=1}^\infty \right), \\
&\quad \vdots
\end{align*}
\]
Consequently, for \( n \geq N \) and \( j \in \mathbb{N} \), we can write
\[
\prod_{i=0}^{k-1} W_{\Omega(t_{n+kj-i})} = \left( \prod_{i=1}^{k} W_{\Omega(t_{k(i+j)+k-1})} \right) \prod_{i=1}^{k} W_{\Omega(t_{k(i+j-1)+k-1})} = \prod_{i=0}^{j} \prod_{l=1}^{k} W_{\Omega(t_{k(i+j)+k-1})}.
\]

(64)

On the other hand, in equation (26), there are different configurations for matrix \( D_{\Omega(t-1)} \) and hence different forms for matrix \( W_{\Omega(t_{k+h+s})} \) in this case. In this case, the phase space of the system is divided into different sub-regions by some borders; see [Monfared & Durstewitz, 2020a,b] for more details. Also, since the system (26) is a linear map in each sub-region, the \( k \) periodic points of \( \Gamma_{k} \) must belong to different sub-regions (at least two different sub-regions). Accordingly, based on (61) and (63), there exists some \( N \in \mathbb{N} \) such that for every \( \hat{n} \geq N \) both \( z_{t_{k+h+s}} \) and \( z_{t_{k+s}} \) belonging to the \( s \)-th sub-region and so the matrices \( W_{\Omega(t_{k+h+s})} \) and \( W_{\Omega(t_{k+s})} \) (\( s \in \{0, 1, 2, \ldots, k-1\} \)) are identical. Hence, for \( n \geq N \), \( \hat{n} \geq N \) and \( j \in \mathbb{N} \), equation (64) becomes
\[
k-1
\prod_{i=0}^{k-1} W_{\Omega(t_{n+kj-i})} = \prod_{i=0}^{j} \prod_{l=1}^{k} W_{\Omega(t_{k(i+j)+k-1})} = \left( \prod_{s=0}^{k-1} W_{\Omega(t_{k+s})} \right)^{j}.
\]

(65)

Therefore, similar to the part (ii), we can prove for every \( z_{1} \in \mathcal{B}_{T_{k}} \), \( \frac{\partial z_{1}}{\partial \theta} \) and \( \frac{\partial C_{r}}{\partial \theta} \) will also remain bounded.

\[\square\]

A.2.2 PROOF OF THEOREM [2] PART (II)

**Proof.** (ii) Let for every \( T > 2 \)
\[
L_{T} := J_{T}^{*} J_{T-1}^{*} \ldots J_{2}^{*}.
\]

(66)

\( \{L_{T}\}_{T \in \mathbb{N}, T > 2} \) is a sequence of matrices \( L_{T} = [j_{ij}(T)]_{1 \leq i,j \leq M} \) and, due to (63), \( \lim_{T \to \infty} \|L_{T}\| = \infty \). Hence, there is at least one sub-sequence \( \{j_{mk}(T_{n})\}_{T_{n} \in \mathbb{N}, T_{n} > 2} \) (for some \( m, k \in \{1, 2, \ldots, M\} \)) such that \( \lim_{T_{n} \to \infty} j_{mk}(T_{n}) = \infty \).

On the other hand
\[
\frac{\partial z_{T}^{*}}{\partial \theta} = \frac{\partial^{+} z_{T}^{*}}{\partial \theta} + \sum_{l=1}^{T-2} \left( \prod_{r=0}^{l-1} J_{r}^{*} \right) \frac{\partial^{+} z_{T-r}^{*}}{\partial \theta}.
\]

(67)

Moreover, there exists some \( N > 2 \) such that (for \( t = T - N + 1 \))
\[
\frac{\partial^{+} z_{T}^{*}}{\partial \theta} \neq 0.
\]

(68)

For \( \theta \) as the \( k \)-th element of a parameter vector \( \theta \) (or belonging to the \( k \)-th row of a parameter matrix \( \Theta \)), the term
\[
\left( \prod_{r=0}^{T-N} J_{r}^{*} \right) \frac{\partial^{+} z_{T-N-1}^{*}}{\partial \theta}
\]

(69)

is a vector in which the \( i \)-th element is \( j_{ik}(T) \frac{\partial^{+} z_{T-N-1}^{*}}{\partial \theta} \).

Since \( \lim_{T_{n} \to \infty} j_{mk}(T_{n}) = \infty \), due to (68) \( \lim_{T_{n} \to \infty} j_{mk}(T_{n}) \frac{\partial^{+} z_{T-N-1}^{*}}{\partial \theta} = \infty \), which implies \( \frac{\partial z_{T}^{*}}{\partial \theta} \) will diverge as \( T \to \infty \). Similarly, by (19), we can prove \( \frac{\partial C_{r}}{\partial \theta} \) is divergent for \( T \to \infty \).

By Oseledec’s multiplicative ergodic Theorem, the results also hold for every \( z_{1} \in \mathcal{B}_{T_{k}} \).
**Proof.** Let \( \Gamma = \{ z_1, z_2, \ldots, z_T, \cdots \} \) be a quasi-periodic attractor. Then, the largest Lyapunov exponent of \( \Gamma \) is

\[
\lambda = \lim_{T \to \infty} \frac{1}{T} \ln \left\| J_T J_{T-1} \cdots J_2 \right\| = \lim_{T \to \infty} \frac{1}{T} \ln \left\| \frac{\partial z_T}{\partial z_1} \right\| = 0. \tag{70}
\]

We prove for every \( 0 < \epsilon < 1 \)

\[
\lim_{T \to \infty} (1 - \epsilon)^T - 1 < \lim_{T \to \infty} \left\| \frac{\partial z_T}{\partial z_1} \right\| < \lim_{T \to \infty} (1 + \epsilon)^T - 1. \tag{71}
\]

For this purpose, we show \( \forall 0 < \epsilon < 1 \)

(I) \( \lim_{T \to \infty} (1 - \epsilon)^T - 1 < \lim_{T \to \infty} \left\| \frac{\partial z_T}{\partial z_1} \right\|, \) and

(II) \( \lim_{T \to \infty} \left\| \frac{\partial z_T}{\partial z_1} \right\| < \lim_{T \to \infty} (1 + \epsilon)^T - 1. \)

Assume for the sake of contradiction that (I) does not hold. Then there exists some \( 0 < \epsilon < 1 \) such that

\[
\lim_{T \to \infty} (1 - \epsilon)^T - 1 \geq \lim_{T \to \infty} \left\| \frac{\partial z_T}{\partial z_1} \right\|. \tag{72}
\]

Therefore

\[
\exists T_0 > 1 \text{ s.t. } \forall T \geq T_0 \implies (1 - \epsilon)^T - 1 \geq \left\| \frac{\partial z_T}{\partial z_1} \right\|. \tag{73}
\]

and so

\[
\exists T_0 > 1 \text{ s.t. } \forall T \geq T_0 \implies \ln(1 - \epsilon)^T - 1 \geq \frac{\ln \left\| \frac{\partial z_T}{\partial z_1} \right\|}{T - 1}. \tag{74}
\]

Consequently, due to (70), for \( T \to \infty \) we have \( \ln(1 - \epsilon) \geq 0. \) This implies \( \epsilon \leq 0, \) which is a contradiction.

Similarly if we assume (II) is not true, then there exists some \( 0 < \epsilon < 1 \) such that

\[
\lim_{T \to \infty} \left\| \frac{\partial z_T}{\partial z_1} \right\| \geq \lim_{T \to \infty} (1 + \epsilon)^T - 1. \tag{75}
\]

Thereby

\[
\exists T_0 > 1 \text{ s.t. } \forall T \geq T_0 \implies \left\| \frac{\partial z_T}{\partial z_1} \right\| \geq (1 + \epsilon)^T - 1, \tag{76}
\]

and thus

\[
\exists T_0 > 1 \text{ s.t. } \forall T \geq T_0 \implies \frac{\ln \left\| \frac{\partial z_T}{\partial z_1} \right\|}{T - 1} \geq \frac{\ln(1 + \epsilon)^T - 1}{T - 1}. \tag{77}
\]

This means \( \ln(1 + \epsilon) \leq 0 \) as \( T \to \infty, \) i.e. \( \epsilon \leq 0, \) which is a contradiction.

Therefore (14) holds for \( \Gamma \) and also, according to Oseledec’s multiplicative ergodic Theorem, for every \( z_1 \) in the basin of attraction of \( \Gamma. \) \( \square \)
A.2.4 PROOF OF THEOREM

**Proof.** Let \( \| \cdot \| \) be any matrix norm satisfying \( \| A_1 A_2 \| \leq \| A_1 \| \| A_2 \| \).

(i) By boundedness of \( \frac{\partial^+ \varepsilon_T}{\partial \theta} \) we have

\[
\left\| \frac{\partial^+ \varepsilon_T}{\partial \theta} \right\| = \left\| \frac{\partial^+ \varepsilon_T}{\partial \theta} + \sum_{t=1}^{T-2} \left( \prod_{r=0}^{t-1} J_{T-r} \right) \frac{\partial^+ \varepsilon_{T-t}}{\partial \theta} \right\|
\leq \xi \left( 1 + \sum_{t=1}^{T-2} \prod_{r=0}^{t-1} \| J_{T-r} \| \right) \leq \xi \left( 1 + \sum_{t=1}^{T-2} \prod_{r=0}^{t-1} \| J_{T-r} \| \right).
\]

Moreover,

\[
\sum_{t=1}^{T-2} \prod_{r=0}^{t-1} \| J_{T-r} \| \leq 1 + \sum_p \| J_p \| + \sum_{p<q} \| J_p \| \| J_q \| + \cdots = (1 + \| J_1 \|) (1 + \| J_{T-1} \|) \cdots (1 + \| J_2 \|) =: \prod_{t=2}^{T} (1 + \| J_i \|).
\]

Since \( \sum_{t=2}^{\infty} \| J_T \| \) converges, according to Wedderburn (1964), the infinite products \( \prod_{t=2}^{\infty} (1 + \| J_T \|) \) in (79) converge to a finite number \( \tilde{K} \neq 0 \). Consequently, by (78) and (79)

\[
\lim_{T \to \infty} \left\| \frac{\partial^+ \varepsilon_T}{\partial \theta} \right\| \leq \tilde{K} < \infty,
\]

which implies \( \frac{\partial^+ \varepsilon_T}{\partial \theta} \) will be bounded for \( T \to \infty \).

Furthermore

\[
\lim_{T \to \infty} \left\| \frac{\partial^+ \varepsilon_T}{\partial \varepsilon_1} \right\| \leq \lim_{T \to \infty} \prod_{t=2}^{\infty} \| J_T \| := \lim_{T \to \infty} \left( \| J_T \| \| J_{T-1} \| \cdots \| J_2 \| \right) \leq \prod_{t=2}^{\infty} (1 + \| J_T \|) \leq \tilde{K},
\]

which completes the proof.

(ii) Since \( \sum_{T=1}^{\infty} \| P_T \| < \infty \), due to Wedderburn (1964) the infinite product

\[
\prod_{t=2}^{\infty} (I + P_T) = \lim_{T \to \infty} J_T =: J_T = J_T J_{T-1} \cdots J_2,
\]

converges to a matrix \( K \neq O \), which implies

\[
0 < \lim_{T \to \infty} \left\| \frac{\partial^+ \varepsilon_T}{\partial \varepsilon_1} \right\| = \| K \| < \infty.
\]

\( \square \)

A.3 ADDITIONAL MATHEMATICAL RESULTS

A.3.1 FURTHER RESULTS AND REMARKS RELATED TO THEOREM

Remark 3. The result of Theorem also holds for unstable orbits \( \{ z_1, z_2, z_3, \cdots \} \) with positive largest Lyapunov exponent. Trivially, for such orbits that diverge to infinity (unbounded latent states) gradients of the loss function will explode as \( T \to \infty \).

Remark 4. For RNNs with ReLU activation functions there are finite compartments in the phase space each with a different functional form. In such a case, to define the largest Lyapunov exponent of \( \Gamma^* \), in the proof of Theorem we assume that \( \Gamma^* \) never maps to the points of the borders.
Based on Theorem 2, we can also formulate the necessary conditions for chaos and diverging gradients in standard RNNs with particular activation functions by considering the norms of their recurrence matrix, for which the following Corollary provides the basis:

**Corollary 1.** Let for a standard RNN

\[
\left\| diag(f'(Wz_{t-1} + Bs_t + h)) \right\| \leq \gamma < \infty.
\]  

(84)

If the RNN is chaotic, then \( \| W \| \gamma > 1 \).

**Proof.** Assume for the sake of contradiction that \( \| W \| \gamma \leq 1 \). From

\[
\left\| \prod_{2 \leq t \leq T} W diag(f'(Wz_{t-1} + Bs_t + h)) \right\| \leq \left\| \prod_{2 \leq t \leq T} W diag(f'(Wz_{t-1} + Bs_t + h)) \right\|
\]

\[
\leq (\| W \| \gamma)^{T-2};
\]  

(85)

it is concluded that \( \lim_{T \to \infty} \left\| \prod_{2 \leq t \leq T} W diag(f'(Wz_{t-1} + Bs_t + h)) \right\| < \infty \), which contradicts (13). This means \( \| W \| \gamma > 1 \) is a necessary condition for the standard RNN to be chaotic.

**Remark 5.** For RNN with the tanh and sigmoid activation functions \( \gamma = 1 \) and \( \gamma = \frac{1}{4} \), respectively. Thus, by Corollary 2 the necessary conditions for chaos in these two cases are \( \| W \| > 1 \) and \( \| W \| > 4 \), respectively.

A.3.2 Further results related to Section 3.5

**Proposition 3.** Let \( O_{z_1} = \{ z_1, z_2, \ldots, z_T, \ldots \} \) be an orbit generated by an RNN \( F_\theta \in \mathcal{R} \) (parameterized by \( \theta \)), and \( \| J_T \| \neq 0 \), \( T \geq 2 \). If \( \sum_{T=2}^{\infty} \ln \| J_T \| \) diverges to \( -\infty \), then the Jacobian \( \frac{\partial z_T}{\partial z_1} \) vanishes as \( T \) tends to infinity.

**Proof.** For \( \| J_T \| \neq 0 \), \( T \geq 2 \), we have

\[
0 \leq \left\| \frac{\partial z_T}{\partial z_1} \right\| \leq \| J_T \| \| J_{T-1} \| \cdots \| J_2 \| = e^{\ln \| J_T \|} e^{\ln \| J_{T-1} \|} \cdots e^{\ln \| J_2 \|} = e^{\sum_{T=2}^{\infty} \ln \| J_T \|}.
\]

(86)

Hence if \( \sum_{T=2}^{\infty} \ln \| J_T \| \to -\infty \), then

\[
\lim_{T \to \infty} \frac{\partial z_T}{\partial z_1} = 0.
\]  

(87)

A.4 Empirical evaluation: Datasets

**Lorenz attractor** The Lorenz system ([Lorenz, 1963]) is a simplified model for atmospheric convection, given by

\[
\begin{align*}
\frac{dx}{dt} &= \sigma(y-x), \\
\frac{dy}{dt} &= x(\rho-z) - y, \\
\frac{dz}{dt} &= xy - \beta z.
\end{align*}
\]  

(88)

The system is of particular interest for its chaotic regime and was studied here for \( \sigma = 16 \), \( \rho = 45.92 \) and \( \beta = 4 \). For these parameters the Lorenz system is known to have a maximal Lyapunov exponent \( \lambda_{\text{max}} = 1.5 \) ([Rosenstein et al., 1993]). To generate a time series, the ODEs were integrated with a step size \( \Delta t = 0.01 \) using `scipy.integrate`. Accordingly, the prediction time is \( \tau_{\text{pred}} = \frac{\ln(2)}{\Delta t \lambda_{\text{max}}} = 46.2 \).
Rössler system Another prime textbook example for a chaotic system is the Rössler system (Rössler 1976) given by:

\[
\begin{align*}
\frac{dx}{dt} &= -y - z, \\
\frac{dy}{dt} &= x + ay, \\
\frac{dz}{dt} &= b + z(x - c).
\end{align*}
\]  

(89)

For the parameters \(a = 0.15\), \(b = 0.2\) and \(c = 10\), the maximal Lyapunov exponent is \(\lambda_{\text{max}} = 0.09\) (Rosenstein et al. 1993). To arrive at a time series, a step size of \(\Delta t = 0.1\) was chosen for integration. This gives us a prediction time of \(\tau_{\text{pred}} = 77.0\) for this system.

Empirical temperature time series This time series was recorded at the Weather Station at the Max Planck Institute for Biogeochemistry in Jena, Germany, spanning the time period between 2009 and 2016, and reassembled by François Chollet for the book Deep Learning with Python. The dataset can be accessed at https://www.kaggle.com/pankrzysiu/weather-archive-jena. To expose the underlying chaotic dynamics of the time series, trends and yearly cycles were removed, and nonlinear noise-reduction was performed (using ghkss from TISEAN, see also Kantz et al. 1993). Fig. 4 (a) shows a snippet of the temperature data in comparison with the de-noised time-series. High-frequency noise was further reduced through Gaussian kernel smoothing (\(\sigma = 200\)), and the resulting time series was sub-sampled (every 5th data point was retained). Fig. 4 (b) clearly reveals a fractional dimension of \(D_{\text{eff}} = 2.8\) for the de-noised and smoothed time-series. This strongly suggests that the dynamics governing the time series are chaotic. We created a time delay embedding (Kantz & Schreiber 2003) with \(m = 5\) (estimated by the false nearest neighbor technique, see Kennel et al. 1992) and delay \(\tau = 500\) (obtained as the first minimum of the mutual information). The first three embedding dimensions are shown in Fig. 4 (c). The maximal Lyapunov exponent of this time series was determined with lyap \(r\) from TISEAN (Hegger et al. 1999) to be \(\lambda_{\text{max}} = 0.016\), see Fig. 3 (a). This value is in close agreement with the literature (Millán et al. 2010). The predictability time of this system is estimated to be \(\tau_{\text{pred}} = 43.3\).

![Figure 4](https://example.com/figure4.png)

Figure 4: (a) Snippet of the original temperature data and de-noised time series. (b) Blue lines show the local slopes of the correlation sums for embedding dimensions \(m \in \{5, \ldots, 10\}\). The convergence of these estimates in \(m\) reveals a fractional dimension indicated by the plateau. (c) First three dimensions of the time-delay embedding series as used for training.

All datasets used were standardized (i.e., centered with unit variance) prior to training.

A.5 Empirical Evaluation: Measures of Reconstruction Quality

Attractor overlap To assess the geometrical similarity of the chaotic attractor produced by the RNN to the one underlying the observations, we calculate the Kullback-Leibler divergence of the ground truth distribution \(p_{\text{true}}(x)\) and the distribution \(p_{\text{gen}}(x | z)\) generated by RNN simulation. To do so in practice, we employ a binning approximation (see Koppel et al. 2019)

\[
D_{\text{step}}(p_{\text{true}}(x), p_{\text{gen}}(x | z)) \approx \sum_{k=1}^{K} \tilde{p}_{\text{true}}^{(k)}(x) \log \left( \frac{p_{\text{true}}^{(k)}(x)}{\tilde{p}_{\text{gen}}^{(k)}(x | z)} \right).
\]
where $K$ is the total number of bins, and $\hat{p}_{\text{true}}^{(k)}(x)$ and $\hat{p}_{\text{gen}}^{(k)}(x \mid z)$ are estimates obtained as relative frequencies through sampling trajectories from the observed time-series and the trained RNN, respectively.

**Power-spectral correlations** Since in DS reconstruction we aim to capture invariant (time-independent) properties of the underlying system, besides the geometrical agreement, we compare the similarity in true and RNN-reconstructed power spectra. To do so, we generate a time series of length 100,000 from the RNN and calculate its power spectrum using the fast Fourier transform (scipy.fft). To reduce the influence of noise we apply Gaussian kernel smoothing and cut off the long high-frequency tails of the spectra. The dimension-wise correlation between observed and generated spectra are then averaged to give the $PSC$.

### A.6 Reconstruction examples: Rössler system

![Rössler attractor and LSTM reconstruction](image)

Figure 5: The Rössler attractor (blue) and reconstruction by a LSTM (orange) trained with a learning interval (a) chosen too small ($\tau = 5$), (b) chosen optimally ($\tau = 30$), and (c) chosen too large ($\tau = 200$).

### A.7 Sparsely forced BPTT

**Loss truncation** One implicit consequence of the teacher forcing, eqn. (16), is the interruption of the hidden-to-hidden connections at these time points. More specifically, if the system is forced at time $t \in T$, then there is no connection between $z_t$ and $z_{t+1}$, that is

$$J_{t+1} = \frac{\partial z_{t+1}}{\partial z_t} = \frac{\partial \text{RNN}(\hat{z}_t)}{\partial z_t} = 0.$$  \hfill (90)

To see how these vanishing Jacobians truncate the loss gradients w.r.t to some parameter $\theta$, let us focus on the loss gradients immediately after the forcing,

$$\frac{\partial L_{t+1}}{\partial \theta} = \frac{\partial L_{t+1}}{\partial z_{t+1}} \sum_{k=1}^{t+1} \frac{\partial z_{t+1}}{\partial z_k} \frac{\partial z_k}{\partial \theta}$$

$$= \frac{\partial L_{t+1}}{\partial z_{t+1}} \left( \frac{\partial^+ z_{t+1}}{\partial \theta} + \sum_{k=1}^{t} \frac{\partial z_{t+1}}{\partial z_k} \frac{\partial^+ z_k}{\partial \theta} \right) = 0,$$

because of (90)

$$= \frac{\partial L_{t+1}}{\partial z_{t+1}} \frac{\partial^+ z_{t+1}}{\partial \theta}. \hfill (91)$$

Eqn. (91) shows that sparsely forced BPTT implicitly truncates the loss gradients because it interrupts the hidden-to-hidden connection from $z_t$ to $z_{t+1}$ for $t \in T$. More generally, defining $\hat{t} := \max\{t' \in T : t' \leq t\}$, the overall loss gradients are truncated to

$$\frac{\partial L}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial z_t} \sum_{k=1}^{t} \frac{\partial z_t}{\partial z_k} \frac{\partial^+ z_k}{\partial \theta}$$

$$\text{tr} := \sum_{t=1}^{T} \frac{\partial L_t}{\partial z_t} \sum_{k=t}^{\hat{t}} \frac{\partial z_t}{\partial z_k} \frac{\partial^+ z_k}{\partial \theta}.$$  \hfill (92)