ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Long Expressive Memory for Sequence Modeling

T. K. Rusch and S. Mishra and N. B. Erichson and M. W. Mahoney

Research Report No. 2021-33
October 2021

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Funding ERC: 770880 COMANFLO

Long Expressive Memory for Sequence Modeling

T. Konstantin Rusch * Siddhartha Mishra * N. Benjamin Erichson †
Michael W. Mahoney ‡

Abstract

We propose a novel method called *Long Expressive Memory* (LEM) for learning long-term sequential dependencies. LEM is gradient-based, it can efficiently process sequential tasks with very long-term dependencies, and it is sufficiently expressive to be able to learn complicated input-output maps. To derive LEM, we consider a system of *multiscale ordinary differential equations*, as well as a *suitable time-discretization* of this system. For LEM, we derive rigorous bounds to show the mitigation of the exploding and vanishing gradients problem, a well-known challenge for gradient-based recurrent sequential learning methods. We also prove that LEM can approximate a large class of dynamical systems to high accuracy. Our empirical results, ranging from image and time-series classification through dynamical systems prediction to speech recognition and language modeling, demonstrate that LEM outperforms state-of-the-art recurrent neural networks, gated recurrent units, and long short-term memory models.

1 Introduction

Learning tasks with sequential data as inputs (and possibly outputs) arise in a wide variety of contexts, including computer vision, text and speech recognition, natural language processing, and time series analysis in the sciences and engineering. While recurrent gradient-based models have been successfully used in processing sequential data sets, it is well-known that training these models to process (very) long sequential inputs is extremely challenging on account of the so-called *exploding and vanishing gradients problem* [32]. This arises as calculating hidden state gradients entails the computation of an iterative product of gradients over a large number of steps. Consequently, this (long) product can easily grow or decay exponentially in the number of recurrent interactions.

Mitigation of the exploding and vanishing gradients problem has received considerable attention in the literature. A classical approach, used in Long Short-Term Memory (LSTM) [18] and Gated Recurrent Units (GRUs) [11], relies on *gating mechanisms* and leverages the resulting additive structure to ensure that gradients do not vanish. However, gradients might still explode, and learning very long-term dependencies remains a challenge for these architectures [27]. An alternative approach imposes constraints on the structure of the hidden weight matrices of the underlying recurrent neural networks (RNNs), for instance by requiring these matrices to be unitary or orthogonal [16, 1, 40, 21]. However, constraining the structure of these matrices might lead to significantly reduced expressivity, i.e., the ability of the model to learn complicated input-output maps. Yet another approach relies on enforcing the hidden weights to lie within pre-specified bounds, leading to control on gradient norms. Examples include Li et al. [27], based on *independent neurons* in each layer, and Rusch and Mishra [36], based on a network of coupled oscillators. Imposing such restrictions on weights might be difficult to enforce, and weight clipping could reduce expressivity significantly.

This brief survey highlights the challenge of *designing recurrent gradient-based methods for sequence modeling which can mitigate the exploding and vanishing gradients problem, while at the same time being sufficiently expressive and possessing the ability to learn complicated input-output maps efficiently.* We seek to address this challenge by proposing a novel gradient-based method.

*Seminar for Applied Mathematics (SAM), D-MATH, ETH Zürich, Switzerland
†Department of Mechanical Engineering and Materials Science, University of Pittsburgh, USA
‡ICSI and Department of Statistics, University of California, Berkeley, USA

The starting point for our method is the observation that realistic sequential data sets often contain information arranged according to multiple (time, length, etc., depending on the data and task) scales. Indeed, if there were only one or two scales over which information correlated, then a simple model with a parameter chosen to correspond to that scale (or, e.g., scale difference) should be able to model the data well. Thus, it is reasonable to expect that a *multiscale model* should be considered to process efficiently such *multiscale data*. To this end, we propose a novel gradient-based architecture, *Long Expressive Memory* (LEM), that is based on a suitable time-discretization of a set of multiscale ordinary differential equations (ODEs). For this novel gradient-based method (proposed in Section 2):

- we derive bounds on the hidden state gradients to prove that LEM mitigates the exploding and vanishing gradients problem (Section 4);
- we rigorously prove that LEM can approximate a very large class of (multiscale) dynamical systems to arbitrary accuracy (Section 4); and
- we provide an extensive empirical evaluation of LEM on a wide variey of data sets, including image and sequence classification, dynamical systems prediction, speech recognition, and language modeling, thereby demonstrating that LEM outperforms or is comparable to state-of-the-art RNNs, GRUs and LSTMs in each task (Section 5).

We also discuss a small portion of the large body of related work (Section 3), and we provide a brief discussion of our results in a broader context (Section 6). Much of the technical portion of our work is deferred to Supplementary Materials.

2 Long Expressive Memory

We start with the simplest example of a system of *two-scale ODEs*,

$$\frac{d\mathbf{y}}{dt} = \tau_y \left(\sigma\left(\mathbf{W}_y\mathbf{z} + \mathbf{V}_y\mathbf{u} + \mathbf{b}_y\right) - \mathbf{y}\right), \quad \frac{d\mathbf{z}}{dt} = \tau_z \left(\sigma\left(\mathbf{W}_z\mathbf{y} + \mathbf{V}_z\mathbf{u} + \mathbf{b}_z\right) - \mathbf{z}\right). \tag{1}$$

Here, $t \in [0, T]$ is the continuous time, $0 < \tau_y \leq \tau_z \leq 1$ are the two time scales, $\mathbf{y}(t) \in \mathbb{R}^{d_y}, \mathbf{z}(t) \in \mathbb{R}^{d_z}$ are the vectors of *slow* and *fast* variables and $\mathbf{u} = \mathbf{u}(t) \in \mathbb{R}^m$ is the *input signal*. For simplicity, we set $d_y = d_z = d$. The dynamic interactions between the neurons are modulated by weight matrices $\mathbf{W}_{y,z}, \mathbf{V}_{y,z}$, bias vectors $\mathbf{b}_{y,z}$ and a *nonlinear* tanh activation function $\sigma(u) = \tanh(u)$. Note that $\odot$ refers to the componentwise product of vectors.

However, two scales (one fast and one slow), may not suffice in representing a large number of scales that could be present in realistic sequential data sets. Hence, we need to generalize (1) to a *multiscale* version. One such generalization is provided by the following set of ODEs,

$$\begin{aligned} \frac{d\mathbf{y}}{dt} &= \hat{\sigma}\left(\mathbf{W}_2\mathbf{y} + \mathbf{V}_2\mathbf{u} + \mathbf{b}_2\right) \odot \left(\sigma\left(\mathbf{W}_y\mathbf{z} + \mathbf{V}_y\mathbf{u} + \mathbf{b}_y\right) - \mathbf{y}\right), \\ \frac{d\mathbf{z}}{dt} &= \hat{\sigma}\left(\mathbf{W}_1\mathbf{y} + \mathbf{V}_1\mathbf{u} + \mathbf{b}_1\right) \odot \left(\sigma\left(\mathbf{W}_z\mathbf{y} + \mathbf{V}_z\mathbf{u} + \mathbf{b}_z\right) - \mathbf{z}\right). \end{aligned} \tag{2}$$

In addition to previously defined quantities, we need additional weight matrices $\mathbf{W}_{1,2}, \mathbf{V}_{1,2}$, bias vectors $\mathbf{b}_{y,z}$ and sigmoid activation function $\hat{\sigma}(u) = 0.5(1 + \tanh(u/2))$. As $\hat{\sigma}$ is monotone, we can set $\mathbf{W}_{1,2} = \mathbf{V}_{1,2} \equiv 0$ and $(\mathbf{b}_1)_j = b_y, (\mathbf{b}_2)_j = b_z$, for all $1 \leq j \leq d$, with $\hat{\sigma}(b_{y,z}) = \tau_{y,z}$ to observe that the two-scale system (1) is a special case of (2). One can readily generalize this construction to obtain many different scales in (2). Thus, we can interpret $\left(\boldsymbol{\tau}_z(\mathbf{y}, t), \boldsymbol{\tau}_y(\mathbf{y}, t)\right) = \left(\hat{\sigma}\left(\mathbf{W}_1\mathbf{y} + \mathbf{V}_1\mathbf{u} + \mathbf{b}_1\right), \hat{\sigma}\left(\mathbf{W}_2\mathbf{y} + \mathbf{V}_2\mathbf{u} + \mathbf{b}_2\right)\right)$ in (2) as input and state dependent gating functions, which endow ODE (2) with *multiple time scales*. These scales can be learned adaptively (with respect to states) and dynamically (in time). Moreover, it turns out that the multiscale ODE system (2) is of the same general form (see **SM**§B) as the well-known Hodgkin-Huxley equations modeling the dynamics of the action potential for voltage-gated ion-channels in biological neurons [19].

In order to realize a concrete algorithm for our model, we propose a time-discretization of the multiscale ODE system (2). As is common with numerical discretizations of ODEs, doing so properly is important

to preserve desirable properties. To this end, we fix $\Delta t > 0$, and we discretize (2) with the following implicit-explicit (IMEX) time-stepping scheme to arrive at LEM, written in compact form as,

$$\begin{aligned}
\mathbf{\Delta t}_n &= \Delta t \hat{\sigma}(\mathbf{W}_1 \mathbf{y}_{n-1} + \mathbf{V}_1 \mathbf{u}_n + \mathbf{b}_1), \\
\overline{\mathbf{\Delta t}}_n &= \Delta t \hat{\sigma}(\mathbf{W}_2 \mathbf{y}_{n-1} + \mathbf{V}_2 \mathbf{u}_n + \mathbf{b}_2), \\
\mathbf{z}_n &= (1 - \mathbf{\Delta t}_n) \odot \mathbf{z}_{n-1} + \mathbf{\Delta t}_n \odot \sigma(\mathbf{W}_z \mathbf{y}_{n-1} + \mathbf{V}_z \mathbf{u}_n + \mathbf{b}_z), \\
\mathbf{y}_n &= (1 - \overline{\mathbf{\Delta t}}_n) \odot \mathbf{y}_{n-1} + \overline{\mathbf{\Delta t}}_n \odot \sigma(\mathbf{W}_y \mathbf{z}_n + \mathbf{V}_y \mathbf{u}_n + \mathbf{b}_y).
\end{aligned} \tag{3}$$

For steps $1 \leq n \leq N$, the hidden states in LEM (3) are $\mathbf{y}_n, \mathbf{z}_n \in \mathbb{R}^d$, with input state $\mathbf{u}_n \in \mathbb{R}^m$. The weight matrices are $\mathbf{W}_{1,2,z,y} \in \mathbb{R}^{d \times d}$ and $\mathbf{V}_{1,2,z,y} \in \mathbb{R}^{d \times m}$ and the bias vectors are $\mathbf{b}_{1,2,z,y} \in \mathbb{R}^d$. We also augment LEM (3) with a linear *output state* $\omega_n \in \mathbb{R}^o$ with $\omega_n = \mathcal{W}_y \mathbf{y}_n$, and $\mathcal{W}_y \in \mathbb{R}^{o \times d}$.

3 Related Work

We start by comparing our proposed model, LEM (3), to the widely used LSTM of Hochreiter and Schmidhuber [18]. Observe that $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ in (3) are similar in form to the *input, forget* and *output* gates in an LSTM (see **SM**§C), and that LEM (3) has exactly the same number of parameters (weights and biases) as an LSTM, for the same number of hidden units. Moreover, as detailed in **SM**§C, we show that by choosing very specific values of the LSTM gates and the $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ terms in LEM (3), the two models are equivalent. However, this analysis also reveals key differences between LEM (3) and LSTMs, as they are equivalent only under very stringent assumptions. In general, as the different gates in both LSTM and LEM (3) are *learned* from data, one can expect them to behave differently. Moreover in contrast to LSTM, LEM stems from a discretized ODE system (2), which endows it with (gradient) stable dynamics.

The use of *multiscale* neural network architectures in machine learning has a long history. An early example was provided in Hinton and Plaut [17], who proposed a neural network with each connection having a fast changing weight for temporary memory and a slow changing weight for long-term learning. More recently, one can view convolutional neural networks as multiscale architectures for processing multiple *spatial* scales in data [4].

The use of ODE-based learning architectures has also received considerable attention in recent years with examples such as *continuous-time* neural ODEs [9, 34, 33] and their recurrent extensions ODE-RNNs [35], as well as RNNs based on discretizations of ODEs [7, 12, 10, 28, 36, 37].

4 Rigorous Analysis of LEM

Here, we present our basic theoretical results for LEM.

Bounds on hidden states. The structure of LEM (3) allows us to prove (in **SM**§D.1) that its hidden states satisfy the following *pointwise bound*.

Proposition 4.1. *Denote $t_n = n\Delta t$ and assume that $\Delta t \leq \frac{1}{2}$. Further assume that the initial hidden states are $\mathbf{z}_0 = \mathbf{y}_0 \equiv 0$. Then, the hidden states $\mathbf{z}_n, \mathbf{y}_n$ of LEM (3) are bounded pointwise as,*

$$\max_{1 \leq i \leq d} \max\{|\mathbf{z}_n^i|, |\mathbf{y}_n^i|\} \leq \sqrt{t_n(1 + \Delta t)}, \quad \forall 1 \leq n \leq N. \tag{4}$$

Upper bounds on hidden state gradients. For any $1 \leq n \leq N$, let $\mathbf{X}_n \in \mathbb{R}^{2d}$ denoted the *concatenated hidden state*, given by $\mathbf{X}_n = \left[\mathbf{z}_n^1, \mathbf{y}_n^1, \ldots\ldots, \mathbf{z}_n^i, \mathbf{y}_n^i, \ldots\ldots, \mathbf{z}_n^d, \mathbf{y}_n^d\right]$. For simplicity of the exposition, we set output state $\omega_n = \mathbf{y}_n$ to consider a *loss function* of the form,

$$\mathcal{E} := \frac{1}{N} \sum_{n=1}^{N} \mathcal{E}_n, \quad \mathcal{E}_n = \frac{1}{2} \sum_{i=1}^{d} |\mathbf{y}_n^i - \overline{\mathbf{y}}_n^i|^2, \tag{5}$$

with $\overline{\mathbf{X}}_n = \left[\overline{\mathbf{z}}_n^1, \overline{\mathbf{y}}_n^1, \ldots\ldots, \ldots, \overline{\mathbf{z}}_n^d, \overline{\mathbf{y}}_n^d\right]$ being the underlying *ground truth*. The training of our proposed model (3) entails computing *gradients* of the loss function (5) with respect to its underlying weights and biases $\theta \in \mathbf{\Theta} = [\mathbf{W_{1,2,y,z}}, \mathbf{V_{1,2,y,z}}, \mathbf{b_{1,2,y,z}}]$, at every step of the gradient descent procedure. In **SM**§D.2, we prove the following upper bounds on this gradient.

Proposition 4.2. *Let $\mathbf{z}_n, \mathbf{y}_n$ be the hidden states generated by LEM* (3)*. We assume that $\Delta t << 1$ is chosen to be sufficiently small. Then, the gradient of the loss function $\mathcal{E}$ in* (5) *with respect to any parameter $\theta \in \mathbf{\Theta}$ is bounded as*

$$\left|\frac{\partial \mathcal{E}}{\partial \theta}\right| \leq \left(3 + \sqrt{3}\hat{X}\right)(3 + 6\eta), \quad \eta = \max\left\{\|\mathbf{W}_1\|_\infty, \|\mathbf{W}_2\|_\infty, \|\mathbf{W}_z\|_\infty, \|\mathbf{W}_y\|_\infty\right\}, \tag{6}$$

and $\hat{X} = \max\limits_{1 \leq n \leq N} \|\overline{\mathbf{X}}_n\|_\infty$.

The upper bound (6) clearly ensures that the hidden state gradient cannot blow up. Thus, the *exploding gradient problem* is mitigated for LEM (3).

On the vanishing gradient problem. Following [32], one needs a more precise characterization of the partial gradient $\frac{\partial \mathcal{E}_n^{(k)}}{\partial \theta} = \frac{\partial \mathcal{E}_n}{\partial \mathbf{X}_n} \frac{\partial \mathbf{X}_n}{\partial \mathbf{X}_k} \frac{\partial^+ \mathbf{X}_k}{\partial \theta}$, which measures the contribution to the hidden state gradient at step n arising from step k of the model, to show mitigation of the vanishing gradient problem. Here, $\frac{\partial^+ \mathbf{X}_k}{\partial \theta}$ refers to taking the partial derivative of $\mathbf{X}_k$ with respect to the parameter θ, while keeping the other arguments constant. In **SM**§D.3, we state and prove proposition D.2, which provides a precise formula for the asymptotics of the partial gradient. Here, we illustrate this formula in a special case as a corollary,

Proposition 4.3. *Let $\mathbf{y}_n, \mathbf{z}_n$ be the hidden states generated by LEM* (3)*. Consistent with the pointwise bounds* (4)*, we assume that $\mathbf{y}_n^j, \mathbf{z}_n^j \sim \mathcal{O}(\sqrt{t_n})$, for all $1 \leq j \leq n$ and $1 \leq j \leq d$ and the ground truth $\overline{\mathbf{y}}_n \sim \mathcal{O}(1)$. Then, for any $k << n$ (long-term dependencies) we have,*

$$\frac{\partial \mathcal{E}_n^{(k)}}{\partial \theta} = \mathcal{O}\left(\Delta t^{\frac{3}{2}}\right) + \mathcal{O}(\Delta t^2). \tag{7}$$

Here, the order notation is defined in **SM** (38) *and the constants in $\mathcal{O}(\Delta t^{\frac{3}{2}})$ depend on only on η* (6) *and $\overline{\eta} = \|\mathbf{W}_y\|_1$ and are independent of n, k.*

This formula (7) shows that although the partial gradient can be small, i.e., $\mathcal{O}(\Delta t^{\frac{3}{2}})$, it is in fact *independent of* k, ensuring that long-term dependencies contribute to gradients at much later steps and mitigating the vanishing gradient problem.

Universal approximation of general dynamical systems. The above bounds on hidden state gradients show that the proposed model LEM (3) mitigates the exploding and vanishing gradients problem. However, this by itself, does not guarantee that it can learn complicated and realistic input-output maps between sequences. To investigate the *expressivity* of the proposed LEM, we will show in the following proposition that it can approximate *any* dynamical system, mapping an input sequence $\mathbf{u}_n$ to an output sequence $\mathbf{o}_n$, of the (very) general form,

$$\boldsymbol{\phi}_n = \mathbf{f}\left(\boldsymbol{\phi}_{n-1}, \mathbf{u}_n\right), \quad \mathbf{o}_n = \mathbf{o}(\boldsymbol{\phi}_n), \quad \forall\ 1 \leq n \leq N, \tag{8}$$

with $\boldsymbol{\phi}_n \in \mathbb{R}^{d_h}, \mathbf{o}_n \in \mathbb{R}^{d_o}$ denoting the *hidden* and *output* states, respectively. The input signal is $\mathbf{u}_n \in \mathbb{R}^{d_u}$ and maps $\mathbf{f} : \mathbb{R}^{d_h} \times \mathbb{R}^{d_u} \mapsto \mathbb{R}^{d_h}$ and $\mathbf{o} : \mathbb{R}^{d_h} \mapsto \mathbb{R}^{d_o}$ are Lipschitz continuous. For simplicity, we set the initial state $\boldsymbol{\phi}_0 = 0$.

Proposition 4.4. *For all $1 \leq n \leq N$, let $\boldsymbol{\phi}_n, \mathbf{o}_n$ be given by the dynamical system* (8) *with input signal $\mathbf{u}_n$. Under the assumption that there exists a $R > 0$ such that $\max\{\|\boldsymbol{\phi}_n\|, \|\mathbf{u}_n\|\} < R$, for all $1 \leq n \leq N$, then for any given $\epsilon > 0$ there exists a LEM of the form* (3)*, with hidden states $\mathbf{y}_n, \mathbf{z}_n \in \mathbb{R}^{d_y}$ and output state $\omega_n = \mathcal{W}_y \mathbf{y}_n \in \mathbb{R}^{d_o}$, for some d_y such that the following holds,*

$$\|\mathbf{o}_n - \omega_n\| \leq \epsilon, \quad \forall 1 \leq n \leq N. \tag{9}$$

From this proposition, proved in **SM**§D.4, we conclude that, in principle, the proposed LEM (3) can approximate a very large class of dynamical systems.

Universal approximation of multiscale dynamical systems. While expressing a general form of input-output maps between sequences, the dynamical system (8) does not explicitly model dynamics at multiple scales. Instead, here we consider the following two-scale *fast-slow* dynamical system of the general form,

$$\boldsymbol{\phi}_n = \mathbf{f}(\boldsymbol{\phi}_{n-1}, \boldsymbol{\psi}_{n-1}, \mathbf{u}_n), \quad \boldsymbol{\psi}_n = \tau \mathbf{g}(\boldsymbol{\phi}_n, \boldsymbol{\psi}_{n-1}, \mathbf{u}_n), \quad \mathbf{o}_n = \mathbf{o}(\boldsymbol{\psi}_n). \tag{10}$$

Here, $0 < \tau << 1$ and 1 are the slow and fast time scales, respectively. The underlying maps $(\mathbf{f}, \mathbf{g}) : \mathbb{R}^{d_h \times d_h \times d_u} \mapsto \mathbb{R}^{d_h}$ are Lipschitz continuous. In the following proposition, proved in **SM**§D.5, we show that LEM (3) can approximate (10) to desired accuracy.

Proposition 4.5. *For any $0 < \tau << 1$, and for all $1 \leq n \leq N$, let $\boldsymbol{\phi}_n, \boldsymbol{\psi}_n, \mathbf{o}_n$ be given by the two-scale dynamical system* (10) *with input signal $\mathbf{u}_n$. Under the assumption that there exists a $R > 0$ such that $\max\{\|\boldsymbol{\phi}_n\|, \|\boldsymbol{\psi}_n\|, \|\mathbf{u}_n\|\} < R$, for all $1 \leq n \leq N$, then for any given $\epsilon > 0$, there exists a LEM of the form* (3)*, with hidden states $\mathbf{y}_n, \mathbf{z}_n \in \mathbb{R}^{d_y}$ and output state $\omega_n \in \mathbb{R}^{d_o}$ with $\omega_n = \mathcal{W}\mathbf{y}_n$ such that the following holds,*

$$\|\mathbf{o}_n - \omega_n\| \leq \epsilon, \quad \forall 1 \leq n \leq N. \tag{11}$$

Moreover, the weights, biases and size (number of neurons) of the underlying LEM (3) *are* independent *of the time-scale τ.*

This argument can be readily generalized to more than two time scales. Hence, we show that, in principle, the proposed model LEM (3) can approximate multiscale dynamical systems, with model size being *independent* of the underlying timescales. These theoretical results for LEM (3) point to the ability of this architecture to learn complicated multiscale input-output maps between sequences, while mitigating the exploding and vanishing gradients problem. Although useful prerequisities, these theoretical properties are certainly not sufficient to demonstrate that LEM (3) is efficient in practice. To do this, we perform several benchmark evaluations, and we report the results below.

5 Empirical results

We present a variety of experiments ranging from long-term dependency tasks to real-world applications as well as tasks which require high expressivity of the model. Details of the training procedure for each experiment can be found in **SM**§F. As competing models to LEM, we choose two different types of architectures—LSTMs and GRUs—as they are known to excel at expressive tasks such as language modeling and speech recognition, while not performing well on long-term dependency tasks, possibly due to the exploding and vanishing gradients problem. On the other hand, we choose state-of-the-art RNNs which are tailor-made to learn tasks with long-term dependencies. Our objective is to evaluate the performance of LEM and compare it with competing models. Code to replicate the experiments can be found at *https://github.com/tk-rusch/LEM*.

Very long adding problem. We start with the well-known adding problem [18], proposed to test the ability of a model to learn (very) long-term dependencies. The input is a two-dimensional sequence of length N, with the first dimension consisting of random numbers drawn from $\mathcal{U}([0, 1])$ and with two non-zero entries (both set to 1) in the second dimension, chosen at random locations, but one each in both halves of the sequence. The output is the sum of two numbers of the first dimension at positions, corresponding to the two 1 entries in the second dimension. We consider three very challenging cases, namely input sequences with length $N = 2000, 5000$ and 10000. The results of LEM together with competing models including state-of-the-art RNNs, which are explicitly designed to solve long-term dependencies, are presented in Fig. 1. We observe in this figure that while LSTM is not able to beat the baseline mean-square error of 0.167 (the variance of the baseline output 1) in any of the three cases, the other models can successively learn the task in the case of $N = 2000$. However for $N = 5000$, only LEM and coRNN, which has been recognized as state-of-the-art (SOTA) on this task so far, are able to beat the baseline. We note that LEM converges significantly earlier than coRNN in this case. In the extreme case of $N = 10000$, only LEM is able to beat the baseline and does so rather quickly. In fact, the experiments suggest that LEM learns the adding problem independently of the length of the underlying sequences by converging very fast and reaching a very low test MSE in all three cases.

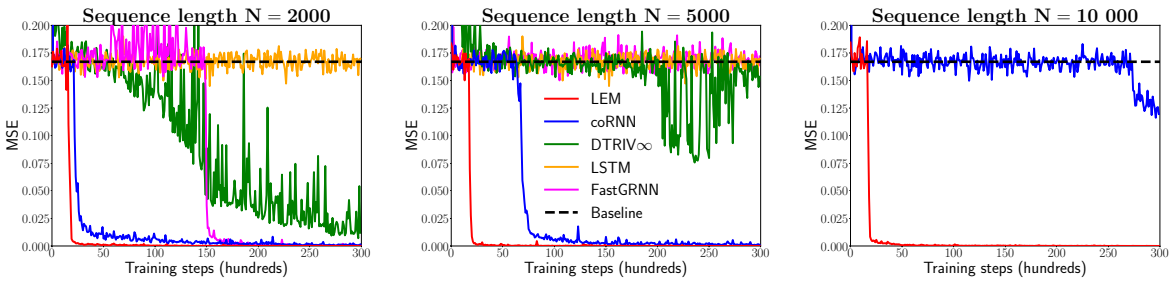

Figure 1: Results on the very long adding problem for LEM, coRNN, DTRIV∞ [6], FastGRNN [25] and LSTM based on three very long sequence lengths N, i.e., $N = 2000$, $N = 5000$ and $N = 10000$.

Sequential image recognition. We consider three experiments based on two widely-used image recognition data sets, i.e., MNIST [26] and CIFAR-10 [22], where the goal is to predict the correct label after reading in the whole sequence. The first two tasks are based on MNIST images, which are flattened along the rows to obtain sequences of length $N = 784$. In sequential MNIST (sMNIST), the sequences are fed to the model one pixel at a time in streamline order, while in permuted sequential MNIST (psMNIST), a fixed random permutation is applied to the sequences, resulting in much longer dependency than for sMNIST. We also consider the more challenging noisy CIFAR-10 (nCIFAR-10) experiment [7], where CIFAR-10 images are fed to the model row-wise and flattened along RGB channels, resulting in 96-dimensional sequences, each of length 32. Moreover, a random noise padding is applied after the first 32 inputs to produce sequences of length $N = 1000$. Hence, in addition to classifying the underlying image, a model has to store this result for a long time. In Table 1, we present the results for LEM on the three tasks together with other SOTA RNNs, which were explicitly designed to solve long-term dependency tasks, as well as LSTM and GRU baselines. We observe that LEM outperforms all other methods on sMNIST and nCIFAR-10. Additionally on psMNIST, LEM performs as well as coRNN, which has been SOTA among single-layer RNNs on this task.

Table 1: Test accuracies on sMNIST, psMNIST and nCIFAR-10, where M denotes the total number of parameters of the corresponding model. Results of other models are taken from the respective original paper referenced in the main text, except that the results for LSTM are taken from Helfrich et al. [15], for GRU from Chang et al. [8] and the results indicated by * are added by us.

Model	MNIST			CIFAR-10	
	sMNIST	psMNIST	# units / M	nCIFAR-10	# units / M
GRU	99.1%	94.1%	256 / 201k	43.8%*	128 / 88k
LSTM	98.9%	92.9%	256 / 267k	11.6%	128 / 116k
anti.sym. RNN	98.0%	95.8%	128 / 10k	48.3%	256 / 36k
Lipschitz RNN	99.4%	96.3%	128 / 34k	57.4%	128 / 46k
expRNN	98.4%	96.2%	360 / 69k	52.9%*	360 / 103k
coRNN	99.3%	**96.6**%	128/ 34k	59.0%	128 / 46k
LEM	**99.5**%	**96.6**%	128 / 68k	**60.5**%	128 / 116k

EigenWorms: Very long sequences for genomics classification. The goal of this task [2] is to classify worms as belonging to either the wild-type or four different mutants, based on 259 very long sequences (length $N = 17984$) measuring the motion of a worm. In addition to the nominal length, it was empirically shown in Rusch and Mishra [37] that the EigenWorms sequences exhibit actual very long-term dependencies (i.e., longer than 10k). Following Morrill et al. [31] and [37], we divide the data into a train, validation and test set according to a $70\%, 15\%, 15\%$ ratio. In Table 2, we present results for LEM together with other models as well as different sub-sampling routines. As the validation and test sets, each consist of only 39 sequences, we report the mean (and standard deviation of) accuracy over 5

Table 2: Test accuracies on EigenWorms using 5 re-trainings of each best performing network (based on the validation set), where all other results are taken from Rusch and Mishra [37] except that the NRDE result is taken from [31].

Model	test accuracy	# units	# params
t-BPTT LSTM	57.9% ± 7.0%	32	5.3k
sub-samp. LSTM	69.2% ± 8.3%	32	5.3k
NRDE	83.8% ± 3.0%	32	35k
expRNN	40.0% ± 10.1%	64	2.8k
IndRNN (2 layers)	49.7% ± 4.8%	32	1.6k
coRNN	86.7% ± 3.0%	32	2.4k
UnICORNN (2 layers)	90.3% ± 3.0%	32	1.5k
LEM	**92.3**% ± **1.8**%	32	5.3k

random initializations to rule out lucky outliers. We observe from this table that LEM outperforms all other methods, even the 2-layer UnICORNN architecture, which has been SOTA on this task.

Healthcare application: Heart-rate prediction. In this experiment, one predicts the heart rate from a time-series of measured PPG data, which is part of the TSR archive [38] and has been collected at the Beth Isreal Deaconess medical center. The data set, consisting of 7949 sequences, each of length $N = 4000$, is divided into a train, validation and test set according to a 70%,15%,15% ratio, [31, 37]. The results, presented in Table 3, show that LEM outperforms the other competing models by having a test L^2 error of 35% less than the SOTA UnICORNN, while being an order of magnitude better than LSTM.

Table 3: Test L^2 error on heart-rate prediction using PPG data. All results are obtained by running the same code and using the same fine-tuning protocol.

Model	test L^2 error	# units	# params
LSTM	9.93	128	67k
expRNN	1.63	256	34k
IndRNN (3 layers)	1.94	128	34k
coRNN	1.61	128	34k
UnICORNN (3 layers)	1.31	128	34k
LEM	**0.85**	128	67k

Multiscale dynamical system prediction. The FitzHugh-Nagumo system [13]

$$v' = v - \frac{v^3}{3} - w + I_{\text{ext}}, \quad w' = \tau(v + a - bw), \tag{12}$$

is a prototypical model for a two-scale fast-slow nonlinear dynamical system, with fast variable v and slow variable w and $\tau << 1$ determining the fast-time scale. This *relaxation-oscillator* is an approximation to the Hodgkin-Huxley model [19] of neuronal action-potentials under an external signal $I_{\text{ext}} \geq 0$. With $\tau = 0.02$, $I_{\text{ext}} = 0.5$, $a = 0.7$, $b = 0.8$ and initial data $(v_0, w_0) = (c, 0)$, with c randomly drawn from $\mathcal{U}([-1, 1])$, we numerically approximate (12) with the explicit Runge-Kutta method of order 5(4) in the interval $[0, 400]$ and generate 128 training and validation and 1024 test sequences, each of length $N = 1000$, to complete the data set. The results, presented in Table 4, show that LEM not only outperforms LSTM by a factor of 6 but also all other methods including coRNN, which is tailormade for oscillatory time-series. This reinforces our theory by demonstrating efficient approximation of multiscale dynamical systems with LEM.

Table 4: Test L^2 error on FitzHugh-Nagumo system prediction. All results are obtained by running the same code and using the same fine-tuning protocol.

Model	error ($\times 10^{-2}$)	# units	# params
LSTM	1.2	16	1k
expRNN	2.3	50	1k
LipschitzRNN	1.8	24	1k
FastGRNN	2.2	34	1k
coRNN	0.4	24	1k
LEM	**0.2**	16	1k

Speech recognition: Google12 (V2) keyword spotting. The Google Speech Commands data set V2 is a widely used benchmark for speech recognition [39], consisting of 35 words, sampled at a rate of 16 kHz from 1 second utterances of 2618 speakers. We focus on the 12-label task (Google12) and follow the pre-defined splitting of the data set into train/validation/test sets and test different sequential models. In order to ensure comparability of different architectures, we do not use performance-enhancing tools such as convolutional filtering or multi-head attention. From Table 5, we observe that both LSTM and GRU, widely used models in this context, perform very well with a test accuracy of 95%. Nevertheless, LEM is able to outperform both on this task and provides the best performance.

Table 5: Test accuracies on Google12. All results are obtained by running the same code and using the same fine-tuning protocol.

Model	test accuracy	# units	# params
tanh-RNN	73.4%	128	27k
LSTM	94.9%	128	107k
GRU	95.2%	128	80k
FastGRNN	94.8%	128	27k
expRNN	92.3%	128	19k
coRNN	94.7%	128	44k
LEM	**95.7**%	128	107k

Language modeling: Penn Tree Bank corpus. Language modeling with the widely used small scale Penn Treebank (PTB) corpus [29], preprocessed by Mikolov et al. [30], has been identified as an excellent task for testing the expressivity of recurrent models [21]. To this end, in Table 6, we report the results of different architectures, with a similar number of hidden units, on the PTB char-level task and observe that RNNs, designed explicitly for learning long-term dependencies, perform significantly worse than LSTM and GRU. On the other hand, LEM is able to outperform both LSTM and GRU on this task by some margin (a test bpc of 1.25 in contrast with approximately a bpc of 1.36). In fact, LEM provides the smallest test bpc among all reported single-layer recurrent models on this task, to the best of our knowledge. This superior performance is further illustrated in Table 7, where the test perplexity for different models on the PTB word-level task is presented. We observe that not only does LEM significantly outperform (by around 40%) LSTM, but it also provides again the best performance among all single layer recurrent models, including the recently proposed TARNN [20]. Moreover, the single-layer results for LEM are better than reported results for multi-layer LSTM models, such as in Gal and Ghahramani [14] (2-layer LSTM, 1500 units each: 75.2 test perplexity) or Bai et al. [3] (3-layer LSTM, 700 units each: 78.93 test perplexity).

Multiscale behavior of LEM. As LEM (3) has been designed to represent multiple scales, it is natural to examine if such multiscale behavior is observed in the experiments. To this end, we evaluate $\boldsymbol{\Delta t}_n, \overline{\boldsymbol{\Delta t}}_n$, which represent effective time scales in (3). In our experiments, we observe that although $\boldsymbol{\Delta t}_n, \overline{\boldsymbol{\Delta t}}_n$ had

Table 6: Test bits-per-character (bpc) on PTB character-level for single layer LEM and other single layer RNN architectures. Other results are taken from the papers cited accordingly in the table, while the results for coRNN are added by us.

Model	test bpc	# units	# params
anti.sym RNN [12]	1.60	1437	1.3M
Lipschitz RNN [12]	1.42	764	1.3M
expRNN [21]	1.51	1437	1.3M
coRNN	1.46	1024	2.3M
nnRNN [21]	1.47	1437	1.3M
LSTM [23]	1.36	1000	5M
GRU [3]	1.37	1024	3M
LEM	**1.25**	1024	5M

Table 7: Test perplexity on PTB word-level for single layer LEM and other single layer RNN architectures.

Model	test perplexity	# units	# params
Lipschitz RNN [12]	115.4	160	76k
FastRNN [20]	115.9	256	131k
LSTM [20]	116.9	256	524k
SkipLSTM [20]	114.2	256	524k
TARNN [20]	94.6	256	524k
LEM	**72.8**	256	524k

values at a single scale before training, they spanned a wide range of scales (several orders of magnitude) after training was completed. See Fig. 2 in **SM**§A for an example of this phenomenon. Moreover, in the experiments, the frequency of values of $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ is distributed as a *power law* of their amplitude (see **SM**§A for a discussion of this theme.) This power-law behavior indicates that a range of scales are indeed expressed, albeit at different frequencies, with smaller scales contributing proportionately to the multiscale dynamics.

6 Discussion

The design of a gradient-based model for processing sequential data that can mitigate the exploding and vanishing gradients problem to learn tasks with long-term dependencies while retaining the ability to learn complicated sequential input-output maps is very challenging [21]. In this paper, we have proposed *Long Expressive Memory* (LEM), a novel recurrent architecture, based on a suitable time-discretization of a specific multiscale system of ODEs (2). By a combination of theoretical arguments and extensive empirical evaluations on a diverse set of learning tasks, we demonstrate that LEM is able to learn long-term dependencies while retaining sufficient expressivity for efficiently solving realistic learning tasks.

It is natural to ask why LEM performs so well. A part of the answer lies in the mitigation of the exploding and vanishing gradients problem. Proofs for gradient bounds (6),(7) reveal a key role played by the smallness of the hyperparameter Δt. We observe from **SM** Table 8 that small values of Δt might be needed for problems with long-term dependencies. The smallest values are necessary for the Eigenworms experiment, with input sequences of length 18K. An ablation study, presented in **SM** Fig. 4, shows that Δt needs to be quite small for this very long sequence task. On the other hand, no fine tuning of the hyperparameter Δt is necessary for tasks such as language modeling and speech recognition, which do not necessarily possess long-term dependencies, and a default value of $\Delta t = 1$ yielded very good performance. However, mitigation of exploding and vanishing gradients problem alone does not explain high expressivity of LEM. In this context, we proved that LEMs can approximate a very large class of multiscale dynamical systems. In fact, as detailed in **SM**§E, one can show that LEMs can emulate

the well-known *Heterogeneous Multiscale Method* (HMMs) (see Kuehn [24] and references therein), for approximating ODEs with several time-scales. Moreover, we provide experimental evidence in **SM**§A to observe that LEM not only expresses a range of scales, as it is designed to do, but also these scales contribute proportionately to the resulting multiscale dynamics. We believe that this combination of gradient stable dynamics, specific model structure, and its multiscale resolution can explain the observed performance of LEM.

We conclude with a comparison of LEM and the widely-used gradient-based LSTM model. In addition to having exactly the same number of parameters for the same number of hidden units, our experiments show that LEMs are better than LSTMs on expressive tasks such as speech recognition and language modeling, while providing significantly better performance on long-term dependencies. This robustness of the performance of LEM with respect to sequence length paves the way for its application to learning many different sequential data sets where competing models might not perform satisfactorily.

Acknowledgements.

The research of TKR and SM was performed under a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 770880). NBE and MWM would like to acknowledge IARPA (contract W911NF20C0035), NSF, and ONR for providing partial support of this work. Our conclusions do not necessarily reflect the position or the policy of our sponsors, and no official endorsement should be inferred.

References

[1] M. Arjovsky, A. Shah, and Y. Bengio. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning*, pages 1120–1128, 2016.

[2] A. Bagnall, H. A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, and E. Keogh. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

[3] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[4] S. Bai, V. Koltun, and J. Z. Kolter. Multiscale deep equilibrium models. In *Advances in Neural Information Processing Systems*, pages 770–778, 2020.

[5] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory.*, 39(3):930–945, 1993.

[6] M. L. Casado. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems*, pages 9154–9164, 2019.

[7] B. Chang, M. Chen, E. Haber, and E. H. Chi. Antisymmetricrnn: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations*, 2018.

[8] S. Chang, Y. Zhang, W. Han, M. Yu, X. Guo, W. Tan, X. Cui, M. Witbrock, M. A. Hasegawa-Johnson, and T. S. Huang. Dilated recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 77–87, 2017.

[9] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6571–6583, 2018.

[10] Z. Chen, J. Zhang, M. Arjovsky, and L. Bottou. Symplectic recurrent neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[11] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014.

[12] N. B. Erichson, O. Azencot, A. Queiruga, and M. W. Mahoney. Lipschitz recurrent neural networks. In *International Conference on Learning Representations*, 2021.

[13] R. Fitzhugh. Mathematical models of threshold phenomena in the nerve membrane. *Bull. Math. Biophysics*, 17:257–278, 1955.

[14] Y. Gal and Z. Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29:1019–1027, 2016.

[15] K. Helfrich, D. Willmott, and Q. Ye. Orthogonal recurrent neural networks with scaled cayley transform. In *International Conference on Machine Learning*, pages 1969–1978. PMLR, 2018.

[16] M. Henaff, A. Szlam, and Y. LeCun. Recurrent orthogonal networks and long-memory tasks. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2034–2042, 2016.

[17] G. Hinton and D. Plaut. Using fast weights to deblur old memories. In *Proceedings of the ninth annual conference of the cognitive science society*, pages 177–186, 1987.

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[19] A. Hodgkin and A. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117:500–544, 1952.

[20] A. Kag and V. Saligrama. Time adaptive recurrent neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15149–15158, June 2021.

[21] G. Kerg, K. Goyette, M. P. Touzel, G. Gidel, E. Vorontsov, Y. Bengio, and G. Lajoie. Non-normal recurrent neural network (nnrnn): learning long time dependencies while improving expressivity with transient dynamics. In *Advances in Neural Information Processing Systems*, pages 13591–13601, 2019.

[22] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[23] D. Krueger, T. Maharaj, J. Kramár, M. Pezeshki, N. Ballas, N. R. Ke, A. Goyal, Y. Bengio, A. C. Courville, and C. J. Pal. Zoneout: Regularizing rnns by randomly preserving hidden activations. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[24] C. Kuehn. *Multiple time scale dynamics*, volume 191. Springer, 2015.

[25] A. Kusupati, M. Singh, K. Bhatia, A. Kumar, P. Jain, and M. Varma. Fastgrnn: A fast, accurate, stable and tiny kilobyte sized gated recurrent neural network. In *Advances in Neural Information Processing Systems*, pages 9017–9028, 2018.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[27] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5457–5466, 2018.

[28] S. H. Lim, N. B. Erichson, L. Hodgkinson, and M. W. Mahoney. Noisy recurrent neural networks. *arXiv preprint arXiv:2102.04877*, 2021.

[29] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[30] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[31] J. Morrill, C. Salvi, P. Kidger, and J. Foster. Neural rough differential equations for long time series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7829–7838. PMLR, 18–24 Jul 2021.

[32] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *ICML'13*, page III–1310–III–1318. JMLR.org, 2013.

[33] A. Queiruga, N. B. Erichson, L. Hodgkinson, and M. W. Mahoney. Compressing deep ode-nets using basis function expansions. *arXiv preprint arXiv:2106.10820*, 2021.

[34] A. F. Queiruga, N. B. Erichson, D. Taylor, and M. W. Mahoney. Continuous-in-depth neural networks. Technical Report Preprint: arXiv:2008.02389, 2020.

[35] Y. Rubanova, R. T. Chen, and D. Duvenaud. Latent odes for irregularly-sampled time series. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 5320–5330, 2019.

[36] T. K. Rusch and S. Mishra. Coupled oscillatory recurrent neural network (cornn): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations*, 2021.

[37] T. K. Rusch and S. Mishra. Unicornn: A recurrent model for learning very long time dependencies. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9168–9178. PMLR, 2021.

[38] C. W. Tan, C. Bergmeir, F. Petitjean, and G. I. Webb. Monash university, uea, ucr time series regression archive. *arXiv preprint arXiv:2006.10996*, 2020.

[39] P. Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[40] S. Wisdom, T. Powers, J. Hershey, J. Le Roux, and L. Atlas. Full-capacity unitary recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 4880–4888, 2016.

Supplementary Material for:
Long Expressive Memory for Sequence Modeling

A Multiscale behavior of LEM

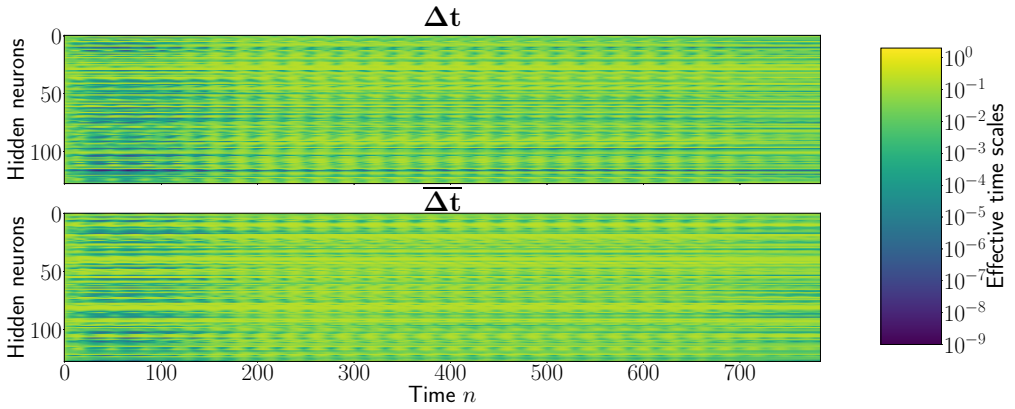

Figure 2: Multiscale behavior of LEM after training on sMNIST.

LEM (3) is designed to represent multiple scales. The terms $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ are explicitly designed to express these multiple scales. To see if this holds in practice, we evaluate $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$, for all $1 \leq n \leq N$, both before and after training, in all our experiments. To illustrate our observations, we consider the sMNIST learning task and plot $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ in Fig. 2 *after training* is completed. We observe from this figure that $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ express a very large range of scales that span over approximately 7 orders of magnitude. This stands in stark contrast to the observation that, *before training*, $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ only assumed values within a single scale, very close to 0.8. Moreover, this variation in scales (after training) is both with respect to sequence position n and state location $1 \leq \alpha \leq d$. Thus, we can conclude that $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ in LEM *learn* to express a range of scales from data, dynamically (in time) and adaptively (with respect to neurons). This phenomenon was also observed in other experiments.

It is interesting to ask if these range of scales, expressed by $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ in LEM, are equally distributed. To answer this, we plot the empirical histograms of the frequency of each scale with respect to the scale amplitude, in order to approximate the underlying distribution. The results, plotted in Fig. 3 (in log-log) show that this frequency of occurrence of different scales for the sMNIST task decays as a power law of the scale amplitude, with exponents ≈ 0.5 for both $\mathbf{\Delta t}_n$ and $\overline{\mathbf{\Delta t}}_n$.

B Relation between LEM and the Hodgkin-Huxley equations

We observe that the multiscale ODEs (2), on which LEM is based, are a special case of the following ODE system,

$$\frac{d\mathbf{z}}{dt} = \mathbf{F}_z\left(\mathbf{y}, t\right) - \mathbf{G}_z(\mathbf{y}) \odot \mathbf{z}, \quad \frac{d\mathbf{y}}{dt} = \mathbf{F}_y\left(\mathbf{z}, t\right) \mathbf{H}\left(\mathbf{y}, t\right) - \mathbf{G}_y(\mathbf{y}) \odot \mathbf{y}. \tag{13}$$

As remarked in the main text, it turns out the well-known Hodgkin-Huxley equations [19], modeling the the dynamics of the action potential of a biological neuron can also be written down in the abstract form (13), with $d_y = 1$, $d_z = 3$ and the variables $\mathbf{y} = y$ modeling the voltage and $\mathbf{z} = (z_1, z_2, z_3)$ modeling the concentration of Potassium activation, Sodium activation and Sodium inactivation channels.

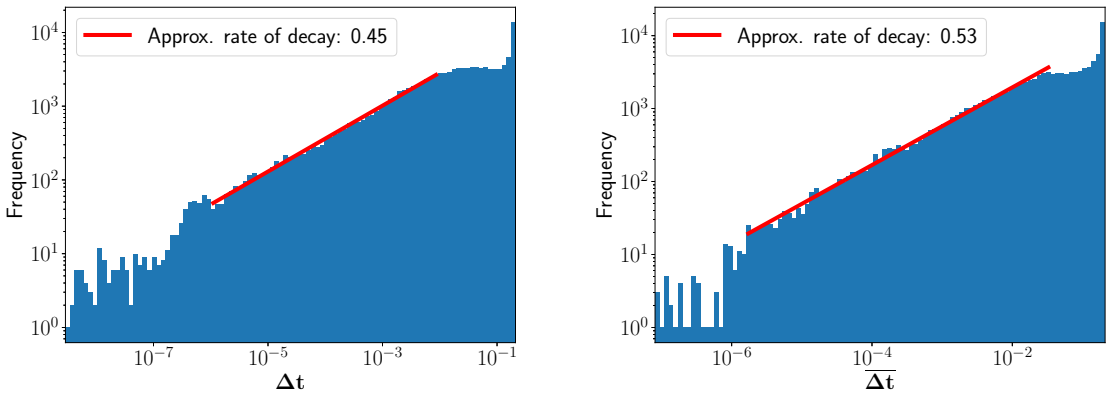

Figure 3: Log of the frequency (y-axis) vs. Log of the amplitude of $\mathbf{\Delta t}_n$ (Left) and $\overline{\mathbf{\Delta t}}_n$ (Right) for LEM trained on sMNIST. Note that the frequency is binned over all sequence positions n and locations in the vectors $\mathbf{\Delta t}_n$ and $\overline{\mathbf{\Delta t}}_n$.

The exact form of the different functions in (13) for the Hodgkin-Huxley equations is given by,

$$
\begin{aligned}
&\mathbf{F}_z(y) = (\alpha_1(y), \alpha_2(y), \alpha_3(y))\,,\\
&\mathbf{G}_z(y) = (\alpha_1(y) + \beta_1(y), \alpha_2(y) + \beta_2(y), \alpha_3(y) + \beta_3(y))\,,\\
&\alpha_1(y) = \frac{0.01(10+\hat{y}-y)}{e^{\frac{10+\hat{y}-y}{10}}-1}, \quad \alpha_2(y) = \frac{0.1(25+\hat{y}-y)}{e^{\frac{25+\hat{y}-y}{10}}-1}, \quad \alpha_3(y) = 0.07e^{\frac{\hat{y}-y}{20}},\\
&\beta_1(y) = 0.125e^{\frac{\hat{y}-y}{80}}, \quad \beta_2(y) = 4e^{\frac{\hat{y}-y}{18}}, \quad \beta_3(y) = \frac{1}{1+e^{1+\frac{\hat{y}-y}{10}}},\\
&\mathbf{F}_y(z,t) = u(t) + z_1^4 + z_2^3 z_3, \quad \mathbf{H}(y) = c_1(\bar{y}-y) + c_2(\bar{y}-y), \quad \mathbf{G}_y(y) = c_3,
\end{aligned}
\tag{14}
$$

with input current u and constants $\hat{y}, \bar{y}, c_{1,2,3}$, whose exact values can be read from [19].

Thus, the multiscale ODEs (2) and the Hodgkin-Huxley equations are special case of the same general family (13) of ODEs. Moreover, the *gating functions* $\mathbf{G}_{y,z}(\mathbf{y})$, that model voltage-gated ion channels in the Hodgkin-Huxley equations, are similar in form to $\mathbf{\Delta t}_n, \overline{\mathbf{\Delta t}}_n$ in (2).

It is also worth highlighting the differences between our proposed model LEM (and the underlying ODE system (2)) and the Hodgkin-Huxley ODEs modeling the dynamics of the neuronal action potential. Given the complicated form of the nonlinearites $\mathbf{F}_{y,z}, \mathbf{G}_{y,z}, \mathbf{H}$ in the Hodgkin-Huxley equations (14), we cannot use them in designing any learning model. Instead, building on the abstract form of (13), we propose *bespoke* non-linearities in the ODE (2) to yield a tractable learning model, such as LEM (3). Moreover, it should be emphasized that the Hodgkin-Huxley equations only model the dynamics of a single neuron (with a scalar voltage and 3 ion channels), whereas the hidden state dimension d of (2) can be arbitrary.

C Relation between LEM and LSTM

The well-known LSTM [18] is given by,

$$
\begin{aligned}
\mathbf{f}_n &= \hat{\sigma}(\mathbf{W}_f\,\mathbf{h}_{n-1} + \mathbf{V}_f\,\mathbf{u}_n + \mathbf{b}_f\,)\\
\mathbf{i}_n &= \hat{\sigma}(\mathbf{W}_i\mathbf{h}_{n-1} + \mathbf{V}_i\mathbf{u}_n + \mathbf{b}_i)\\
\mathbf{o}_n &= \hat{\sigma}(\mathbf{W}_o\mathbf{h}_{n-1} + \mathbf{V}_o\mathbf{u}_n + \mathbf{b}_o)\\
\mathbf{c}_n &= \mathbf{f}_n \odot \mathbf{c}_{n-1} + \mathbf{i}_n \odot \sigma(\mathbf{W}\mathbf{h}_{n-1} + \mathbf{V}\mathbf{u}_n + \mathbf{b})\\
\mathbf{h}_n &= \mathbf{o}_n \odot \sigma(\boldsymbol{\psi}_n).
\end{aligned}
\tag{15}
$$

Here, for any $1 \leq n \leq N$, $\mathbf{h}_n \in \mathbb{R}^d$ is the hidden state and $\mathbf{c}_n \in \mathbb{R}^d$ is the so-called *cell state*. The vectors $\mathbf{i}_n, \mathbf{f}_n, \mathbf{o}_n \in \mathbb{R}^d$ are the *input, forget* and *output* gates, respectively. $\mathbf{u}_n \in \mathbb{R}^m$ is the input signal and the weight matrices and bias vectors are given by $\mathbf{W}, \mathbf{W}_{f,i,o} \in \mathbb{R}^{d\times d}, \mathbf{V}, \mathbf{V}_{f,i,o} \in \mathbb{R}^{m\times d}$ and $\mathbf{b}, \mathbf{b}_{f,i,o} \in \mathbb{R}^d$, respectively.

It is straightforward to relate LSTM (15) and LEM (3) by first setting the cell state $\mathbf{c}_n = \mathbf{z}_n$, for all $1 \leq n \leq N$ and the hidden state $\mathbf{h}_n = \mathbf{y}_n$.

We further need to assume that the input state $\mathbf{i}_n = \boldsymbol{\Delta t}_n$ and the forget state has to be $\mathbf{f}_n = 1 - \boldsymbol{\Delta t}_n$. Finally, the output state of the LSTM (15) has to be

$$\mathbf{o}_n = \overline{\boldsymbol{\Delta t}}_n = 1, \quad \forall 1 \leq n \leq N.$$

Under these assumptions and by setting $\Delta t = 1$, we can readily observe that the LEM (3) and LSTM (15) are equivalent.

D Supplement to the rigorous analysis of LEM

In this section, we will provide detailed proofs of the propositions in Section 4 of the main article. We start with the following simplifying notation for various terms in LEM (3),

$$\begin{aligned}
\mathbf{A}_{n-1} &= \mathbf{W}_1 \mathbf{y}_{n-1} + \mathbf{V}_1 \mathbf{u}_n + \mathbf{b}_1, \\
\mathbf{B}_{n-1} &= \mathbf{W}_2 \mathbf{y}_{n-1} + \mathbf{V}_2 \mathbf{u}_n + \mathbf{b}_2, \\
\mathbf{C}_{n-1} &= \mathbf{W}_z \mathbf{y}_{n-1} + \mathbf{V}_z \mathbf{u}_n + \mathbf{b}_z, \\
\mathbf{D}_n &= \mathbf{W}_y \mathbf{z}_n + \mathbf{V}_y \mathbf{u}_n + \mathbf{b}_y.
\end{aligned}$$

Note that for all $1 \leq n \leq N$, $\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n, \mathbf{D}_n \in \mathbb{R}^d$. With the above notation, LEM (3) can be written componentwise, for each component $1 \leq i \leq d$ as,

$$\begin{aligned}
\mathbf{z}_n^i &= \mathbf{z}_{n-1}^i + \Delta t \hat{\sigma}(\mathbf{A}_{n-1}^i)\sigma(\mathbf{C}_{n-1}^i) - \Delta t \hat{\sigma}(\mathbf{A}_{n-1}^i)\mathbf{z}_{n-1}^i, \\
\mathbf{y}_n^i &= \mathbf{y}_{n-1}^i + \Delta t \hat{\sigma}(\mathbf{B}_{n-1}^i)\sigma(\mathbf{D}_n^i) - \Delta t \hat{\sigma}(\mathbf{B}_{n-1}^i)\mathbf{y}_{n-1}^i.
\end{aligned} \tag{16}$$

Note that the techniques of proof in this following three sub-sections burrow heavily from those introduced in Rusch and Mishra [36].

D.1 Proof of Proposition 4.1 of main text.

Proposition 4.1 yields the bound (4) for the hidden states of LEM. We provide the proof below,

Proof. We will use the following elementary identity,

$$b(a-b) = \frac{a^2}{2} - \frac{b^2}{2} - \frac{1}{2}(a-b)^2, \tag{17}$$

for any $a, b \in \mathbb{R}$.

We fix $1 \leq i \leq d$ and multiply the first equation in (16) with $\mathbf{z}_{n-1}^i$ and apply (17) to obtain,

$$
\begin{aligned}
\frac{(\mathbf{z}_n^i)^2}{2} &= \frac{(\mathbf{z}_{n-1}^i)^2}{2} + \Delta t\hat{\sigma}(\mathbf{A}_{n-1}^i)\sigma(\mathbf{C}_{n-1}^i)\mathbf{z}_{n-1}^i - \Delta t\hat{\sigma}(\mathbf{A}_{n-1}^i)(\mathbf{z}_{n-1}^i)^2 + \frac{(\mathbf{z}_n^i - \mathbf{z}_{n-1}^i)^2}{2} \\
&= \frac{(\mathbf{z}_{n-1}^i)^2}{2} + \Delta t\hat{\sigma}(\mathbf{A}_{n-1}^i)\sigma(\mathbf{C}_{n-1}^i)\mathbf{z}_{n-1}^i - \Delta t\hat{\sigma}(\mathbf{A}_{n-1}^i)(\mathbf{z}_{n-1}^i)^2 \\
&\quad + \frac{\Delta t^2}{2}\left(\hat{\sigma}(\mathbf{A}_{n-1}^i)\sigma(\mathbf{C}_{n-1}^i) - \hat{\sigma}(\mathbf{A}_{n-1}^i)\mathbf{z}_{n-1}^i\right)^2, \quad \text{(from (16))} \\
&\leq \frac{(\mathbf{z}_{n-1}^i)^2}{2} + \Delta t\hat{\sigma}(\mathbf{A}_{n-1}^i)\sigma(\mathbf{C}_{n-1}^i)\mathbf{z}_{n-1}^i - \Delta t\hat{\sigma}(\mathbf{A}_{n-1}^i)(\mathbf{z}_{n-1}^i)^2 \\
&\quad + \Delta t^2(\hat{\sigma}(\mathbf{A}_{n-1}^i)\sigma(\mathbf{C}_{n-1}^i)^2 + \Delta t^2(\mathbf{A}_{n-1}^i)(\mathbf{z}_{n-1}^i)^2 \quad (\text{as } (a-b)^2 \leq 2(a^2+b^2)) \\
\Rightarrow \quad \frac{(\mathbf{z}_n^i)^2}{2} &\leq \frac{(\mathbf{z}_{n-1}^i)^2}{2} + \frac{\Delta t}{2}(\sigma(\mathbf{C}_{n-1}^i)^2 + \Delta t^2(\hat{\sigma}(\mathbf{A}_{n-1}^i))^2(\sigma(\mathbf{C}_{n-1}^i)^2 \\
&\quad + \left[\Delta t^2(\hat{\sigma}(\mathbf{A}_{n-1}^i))^2 - \frac{\Delta t}{2}\hat{\sigma}(\mathbf{A}_{n-1}^i)\right](\mathbf{z}_{n-1}^i)^2.
\end{aligned}
$$

Using the fact that $0 \leq \hat{\sigma}(x) \leq 1$ for all $x \in \mathbb{R}$ and $(\sigma(x))^2, (\hat{\sigma}(x))^2$ and the assumption that $\Delta t \leq 1/2$, we obtain from the last line of the previous equation that,

$$
(\mathbf{z}_n^i)^2 \leq (\mathbf{z}_{n-1}^i)^2 + (\Delta t + 2\Delta t^2), \quad \forall 1 \leq n \leq N.
$$

Iterating the above estimate over $n = 1, \ldots, \bar{n}$, for any $1 \leq \bar{n} \leq N$ and setting $\bar{n} = n$ yields,

$$
\begin{aligned}
&(\mathbf{z}_n^i)^2 \leq (\mathbf{z}_0^i)^2 + n\left(\Delta t + 2\Delta t^2\right), \\
\Rightarrow \quad &(\mathbf{z}_n^i)^2 \leq t_n(1 + 2\Delta t), \quad \text{as } \mathbf{z}_0^i = 0,\ t_n = n\Delta t.
\end{aligned}
$$

Taking a square root in the above inequality yields the desired pointwise bound (4) for $\mathbf{z}_n^i$. The desired pointwise bound (4) for $\mathbf{y}_n^i$ follows completely analogously by repeating the steps of the above proof for the second line of Eqn (16). □

D.2 Proof of Proposition 4.2 of main text.

Proof. We can apply the chain rule repeatedly (for instance as in [32]) to obtain,

$$
\frac{\partial \mathcal{E}_n}{\partial \theta} = \sum_{1 \leq k \leq n} \underbrace{\frac{\partial \mathcal{E}_n}{\partial \mathbf{X}_n}\frac{\partial \mathbf{X}_n}{\partial \mathbf{X}_k}\frac{\partial^+ \mathbf{X}_k}{\partial \theta}}_{\frac{\partial \mathcal{E}_n^{(k)}}{\partial \theta}}. \tag{18}
$$

Here, the notation $\frac{\partial^+ \mathbf{X}_k}{\partial \theta}$ refers to taking the partial derivative of $\mathbf{X}_k$ with respect to the parameter θ, while keeping the other arguments constant.

A straightforward application of the product rule yields,

$$
\frac{\partial \mathbf{X}_n}{\partial \mathbf{X}_k} = \prod_{k < \ell \leq n} \frac{\partial \mathbf{X}_\ell}{\partial \mathbf{X}_{\ell-1}}. \tag{19}
$$

For any $k < \ell \leq n$, a tedious yet straightforward computation yields the following representation formula,

$$
\frac{\partial \mathbf{X}_\ell}{\partial \mathbf{X}_{\ell-1}} = \mathbf{I}_{2d \times 2d} + \Delta t \mathbf{E}^{\ell,\ell-1} + \Delta t^2 \mathbf{F}^{\ell,\ell-1}. \tag{20}
$$

Here $\mathbf{E}^{\ell,\ell-1} \in \mathbb{R}^{2d\times 2d}$ is a matrix whose entries are given below. For any $1 \le i \le d$, we have,

$$\begin{aligned}
\mathbf{E}^{\ell,\ell-1}_{2i-1,2j-1} &\equiv 0, \quad j \neq i \\
\mathbf{E}^{\ell,\ell-1}_{2i-1,2i-1} &= -\hat{\sigma}(\mathbf{A}^i_{\ell-1}), \\
\mathbf{E}^{\ell,\ell-1}_{2i-1,2j} &= (\mathbf{W}_1)_{i,j}\hat{\sigma}'(\mathbf{A}^i_{\ell-1})\left(\sigma(\mathbf{C}^i_{\ell-1}) - \mathbf{z}^i_{\ell-1}\right) + (\mathbf{W}_z)_{i,j}\hat{\sigma}'(\mathbf{A}^i_{\ell-1})\sigma'(\mathbf{C}^i_{\ell-1}),\ \forall 1 \le j \le d \\
\mathbf{E}^{\ell,\ell-1}_{2i,2j-1} &= (\mathbf{W}_y)_{i,j}\hat{\sigma}(\mathbf{B}^i_{\ell-1})\sigma'(\mathbf{D}^i_{\ell}),\ \forall 1 \le j \le d \\
\mathbf{E}^{\ell,\ell-1}_{2i,2j} &= (\mathbf{W}_2)_{i,j}\hat{\sigma}'(\mathbf{B}^i_{\ell-1})\left(\sigma(\mathbf{D}^i_{\ell}) - \mathbf{y}^i_{\ell-1}\right), \quad j \neq i \\
\mathbf{E}^{\ell,\ell-1}_{2i,2i} &= -\hat{\sigma}(\mathbf{B}^i_{\ell-1}) + (\mathbf{W}_2)_{i,i}\hat{\sigma}'(\mathbf{B}^i_{\ell-1})\left(\sigma(\mathbf{D}^i_{\ell}) - \mathbf{y}^i_{\ell-1}\right).
\end{aligned} \tag{21}$$

Similarly, $\mathbf{F}^{\ell,\ell-1} \in \mathbb{R}^{2d\times 2d}$ is a matrix whose entries are given below. For any $1 \le i \le d$, we have,

$$\begin{aligned}
\mathbf{F}^{\ell,\ell-1}_{2i-1,j} &\equiv 0, \quad \forall\ 1 \le j \le 2d, \\
\mathbf{F}^{\ell,\ell-1}_{2i,2j-1} &= -(\mathbf{W}_y)_{i,j}\hat{\sigma}(\mathbf{A}^i_{\ell-1})\hat{\sigma}(\mathbf{B}^i_{\ell-1})\sigma'(\mathbf{A}^i_{\ell-1}), \quad 1 \le j \le d, \\
\mathbf{F}^{\ell,\ell-1}_{2i,2j} &= \hat{\sigma}(\mathbf{B}^i_{\ell-1})\sigma'(\mathbf{D}^i_{\ell})\sum_{\lambda=1}^{d}(\mathbf{W}_y)_{i,\lambda}\left(\left(\sigma(\mathbf{C}^\lambda_{\ell-1}) - \mathbf{z}^\lambda_{\ell-1}\right)\hat{\sigma}'(\mathbf{A}^\lambda_{\ell-1})(\mathbf{W}_1)_{\lambda,j} + \hat{\sigma}(\mathbf{A}^\lambda_{\ell-1})\sigma'(\mathbf{C}^\lambda_{\ell-1})(\mathbf{W}_z)_{\lambda,j}\right).
\end{aligned} \tag{22}$$

Using the fact that,

$$\sup_{x\in\mathbb{R}} \max\left\{|\sigma(x)|, |\sigma'(x)|, |\hat{\sigma}(x)|, |\hat{\sigma}'(x)\right\} \le 1,$$

the pointwise bounds (4), $t_n < 1$, the definition of η (6) and the definition of matrix norms, we obtain that,

$$\|\mathbf{E}^{\ell,\ell-1}\|_\infty \le \max\left\{1 + \|\mathbf{W}_z\|_\infty + 2\|\mathbf{W}_1\|_\infty, 1 + \|\mathbf{W}_y\|_\infty + 2\|\mathbf{W}_2\|_\infty\right\} \le 1 + 3\eta. \tag{23}$$

By similar calculations, we obtain,

$$\|\mathbf{F}^{\ell,\ell-1}\|_\infty \le \|\mathbf{W}_y\|_\infty\left(1 + 2\|\mathbf{W}_1\|_\infty + \|\mathbf{W}_z\|_\infty\right) \le \eta + 3\eta^2. \tag{24}$$

Applying (23) and (24) in the representation formula (20), we obtain.

$$\begin{aligned}
\left\|\frac{\partial \mathbf{X}_\ell}{\partial \mathbf{X}_{\ell-1}}\right\|_\infty &\le 1 + (1+3\eta)\Delta t + \eta(1+3\eta)\Delta t^2, \\
&\le 1 + (1+3\eta)\Delta t + \frac{((1+3\eta)\Delta t)^2}{2}.
\end{aligned}$$

Denote $\Delta\tau = (1+3\eta)\Delta t$ and using the expression (19) with the above inequality yields,

$$\left\|\frac{\partial \mathbf{X}_n}{\partial \mathbf{X}_k}\right\|_\infty \le \left(1 + \Delta\tau + \frac{\Delta\tau^2}{2}\right)^{n-k}. \tag{25}$$

Next, we choose Δt small enough such that $\Delta\tau << 1$ and the following holds,

$$\left(1 + \Delta\tau + \frac{\Delta\tau^2}{2}\right)^{n-k} \le 1 + 2(n-k)\Delta\tau, \tag{26}$$

for any $1 \le k < n \le N$.

Hence applying (26) in (25), we obtain,

$$\left\|\frac{\partial \mathbf{X}_n}{\partial \mathbf{X}_k}\right\|_\infty \le 1 + 2(n-k)\Delta\tau. \tag{27}$$

For the sake of definiteness, we fix any $1 \leq \alpha, \beta \leq d$ and set $\theta = (\mathbf{W}_y)_{\alpha,\beta}$ in the following. The following bounds for any other choice of $\theta \in \Theta$ can be derived analogously. Given this, it is straightforward to calculate from the structure of LEM (3) that entries of the vector $\frac{\partial^+ \mathbf{X}_k}{\partial (\mathbf{W}_z)_{\alpha,\beta}}$ are given by,

$$\begin{aligned} \left(\frac{\partial^+ \mathbf{X}_k}{\partial (\mathbf{W}_z)_{\alpha,\beta}}\right)_j &\equiv 0, \quad \forall\ j \neq 2\alpha, \\ \left(\frac{\partial^+ \mathbf{X}_k}{\partial (\mathbf{W}_z)_{\alpha,\beta}}\right)_{2\alpha} &= \Delta t \hat{\sigma}(\mathbf{B}^\alpha_{k-1}) \sigma'(\mathbf{D}^\alpha_k) \mathbf{z}^\beta_k. \end{aligned} \tag{28}$$

Hence, by the pointwise bounds (4), we obtain from (28) that

$$\left\| \frac{\partial^+ \mathbf{X}_k}{\partial (\mathbf{W}_z)_{\alpha,\beta}} \right\|_\infty \leq \Delta t \sqrt{k(\Delta t + 2\Delta t^2)} \leq \Delta t \sqrt{3k\Delta t}. \tag{29}$$

Finally, it is straightforward to calculate from the loss function (5) that

$$\frac{\partial \mathcal{E}_n}{\partial \mathbf{X}_n} = \left[0, \mathbf{y}^1_n - \bar{\mathbf{y}}^1, \ldots\ldots, 0, \mathbf{y}^d_n - \bar{\mathbf{y}}^d\right]. \tag{30}$$

Therefore, using the pointwise bounds (4), we obtain

$$\left\| \frac{\partial \mathcal{E}_n}{\partial \mathbf{X}_n} \right\| \leq \hat{\mathbf{X}} + \sqrt{3n\Delta t}. \tag{31}$$

Applying (27), (29) and (31) in the definition (18) yields,

$$\left| \frac{\partial \mathcal{E}^{(k)}_n}{\partial (\mathbf{W}_z)_{\alpha,\beta}} \right| \leq \Delta t \sqrt{3k\Delta t} (\hat{\mathbf{X}} + \sqrt{3n\Delta t}) \left(1 + 2(n-k)\Delta\tau\right). \tag{32}$$

Substituting $t_n = n\Delta t \leq 1$, $1 \leq k \leq n$ in (32) yields,

$$\begin{aligned} \left| \frac{\partial \mathcal{E}^{(k)}_n}{\partial (\mathbf{W}_z)_{\alpha,\beta}} \right| &\leq \Delta t \sqrt{3k\Delta t} (\hat{\mathbf{X}} + \sqrt{3n\Delta t}) \left(1 + 2(n-k)\Delta\tau\right), \\ &\leq \Delta t \sqrt{3n\Delta t} (\hat{\mathbf{X}} + \sqrt{3n\Delta t}) \left(1 + 2n\Delta\tau\right) \\ &\Delta t \sqrt{3t_n} (\hat{\mathbf{X}} + \sqrt{3t_n}) \left(1 + 2n\Delta\tau\right) \\ &\leq \sqrt{3} \left(\sqrt{3} + \hat{\mathbf{X}}\right) (\Delta t + 2\Delta\tau). \end{aligned}$$

Recalling that $\Delta\tau = (1 + 3\eta)\Delta t$ in the above expression leads to,

$$\left| \frac{\partial \mathcal{E}^{(k)}_n}{\partial (\mathbf{W}_z)_{\alpha,\beta}} \right| \leq \left(3 + \sqrt{3}\hat{\mathbf{X}}\right) (3 + 6\eta)\, \Delta t, \quad \forall 1 \leq k \leq n. \tag{33}$$

Applying (33) in (18) leads to,

$$\begin{aligned} \left| \frac{\partial \mathcal{E}_n}{\partial (\mathbf{W}_z)_{\alpha,\beta}} \right| &\leq \sum_{k=1}^n \left| \frac{\partial \mathcal{E}^{(k)}_n}{\partial (\mathbf{W}_z)_{\alpha,\beta}} \right| \\ &\leq \left(3 + \sqrt{3}\hat{\mathbf{X}}\right) (3 + 6\eta) \sum_{k=1}^n \Delta t, \\ &\leq \left(3 + \sqrt{3}\hat{\mathbf{X}}\right) (3 + 6\eta)\, t_n. \end{aligned} \tag{34}$$

Finally, recalling that $t_N = N\Delta t \leq 1$, we obtain from (34) that,

$$\left| \frac{\partial \mathcal{E}}{\partial (\mathbf{W}_z)_{\alpha,\beta}} \right| \leq \left(3 + \sqrt{3}\hat{\mathbf{X}}\right) (3 + 6\eta), \tag{35}$$

which is the desired bound (6) for the scalar parameter $\theta = (\mathbf{W}_y)_{\alpha,\beta}$. □

Remark D.1. *The upper bound* (6) *holds under the assumption that the* time step Δt *is* sufficiently small. *Apart from the assumption that* $\Delta t \leq \frac{1}{2}$*, which suffices for the derivation of the pointwise bounds* (4)*, the derivation of the gradient upper bound* (6)*, used the fact that* $T = t_N = n\Delta t = 1$*. This assumption is simply a convenient normalization and we can readily relax it leading to a gradient upper bound of the form;*

$$\left|\frac{\partial \mathcal{E}}{\partial \theta}\right| \leq \left(3T + \sqrt{3T}\hat{\mathbf{X}}\right)(1 + 2T(1+3\eta))\,T. \tag{36}$$

Another source for the smallness requirement on Δt *stems from the assumption that* Δt *be small enough such that the bound* (26) *will hold. One can also relax this smallness assumption on* Δt *by recognizing that*

$$\left(1 + \Delta\tau + \frac{\Delta\tau^2}{2}\right)^{n-k} \leq \left(e^{\Delta\tau}\right)^{n-k} = e^{(n-k)\Delta\tau}.$$

Note that the above bound holds for any value of $\Delta t > 0$*. One can apply this bound and repeat all the calculations from* (32)-(35) *to conclude that the bound on the gradient (with* $t_N = T = 1$*) is,*

$$\left|\frac{\partial \mathcal{E}}{\partial \theta}\right| \leq \left(3 + \sqrt{3}\hat{\mathbf{X}}\right)\left(1 + e^{1+3\eta}\right). \tag{37}$$

D.3 Proof of Proposition 4.3 of main text.

To mitigate the vanishing gradient problem, we need to obtain a more precise characterization of the gradient $\frac{\partial \mathcal{E}_n^{(k)}}{\partial \theta}$ defined in (18). To this end, we introduce the following *order*-notation for matrices,

$$\begin{aligned} &\beta = \mathcal{O}(\alpha), \text{for } \alpha, \beta \in \mathbb{R}_+ \quad \text{if there exists constants } \overline{C}, \underline{C} \text{ such that } \underline{C}\alpha \leq \beta \leq \overline{C}\alpha. \\ &\mathbf{M} = \mathcal{O}(\alpha), \text{for } \mathbf{M} \in \mathbb{R}^{d_1 \times d_2}, \alpha \in \mathbb{R}_+ \quad \text{if there exists constant } \overline{C} \text{ such that } \|\mathbf{M}\| \leq \overline{C}\alpha. \end{aligned} \tag{38}$$

For the sake of definiteness, we fix any $1 \leq \alpha, \beta \leq d$ and set $\theta = (\mathbf{W}_y)_{\alpha,\beta}$ in the following. The following formulas for any other choice of $\theta \in \Theta$ can be derived analogously. Moreover, for simplicity of notation, we set the target function $\bar{\mathbf{X}}_n \equiv 0$.

Proposition 4.3 is a straightforward corollary of the following,

Proposition D.2. *Let* $\mathbf{y}_n, \mathbf{z}_n$ *be the hidden states generated by LEM* (3)*, then we have the following representation formula for the hidden state gradient,*

$$\begin{aligned} \frac{\partial \mathcal{E}_n^{(k)}}{\partial \theta} &= \Delta t \hat{\sigma}(\mathbf{B}_{k-1}^{\alpha}) \sigma'(\mathbf{D}_k^{\alpha}) \mathbf{z}_k^{\beta} \left(\mathbf{y}_n^{\alpha} - \bar{\mathbf{y}}_n^{\alpha}\right) \\ &+ \Delta t^2 \hat{\sigma}(\mathbf{B}_{k-1}^{\alpha}) \sigma'(\mathbf{D}_k^{\alpha}) \mathbf{z}_k^{\beta} \left[\sum_{j=1}^{d} \left(\mathbf{y}_j^n - \bar{\mathbf{y}}_j^n\right) \sum_{\ell=k+1}^{n} \hat{\sigma}'(\mathbf{B}_{\ell-1}^{j}) \left(\sigma(\mathbf{D}_{\ell}^{j}) - \mathbf{y}_{\ell-1}^{j}\right) (\mathbf{W}_2)_{j,2\alpha}\right] \\ &+ \Delta t^2 \hat{\sigma}(\mathbf{B}_{k-1}^{\alpha}) \sigma'(\mathbf{D}_k^{\alpha}) \mathbf{z}_k^{\beta} \left[\sum_{\ell=k+1}^{n} \hat{\sigma}(\mathbf{B}_{\ell-1}^{\alpha}) \left(\mathbf{y}_{\alpha}^n - \bar{\mathbf{y}}_{\alpha}^n\right)\right] + \mathcal{O}(\Delta t^3). \end{aligned} \tag{39}$$

Here, the constants in $\mathcal{O}$ *could depend on* η *defined in* (6) *(main text).*

Proof. The starting point for deriving an asymptotic formula for the hidden state gradient $\frac{\partial \mathcal{E}_n^{(k)}}{\partial \theta}$ is to observe from the representation formula (20), the bound (24) on matrices $\mathbf{F}^{\ell,\ell-1}$ and the order notation (38) that,

$$\frac{\partial \mathbf{X}_\ell}{\partial \mathbf{X}_{\ell-1}} = \mathbf{I}_{2d\times 2d} + \Delta t \mathbf{E}^{\ell,\ell-1} + \mathcal{O}(\Delta t^2), \tag{40}$$

as long as η is independent of Δt.

By using induction and the bounds (23),(24), it is straightforward to calculate the following representation formula for the product,

$$\frac{\partial \mathbf{X}_n}{\partial \mathbf{X}_k} = \prod_{k<\ell\leq n} \frac{\partial \mathbf{X}_\ell}{\partial \mathbf{X}_{\ell-1}} = \mathbf{I}_{2d\times 2d} + \Delta t \sum_{\ell=k+1}^{n} \mathbf{E}^{\ell,\ell-1} + \mathcal{O}(\Delta t^2). \tag{41}$$

Recall that we have set $\theta = (\mathbf{W}_z)_{\alpha,\beta}$. Hence, by the expressions (30) and (28), a direct but tedious calculation leads to,

$$\frac{\partial \mathcal{E}_n}{\partial \mathbf{X}_n}\mathbf{I}_{2d\times 2d}\frac{\partial^+ \mathbf{X}_k}{\partial \theta} = \Delta t\hat{\sigma}(\mathbf{B}^\alpha_{k-1})\sigma'(\mathbf{D}^\alpha_k)\mathbf{z}^\beta_k\left(\mathbf{y}^\alpha_n - \bar{\mathbf{y}}^\alpha_n\right), \tag{42}$$

$$\sum_{\ell=k+1}^{n}\frac{\partial \mathcal{E}_n}{\partial \mathbf{X}_n}\mathbf{E}^{\ell,\ell-1}\frac{\partial^+ \mathbf{X}_k}{\partial \theta} = \tag{43}$$

$$\Delta t\hat{\sigma}(\mathbf{B}^\alpha_{k-1})\sigma'(\mathbf{D}^\alpha_k)\mathbf{z}^\beta_k\left[\sum_{j=1}^{d}\left(\mathbf{y}^n_j - \bar{\mathbf{y}}^n_j\right)\sum_{\ell=k+1}^{n}\hat{\sigma}'(\mathbf{B}^j_{\ell-1})\left(\sigma(\mathbf{D}^j_\ell) - \mathbf{y}^j_{\ell-1}\right)(\mathbf{W}_2)_{j,2\alpha} - \sum_{\ell=k+1}^{n}\hat{\sigma}(\mathbf{B}^\alpha_{\ell-1})\left(\mathbf{y}^n_\alpha - \bar{\mathbf{y}}^n_\alpha\right)\right]. \tag{44}$$

Therefore, by substituting the above expression into the representation formula (41) yields the desired formula (39).

In order to prove the formula (7) (see Proposition 4.3 of main text), we focus our interest on long-term dependencies i.e., $k << n$. Then, a closer perusal of the expression in (42), together with the pointwise bounds (4) which implies that $\mathbf{y}_{k-1} \approx O(\sqrt{\Delta t})$, results in the following,

$$\frac{\partial \mathcal{E}_n}{\partial \mathbf{X}_n}\mathbf{I}_{2d\times 2d}\frac{\partial^+ \mathbf{X}_k}{\partial \theta} = O\left(\Delta t^{\frac{3}{2}}\right). \tag{45}$$

Similarly, we also obtain,

$$\Delta t\sum_{\ell=k+1}^{n}\frac{\partial \mathcal{E}_n}{\partial \mathbf{X}_n}\mathbf{E}^{\ell,\ell-1}\frac{\partial^+ \mathbf{X}_k}{\partial \theta} = O\left(\Delta t^{\frac{3}{2}}\right). \tag{46}$$

Combining (45) and (46) results in the desired asymptotic bound (7). □

D.4 Proof of Proposition 4.4

Proof. To prove this proposition, we have to construct hidden states $\mathbf{y}_n, \mathbf{z}_n$, output state ω_n, weight matrices $\mathbf{W}_{1,2,y,z}, \mathcal{W}_{y,z}, \mathbf{V}_{1,2,y,z}$ and bias vectors $\mathbf{b}_{1,2,y,z}$ such that LEM (3) with output state $\omega_n = \mathcal{W}_y\mathbf{y}_n$ approximates the dynamical system (8).

Let $\mathbb{R}^* > R >> 1$ and $\epsilon^* < \epsilon$, be parameters to be defined later. By the theorem for universal approximation of continuous functions with neural networks with the tanh activation function $\sigma = \tanh$ [5], given ϵ^*, there exist weight matrices $W_1 \in \mathbb{R}^{d_1\times d_h}, V_1 \in \mathbb{R}^{d_1\times d_h}, W_2 \in \mathbb{R}^{d_h\times d_1}$ and bias vector $b_1 \in \mathbb{R}^{d_1}$ such that the tanh neural network defined by,

$$\mathcal{N}_1(h,u) = W_2\sigma\left(W_1h + V_1u + b_1\right), \tag{47}$$

approximates the underlying function $\mathbf{f}$ in the following manner,

$$\max_{\max(\|h\|,\|u\|)<R^*}\|\mathbf{f}(h,u) - \mathcal{N}_1(h,u)\| \leq \epsilon^*. \tag{48}$$

Similarly, one can readily approximate the identity function $\mathbf{g}(h,u) = h$ with a tanh neural network of the form,

$$\bar{\mathcal{N}}_2(h) = \bar{W}_2\sigma\left(\bar{W}_1h\right), \tag{49}$$

such that

$$\max_{\|h\|,\|u\|<R^*}\|\mathbf{g}(h) - \mathcal{N}_2(h)\| \leq \epsilon^*. \tag{50}$$

Next, we define the following dynamical system,

$$\begin{aligned}\bar{\mathbf{z}}_n &= W_2\sigma\left(W_1\bar{\mathbf{y}}_{n-1} + V_1\mathbf{u}_n + b_1\right),\\ \bar{\mathbf{y}}_n &= \bar{W}_2\sigma\left(\bar{W}_1\bar{\mathbf{z}}_n\right),\end{aligned} \tag{51}$$

with initial states $\bar{\mathbf{z}}_0 = \bar{\mathbf{y}}_0 = 0$.

Using the approximation bound (48), we derive the following bound,

$$\begin{aligned}\|\boldsymbol{\phi}_n - \bar{\mathbf{y}}_n\| &= \|\mathbf{f}\left(\boldsymbol{\phi}_{n-1}, \mathbf{u}_n\right) - \bar{\mathbf{z}}_n + \bar{\mathbf{z}}_n - \bar{\mathbf{y}}_n\| \\ &\leq \|\mathbf{f}\left(\boldsymbol{\phi}_{n-1}, \mathbf{u}_n\right) - W_2\sigma\left(W_1\bar{\mathbf{y}}_{n-1} + V_1\mathbf{u}_n + b_1\right)\| + \|\mathbf{g}(\bar{\mathbf{z}}_n) - \bar{W}_2\sigma\left(\bar{W}_1\bar{\mathbf{z}}_n\right)\| \\ &\leq \|\mathbf{f}\left(\boldsymbol{\phi}_{n-1}, \mathbf{u}_n\right) - \mathbf{f}\left(\bar{\mathbf{y}}_{n-1}, \mathbf{u}_n\right)\| + \|\mathbf{f}\left(\bar{\mathbf{y}}_{n-1}, \mathbf{u}_n\right) - W_2\sigma\left(W_1\bar{\mathbf{y}}_{n-1} + V_1\mathbf{u}_n + b_1\right)\| \\ &+ \|\mathbf{g}(\bar{\mathbf{z}}_n) - \bar{W}_2\sigma\left(\bar{W}_1\bar{\mathbf{z}}_n\right)\| \\ &\leq \mathrm{Lip}(\mathbf{f})\|\boldsymbol{\phi}_n - \bar{\mathbf{y}}_n\| + 2\epsilon^* \quad \text{(from (48), (50))}.\end{aligned}$$

Here, $\mathrm{Lip}(\mathbf{f})$ is the Lipschitz constant of the function $\mathbf{f}$ on the compact set $\{(h,u) \in \mathbb{R}^{d_h \times d_u} : \|h\|, \|u\| < \mathbb{R}^*\}$. Note that one can readily prove using the fact that $\bar{\mathbf{y}}_0 = \bar{\mathbf{z}}_0 = 0$, bounds (48), (50) and the assumption $\|\boldsymbol{\phi}_n\|, \|\mathbf{u}_n\| < R$, that $\|\bar{\mathbf{z}}_n\|, \|\bar{\mathbf{y}}_n\| < R^* = 2R$.

Iterating the above inequality over n leads to the bound,

$$\|\boldsymbol{\phi}_n - \bar{\mathbf{y}}_n\| \leq 2\left(\mathrm{Lip}(\mathbf{f})\right)^n \epsilon^*. \tag{52}$$

Hence, using the Lipschitz continuity of the output function $\mathbf{o}$ in (8), one obtains,

$$\|\mathbf{o}_n - \mathbf{o}(\bar{\mathbf{y}}_n)\| \leq 2(\mathrm{Lip}(\mathbf{o}))\left(\mathrm{Lip}(\mathbf{f})\right)^n \epsilon^*, \tag{53}$$

with $\mathrm{Lip}(\mathbf{o})$ being the Lipshitz constant of the function $\mathbf{o}$ on the compact set $\{h \in \mathbb{R}^{d_h} : \|h\| < \mathbb{R}^*\}$.

Next, we can use the universal approximation theorem for neural networks again to conclude that given a tolerance $\bar{\epsilon}$, there exist weight matrices $W_3 \in \mathbb{R}^{d_2 \times d_h}, W_4 \in \mathbb{R}^{d_h \times d_2}$ and bias vector $b_2 \in \mathbb{R}^{d_2}$ such that the tanh neural network defined by,

$$\mathcal{N}_3(h) = W_4\sigma\left(W_3 h + b_2\right), \tag{54}$$

approximates the underlying output function $\mathbf{o}$ in the following manner,

$$\max_{\|h\| < R^*} \|\mathbf{o}(h) - \mathcal{N}_3(h)\| \leq \bar{\epsilon}. \tag{55}$$

Now defining,

$$\bar{\omega}_n = W_4\sigma\left(W_3\bar{\mathbf{y}}_n + b_2\right), \tag{56}$$

we obtain from (55) and (53) that,

$$\|\mathbf{o}_n - \bar{\omega}_n\| \leq \bar{\epsilon} + 2(\mathrm{Lip}(\mathbf{o}))\left(\mathrm{Lip}(\mathbf{f})\right)^n \epsilon^*. \tag{57}$$

Next, we introduce the notation,

$$\tilde{\mathbf{z}}_n = \sigma\left(W_1\bar{\mathbf{y}}_{n-1} + V_1\mathbf{u}_n + b_1\right), \quad \tilde{\mathbf{y}}_n = \sigma\left(\bar{W}_1\bar{\mathbf{z}}_n\right). \tag{58}$$

From (51), we see that

$$\bar{\mathbf{z}}_n = W_2\tilde{\mathbf{z}}_n, \quad \bar{\mathbf{y}}_n = \bar{W}_2\tilde{\mathbf{y}}_n \tag{59}$$

Thus from (59) and (57), we have

$$\begin{aligned}\bar{\omega}_n &= W_4\sigma\left(W_3W_2\tilde{\mathbf{y}}_n + b_2\right), \\ &= W_4\sigma\left(W_3W_2\sigma\left(\bar{W}_1W_2\tilde{\mathbf{z}}_n\right) + b_2\right).\end{aligned} \tag{60}$$

Define the function $\mathcal{R} : \mathbb{R}^{d_h} \times \mathbb{R}^{d_u} \mapsto \mathbb{R}^{d_o}$ by,

$$\mathcal{R}(z) = W_4\sigma\left(W_3W_2\sigma\left(\bar{W}_1W_2z\right) + b_2\right). \tag{61}$$

The function, defined above, is clearly Lipschitz continuous. We can apply the universal approximation theorem for tanh neural networks to find, for any given tolerance $\tilde{\epsilon}$, weight matrices $W_5 \in \mathbb{R}^{d_3 \times d_4}, W_6 \in \mathbb{R}^{d_o \times d_3}, V_2 \in \mathbb{R}^{d_3 \times d_u}$ and bias vector $b_3 \in \mathbb{R}^{d_3}$ such that the following holds,

$$\max_{\max(\|z\|) < R^*} \|\mathcal{R}(z) - W_6\sigma(W_5z + b_3)\| \leq \tilde{\epsilon}. \tag{62}$$

Denote $\tilde{\omega}_n = W_6\sigma(W_5\tilde{\mathbf{z}}_n + b_3)$, then from (62) and (60), we obtain that

$$\|\bar{\omega}_n - \tilde{\omega}_n\| \leq \tilde{\epsilon}.$$

Combining this estimate with (57) yields,

$$\|\mathbf{o}_n - \tilde{\omega}_n\| \leq \tilde{\epsilon} + \bar{\epsilon} + 2(\mathrm{Lip}(\mathbf{o}))\,(\mathrm{Lip}(\mathbf{f}))^n\,\epsilon^*. \tag{63}$$

Now, we collect all the ingredient to define the LEM that can approximate the dynamical system (8). To this end, we define hidden states $\mathbf{z}_n, \mathbf{y}_n \in \mathbb{R}^{2d_h}$ as

$$\mathbf{z}_n = [\tilde{\mathbf{z}}_n, \hat{\mathbf{z}}_n]\,, \quad \mathbf{y}_n = [\tilde{\mathbf{y}}_n, \hat{\mathbf{y}}_n]\,,$$

with $\tilde{\mathbf{z}}_n, \hat{\mathbf{z}}_n, \tilde{\mathbf{y}}_n, \hat{\mathbf{y}}_n \in \mathbb{R}^{d_h}$. These hidden states are evolved by the dynamical system,

$$\begin{aligned}
\mathbf{z}_n^{\perp} &= \sigma\left(\begin{bmatrix} W_1\bar{W}_2 & 0 \\ 0 & 0 \end{bmatrix}\mathbf{y}_{n-1}^{\perp} + [V_1\mathbf{u}_n, 0]^{\perp} + [b_1, 0]^{\perp}\right), \\
\mathbf{y}_n^{\perp} &= \sigma\left(\begin{bmatrix} \bar{W}_1W_2 & 0 \\ W_5 & 0 \end{bmatrix}\mathbf{z}_n^{\perp} + [0, 0]^{\perp} + [0, b_3]^{\perp}\right)
\end{aligned} \tag{64}$$

and the output state is calculated by,

$$\omega_n^{\perp} = [0, W_6]\mathbf{y}_n^{\perp}. \tag{65}$$

Finally, we can recast the dynamical system (64), (65) as a LEM of the form (3) for the hidden states $\mathbf{y}_n, \mathbf{z}_n$, defined in (64), with the following parameters, Now, define the hidden states $\bar{\mathbf{y}}_n, \bar{\mathbf{z}}_n \in \mathbb{R}^{d_h}$ for all $1 \leq n \leq N$ by the LEM (3) with the following parameters,

$$\begin{aligned}
&\Delta t = 1, \quad d_y = 2d_h, \\
&\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{V}_1 = \mathbf{V}_2 = 0 \\
&\mathbf{b}_1 = \mathbf{b}_2 = \mathbf{b}_\infty, \\
&\mathbf{W}_z = \begin{bmatrix} W_1\bar{W}_2 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{V}_z = [V_1, 0], \quad \mathbf{b}_z = [b_1, 0] \\
&\mathbf{W}_y = \begin{bmatrix} \bar{W}_1W_2 & 0 \\ W_5 & 0 \end{bmatrix}, \quad \mathbf{V}_y = 0, \quad \mathbf{b}_z = [0, b_3]. \\
&\mathcal{W}_y = [0, W_6].
\end{aligned} \tag{66}$$

Here, $\mathbf{b}_\infty \in \mathbb{R}^{d_h}$ is defined as

$$\mathbf{b}_\infty = [b_\infty, b_\infty, \ldots, \ldots, b_\infty],$$

with $1 << b_\infty$ is such that

$$|1 - \hat{\sigma}(b_\infty)| \leq \delta. \tag{67}$$

The nature of the sigmoid function guarantees the existence of such a b_∞ for any δ. As δ decays exponentially fast, we set it to 0 in the following for notational simplicity.

It is straightforward to verify that the output state of the LEM (3) with parameters given in (66) is $\omega_n = \tilde{\omega}_n$.

Therefore, from (63) and by setting $\bar{\epsilon} < \frac{\epsilon}{3}$, $\tilde{\epsilon} < \frac{\epsilon}{3}$ and

$$\epsilon^* < \frac{\epsilon}{6\mathrm{Lip}(\mathbf{g})(\mathrm{Lip}(\mathbf{f}))^N},$$

we prove the desired bound (9).

□

D.5 Proof of Proposition 4.5

Proof. The proof of this proposition is based heavily on the proof of Proposition 4.4. Hence, we will highlight the main points of difference.

As the steps for approximation of a general Lipschitz continuous output map are identical to the corresponding steps in the proof of proposition 4.4 (see the steps from Eqns. (53) to (63)), we will only consider the following linear output map for convenience herein,

$$\mathbf{o}(\boldsymbol{\psi}_n) = \mathcal{W}_c \boldsymbol{\psi}_n. \tag{68}$$

Let $\mathbb{R}^* > R >> 1$ and $\epsilon^* < \epsilon$, be parameters to be defined later. By the theorem for universal approximation of continuous functions with neural networks with the tanh activation function $\sigma = \tanh$, given ϵ^*, there exist weight matrices $W_1^f, W_2^f \in \mathbb{R}^{d_1 \times d_h}, V_1^f \in \mathbb{R}^{d_1 \times d_u}, W_3^f \in \mathbb{R}^{d_h \times d_1}$ and bias vector $b_1^f \in \mathbb{R}^{d_1}$ such that the tanh neural network defined by,

$$\mathcal{N}_f(h, c, u) = W_3^f \sigma\left(W_1^f h + W_2^f c + V_1^f u + b_1^f\right), \tag{69}$$

approximates the underlying function $\mathbf{f}$ in the following manner,

$$\max_{\max(\|h\|, \|c\|, \|u\|) < R^*} \|\mathbf{f}(h, c, u) - \mathcal{N}_f(h, c, u)\| \leq \epsilon^*. \tag{70}$$

Next, we define the following map,

$$\mathbf{G}(h, c, u) = \mathbf{g}(h, c, u) - \left(1 - \frac{1}{\tau}\right) c, \tag{71}$$

for any $\tau > 0$.

By the universal approximation theorem, given ϵ^*, there exist weight matrices $W_1^g, W_2^g \in \mathbb{R}^{d_2 \times d_h}, V_2^f \in \mathbb{R}^{d_2 \times d_u}, W_3^g \in \mathbb{R}^{d_h \times d_2}$ and bias vector $b_1^g \in \mathbb{R}^{d_2}$ such that the tanh neural network defined by,

$$\mathcal{N}_g(h, c, u) = W_3^g \sigma\left(W_1^g h + W_2^g c + V_1^g u + b_1^g\right), \tag{72}$$

approximates the function $\mathbf{G}$ (71) in the following manner,

$$\max_{\max(\|h\|, \|c\|, \|u\|) < R^*} \|\mathbf{f}(h, c, u) - \mathcal{N}_f(h, c, u)\| \leq \epsilon^*. \tag{73}$$

Note that the sizes of the neural network $\mathcal{N}_g$ can be made independent of the small parameter τ by simply taking the difference of the neural networks approximating the functions g and $\hat{g}(h, c, u) = c$ with Tanh neural networks. As neither of these functions depend on the small parameter τ, the sizes of the corresponding neural networks are independent of the small parameter too.

Next, as in the proof of proposition 4.4, one can readily approximate the identity function $\hat{f}(h, c, u) = h$ with a tanh neural network of the form,

$$\bar{\mathcal{N}}_f(h) = \bar{W}_2 \sigma\left(\bar{W}_1 h\right), \tag{74}$$

such that

$$\max_{\|h\|, \|c\|, \|u\| < R^*} \|\hat{f}(h, c, u) - \mathcal{N}_f(h)\| \leq \epsilon^*, \tag{75}$$

and with the same weights and biases, one can approximate the identity function $\hat{g}(h, c, u) = c$ with the Tanh neural network,

$$\bar{\mathcal{N}}_g(c) = \bar{W}_2 \sigma\left(\bar{W}_1 c\right), \tag{76}$$

such that

$$\max_{\|h\|, \|c\|, \|u\| < R^*} \|\hat{g}(h, c, u) - \mathcal{N}_g(c)\| \leq \epsilon^*. \tag{77}$$

Next, we define the following dynamical system,

$$\begin{aligned}
\hat{\mathbf{z}}_n &= W_3^f\,\sigma\left(W_1^f\,\tilde{\mathbf{y}}_{n-1} + W_2^f\,\hat{\mathbf{y}}_{n-1} + V_1^f\,\mathbf{u}_n + b_1^f\right),\\
\tilde{\mathbf{z}}_n &= \bar{W}_2\sigma\left(\bar{W}_1\hat{\mathbf{y}}_{n-1}\right),\\
\hat{\mathbf{y}}_n &= (1-\tau)\hat{\mathbf{y}}_{n-1} + \tau W_3^g\sigma\left(W_1^g\hat{\mathbf{z}}_n + W_2^g\tilde{\mathbf{z}}_n + V_1^g\mathbf{u}_n + b_1^g\right),\\
\tilde{\mathbf{y}}_n &= \bar{W}_2\sigma\left(\bar{W}_1\hat{\mathbf{z}}_n\right),
\end{aligned} \tag{78}$$

with hidden states $\hat{\mathbf{z}}_n, \tilde{\mathbf{z}}_n, \hat{\mathbf{y}}_n, \tilde{\mathbf{y}}_n \in \mathbb{R}^{d_h}$ and with initial states $\hat{\mathbf{z}}_0 = \tilde{\mathbf{z}}_0 = \hat{\mathbf{y}}_0 = \tilde{\mathbf{y}}_0 = 0$.

We derive the following bounds,

$$\begin{aligned}
\|\boldsymbol{\phi}_n - \hat{\mathbf{z}}_n\| &= \|\mathbf{f}(\boldsymbol{\phi}_{n-1}, \boldsymbol{\psi}_{n-1}, \mathbf{u}_n) - W_3^f\,\sigma\left(W_1^f\,\tilde{\mathbf{y}}_{n-1} + W_2^f\,\hat{\mathbf{y}}_{n-1} + V_1^f\,\mathbf{u}_n + b_1^f\right)\|\\
&\le \|\mathbf{f}(\boldsymbol{\phi}_{n-1}, \boldsymbol{\psi}_{n-1}, \mathbf{u}_n) - \mathbf{f}(\tilde{\mathbf{y}}_{n-1}, \hat{\mathbf{z}}_{n-1}, \mathbf{u}_n)\|,\\
&+ \|\mathbf{f}(\tilde{\mathbf{y}}_{n-1}, \hat{\mathbf{z}}_{n-1}, \mathbf{u}_n) - W_3^f\,\sigma\left(W_1^f\,\tilde{\mathbf{y}}_{n-1} + W_2^f\,\hat{\mathbf{y}}_{n-1} + V_1^f\,\mathbf{u}_n + b_1^f\right)\|\\
&\le \mathrm{Lip}(\mathbf{f})\left(\|\boldsymbol{\phi}_{n-1} - \hat{\mathbf{z}}_{n-1}\| + \|\tilde{\mathbf{y}}_{n-1} - \hat{\mathbf{z}}_{n-1}\| + \|\boldsymbol{\psi}_{n-1} - \bar{\mathbf{y}}_{n-1}\|\right) + \epsilon^* \quad \text{(by (73))}\\
&\le \mathrm{Lip}(\mathbf{f})\left(\|\boldsymbol{\phi}_{n-1} - \hat{\mathbf{z}}_{n-1}\| + \|\boldsymbol{\psi}_{n-1} - \bar{\mathbf{y}}_{n-1}\|\right) + (1 + \mathrm{Lip}(\mathbf{f}))\,\epsilon^* \quad \text{(by (75), (78))},
\end{aligned}$$

and

$$\begin{aligned}
\|\boldsymbol{\psi}_n - \hat{\mathbf{y}}_n\| &= \|(1-\tau)(\boldsymbol{\psi}_{n-1} - \bar{\mathbf{y}}_{n-1}) + \tau\left(\mathbf{G}(\boldsymbol{\phi}_n, \boldsymbol{\psi}_{n-1}, \mathbf{u}_n) - W_3^g\sigma\left(W_2^g\tilde{\mathbf{z}}_n + W_1^g\hat{\mathbf{z}}_n + V_1^g\mathbf{u}_n + b_1^g\right)\right)\|\\
&\le \|(\boldsymbol{\psi}_{n-1} - \bar{\mathbf{y}}_{n-1})\| + \tau\|\mathbf{G}(\boldsymbol{\phi}_n, \boldsymbol{\psi}_{n-1}, \mathbf{u}_n) - \mathbf{G}(\bar{\mathbf{z}}_n, \tilde{\mathbf{z}}_n, \mathbf{u}_n)\|\\
&\quad + \tau\|\mathbf{G}(\bar{\mathbf{z}}_n, \tilde{\mathbf{z}}, \mathbf{u}_n) - W_3^g\sigma\left(W_2^g\tilde{\mathbf{z}}_n + W_1^g\hat{\mathbf{z}}_n + V_1^g\mathbf{u}_n + b_1^g\right)\|\\
&\le \|\boldsymbol{\psi}_{n-1} - \bar{\mathbf{y}}_{n-1})\| + \tau\mathrm{Lip}(\mathbf{G})\left(\|\boldsymbol{\phi}_n - \hat{\mathbf{z}}_n\| + \|\tilde{\mathbf{z}}_n - \hat{\mathbf{y}}_{n-1}\| + \|\boldsymbol{\psi}_{n-1} - \hat{\mathbf{y}}_{n-1}\|\right) + \tau\epsilon^*,\\
&\le (1 + \tau\mathrm{Lip}(\mathbf{G}))(1 + \mathrm{Lip}(\mathbf{f}))\|\boldsymbol{\psi}_{n-1} - \hat{\mathbf{y}}_{n-1}\| + \tau\mathrm{Lip}(\mathbf{G})\mathrm{Lip}(\mathbf{f})\|\boldsymbol{\phi}_{n-1} - \hat{\mathbf{z}}_{n-1}\|\\
&\quad + \tau(1 + \mathrm{Lip}(\mathbf{G})(2 + \mathrm{Lip}(\mathbf{f})))\epsilon^*,
\end{aligned}$$

where the last inequality follows by using the previous inequality together with (78) and (77).

As $\tau < 1$, it is easy to see from (71) that $\mathrm{Lip}(\mathbf{G}) < \mathrm{Lip}(\mathbf{g}) + \frac{2}{\tau}$. Therefore, the last inequality reduces to,

$$\begin{aligned}
\|\boldsymbol{\psi}_n - \hat{\mathbf{y}}_n\| \le\ & (3 + \tau\mathrm{Lip}(\mathbf{g}))(1 + \mathrm{Lip}(\mathbf{f}))\|\boldsymbol{\psi}_{n-1} - \hat{\mathbf{y}}_{n-1}\| + (2 + \tau\mathrm{Lip}(\mathbf{g}))\mathrm{Lip}(\mathbf{f})\|\boldsymbol{\phi}_{n-1} - \hat{\mathbf{z}}_{n-1}\|\\
&+ (\tau + (2 + \tau\mathrm{Lip}(\mathbf{g}))(2 + \mathrm{Lip}(\mathbf{f}))\,\epsilon^*.
\end{aligned}$$

Adding we obtain,

$$\|\boldsymbol{\phi}_n - \hat{\mathbf{z}}_n\| + \|\boldsymbol{\psi}_n - \hat{\mathbf{y}}_n\| \le C^*\left(\|\boldsymbol{\phi}_{n-1} - \hat{\mathbf{z}}_{n-1}\| + \|\boldsymbol{\psi}_{n-1} - \hat{\mathbf{y}}_{n-1}\|\right) + D^*\epsilon^*, \tag{79}$$

where,

$$\begin{aligned}
C^* &= \max\{(3 + \mathrm{Lip}(\mathbf{g}))\mathrm{Lip}(\mathbf{f}), \mathrm{Lip}(\mathbf{f})(3 + \mathrm{Lip}(\mathbf{g}))(1 + \mathrm{Lip}(\mathbf{f}))\},\\
D^* &= 1 + \mathrm{Lip}(\mathbf{f}) + (1 + \mathrm{Lip}(\mathbf{g})(2 + \mathrm{Lip}(\mathbf{f}))).
\end{aligned} \tag{80}$$

Iterating over n leads to the bound,

$$\|\boldsymbol{\phi}_n - \hat{\mathbf{z}}_n\| + \|\boldsymbol{\psi}_n - \hat{\mathbf{y}}_n\| \le n(C^*)^n D^*\epsilon^*. \tag{81}$$

Here, $\mathrm{Lip}(\mathbf{f}), \mathrm{Lip}(\mathbf{g})$ are the Lipschitz constants of the functions $\mathbf{f}, \mathbf{g}$ on the compact set $\{(h, c, u) \in \mathbb{R}^{d_h \times d_h \times d_u} : \|h\|, \|c\|, \|u\| < \mathbb{R}^*\}$. Note that one can readily prove using the zero values of initial states, the bounds (75), (77) and the assumption $\|\boldsymbol{\phi}_n\|, \|\boldsymbol{\psi}_n\|, \|\mathbf{u}_n\| < R$, that $\|\hat{\mathbf{z}}_n\|, \|\tilde{\mathbf{z}}_n\|, \|\hat{\mathbf{y}}_n\|, \|\tilde{\mathbf{y}}_n\| < R^* = 2R$.

Using the definition of the output function (11) and the bound (81) that,

$$\|\mathbf{o}_n - \mathbf{o}(\hat{\mathbf{y}}_n)\| \le \|\mathcal{W}_c\| n(C^*)^n D^*\epsilon^*. \tag{82}$$

Defining the dynamical system,

$$\begin{aligned}
\mathbf{z}_n^* &= \sigma\left(W_1^f \bar{W}_2 \bar{\mathbf{y}}_{n-1} + W_2^f W_3^g \mathbf{y}_{n-1}^* + V_1^f \mathbf{u}_n + b_1^f\right) \\
\bar{\mathbf{z}}_n &= \sigma\left(\bar{W}_1 W_3^g \mathbf{y}_{n-1}^*\right) \\
\mathbf{y}_n^* &= (1-\tau)\mathbf{y}_{n-1}^* + \tau\sigma\left(W_1^g W_3^f \mathbf{z}_n^* + W_2^g \bar{W}_2 \bar{\mathbf{z}}_n + V_1^g \mathbf{u}_n + b_1^g\right), \\
\bar{\mathbf{y}}_n &= \sigma\left(\bar{W}_1 W_3^f \mathbf{z}_n^*\right).
\end{aligned} \tag{83}$$

By multiplying suitable matrices to (78), we obtain that,

$$\hat{\mathbf{z}}_n = W_3^f \mathbf{z}_n^*, \quad \tilde{\mathbf{z}}_n = \bar{W}_2 \bar{\mathbf{z}}_n, \quad \hat{\mathbf{y}}_n = W_3^g \mathbf{y}_n^*, \quad \tilde{\mathbf{y}}_n = \bar{W}_2 \bar{\mathbf{y}}_n. \tag{84}$$

Finally, in addition to b_∞ defined in (52), for any given $\tau \in (0,1]$, we introduce $b_\tau \in \mathbb{R}$ defined by

$$\hat{\sigma}(b_\tau) = \tau. \tag{85}$$

The existence of a unique b_τ follows from the fact that the sigmoid function $\hat{\sigma}$ is monotone. Next, we define the two vectors $\mathbf{b}_\infty, \mathbf{b}_\tau \in \mathbb{R}^{2d_h}$ as

$$\begin{aligned}
\mathbf{b}_\infty^i &= b_\infty, \quad \forall\ 1 \le i \le 2d_h, \\
\mathbf{b}_\tau^i &= b_\tau, \quad \forall\ 1 \le i \le d_h, \\
\mathbf{b}_\tau^i &= b_\infty, \quad \forall\ d_h + 1 \le i \le 2d_h.
\end{aligned} \tag{86}$$

We are now in a position to define the LEM of form (3), which will approximate the two-scale dynamical system (10). To this end, we define the hidden states $\mathbf{z}_n, \mathbf{y}_n \in \mathbb{R}^{2d_h}$ such that $\mathbf{z}_n = [\mathbf{z}_n^*, \bar{\mathbf{z}}_n^*]$ and $\mathbf{y}_n = [\mathbf{y}_n^*, \bar{\mathbf{y}}_n^*]$. The parameters for the corresponding LEM of form (3) given by,

$$\begin{aligned}
&\Delta t = 1, d_y = 2d_h \\
&\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{V}_1 = \mathbf{V}_2 \equiv 0, \\
&\mathbf{b}_1 = \mathbf{b}_\infty, \quad \mathbf{b}_2 = \mathbf{b}_\tau, \\
&\mathbf{W}_z = \begin{bmatrix} W_2^f W_3^g & W_1^f \bar{W}_2 \\ \bar{W}_1 W_3^g & 0 \end{bmatrix}, \quad \mathbf{V}_z = [V_1^f\, 0], \quad \mathbf{b}_z = [b_1^f, 0], \\
&\mathbf{W}_y = \begin{bmatrix} W_1^g W_3^f & W_2^g \bar{W}_2 \\ \bar{W}_1 W_3^f & 0 \end{bmatrix}, \quad \mathbf{V}_z = [V_1^g 0], \quad \mathbf{b}_z = [b_1^g, 0],
\end{aligned} \tag{87}$$

and with following parameters defining the output states,

$$\mathcal{W}_y = \begin{bmatrix} \mathcal{W}_c W_3^g & 0 \end{bmatrix}, \tag{88}$$

yields an output state $\omega_n = \mathcal{W}_y \mathbf{y}_n$.

It is straightforward to observe that $\omega_n \equiv \mathbf{o}(\hat{\mathbf{y}}_n)$. Hence, the desired bound (11) follows from (81) by choosing,

$$\epsilon^* = \frac{\epsilon}{N(C^*)^N D^*}.$$

□

E LEMs emulate Heterogeneous multiscale methods for ODEs

Following [24], we consider the following prototypical example of a fast-slow system of ordinary differential equations,

$$\begin{aligned}
\boldsymbol{\phi}'(t) &= \frac{1}{\tau}\left(f(\boldsymbol{\psi}) - \boldsymbol{\phi}\right), \\
\boldsymbol{\psi}'(t) &= g(\boldsymbol{\phi}, \boldsymbol{\psi}).
\end{aligned} \tag{89}$$

Here $\boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbb{R}^m$ are the fast and slow variables respectively and $0 < \tau << 1$ is a small parameter. Note that we have rescaled time and are interested in the dynamics of the slow variable $\boldsymbol{\psi}(t)$ in the time interval $[0, T]$.

A naive time-stepping numerical scheme for (89) requires a time step size $\delta t \sim O(\tau)$. Thus, the computation will entail time updates $N \sim O(1/\tau)$. Hence, one needs a multiscale ODE solver to approximate the solutions of the system (89). One such popular ODE solver can be derived by using the Heterogenous multiscale method (HMM); see Kuehn [24] and references therein. This in turns, requires using two time stepping schemes, a *macro* solver for the slow variable, with a time step Δt of the form,

$$\boldsymbol{\psi}_n = \boldsymbol{\psi}_{n-1} + \tilde{\Delta t} g(\boldsymbol{\phi}_n, \boldsymbol{\psi}_{n-1}). \tag{90}$$

Here, the time step $\tilde{\Delta t} < 1$ is independent of the small parameter τ.

Moreover, the fast variable is updated using a *micro* solver of the form,

$$\begin{aligned} \boldsymbol{\phi}_{n-1}^{(k)} &= \boldsymbol{\phi}_{n-1}^{(k-1)} - \delta t (f(\boldsymbol{\psi}_{n-1}) - \boldsymbol{\phi}_{n-1}^{(k-1)}), \quad 1 \le k \le K. \\ \boldsymbol{\phi}_n &= \boldsymbol{\phi}_{n-1}^{K}, \\ \boldsymbol{\phi}_{n-1}^{(0)} &= \boldsymbol{\phi}_{n-1}. \end{aligned} \tag{91}$$

Note that the micro time step size δt and the number of micro time steps K are assumed to independent of the small parameter τ.

It is shown in [24] (Chapter 10.8) that for any given small tolerance $\epsilon > 0$, one can choose a macro time step $\tilde{\Delta t}$, a micro time step δt, the number K of micro time steps, the number N of macro time steps, independent of τ, such that the discrete states $\boldsymbol{\psi}_n$ approximate the slow-variable $\boldsymbol{\psi}(t_n)$ (with $t_n = n\tilde{\Delta t}$) of the fast-slow system (89) to the desired accuracy of ϵ.

Our aim is to show that we can construct a LEM of the form (3) such that the states $\boldsymbol{\phi}_n, \boldsymbol{\psi}_n$, defined in (90), (91) can be approximated to arbitrary accuracy. By combining this with the accuracy of HMM, we will prove that LEMs can approximate the solutions of the fast-slow system (89) to desired accuracy, independent of the small parameter τ in (89).

Proposition E.1. *Let $\boldsymbol{\phi}_n, \boldsymbol{\psi}_n \in \mathbb{R}^m$, for $1 \le n \le N$, be the states defined by the HMM dynamical system* (90), (91). *For any given $\epsilon > 0$, there exists a LEM of the form* (3) *with hidden states $[\mathbf{z}_n, \mathbf{y}_n]$, where $\mathbf{z}_n, \mathbf{y}_n \in \mathbb{R}^{d_m}$ and output states ω_n^h, ω_n^c such that the following holds,*

$$\max\left\{\|\boldsymbol{\phi}_n - \omega_n^h\|, \|\boldsymbol{\psi}_n - \omega_n^c\|\right\} \le \epsilon, \quad \forall 1 \le n \le N. \tag{92}$$

Proof. We start by using iteration on the micro solver (90) from $k = 1$ to $k = K$ to derive the following,

$$\begin{aligned} \boldsymbol{\phi}_n &= \overline{\delta t}\boldsymbol{\phi}_{n-1} + (1 - \overline{\delta t}) f(\boldsymbol{\psi}_{n-1}), \\ \overline{\delta t} &= (1 - \delta t)^K. \end{aligned} \tag{93}$$

As $\delta t < 1$, we have that $\overline{\delta t} < 1$.

By the universal approximation theorem for Tanh neural networks, for any given tolerance ϵ^*, there exist weight matrices $W_1^f \in \mathbb{R}^{d_1 \times m}, W_2^f \in \mathbb{R}^{m \times d_1}$ and bias vector $b_1^f \in \mathbb{R}^{d_1}$ such that the tanh neural network defined by,

$$\mathcal{N}_f(c) = W_2^f \sigma\left(W_1^f c + b_1^f\right), \tag{94}$$

approximates the underlying function $\mathbf{f}$ in the following manner,

$$\max_{\|c\| < R^*} \|\mathbf{f}(c) - \mathcal{N}_f(c)\| \le \epsilon^*. \tag{95}$$

Next, we define the following map,

$$\mathbf{G}(h, c) = \mathbf{g}(h, c) + c, \tag{96}$$

By the universal approximation theorem, given ϵ^*, there exist weight matrices $W_1^g, W_2^g \in \mathbb{R}^{d_2 \times m}, W_3^g \in \mathbb{R}^{m \times d_2}$ and bias vector $b_1^g \in \mathbb{R}^{d_2}$ such that the tanh neural network defined by,

$$\mathcal{N}_g(h, c) = W_3^g \sigma\left(W_1^g h + W_2^g c + b_1^g\right), \tag{97}$$

approximates the function $\mathbf{G}$ (96) in the following manner,

$$\max_{\max(\|h\|,\|c\|)<R^*} \|\mathbf{G}(h,c) - \mathcal{N}_g(h,c)\| \leq \epsilon^*. \tag{98}$$

Next, as in the proof of propositions 4.4 4.5, one can readily approximate the identity function $\hat{f}(h,c) = h$ with a tanh neural network of the form,

$$\bar{\mathcal{N}}_f(h) = \bar{W}_2\sigma\left(\bar{W}_1 h\right), \tag{99}$$

such that

$$\max_{\|h\|,\|c\|<R^*} \|\hat{f}(h,c) - \mathcal{N}_f(h)\| \leq \epsilon^*, \tag{100}$$

and with the same weights and biases, one can approximate the identity function $\hat{g}(h,c) = c$ with the Tanh neural network,

$$\bar{\mathcal{N}}_g(c) = \bar{W}_2\sigma\left(\bar{W}_1 c\right), \tag{101}$$

such that

$$\max_{\|h\|,\|c\|<R^*} \|\hat{g}(h,c) - \mathcal{N}_g(c)\| \leq \epsilon^*. \tag{102}$$

Then, we define the following dynamical system,

$$\begin{aligned}
\hat{\mathbf{z}}_n &= \overline{\delta t}\hat{\mathbf{z}}_n + (1-\overline{\delta t})W_2^f\,\sigma\left(W_1^f\,\hat{\mathbf{y}}_{n-1} + b_1^f\right),\\
\tilde{\mathbf{z}}_n &= \bar{W}_2\sigma\left(\bar{W}_1\hat{\mathbf{y}}_{n-1}\right),\\
\hat{\mathbf{y}}_n &= (1-\tilde{\Delta t})\hat{\mathbf{y}}_{n-1} + \tilde{\Delta t}W_3^g\sigma\left(W_1^g\hat{\mathbf{z}}_n + W_2^g\tilde{\mathbf{z}}_n + b_1^g\right),\\
\tilde{\mathbf{y}}_n &= \bar{W}_2\sigma\left(\bar{W}_1\hat{\mathbf{z}}_n\right),
\end{aligned} \tag{103}$$

with hidden states $\hat{\mathbf{z}}_n, \tilde{\mathbf{z}}_n, \hat{\mathbf{y}}_n, \tilde{\mathbf{y}}_n \in \mathbb{R}^m$ and with initial states $\hat{\mathbf{z}}_0 = \tilde{\mathbf{z}}_0 = \hat{\mathbf{y}}_0 = \tilde{\mathbf{y}}_0 = 0$.

Completely analogously as in the derivation of (81), we can derive the following bound,

$$\|\boldsymbol{\phi}_n - \hat{\mathbf{z}}_n\| + \|\boldsymbol{\psi}_n - \hat{\mathbf{y}}_n\| \leq C^*\epsilon^*, \tag{104}$$

with constant $C^* = C^*(n, \mathrm{Lip}(f), \mathrm{Lip}(g))$.

Defining the dynamical system,

$$\begin{aligned}
\mathbf{z}_n^* &= \overline{\delta t}\mathbf{z}_n^* + (1-\overline{\delta t})\sigma\left(W_1^f\,W_3^g\hat{\mathbf{y}}_{n-1} + b_1^f\right)\\
\bar{\mathbf{z}}_n &= \sigma\left(\bar{W}_1 W_3^g\mathbf{y}_{n-1}^*\right)\\
\mathbf{y}_n^* &= (1-\tilde{\Delta t})\mathbf{y}_{n-1}^* + \tilde{\Delta t}\sigma\left(W_1^g W_3^f\,\mathbf{z}_n^* + W_2^g\bar{W}_2\tilde{\mathbf{z}}_n + b_1^g\right)\\
\bar{\mathbf{y}}_n &= \sigma\left(\bar{W}_1 W_2^f\,\mathbf{z}_n^*\right).
\end{aligned} \tag{105}$$

By multiplying suitable matrices to (105), we obtain that,

$$\hat{\mathbf{z}}_n = W_2^f\,\mathbf{z}_n^*, \quad \tilde{\mathbf{z}}_n = \bar{W}_2\bar{\mathbf{z}}_n, \quad \hat{\mathbf{y}}_n = W_3^g\mathbf{y}_n^*, \quad \tilde{\mathbf{y}}_n = \bar{W}_2\bar{\mathbf{y}}_n. \tag{106}$$

In addition to b_∞ defined in (52), for $\overline{\delta t} \in (0,1]$, we introduce $b_\delta \in \mathbb{R}$ defined by

$$\hat{\sigma}(b_\delta) = \overline{\delta t}. \tag{107}$$

Similarly for $\tilde{\Delta t} \in (0,1]$, we introduce $b_\Delta \in \mathbb{R}$ defined by

$$\hat{\sigma}(b_\Delta) = 1 - \tilde{\Delta t}. \tag{108}$$

The existence of unique b_δ and b_Δ follows from the fact that the sigmoid function $\hat{\sigma}$ is monotone.

Next, we define the two vectors $\mathbf{b}_\infty, \mathbf{b}_\delta, \mathbf{b}_\Delta \in \mathbb{R}^{2m}$ as

$$\begin{aligned} \mathbf{b}_\delta^i &= b_\delta, \quad \forall\ 1 \leq i \leq m, \\ \mathbf{b}_\delta^i &= b_\infty, \quad \forall\ m+1 \leq i \leq 2m, \\ \mathbf{b}_\Delta^i &= b_\Delta, \quad \forall\ 1 \leq i \leq m, \\ \mathbf{b}_\Delta^i &= b_\infty, \quad \forall\ m+1 \leq i \leq 2m. \end{aligned} \tag{109}$$

We define the LEM of form (3), which will approximate the HMM (90),(91). To this end, we define the hidden states $\mathbf{z}_n, \mathbf{y}_n \in \mathbb{R}^{2m}$ such that $\mathbf{z}_n = [\mathbf{z}_n^*, \bar{\mathbf{z}}_n^*]$ and $\mathbf{y}_n = [\mathbf{y}_n^*, \bar{\mathbf{y}}_n^*]$. The parameters for the corresponding LEM of form (3) given by,

$$\begin{aligned} &\Delta t = 1, d_y = 2m \\ &\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{V}_1 = \mathbf{V}_2 \equiv 0, \\ &\mathbf{b}_1 = \mathbf{b}_\delta, \quad \mathbf{b}_2 = \mathbf{b}_\Delta, \\ &\mathbf{W}_z = \begin{bmatrix} W_1^f W_3^g & 0 \\ \bar{W}_1 W_3^g & 0 \end{bmatrix}, \quad \mathbf{V}_z = 0, \quad \mathbf{b}_z = [b_1^f, 0], \\ &\mathbf{W}_y = \begin{bmatrix} W_1^g W_3^f & W_2^g \bar{W}_2 \\ \bar{W}_1 W_2^f & 0 \end{bmatrix}, \quad \mathbf{V}_z = 0, \quad \mathbf{b}_z = [b_1^g, 0]. \end{aligned} \tag{110}$$

The output states are defined by,

$$\omega_n^h = W_2^f \mathbf{z}_n^*, \quad \omega_n^h = W_3^g \mathbf{y}_n^* \tag{111}$$

It is straightforward to observe that $\omega_n^h = \hat{\mathbf{z}}_n,\ \omega_n^c = \hat{\mathbf{y}}_n$. Hence, the desired bound (92) follows from (104) by choosing,

$$\epsilon^* = \frac{\epsilon}{C^*}.$$

□

F Training details

All experiments were run on CPU, namely Intel Xeon Gold 5118 and AMD EPYC 7H12, except for Google12, PTB character-level and PTB word-level, which were run on a GeForce RTX 2080 Ti GPU. All weights and biases of LEM (3) are initialized according to $\mathcal{U}(-1/\sqrt{d}, 1/\sqrt{d})$, where d is the number of hidden units.

Table 8: Rounded hyperparameters of the best performing LEM architecture for each experiment. If no value is given for Δt, it means that Δt is fixed to 1 and no fine-tuning is performed on this hyperparameter.

experiment	learning rate	batch size	Δt
Adding ($N = 10000$)	2.6×10^{-3}	50	2.4×10^{-2}
sMNIST	1.8×10^{-3}	128	2.1×10^{-1}
psMNIST	3.5×10^{-3}	128	1.9×10^{0}
nCIFAR-10	1.8×10^{-3}	120	9.5×10^{-1}
EigenWorms	2.3×10^{-3}	8	1.6×10^{-3}
Healthcare	1.6×10^{-3}	32	1.9×10^{-1}
FitzHugh-Nagumo	9.0×10^{-3}	32	/
Google12	8.9×10^{-4}	100	/
PTB character-level	6.6×10^{-4}	128	/
PTB word-level	6.8×10^{-4}	64	/

The hyperparameters are selected based on a random search algorithm, where we present the rounded hyperparameters for the best performing LEM model (*based on a validation set*) on each task in Table 8.

We base the training for the PTB experiments on the following language modelling code: https://github.com/deepmind/lamb, where we fine-tune, based on a random search algorithm, only the learning rate, input-, output- and state-dropout, L^2-penalty term and the maximum gradient norm.

We train LEM for 100 epochs on sMNIST, psMNIST and nCIFAR-10, after which we decrease the learning rate by a factor of 10 and proceed training for 20 epochs. Moreover, we train LEM for 50, 60 as well as 400 epochs on EigenWorms, Google12 and FitzHugh-Nagumo. We decrease the learning rate by a factor of 10 after 50 epochs on Google12. On the Healthcare task, we train LEM for 250 epochs, after which we decrease the learning rate by a factor of 10 and proceed training for 250 epochs.

G Further experimental results

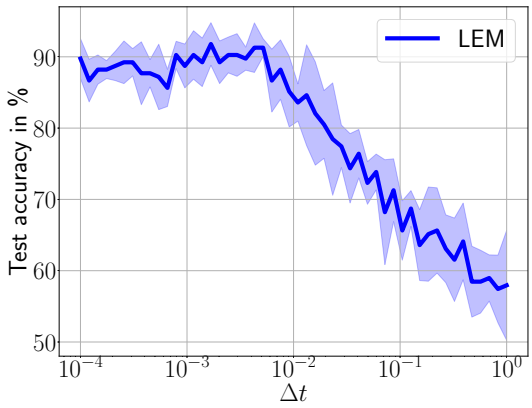

Figure 4: Ablation study on hyperparameter Δt in (3) using the EigenWorms experiment.

On the smallness of Δt. As mentioned in the main text, the parameter Δt in LEM (3) appeared to play a role in the proofs of Propositions 4.2 and D.2. In particular, one needed Δt to be small enough for the gradients to be bounded in (6). In practice, it turns out that a default value of $\Delta t = 1$ suffices for yielding very good empirical performance on many of our experiments, see Table 8. These experiments were based on data with sequence lengths that are not too long. On the other hand, data sets with long sequence lengths did require moderately small values of $\Delta t \approx \mathcal{O}(0.1)$. The only exception was the Eigenworms experiment with a very long sequence length of 17984. In this case, the best performing LEM corresponded to a $\Delta t = 1.6 \times 10^{-3}$. We further investigate the variation of accuracy in the Eigenworms experiment with respect to changing Δt with an ablation study, whose result is presented in Fig. 4. From this figure, we do observe that the accuracy of LEM is rather poor when $\Delta t \approx 1$. The accuracy improves considerably as Δt is reduced, till it reach a saturation around a $\Delta t \approx 10^{-2}$. Thus, we conclude that small values of Δt are needed in both theory and practice if the underlying sequences are very long.