Understanding and Optimizing Temporal Credit Assignment in Biological and Artificial Neural Networks using Dynamical Systems Theory

Rainer Engelken (re2365@columbia.edu) & **Larry Abbott (lfa2103@columbia.edu)**
Zuckerman Mind Brain Behavior Institute
Columbia University, New York, USA

Abstract

Finding temporal associations across long timescales in cognitively demanding tasks, such as delayed match to sample and parametric working memory, is challenging for both biological and artificial neural networks. Gradient-based training of recurrent neural circuit models for temporal tasks with long time horizons presents challenges that potentially lead to vanishing or exploding gradients. We leverage dynamical systems theory to understand the learning dynamics and solution space of such temporal credit assignment problems in spiking and firing rate networks. Specifically, we connect this issue to the Lyapunov exponents of the forward dynamics, describing how perturbations grow or shrink during forward passes.

We propose "gradient flossing", a method to address gradient instability in recurrent spiking and firing rate networks by controlling the Lyapunov exponents of the forward dynamics throughout learning. We regularize Lyapunov exponents towards zero, ensuring that the corresponding directions in tangent space grow or shrink only slowly to facilitate more robust propagation of learning signals over long time horizons.

This approach improves RNN stability and training success in temporal cognitive tasks by regulating the norm and dimensionality of the gradient signal in backpropagation through the dynamic adjustment of Lyapunov exponents.

Keywords: Temporal Credit Assignment; Learning Dynamics; Spiking Networks; Rate Networks; Backpropagation Through Time; Exploding/Vanishing Gradients

Models

For rate networks, we consider a conventional model of N nonlinear rate neurons obeying the dynamics $\tau \frac{\mathrm{d}h_i}{\mathrm{d}t} = -h_i + \sum_{j=1}^{N} J_{ij}\phi(h_j) + I_i(t)$, (Eq 1). Here, the coupling weights J_{ij} are drawn independently from a Gaussian distribution with zero mean and variance g^2/N, with g being a gain parameter that controls weight heterogeneity. We use a $\tanh$ transfer function $\phi(x) = \tanh(x)$. $I_i(t)$ is the external input to each neuron. In the case of spiking networks, we examine a balanced network of leaky integrate-and-fire (QIF) neurons governed by the voltage equation $\tau_m \frac{\mathrm{d}V_i}{\mathrm{d}t} = -V_i + I_i(t) + I_{ext}$ with exponentially decaying synapic currents $\tau_s \frac{\mathrm{d}I_i}{\mathrm{d}t} = -I_i + \tau_s \sum_j J_{ij}\delta(t - t_j^s)$. The membrane potential V_i of individual neurons resets from $V_{th} = 1$ to $V_{re} = 0$ after a spike. J_{ij} is initially a sparse directed Erdös-Rényi graph of size N with K synapses per neuron. All non-zero weights are set to $-\frac{J_0}{\sqrt{K}}$. Thus, the network settles into a balanced state characterized by asynchronous irregular activity(van Vreeswijk & Sompolinsky, 1996, 1998; Brunel, 2000; Renart et al., 2010). Recurrent inhibition dynamically cancels the positive external current, $I_{ext} = I_0\sqrt{K}$. We solved the spiking network dynamics by either temporally discretizing it or through event-based simulations (Engelken, 2023b), and trained the spiking network using surrogate gradients (Neftci, Mostafa, & Zenke, 2019).

Additional Details: Dynamic Stability & Propagation of Learning Signals in Recurrent Neural Networks: Our approach is to exploit a link between dynamic stability and the propagation of learning signals: Lyapunov exponents (LEs) give the average exponential growth rates of infinitesimal perturbations in the tangent space of the forward dynamics of an RNN, which also constrains the signal propagation in backpropagation over long time horizons. Mathematically speaking, for the discretized RNN dynamics $\mathbf{h}_{s+1} = \mathbf{f}_\theta(\mathbf{h}_s, \mathbf{x}_{s+1})$, with recurrent network state $\mathbf{h}$, external input $\mathbf{x}$, and parameters θ, the gradient of the loss with respect to θ is evaluated by unrolling the network dynamics in time. The resulting expression for the gradient is given by: $\frac{\partial \mathcal{L}_t}{\partial \theta} = \frac{\partial \mathcal{L}_t}{\partial \mathbf{h}_t} \sum_\tau \mathbf{T}_t(\mathbf{h}_\tau) \frac{\partial \mathbf{h}_\tau}{\partial \theta}$ where $\mathbf{T}_t(\mathbf{h}_\tau)$ is composed of a product of Jacobians $\mathbf{T}_t(\mathbf{h}_\tau) = \prod_{\tau'=\tau}^{t-1} \frac{\partial \mathbf{h}_{\tau'+1}}{\partial \mathbf{h}_{\tau'}}$. Due to the chain of matrix multiplications in $\mathbf{T}_t$, gradients tend to vanish or explode exponentially with time (Hochreiter & Jrgen Schmidhuber, 1997; Pascanu, Mikolov, & Bengio, 2013). Recently, it has been pointed out that $\mathbf{T}_t(\mathbf{h}_\tau)$ is closely related to LE exponents. More specifically, LE are defined as the asymptotic time-averaged logarithms of the singular values of the long-term Jacobian $\lambda_i = \lim_{t\to\infty} \frac{1}{t-\tau} \log(\sigma_{i,t})$ where $\sigma_{i,t}$ denotes the ith singular value of $\mathbf{T}_t(\mathbf{h}_\tau)$ with $\sigma_{1,t} \geq \sigma_{2,t} \ldots \sigma_{N,t}$ (Engelken, 2023a; Park, Sgodi, & Sok, 2023; Lindner, 2021; Vogt, Puelma Touzel, Shlizerman, & Lajoie, 2022; Engelken, Wolf, & Abbott, 2020, 2023). Thus, positive LEs imply exponentially growing gradient modes, while negative ones correspond to exponentially vanishing gradient modes.

Gradient Flossing: Idea and Algorithm We introduce *gradient flossing*, a method that regularizes the singular values of the long-term Jacobian, thereby enhancing the information propagation of learning signals. To prevent exploding and vanishing gradients, we constrain Lyapunov exponents to be close to zero. This ensures that the corresponding directions in tangent space grow or shrink slowly on average. This is achieved by incorporating an additional term in the loss function, based on the squares of the k largest LEs: $\mathcal{L}_{\text{flossing}} = \sum_{i=1}^{k} \lambda_i^2$ which is computed using differentiable linear algebra. In Fig. 1C, we show how *gradient flossing* can modify the first LE of randomly initialized RNNs to match a desired target value. In Fig 1D, we then show the same for multiple LEs.

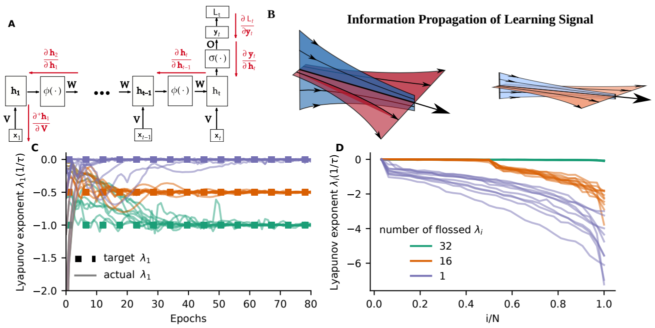

Figure 1: ***Gradient flossing* controls Lyapunov exponents and information propagation of learning signals**
A) Exploding and vanishing gradients in backpropagation through time arise from amplification or attenuation of the product of Jacobians that form the long-term Jacobian $\mathbf{T}_t(\mathbf{h}_\tau)=\prod_{\tau'=\tau}^{t-1}\frac{\partial \mathbf{h}_{\tau'+1}}{\partial \mathbf{h}_{\tau'}}$. This is closely related to Lyapunov exponents of the forward dynamics that measure average exponential rates of divergence of nearby trajectories. **B)** To improve the information propagation of learning signals, we introduce *gradient flossing*, which regularizes the tangent dynamics of RNNs. **C)** First Lyapunov exponent of recurrent networks over the course of training. Minimizing the mean squared error between the estimated first Lyapunov exponent and target Lyapunov exponent $\lambda_1=-1,-0.5,0$ by gradient descent. Ten recurrent networks were initialized with Gaussian recurrent weights $J_{ij}\sim\mathcal{N}(0,g^2/N)$ where values of g were drawn $g\sim\mathrm{Unif}(0,1)$. **D)** Full Lyapunov spectrum of recurrent network after a different number of Lyapunov exponents are regularized towards zero via *gradient flossing*.

Additional Results on Challenging Cognitive Tasks We show that *gradient flossing* enhances both training speed and success rate across three challenging cognitive tasks (Fig. **??**). First, we show that in a parametric working memory task where a variable number of items must be stored and reproduced by the recurrent network after a delay of d, more memory items can be successfully recalled with gradient flossing (Fig. 2A, C). Additionally, in a delayed context-dependent go-nogo discrimination task, gradient flossing enables the recurrent circuit to bridge a longer time horizon, thus facilitating a more challenging task (Fig. 2B, D). We study two different versions of the tasks: In the first version, the stimulus and context signal are presented sequentially, we call this the temporal context-dependent go-nogo task (Fig. 2E, G). In the second version, the context signal is presented at the same time as the task stimulus, we call this the spatial context-dependent go-nogo task (Fig. 2F, H). Furthermore, we show that gradient flossing during training can further enhance performance (green line in Fig. 2).

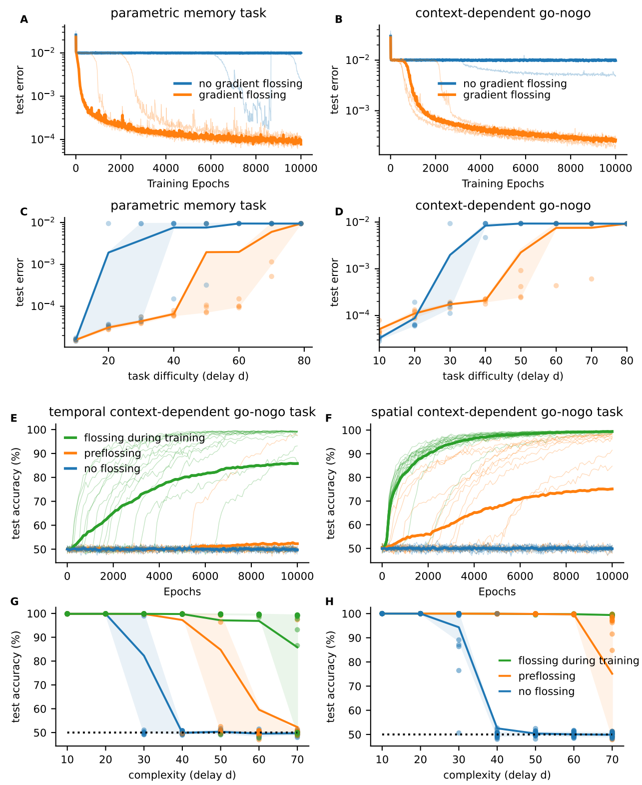

Figure 2: ***Gradient flossing* improves learning on cognitive tasks that involve bridging long time horizons**
A) Test error for recurrent rate networks trained on parametric memory task $y_t=x_{t-d}$ for $d=40$ with and without *gradient flossing*. Solid lines are medians across 5 network realizations. **B)** Same as **A** for the spatial delayed context-dependent go-nogo discrimination task. **C)** Mean final test error as a function of task difficulty (delay d) for parametric working memory task. **D)** Mean final test error as a function of task difficulty (delay d) for delayed context-dependent go-nogo task. **E)** Test accuracy for recurrent network trained on temporal version of context-dependent go-nogo task with *gradient flossing* during training (green), preflossing (*gradient flossing* before training) (orange), and with no *gradient flossing* (blue) for $d=70$. **F)** Same as **E** for spatial version of context-dependent go-nogo task. **G)** Test accuracy as a function of task difficulty (delay d) for temporal version of context-dependent go-nogo discrimination task. **H)** Test accuracy as a function of task difficulty (delay d) for spatial version of context-dependent go-nogo discrimination task.

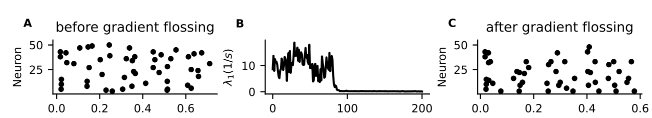

Figure 3: ***Gradient flossing* in spiking network A)** Spike raster before *gradient flossing*. **B)** Maximum Lyapunov exponent of spiking network is pushed towards zero during *gradient flossing*. **C)** Spike raster after *gradient flossing*.

References

Brunel, N. (2000, May). Dynamics of Sparsely Connected Networks of Excitatory and Inhibitory Spiking Neurons. *Journal of Computational Neuroscience*, *8*(3), 183–208. Retrieved 2012-11-19, from http://link.springer.com/article/10.1023/A%3A1008925309027 doi: 10.1023/A:1008925309027

Engelken, R. (2023a, May). Gradient Flossing: Improving Gradient Descent through Dynamic Control of Jacobians. Retrieved 2023-08-28, from https://openreview.net/forum?id=jEQRoJzDx8&referrer=%5BAuthor%20Console%5D(%2Fgroup%3Fid%3DNeurIPS.cc%2F2023%2FConference%2FAuthors%23your-submissions)

Engelken, R. (2023b, May). SparseProp: Efficient Event-Based Simulation and Training of Sparse Recurrent Spiking Neural Networks. Retrieved 2023-08-28, from https://openreview.net/forum?id=yzZbwQPkmP&referrer=%5BAuthor%20Console%5D(%2Fgroup%3Fid%3DNeurIPS.cc%2F2023%2FConference%2FAuthors%23your-submissions)

Engelken, R., Wolf, F., & Abbott, L. F. (2020, June). Lyapunov spectra of chaotic recurrent neural networks. *arXiv:2006.02427 [nlin, q-bio]*. Retrieved 2020-06-14, from http://arxiv.org/abs/2006.02427 (arXiv: 2006.02427)

Engelken, R., Wolf, F., & Abbott, L. F. (2023, October). Lyapunov spectra of chaotic recurrent neural networks. *Physical Review Research*, *5*(4), 043044. Retrieved 2023-10-19, from https://link.aps.org/doi/10.1103/PhysRevResearch.5.043044 doi: 10.1103/PhysRevResearch.5.043044

Hochreiter, S., & Jrgen Schmidhuber. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780.

Lindner, J. (2021). *Investigating the exploding and vanishing gradients problem with Lyapunov exponents*. Unpublished master's thesis, RWTH Aaachen, Aaachen/Juelich.

Neftci, E. O., Mostafa, H., & Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, *36*(6), 51–63.

Park, I. M., Sgodi, ., & Sok, P. A. (2023, August). Persistent learning signals and working memory without continuous attractors. *ArXiv*, arXiv:2308.12585v1.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013, June). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* (pp. III–1310–III–1318). Atlanta, GA, USA: JMLR.org.

Renart, A., Rocha, J. d. l., Bartho, P., Hollender, L., Parga, N., Reyes, A., & Harris, K. D. (2010, January). The Asynchronous State in Cortical Circuits. *Science*, *327*(5965), 587–590. Retrieved 2012-12-20, from http://www.sciencemag.org/content/327/5965/587 doi: 10.1126/science.1179850

van Vreeswijk, C., & Sompolinsky, H. (1996, December). Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity. *Science*, *274*(5293), 1724 –1726. Retrieved 2011-12-11, from http://www.sciencemag.org/content/274/5293/1724.abstract doi: 10.1126/science.274.5293.1724

van Vreeswijk, C., & Sompolinsky, H. (1998). Chaotic Balanced State in a Model of Cortical Circuits. *Neural Computation*, *10*(6), 1321–1371. Retrieved from http://dx.doi.org/10.1162/089976698300017214 doi: 10.1162/089976698300017214

Vogt, R., Puelma Touzel, M., Shlizerman, E., & Lajoie, G. (2022). On Lyapunov Exponents for RNNs: Understanding Information Propagation Using Dynamical Systems Tools. *Frontiers in Applied Mathematics and Statistics*, *8*. Retrieved 2023-04-08, from https://www.frontiersin.org/articles/10.3389/fams.2022.818799