Learning Dynamics of Linear Recurrent Neural Networks

Alexandra M. Proca (a.proca22@imperial.ac.uk)
Department of Computing, Imperial College London, 180 Queen's Gate
London, United Kingdom SW7 2RH

Murray Shanahan (m.shanahan@imperial.ac.uk)
Department of Computing, Imperial College London, 180 Queen's Gate
London, United Kingdom SW7 2RH

Pedro A.M. Mediano (pmediano@imperial.ac.uk)
Department of Computing, Imperial College London, 180 Queen's Gate
London, United Kingdom SW7 2RH

Abstract

Recurrent neural networks (RNNs) are widely used in neuroscience to model neural dynamics and learn tasks with temporal dependencies, and have been shown to utilize complex dynamical structures. However, it is still unknown how such structures emerge during training. Here, we aim to develop a better theoretical understanding of learning dynamics in RNNs by analyzing their linear counterparts analytically. Despite the absence of nonlinearity, deep linear networks are known to exhibit nonlinear learning dynamics. We show that the effect of exploding gradients acts as an effective regularizer of both recurrent and input-output weights and derive exact solutions of the nonlinear learning dynamics of the input-output connectivity modes, verified in simulation. Finally, we study the loss landscape and gradients for data with different temporal structures, revealing (un)learnable data dynamics and their solutions, criteria for generalization across trajectory lengths, and the existence of a bifurcation leading parameters towards either the global minimum or suboptimal solutions. Our work provides a first analytical treatment of the relationship between temporally-evolving data and learning dynamics in linear RNNs and builds a basis from which we can better understand how complex dynamic behavior emerges in cognitive models.

Keywords: learning dynamics; RNNs; linear networks

In the age of growing interdisciplinary exchange between cognitive neuroscience and machine learning, recurrent neural networks (RNNs) have become a popular choice for cognitive models of neural dynamics, as they not only replicate recurrent dynamics recorded in animals but are also capable of performing cognitive tasks with temporal dependencies. It's also been shown that RNNs reuse dynamical motifs, such as line attractors and limit cycles, for similar computations across different sets of tasks (Driscoll, Shenoy, & Sussillo, 2022). However, despite their widespread use and known complex computational abilities, there is still limited theoretical understanding of these models and how their underlying functional structures emerge.

One line of previous work has focused on using deep linear networks to analyze learning dynamics (Saxe, McClelland, & Ganguli, 2014, 2019). Although unable to solve nonlinear problems, these networks exhibit complex nonlinear learning dynamics and are analytically tractable, providing a useful framework for theoretical investigation. However, the analytical treatment of learning dynamics in linear networks has primarily remained in the domain of feedforward networks.

In this work, we're interested in studying the learning dynamics of linear RNNs to better understand the influence of temporal data on learning in recurrent cognitive systems.

Model and Results

We consider a RNN parameterized by matrices $W_x \in \mathbb{R}^{N_h \times N_x}, W_h \in \mathbb{R}^{N_h \times N_h}, W_y \in \mathbb{R}^{N_y \times N_h}$ with a hidden state $h_t \in \mathbb{R}^{N_h}$ that receives an input $x_t \in \mathbb{R}^{N_x}$ at each timestep t and updates its hidden state, until the final timestep T where it produces an output $\hat{y}_T \in \mathbb{R}^{N_y}$. We additionally initialize h_1 as a vector of zeros, yielding

$$\begin{aligned} h_{t+1} &= W_h h_t + W_x x_t \\ &= \sum_{i=1}^{t} W_h^{t-i} W_x x_i \\ \hat{y}_T &= W_y h_{T+1} \end{aligned}$$

Our goal is to train the network on a set of P trajectories $\{x_{1,\rho}, x_{2,\rho}, \dots, x_{T,\rho}, y_{T,\rho}\}_{\rho=1}^{P}$ by gradient descent on the squared error $\mathcal{L} = \sum_{\rho=1}^{P} \|y_{T,\rho} - W_y(\sum_{i=1}^{T} W_h^{T-i} W_x x_{i,\rho})\|^2$.

We make several assumptions to simplify our form, namely: (1) inputs are uncorrelated and whitened, with mean 0; (2) the data correlation matrix of the input x_t at timestep t and final output y_T has constant left and right singular matrices across all timesteps such that $\Sigma^{YX_t} = \sum_{\rho=1}^{P} y_T x_t^\top = U_y S_t V_x^\top$ for all t; (3) weight matrices W_x, W_h, W_y are diagonalizable at initialization by some set of orthogonal matrices and the left and right singular matrices U_y, V_x of the data correlation matrix. The gradient updates can then be written in terms of the singular values of the data correlation matrices over time $s_{\alpha,t}$, which are decoupled for each dimension α,

$$\begin{aligned} \frac{d}{dt} a_\alpha &= \sum_{i=1}^{T} c_\alpha b_\alpha^{T-i} (s_{\alpha,i} - c_\alpha b_\alpha^{T-i} a_\alpha) \\ \frac{d}{dt} b_\alpha &= \sum_{i=1}^{T} c_\alpha b_\alpha^{T-i-1} a_\alpha (s_{\alpha,i} - c_\alpha b_\alpha^{T-i} a_\alpha) \\ \frac{d}{dt} c_\alpha &= \sum_{i=1}^{T} b_\alpha^{T-i} a_\alpha (s_{\alpha,i} - c_\alpha b_\alpha^{T-i} a_\alpha) \end{aligned}$$

which arise from gradient descent on the energy function

$$E = \frac{1}{2} \sum_{i=1}^{T} (s_{\alpha,i} - c_\alpha b_\alpha^{T-i} a_\alpha)^2$$

where $c_\alpha, b_\alpha, a_\alpha$ are the α^{th} diagonal entries of the diagonalized matrices $\overline{W}_y, \overline{W}_h, \overline{W}_x$, respectively, also known as *connectivity modes*. We refer to b as the *recurrent* connectivity mode and ac as the *input-output* connectivity mode. To

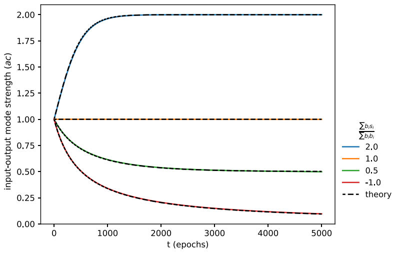

Figure 1: **Learning dynamics of input-output connectivity modes in a linear RNN**. The colored lines are simulations for different ratios of $\beta_s : \beta_b$ and dashed lines are the corresponding theoretical predictions for a trajectory of $T = 10$.

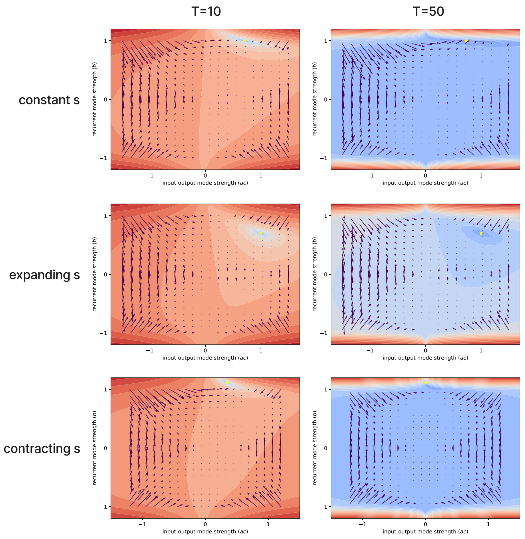

Figure 2: **Loss landscape and gradient vector field of connectivity modes for different data singular value dynamics**. Low to high loss corresponds to blue to red coloring and the global optimum is marked with a yellow star.

ease notation, we omit specifying α in the rest of the paper, although note that all terms still refer to a particular singular value dimension α. A first observation from the gradient functions is the effect of the well-known problem of vanishing (exploding) gradient, given by $|b| < 1$ ($|b| > 1$) as $T \rightarrow \infty$. For our analysis, we consider the case where $|b(0)| < 1$, as its counterpart is unstable. Then, as $T \rightarrow \infty$, the energy function converges to

$$E = \frac{1}{2}\Big(\sum_{i=1}^{T} s_i(s_i - 2cb^{T-i}a)\Big) - \frac{1}{2}c^2a^2\Big(\frac{1}{1-b} + \frac{1}{1+b} - 2\Big)$$

In this form, the second term drives the connectivity modes a, b, c towards 0, and specifically pushes b away from $1, -1$. This interplay leads to an effective regularization and sparsification of the network weights by keeping recurrent connectivity modes between the range of $-1, 1$ and favoring input-output connectivity modes centered around 0. For large T, b will generally not increase to become greater than one, unless a or c are close to 0. The space of learnable (linear) functions is thus constrained to data that can be modelled with $|b| \leq 1$.

As a first attempt at deriving exact solutions for the learning dynamics of the RNN connectivity modes, we assume $a(0) = c(0)$. Letting $\beta_s = \sum_{i=1}^{T} b^{T-i} s_i$, $\beta_b = \sum_{i=1}^{T} b^{2(T-i)}$, $\tau = \frac{1}{\text{LR}}$, the solution for the learning dynamics of ac are:

$$a(t)c(t) = \frac{e^{2t\beta_s/\tau}\beta_s}{\beta_s/(a(0)c(0)) - \beta_b + e^{2t\beta_s/\tau}\beta_b}$$

We verify this solution with simulations in Fig. 1 without updating b for different ratios of $\beta_s : \beta_b$.

Finally, we visualize the loss landscape and gradient vector field for b and ac for different data singular value dynamics $s_{1:T}$ and trajectory lengths in Fig. 2. We consider cases where $s_{1:T}$ are either constant ($s_t = 0.7$), expanding ($s_t = 0.7^{T-t}$), or contracting ($s_t = 0.9^t$), for $T = 10$ or $T = 50$. These cases correspond to output y_T being a sum of inputs $x_{1:T}$, where x_t is weighted by $s_{\alpha,t}$ for every singular value dimension. Thus, the expanding s case corresponds to weighting x_t in ascending order, contracting in descending, and constant equally.

For constant $s_t = 0.7$, the global minimum exists at $b = 1, ac = s = 0.7$, as b remains constant at 1 and ac learn the appropriate 'weighting.' Instead, for expanding s_t, the global minimum is found at $b = s_{T-1} = 0.7, ac = 1$, as the trajectory of s corresponds to the evolution of b given by b^{T-t}. In this case, b 'weights' inputs in ascending order, with ac remaining constant at 1. In both settings where data singular value dynamics are matched to RNN dynamics for any trajectory length T, the global minimum is the same for all trajectory lengths, indicating that learned optimal parameters should generalize for different trajectories. Interestingly, we also observe a bifurcation around $b = 0$, which pushes b towards either positive or negative values and subsequently, optimal or suboptimal solutions.

Finally, in the case of s_t contracting from $s_1 = 0.9$, the global minimum is $b = \frac{1}{s_1} \approx 1.11, ac = s_1^T \approx 0.359$ for $T = 10$ and $b = \frac{1}{s_1} \approx 1.11, ac = s_1^T \approx 0.00515$ for $T = 50$. In this setting, b produces contracting 'weighting' by being greater than 1 and scaled according to s_1. Because the 'weighting' b performs at each timestep is dependent on trajectory length, ac scales the contraction of b according to T to match that of the data singular values. Therefore, for small T, learned optimal parameters will not perfectly generalize to trajectory lengths that differ substantially, as the value of ac is dependent on T.

From these initial case studies, we predict that for large trajectories T, (1) all constant data singular values s are learnable, as $b = 1$ and $ac = s$, (2) most, if not all, s expanding according to s^t for $s > 1$ or s^{T-t} for $s < 1$ are learnable, and (3) for the most part, contracting s are not learnable, as their solutions lie outside of $|b| \leq 1$, which is unstable. We also predict that an optimal network's ability to generalize to different trajectory lengths will depend critically on whether the data singular value dynamics are dependent on trajectory length in a way that differs from the RNN connectivity modes.

Conclusion

We present here an initial theoretical study of the learning dynamics of linear RNNs and analyze how the temporal structure of data influences learning. We derive equations for the gradients of different parameters and the energy function in terms of the data singular values, as well as exact solutions for the learning dynamics of input-output connectivity modes which we verify in simulation. Our analysis of the loss landscape and gradient vector field reveals surprisingly complex dynamics which differ from those of linear feedforward networks, and lead us to make predictions about the learnability and generalizability of data with different temporal properties.

Acknowledgments

A.M.P. is funded by the Imperial College London President's PhD Scholarship.

References

Driscoll, L., Shenoy, K., & Sussillo, D. (2022). Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *bioRxiv*. doi: 10.1101/2022.08.15.503870

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations*.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537-11546. doi: 10.1073/pnas.1820226116