Partial observation can induce mechanistic mismatches in data-constrained RNNs

William Qian (williamqian@g.harvard.edu)
Biophysics Graduate Program
and
Kempner Institute for the Study of Natural and Artificial Intelligence,
Harvard University, 29 Oxford Street
Cambridge, MA 02138 USA

Jacob A. Zavatone-Veth (jzavatoneveth@g.harvard.edu)
Department of Physics,
Center for Brain Science,
and
John A. Paulson School of Engineering and Applied Sciences,
Harvard University, 29 Oxford Street
Cambridge, MA 02138 USA

Benjamin S. Ruben (benruben@g.harvard.edu)
Biophysics Graduate Program,
Harvard University, 29 Oxford Street
Cambridge, MA 02138 USA

Cengiz Pehlevan (cpehlevan@seas.harvard.edu)
John A. Paulson School of Engineering and Applied Sciences,
Center for Brain Science,
and
Kempner Institute for the Study of Natural and Artificial Intelligence,
Harvard University, 29 Oxford Street
Cambridge, MA 02138 USA

Abstract

One of the central goals of computational neuroscience is to understand how the dynamics of neural circuits give rise to their observed function. A popular approach towards this end is to train recurrent neural networks (RNNs) to reproduce experimental recordings of neural activity. These trained RNNs are then treated as surrogate models of biological neural circuits, whose properties can be dissected via dynamical systems analysis. While recent advances in population-level recording technologies have allowed simultaneous recording of up to tens of thousands of neurons, this represents only a tiny fraction of most cortical circuits. Here we show that partial observation can create mechanistic mismatches between a simulated teacher network and a data-constrained student, even when the two networks have otherwise matching architectures. In particular, we show that partial observation of models of working memory in cortex based on functionally feedforward or low-rank connectivity can lead to surrogate models with spurious attractor structure.

Keywords: Recurrent neural networks; short-term memory; dynamical systems; data-driven modeling

Introduction

Data-driven models of neural population dynamics are constructed under a number of less-than-ideal conditions, including partial observation of the target neural population, neuronal and measurement noise, and significant architecture mismatch between model and biology. This makes it virtually impossible to accurately reconstruct local information, such as synaptic weights (Das & Fiete, 2020). Nonetheless, a reasonable hope is that data-constrained RNNs should be able to at least capture the dynamical properties of ground truth circuits at a qualitative level–that is, recapitulate dynamical phenomena such as slow time scales, unstable directions, oscillatory dynamics, and attractors (Khona & Fiete, 2022; Vahidi, Sani, & Shanechi, 2024).

In particular, approximate line attractors—sets of stable fixed points organized along lines in neural activity space—could in principle be discovered by linearizing fitted RNN dynamics around a fixed point, and then observing whether the eigenspectra of the Jacobian contains just a single eigenvalue with real part near zero (Sussillo & Barak, 2013; Maheswaranathan, Williams, Golub, Ganguli, & Sussillo, 2019). Indeed, recent work has used data-constrained models in this fashion to propose that line attractors underlie the accumulation of internal drives to perform complex behaviors like aggression and mating (Nair et al., 2023; Liu, Nair, Linderman, & Anderson, 2023; Mountoufaris, Nair, Yang, Kim, & Anderson, 2023).

However, whether this procedure correctly recovers attractor mechanisms under partial observation remains unknown. To explore when data-driven RNN modeling and dynamical systems analysis can uncover spurious attractor structure, we study a student-teacher learning setup subject to partial observation and process noise.

Results

Problem setup

We consider a student-teacher setup where the activity of units in the student RNN obey the discretized rate-based dynamics

$$\mathbf{x}_t = (1-\alpha)\mathbf{x}_{t-1} + \alpha A\phi(\mathbf{x}_{t-1}) + \alpha\eta_t \tag{1}$$

where $\alpha = \Delta t/\tau$ is the discretization scale, $A \in \mathbb{R}^{d\times d}$ is the dynamics matrix, and $\eta_t \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_d)$ represents isotropic Gaussian noise. Suppose we fit the dynamics matrix A to observed data $\{\mathbf{x}_t^o\}_{t\in[T]}$ generated via partial observations of a teacher RNN of size $D > d$:

$$\mathbf{z}_t = (1-\alpha)\mathbf{z}_{t-1} + \alpha B\phi(\mathbf{z}_{t-1}) + \alpha\xi_t \tag{2}$$

$$\mathbf{x}_t^o = \mathbb{P}\mathbf{z}_t, \quad \mathbb{P} = \begin{pmatrix} I_{d\times d} & \mathbf{0}_{d\times(D-d)} \end{pmatrix} \tag{3}$$

where $B \in \mathbb{R}^{D\times D}$, and $\xi_t \sim \mathcal{N}(\mathbf{0}, \sigma_\xi^2 I_D)$. Under this model, the MAP estimate of the inferred dynamics matrix can then be described solely in terms of properties of the teacher RNN:

$$\hat{A} = \alpha^2\mathbb{P}(B\hat{C} + \textstyle\sum_{t=1}^T \xi_t\phi(\mathbf{z}_{t-1})^\top)\mathbb{P}^\top(\rho I_d + \alpha^2\mathbb{P}\hat{C}\mathbb{P}^\top)^{-1} \tag{4}$$

where $\hat{C} = \sum_{t=1}^T \phi(\mathbf{z}_{t-1})\phi(\mathbf{z}_{t-1})^\top$ is the empirical covariance.

To understand whether the dynamics of the teacher are qualitatively recovered, a natural question arises: When is the eigenspectrum of $\hat{A}$ qualitatively similar to that of B? Here we focus on the linear case $\phi(\mathbf{x}) = \mathbf{x}$ and the long time limit $T \to \infty$, which is amenable to analytical study. A statistic of particular interest is the gap between the two largest timescales τ_1 and τ_2 of the linear dynamics, which determines the "line attractor score" $\log_2(\tau_1/\tau_2)$ of Nair et al. (2023). There, scores of order unity were interpreted as approximate line attractors.

Normal teacher connectivity

First, consider the case in which B is a normal matrix ($BB^\top = B^\top B$), as in classical models of attractor dynamics (Seung, 1996). Ordering the eigenvalues of the teacher dynamics matrix B as $\Re(\lambda_1) \geq \Re(\lambda_2) \geq \cdots \geq \Re(\lambda_D)$, we show that the eigenvalues $\hat{\lambda}_i$ of the learned student dynamics matrix satisfy $\Re(\hat{\lambda}_i) \in [\Re(\lambda_D), \Re(\lambda_1)]$. As a consequence, spurious discovery of long persistent timescales of dynamics cannot occur. Further, we show that if B is symmetric and supports an approximate line attractor ($\lambda_1 = 1-\varepsilon$, $\varepsilon, \lambda_2 \ll 1$) of random orientation (Seung, 1996), the learned eigenvalues satisfy $\hat{\lambda}_1 \geq \lambda_1 - O(\frac{\varepsilon D}{d})$ and $\hat{\lambda}_2 \leq \lambda_2$. Thus, an approximate line attractor is recovered so long as the subsampling fraction does not overwhelm ε. We illustrate an example of this successful recovery in Fig. 1.

Non-normal teacher connectivity

If B is non-normal, severe overestimation of persistent timescales of dynamics can occur as a consequence of non-normal amplification. We demonstrate this explicitly for two neuroscience-inspired connectivity structures:

Functionally feedforward networks: We next consider teacher connectivity of the form $B = UTU^\top$, where T is

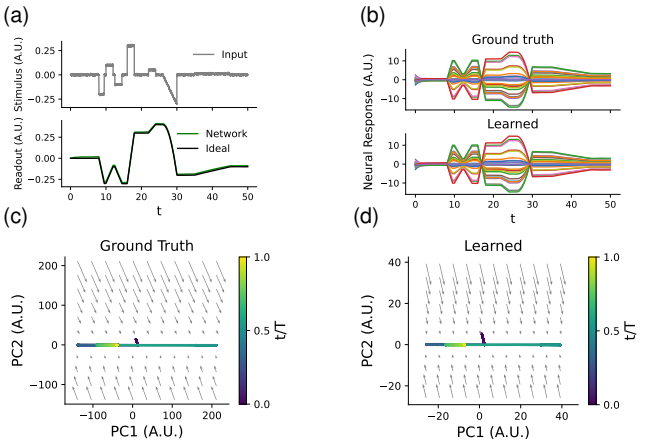

Figure 1: Successful recovery of a line attractor from recordings of its activity. (a). Input signal to be integrated, shown alongside the output for the true network and an ideal integrator. (b). Comparison of learned and ground truth neural activity. (c-d). Projection of flow fields of true and fitted models onto the top two principle components of their activity. The line attractor score of the fitted model is 6.67.

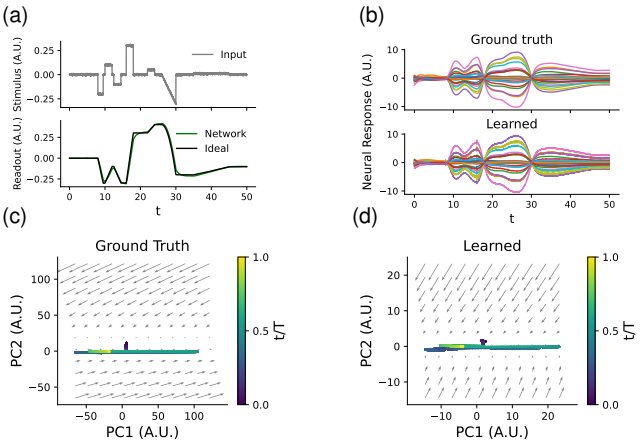

Figure 2: Spurious discovery of a line attractor from recordings of a functionally feedforward network. (a). Input signal to be integrated, shown alongside the output for the true network and an ideal integrator. (b). Comparison of learned and ground truth neural activity. (c-d). Projection of flow fields of true and fitted models onto the top two principle components of their activity. In the true model, the slowest points of dynamics are misaligned with activity. The line attractor score of the fitted model is 6.13.

a strictly upper triangular matrix satisfying $T_{ij} = \delta_{i+1,j} + \beta\delta_{i,1}(1-\delta_{1,j})$, and U is an orthonormal matrix. Although connectivity is fully recurrent, it is functionally feedforward in that it supports sequential activation of orthogonal modes of activity. Here, β controls the strength of skip connections that further amplify the output mode of activity. Like line attractors, connectivity of this form can be used to maintain memory of and integrate external inputs over time. However, this memory is achieved without long persistent timescales generated via large eigenvalues (Goldman, 2009; Ganguli, Huh, & Sompolinsky, 2008). These functionally feedforward chains have been proposed to underlie short-term memory storage in cortex (Daie, Svoboda, & Druckmann, 2021; Daie, Fontolan, Druckmann, & Svoboda, 2023). In Fig. 2, we show that fitting a latent linear dynamical system (LDS) to partially observed activity of a functionally feedforward network performing a simple 1D integration task can spuriously yield a line attractor.

Low-rank networks: Suppose the teacher dynamics matrix is both non-normal and low-rank, as in recently-proposed models for cortical computations (Dubreuil, Valente, Beiran, Mastrogiuseppe, & Ostojic, 2022). We address a minimal case where $B = MN^\top$, $M \in \mathbb{R}^{D\times r}$, $N \in \mathbb{R}^{D\times r}$. If $N^\top M = \mathbf{0}_{r\times r}$ and $N^\top N = M^\top M = \gamma^2 I_r$, then B has all 0 eigenvalues, but is non-normal. Here γ^2 is a scale parameter that controls the degree of non-normality. If one selects $\gamma^2 \sim O(\frac{D}{\sqrt{r}})$, then the elements of B are $O(1)$. In this case, when $d << D$, we show that the r eigenvalues of the learned dynamics matrix $\hat{\lambda}_i$ approach 1. We illustrate via simulations that such overestimation of eigenvalues qualitatively extends to low-rank teachers with non-trivial eigenspectra and to the nonlinear setting, thereby causing spurious discovery of attractors (Fig. 3).

Conclusions and extensions

We have shown that partial observation can result in spurious attractor dynamics in data-constrained RNN models. Though we do not show the results here, we have demonstrated that alternative fitting approaches also suffer from overestimation of eigenvalues as a consequence of partial observation (Nair et al., 2023; Dinc, Shai, Schnitzer, & Tanaka, 2023). In total, our results illustrate how partial observation can give rise to mechanistic mismatches between the circuits generating neural activity and data-constrained models.

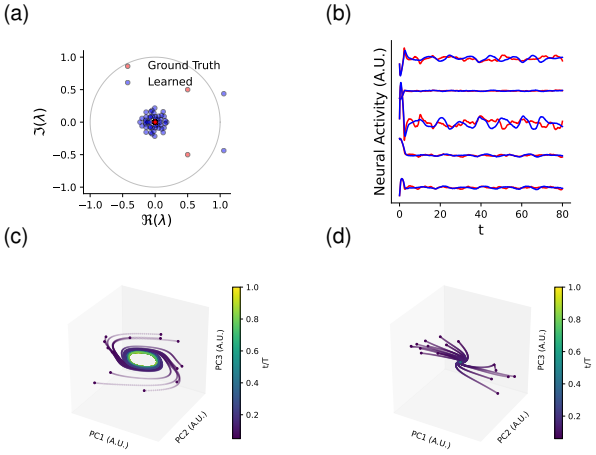

Figure 3: Spurious discovery of a limit cycle from a non-normal rank-2 teacher in the nonlinear setting. (a). Learned and ground truth dynamics matrix eigenvalues. (b) Sample snippets of true and fitted neural activity. (c-d) Randomly sampled trajectories from the (c) learned and (d) ground truth dynamics, projected onto the top three principal components.

Acknowledgements

JAZ-V, BSR, and CP were supported by NSF Award DMS-2134157 and NSF CAREER Award IIS-2239780. CP received additional support from a Sloan Research Fellowship. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

References

Daie, K., Fontolan, L., Druckmann, S., & Svoboda, K. (2023). Feedforward amplification in recurrent networks underlies paradoxical neural coding. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2023/08/07/2023.08.04.552026 doi: 10.1101/2023.08.04.552026

Daie, K., Svoboda, K., & Druckmann, S. (2021, Feb 01). Targeted photostimulation uncovers circuit motifs supporting short-term memory. *Nature Neuroscience*, *24*(2), 259-265. Retrieved from https://doi.org/10.1038/s41593-020-00776-3 doi: 10.1038/s41593-020-00776-3

Das, A., & Fiete, I. R. (2020, Oct 01). Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nature Neuroscience*, *23*(10), 1286-1296. Retrieved from https://doi.org/10.1038/s41593-020-0699-2 doi: 10.1038/s41593-020-0699-2

Dinc, F., Shai, A., Schnitzer, M., & Tanaka, H. (2023). Cornn: Convex optimization of recurrent neural networks for rapid inference of neural dynamics. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 51273–51301). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/a1035297387069793317783777f2d5864-Paper-Conference.pdf

Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., & Ostojic, S. (2022, Jun 01). The role of population structure in computations through neural dynamics. *Nature Neuroscience*, *25*(6), 783-794. Retrieved from https://doi.org/10.1038/s41593-022-01088-4 doi: 10.1038/s41593-022-01088-4

Ganguli, S., Huh, D., & Sompolinsky, H. (2008). Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, *105*(48), 18970-18975. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.0804451105 doi: 10.1073/pnas.0804451105

Goldman, M. S. (2009). Memory without feedback in a neural network. *Neuron*, *61*(4), 621-634. Retrieved from https://www.sciencedirect.com/science/article/pii/S0896627308010830 doi: https://doi.org/10.1016/j.neuron.2008.12.012

Khona, M., & Fiete, I. R. (2022, 12 01). Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, *23*(12), 744-766. Retrieved from https://doi.org/10.1038/s41583-022-00642-0 doi: 10.1038/s41583-022-00642-0

Liu, M., Nair, A., Linderman, S. W., & Anderson, D. J. (2023). Periodic hypothalamic attractor-like dynamics during the estrus cycle. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2023/05/22/2023.05.22.541741 doi: 10.1101/2023.05.22.541741

Maheswaranathan, N., Williams, A., Golub, M., Ganguli, S., & Sussillo, D. (2019). Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2019/file/d921c3c762b1522c475ac8fc0811bb0f-Paper.pdf

Mountoufaris, G., Nair, A., Yang, B., Kim, D.-W., & Anderson, D. J. (2023). Neuropeptide signaling is required to implement a line attractor encoding a persistent internal behavioral state. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2023/11/05/2023.11.01.565073 doi: 10.1101/2023.11.01.565073

Nair, A., Karigo, T., Yang, B., Ganguli, S., Schnitzer, M. J., Linderman, S. W., ... Kennedy, A. (2023). An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell*, *186*(1), 178-193.e15. Retrieved from https://www.sciencedirect.com/science/article/pii/S0092867422014714 doi: https://doi.org/10.1016/j.cell.2022.11.027

Seung, H. S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, *93*(23), 13339-13344. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.93.23.13339 doi: 10.1073/pnas.93.23.13339

Sussillo, D., & Barak, O. (2013, 03). Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation*, *25*(3), 626-649. Retrieved from https://doi.org/10.1162/NECO_a_00409 doi: 10.1162/NECO_a_00409

Vahidi, P., Sani, O. G., & Shanechi, M. M. (2024). Modeling and dissociation of intrinsic and input-driven neural population dynamics underlying behavior. *Proceedings of the National Academy of Sciences*, *121*(7), e2212887121. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2212887121 doi: 10.1073/pnas.2212887121