Optimal encoding of prior information in noisy working memory systems

Hua-Dong Xiong (hdx@email.arizona.edu)
Department of Psychology, The University of Arizona,
Tucson, AZ 85721, United States

Xue-Xin Wei (weixx@utexas.edu)
Departments of Neuroscience and Psychology, The University of Texas at Austin,
Austin, TX 78712, United States

Abstract:

The brain adapts to the statistical regularities of the environments to improve behavior. While there are many theories of efficient coding for feedforward processing, little is known about how prior information is encoded through recurrent computation in the neural systems, which is critical for cognition. Here we investigate this question in the context of working memory. By optimizing recurrent neural networks (RNNs) to perform a working memory (WM) task with different noise levels and stimulus priors. We found that, with increasing neural noise, the attractor dynamics in RNNs transform from continuous to discrete. Moreover, to encode stimulus statistics, RNNs generally allocate more attractor states for more frequent stimuli, leading to an increased encoding precision. The resulting neural representations exhibit systematic deviations from previous theories of efficient coding. Our results reveal novel mechanistic insights into how prior information is encoded through recurrent computations.

Keywords: working memory; efficient coding; stimulus statistics; recurrent neural network

Introduction

The efficient coding hypothesis (Barlow, 1961) proposed that the brain should adapt to the statistical properties of the environment to optimize information transmission. While there are many theories on how noisy neural representations should optimally encode the stimulus (reviewed in Kriegeskorte & Wei, 2021; Simoncelli & Olshausen, 2001), previous studies mainly addressed optimal representations problems in feedforward processing with a focus on a static view. However, cognition often involves holding or integrating information over time, which requires recurrent computation. The recurrent computation may impose constraints on the efficient processing, leading to different efficient optimal coding solutions (Bredenberg & Simoncelli, 2020). So far, this remains an open problem.

We investigate this question by studying how working memory (WM) systems should best adapt to the stimulus statistics. Previous research suggested that WM incorporated prior information about stimuli statistics for inference (Honig et al., 2020; Panichello et al., 2019). Here we optimize RNNs to solve a WM task and study how RNNs solve it. It could provide new understandings of how noisy working memory systems could optimally encode stimulus prior.

We found the trained RNNs generally allocate attractor states according to stimulus prior. Furthermore, neural noise promotes the RNNs to develop discrete attractors to achieve a better bias-variance tradeoff. Our results lead to novel predictions of WM at the levels of neural representation, network mechanisms, and behavioral characteristics.

Method

We trained RNNs to perform the delay estimation task (Fig 1). The network activity $\boldsymbol{r}$ follows a dynamical equation:

$$\tau \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = -\mathbf{r} + f\left(W^{\mathrm{rec}}\mathbf{r} + W^{\mathrm{in}}\mathbf{u} + \mathbf{b} + \sqrt{2\tau\sigma_{\mathrm{rec}}^2}\boldsymbol{\xi}\right) \quad (1)$$

where $\mathbf{u}$ is the input to the network, $\mathbf{b}$ is the bias, $\boldsymbol{\xi}$ are independent Gaussian noise processes with zero mean and unit variance and σ_{rec} is the strength of the noise, and $f(\cdot)$ is a nonlinear Sigmoid activation function. We implemented the time-discretized version of Eq (1) with N_{rec} = 128 units, $\mathrm{d}t$ = 75ms and neuronal time constant τ = 100ms.

The stimuli orientation θ is encoded with an array of orientation-selective neurons with von Mises (circular Gaussian) tuning curves. The output of the RNN is a linear readout of the recurrent units. The loss function is computed by the mean squared errors between the network output and the true stimulus. We impose L2 regularizations on recurrent weights and rates to mimic biological resource constraints.

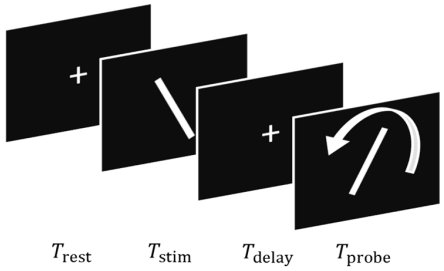

Fig. 1 Working memory task. The agents remember the orientation of the bar when stimulus presents, then maintain the memory and reproduce the memorized orientation.

Results

Continuous v.s. discrete attractors

We trained RNNs to perform the WM task under different noise levels and delay lengths. We found low-dimensional neural dynamics in our trained RNNs (Fig. 2a, g). The neural dynamics resemble a continuous ring attractor (Burak & Fiete, 2012; Compte et al., 2000) when the effective neural

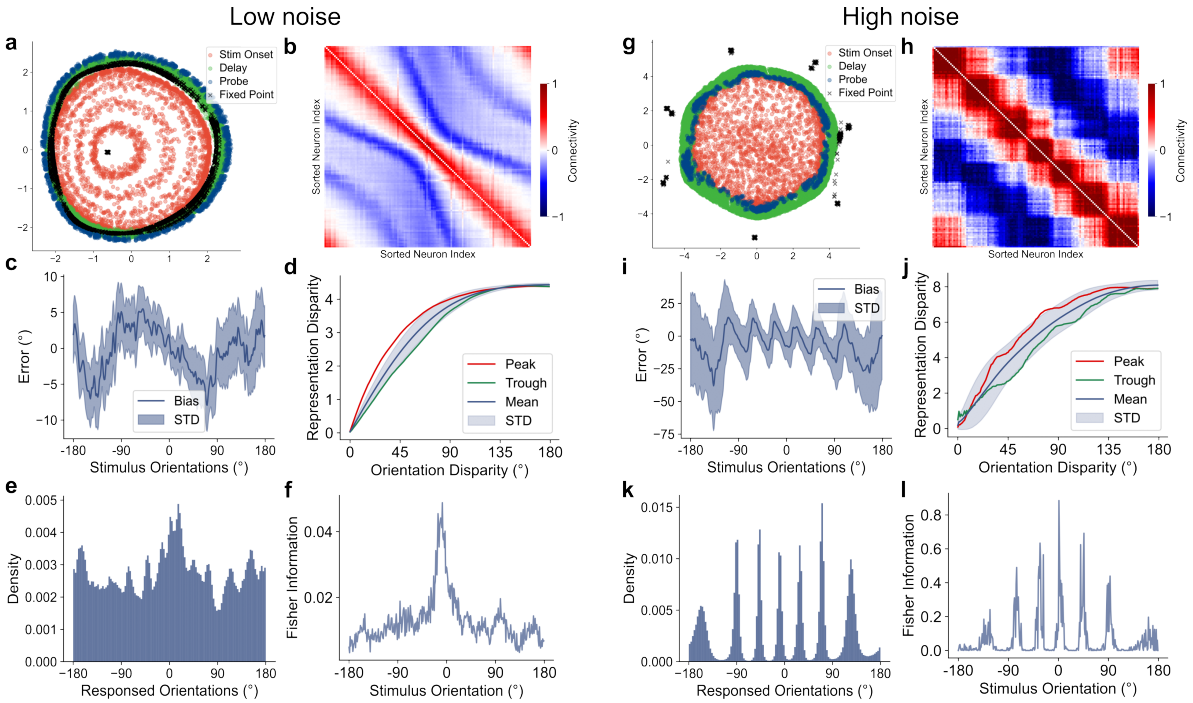

Fig. 2 RNNs under low and high noise trained with data from a von Mises distribution centered at 0 deg and tested with data from a uniform distribution. **a**, **g)** Neural activities in the 2D subspace defined by the first two principal components of PCA. (Variance explained, **a**: 77%, **g**: 93%.) red: stimulus presents; blue: delay period; green: probe phase; black: fixed points. **b**, **h)** Recurrent weights matrix. **c**, **i)** Behavioral bias as a function of the stimulus. **d**, **j)** Representational distance as a function of stimulus disparity. **e**, **k)** Distribution of reported memory. **f**, **l)** Fisher information of stimulus representations.

noise is small (Fig. 2a).

Interestingly, with increased noise or length of the delay period during training, the resulting RNNs exhibit more discrete attractors (Fig. 2g, k). This suggests that RNNs learn to store memories with finite stable states to mitigate noise. The discrete solution makes the WM representation robust at the cost of introducing biases in memory. Overall, the RNNs solve the optimal bias-variance tradeoff and generate an increasingly more discrete solution with increased noise.

Mechanisms of encoding prior information

To understand how RNNs encode a non-uniform prior, we trained them to remember stimuli generated from a von Mises distribution centered at 0 deg and test them with uniform data. We found that RNNs leverage stimulus statistics in training data to optimize behaviors. The squared error of the more frequent stimuli is significantly smaller than the less frequent stimuli in most RNN and task settings. However, the error exhibits a complex pattern at a finer scale. It is not inversely proportional to the prior density of the training stimulus (Fig. 2c, e, i, k), as predicted by previous efficient coding theory (Wei & Stocker, 2015). This pattern is due to the discreteness of the solutions, as each attractor could induce characteristic bias-variance patterns locally.

We identified the fixed points (Sussillo & Barak, 2013) and visualized them in the 2D state space (Fig. 2a, g). We found that the attractors concentrate on the most frequent stimuli. This pattern is also reflected in the bias (Fig. 2c, i) and the distribution (Fig. 2e, k) of reported memory. We then examined the representational geometry (Kriegeskorte & Wei, 2021) of different stimuli by quantifying the representational distance from one stimulus orientation to all other orientations. Our result shows that the more frequent stimuli have more distance from other stimuli (Fig. 2d, j) and tend to have larger Fisher information (Fig. 2f, l). Both results suggest that the representations of more frequent stimuli have better discriminability.

Discussion

Our results suggest that the noisy recurrent circuits could adapt to the stimulus statistics and noise by flexibly distributing the attractor states. The behavioral pattern of our trained RNNs is reminiscent of experimental reports on a similar WM task (Bae et al., 2015; Honig et al., 2020; Panichello et al., 2019).

Our model leads to multiple novel predictions. One such prediction is that the behavioral output will be heavily biased toward the discrete attractors, even when the prior is uniform. This prediction deviates substantially from previous efficient coding theories (Brunel & Nadal, 1998; Wei & Stocker, 2015). To this end, our work highlights the importance of explicitly incorporating circuit constraints when formulating normative theories of brain function.

References

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology. General*, *144*(4), 744–763.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, *1*(01).

Bredenberg, & Simoncelli. (2020). Learning efficient task-dependent representations with synaptic plasticity. *Advances in Neural Information Processing Systems*.

Brunel, N., & Nadal, J. P. (1998). Mutual information, Fisher information, and population coding. *Neural Computation*, *10*(7), 1731–1757.

Burak, Y., & Fiete, I. R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(43), 17645–17650.

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex* , *10*(9), 910–923.

Honig, M., Ma, W. J., & Fougnie, D. (2020). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8391–8397.

Kriegeskorte, N., & Wei, X.-X. (2021). Neural tuning and representational geometry. *Nature Reviews. Neuroscience*, *22*(11), 703–718.

Panichello, M. F., DePasquale, B., Pillow, J. W., & Buschman, T. J. (2019). Error-correcting dynamics in visual working memory. *Nature Communications*, *10*(1), 1–11.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience*, *18*(10), 1509–1517.