Performing decision-making tasks through dynamics of recurrent neural networks trained with reinforcement learning

Conference Paper · September 2023

DOI: 10.1109/DCNA59899.2023.10290321

2 authors:

Roman Kononov
Institute of Applied Physics of Russian Academy of Sciences
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Oleg Maslennikov
Institute of Applied Physics of Russian Academy of Sciences, Nizhny Novgorod, Ru…
39 PUBLICATIONS **321** CITATIONS
SEE PROFILE

2023 7th Scientific School Dynamics of Complex Networks and their Applications (DCNA) | 979-8-3503-0574-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/DCNA59899.2023.10290321

Performing decision-making tasks through dynamics of recurrent neural networks trained with reinforcement learning

Roman Kononov
Nonlinear Dynamics Dept.
Institute of Applied Physics of the RAS
Nizhny Novgorod, Russia
r.kononov@ipfran.ru

Oleg Maslennikov
Nonlinear Dynamics Dept.
Institute of Applied Physics of the RAS
Nizhny Novgorod, Russia
olmaov@ipfran.ru

***Abstract*—In this work, recurrent neural networks are considered as functional models that perform target tasks inspired by cognitive neuroscience experiments. These networks are trained with reinforcement learning methods, after that structure and dynamics mechanisms underlying their behaviors are studied. We use two versions—with and without a context signal—of the cognitive perceptual decision-making task. Population and single-neuron dynamics are studied resulting in stimulus processing as well as successful completing the target tasks. Functionally specialized neurons as well as cluster structure of the trained networks are found and investigated. Similarities and differences between the model neural networks and biological prototypes are discussed.**

***Index Terms*—recurrent neural networks, nonlinear dynamics, reinforcement learning, decision-making tasks, computation through dynamics**

I. Introduction

Computation through population dynamics is a modern framework in mathematical neuroscience aiming at explaining mechanisms of collective neural activity underlying sensor-motor and cognitive phenomena [1]–[3]. A methodology of this framework which uses recent achievements in machine learning is the top-down approach [4], [5]. The starting—top—point is a cognitive task mathematically defined in a properly reduced form. The end—bottom—point is an artificial neural network properly initialized and is trained to perform the task [6]–[9]. There are various ways to come from the top to the bottom, namely, one can choose different architectures, neuron models, and training techniques to design a neural net model. Most of the research in this field use supervised learning methods which are basically biologically inconsistent. Here we use reinforcement learning paradigm as a tool to train recurrent networks of rate-based neurons and focus on two versions of the perceptual decision-making task as target functions. For relating neural dynamics and cognitive activity, the resulting neural networks are considered as multi-dimensional dynamical systems. Population and single-neuron dynamics are studied resulting in stimulus processing as well as successful completing the target tasks. Clustering structure of the trained networks as well as functional specialization of different neurons and clusters are studied.

This research was supported by the Russian Science Foundation (project 23-72-10088).

II. Reinforcement learning

The basic idea of reinforcement learning is the agent–environment setting. Environment is a system that generate observation and reward signals while an agent is a decision-making system. The agent makes actions based on observations and learns to maximize its accumulated reward received from the environment.

At time moment t, the environment is in state $\mathbf{s}_t$ and generates an observation signal $\mathbf{x}_t \in \mathbb{R}^n$ and a reward signal $r_t \in \mathbb{R}$. The agent makes an action $a_t \in \mathbb{Z}$ based on observations and rewards thus being able to change the state of the environment. A chain of actions made by the agent is determined by a (stochastic) function called policy π which defines a probability $\pi(a|\mathbf{s})$ to make the action a in the state $\mathbf{s}$. During reinforcement learning, the policy is subject to change in order to maximize the accumulated reward. In this work, we use the proximal policy optimization (PPO) method of reinforcement learning [10]. Its basic idea is that after the modification of control parameters during training, the new updated policy should not be too far from the old policy. For making this possible, PRO uses so-called clipping to avoid too large updates.

III. Network model

In this work, we use the actor–critic architecture of the neural network where each subnetwork of the actor and critic have the identical structure but two different roles. The actor subnetwork receives the inputs (observations) and produces the outputs (actions) while the critic subnetwork estimates the values of the actions made by the actor. Both subnetworks consist of rate-based artificial neurons described by the rectified linear unit activation function. The following system describes each of the neural subnetworks in the model:

$$\begin{cases} \mathbf{h}_t = relu(\mathbf{x}_t W_{ih}^T + \mathbf{h}_{t-1} W_{hh}^T) \\ \mathbf{out}_t = \mathbf{h}_t W_{ho}^T, \end{cases} \quad (1)$$

where t is discrete time, $\mathbf{x} \in \mathbb{R}^K$ are the observation signals (the input vector) from the environment, K is the dimension of the input vector,

$\mathbf{h}_t \in \mathbb{R}^{50}$ is hidden activity or the network state at time moment t ($\mathbf{h}_{-1} = \mathbf{0}$),

$\mathbf{out}_t \in \mathbb{R}^M$ is the output activity of the network at t,

$W_{ih} \in \mathbb{R}^{50\times K}$ is the matrix of input weights,

$W_{hh} \in \mathbb{R}^{50\times 50}$ is the matrix of recurrent weights,

$W_{ho} \in \mathbb{R}^{M\times 50}$ is the matrix of output weights.

Parameters $M = 1$ for the critic subnetwork and $M = 3$ for the actor subnetwork.

The actor output is a non-normalized probability distribution. The normalization procedure is as follows:

$$\pi(\mathbf{s}_t) = \mathbf{out} - \ln \sum_{j=0}^{M} e^{out_{tj}}$$

The policy $\pi(\mathbf{s}_t)$ allows to produce a choice of action a_t.

The critic output is an estimation of the value of the state for the previous policy. When applying PRO, one uses the critic output to formulate a loss function subject to minimization based on conventional gradient descent methods.

IV. Target task I: Perceptual decision-making

A. Experiment prototype and model

The prototype experiments used here for choosing target tasks, were conducted on monkeys. The animals were shown moving dots on the screen and trained to define to which direction most of the dots were moving [11].

The inputs in this task are modeled by vector $\mathbf{x} \in \mathbb{R}^3$. The first component is a fixation signal which separates the trials: $x_1 = 1$ always except of the decision moment when $x_1 = 0$. The components x_2, x_3 model perceptual variables, see Fig. 1(a). They define the fraction of dots moving towards each of two possible directions and satisfy the condition $x_2 + x_3 = 1$ in the deterministic case without noise.

The single parameter of a trial is coherence $coh = (x_2 - x_3) * 100$ where x_2 and x_3 are taken into account without noise.

Each trial can be separated into four periods:

Fixation period: the fixation $x_1 = 1$, the inputs $x_2, x_3 = 0$.

Stimulus period: the fixation $x_1 = 1$, the inputs $x_2, x_3 \neq 0$.

Delay period: the fixation $x_1 = 1$, the inputs $x_2, x_3 = 0$.

Decision period: the whole vector input $\mathbf{x} = \mathbf{0}$. In this period, the agent produces its meaningful output making its decision.

B. Dynamics of the neural network

To analyze the dynamics of the trained network capable of successful performing the target task, we study projections of its multidimensional activity $\mathbf{h}$ to the principal component plane (PCA method). We found that, during the fixation period there are two fixed points in phase plane each of which can attract neural network trajectories. These fixed points labeled as D_1 and D_2 in Fig.3, encode two alternative decisions

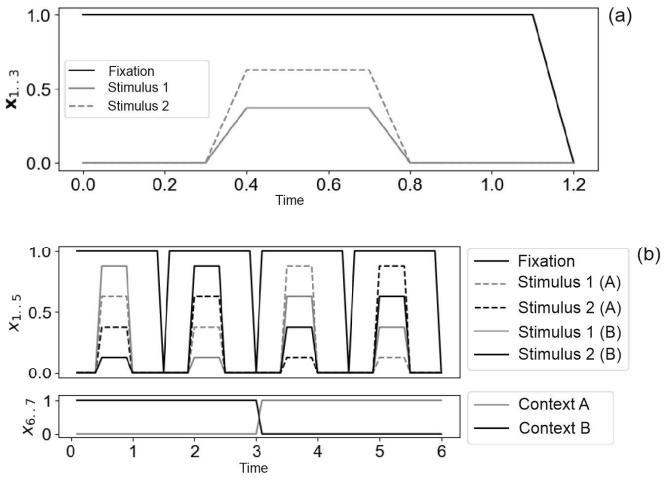

Fig. 1. Structure of input and output signals for neural networks performing target tasks of (a) perceptual decision-making, (b) context-dependent decision-making.

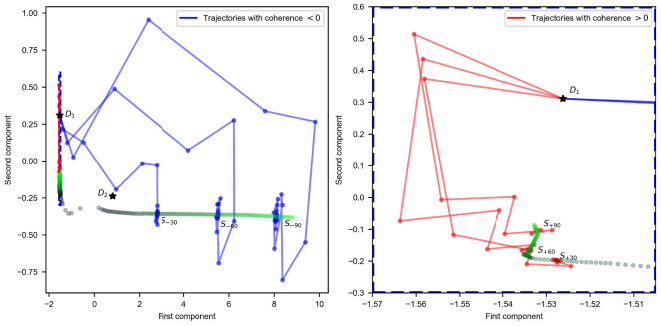

Fig. 2. Projections of the neural network trajectories into the plane of two principal components during the stimulus period: red trajectories for positive values of the coherence, blue ones for negative values. The curve of green-colored points is a projection of the fixed points induced by the stimuli with different values of the coherence.

and can be regarded as a dynamical mechanism of working memory.

When input stimuli switch on, the fixed points disappear. Every pair of stimuli characterized by some coherence induce a new fixed point S_{coh} in the neural phase space. The set of possible values of input stimuli form a special curve of fixed points which makes the basis for network encoding. Thus, a new pair of stimuli results in network activity attracting to a particular fixed point, see Fig.2. Note that during the fixation period, the neural trajectories are attracted by D_1, and after the stimuli appear, they tend to a new fixed point S_{coh}.

The points from the vicinity of S_{coh} corresponding to the positive coherence, when the stimuli switch off, are attracted to D_1 matching choice 2. Those from the vicinity of S_{coh} corresponding to the negative coherence are attracted to D_2 matching choice 1, see Fig.3.

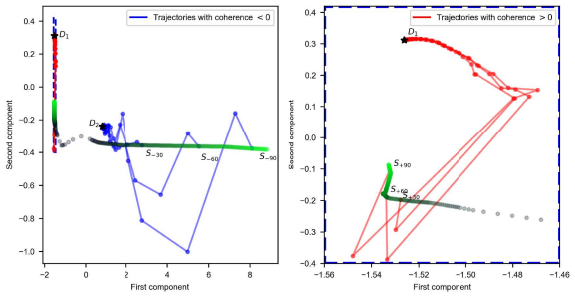

Fig. 3. Projections of the neural network trajectories into the plane of two principal components during the delay period: red trajectories for positive values of the coherence, blue ones for negative values. The curve of green-colored points is a projection of the fixed points induced by the stimuli with different values of the coherence.

C. Activity of neurons

To study functional properties of the neural network after training, we apply an agglomeration algorithm of clustering to the matrix of recurrent weights. After that, one obtain a dendrogram exemplified in Fig.4 which shows the neural network activity during two trials with opposite signs of the coherence. It has been found that most of the neurons were divided between the clusters which are active only during the trials with a certain sign of coherence.

The dynamic mechanism described above is qualitatively preserved when training networks with a different random initialization. The figures show the activity of a particular network with asymmetry in activity amplitude for trials with different coherence signs, but during training, networks with the uniform activity amplitudes were also obtained.

V. Target task II: Perceptual decision-making with context

A. Experiment prototype and model

During another version of decision-making cognitive task, the monkeys were shown moving dots colored with two different colors on the screen. The animals should define depending an a context signal, whether which direction or which color is prevalent in the cloud of dots [12].

The inputs in this task are modeled by vector $\mathbf{x} \in \mathbb{R}^7$ where $x_{1,2,3}$ have the same meaning as in the previous task, $x_{4,5}$ are a pair of stimuli modeling a color of dots, $x_{6,7}$ are the context signals. Only one of $x_{6,7}$ equals 1 while another equals 0 during trials, see Fig. 1(b).

B. Dynamics of the neural network

To analyze the dynamics of the trained network capable of successful performing the target task, we study projections of its multidimensional activity $\mathbf{h}$ to the principal component space. We found that, during the fixation period, there appear two periodical trajectories one of each can attract neural network activity. Thus in this case of the target task, two periodic trajectories encode a decision of the network and comprise a dynamical mechanism of working memory [12].

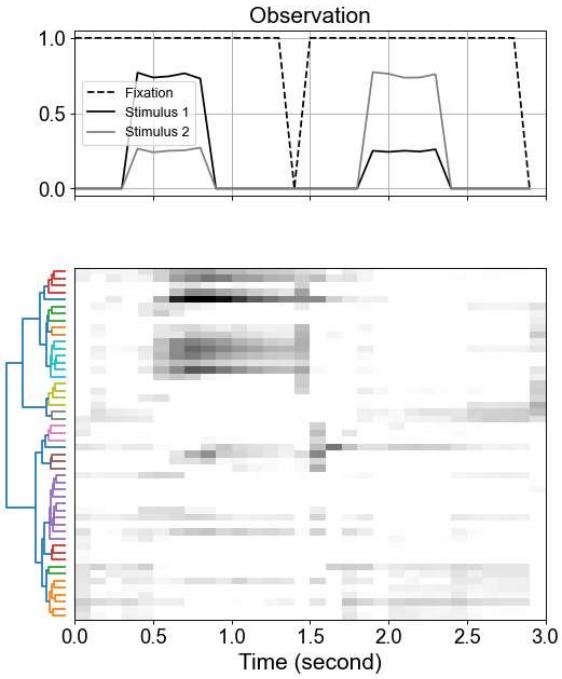

Fig. 4. The input stimuli during two trials (top) with different coherence signs. Corresponding neural network activity for the sorted system according to the agglomeration algorithm of clustering (bottom).

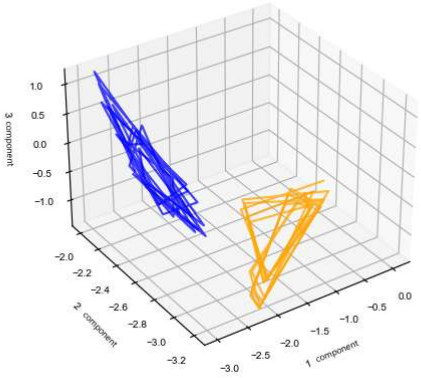

Fig. 5. Periodical trajectories encode a decision of the neural network show in the principal component space.

When stimuli appear, one of the periodic trajectories vanishes and the network activity attracts to the remaining. When stimuli disappear, the network activity trajectories continue moving in the vicinity of the current periodical trajectory. The neural network decision is determined according to which of the two periodic solutions is reached by phase trajectories at the moment of decision making (when $x_0 = 0$).

C. Activity of neurons

The dynamic mechanism described above is qualitatively preserved when training networks with a different random initialization. The figures show the activity of a particular network with asymmetry in activity amplitude for trials with different coherence signs, but during training, networks with the uniform activity amplitudes were also obtained.

To study functional properties of the neural network after training, we apply an agglomeration algorithm of clustering to the matrix of recurrent weights. After that, one obtain a dendrogram exemplified in Fig. 6 which shows the neural network activity during two trials with opposite signs of the coherence for the first context. It has been found that there appear clusters that are active during particular period of the trial for particular contexts. It follows from Fig. 6 that during the fixation period, the trajectories move in the vicinity of the first periodical trajectory.

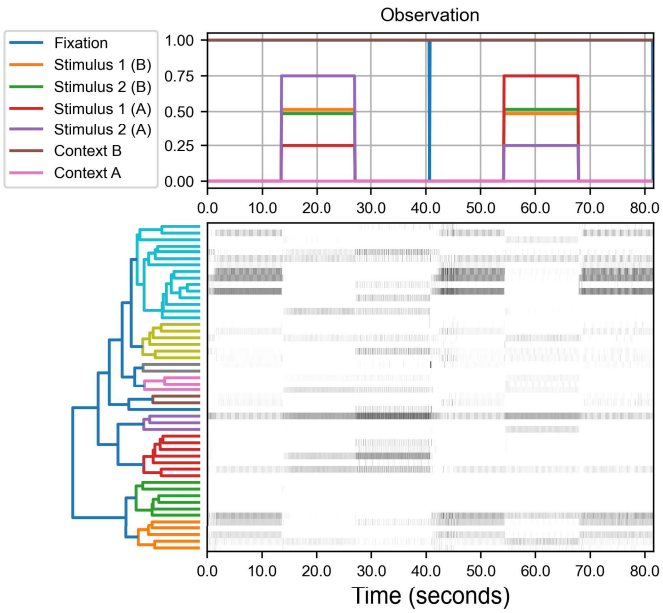

Fig. 6. The input stimuli during two trials (top) with different coherence signs for the same contex signal. Corresponding neural network activity for the sorted system according to the agglomeration algorithm of clustering (bottom).

The figures show the activity of a particular network, but the described dynamic mechanisms are qualitatively preserved when training a network with another random initialization.

VI. Conclusion

We have studied recurrent neural networks trained to perform two target tasks inspired by cognitive neuroscience experiments on perceptual decision making with and without context signals. The training has been performed in the framework of reinforcement learning using the actor-critic coupling architecture and the proximal policy optimization algorithm. We uncovered dynamical and structure mechanisms underlying successful completing of the target tasks. We found special trajectories in the phase space of neural activity capable of working memory formation and show the existence of functionally specialized neurons and neural clusters which are selective to particular periods of trials or particular values of stimuli.

References

[1] O. Sporns, G. Tononi, and R. Kötter, "The human connectome: a structural description of the human brain," *PLoS computational biology*, vol. 1, no. 4, p. e42, 2005.

[2] S. Vyas, M. D. Golub, D. Sussillo, and K. V. Shenoy, "Computation through neural population dynamics," *Annual Review of Neuroscience*, vol. 43, pp. 249–275, 2020.

[3] K. Anokhin, "Cognitome: In search of fundamental neuroscience theory of consciousness," *Zhurnal vysshei nervnoi deyatelnosti imeni I.P. Pavlova*, vol. 71, no. 1, pp. 39–71, 2021.

[4] O. Barak, "Recurrent neural networks as versatile tools of neuroscience research," *Current opinion in neurobiology*, vol. 46, pp. 1–6, 2017.

[5] D. Sussillo, "Neural circuits as computational dynamical systems," *Current opinion in neurobiology*, vol. 25, pp. 156–163, 2014.

[6] O. V. Maslennikov, M. M. Pugavko, D. S. Shchapin, and V. I. Nekorkin, "Nonlinear dynamics and machine learning of recurrent spiking neural networks," *Physics-Uspekhi*, vol. 65, no. 12, 2022.

[7] O. V. Maslennikov and V. I. Nekorkin, "Stimulus-induced sequential activity in supervisely trained recurrent networks of firing rate neurons," *Nonlinear Dynamics*, vol. 101, no. 2, pp. 1093–1103, 2020.

[8] O. V. Maslennikov, "Dynamics of an artificial recurrent neural network for the problem of modeling a cognitive function," *Izvestiya VUZ. Applied Nonlinear Dynamics*, vol. 29, no. 5, pp. 799–811, 2021.

[9] M. M. Pugavko, O. V. Maslennikov, and V. I. Nekorkin, "Dynamics of spiking map-based neural networks in problems of supervised learning," *Communications in Nonlinear Science and Numerical Simulation*, vol. 90, p. 105399, 2020.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[11] K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon, "The analysis of visual motion: a comparison of neuronal and psychophysical performance," *Journal of Neuroscience*, vol. 12, no. 12, pp. 4745–4765, 1992.

[12] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex," *nature*, vol. 503, no. 7474, pp. 78–84, 2013.