Neurological Based Timing Mechanism for Reinforcement Learning

Extended Abstract

Michael J. Tarlton
Oslo Metropolitan University
Oslo, Norway
michaelt@oslomet.no

Gustavo B. Mello
Oslo Metropolitan University
Oslo, Norway
gustavom@oslomet.no

Anis Yazidi
Oslo Metropolitan University
Oslo, Norway
anisy@oslomet.no

ABSTRACT

The inherently time-dependent dynamics which underly the neuronal spiking communication, are ubquitous throughout brain, and yet are not fully understood. Likewise time-based mechanisms are underdeveloped in the field of Machine and Reinforcement Learning (RL) [7]. The complexity-rich and multi-dimensional dynamics observed in the brain offer potential advancements in Machine Learning (ML), and development of Artificial Generalized Intelligence.

It is in our interests to model known time-mechanisms of neuronal spiking communication, and reproduce the emergent properties of complex timing and learning in assemblies. If neuronal temporal dynamics can be understood, a new field of possibilities will open for *in-situ* models which learn in complex real-time environments. A key challenge for these models is correctly identifying associations of actions and stimulus at variable time separations. In this article, we bring the flexible time representation mechanisms from neuroscience to the field of automata and RL, to explore its potential.

KEYWORDS

Neuro; AI; RL; Automata; Periodicity; Striatal Beat Frequency

ACM Reference Format:
Michael J. Tarlton, Gustavo B. Mello, and Anis Yazidi. 2024. Neurological Based Timing Mechanism for Reinforcement Learning: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTERVAL TIMING

Interval Timing (IT): the internal process by which animals perceive and estimate the duration of time between events. IT functions as an internal clock that helps gauge how much time has passed and enables organisms to not know what to do but when to do it, even in the absence of cues. IT is crucial for causal inference, decision making, and estimation of reward value; consequently, it is fundamental to understanding how RL operates in living organisms [7]. IT is the subject of extensive research within neuroscience, and multiple models have been developed to explain how ensembles of neurons operating at millisecond scale can flexibly scale their activity to represent a wide range of intervals scaling from seconds to several minutes.

Researchers have devised several behavioral experiments to access IT. Perhaps the most widely used is the Fixed-Interval (FI) task [8] [2], where an animal is conditioned to perform an action (e.g., press a lever) after a target interval of time (the **criterion time**) to receive a reward. Neurophysiological studies have shown that this scaling property is present in the activity of brain cells. This reveals the potential for flexible, efficient coding. By implementing these computational models of timing in automata, we would expect an artificial agent with similar timing properties. That is, able to reproduce the activity patterns observed in animal behavior in this task. Namely **temporal rescaling** of the neuronal **receptive fields**: changes in the mean peak and variance of motor neuron activity over time, when learning a new criterion time.

2 THE SBF MODEL

The Striatal Beat Frequency (SBF) model [4] [5], is a neuroscientific theory for explaining IT in the brain. The SBF encodes associations of events in time, on the state in the neural ensemble with a reward mediated mechanism. These make the model apt for use in artificial neural networks and RL frameworks. The SBF model has been implemented with real-time asynchronous learning systems such as Spiking Neural Networks and Spike-Timing Dependent-Plasticity [9]. The SBF model is well-supported by computational simulations [1] and has been considered as a viable model for flexible and distributed time-information encoding.

To the best of our knowledge, the SBF model has not yet been introduced to the ML domain. If successfully introduced, the SBF model can be applied to problems of periodic event learning in real-time environments, e.g. Web-Crawling Optimization [3]. The original SBF model as described by the authors was created in simulation with a neurobiologically plausible ensemble of thousands of neurons [1]. In introducing this method to ML, we simplify the model in a naïve RL-automata framework. We represent large neural population activities with single units. These units are active with unique, fixed, oscillating activity. Using multiple units ("***oscillators***"), each with a unique period to its oscillation. The activity on the oscillators is then projected to an "executive" unit, which "decides" whether to act or not. The weighted strengths between the oscillators and the executive unit allow the periodic activity to act as predictive stimuli of meaningful events in the environment. Each time the executive unit acts, the weights change in response to the presence or absence of a reward. This is done continuously until the weights converge to a solution. The weights reflect the

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

contribution each oscillator makes towards predicting the correct periodicity. Because SBF model doesn't rely on a centralized "clock", nor explicitly keep time information in memory. Instead, SBF leverages distributed, ongoing sensiomotor activity, to make and keep inferences about time [1]. This is consistent with the perception that timing mechanisms are not a singular, modular aspect of the brain, but foundational to all neuronal communication [2].

3 THE SBF-AUTOMATA

We introduce the SBF-Automata (SBF-A) model, a naïve RL model which replicates features of the SBF. The SBF-A is composed of two main parts: 1. the **oscillator block**, containing a set of oscillator units, each of which "peaks" in activity at a unique periodicity; and 2. the **executive unit**, which integrates the activity of the oscillator block and decides whether to act or not. Where the probability of taking action is the summed weights of active units: $P = \sum w_a$.

The oscillator units act discretely; unit activity is "on" at a particular time-step in the cycle coinciding with its period, and otherwise "off". For our experiments, the oscillators' cycle periods are selected from a uniform distribution with a minimum period zero, to a maximum period of 0.9 times the criterion time (to prevent having an oscillator with an exact solution). The "zero-period" oscillator acts as a broad inhibitory signal, which votes against acting on every time-step. This absorbs excess weight in the system, which could force hyperactivity.

When an action is taken in a correct time-step (corresponding to the cyclic criterion time), a feedback signal triggers a "reward" update to the weights. Oscillators whose activity contributed to taking the action have their weight (w_a) strengthened, $w_a = w_a + \alpha \frac{\sum w_i}{n_a}$; while oscillators who were inactive have their weight (w_i) weakened $w_i = w_i \times (1 - \alpha)$. Additionally, to balance exploration and exploitation, whenever the executive unit takes action on an incorrect time-step, weights to the oscillators that contribute to the wrong action are penalized and others are strengthened. Both of these methods enable faster convergence towards a solution.

3.1 Experiment

We implemented a version of the FI task [2], in which the environment consists of discrete time-steps, and the reward is periodically available on time-steps corresponding to the criterion time interval. The SBF-A must learn the periodicity using only feedback from previous actions. Experiments used oscillator sets of sizes: [100, 600, 900] and criterion times: [1000, 6000, 10,000]. Results for each set of conditions were averaged over 20 experiments.

3.2 Results

For each interval of the criterion time, we compare the total number of actions taken against the number of correct actions taken as a measure of accuracy, and cross compare performance for different initializations of oscillator size and criterion times (Figure 1). In a broader trend, the reward retrieved seems to depend on the average actions taken. As the size of the oscillator set increases, both the reward and the action rate decrease drastically; with the variance of the action rate increasing with larger criterion times. While this might imply instability, the distribution of oscillator weights over the lifetime of each experiment shows the automata consistently arrives at a unique solution for each set of conditions. We include an example in Figure 2 The final distribution of the weights prefers a few select oscillators, while the majority of weights approach zero.

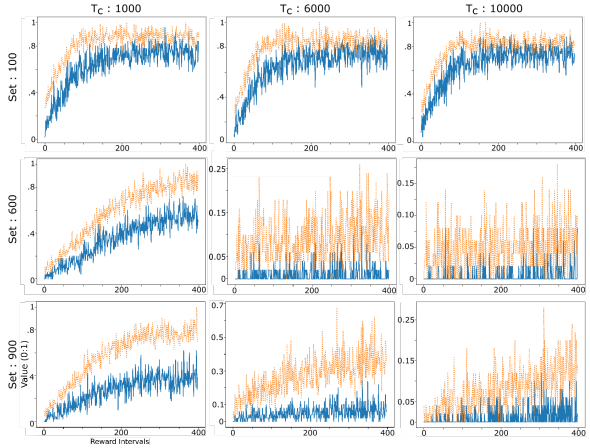

Figure 1: Blue: Average amount of reward recovered for each interval. Orange: Average number of actions per total of each interval. Rows are the size of the oscillator set. Columns are the criterion time (T_c) used.

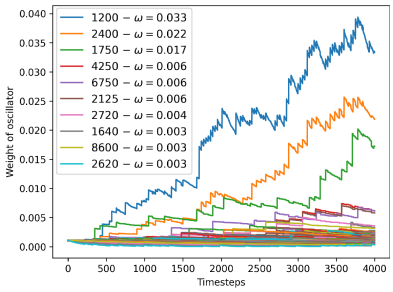

Figure 2: Oscillator weights over the lifetime of experiment (T_c = 10000, set size: 900). Oscillator periods are denoted by color, the legend is the top-ten highest weighted oscillators at the end of the experiment, sorted by weight.

3.3 Conclusion

These results show that the SBF-A is not particularly effective in the sub-thousands of oscillators regime. We even see a trend opposite of our expectation, where increasing the number of oscillators in a set leads to less reward being recovered on average. However, the decreased and varied average number of actions per interval in experiments with larger oscillator sets could indicate a more "accurate" (at least in the sense of when to not act) performance.

This is not entirely unexpected as we use extremely small populations of oscillators with nonoverlapping domains (comparatively [1] uses 15,000, normally distributed oscillators). We still see promise in that the SBF-A still arrives at some unique solution for each problem; however, as the oscillator activity is discrete, this causes the SBF-A to favor a few large oscillators with periods which evenly divide into multiples of the criterion time.

This model is oversimplified for the sake of clarity and cannot fully exhibit the performance we would expect to see from other SBF simulations [1] [6]. Even so, we can observe it to be sensitive to the conditions of the environment and capable of "learning" in the weight distribution. Further development of the SBF-A will use phasic oscillator activity domains to emulate large neuron populations.

REFERENCES

[1] Melissa J. Allman and Warren H. Meck. 2012. Pathophysiological Distortions in Time Perception and Timed Performance. *Brain* 135, Pt 3 (March 2012), 656–677. https://doi.org/10.1093/brain/awr210

[2] Gustavo Borges Moreno e Mello. 2016. *Neural and Behavioral Mechanisms of Interval Timing in the Striatum.* Ph.D. Dissertation.

[3] Shuguang Han, Michael Bendersky, Przemek Gajda, Sergey Novikov, Marc Najork, Bernhard Brodowsky, and Alexandrin Popescul. 2020. Adversarial Bandits Policy for Crawling Commercial Web Content. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 407–417. https://doi.org/10.1145/3366423.3380125

[4] Matthew S. Matell and Warren H. Meck. 2000. Neuropsychological Mechanisms of Interval Timing Behavior. *BioEssays* 22, 1 (2000), 94–103. https://doi.org/10.1002/(SICI)1521-1878(200001)22:1<94::AID-BIES14>3.0.CO;2-E

[5] Matthew S. Matell and Warren H. Meck. 2004. Cortico-Striatal Circuits and Interval Timing: Coincidence Detection of Oscillatory Processes. *Brain Research. Cognitive Brain Research* 21, 2 (Oct. 2004), 139–170. https://doi.org/10.1016/j.cogbrainres.2004.06.012

[6] Sorinel A. Oprisan, Dereck Novo, Mona Buhusi, and Catalin V. Buhusi. 2022. Resource Allocation in the Noise-Free Striatal Beat Frequency Model of Interval Timing. *Timing & Time Perception* 11, 1-4 (July 2022), 103–123. https://doi.org/10.1163/22134468-bja10056

[7] Elijah A. Petter, Samuel J. Gershman, and Warren H. Meck. 2018. Integrating Models of Interval Timing and Reinforcement Learning. *Trends in Cognitive Sciences* 22, 10 (Oct. 2018), 911–922. https://doi.org/10.1016/j.tics.2018.08.004

[8] Joshua E. Swearingen and Catalin V. Buhusi. 2010. The Pattern of Responding in the Peak-Interval Procedure with Gaps: An Individual-Trials Analysis. *Journal of experimental psychology. Animal behavior processes* 36, 4 (Oct. 2010), 443–455. https://doi.org/10.1037/a0019485

[9] Wei Xu and Stuart N. Baker. 2016. Timing Intervals Using Population Synchrony and Spike Timing Dependent Plasticity. *Frontiers in Computational Neuroscience* 10 (2016).