A bio-inspired minimal model for non-stationary K-armed bandits

Krubeal Danieli, Mikkel Elle Lepperød

December 16, 2024

Contents

1 Introduction

The ability to make decisions for long-term reward maximization is a fundamental aspect of cognition. The brain is has evolved specialized and interconnected regions to implement this behaviour under the constraints of biology.

Well-studied ecological settings of decision-making are foraging tasks, such as food search. In this problems, the agent is usually asked to choose between different options to maximize an expected reward. In nature, animals have been shown to exhibit different strategies depending on context. *Matching behaviour* is a well-known phenomenon in which the animal's decision patterns are proportional to the reward probability of the available options. Such behaviour is thought to result from the trade-off between exploration and exploitation [1, 2]. In fact, this is a well known phenomenon in the reinforcement learning literature, in which an agent is faced with the dilemma of exploring new alternatives, potentially more rewarding, or exploiting known options, despite being possibly sup-optimal.

A popular formalization of these type of tasks is the *multi-armed bandit problem* (MABP) [3]. This setting is usually described in terms of a slot machine endowed with K distinct levers, also called arms. During a round, the agent selects one of the levers and collects a reward R according to an unknown reward probability specific to the chosen lever. The goal is simply to maximizes the total reward after a given number of steps, which is achieved by effectively updating a selection policy after each round. This problem has been extensively studied in the context of reinforcement learning, and it is considered a fundamental building block for more complex tasks [1].

There exist various flavours of this problem, with the simplest having a stationary reward distribution. Over the years, several algorithms have been proposed, alongside with their theoretical guarantees. In this regard, Thompson sampling is a popular algorithm that has been shown to achieve near-optimal regret bounds in the stochastic setting [4, 5]. This approach relies on Bayesian optimization, where the goal is to maintain a posterior distribution over the reward probabilities of the actions, and select actions accordingly. Another popular algorithm is Upper Confidence Bound (UCB), which has been shown to achieve near-optimal regret bounds in the adversarial setting [6]. The approach is based on the idea of maintaining an upper limit on the reward probabilities of the actions, and select actions accordingly. Other successful algorithms are ϵ-greedy and VDBE [7, 8, 9, 10]. Nonetheless, despite their success they have little resemblance to neural dynamics nor clear functional similarity to brain regions.

In this work, we propose a biologically plausible algorithm using rate neurons, whose hyper-parameters are optimized through an evolution search. The benchmarks we chose are stochastic bandit problems, more challenging variants

of the original task endowed with *concept drift*, where the reward distribution changes over time [11, 12, 13].

The architecture of our model consists of two connected neuronal layers, both with as many neurons as the arms of the bandit task. The first layer is inspired by the functionality of the orbitofrontal cortex (OFC), and its scope is to maintain an active representation of the arms weighted by the input from the second layer. Moreover, this are is thought to be involved in motivation and representation of the expected value of the actions, either positive or negative [14, 15, 16], action selection in uncertain environments [17], and contextual processing [18]. The second layer is instead modeled after the ACC, and it is meant to represent the value of the arms. Its input connections are updated through a learning rule dependant on the reward history and current connectivity pattern.

Our model features two important aspects of the brain during decision making. Firstly, the option selection process itself is implemented as a dynamical interaction between neural populations, similarly to bump attractor networks for perceptual cognition [19, 20]. The final choice of the arm is achieved by the agreement or disagreement between the two populations, and it depends on their underlying value representation [21, 22].

Secondly, plasticity is based on a non-associative learning rule, endowed with a non-linear kernel for the weight update term. Behind this design choice there is our hypothesis that the scale of the synaptic update should vary non-linearly according to its magnitude. This consideration is aligned with the idea that the learning rate is a parameter specific to each neuron. This synapse-type specific plasticity (STSP) [23] is a unction of the resources available at the synaptic button and its state, including the size [24, 25, 26]. This approach has been already adopted in several computational architectures, for instance in spiking neural networks [27] and for synaptic metaplasticity [28]. Lastly, there is experimental evidence that this adaptation function might be covered by dopamine [29]. Indeed, its involvement in calculating prediction errors and reward signaling is well established [30], as well with its modulation of high-level cortical networks like the PFC [31, 32, 33].

2 Methods

The following section is organized as follows. First, we introduce a formalization of general problem setting, together with the variants considered in this work. Then, we outline the architecture of our model and how it can be mapped to neurobiology. Finally, we describe the learning procedure, and showcase its dynamics in a simple example.

2.1 Binomial K-armed bandit problem

The standard formulation of the task is structured as a set of K arms (or levers) $\mathcal{A}_K = \{a_1 \dots a_K\}$, with an associated reward distribution $\mathbf{p} = \{p_1, \dots p_K\}$. At

each iteration, the agent pulls an arm and collects a possible reward drawn as a Bernoulli variable $R \sim \mathcal{B}(\{0,1\}, p_k)$. The agent's objective is maximizing the total reward $\sum_t^T R_t$, after a certain number T of rounds, also called horizon. Importantly, the agent is unaware of the true reward probabilities, and thus has to make its decisions following a certain policy, denoted as π. In the reinforcement learning literature, the policy is often defined as a distribution over actions, here the arms $\mathcal{A}_K$, given the current state at time t. In the bandit problem, the state can be taken to correspond to the history h_t of past actions and rewards in the period $(0 \dots t]$, and the policy as a function that return a selected arm $\pi(h_t) = a_t$ [34].

Given the inherent stochasticity of the feedbacks from the environment, the policy is affected by the so-called exploration-exploitation trade-off, which here is phrased as the contrast between the option of the arm with the estimated highest expected reward versus the option to explore other arms, so to gather more information. A common approach is the $\epsilon-$greedy policy, where the choice to explore is selected with a probability ϵ. Moreover, it is often preferable to have a more explorative behaviour early during the training, with the intent to have a good sample size for the empirical reward distribution, which can be later exploited for maximizing reward.

Another important concept in multi-armed bandit problems is *regret*. Intuitively, it quantifies the loss of reward due to following a certain policy, and it is determined by the difference beween the collected reward and the theoretical optimal, obtained by choosing the best arm at each round. Formally, defined a function $r(\pi)$ which returns the expected reward while following policy π, then the regret over an horizon T can be formulated as:

$$\rho = \frac{1}{T} \sum_t^T p_t^* - r(\pi(h_t)) \tag{1}$$

where p_t^* is the expected reward of the optimal arm at time t, which correspond to its probability since it is a Bernoulli distribution. The goal of the agent is to minimize the regret, and thus maximize the total reward.

2.2 Model description

The model is constructed as a rate network of two populations of neurons U and V, the former representing the memory trace of the K available options (*i.e.* the bandits), and the latter the value of the options under the current policy. More formally, the model is defined by a set of coupled ordinary differential equations (ODEs). The first equation tracks the evolution of the neural activity $\mathbf{u}$ of population U, while the second tracks the activity $\mathbf{v}$ of the population V. The time constant τ is the same for both equations and it is set to 10ms.

$$\begin{aligned} \tau \dot{\mathbf{u}} &= -\mathbf{u} + \mathbf{W}^{VU} \phi_v(\mathbf{v}) + \mathbf{I}_{\text{ext}} \\ \tau \dot{\mathbf{v}} &= -\mathbf{v} + \tilde{\mathbf{W}}^{UV} \phi_u \mathbf{u} \end{aligned} \tag{2}$$

The external input $\mathbf{I}_{\text{ext}}$ is a constant input that is used to set the initial conditions of the neural activity $\mathbf{u}$. The activation functions ϕ_v, ϕ_u are applied to population v, and represent two distinct neural response function tailored to each population. They have been chosen to be a step-function with threshold θ_v, θ_v applied to a generalized sigmoid with gain g_v, g_u and offset s_v, s_u.

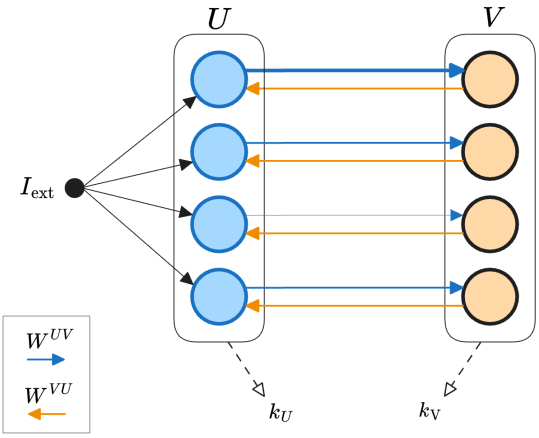

Figure 1: Model architecture - *The model is composed of a layer U (blue), receiving a feedfoward input I_{ext}, a layer V (orange), and connections $\boldsymbol{W}^{UV}$ and $\boldsymbol{W}^{VU}$. Additionally, two indexes k_U, k_V can be extracted from the layers and corresponds to the selection made by the two populations as $k_U = argmax_k\{\boldsymbol{u}\}$, $k_V = argmax_k\{\boldsymbol{v}\}$.*

Importantly, the two layers are not fully connected and the matrices are diagonal. More in detail, the weight matrix $\mathbf{W}^{VU}$ is simply made of 1s, while $\tilde{\mathbf{W}}^{UV}$ is a function of the actual weights $\Phi_v(\mathbf{W}^{UV})$ and it represents the contribution of the active options $\mathbf{u}$ to the value representation $\mathbf{v}$, it is thus referred to as *option value function*. The function Φ_v is defined a weighted sum of a generalized sigmoid and a Gaussian, whose shape is characterized by a bell curve smoothly settling to a constant value. For details see in the appendinx 5. Since the exact values of the model hyperparameters are optimized through an evolution, our motivation behind the structure of Φ_v is to be agnostic about its final form, and allow competition or integration of two distinct traits of the function shape. One is a smooth transition to a plateau value with a certain steepness (or gain), which can represent a saturation once a threshold is crossed, such feature has been reporeted for both biological and artificial neurons [35, 36]. The other is a bell-shaped curve with a defined center and width, which can allow for placing emphasis on values only within a given window and modulate information transfer [37].

2.2.1 Option selection

The decision-making process within a single round is structured in two distinct phases. Initially, the model receives a constant external input targeting all neurons in the memory population U equally. During this phase, $\mathbf{I}_{\text{ext}}$ works as an equilibrium value while the reciprocal interactions with population V push $\mathbf{u}$ to different values, depending on the current policy encoded in $\tilde{\mathbf{W}}^{UV}$. Importantly, the weights $\mathbf{W}^{UV}$ are initialized to zero, and thus the input from U to V is uniform. This approach ensures the absence of biases towards any arm by having all weights equal, and corresponds to a completely untrained network. After a fixed amount of time $\sim$ 2s, the second phase begins. Here, the external input is removed and the model is left to evolve autonomously, and since there are no recurrent connections in neither population the dynamics are entirely driven by their coupling. A selection k is sampled after another fixed amount of time $\sim$ 5s, and it is defined according to the following rule:

$$k = \begin{cases} \text{argmax}_k\{\mathbf{v}\} & \textit{if } \text{argmax}_k\{\mathbf{v}\} = \text{argmax}_k\{\mathbf{u}\} \\ \text{random}(K) & \textit{otherwise} \end{cases}$$

The selection rule is simple: if the value representation $\mathbf{v}$ is in agreement with the memory trace $\mathbf{u}$, then the option with the highest value is selected. Otherwise, a random option is chosen. This rule is a way to express the exploration-exploitation trade-off, and it is dependent on the current policy $\tilde{\mathbf{W}}^{UV}$.
Below 2.2.1, is reported the pseudo-code for algorithm behind the selection process, which is applied during each round t.

Algorithm 1: Two-phases option selection process

Input: External input $\mathbf{I}_{\text{ext}}$, population $\mathbf{u}$, population $\mathbf{v}$, weights $\tilde{\mathbf{W}}^{UV}$
Output: Selected action k
Phase 1: *external input* ; `// Duration: ~2s`
Define constant $\mathbf{I}_{\text{ext}}$;
Update populations $\mathbf{u}, \mathbf{v}$ according to 2.2;
Phase 2: *autonomous evolution* ; `// Duration: ~2s`
Remove external input $\mathbf{I}_{\text{ext}}$;
Let system evolve through population coupling according to 2.2;
Selection process:;
$k_u \leftarrow \text{argmax}_k\{\mathbf{u}\}$;
$k_v \leftarrow \text{argmax}_k\{\mathbf{v}\}$;
if $k_u = k_v$ **then**
  $k \leftarrow k_v$; `// Exploitation`
else
  $k \leftarrow \text{random}(K)$; `// Exploration`
end
return k

In figure 2 it is shown the history of selections over three trials. The initial rounds feature higher variability. In particular, it can be noted how the policy adopted by the model encounters periods of exploration and successive settling over an explotative strategy, which can be reverted in case of a change in the environment's reward distribution.

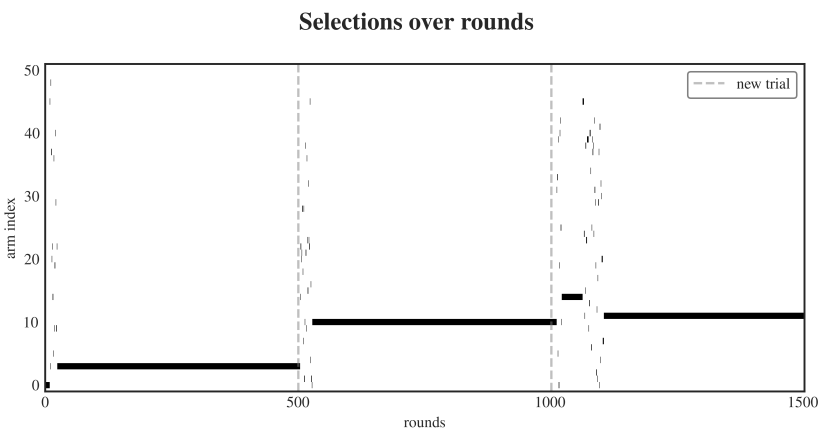

Figure 2: SELECTION EVOLUTION OVER ROUNDS - *the y-axis represents the available arms, while the x-axis the number of rounds, with the dotted vertical lines indicating the start of a new trial with 300 rounds each. The model selections are the black vertical lines for an arm and a round. The red horizontal lines signal the arm with the highest reward probability, representing the best option.*

2.3 Learning

Given a selected option k, the environment (set of bandits) samples and returns a reward $R \in \{0, 1\}$ with probability p_k. Then, the connections $\mathbf{W}^{UV}$ for the neuron corresponding to the option k are updated according to the following plasticity rule:

$$\Delta \mathbf{W}_k^{UV} = \tilde{\eta}_k \left(R \cdot w^+ - \mathbf{W}_k^{UV} \right) \tag{3}$$

where w^+ is a constant maximum synaptic weight, while $\tilde{\eta}_k$ is the learning rate for the option k determined by a function Φ_η of the current weights $\mathbf{W}_k^{UV}$, referred to as *learning rate function*.

The shape of Φ_η is again a Gaussian-sigmoid but with different parameters, giving evolution the opportunity to combine the shape traits of plateau and bell-shaped tuning in a task-efficient manner. In particular, these features can be combined so to define mechanisms of synapse-type specific plasticity as a function of the current synaptic strenght [23], as well the application of other

useful homeostatic constraints with computational advantages, such as synaptic scaling and proportional updates [38, 39, 40].

2.4 Evolution search

The optimization of the hyper-parameters was performed using the Covariance Matrix Adaptation evolutionary strategy algorithm (CMA-ES) [41]. The search was run with a population of 256 individuals (unique set of genomes corresponding to a sample in parameter space) for 80 generations. The fitness function was defined as the average reward obtained by an individual over 3 different non-stationary bandit environments, each for $K = \{50, 300\}$, and all averaged over 2 iterations. The results are summarized below in figure 3.

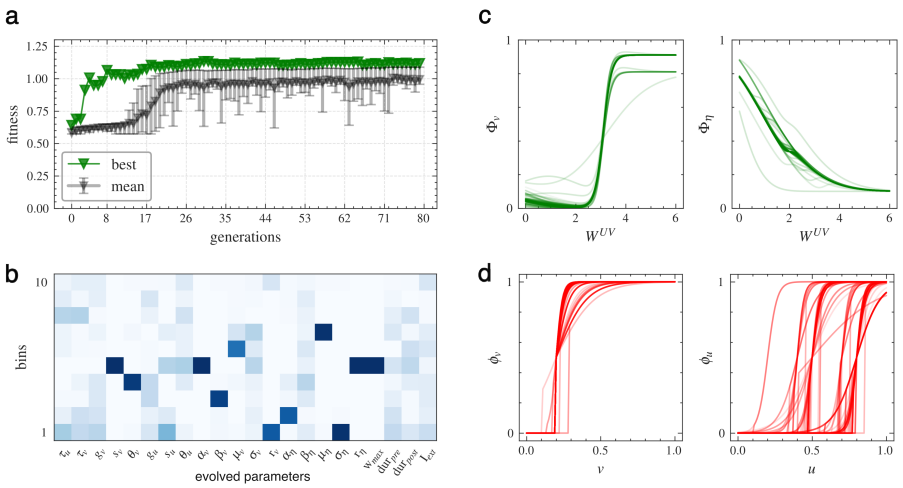

Figure 3: EVOLUTION RESULTS - **a**: *top fitness, mean and standard deviation (as 16-84 percentile) of the population over generations.* - **b**: *heatmap of the evolved parameters (rows) as histogram bins (y-axis) calculated from the 50 percentile of the population of the last generation; higher density is in dark blue* - **c-d**: *Gaussian-sigmoid [c] and neural response functions [d] of the top-half of the population, color intensity proportional to the fitness*

The evolution results show a steady improvement of the fitness over generations, before hitting a plateau corresponding to the theoretical optimal (~ 0.9) of the chosen simulations.

TELL ME ABOUT THE RESPONSE FUNCTIONS, LIKE IS IT A TYPE II NEURON FR? Regarding the evolved functions, the option value function Φ_v is characterized by a steep sigmoid curve, with a marked. This is consistent with the idea that the input of population U to population V is weighted maximally for high option values (strong synapses), whereas for weaker estimates the contributions are low or close to zero, allowing for more

exploration.

The learning rate function Φ_η is instead characterized by a marked bell-shaped curve, given a parameter $r = 0.06$. The associated Gaussian has a positive mean located at $\mu = 1.$, which aligns approximately with the local valley of the weight function Φ_v. A possible interpretation is that it serves as a mechanism to ensure that the learning rate is high when the value options are more uncertain, and low otherwise, thus preventing overshooting and oscillations in the weight updates. This adaptive behaviour is in line with known neuronal dynamics such as homeostatic plasticity, which works towards a stabilization of synapses, for instance through synaptic scaling and proportional updates [38]. A variable learning rate is an important feature of several plasticity rules, from the more biologically plausible like the Oja [42] to deep learning optimizers like Adam [43].

2.5 Bio-inspired features

The model is inspired by the functioning of the prefrontal cortex (PFC) and its importance in decision-making processes. In particular, the two population U, V of the model can be related to the orbito-frontal cortex (OFC) and anterior cingulate cortex (ACC), respectively. More specifically, the OFC is known to be involved in the representation of the state different options and update their value with respect to rewarding outcomes and their history [44, 45]. The ACC has been associated to action values and influencing the exploration-exploitation assessment [46]. Further, its dynamic interplay with the OFC is observed to elicit transient pre-stimulus activation, which biases the decision towards the most valuable option [47, 48, 49].

In the model, the first layer represents the available options, while the learned connections with the second layer encode their values based on the recent reward history. Another similarity with this particular pre-frontal circuit is the realization of a choice as a sample of the network state after a period of autonomous neural activity, where the stability of the neural activations depend on the strength and reliability of the highest option value [50, 51]. Moreover, the application of the function Φ_v on the connections $\mathbf{W}^{UV}$ can be regarded as meta-plasticity, mediated by a neuromodulator [52].

3 Experiments

The model has been tested in a series of benchmark environments, each with a different number of arms and reward distributions. The performance has been compared with the following algorithms: Random Baseline, Upper-Confidence Bound (UCB), Thompson Sampling, and Epsilon-Greedy.

3.1 Game variants

Our goal in this work is to investigate the performance of the agent in a non-stationary environment with Binomial reward distributions, meaning that its underlying distribution changes over time [1] We choose this setting as it resembles an ecological scenario in which an animal has to forage in an environment with food (reward) is distributed over a set of fixed locations, but whose occurrence probability can change over time. More specifically, we used four different variants obtained by introducing different types of non-stationarity: piecewise constant, uniformly changing, sinusoidally changing, and sinusoidally changing plus piecewise constant. The reason for this choices is to test the model performance under different speed and uniformity of the distribution changes. Figure 4 visually illustrates their specificities.

Piecewise stationary environment [KAB-P]
Within a trial the reward distribution is stationary and it is drawn from a uniform $\mathbf{p} = \mathcal{U}(0,1)^K$. At the end of each trial i it is drawn a new distribution $\mathbf{p}_i \rightarrow \mathbf{p}_{i+1}$ [34].

Piecewise stationary environment with drift [KAB-D]
At the very beginning, the reward distribution $\mathbf{p}$ is sampled from a uniform $\mathbf{p} = \mathcal{U}(0,1)^K$. Then, it changes gradually over the rounds, tracked as time t, such that its values tend towards a target distribution $\mathbf{q}_i$ as $\tau_p \dot{\mathbf{p}}_t = \mathbf{q}_i - \mathbf{p}_t$. Here, $\dot{\mathbf{p}}$ is the time derivative of the distribution and τ_p is its time constant. Once the distance is below a threshold δ as $|\mathbf{q}_i - \mathbf{p}_t| < \delta$, the target distribution is changed to a new one $\mathbf{q}_i \rightarrow \mathbf{q}_{i+1}$. In this variant, there are no proper trials but the target distribution keep changing until a maximum number of rounds is reached.

Sinusoidal distribution shift [KAB-sin]
The reward distribution changes over rounds, with the probability of each arm following a sine wave with a specific frequency f_k, phase λ_k and amplitude 1. At any given time t, the distribution is $\mathbf{p}_t = \{\sin(2\pi f_k t + \lambda_k) \text{ for } k = 1 \dots K\}$.

Partial sinusoidal distribution shift [KAB-sinP]
Identical to the sinusoidal distribution shift, but only a subset of the arms changes sinusoidally while the rest is kept at a constant value and the distribution is not normalized.

3.2 Environment variants and number of arms

The model has been tested and compared with the other algorithms: Thompson Sampling, Epsilon-Greedy, and UCB, in the four different variants of the K-

[1] Since the arm probabilities are not normalized to 1, it is technically improper to call them *probability distributions*; we will therefore refer to either *probability* or *distribution* separately at any given time for avoiding confusion.

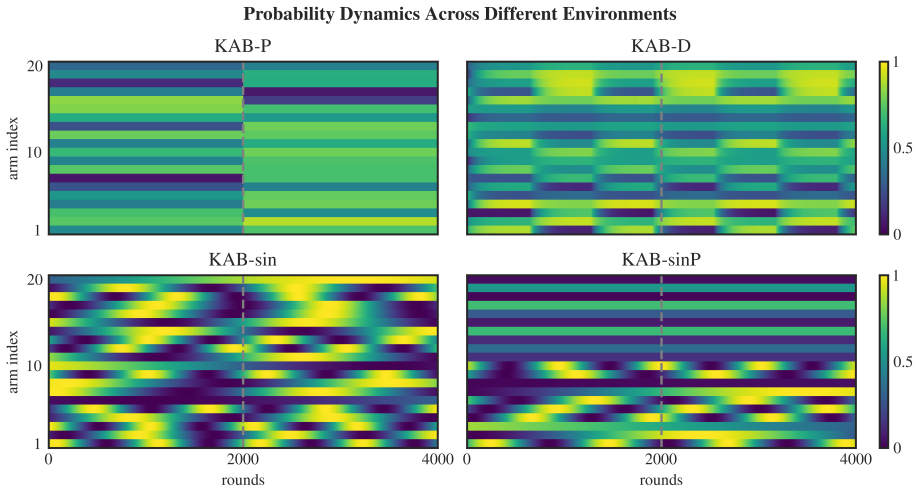

Figure 4: Reward distribution for the four game variants - *The reward distribution for each arm and environment is plotted over two trials of 2000, demarcated by a dotted grey line.*

armed bandit problem. In figure 5, it is reported their results over a different number of arms, ranging from 5 to 1000. Overall, our model displayed a good performance over all environments and arm numbers, suffering only when the latter reached 1000.

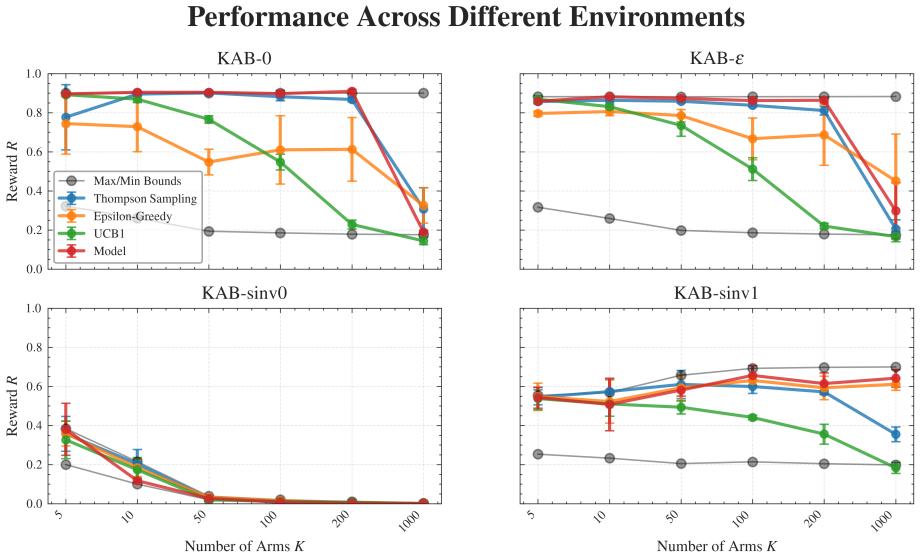

Figure 5: Performance comparison for different values of K and game variants - *The models are evaluated on the four variants of the bandit problem, and their performance is measured as the average reward obtained over 2 trials of 2000 rounds each.*

3.3 Decision-making dynamics

3.3.1 Entropy analysis

For a better understanding of the qualitative differences between the models, we analyzed the progress over the rounds by tracking the selected arms in a simple piecewise stationary distribution environment. The simulation was ran for 3 trials with 2000, and averaged over 5 iterations. Additionally, in order to quantify the variability of the decision policy at a given time and highlight the particularity of each decision-making behaviour, we calculated the entropy of the probability distribution p of chosen arms, calculated over a window of 20 rounds, as $H = -\sum_i^K p_i \log(p_i)$. The unit of entropy is in nats, and it ranges from 0 (no uncertainty) to $\log_e(K)$ (maximum uncertainty). In figure 6, it is plotted for each model the raster plot of selected arms together with its level of entropy. The reward probability distribution over the arms has an average of $H = 2.02$.

As expected, the shape of the entropy curve expresses the inherent strategy adopted by each model. In particular, the UCB algorithm showed the highest variability, marked by a persistent exploratory behaviour throughout the trials despite converging to reward options. Thompson Sampling was able to reach most solutions, although with difficulty in adapting to new reward distributions leading to high entropy levels. $\epsilon-$Greedy also showed a good performance quite reliably, with the greedy strategy assuring low entropy for most of the rounds.

Similar behaviour was observed for our model, which was able to reach the optimal policy and maintain it over time, with entropy peaking mostly at the beggining of the trials and being, on average, the lowest among all models. Indeed, the dynamics of our model make it particularly suited for the task of non-stationary K-armed bandits, as it is able to quickly adapt to new reward distributions and firmly maintain a greedy policy.

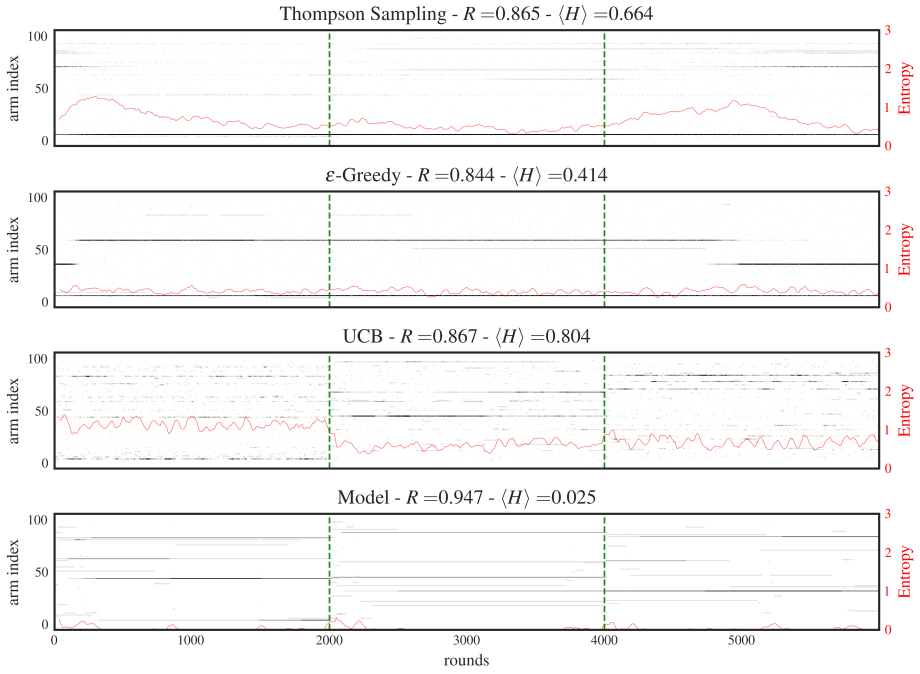

Figure 6: Decision-making dynamics for different models *Each plot display the results from one model. The raster plots (black dots) show the arms selected at each round. The red lines represent the entropy level, calculated from the distribution of selections over the preceeding 20 rounds, smoothed with a 30-steps moving average. In the plot titles, the total reward and average entropy over all trials are also reported.*

3.3.2 Robustness

Then, we sought to investigate the robustness of the model, quantified as the capacity to endure increasing levels of entropy in the reward distribution. The simulation was done in a piecewise stationary environment with $K = 50$, averaged over 128 independent runs. The results report how all models are capable of robust performance even in the presence of high uncertainty. In the top row of figure 3.3.2, it is plotted the average reward obtained by each model against the reward distribution entropy in two trials. In the second trial however, there is a ubiquitous and clear decline in rounds with elevated entropy. This can

be explained by the greater challenge of switching arms when numerous options appear similarly good. In general, our model shows to perform as good as UCB, and better than ϵ-Greedy and Thompson Sampling. Further, the latter seemed to suffer the most, probably due to its conservative approach and difficulty of disengaging from an previously rewarding arm, as underlined also in figure 6. In the same figure, the regret (dashed lines) tells the same story, and remarks the robust performance over uncertainty.

Another perspective to this analysis is given by the plots in the bottom row, which show the average entropy of the selections. Overall, there is the not suprising trend of increasing selection entropy with the entropy of the reward distribution. However, striking is the exception of Epsilon-Greedy, which maintain a constant level throughout. On the one hand, UCB shows a marked and gradual increase, while Thompson Sampling follows with some delay. On the other hand, our model display a more abrupt change, **at around 2.43 nats**, going from a state of very low to very high entropy. For more details about the distribution see the appendix 5.4.

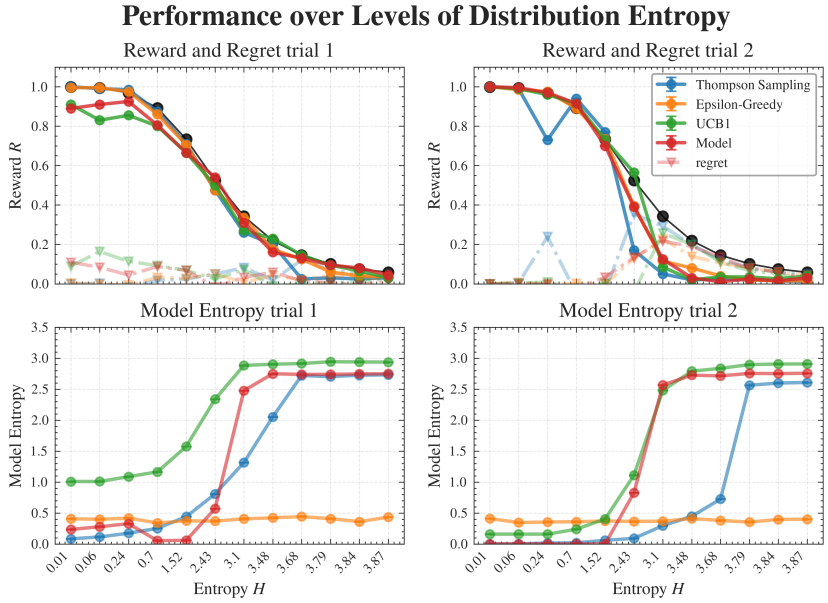

Figure 7: ENTROPY ANALYSIS FOR THE MODEL IN A STATIONARY SETTING - Top row: *trial 1 and 2 have been divided into two columns. A solid line represents the average reward obtained by a model for increasing levels of entropy (in nats) in the reward distribution; a dashed line instead reports the regret with respect to the upper bound (black solid line)* - Bottom row: *average entropy of the selections for the first and second trial of the simulation, each with 2000 rounds each (as calculated in 3.3.1).*

4 Discussion

The process of making decision in uncertain settings is a remarkable aspect of cognition. For instance, such behaviour is implemented in animals during foraging and matching behaviour. In the context of humans, it has been observed that the pool of adopted policies vary considerably [53]. Nevertheless, the subjects seems able to integrate environmental uncertainty and trial generalization in their strategy, and Bayesian algorithms are generally a good fit for the observed policies [54, 55]. A useful formalization of such tasks is the multi-armed bandit problem, which has been extensively studied in the context of reinforcement learning [1]. Although several algorithms have been proposed to solve the problem with robust theoretical guarentees, there is a general lack of biological plausibility of the architecture and dynamics.

In this work, we introduced a model based on two interactive population of rate neurons to address the binomial K-armed bandit problem in non-stationary environments. Our goal was to design an architecture that resemble the functional role of the orbitofrontal cortex (OFC) and anterior cingulate cortex (ACC), together with biologically plausible neuronal dynamics based on synaptic plasticity. The results obtained report how it is able to succesfully adapt to changing reward distributions and maintain a near-optimal policy over time, achieving equally well when compared to the standard algorithms. The assessment was done over four different variants of the bandit problem and a wide range of number of arms, corroborating the robstness of the model.

Further analysis involved the evaluation of the model's behaviour in situations with variable levels of entropy in the reward distribution. One insight was that in situation with low uncertainty, the model is almost always capable of quickly switching to the optimal option and settling to a greedy strategy, similarly to Thompson Sampling but unlike UCB, which is used to persevere in a noticeable exploratory behaviour. When the uncertainty increases also the model's entropy grows, which however does not necessarily hinder performance, except for switching arm in new trials. Here, the model's approach becomes more similar to UCB's than Thompson's.

The strengths of the model can be traced both in the architecture and in the learning paradigm, whose hyperparameters were optimized through an evolutionary process. On one hand the attractor dynamics, which rely on plastic connections and a consensus-like selection process. Particularly important was the choice of modulating the afferent connections to the value population V according to a non-linear function dependant on the synaptic weight itself. In so doing, it was possible to evolve implicitly an effective option-value policy for the tradeoff between exploration and exploitation. This approach can be seen as a form of meta-plasticity implemented through neuromodulation [52], where a region external to the network affects the synaptic connections withough alterning their actual weights; dopamine is a well-suited candidate [29, 56, 57]. On another hand, learning was structured as a non-associative plasticity rule based on the reward. Similarly to before, a non-linear function of the synaptic weights played a critical role, specifically in defining the synapse-specific learn-

ing rate [23]. Again, this mechanism can be considered a form of meta-learning, with evolution leading to the emergence of hyper-parameter encoding important inductive biases [27, 28].

Despite the promising results, there are some limitations to the model. First and foremost, the great level of abstraction in the neuronal details, as we considered simple point neurons with synapses modeled with relatively elementary functions. In particular, the model does not account for the presence of noise in the neural dynamics, which is a well-known feature of biological neurons [58]. Further, the functional association with the pre-frontal cortical region is only moderate. On the computational side, since our interested lied in the biological plausibility and evolution of adaptive meta-learning solutions, we used as reference only a few well established and relatively simple algorithms, and not taken into account more advanced variants [9, 10]. Future work could involve the comparison with more complex algorithms, and the introduction of more realistic neural dynamics, such as spiking neurons [59].

Acknowledgements & Statements
The authors declare no competing interests.
The code is publicly available and can be found at `https://github.com/iKiru-hub/minBandit.git`.
This research was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement Nº 945371 and the University of Oslo.
The research presented in this paper has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

In advance, thanks Kosio.

References

[1] Richard S. Sutton and Andrew G. Barto. The Reinforcement Learning Problem. In *Reinforcement Learning: An Introduction*, pages 51–85. MIT Press, 1998.

[2] Yael Niv, Daphna Joel, Isaac Meilijson, and Eytan Ruppin. Evolution of Reinforcement Learning in Uncertain Environments: A Simple Explanation for Complex Foraging Behaviors. *International Society for Adaptive Behavior*, 2002.

[3] Bruno B. Averbeck. Theory of Choice in Bandit, Information Sampling and Foraging Tasks. *PLoS Computational Biology*, 11(3):e1004164, March 2015.

[4] Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 39.1–39.26. JMLR Workshop and Conference Proceedings, June 2012.

[5] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Asymptotically Optimal Finite Time Analysis, July 2012.

[6] Peter Auer and Nicolo Cesa-Bianchi. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 2002.

[7] J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

[8] Yikun Ban, Jingrui He, and Curtiss B. Cook. Multi-facet Contextual Bandits: A Neural Network Perspective, June 2021.

[9] Michel Tokic. Adaptive ε-Greedy Exploration in Reinforcement Learning Based on Value Differences. In Rüdiger Dillmann, Jürgen Beyerer, Uwe D. Hanebeck, and Tanja Schultz, editors, *KI 2010: Advances in Artificial Intelligence*, volume 6359, pages 203–210. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[10] Michel Tokic and Günther Palm. Value-Difference Based Exploration: Adaptive Control between Epsilon-Greedy and Softmax. In Joscha Bach and Stefan Edelkamp, editors, *KI 2011: Advances in Artificial Intelligence*, volume 7006, pages 335–346. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[11] Aurélien Garivier and Eric Moulines. On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems, May 2008.

[12] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[13] Emanuele Cavenaghi, Gabriele Sottocornola, Fabio Stella, and Markus Zanker. Non Stationary Multi-Armed Bandit: Empirical Evaluation of a New Concept Drift-Aware Algorithm. *Entropy*, 23(3):380, March 2021.

[14] J. O'Doherty, M. L. Kringelbach, E. T. Rolls, J. Hornak, and C. Andrews. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience*, 4(1):95–102, January 2001.

[15] Justin S. Riceberg and Matthew L. Shapiro. Reward Stability Determines the Contribution of Orbitofrontal Cortex to Adaptive Behavior. *Journal of Neuroscience*, 32(46):16402–16409, November 2012.

[16] Léon Tremblay and Wolfram Schultz. Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708, April 1999.

[17] Rebecca Elliott, Raymond J. Dolan, and Chris D. Frith. Dissociable Functions in the Medial and Lateral Orbitofrontal Cortex: Evidence from Human Neuroimaging Studies. *Cerebral Cortex*, 10(3):308–317, March 2000.

[18] Michael J. Frank and Eric D. Claus. Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, 113(2):300–326, April 2006.

[19] Sam Carroll, Krešimir Josić, and Zachary P. Kilpatrick. Encoding certainty in bump attractors. *Journal of Computational Neuroscience*, 37(1):29–48, August 2014.

[20] Jose M. Esnaola-Acebes, Alex Roxin, and Klaus Wimmer. Bump attractor dynamics underlying stimulus integration in perceptual estimation tasks, March 2021.

[21] Bilal A. Bari and Jeremiah Y. Cohen. Dynamic decision making and value computations in medial frontal cortex. *International review of neurobiology*, 158:83–113, 2021.

[22] Alasdair I. Houston, Pete C. Trimmer, and John M. McNamara. Matching Behaviours and Rewards. *Trends in Cognitive Sciences*, 25(5):403–415, May 2021.

[23] Rylan S Larsen and P Jesper Sjöström. Synapse-type-specific plasticity in local circuits. *Current opinion in neurobiology*, 35:127–135, December 2015.

[24] Arne V. Blackman, Therese Abrahamsson, Rui Ponte Costa, Txomin Lalanne, and P. Jesper Sjöström. Target-cell-specific short-term plasticity in local circuits. *Frontiers in Synaptic Neuroscience*, 5:11, December 2013.

[25] Thomas M. Bartol, Cailey Bromer, Justin Kinney, Michael A. Chirillo, Jennifer N. Bourne, Kristen M. Harris, and Terrence J. Sejnowski. Hippocampal Spine Head Sizes Are Highly Precise, March 2015.

[26] Pablo Ariel, Michael B. Hoppa, and Timothy A. Ryan. Intrinsic variability in Pv, RRP size, Ca(2+) channel repertoire, and presynaptic potentiation in individual synaptic boutons. *Frontiers in Synaptic Neuroscience*, 4:9, 2012.

[27] Jeffrey B. Inglis, Vivian V. Valentin, and F. Gregory Ashby. Modulation of Dopamine for Adaptive Learning: A Neurocomputational Model. *Computational brain & behavior*, 4(1):34–52, March 2021.

[28] Kiyohito Iigaya. Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system. *eLife*, 5:e18073, August 2016.

[29] Philippe N. Tobler, Christopher D. Fiorillo, and Wolfram Schultz. Adaptive Coding of Reward Value by Dopamine Neurons. *Science*, 307(5715):1642–1645, March 2005.

[30] Wolfram Schultz, Peter Dayan, and P. Read Montague. A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599, March 1997.

[31] Danila Di Domenico and Lisa Mapelli. Dopaminergic Modulation of Prefrontal Cortex Inhibition. *Biomedicines*, 11(5):1276, May 2023.

[32] Sweyta Lohani, Adria K. Martig, Karl Deisseroth, Ilana B. Witten, and Bita Moghaddam. Dopamine Modulation of Prefrontal Cortex Activity Is Manifold and Operates at Multiple Temporal and Spatial Scales. *Cell Reports*, 27(1):99–114.e6, April 2019.

[33] Kimberlee D'Ardenne, Neir Eshel, Joseph Luka, Agatha Lenartowicz, Leigh E. Nystrom, and Jonathan D. Cohen. Role of prefrontal cortex and the midbrain dopamine system in working memory updating. *Proceedings of the National Academy of Sciences*, 109(49):19900–19909, December 2012.

[34] Han Qi, Fei Guo, and Li Zhu. Forced Exploration in Bandit Problems, December 2023.

[35] Gabriel Koch Ocker and Michael A. Buice. Flexible neural connectivity under constraints on total connection strength, January 2020.

[36] Andrea Apicella, Francesco Donnarumma, Francesco Isgrò, and Roberto Prevete. A survey on modern trainable activation functions. *Neural Networks*, 138:14–32, June 2021.

[37] Paul Miller and Jonathan Cannon. Combined mechanisms of neural firing rate homeostasis. *Biological Cybernetics*, 113(1):47–59, 2019.

[38] Ami Citri and Robert C. Malenka. Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms. *Neuropsychopharmacology*, 33(1):18–41, January 2008.

[39] Mary B. Kennedy. Synaptic Signaling in Learning and Memory. *Cold Spring Harbor Perspectives in Biology*, 8(2):a016824, February 2016.

[40] Mohammad Samavat, Thomas M. Bartol, Kristen M. Harris, and Terrence J. Sejnowski. Synaptic Information Storage Capacity Measured With Information Theory. *Neural Computation*, 36(5):781–802, April 2024.

[41] Christian Igel, Nikolaus Hansen, and Stefan Roth. Covariance Matrix Adaptation for Multi-objective Optimization. *Evolutionary Computation*, 15(1):1–28, March 2007.

[42] Erkki Oja. Oja learning rule. *Scholarpedia*, 3(3):3612, March 2008.

[43] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017.

[44] Chung-Hay Luk and Jonathan D. Wallis. Choice Coding in Frontal Cortex during Stimulus-Guided or Action-Guided Decision-Making. *Journal of Neuroscience*, 33(5):1864–1871, January 2013.

[45] Steven W. Kennerley and Mark E. Walton. Decision Making and Reward in Frontal Cortex. *Behavioral Neuroscience*, 125(3):297–317, June 2011.

[46] Mehdi Khamassi, Pierre Enel, Peter Ford Dominey, and Emmanuel Procyk. Chapter 22 - Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. In V. S. Chandrasekhar Pammi and Narayanan Srinivasan, editors, *Progress in Brain Research*, volume 202 of *Decision Making*, pages 441–464. Elsevier, January 2013.

[47] Shintaro Funahashi. Prefrontal Contribution to Decision-Making under Free-Choice Conditions. *Frontiers in Neuroscience*, 11, July 2017.

[48] Encarni Marcos and Aldo Genovesio. Determining Monkey Free Choice Long before the Choice Is Made: The Principal Role of Prefrontal Neurons Involved in Both Decision and Motor Processes. *Frontiers in Neural Circuits*, 10, September 2016.

[49] Zuzanna Z. Balewski, Thomas W. Elston, Eric B. Knudsen, and Joni D. Wallis. Value dynamics affect choice preparation during decision-making. *Nature neuroscience*, 26(9):1575–1583, September 2023.

[50] Lars Bäckman, Lars Nyberg, Anna Soveri, Jarkko Johansson, Micael Andersson, Erika Dahlin, Anna S. Neely, Jere Virta, Matti Laine, and Juha O. Rinne. Effects of Working-Memory Training on Striatal Dopamine Release. *Science*, 333(6043):718–718, August 2011.

[51] Pierre Enel, Joni D Wallis, and Erin L Rich. Stable and dynamic representations of value in the prefrontal cortex. *eLife*, 9:e54313, July 2020.

[52] Jane X Wang. Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, 38:90–95, April 2021.

[53] Mark Steyvers, Michael D. Lee, and Eric-Jan Wagenmakers. A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3):168–179, June 2009.

[54] Eric Schulz, Nicholas T. Franklin, and Samuel J. Gershman. Finding structure in multi-armed bandits. *Cognitive Psychology*, 119:101261, June 2020.

[55] Shunan Zhang and Angela J Yu. Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[56] Matthew R. Roesch, Donna J. Calu, and Geoffrey Schoenbaum. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, 10(12):1615–1624, December 2007.

[57] Roshan Cools. Chemistry of the Adaptive Mind: Lessons from Dopamine. *Neuron*, 104(1):113–131, October 2019.

[58] A. Aldo Faisal. Noise in Neurons and Other Constraints. In N. Le Novère, editor, *Computational Systems Neurobiology*, pages 227–257. Springer Netherlands, Dordrecht, 2012.

[59] João D. Nunes, Marcelo Carvalho, Diogo Carneiro, and Jaime S. Cardoso. Spiking Neural Networks: A Survey. *IEEE Access*, 10:60738–60764, 2022.

5 Appendix

5.1 Neural response function

The activation functions applied to the two neuronal population are defined as a step-function composed with a generalized sigmoid as follows:

$$f(x; g, o, \theta) = \begin{cases} \left[1 + e^{-g(x-o)}\right]^{-1} & \text{if } \left[1 + e^{-g(x-o)}\right]^{-1} > \theta \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Where:

- x is the neuron pre-activation value
- g is the gain
- o is the offset
- θ is the threshold

Each population has its own set of parameters, which are optimized through evolutionary search.

5.2 Gaussian-sigmoid function

The function $\Phi_\cdot$ is defined by combining a generalized version of the sigmoid, namely with a gain $\beta \neq 1$ and offset $\alpha \neq 0$, and a Gaussian with mean μ and variance σ^2. Their contributions are weighted by as r and $1 - r$ ($r \in (0, 1)$) respectively.

$$\Phi_v(x) = r\left(1 + \exp^{-\beta(x-\alpha)}\right)^{-1} + (1 - r)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The motivation behind this choice is to express a function that possesses a bounded region (depending on μ, σ) at a high/low peak (depeding on the value of γ_2), and a continuous transition to a constant value (depending on the steepness of the sigmoid β, shift α, and intensity γ_1).

5.3 Evolution search

The optimization was carried out over several parameters concerning the model architecture and dynamics:
Network parameters

- τ_u: time constant of population U
- τ_v: time constant of population V
- g_u: gain of the neural response function of population U
- g_v: gain of the neural response function of population V

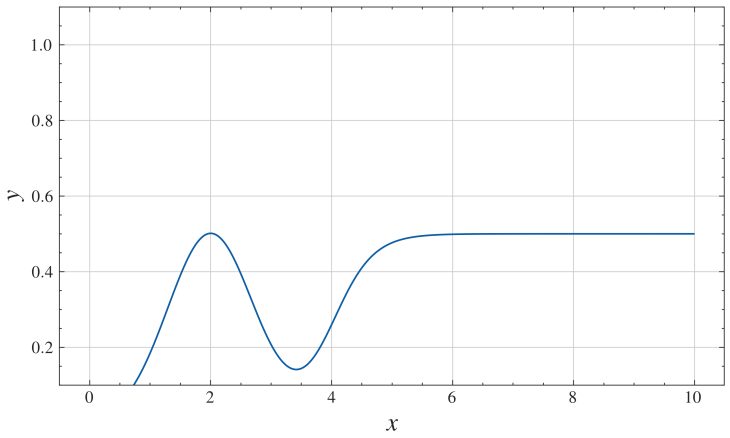

Figure 8: ACTIVATION FUNCTION Φ_v - *Parameters* $\beta = 10$, $\alpha = 1$, $\mu = 1$, $\sigma = 1$, *and* $r = 0.5$.

- o_u: offset of the neural response function of population U
- o_v: offset of the neural response function of population V
- θ_u: threshold of the neural response function of population U
- θ_v: threshold of the neural response function of population V
- W^+: maximal weight value for the weights $\mathbf{W}^{UV}$

Option value function parameters

- β_v: steepness of the sigmoid
- α_v: shift of the sigmoid
- μ_v: mean of the Gaussian
- σ_v: variance of the Gaussian
- r_v: weight of the sigmoid

Learning rate function parameters

- β_η: steepness of the sigmoid
- α_η: shift of the sigmoid
- μ_η: mean of the Gaussian
- σ_η: variance of the Gaussian
- r_η: weight of the sigmoid

Each individual has been evaluated over environment the following environments:

- MAB-0: average reward distribution entropy $\langle H \rangle = 2.05$
- KAB-sinP: average reward distribution entropy $\langle H \rangle = 2.1$, given K arm frequencies f_k as an equally spaced set $\{0.1 \dots i \dots 0.4\}$, phases λ_k drawn from an uniform $\sim \mathcal{U}(0, 2\pi)$, and half of the arms have been set to constant values drawn from another uniform $\sim \mathcal{U}(0.1, 0.7)$; the final reward distribution was not normalized.

The number of arms was $K = 10$ and 150, and lasted for 2 trials with 2000 rounds each. The final fitenss was the average over 2 iterations.

The optimization has been implemented in Python using the `DEAP` library, and the algorithm used was the `CMA-ES` algorithm. The optimization involved 40 generations with a population size of 256 individuals. The mutation rate was set to 0.5 with a sigma of 0.8, the cross-over rate was set to 0.4. The run were carried out on a 256-core AMD EPYC 7763 with 2TB of RAM.

5.4 Reward distribution entropy

The calculation of a set of N reward probability distribution $\mathbf{p}_i$ for $i \dots N$ for K values with a progressively decreasing levels of entropy $\mathbf{h}_i$ for $i \dots N$ has been obtained by the following algorithm:

Algorithm 2: Reward Probability Distribution Generation

Input: Number of distributions N, dimension K
Output: Set of probability distributions $\mathbf{p}_i$ with decreasing entropy
Initial Setup: Define set $B = \{1.5^x \mid x = 1, \dots, 7\}$;
for $i \leftarrow 1$ *to* N **do**
  $\mathbf{z} \leftarrow \text{RandomVector}(0, 1)^K$;
  $j \leftarrow \text{RandomIndex}(K)$;
  $\mathbf{z}_j \leftarrow 1$;
  $\beta_i \leftarrow$ Sample index=i from (B) ; `// Sample temperature from` B
  $\mathbf{p}_i \leftarrow \frac{\exp(\beta_i \mathbf{z})}{\sum_j \exp(\beta_i \mathbf{z}_j)}$; `// Softmax with temperature`
end
return $\mathbf{p}_i$

5.5 Table of results

Note: *table to update*

5.6 Weight update dynamics

We also analyzed the weight update dynamics of the model over the rounds. In figure 9, we plotted the evolution of the total weight ΔW^{UV} over time, averaged over 20 simulations and smoothed over 30 rounds. The results show that the model is able to quickly adapt to new reward distributions. It is also able to maintain the optimal policy over time, with the weights remaining approximately stable. The update quantity ΔW_k^{UV}, which at each round is applied to one connection k, changes sign according to the collected reward, with its magnitude being higher at the beginning of the trials. Initially, the sign is mostly positive (potentiation) since the weights start at zero, and after some uncertainty a consistently preferred arm emerges. However, when the reward distribution switches a regular series of sub-optimal choices with respect to the new distribution is made, leading to zero reward. This causes an accumulation of weight updates with negative sign (depression), eventually bringing the value of the preferred arm to drop. In the meantime, other options are probed until another sequence of choices converges to another arm, promoted by a trail of positive weight updates.

This behaviour is consistent with the low entropy levels observed in the previous analysis.

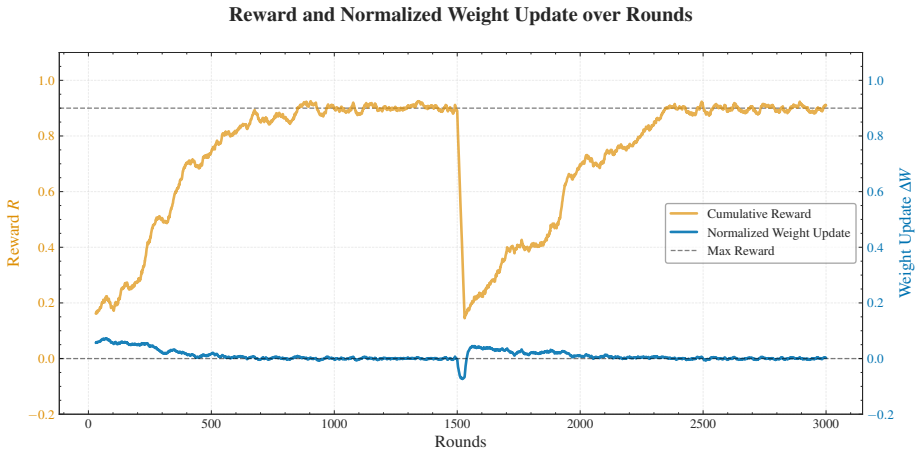

Figure 9: Weight update developement for the model *The plot displays the weight update quantity ΔW_k^{UV} for each round (blue line), smoothed as a 20-steps moving average. It is also reported the average reward in a window of 30 rounds (orange line). The results have been obtained averaging over 20 iterations.*