The contribution of distributional activities of dopamine in the exploration

Mien Brabeeba Wang * Nancy Lynch † Michael Halassa ‡

August 1, 2023

Decision making in natural settings requires efficient exploration to handle uncertainty. Since associations between actions and outcomes are uncertain, animals need to balance the explorations and exploitation to select the actions that lead to maximal rewards. The computational principles by which animal brains explore during decision-making are poorly understood. Our challenge here was to build a biologically plausible neural network that efficiently explores an environment and understands its effectiveness mathematically.

One of the most evolutionarily conserved and important systems in decision making is basal ganglia (BG)[1]. In particular, the dopamine activities (DA) in BG is thought to represent reward prediction error (RPE) to facilitate reinforcement learning[2]. Therefore, our starting point is a cortico-BG loop motif[3]. This network adjusts exploration based on neuronal noises and updates its value estimate through RPE. To account for the fact that animals adjust exploration based on experience, we modified the network in two ways. First, it is recently discovered that DA does not simply represent the scalar RPE value; rather it represents RPE in a distribution[4]. We incorporated the distributional RPE framework and further the hypothesis, allowing an RPE distribution to update the posterior of action values encoded by cortico-BG connections. Second, it is known that the firing in the layer 2/3 of cortex fires is variable and sparse[5]. Our network thus included a random sparsification of cortical activity as a mechanism of sampling from this posterior for experience-based exploration. Combining these two features, our network is able to take the uncertainty of our value estimates into account to accomplish efficient exploration in a variety of environments.

Additional Details Our models connect to both biological correlates and normative theories and excel at multi-armed bandit tasks in various environments. To measure the performance of each model, we considered the regret of a model, which is the expected difference of rewards between a model's actions and the best actions. This network has comparable or better performance on bandit tasks to Thompson sampling, a widely used algorithm in practice with tight

*CSAIL, MIT, Cambridge, Massachusetts, USA. Email: brabeeba@mit.edu.

†CSAIL, MIT, Cambridge, Massachusetts, USA. Email: lynch@csail.mit.edu.

‡School of Medicine, Tufts University, Boston, Massachusetts, USA. Email: michael.halassa@tufts.edu.

[1]Stephenson-Jones et al., "Evolutionary conservation of the basal ganglia as a common vertebrate mechanism for action selection" (2011); Grillner et al., "The basal ganglia downstream control of brainstem motor centres–an evolutionarily conserved strategy" (2015).

[2]Schultz et al., "A neural substrate of prediction and reward" (1997); Niv, "Reinforcement learning in the brain" (June 2009).

[3]Soltani et al., "A biophysically based neural model of matching law behavior: melioration by stochastic synapses" (2006).

[4]Dabney et al., "A distributional code for value in dopamine-based reinforcement learning" (2020).

[5]Kerr et al., "Spatial organization of neuronal population responses in layer 2/3 of rat barrel cortex" (2007).

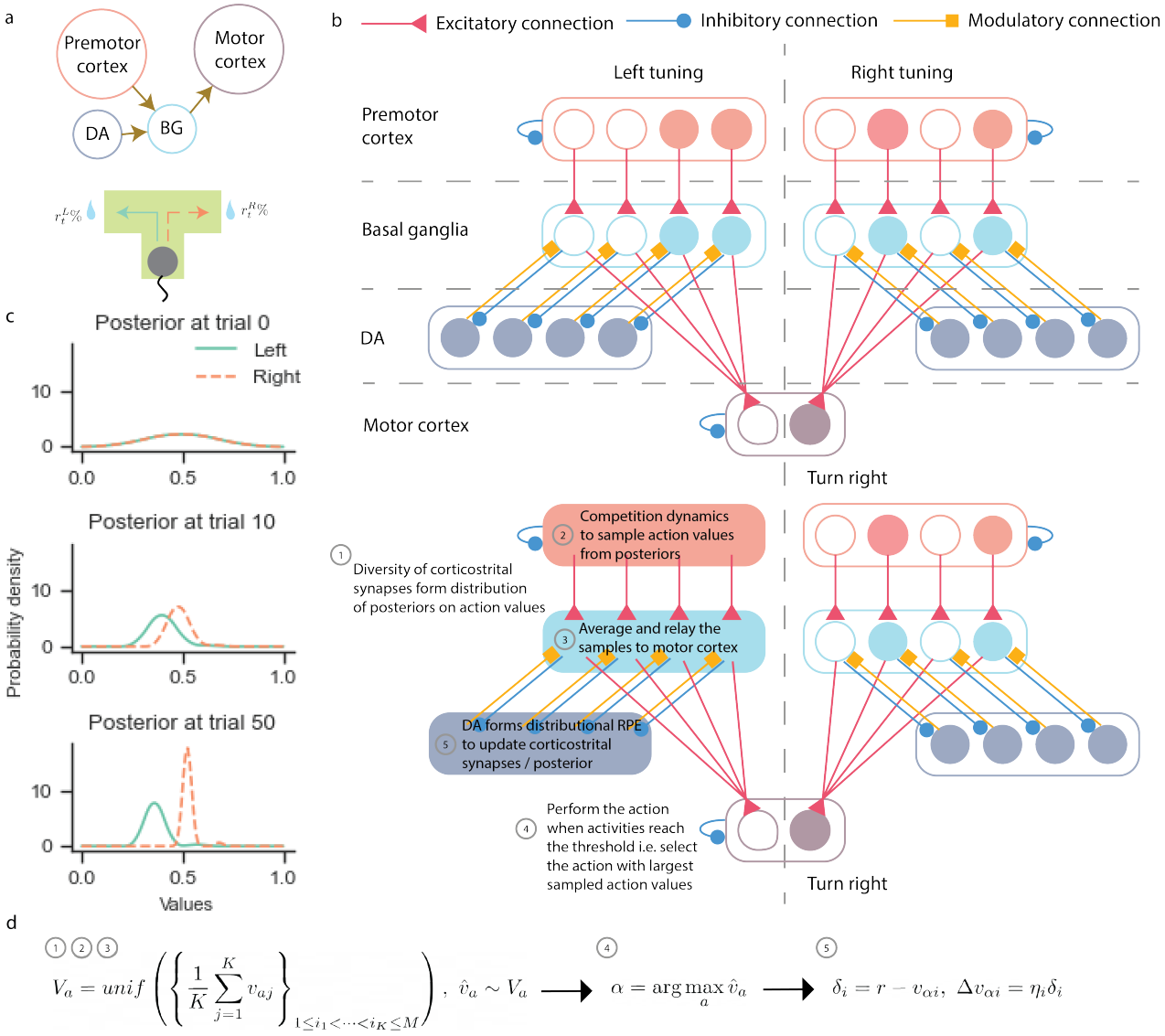

Figure 1: A distributional RPE-based corticostriatal model. a. The system and the task we model. **b.** Details schematic of the model. **c.** Sample posteriors from the model. **d.** The normative theory which corresponds to the model.

theoretical guarantee[6], and outperforms the traditional scalar RPE-based neural network which explores based on noises. To understand its performance, we further approximated the circuit with a normative theory which we mathematically prove to have near-optimal performance $O(\sqrt{AT \log AT})$, only a logarithmic factor away from the known lower bound of $\Omega(\sqrt{AT})$[7] where A is the number of arms and T is the number of trials taken. By perturbing the model, we identify that the diversity of synapses, large ensemble size and moderate sparsity are crucial for the network to build posterior that allows efficient exploration.

[6]Korda et al., "Thompson Sampling for 1-Dimensional Exponential Family Bandits" (2013).
[7]Auer et al., "Gambling in a rigged casino: The adversarial multi-armed bandit problem" (1995).

Model We consider neural networks of premotor cortex, basal ganglia and motor cortex (Figure 1 a) with neurons of the form $\tau \frac{dx}{dt} = -x + f(Wx + I)$. Specifically, within premotor and basal ganglia, there are A ensembles that each tune to a different action. Let the ensemble of cortex-BG synapse that tuned to action a be $\{v_{ai}\}_{i\in[M]}$ and we set all the strength of BG-cortical synapses to be $\frac{1}{K}$. We set the recurrent weight of premotor cortex in each ensemble to do K-WTA while setting the recurrent weight of the motor cortex to do WTA. Once the activity of a motor neuron is above a certain threshold, the corresponding action α is performed to receive reward r. Then, the DA activities form a distributional RPE $\delta_i = r - v_{\alpha i}$ to update the synapse $v_{\alpha i} \leftarrow v_{\alpha i} + \eta_i \delta_i$. Crucially, each corticostriatal synapse has a different initial weight $\bar{v}_i$ and a different learning rate η_i (Figure 1 b).

To amend for mathematical analysis, we consider the following normative theory that approximates the above neural network model. First, the K-WTA dynamics in the premotor cortex samples $\{v_{ai_j}\}_{j\in[K],\ 1\le i_1<\cdots<i_K\le M}$ uniformly which is then averaged $\hat{v}_a = \frac{1}{K}\sum_{j=1}^{K} v_{ai_j}$ and relayed to motor neurons by BG. The WTA dynamics in motor cortex then selects action $\alpha = \text{argmax}_a\, \hat{v}_a$ to receive reward r and form distributional RPE δ_i which in terms update the synapses $v_{\alpha i} \leftarrow v_{\alpha i} + \eta_i \delta_i$ (Figure 1 d).

Intuitively, the model works by sampling two posteriors and selecting the action with larger sampled values. When the model is not confident at its value estimate, two posteriors will have large overlapping which results in exploration while when the model is confident at its estimate, two posteriors will become narrower and well separated which results in exploitation. (Figure 1 c).

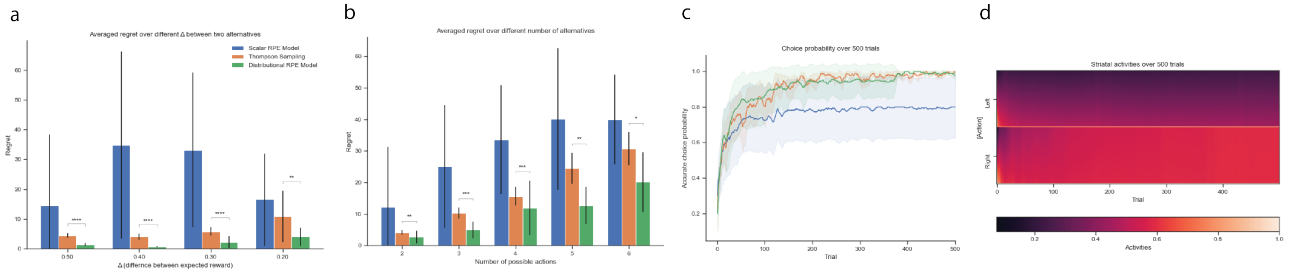

Figure 2: Distributional RPE model explores efficiently in various environment. a. Comparison of models in two-armed bandit tasks with different Δ between expected rewards of two arms. **b.** Comparison of models in environments with a different number of alternatives **c.** Choice probability of each model in a two-armed bandit task. **d.** The average striatal activities in the distributional RPE model

Result We run the model in various multi-armed bandit tasks with Bernoulli rewards and compare its performance with Thompson sampling and a scalar RPE model. We first vary the difference between expected values in a two-arm bandit task. While the traditional scalar RPE model fails to identify the correct arm frequently resulting in large regret and variance in performance, our distributional model outperforms both the scalar RPE model and Thompson sampling in all environments (Figure 2 a). We then test our models by varying the number of arms in an environment. Again, our distributional model outperforms both the scalar RPE model and Thompson sampling in all environments and as the number of arms increases, the average regret increases (Figure 2 b). Next, we look at the trial averaged correct choice probability in a two-armed bandit task. The scalar RPE model fails to reach probability 1

which indicates that it sometimes fails to identify the preferred action while both Thompson sampling and the distributional RPE model reach probability 1. In particular, in the first 100 trials, the distributional RPE model identifies the preferred arms and commits to the preferred arm faster than Thompson sampling (Figure 2 c). The striatal activities in BG also show signatures of efficient exploration. The neurons tuned to the correct action (right) quickly narrow the distribution and converge to the correct estimates while the neurons tuned to the less preferred action still show a gradient of activities distinct from that of the preferred action which indicates the posterior is wide but well-separated from the posterior of the correct action (Figure 2 d).

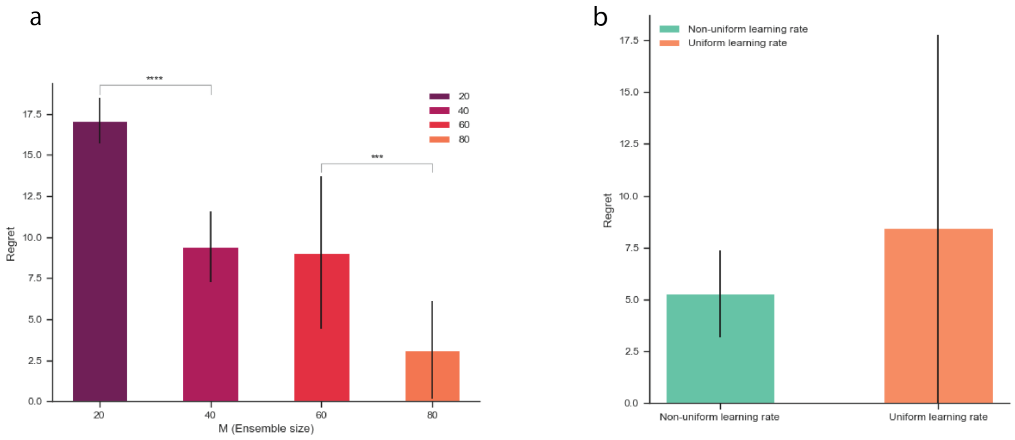

Figure 3: Larger ensemble size and non-uniform learning rate is necessary for efficient exploration. a. Comparison of models by varying the size of the ensemble **b.** Comparison between model with uniform learning and non-uniform learning rate

To understand how different parameters influences the efficient exploration, we first vary the ensemble size M and we observe that the larger the ensemble size is, the lower the regret is (Figure 3 a). We conjecture that the model with larger ensemble size can encode the posteriors in finer resolution and therefore explore more efficiently. Second, we compare how the diversity of learning rates influence efficient exploration and we observe that the model with non-uniform learning rates perform better than that with uniform learning rate (Figure 3 b). We conjecture that the diversity of learning rate forms a diverse ensemble of estimates to represent the posterior and therefore allow the model to explore more efficiently. To further understand its performance, we prove the following theorem.

Theorem 1. *If we choose the sparsity K, initial weight $\{\hat{v}_{ai}\}_{a\in[A],i\in[M]}$, the learning rate $\{\eta_i\}_{i\in[M]}$ appropriately, then the regret of the normative theory after T trials is bounded by $\sqrt{600AT\log(AT)}$. In particular, this means that our model has nearly-optimal regret, only a logarithmic factor away.*

The basic idea of the analysis is to bound the expected number when a sub-optimal arm is chosen. One can separate this term into two situations, when the estimated value of an arm is smaller than the optimal value minus a small constant or when it is larger, and bound them separately. The first term intuitively corresponds to how many samples one needs to be confident in the value estimates of the optimal arm while the second term corresponds to the exploration of the sub-optimal arms.