Reward Prediction Error Neurons Implement an Efficient Coding for Future Reward in Multidimensional Probabilistic Map

Jialin Li [1]

[1]New York University

Motivation

Advances on artificial intelligence shed light on the heterogeneous response of dopamine neurons in both reward magnitude [1] and time scale [3] compared to traditional reinforcement learning.

From the perspective of efficient coding[2, 4], we proposed a multidimensional efficient coding model for reward prediction errors neurons(RPEs) that have several characteristics:

- Jointly encodes **reward magnitude** and **temporal delay**.
- Accounts for **temporally decaying firing budget** (natural discounting).
- Grounded in **information maximization principles**.

Efficient Coding Model for Multidimensional Probabilistic Map

We assume the instantaneous firing rate of the neuron i in response to a reward of magnitude r delivered at delay t(relative to the cue or an expected time) is given by a sigmoidal tuning function σ **(Fig. 1)** modulated by a time-dependent scaling:

$$y_i(r,t) \;=\; F_{\max}\,\sigma\!\left(\frac{g(\gamma_i,\,t)\,r \;-\; \theta_i}{b_i}\right) \;-\; F_{\text{base}} \tag{1}$$

- θ_i determines the range of reward magnitudes the neuron is sensitive to.
- γ_i determines how strongly the neuron responds to delayed rewards.
- b_i determines the slope of the tuning curve.
- $F_{\max}$ and F_{base} determine the maximal and spontaneous firing rate respectively.

The term $g(\gamma_i,\,t)$ implements temporal scaling of the effective input, reflecting how the neuron "values" a reward delivered after delay t **(Fig. 2)**. A common function to use is the exponential discounting (McClure et al., 2004) or the hyperbolic discounting (Kable & Glimcher, 2007):

$$g(\gamma_i,t) \;=\; \gamma_i^{\,t} \text{ or } g(\gamma_i,t) = \frac{1}{1+\gamma_i t} \tag{2}$$

Importantly, we assume that the neuron population encodes the environment with the efficient coding principle. Thus, the joint distribution of neural parameters D(θ, γ) should be allocated proportionally to the prior **(Fig. 3)**:

$$D(\theta,\gamma) \propto \frac{1}{\Delta\theta\,\Delta\gamma} \iint_{\substack{r\approx\theta/g(\gamma,t)\\ t:\text{appropriate}}} p(r,t)\,dr\,dt, \tag{3}$$

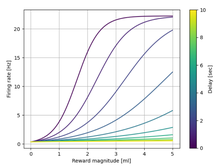

Figure 1. Marginal reward response

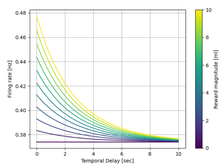

Figure 2. Marginal time response

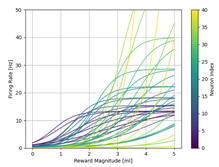

Figure 3. Population tuning curve

Methods

We proposed to used the data [5], which recorded 43 optogenetically identified dopamine neurons from 6 mice. The variable reward magnitude and time delay task involve five different trial types**(Table 1)**, which were randomly interleaved throughout the session.

Trial Type	Odor Duration	Delay Duration	Reward Type	Reward Amount (µL)
1	1 s	0 s	Fixed	4.5 µL
2	1 s	1.5 s	Fixed	4.5 µL
3	1 s	3 s	Fixed	4.5 µL
4	1 s	6 s	Fixed	4.5 µL
5	1 s	3 s	Probabilistic	1 (0.25), 2.75 (0.167), 4.5 (0.167), 6.25 (0.167), 8 (0.25)

Table 1. Summary of trial types used in Sousa et al. (2023)

- Trial timing was sampled from exponential distribution to ensure approximately **constant reward rate** and **uniform hazard**
- Odor stimuli were randomized across animals to **prevent odor-reward associations**

To fit the parameters of each dopamine neuron, we can use maximum likelihood estimation based on Poisson noise[4], which minimizes the difference between measured firing rate r_j and the predicted firing rate $\sigma\left(R_j;\theta\right)$.:

$$\hat{\theta} \;= \text{argmin}_\theta \sum_{j=1}^{T} \left(r_j\; log\left(\sigma\left(R_j;\theta\right)\right) \;-\; \sigma\left(R_j;\theta\right)\right) \tag{4}$$

Results: Heterogenous Activity of Dopamine Neuron Response

This model can exhibit the heterogeneous neural activity of dopamine neurons because of the flexible setting of parameters. Here we simulated an example neuron with threshold $\theta = 1.83$ and discount factor $\gamma = 0.69$. The neuron exhibits the characteristics below**(Fig. 4)**:

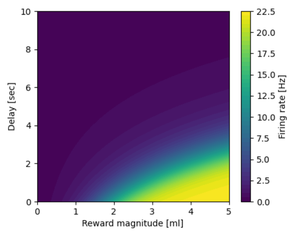

Figure 4. Example neuron's firing rate as a function of reward magnitude and delay

For individual neuron with specific threshold θ and discount factor γ:

1. The firing rate is increased when the reward magnitude is higher.
2. The firing rate is decreased when there is longer time delay to deliver the reward.
3. The transition from silence to firing occurs along a curved ridge, $\gamma^t r \;\approx\; \theta$, which indicates that the discounted reward equals the neuron's threshold

For neuron population with different combinations of threshold θ and discount factor γ:

1. Neurons with higher θ have larger maximal firing rate
2. Neurons with smaller γ decay more quickly

Relationship with Distributional Reinforcement Learning

In distributional reinforcement learning model[1], it proposes that different channels have different relative scaling for positive(α^+) and negative (α^-) RPEs:

$$V_i(t) \leftarrow \begin{cases} V_i(t) + \alpha_i^+ \delta_i(t), & \delta_i(t) > 0 \\ V_i(t) + \alpha_i^- \delta_i(t), & \delta_i(t) \le 0 \end{cases}$$

An imbalance between α^+ and α^- causes each channel to learn a different value prediction, which leads to the heterogeneity of neural activity in our brain. Compared with their model, there are two interesting connections.

Curvature of tuning curve

In distributional RL, the reversal point is defined as the reward magnitude where RPE shifts from negative to positive. In our model, neural tuning curves not only differ in curvature but also evolve over time due to the discount factor (γ), resulting in more diverse responses **(Fig. 3)**.

Slope changes with threshold

Similar to [4], neurons with high thresholds typically have flatter slopes because they should have wider responses towards reward magnitudes. This prediction preserves the important features of efficient coding in terms of constraint firing rate budget.

Results: Neuron Population Reconstruct the Environmental Statistics

In our model, each neuron's threshold θ and discount factor γ is placed on a specific quantile in the reward magnitude and time delay distribution. Thus, the neuron population should in principle encode the original environmental statistics **(Fig. 5)**.

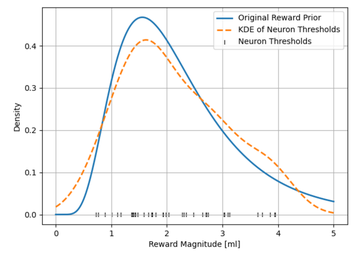

Figure 5. Reconstructed distribution of neuron population

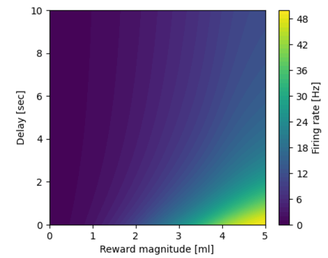

Figure 6. Neural population activity

Additionally, just like each individual neuron, the neuron population exhibits similar characteristics on how the overall firing rate changes with the reward magnitude and time delay **(Fig.6)**.

References

[1] Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675, 2020.

[2] Deep Ganguli and Eero P. Simoncelli. Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural Computation*, 26(10):2103–2134, 2014.

[3] Paul Masset, Pablo Tano, HyungGoo R. Kim, Athar N. Malik, Alexandre Pouget, and Naoshige Uchida. Multi-timescale reinforcement learning in the brain. *bioRxiv*, 2023.

[4] Heiko H. Schütt, Dongjae Kim, and Wei Ji Ma. Reward prediction error neurons implement an efficient code for reward. *Nature Neuroscience*, 27(7):1333–1339, 2024.

[5] Margarida Sousa, Pawel Bujalski, Bruno F. Cruz, Kenway Louie, Daniel McNamee, and Joseph J. Paton. Dopamine neurons encode a multidimensional probabilistic map of future reward. bioRxiv, 2023.