Distributional Reinforcement Learning in Dopamine Neurons

Neuroscience Project, Fall 2025

Ronghao Huang[1*] Zeyu Cui[1*] Haoyang Chen[1*]
{231880191, 231880019, 23188293}@smail.nju.edu.cn

Abstract

Dopamine neurons are thought to encode reward prediction errors, but whether they represent only average rewards or full distributions remains debated. We compare classical and distributional temporal difference (TD) models using a simulated variable-reward task and recorded mouse dopamine data. The distributional TD model adapts faster to environmental changes and better captures the diversity of neuronal responses. A decoding analysis further shows that only distributional TD can recover the true reward distribution. These findings suggest that dopamine neurons may implement a distributional reinforcement learning strategy. Code is available at Our GitHub repository.

1. introduction

Reinforcement Learning (RL) models how agents learn from reward feedback. Classical algorithms like temporal difference (TD) learning estimate the expected value of future rewards but fail to capture real-world uncertainty.

Recent neuroscience suggests that dopamine (DA) neurons, long linked to reward prediction errors (RPEs), may encode richer information. Distributional RL (DistRL) proposes that agents represent full reward distributions via a population of asymmetric value predictors, each corresponding to a different expectile [1].

This perspective is supported by empirical evidence showing that dopamine neurons exhibit varied RPE signals under uncertainty [2]—some respond more strongly to positive outcomes, others to negative ones, and some remain relatively neutral.

In this project, we test whether DistRL better accounts for DA activity than classical TD. We simulate a variable reward task, analyze mouse DA recordings, and apply a decoding framework to reconstruct reward distributions from neural data. Our results support the view that DA neurons implement a distributional code for value.

2. Cat - Car Task Simulation

To compare classical and distributional TD models under uncertainty, we simulate a two-cue task inspired by [2] [3].

2.1. Task Design

We design a two-stage task with two image categories:

- **Car**: always yields $r = +1$
- **Cat**: $r \in \{-3, +5\}$ with equal probability

Both have $\mathbb{E}[r] = +1$, but only Cat exhibits variance. In Stage 2 (after 200 epochs), Cat's reward shifts deterministically to $r = +5$, testing model adaptability to changing reward distributions.

2.2. Model Setup

The classical TD agent learns via a symmetric update rule:

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot \delta_t, \quad \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (1)$$

The distributional TD agent maintains multiple value channels $V^{(i)}(s)$, each with asymmetric updates:

$$V^{(i)}(s_t) \leftarrow V^{(i)}(s_t) + \begin{cases} \alpha_+^{(i)} \cdot \delta_t^{(i)} & \text{if } \delta_t^{(i)} > 0 \\ \alpha_-^{(i)} \cdot \delta_t^{(i)} & \text{if } \delta_t^{(i)} < 0 \end{cases} \quad (2)$$

where each $\delta_t^{(i)}$ is the TD error for unit i. This enables learning of distinct expectiles across the reward distribution.

2.3. Results and Analysis

In Stage 1, both models successfully learn the expected reward values. And after the rule switch in Stage 2, the distributional TD model adapts rapidly because its expectile-based channels have already represented high-reward outcomes and only need to adjust their weights. In contrast, the classical TD model must relearn the new expected value from scratch, leading to a slower adjustment.

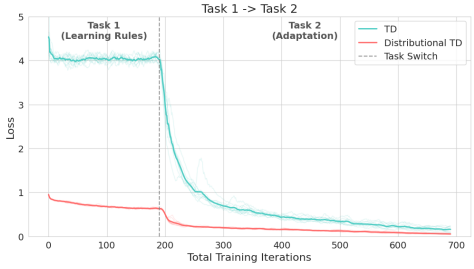


Figure 1. **Loss curves for Cat-Car task.**

These findings demonstrate that distributional TD more efficiently leverages prior uncertainty representations to handle nonstationary environments.

3. Neural Data Analysis

To test whether distributional TD learning matches biological data, we analyze spike activity from midbrain dopamine (DA) neurons in mice performing variable-reward tasks, using the dataset in [2]. We consider two paradigms: one with variable reward magnitudes and one with variable reward probabilities.

3.1. Reversal Points in Dopaminergic Responses

In the variable-magnitude task, mice received water rewards whose volume on each trial was drawn from a discrete set. For each DA neuron, we measured firing rates as a function of reward size and defined its *reversal point* as the reward magnitude at which the response switched from net suppression to net excitation; this value was used to sort cells.

Figure 2 compares reversal tuning curves across models and data. Classical TD units (left) show tightly aligned reversal points, consistent with a single scalar prediction error. Distributional TD units (middle) show a wider spread across cells. Real DA neurons (right) display even stronger heterogeneity, consistent with a more distributional code.

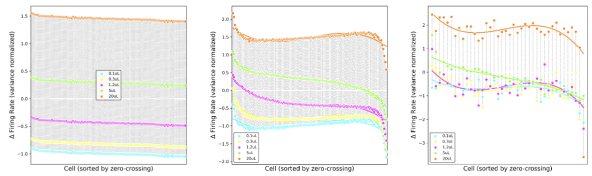


Figure 2. Reversal tuning curves. Left: classical TD model units. Middle: distributional TD model units. Right: DA neuron recordings. Cells are sorted by their reversal point.

3.2. Optimistic and Pessimistic Subpopulations

In the variable-probability task, cues signaled reward probabilities of 10%, 50%, or 90%. Under the 50% condition, cue-evoked DA responses varied systematically: some neurons responded above the population mean ("optimistic"), whereas others responded below it ("pessimistic").

We quantified this effect by computing a t-statistic for each neuron's deviation from the population mean at 50%. As shown in Figure 3, DA neurons exhibit an approximately bimodal distribution, indicating optimistic and pessimistic subpopulations. The distributional TD model reproduces this pattern, whereas the classical TD model yields a more unimodal distribution. GABAergic neurons show a similarly broad, multi-peaked distribution.

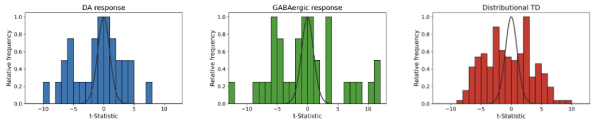


Figure 3. t-statistic distributions of neural and model responses under 50% reward probability. DA and GABA neurons show multi-peaked distributions similar to the distributional TD model, while the classical TD model is more unimodal.

Overall, these results suggest that dopaminergic populations are better described as encoding a distribution of prediction errors, rather than a single scalar value, in line with the distributional TD framework.

4. Decoding and Model Comparison

To further test the distributional reinforcement learning hypothesis, we performed a decoding analysis to determine whether the underlying reward distribution can be reconstructed from dopamine neuron activity.

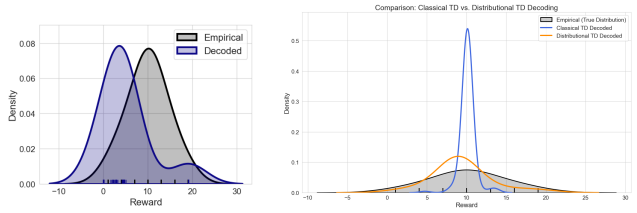
We adopted a decoding framework based on expectile regression, assuming that each neuron encodes a specific expectile $\tau \in (0, 1)$ of the reward distribution. The decoder estimates a reward density that best explains the observed firing patterns across a population. The core loss function for expectile regression is defined as:

$$\mathcal{L}_\tau(y, \hat{y}) = |\tau - \mathbb{I}(y < \hat{y})| \cdot (y - \hat{y})^2 \tag{3}$$

where y is the true reward and $\hat{y}$ is the prediction. The indicator function $\mathbb{I}(\cdot)$ creates an asymmetric penalty, causing each neuron with a unique τ to converge to a different expectile of the reward distribution.

We compared the decoding performance of classical and distributional TD models. The classical TD model, with its symmetric update rule, converged to a single mean value. In contrast, the distributional TD model, leveraging varied asymmetry coefficients, successfully reconstructed the entire reward distribution.

We validated this approach on real dopamine neuron recordings, where the decoded distribution accurately matched the empirical reward distribution, providing strong biological support for the distributional hypothesis.



(a) Run Decoding Example (b) Classical vs Distribution

Figure 4. **Decoding reward distributions from dopamine neuron responses.** The figure illustrates the decoding performance.

These results suggest that distributional TD not only explains heterogeneity in reward prediction but also enables accurate recovery of latent reward statistics—something classical TD fails to achieve.

5. Conclusion and Discussion

Our results support the hypothesis that DA encode distributions over future rewards, not just their mean. Compared to classical TD, the distributional model better captures neuronal diversity and explains learning under uncertainty.

Decoding analyses further show that population activity can reconstruct the true reward distribution, highlighting the biological plausibility of distributional reinforcement learning. This framework offers a promising basis for understanding adaptive behavior in uncertain environments.

References

[1] Margarida Sousa, Pawel Bujalski, Bruno F. Cruz, Kenway Louie, Daniel C. McNamee, and Joseph J. Paton. A multi-dimensional distributional map of future reward in dopamine neurons. *Nature*, 642:691–699, 2025. 1

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

[3] Mark Rowland, Marc G Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR, 2019. 1