MODELING DOPAMINE USING TD-LEARNING

Reinforcement Learning

Quantitative Life Sciences

Juan S. Rojas.

Outline

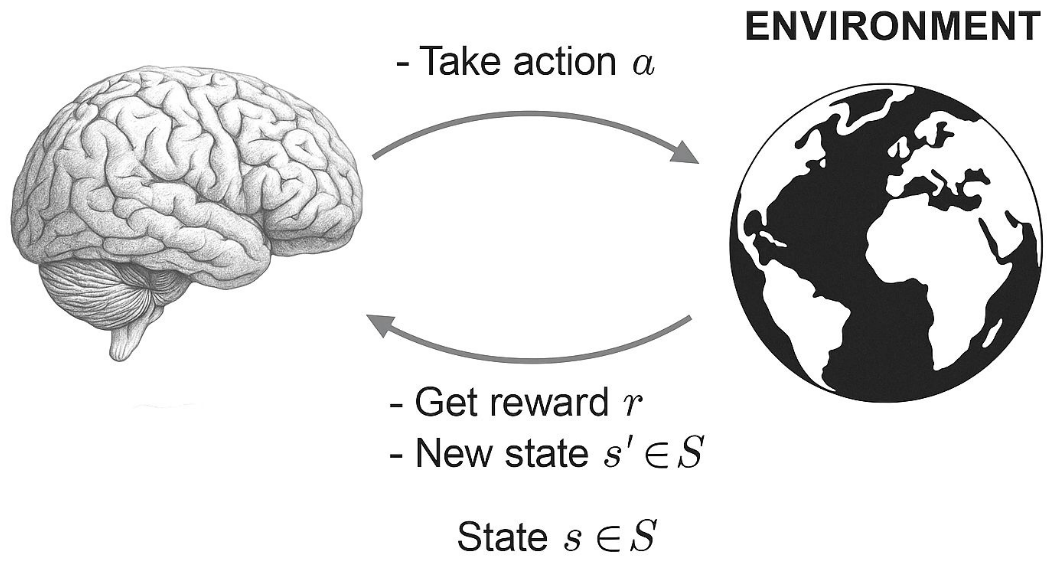
"What we observe is not nature itself, but nature exposed to our method of questioning."
- Werner Heisenberg

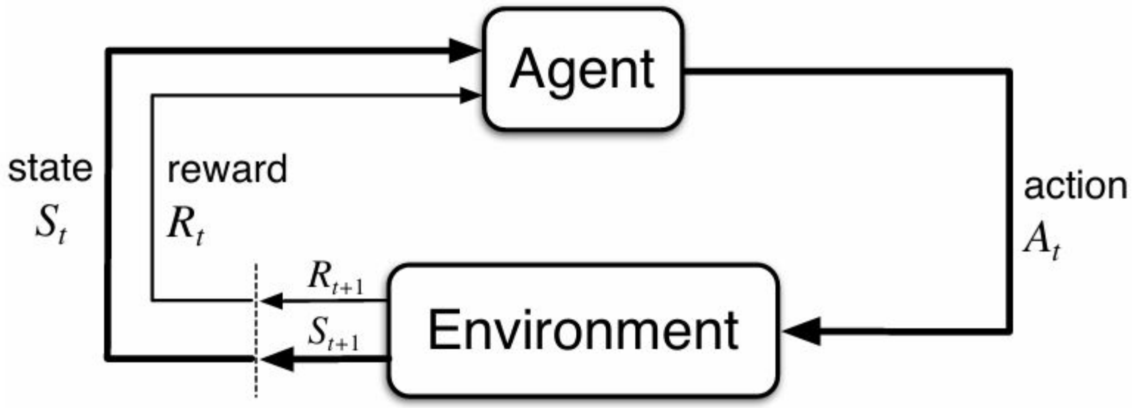
1. Motivation: Brain & Reinforcement Learning
2. Classical TD Learning
3. Observations
4. Distributional TD Learning
5. Neuroscience Validation

Neuromatch Academy. (2020). Available at: https://compneuro.neuromatch.io, The Universe Is a Machine That Keeps Learning, Scientists Say

Motivation: **Brain & Reinforcement Learning**

- The connection between RL and Brain is intuitive.

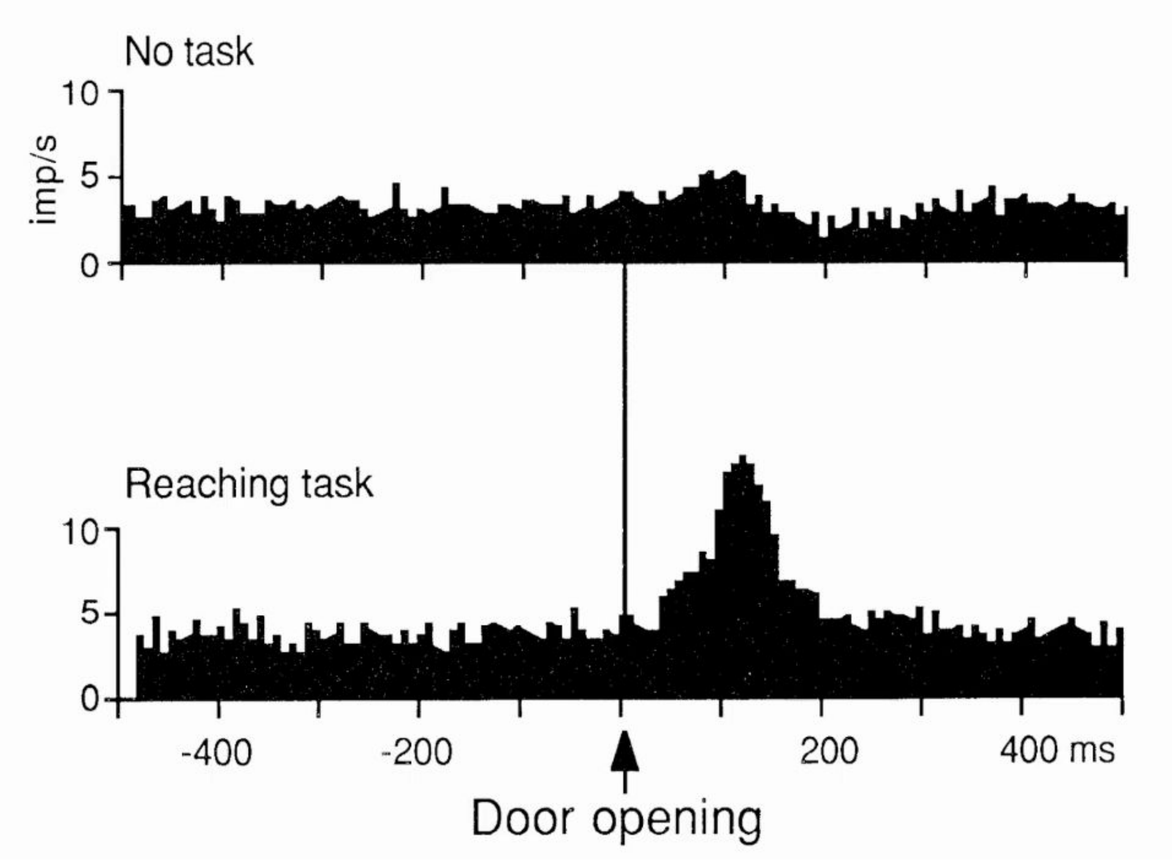


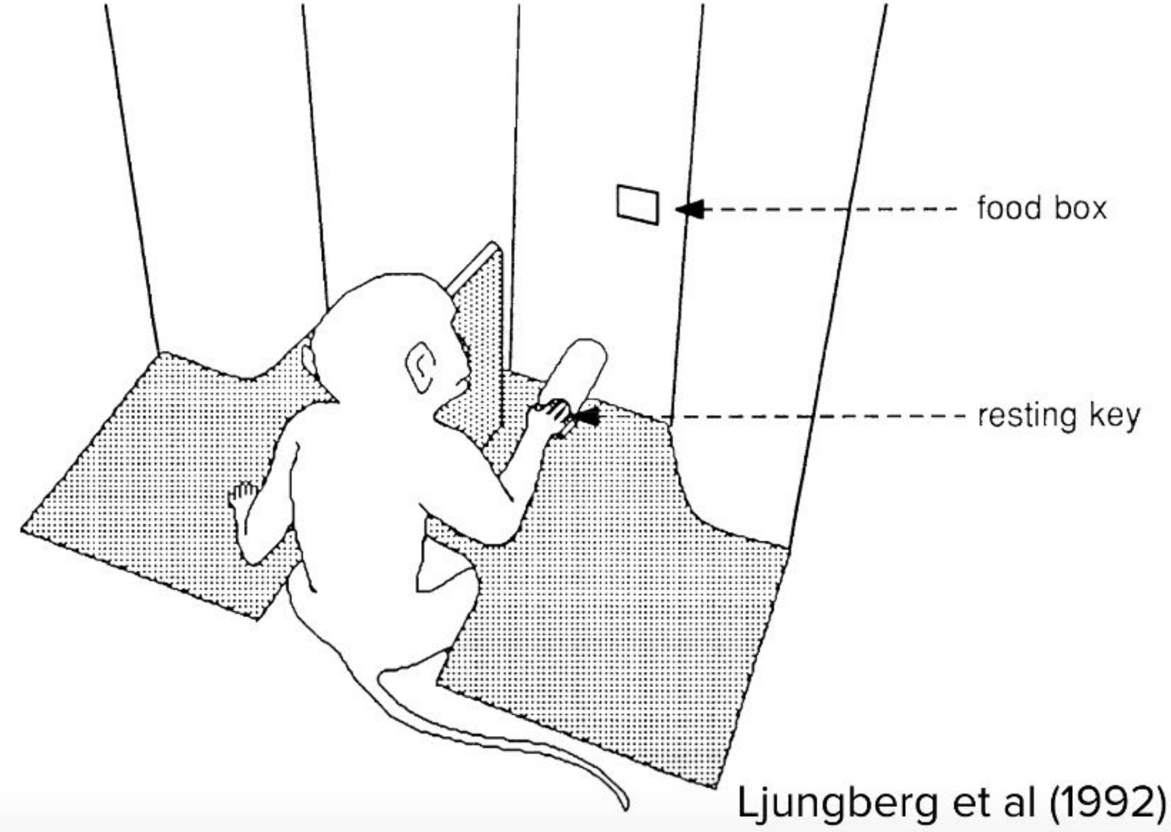


- This project replicates two foundational studies:
  - **Schultz et al. (1997):** Showed that dopamine neurons encode reward prediction errors.
  - **Dabney et al. (2020):** Revealed that dopamine neurons encode not just a single expected value, but a *distribution* of possible rewards.
- There are more ideas, model-based & model-free. **Hypothesis, our hippocampus develops models that the cortex trains while we are sleeping.**

Neuromatch Academy. (2020). Available at: https://compneuro.neuromatch.io

1) The agent goes through episodes with no need to take actions.
2) Each episode is a fixed sequence of states: S0→S1→S2→···→Sn. In each episode: A **Conditioned Stimulus (CS)** is shown early and an **Unconditioned Stimulus (US)** or reward.
3) The agent's task is to **learn to predict future rewards** based on the current state.





Ljungberg et al (1992)

Neuromatch Academy. (2020). Available at: https://compneuro.neuromatch.io

Classical TD-Learning: Framework

Return:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

Value Function:

$$V_\pi(s) = E[G_t \mid s_t = s] \qquad V_\pi(s) = E[r_{t+1} + \gamma V_\pi(s_{t+1}) \mid s_t = s]$$

$$V_\pi(s) = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s)\,[r + \gamma V_\pi(s')]$$

TD-error & Update:

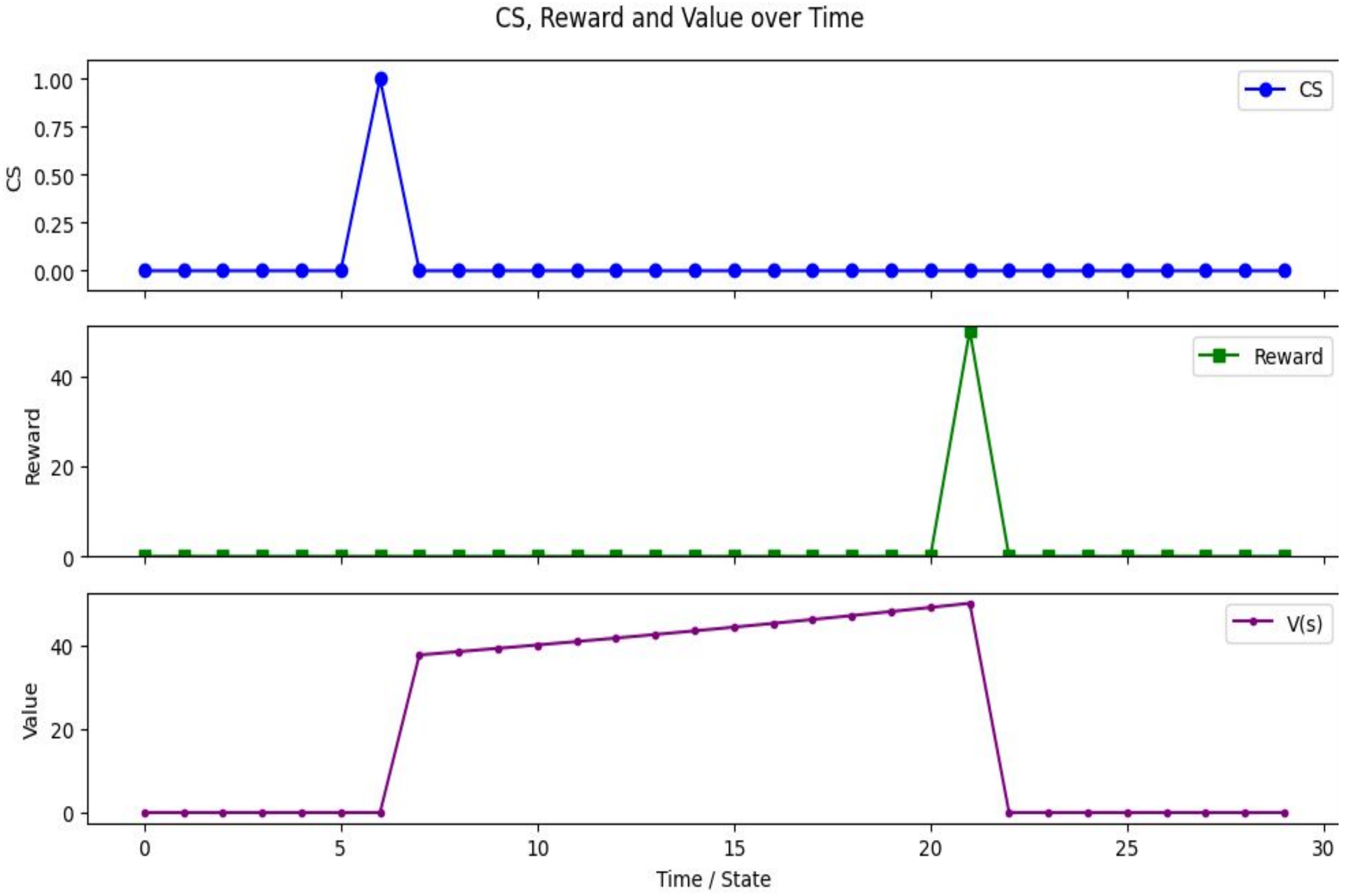
$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

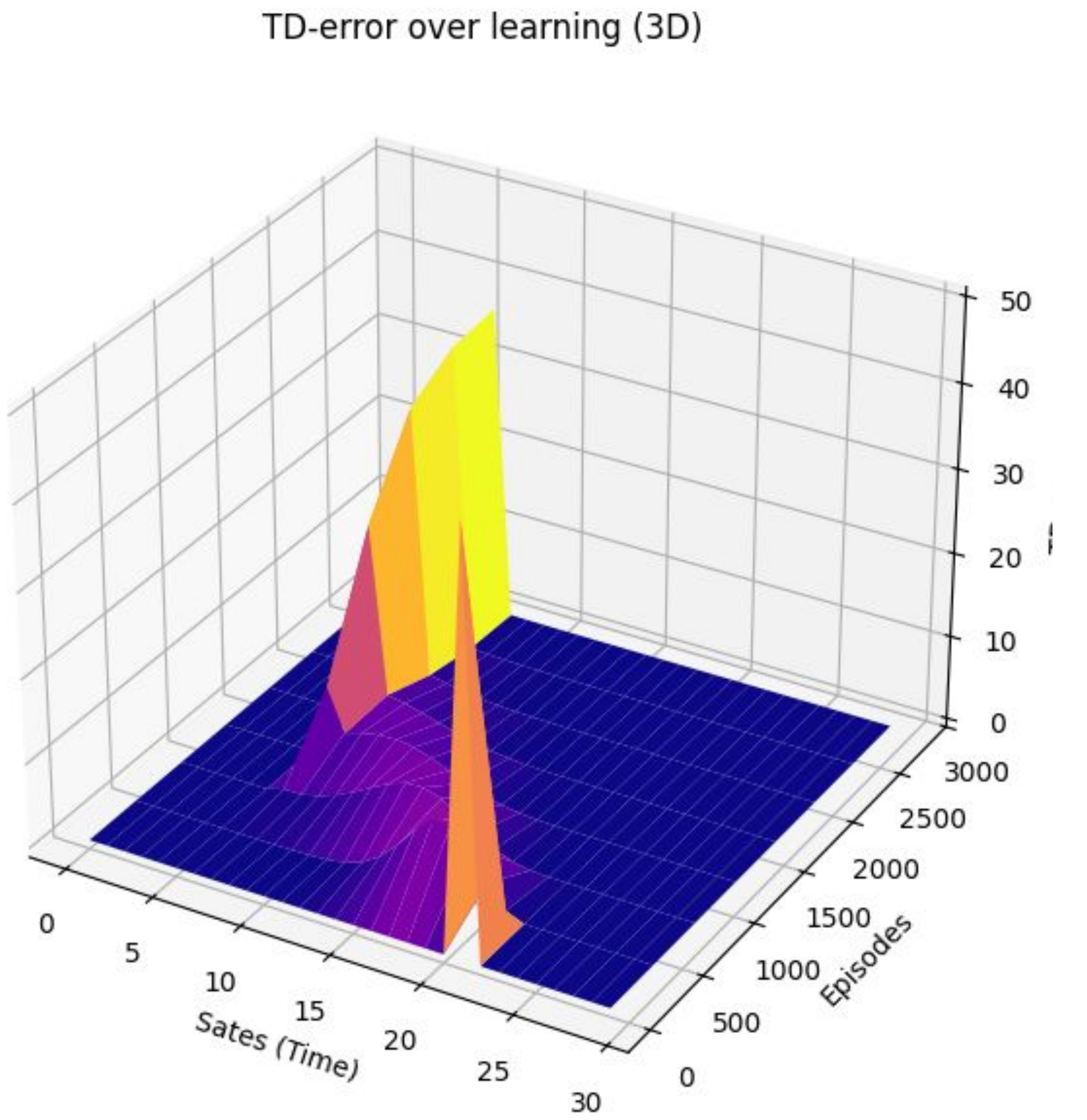
$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

In essence, first δ_t measures TD-error measures the discrepancy between values at time t and t+1, then we update V.

Neuromatch Academy. (2020). Available at: https://compneuro.neuromatch.io

TD-Learning: Case 1 guaranteed rewards





After many episodes, CS starts to predict the future reward, so when the reward arrives you were already expecting the reward and there is no prediction error.

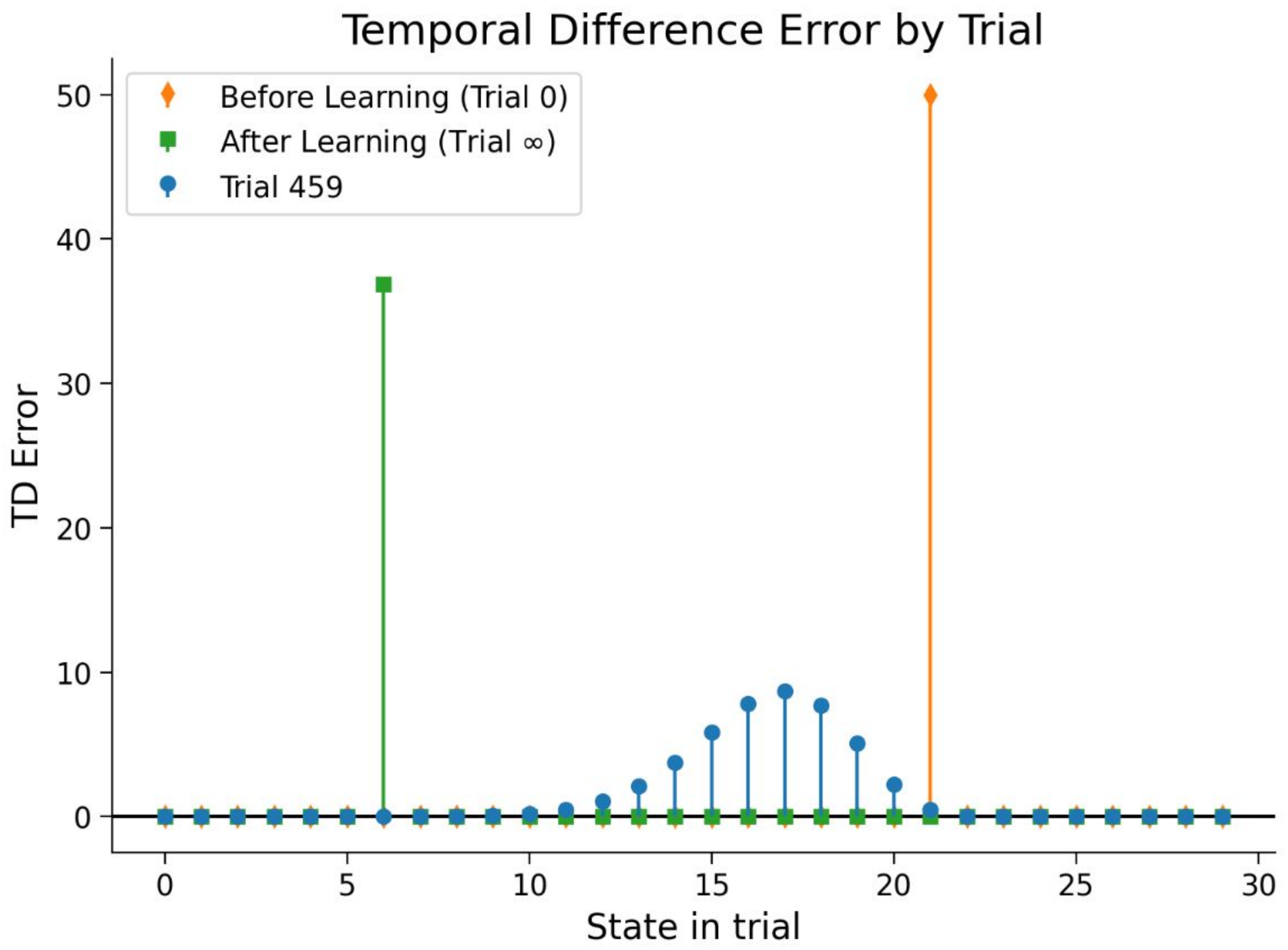
Neuromatch Academy. (2020). Available at: https://compneuro.neuromatch.io

On TD-learning,

- α Controls learning rate: How much to trust each TD-error.
- γ Discounts future rewards: Higher values mean more future-oriented learning.

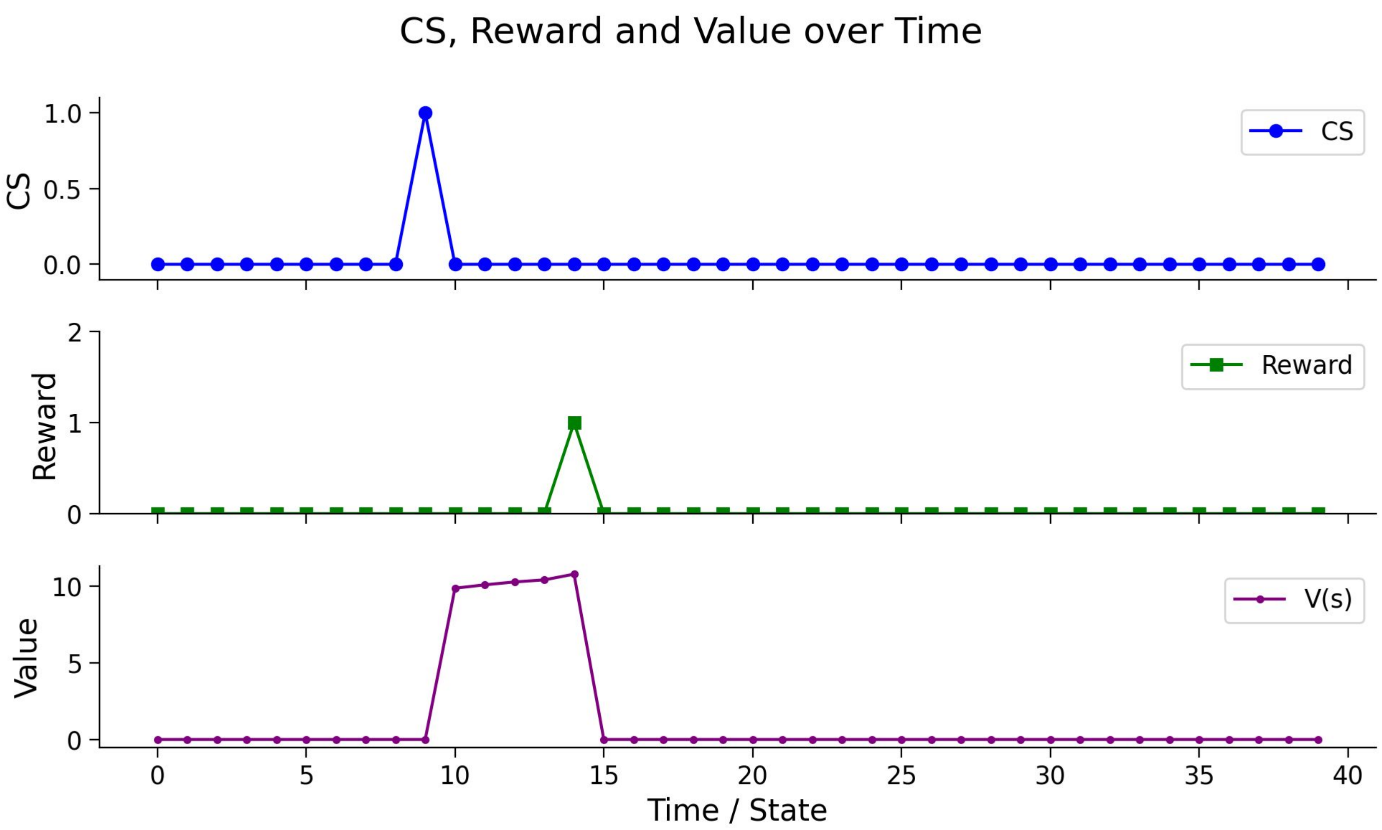
$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

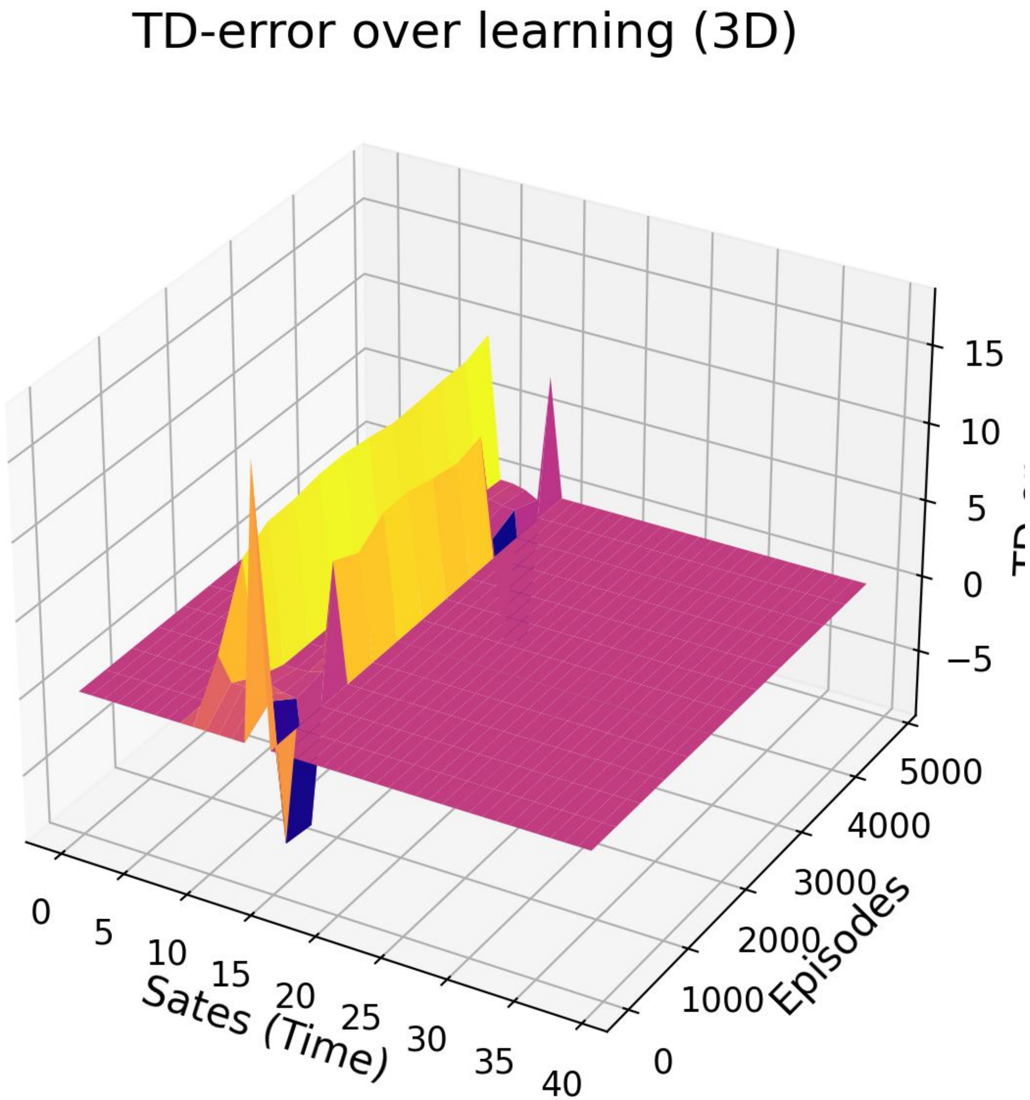
$$V(s_t) \leftarrow V(s_t) + \alpha \delta_t$$



Neuromatch Academy. (2020). Available at: https://compneuro.neuromatch.io

TD-Learning: Case 2 Multiple Reward Magnitudes

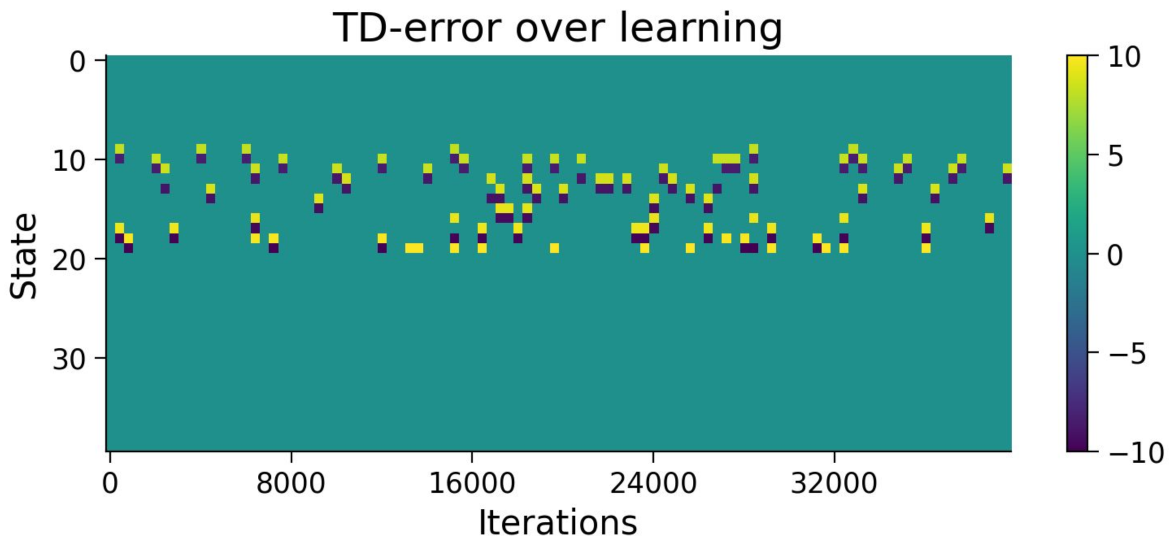


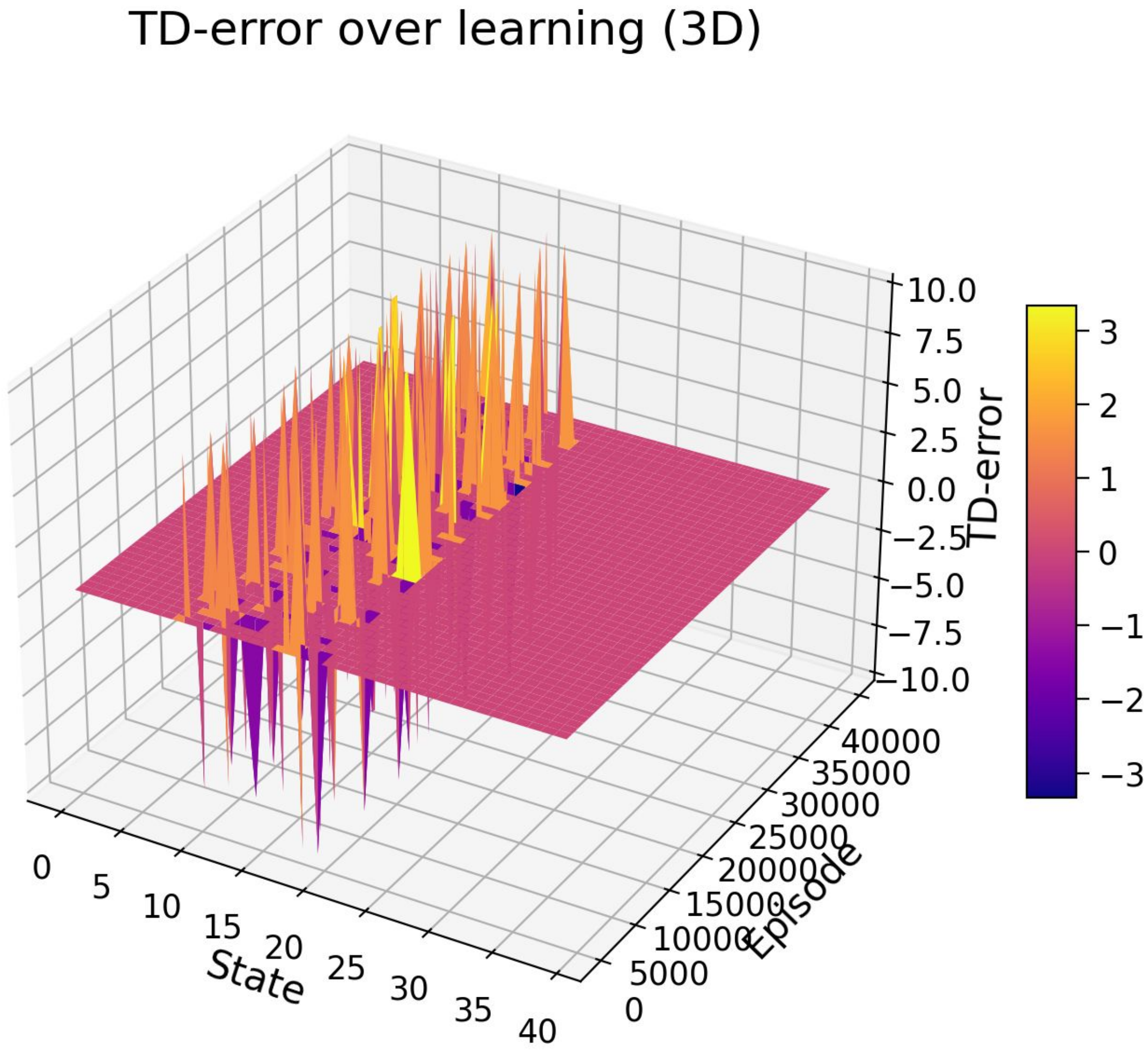


New Phenomena: Negative TD-Prediction with Variable Rewards:

$$\delta_t = 6 + \gamma \cdot 0 - 14 = -8$$

Neuromatch Academy. (2020). Available at: https://compneuro.neuromatch.io





Curiously, (p, α) are highly relevant for the learning process.
Let's consider a macroscopic measures that captures the total variation in learning.

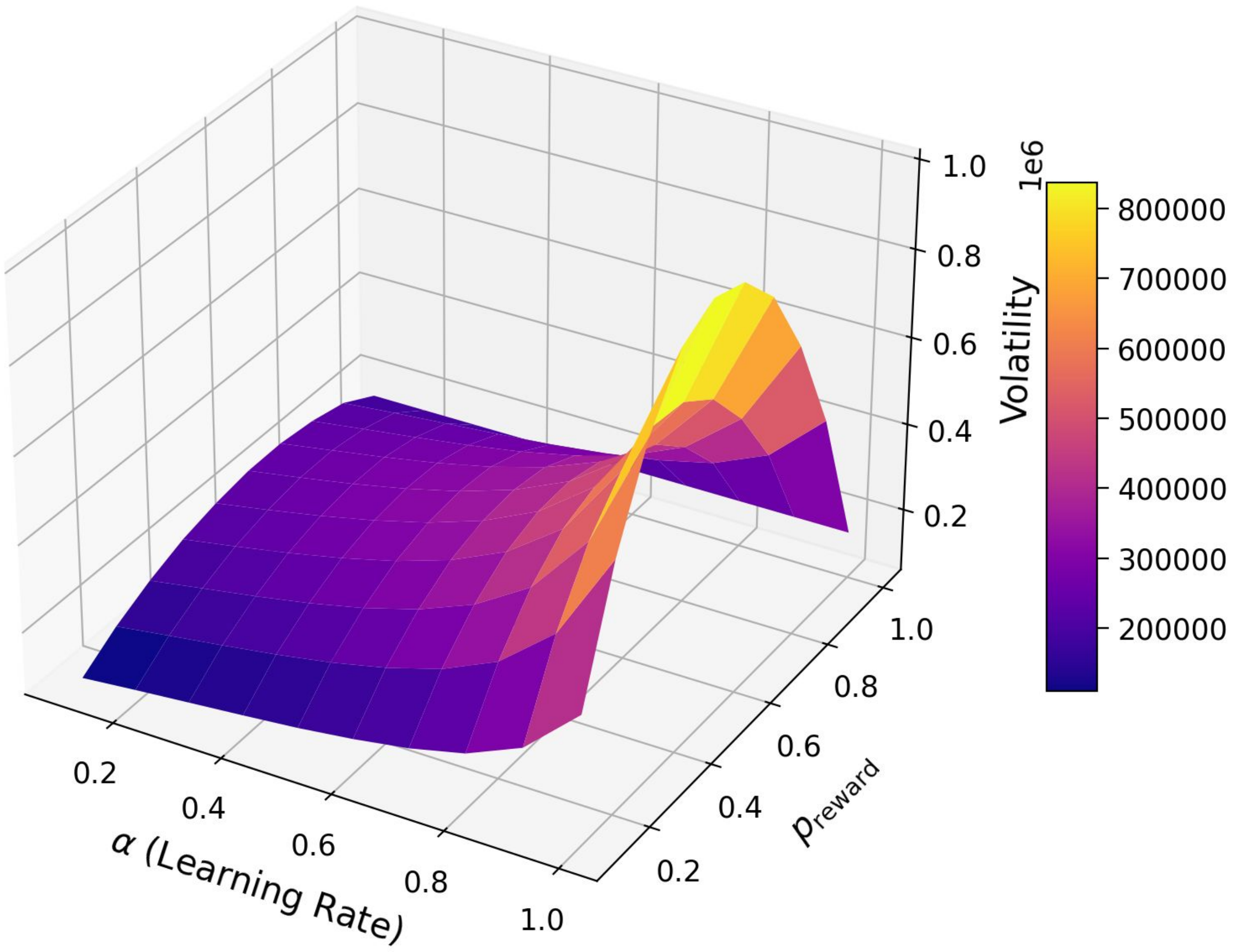
TD-Learning: Case 3 Probabilistic Rewards

$$\textbf{Cumulative Volatility} = \sum_e \sum_t \left|\delta_t^{(e)}\right|$$

🧠 Conclusions:

- Stochastic and deterministic rewards with the same expectation lead to equivalent TD updates.
- Cumulative volatility peaks when learning is fast (α high) and the environment is uncertain (p≈0.5).
- A population with heterogeneous (α,p) can model learning society.
- There may exist a nonlinear function p(α) reflecting how our rewards society reacts to your learning ability.

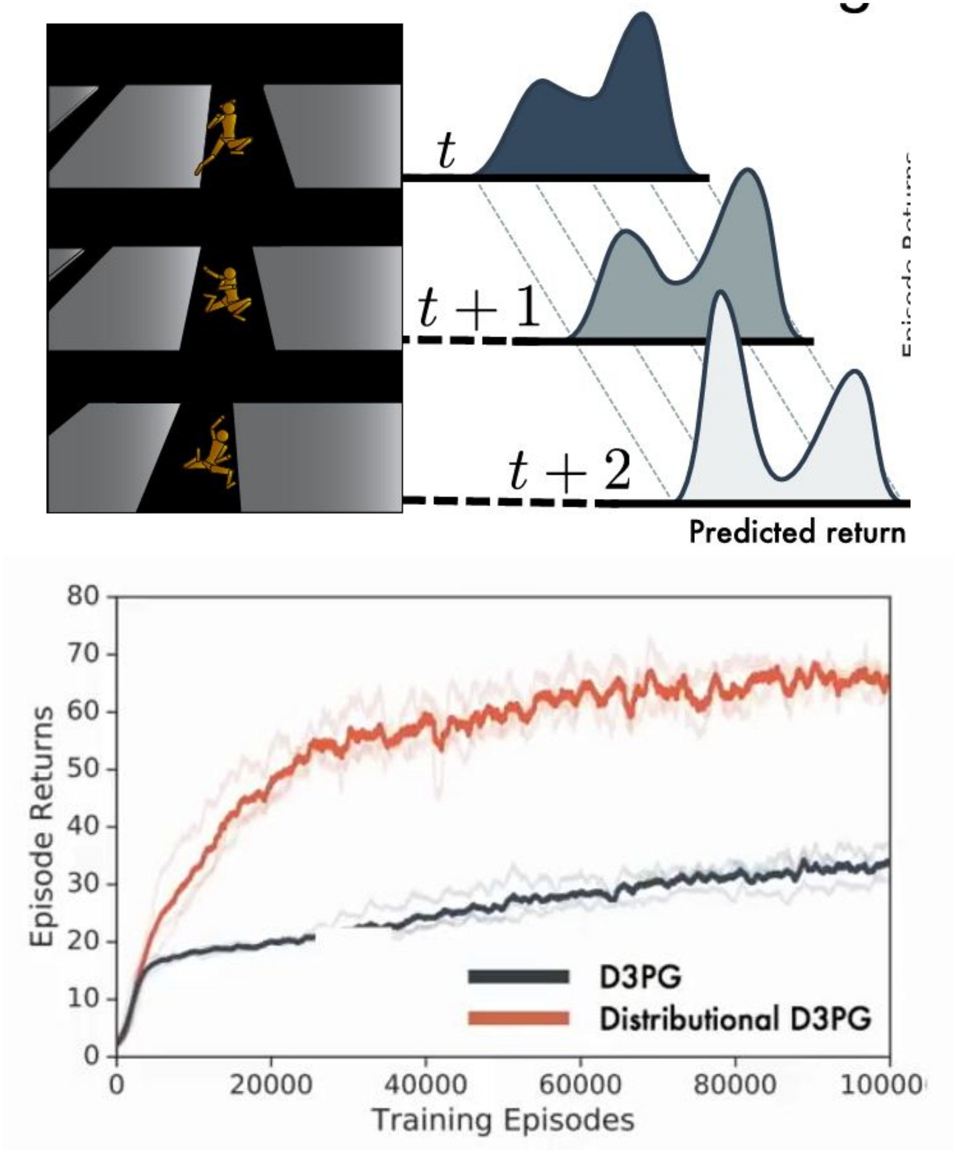
3D Surface: Cumulative TD-error Volatility Across α and p_{reward}



Neuromatch Academy. (2020). Available at: https://compneuro.neuromatch.io

Is Learning Just a Single Value?

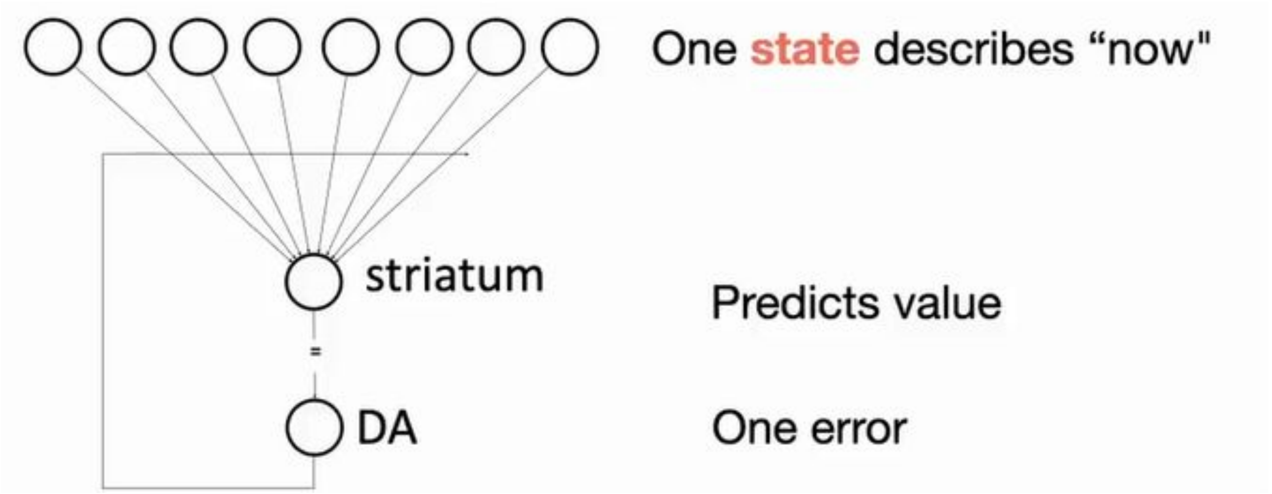
- In standard TD-learning, learning is reduced to estimating a **single expected value** of future rewards. However, in many real-world scenarios, rewards are variable and uncertain—better represented by a **distribution of predicted outcomes** rather than a single average. This idea is illustrated in the figure, adapted from [2].

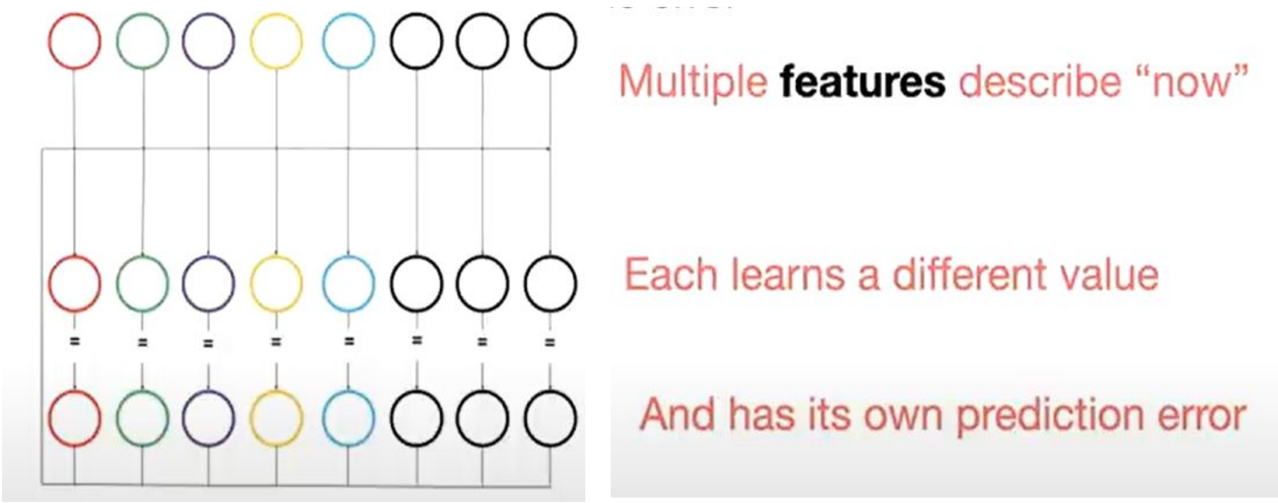


- Notably, distributional RL can increase performance in deep learning systems.

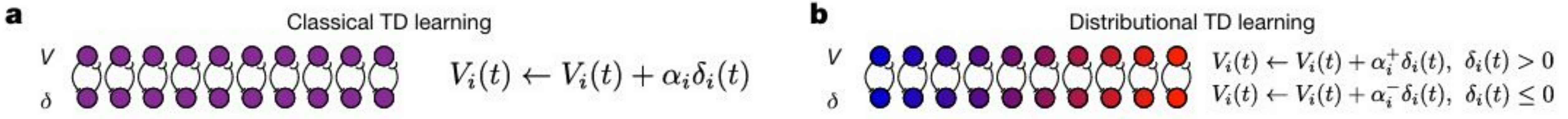
[2] A distributional code for value in dopamine-based RL. Dabney et al., Nature, 2020.

Standard TD-Learning vs Distribution TD-Learning





Insight: In essence each cell predicts different state



[2] A distributional code for value in dopamine-based RL. Dabney et al., Nature, 2020.

Distribution TD-Learning: Framework

🧠 **Standard TD Learning with Utility Function**

$$U(r) = \frac{f_{\max} \cdot \mathrm{sign}(r) \cdot |r|^{0.5}}{|r|^{0.5} + \sigma^{0.5}}$$

🧠 **TD-error & Update**

$$\delta = U(r) - V_i$$

$$V_i \leftarrow V_i + \alpha_i \cdot \eta \cdot \delta$$

🧠 **Spiking Reward prediction error**

$$R_{i,j} = \alpha_i \cdot (U(r_j) - V_i)$$

$$R_{i,j}^{\text{norm}} = \frac{R_{i,j}}{\mathrm{std}(R_{i,:})}$$

🧠 **Distribution TD Learning, Utility Function:**

$$U(r) = \frac{f_{\max} \cdot \mathrm{sign}(r) \cdot |r|^{0.5}}{|r|^{0.5} + \sigma^{0.5}}$$

🧠 **TD-error & Update using "Quantile":**

$$\delta_i = U(r) - Z_i$$

$$Z_i \leftarrow Z_i + \eta \cdot \left(\text{valence}_i \cdot \alpha_i^- + (1 - \text{valence}_i) \cdot \alpha_i^+\right) \cdot \delta_i$$

$$\text{valence}_i = \begin{cases} 1 & \text{if } \delta_i \leq 0 \quad \text{(negative error)} \\ 0 & \text{if } \delta_i > 0 \quad \text{(positive error)} \end{cases}$$

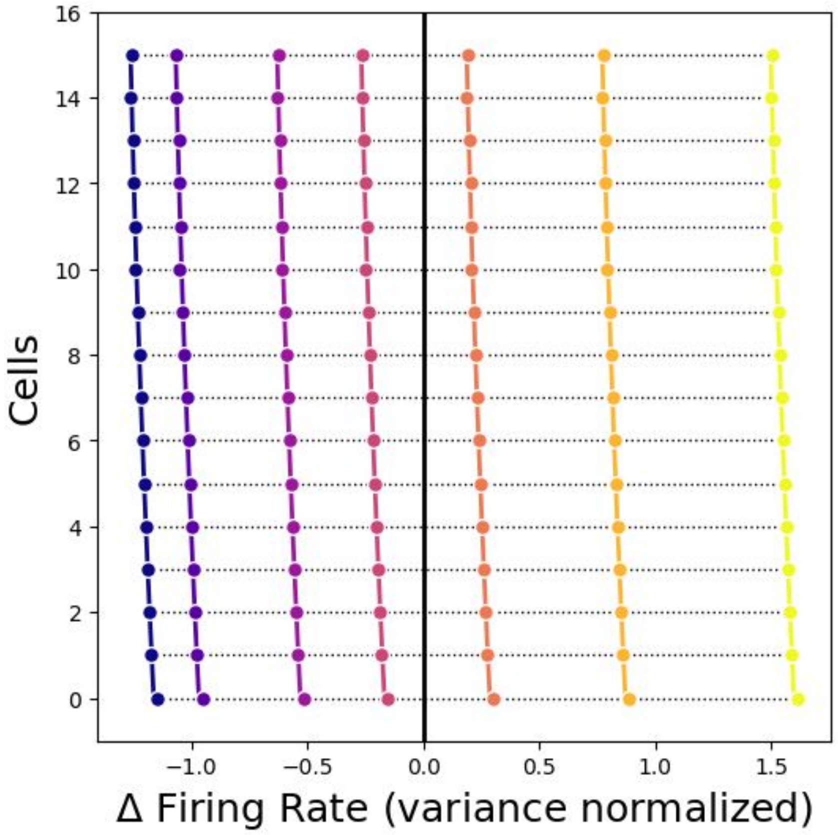
🧠 **Spiking Reward prediction error**

$$\delta_{i,j} = U(r_j) - Z_i$$

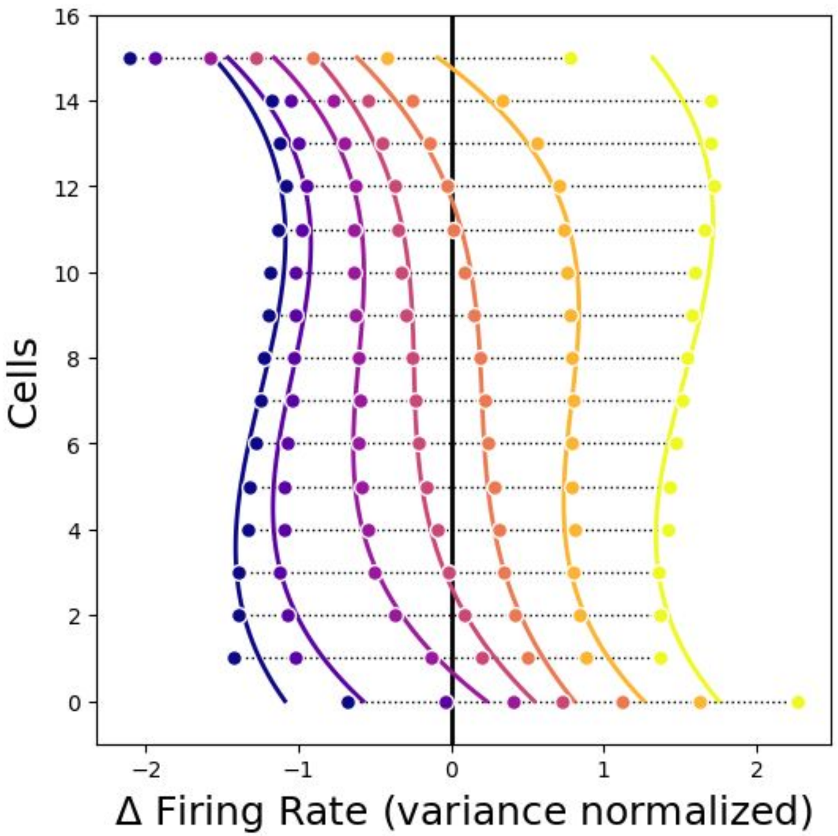
$$R_{i,j} = \left(\alpha_i^- \cdot 1[\delta_{i,j} \leq 0] + \alpha_i^+ \cdot 1[\delta_{i,j} > 0]\right) \cdot \delta_{i,j}$$

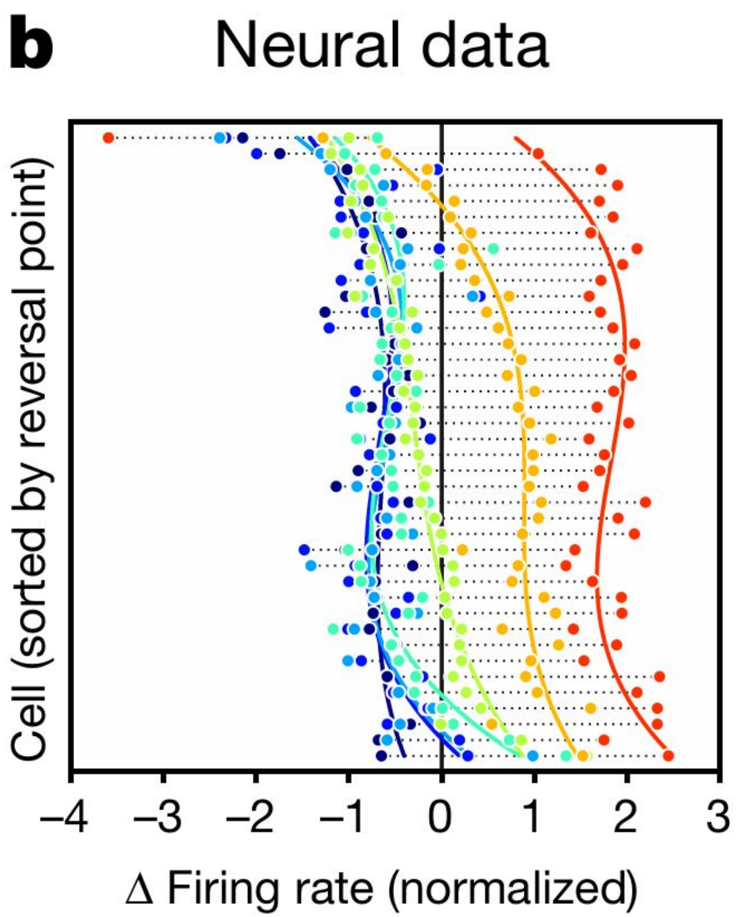
[2] A distributional code for value in dopamine-based RL. Dabney et al., Nature, 2020.

Standard TD-Learning



Distribution TD-Learning





Measured dopaminergic neurons in the Ventral Tegmental Area (VTA)

[2] A distributional code for value in dopamine-based RL. Dabney et al., Nature, 2020.

Conclusions :

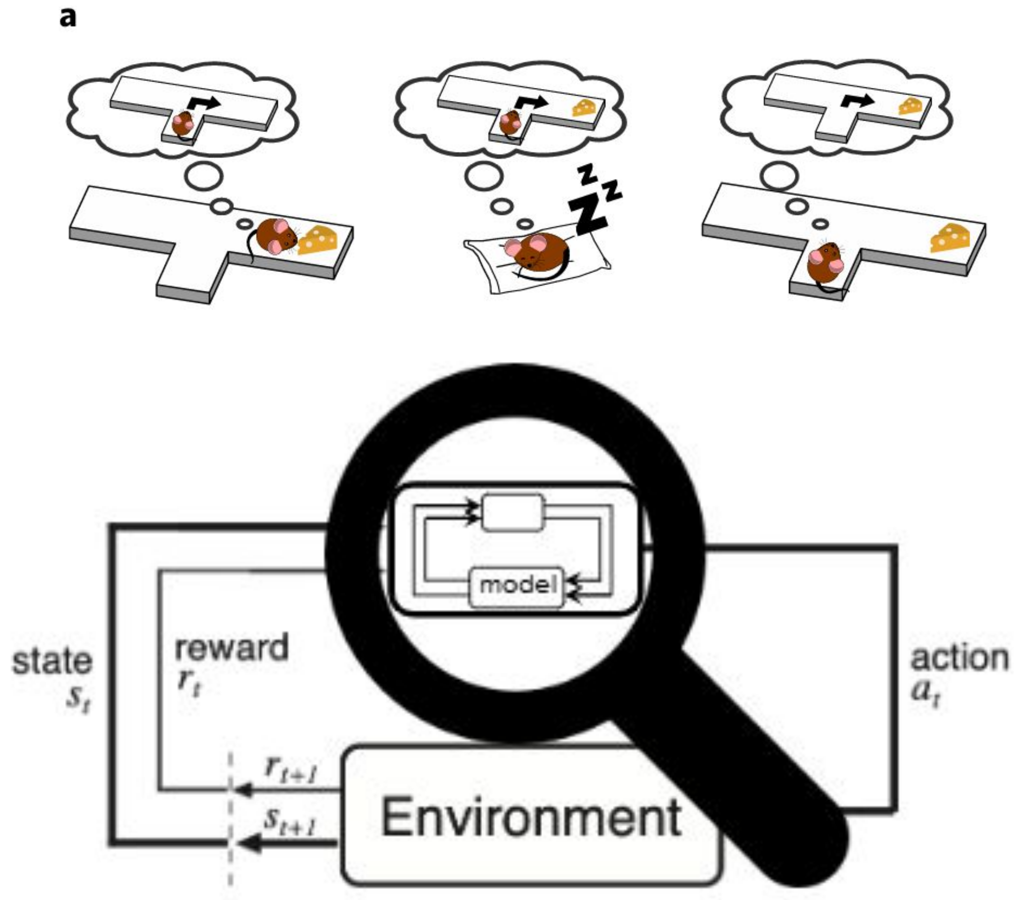
- **Classic TD model**:
  All neurons converge to similar responses; small variations are just noise.

- **Distributional TD model**:
  Neurons specialize — some are **optimistic**, some **pessimistic**, and others **neutral**.

- **Neural data (VTA dopamine neurons)**:
  This diversity is real and matching distributional predictions.

Other ways to see RL & Brain:

- **Model Free & Model Based.**



Prioritized memory access explains planning and hippocampal replay, Marcelo G. Mattar \& Nathaniel D. Daw, Nature Neuroscience, 2018.


THANKS