智能科学与技术学院
School of Intelligence Science and Technology

Distributional Reinforcement Learning in Dopamine Neurons

Neuroscience Project, Fall 2025

汇报人：黄荣昊 崔泽宇 陈浩扬

智能科学与技术学院
School of Intelligence Science and Technology

Overview

- Background
- Cat-Car Task Simulation
- Neural Evidence for Distributional Reward Prediction
- Decoding and Model Comparison
- Conclusion
- References

Pavlov's Experiment

Pavlov and His Dog

一百多年前，巴甫洛夫做了一个实验：

他给狗听一个蜂鸣声，然后过一会儿给它食物。几轮训练之后，狗一听到蜂鸣就开始流口水——尽管食物还没来。

——狗学会了预测奖励！

Dopamine and Reward Prediction

- 多巴胺神经元 ≈ 奖励预测误差信号；
- 这与人工智能中的「时序差分学习（TD learning）」高度一致：

  ——“预测误差”用来更新模型，让它学得越来越准；
- 那么核心问题来了：

  → 多巴胺神经元只编码平均奖励，还是完整的奖励分布?

Classical TD

我们通过“当前的预测值”和“实际的回报”之间的差值（prediction error），来更新模型。

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

这个 δ 是预测误差（prediction error），用来调整我们对当前状态的价值判断。

但是仅仅估计期望值够吗？

Distributional TD

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

我们可以把这些不同的单元理解为“大脑中对未来持不同态度的神经元”

- 这种机制构建了一个异质性的价值表示系统
- 在神经实验中也观察到类似现象：不同DA神经元对应不同反应模式（乐观、中性、悲观）

为了比较经典TD和分布式TD模型的差异，
我们设计了一个具有 **“猫猫 vs 汽车”** 风格的图像分类任务，
来模拟奖励结构的不确定性和变化。

Experiment Setup

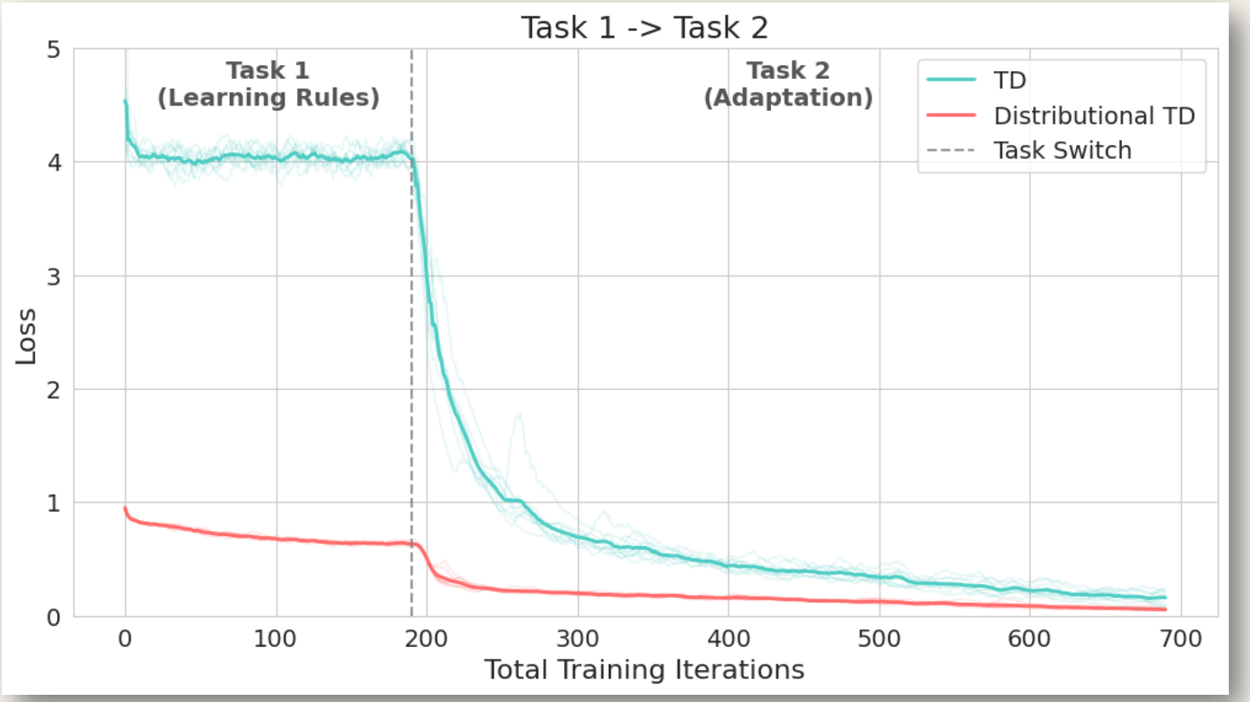
- **Car：** 代表稳定的奖励来源，始终给予 **+1** 分。
- **Cat：** 代表具有不确定性的刺激，前期有 **50%** 概率得到 **+5** 分，**50%** 概率得到 **–3** 分，虽然平均也是 +1 分，但波动很大。

之后，我们在任务第2阶段引入突变：
——猫图像不再是波动奖励，而是 **100%** **给予** **+5** **分**。

- **Classical TD：** 输出一个标量，学习平均值 。
- **Distributional TD：** 输出多个值，代表奖励的不同分位点。

Result

- **Task 1 学习阶段：** 两个模型都迅速收敛
- **Task 2 迁移阶段：** 红线（分布式）迅速下降，
  绿线（经典）缓慢下降



- **经典 TD 模型**需要慢慢重新学习新的平均值；
- **分布式 TD 模型**因为之前已经“知道”+5 分存在，
  只需微调概率权重即可完成迁移，**表现出显著更快的适应速度**。

-> 分布式模型更擅长应对环境的不确定性和动态变化。

Neural Evidence for Distributional Reward Prediction

多巴胺（DA）神经元是否在编码“奖励分布”，而不只是平均值?

我们提出的分布式 TD（Distributional TD）首先是一个假设模型。

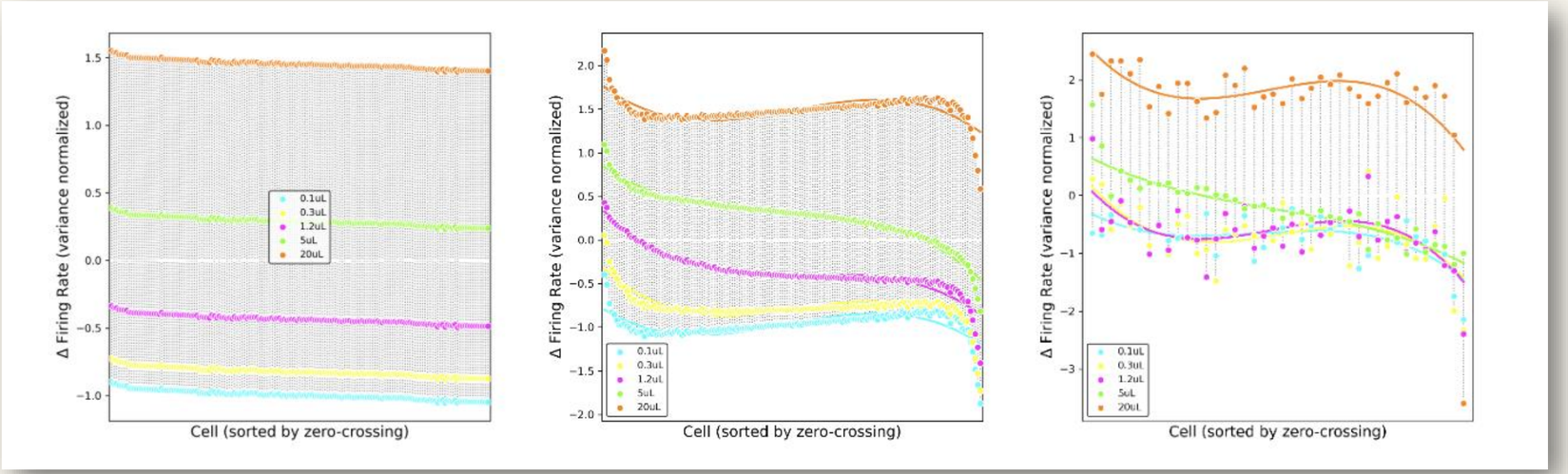
- 下一步关键是验证：它是否更贴近真实的生物神经系统?
- 为此我们分析小鼠 DA 神经元在两类任务中的放电数据，寻找“编码奖励分布”的证据。

- 实验1：变量奖励体积任务 → 反转点（Reversal Point）
- 实验2：变量奖励概率任务 → 乐观/悲观分化（t 统计量双峰）

Experiment 1

任务示意

- 提示音（cue）→ 等待（delay）→ 获得液体奖励
- 奖励体积从 5 个不同大小中随机抽取
- 数据中记录了 40 个 DA 神经元的反应



图： 左：经典 TD 单元　　中：分布式 TD 单元　　右：真实 DA 神经元

DA 神经元的反转点几乎覆盖整个奖励区间
——有人偏“乐观”，有人偏“悲观”，有人居中。

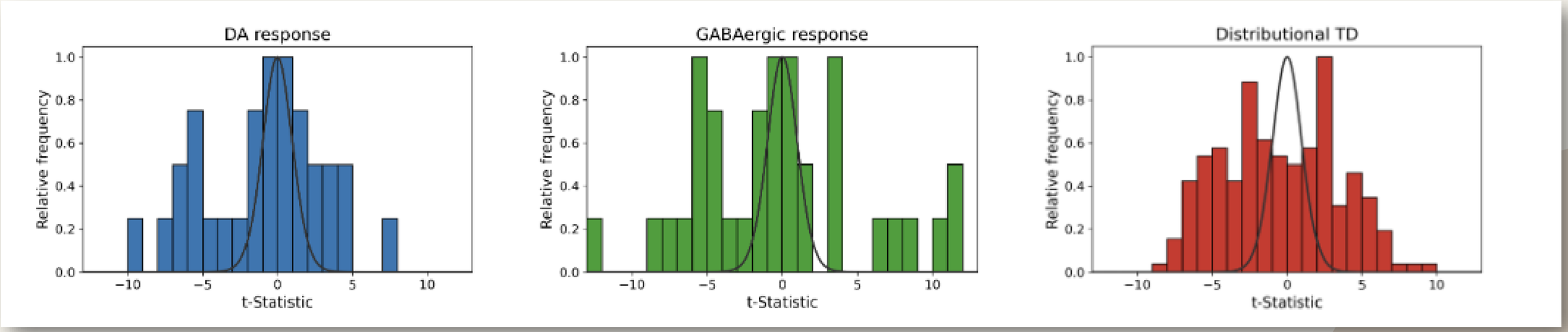
Experiment 2

任务示意

- cue 表示奖励概率：10% / 50% / 90%
- 我们重点观察 50% cue（最模糊、最能区分模型）
- 如果只编码均值：神经元对 50% cue 应该趋同
- 如果编码分布：可能出现乐观/悲观两类偏向

预测对比

- 经典 TD 预测：50% cue 反应应围绕“平均预测”，差异小
- 分布式 TD 预测：50% cue 下会分裂为偏向高概率/低概率的不同单元



图： 左：真实DA响应　　中：真实GAB响应　　右：模拟的分布式TD分布

真实 DA 数据的 t 统计量呈“双峰/两团”

Comparison Conclusion

生物证据支持“分布式奖励编码”：

对比维度	经典 TD	分布式 TD	真实 DA 神经元
编码内容	主要是均值	多通道表示分布（不同位置/分位点）	反应多样、方向各异
神经元多样性	低	高	高
实验1：体积反转点	倾向集中	覆盖整个区间	覆盖整个区间
实验2：50% cue	单峰/趋同	乐观-悲观分化	乐观-悲观分化（双峰）

- 结论：DA 神经元群体更像“分布式预测单元集合”，而不是“同一个均值预测器的复制品”。

因此，两项任务共同指向：DA神经元群体存在稳定的多样性，符合一种 ‘分布式编码奖励分布’ 的机制。这为Distributional TD 式单元提供了生物学层面的支持证据。

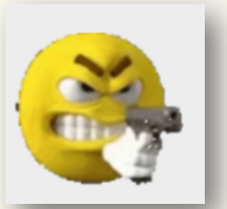
Decoding & Model Comparison

本部分的核心目标，是通过计算建模来论证：

分布式强化学习模型比传统TD模型能更好地解释多巴胺神经元的编码机制

为实现这一目标，我们尝试构建了一种解码分析工具。

分布式TD　vs　经典TD

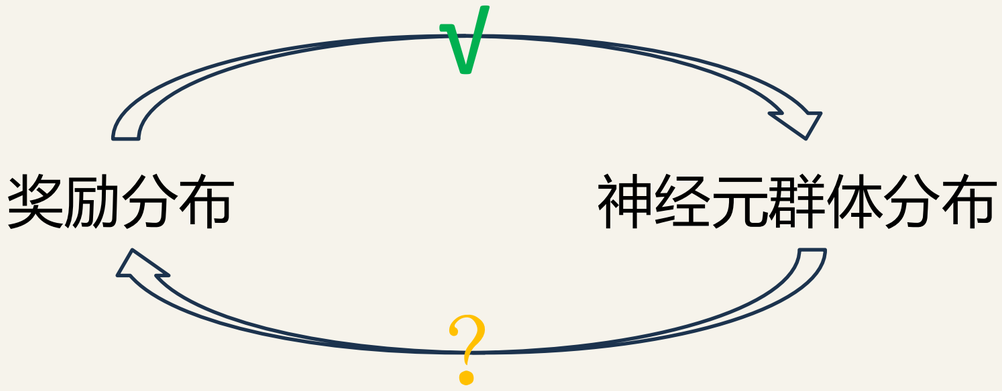


Reward Decoding

| 模型是否只是拟合数据？我们还能做得更进一步：

我们尝试使用解码分析（decoding analysis）从神经元群体活动中反向还原奖励分布，进一步测试 Distributional TD 的生物合理性。

- Decoding函数
- 真实神经元数据
- 人工制造奖励



Decoding Function

解码 (Decoding) 的意义：

如果我们能从一个神经元群体的活动中，反向推断出它们所编码的完整奖励分布，就能为分布式假说提供有力证据。

挑战

在分布式假说当中，神经元并非“客观”的传感器，它们有自己的“偏好”或不对称性（Asymmetry)。

Decoding Function

我们假设每个神经元编码的是奖励分布的一个**期望分位数 (Expectile)**，由其不对称性参数τ决定。
解码器通过最小化一个特定的损失函数来寻找最佳的奖励分布。

期望分位数回归损失函数:

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

y是真实的奖励值，$\hat{y}$是神经元的预测。
$(y - \hat{y})^2$是标准的预测误差。
τ是神经元的乐观悲观程度 (0到1之间，0.5代表中立)。
$|\tau - I(y < \hat{y})|$是**不对称惩罚项**。它使得当预测偏高或偏低时，受到的“惩罚”是不同的，这正是这个公式的关键。

Decoding Function

run_expectile_decoding：

- 接收一群神经元的“读数” (reversal_points) 和它们的“个性” (τ)。
- 内部定义了 expectile_loss_fn 来计算当前预测的“离谱指数”。
- 通过 scipy.optimize.minimize 算法，不断调整一个猜测的奖励分布，直到找到那个能让“离谱指数”最小化的最佳分布。

```
for _ in range(max_epochs):
    samples = np.random.uniform(minv, maxv, size=(max_samples, N))
    fvalues = np.array([expectile_loss_fn(points, tau, x0) for x0 in samples])
    x0 = np.sort(samples[fvalues.argmin()])
    result = scipy.optimize.minimize(
        lambda x: expectile_loss_fn(points, tau, x),
        method=method,
        bounds=[(minv, maxv)] * len(x0),
        x0=x0
    )
```

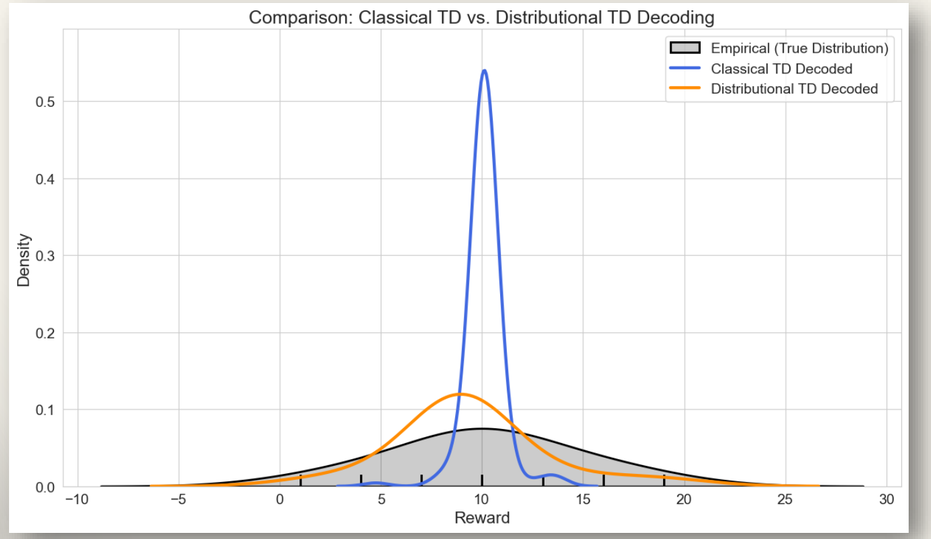
Classical VS Distributional

假设每个 DA 神经元代表一个特定 expectile，我们利用这些神经元对奖励的响应，用回归方式尝试重建整个奖励分布。

我们对比了三种来源的奖励分布：

- 经典 TD 模型模拟数据
- 分布式 TD 模型模拟数据
- 真实奖励分布

得到的分布图像如下：



经典TD下模拟的分布只还原出单一平均值分布，丢失结构。

分布式TD模拟的分布结构高度吻合 ground truth。

只有 Distributional TD 架构能极大程度上还原出奖励信号的真实分布

Conclusion

我们通过一个完整链条的分析：

行为任务 → 模型仿真 → 神经记录 → 反向解码

发现：

- 多巴胺神经元不只是编码平均值，它们可能携带**整个奖励分布的编码信息**；
- Distributional TD 在生物解释力、学习效率、解码能力上**均优于经典 TD**；
- 这为理解 DA 系统的灵活性与不确定性下的学习策略提供了新视角。

References

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

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

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

智能科学与技术学院
School of Intelligence Science and Technology

THANK YOU

For your attention