確率的報酬課題におけるドーパミンニューロンの活動度の解釈
Interpreting the dopamine activities in stochastic reward tasks

朝比奈 亜貴代 (PY)†, 平山 淳一郎‡, 石井 信‡†
Akiyo Asahina (PY), Jun-ichiro Hirayama, and Shin Ishii
† 奈良先端科学技術大学院大学 情報科学研究科　　‡ 京都大学 大学院情報学研究科
akiyo-a@is.naist.jp, junich-h@sys.i.kyoto-u.ac.jp, ishii@i.kyoto-u.ac.jp

Abstract— The phasic activations of dopamine (DA) neurons in the primate midbrain have been considered as representing temporal difference (TD) errors from a computational perspective. Recently, several studies have reported that, in stochastic reward tasks, the DA activity can gradually increase before receiving actual rewards, which is not well explained by the simple TD model. We propose an alternative model based on a probabilistic formulation of the stochastic reward task. In a simulation experiment, our model well described the gradually increasing DA activities during a wait period even for a single trial.

Keywords— **dopamine neurons, stochastic reward task, uncertainty, temporal difference error**

1　まえがき

動物が適切に行動するには，将来得られるであろう報酬量をできるだけ正確に予測することが重要である．中脳のドーパミン (DA) 作動性細胞は，報酬予測誤差に対応して活動することから，従来，時間差分 (TD) 誤差を表現して強化学習に関わるという計算論的仮説が提案されてきた [1]．しかし，報酬が確率的に与えられる状況などにおいて，単純な TD モデルでは説明できないような DA 細胞の挙動が報告され，その解釈について議論が行われている．特に，報酬を待つまでの間の DA 細胞の活動度上昇 (gradual increase; GI) は，従来の Niv らのモデル [4] では TD 誤差の試行平均を取らないとならず，Fiorillo らの実験 [5] における 1 試行中でも見られる GI 活動は説明できていない．本研究では，確率的報酬課題を単純な確率モデルにより定式化し，それに基づいて DA 細胞の GI を比較的良く説明しうることを示す．

2　確率的報酬課題における DA 信号の挙動

DA 信号が TD 誤差と類似することを示した Shultz らの研究 [1] によれば，サルを用いた遅延報酬課題において，DA 細胞は，学習と共に報酬そのものではなく報酬を予測する刺激に対して反応するようになり，また，予期した報酬が得られない場合には発火率が自発レートよりも下がる．これらの結果は，報酬が刺激後に必ず与えられる決定論的報酬課題におけるものであったが，近年 Fiorillo らや Tobler らは，確率的報酬課題での DA 細胞の挙動を報告している [2, 3]．

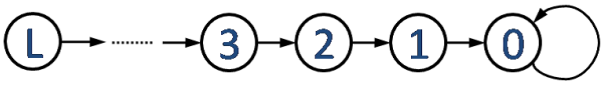

図 1: 状態遷移図

Fiorillo らの課題では，課題の始まりを表す合図として視覚刺激を 2 秒間見せ，視覚刺激が消えたと同時に，その視覚刺激に対応する確率で報酬が与えられる．視覚刺激は 5 種類あり，それぞれが報酬確率 $p = 0, 0.25, 0.5, 0.75, 1.0$ に対応する．また Tobler らの課題では，報酬確率は $p = 0.5$ と固定であるが，視覚刺激は 3 種類あり，それぞれが報酬量 r=0.05, 0.15, 0.50 ml に対応する．サルは $p = 0.5$ で，刺激に対応する量の報酬，または 0 ml (無報酬) を得る．Fiorillo らの実験では，確率 $p = 0.5$ で視覚刺激から報酬までの時間での DA 細胞の GI が最大となるため，DA 細胞がコードするのは予測誤差のみならず，“ 報酬獲得に対する不確かさ ”であると主張された [2, 3]．Niv ら [4] は，TD 誤差において，DA 細胞の発火率が自発レートより上がる正の活性と，下がる負の活性との間で非対称性があることを考慮した上で，試行間で平均を取ることを行えば，TD 学習のアーチファクトとして GI が説明可能であるというモデル研究を行った．それに対し Fiorillo ら [5] は，GI は，課題 1 試行でも起こるものであり，DA 信号は確率的報酬に対する不確かさが関与することを述べている．

3　提案モデル

確率的報酬課題を以下のように定式化する．あるエピソード中の離散時刻を $t \in \{1, 2, \ldots, t_{\max}\}$ で表す．$t = 1$ は刺激提示時刻であり，$t_{\max}$ はエピソードの長さとする．また，報酬が (確率的に) 与えられる時刻を t_{reward} とする．時刻 t における報酬量を r_t で表す．過去の報酬系列 $\mathcal{H}$ に基づき，以下の期待収益を適切に推定することを目的とする．

$$v_t = E\left[\sum_{u=0}^{t_{\max}-t} \gamma^u r_{t+u} \middle| \mathcal{H}\right] \qquad (1)$$

ここで，$\gamma \in (0, 1)$ は割引率である．

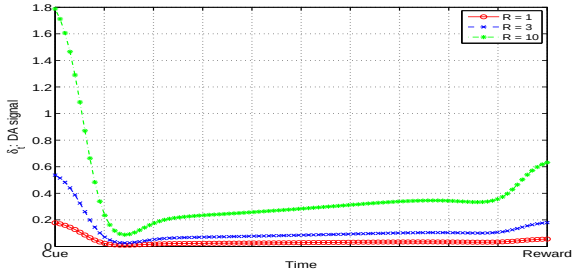

図 2: 報酬量を変化させた際の GI

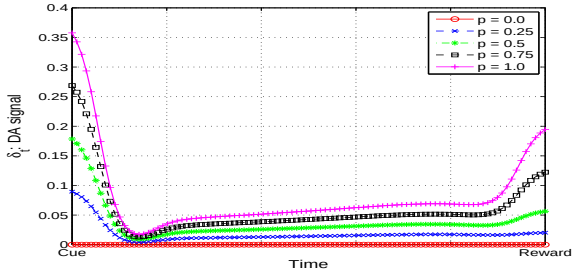

図 3: 確率を変化させた際の GI

本研究では，サルは外界に対する次のような生成モデルに基づいて期待収益を推定すると仮定する．ここで，報酬時刻 t_{reward} はサルにとっては未知であり，実験者が固定していたとしても，認知的には確率的にゆらぐと考えられる．エピソード中の (認知的) 状態遷移をモデル化するために，図 1 のような吸収状態を持つマルコフ鎖を仮定する．状態変数を $z_t \in \{0, 1, 2, \ldots, L\}$ とし，$z_t = 1$ で報酬イベントが起こるものとする．また，$z_t = 0$ を他状態への遷移が起こらない吸収状態とする．さらに，$z_{t-1} = l$ $(l \geq 1)$ のとき，確率 1 で $z_t = l - 1$ に遷移するものとする．初期状態は確率分布 $\phi(l)$ にしたがって生成されるものとする (ただし $\phi(0) = 0$)．このとき，$z_t \in \{1, 2, \ldots, L\}$ は報酬イベントまでの残りステップ数，$z_t = 0$ を報酬イベント終了後の状態とみなすことができる．これらの間の遷移が決定論的に起こるのに対し，初期状態確率 $\phi(l)$ のみが確率的であり，これは報酬時刻 t_{reward} の分布に対応する．これらの各状態に依存して，観測 r_t の条件つき分布を以下で与える．

$$p(r_t|z_t) = \begin{cases} N(r_t|\mu, \sigma^2) & (z_t = 1) \\ N(r_t|0, \epsilon^2) & (z_t \neq 1) \end{cases} \tag{2}$$

μ, σ^2 は報酬の平均値と分散を表すモデルパラメータであり，ϵ^2 は十分小さな定数とする．

いま，モデルパラメータ μ，σ^2 は適切に学習済みとする．このとき，式 (1) は次のように計算される．

$$v_t = \sum_{z_t} p(z_t \mid \mathbf{r}_{1:t}) V(z_t) \tag{3}$$

$V(\cdot)$ は状態価値関数であり，以下で定義される．

$$V(z_t) = \sum_{\mathbf{z}_t^+} p(\mathbf{z}_t^+ \mid z_t) \sum_{u=0}^{t_{\max}-t} \gamma^u m(z_{t+u}) \tag{4}$$

ここで，$\mathbf{z}_t^+ = (z_{t+1}, \ldots, z_{t_{\max}})$ とおいた．また，

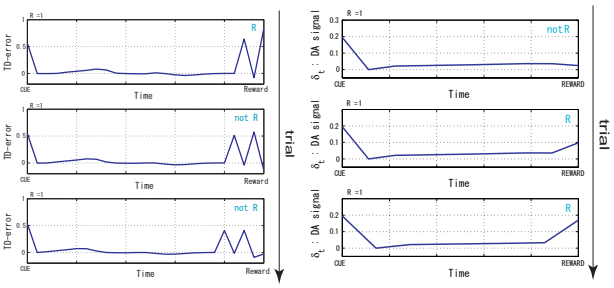

図 4: 1 試行での GI; 左: Niv らのモデル, 右: 提案モデル

$m(z_t) = \mu$ $(z_t = 1)$，$m(z_t) = 0$ $(z_t \neq 1)$ である．さらに，状態遷移が決定論的であることを用いると，価値関数は，以下の単純な形で書くことができる．

$$V(l) = \begin{cases} 0 & (z_t = 0) \\ \gamma^{l-1}\mu & (z_t \neq 0) \end{cases} \tag{5}$$

各時刻における期待収益の推定値 v_t について，その差分 $\delta_t = v_t - v_{t-1}$ を定義すると，δ_t は TD 学習における TD 誤差に類似した量となる．本研究ではこれを DA 信号と解釈する．

4 計算機実験とまとめ

計算機実験により，提案モデルにおける DA 信号 δ_t の基本的な挙動を確認した．この実験では，報酬時刻の認知的ゆらぎを再現するために，報酬生成のタイミングを実際の行動実験とは異なり確率的に与えた．そのために，まず，真の状態系列 $z_1, z_2, \ldots, z_{t_{\max}}$ を前節のモデルに基づき生成した．報酬生成については，式 (2) を用いる代わりに，より実験設定に近い形で与えた．つまり，$z_t \neq 1$ のとき常に $r_t = 0$ とし，$z_t = 1$ のとき確率 p で $r_t = R$，$1 - p$ で $r_t = 0$ とした．報酬確率 p と報酬量 R はタスク中は固定とした．計算機実験の結果より，GI は推定した状態価値関数の割り引きによって，説明できることが確認できた (図 2, 図 3)．加えて，1 試行での GI は，Niv らのモデルでは再現できなかった (図 4 左) のに対し，提案モデルでは説明可能である (図 4 右)．以上より，価値関数の推定に対する時間差分が，DA 信号の GI に相当する可能性が示された．

参考文献

[1] Schultz et al. (1997) “A neural substrate of prediction and reward.” Science, **275**, 1593–1599.

[2] Fiorillo et al. (2003) “Discrete coding of reward probability and uncertainty by dopamine neurons.” Science, **299**, 1898–1902.

[3] Tobler et al. (2005) “Adaptive coding of reward value by dopamine neurons.” Science, **307**, 1642–1645.

[4] Niv et al. (2005) “Dopamine, uncertainty and TD learning.” Behavioral and Brain Functions, **1**:6.

[5] Fiorillo et al. (2005) “Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating TD errors.” Behavioral and Brain Functions, **1**:7.