2N5-OS-03b-4

線条体ニューロンの持続的発火と強化学習

A new reinforcement learning model inspired by firing prolongation in the striatum

太田 宏之[*1] 甲野 佑[*2] 高橋 達二[*2]
Hiroyuki Ohta Yu Kohno Tatsuji Takahashi

[*1] 防衛医科大学校生理学講座
National Defense Medical College, Department of Physiology

[*2] 東京電機大学
Tokyo Denki University

We propose a new reinforcement learning model which was inspired by the prolonged firing observed in the rat striatal neurons. The proposed model has "QTimer" which counts the remaining steps to an update of state-action value, based on the Monte Carlo method, and which is also used as a kind of eligibility trace. The state-action value is asynchronously and intermittently updated, and reference to the previous state-action values at each step is not required. This feature contributes to keep absolute changes of the state-action value small. Moreover, the model is extendable for the parallel processing for adapting to a non-Markovian environment.

1. はじめに

本稿では，ラット大脳基底核線条体のニューロンにおいて観察された持続的発火及びその残存現象にヒントを得て，環境の状態遷移に関するマルコフ性が仮定できないタスクに対応できる強化学習モデルを提案する．

学習試行中における大脳基底核のドーパミンニューロンの挙動[Schultz 95]が TD 誤差(Temporal Difference Error) [Barto 95]に似ていることなどから，大脳基底核は強化学習の座と目されている[Houk 95]．また，線条体ニューロンの一部は行動価値関数(Q 値)を表現しているとの報告もあり[Samejima 05]，大脳基底核の挙動を理解する目的で強化学習理論が参照されている．

一般的な強化学習理論の枠組みにおいては，環境の状態遷移と行動，報酬に関してマルコフ決定過程が前提とされ，遅延報酬課題を解く場合には，適格度トレース(eligibility trace) [Klopf 72]と呼ばれる仕組みが導入されたモデルが提案されている．

我々は，大脳基底核線条体のニューロンが 8 秒～20 秒程度の頻度の繰り返し刺激入力によって数秒間持続的に発火し，かつ繰り返しによって徐々にその発火期間が延長していく，という現象を発見した（論文投稿中）．このことから線条体が何らかの時間方向の積分器的能力を持っている可能性が想定される．この現象をヒントに，個々の状態－行動対（線条体ニューロン）における自己完結的な適格度トレースと可塑性の存在を仮定した，環境のマルコフ性を前提としない強化学習モデルを提案する．

2. 線条体の持続的発火及び残存現象

青色光刺激によって脱分極を誘発可能なチャネルロドプシン 2 が発現したトランスジェニックラット[Tomita 09][Ji 12][Ohta 13]の線条体の急性スライスを作成し，顕微鏡下で青色光刺激時の活動電位（発火）を細胞外で記録した．1 秒間の光刺激を 8 秒に 1 回の頻度で 5 回繰り返したところ，発火数と発火期間が徐々に増加した．また，繰り返し光刺激を 5 回行った後，休止期間をおいて再度光刺激を 1 回加えた際の残存スパイク数は，休止期間 20 秒までであれば 5 回目の光刺激時のスパイク数と同程度であり，それ以上の休止期間を経ると減衰した．また，この時間発展する現象は，カルシウム濃度増大に関するポジティブフィードバックに依存していることがわかった．

このカルシウム依存性の持続的現象は，適格度トレースのような持続的変数の根拠の候補の一つとなり得るものであり，また同時にニューロン単位の自己完結的な Q 値の更新プロセスの存在を予期させる．この現象をヒントに，以下に示すような，ステ

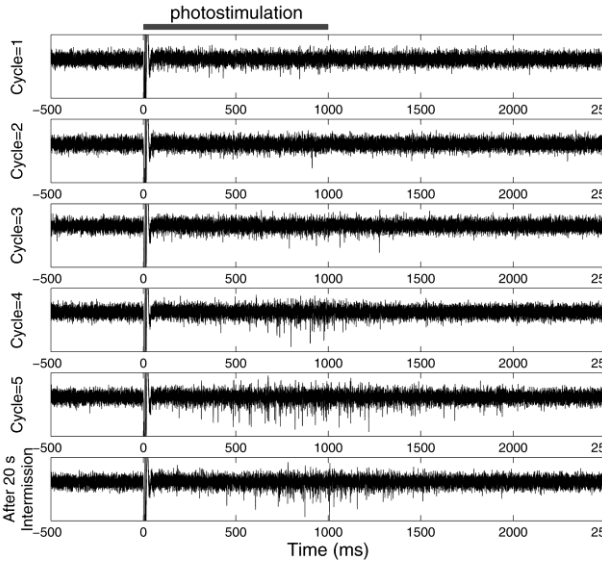

図 1 チャネルロドプシン発現ラット線条体ニューロンの光刺激応答．8 秒に 1 回の頻度で 1 秒間の光刺激を 5 回与えたところ，活動電位の数が増えると同時にその発生期間が徐々に延長した（上から 5 段目まで）．その後，20 秒間の休止期間を経て再度光刺激入力を行ったところ，時間発展した持続的発火が残存していた（5 段目）．このことは，10 秒単位の短期的な積算機能が線条体ニューロンに存在していることを示している．平均すると，この残存発火（5 回目）は 20 秒までの休止期間であれば最後の発火数（5 回目）と同程度であり，30 秒で半減，60 秒以上でほぼ消失する．

連絡先：太田宏之，防衛医科大学校生理学講座，埼玉県所沢市並木3－2，042-995-1225(内線 2227)，ohta@ndmc.ac.jp

ップ毎の Q 値更新と状態－行動対管理担体(ニューロン)の外部からの Q 値参照が不要な強化学習モデルを提案する.

3. QTimer Model 1.0

本研究で提案する QTimer モデルは Sarsa(λ)等の一般的な強化学習モデル[Watkins 92]と同様に, 行動価値関数Q(s, a)を用いて, 各ステップの観測状態 s 毎に行動 a をε -greedy 方策で選択する($\varepsilon = 0.1$). ただし更新方法が異なり, Q(s, a)は状態 s を初回訪問したときに起動されるタイマー(QTimer)が終了した後に, その間に得た収益に対してモンテカルロ法と同様の更新がされる. 各状態行動対に対する QTimer の残り時間$T(s_i, a_j)$は次のように定義し, ステップ毎に更新される.

$$T(s_i, a_j) = \begin{cases} T_{max}, & s_i = s_t \wedge a_j = a_t \wedge T(s_i, a_j) = 0 \\ T(s_i, a_j) - 1, & T(s_i, a_j) > 0 \end{cases}$$

T_{max} は QTimer の最大値である. QTimer の残り時間$T(s_i, a_j)$が非 0 の時はそのステップで得られた報酬r_tは, 残り時間に依存した重みを付けて QTimer 起動中の収益$R(s_i, a_j)$に加算して保存される.

$$R(s_i, a_j) = \begin{cases} R(s_i, a_j) + \gamma^{(T_{max} - T(s_i, a_j))} r_t, & T(s_i, a_j) \neq 0 \\ 0, & otherwise \end{cases}$$

QTimer の残り時間 $T(s_i, a_j)$が非 0 から 0 になった時, それまでに得た報酬に関する決算として Q 値の更新が個々の$Q(s_i, a_j)$において行われる. その状態行動対への初回訪問からT_{max}ステップ後まで更新を留保する点から QTimer モデルは半オンライン的な学習を行うモデルである.

$$Q(s_i, a_j) = Q(s_i, a_j) + \alpha \left(R(s_i, a_j) - Q(s_i, a_j) \right)$$

ここで, αは学習率, γは QTimer の残ステップ数を Q 値の更新量に反映させる係数(割引率)である. タイマー起動中の収益を用いて行動価値の更新を行う点から, QTimer model は$T_{max} = \infty$ とすると強化学習における初回訪問モンテカルロ法と一致する.

4. 遅延報酬課題

提案モデルを Sarsa(0)及び Sarsa(λ=0.9)と比較するため, 次のような課題を用いた.

状態数は 10 個(s0~s9)あり, ループ上に連なっている(図2). それぞれの状態につきエージェントが実行可能な行動選択肢は,「右に行く(例:s0→s1)」,「左に行く(例:s0→s9)」,「留まる(例:s0→s0)」の 3 種である. 状態には特殊な状態である原因状態(s2)と報酬状態(s7)が存在する. 報酬は s2 を訪問した後, 10 step 以内に s7 を訪れたときに発生する. エージェントが報酬を得た時の状態遷移は通常と異なり, 初期状態(s0)に強制的に戻される. 報酬の原因となる「状態 s2 を過去 10 step 以内に通ったか」という情報はエージェントに記憶されないため, エージェントにおいてはマルコフ性の無い課題になる. 原因状態 s2 の後に報酬状態 s7 を通る事で報酬が得られるこの課題の最適な行動は, 初期状態(s0)から常に右に進み続ける事である. 即ち, 実際に報酬が得られる s7 に執着した左回りや, どちらの方向でも良いからとにかく 10 個の状態を一周するという選択では正しい学習ができていない事になる.

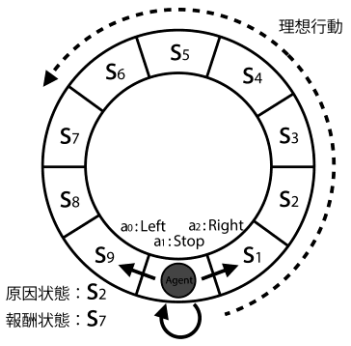

図 2 提案モデルを評価するためのタスク

5. シミュレーション結果

ランダム状態遷移, Sarsa(0), Sarsa(λ), QTimer の獲得した報酬の総和と 1 ステップあたりの報酬量, Sarsa(λ)と QTimer の Q 値更新頻度・更新量をプロットしたものを図3に示す. ここでいずれのモデルも学習率 $\alpha = 0.05$, 割引率 $\gamma = 0.9$ とし,

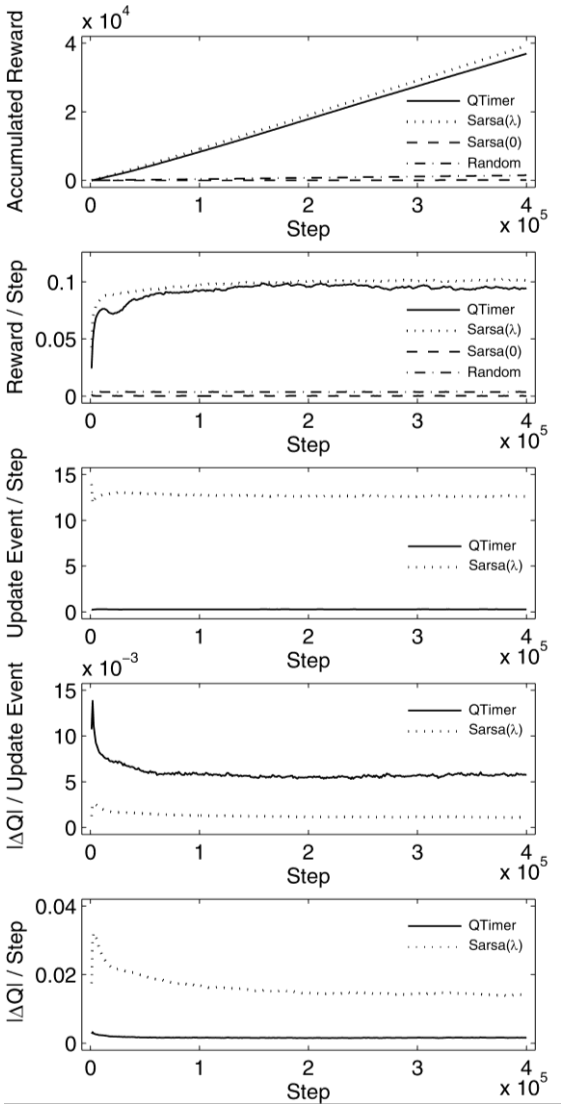

図 3 報酬の総和の発展曲線, 1 ステップあたりの報酬, 1 ステップあたりの Q 値更新回数. 1 回の更新あたりの Q 値の変化量の絶対値, 1 ステップあたりの Q 値の変化量の絶対値. すべて 10,000 個体の平均値を示す. 原理的に Sarsa(λ)の Q 値の更新回数は適格度トレースの値$e(s_i, a_j)$が非 0 である状態行動対の数だけ行われるが(本シミュレーションでは最大 30 回), それでは原理的に常に過去訪問した全ての状態行動対に対して, ほとんど影響はなくとも更新が行われてしまう. 本研究では有効な更新を対象とするため, 適格度トレース値 $e(s_i, a_j)$ が 0.001 を上回っている場合に限って 1 回とカウントした.

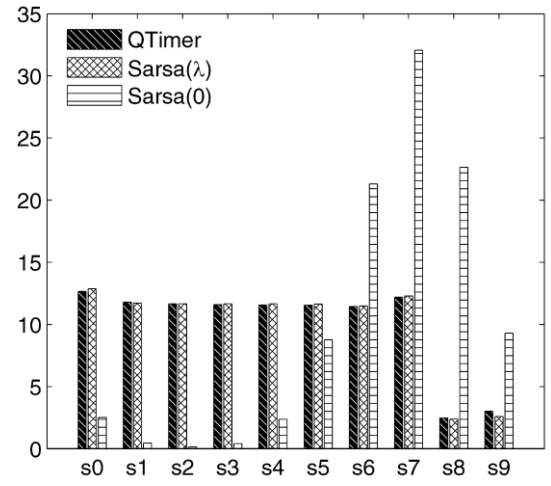

図 4　各状態への訪問率．QTimer と SARSA(0.9)は，報酬の得られる s7 の左側に寄っており，s8, s9 へは訪問しないが，SARSA(0)は s7 の周囲に左右対称に訪問率が上がっている．

QTimer の長さ $T_{max} = 30$ とした．また，比較対象である Sarsa(λ)の適格度トレース減衰パラメータ $\lambda = 0.9$ とした．QTimer モデルは，学習の初期に遅れがあるものの，概ね Sarsa(λ)に近い性能を持っている事がわかる．1 ステップあたりの Q 値の更新量は，QTimer モデルが半オフライン的な学習を行うため，毎ステップ更新する Sarsa(λ)と比較して少なく抑えられている．

QTimer model における各状態への訪問率(%)を図4に示す．理想的に学習したエージェントは初期状態 s0 から s7 へ迷い無く移動する．つまり正常に学習されていれば，報酬獲得に関係ない状態 s8, s9 には訪問せず，それ以外には均等に訪問する．行動選択は右方向へ顕著偏る事が予測される．各状態における行動選択率を図5に示す．

Sarsa(0)は，s7 に多く訪問しているが s7 に対して対称的な訪問比率を有するため，左右ループには差がない．よって学習は上手くできていない．行動選択も左右まばらである．

Sarsa(λ)及び QTimer は共に，初期状態 s0 に最も多く訪問し，次いで報酬状態 s7 に訪問している．それ以外の s0 と s7 の間の状態は，ほぼ同じ訪問率を持ち，報酬獲得に必要ない s8 と s9 の訪問率が低い．また Sarsa(λ)及び QTimer は共に，s0 から s6 までの状態においての行動は右向きが多く選択されている．s8 と s9 において Sarsa(λ)は左向きの選択率が高く，この s7 へと戻る行動選択が結果として QTimer と比較してやや高い総報酬につながっているものと考えられる．

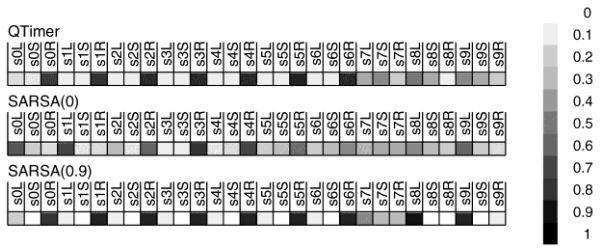

図 5　各状態(s0～s9)における行動選択率(L=左，S=停止，R=右)．各セルの濃さは選択率(0～1)を示す．

6. 考察

6.1 マルコフ性と適格度トレース

一般的に，強化学習モデルにおいては離散マルコフ決定過程が仮定される．すなわち，環境の次状態は前状態と離散的に選ばれた行動によって決定され，報酬は状態に依存して与えられる．報酬の遅延を想定する場合，報酬は状態毎に与えられず，特定の状態に至って初めて与えられるため，遅延報酬を過去の状態－行動対に反映させるための仕組みが必要となる．

ここで，マルコフ性を仮定することの意義を検討する．強化学習モデルには複数の仕組み・前提が絡まりあっており，議論を簡潔にする目的で，まず報酬に遅延が無い場合を考える．このとき，マルコフ性を仮定する意義は，報酬の原因となる前状態の推測が一意に決定できる，という利点にある．動物の置かれた実環境においては n ステップ前のどの状態群（単一の表象が不可能な場合も含む[Ohta 12][太田 13]）に依存しているか否か本質的には不明であり，原因となる状態の推定を行う必要がある．これに対してマルコフ性を仮定した場合，その必要性がなくなる．これがマルコフ性を仮定する利点である．ところが報酬に遅延がある場合にはこの利点が解消されてしまう．

報酬に遅延がある場合について検討する．報酬に関するマルコフ性を仮定すると，報酬は現在の状態にのみ依存しているため，報酬の遅延は n ステップの状態遷移の結果として表現される．一般的な強化学習タスクにおいて，各状態の定義は過去のすべての状態－行動対から成るベキ集合とはなっておらず，複数の経路・行動から現在の状態に至ることが可能であるため，状態―行動対に関する一意性は成立していない．そのため，報酬及び一階の状態－行動対に関するマルコフ性の仮定のみでは，報酬の原因となった n ステップ前の状態－行動対を特定することができない．適格度トレース(eligibility trace)[Klopf 72]は，この問題を疑似的に解決するロバストな仕組みであると言える．状態遷移そのものとは別に，通過した状態―行動対にフラグを付け，連続的にその値を減じることで時間的な価値を表現し，報酬の時点から遡って ad hoc に報酬の原因の推定を行うことができる．

つまり，マルコフ性の仮定によって原因となる状態の推定の必要性は無いものの，遅延報酬を前提とした場合，別途原因となる状態の推定の仕組みとして適格度トレースが必要とされている．

また一般に Q 値の差分と報酬(TD 誤差)を Q 値更新に利用する手法が採られているが，提案モデルは，TD 誤差を用いず，適格度トレースのみに依存して Q 値を更新する．

シミュレーション結果は，Sarsa(λ)の適格度トレースの持つ非マルコフ的課題に関する学習能力の高さを示すと同時に，提案モデルが Sarsa(λ)と同程度の能力を持っていることを示している．

6.2 Q 値更新に必要な計算プロセスとパラメータ

提案モデルは QTimer が終了したタイミングでのみ Q 値を更新するという特徴を持つ．TD 誤差を用いた強化学習モデルにおいては，ステップ毎に前後の Q 値を参照して TD 誤差を算出し Q 値を更新するが，その差分の計算のために，前ステップの Q 値と次ステップの Q 値を同期させる必要がある．大脳基底核においてそのような同期機構は，未だ見つかっていない．しかし，個々の線条体ニューロンにおいて Q 値が保持されており[Samejima 05]，かつ本モデルのように個々のニューロンに Q 値を更新するための持続的なプロセスを仮定するならば，Q 値を外部から同期的に参照することなく報酬のみから個別に更新が可能である．なお，グルタミン酸性シナプス入力をきっかけにし

たカルシウム濃度上昇, cAMP, DRAPP-32, STEP, ERK のリン酸化等の細胞内プロセスによってシナプス入力とドーパミン入力の履歴が蓄積されている可能性が検討されているが[Houk 95][Nakano 10][Shifleett 10][Shiflett 11] , 本モデルの QTimer はそれらの持続的な細胞内シグナリングをイメージしている.

また, 一般的な強化学習モデルのように Q 値をステップ毎に更新することは, 過去に学習した内容を忘却するリスクがある. 複数の条件が揃うことで初めて線条体ニューロンの可塑的変化が誘導されることを考えれば[Reynolds 02], Q 値の更新がステップ毎に発生することは考えにくい. 提案モデルでは, 状態と行動が揃ったときに起動される時限付き細胞内シグナリングの終了時点における可塑的変化の誘導を想定しており、Q 値更新のタイミングは間欠的である. そのため, 1 ステップあたりの Q 値更新量は低く抑えられている(図3).

以上のように, 提案モデルは生理学的に見て自然なモデルとなっており, 大脳基底核の学習機能に関する理解の助けになるものと考えられる.

また, 非マルコフ的な環境に適応するためには, 大脳皮質からの複数のステップに渡る入力を並列的に処理する必要がある. 提案したモデルは, 非同期で独立した Q 値の更新機構があるため, そのような並列的な処理に適した拡張性を持っている. 例えば, 線条体ニューロンの大脳皮質に対する入力次元(receptive region/receptive field)がドーパミンによって変化するようなニューラルネットワークモデル[Schultz 95] [Nakahara 02]と本提案モデルは親和性が高いものと考えられ, 強化学習モデルに状態及び時間を空間的に扱う能力を付与する基礎となりうる.

7. まとめ

大脳基底核線条体ニューロンの持続的に発火する性質をヒントに, 状態－行動対の訪問時に起動されるタイマーに基づいた非同期な行動価値関数更新アルゴリズムを持った強化学習モデルを提案した. 提案モデルは, 非マルコフ的なタスクにおいて, Sarsa(λ)と概ね同程度の行動選択パターンと学習性能を持っていることがわかった.

謝辞

本研究は科研費(24700200), 防衛医科大学校特別研究費, 光科学技術研究振興財団, 東北大学電気通信研究所共同プロジェクトの助成を受けたものである.

参考文献

[Barto 95] A. Barto: Adaptive critics and the basal ganglia. Models of Information Processing in the Basal Ganglia, in Models of Information Processing in the Basal Ganglia, J. C. Houk, J. Davis, and D. Beiser, Eds. Cambridge, MA: MIT Press, 1995, pp. 215–232.

[Houk 95] J. C. Houk, J. L. Adams, A. Barto: A model of how the basal ganglia generate and use neural signals that predict reinforcement," in Models of Information Processing in the Basal Ganglia, J. C. Houk, J. L. Davis, and D. G. Beiser, Eds. Cambridge, MA: MIT Press, 1995, pp. 249–270.

[Ji 12] Z.-G. Ji, S. Ito, T. Honjoh, H. Ohta, T. Ishizuka, Y. Fukazawa, H. Yawo: Light-evoked somatosensory perception of transgenic rats that express channelrhodopsin-2 in dorsal root ganglion cells, PLoS One, vol. 7, no. 3, p. e32699, Mar. 2012.

[Klopf 72] A. Klopf: Brain function and adaptive systems: a heterostatic theory, AIR FORCE CAMBRIDGE Res. LAORATORIES, no. 133, 1972.

[Nakahara 02] H. Nakahara, S. Amari,O. Hikosaka: Self-organization in the basal ganglia with modulation of reinforcement signals, Neural Comput., vol. 844, pp. 819–844, 2002.

[Nakano 10] T. Nakano, T. Doi, J. Yoshimoto, K. Doya, A kinetic model of dopamine-and calcium-dependent striatal synaptic plasticity, PLoS Comput. Biol., vol. 6, no. 2, pp. 1–16, 2010.

[Ohta 12] H. Ohta, D. Uragami, Y. Nishida, J. C. Houk: Presynaptic inhibition balances the trade-off between differential sensitivity and reproducibility, Proc. of 6th Int. Conf. Soft Comput. Intell. Syst. 13th Int. Symp. Adv. Intell. Syst., pp. 1172–1175, Nov. 2012.

[太田 13] 太田宏之, 西田育弘: 神経可塑性と状態の生成, 人工知能学会全国大会(第 27 回)論文集, 2L4-OS-24d-5, 2013.

[Ohta 13] H. Ohta, S. Sakai, S. Ito, T. Ishizuka, Y. Fukazawa, M. Tandai-hiruma, S. Maruyama, H. Mushiake, H. Yawo, Y. Nishida, Spike timing- dependent retrograde plasticity of the CA3 excitability in the rat hippocampus, Neurosci. Lett. 534, pp. 182-7, 2013

[Reynolds 02] J. N. J. Reynolds, J. R. Wickens: Dopamine-dependent plasticity of corticostriatal synapses, Neural Netw., vol. 15, no. 4–6, pp. 507–21, 2002.

[Samejima 05] K. Samejima, Y. Ueda, K. Doya, M. Kimura: Representation of action-specific reward values in the striatum, Science, vol. 310, no. 5752, pp. 1337–40, Nov. 2005.

[Schultz 95] W. Schultz, R. Romo, T. Ljungberg, J. Mirenowicz, J. R. Hollerman, and A. Dickinson: Reward-related signals carried by dopamine neurons, in Models of information processing in the basal ganglia, vol. 12, J. C. Houk, J. L. Davis, and D. G. Beiser, Eds. MIT Press, 1995, pp. 233–248.

[Shiflett 10] M. Shiflett: Acquisition and performance of goal-directed instrumental actions depends on ERK signaling in distinct regions of dorsal striatum in rats, J. ..., vol. 30, no. 8, pp. 2951–2959, 2010.

[Shiflett 11] M. Shiflett, B. Balleine: Contributions of ERK signaling in the striatum to instrumental learning and performance, Behav. Brain Res., vol. 218, no. 1, pp. 240–7, Mar. 2011.

[Tomita 09] H. Tomita, E. Sugano, Y. Fukazawa, H. Isago, Y. Sugiyama, T. Hiroi, T. Ishizuka, H. Mushiake, M. Kato, M. Hirabayashi, R. Shigemoto, H. Yawo, M. Tamai: Visual properties of transgenic rats harboring the channelrhodopsin-2 gene regulated by the thy-1.2 promoter, PLoS One, vol. 4, no. 11, p. e7679, Jan. 2009.

[Watkins 92] C. Watkins, P. Dayan: Q-learning, Mach. Learn., vol. 292, pp. 279–292, 1992.