1M2-1

# 線条体の動作に触発された習慣形成の強化学習モデル

Habit Forming Reinforcement learning model that has been inspired by the behavior of the striatum

甲野 佑 *1 (Yu Kohno)　水戸 亜友美 *2 (Ayumi Mito)　太田 宏之 *3 (Hiroyuki Ohta)　高橋 達二 *2 (Tatsuji Takahashi)　笹川 隆史 *2 (Takashi Sasagawa)

*1東京電機大学大学院
Graduate School of Tokyo Denki University

*2東京電機大学理工学部
School of Science and Technology, Tokyo Denki University

*3防衛医科大学校生理学講座
National Defense Medical College, Department of Physiology

The movement is decided by small time width In reinforcement learning to get expedient actions. On the other hand, We automate most of actions As a habitual subroutine to achieve a purpose. In addition, When brain learns a custom, It is revealed that striatum behave on a long time scale in comparison with cortex, because it unifies primitive movement. In this study, We devised a reinforcement learning model to treat by habit by bundling up primitive movement by compressing and extracting chronological order, based on these knowledge.

## 1. はじめに

我々の行動の大半はあらかじめ形成された習慣的行動であり，それは細かな筋肉運動などを連続した動作系列としてまとめられ自動的に行われている．例えばドアを開ける程度であれば, 学習済みの “(ドアを開ける一連の) 動作コマンド” を実行し, 次の動作決定まで持続する．しかし, 合目的的行動の学習を担う既存の強化学習モデルでは非常に細かな時間幅での動作決定が一般的である．脳における強化学習・習慣学習には大脳基底核・線条体が関与しており，最近, 線条体は細かな動作を統合し, 大脳皮質と比べて長い時間幅で機能する事が判ってきた [Ohta 13][太田 13]．これによって細かな動作単位を束ねて習慣を形成する過程に線条体の時間的持続特性が関与している可能性が出てきた．本研究ではこのような線条体の特性に習い，持続する機構によって細かな動作単位を束ねる強化学習モデルを, 実際にアルゴリズムとして動作可能なレベルで実装・提案する．

## 2. 強化学習と脳

強化学習とは報酬を獲得するための試行錯誤的な動物の行動学習に端を発した枠組みである．強化学習課題には現在の状態における行動の評価 (行動価値関数) に基づき行動し, その行動評価を意味する TD 誤差を用いて学習する TD 学習と呼ばれる手法が一般的に用いられる [Sutton 00]．TD 学習は単なるアルゴリズムとしてだけでなく，学習中の大脳基底核の挙動 [Schultz 95] が TD 誤差に類似することなどから, 大脳基底核と強化学習のモデルの基礎ともなっている [Houk 95]．また, 線条体の一部が行動価値関数を表現しているとの報告があり, 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．強化学習に基づいた大脳基底核理論では, 皮質＝感覚・環境情報, 線条体＝それに対応した (Q 値に相当する) 行動価値を表すとされている．それに対して，線条体は入力に対する受付時間が長いという特徴が新たに発見された [Ohta 13][太田 13]．そのため線条体は皮質から入力される動作単位＋細かい感覚とドーパミン (報酬) を, その長い受付時間によって報酬獲得のスイッチとなる感覚，運動情報群としてまとめている (習慣形成) のではないかと考えられる．

連絡先: 高橋達二, 東京電機大学, 350-0394 埼玉県比企郡鳩山町石坂, 049-296-5416, tatsujit@mail.dendai.ac.jp

## 3. 習慣の学習と時系列の圧縮抽出

本研究で扱う習慣とは, 例えば部屋から出たい場合, その場からドアまで歩き, ドアを開けて出て行く, つまり部屋を出ようと思った時から, 出るまでの無意識的な一連の動作の事を指す．ラットの実験において, 習慣形成前では課題中ほとんど常に線条体ニューロンが活発に活動するのだが, 習慣形成後では行動の開始時と報酬獲得時のみの活動する．すなわち形成された習慣行動中は状態観測や意思決定に注意が向いていない事が知られている [Smith 13]．人間は (1) 新しい行動の吟味, (2) 習慣の形成, (3) 習慣の脳への刷り込み という 3 つの段階を経て習慣の学習を行うとされている.[Smith 13]．通常の強化学習は (1) のみに相当し，本研究では (2)　(3) についてのモデル化を行う．つまり習慣の獲得のためには, 一連の動作とその開始条件 (スイッチ) となる状態を報酬によって適切に関連づける必要がある．そのために報酬を獲得した (状態と行動で記述される非同期な) 一連の時系列から関係ない状態を排除して整理できなくてはならない．そこで本研究では神経生理に習った時系列から必要な要素を圧縮抽出という形式で習慣の学習を行う．

### 3.1 動作コマンドの学習

一般的に言われる習慣とは前述した形成過程以外に，前述の, “その場からドアまで歩く”, “ドアを開けて出て行く” ようなプリミティブな行動を組み合わせた動作コマンドを学習する必要がある．本研究では次節で述べる時系列の圧縮に焦点をおくため，プリミティブな行動 (特定の方向に移動等) を慣性的に行い続ける事を動作コマンドとして扱っている．

### 3.2 時系列の圧縮抽出

習慣の獲得のためには, 前述した動作コマンドの学習がある習慣の一連の動作そのものを形成の対として，動作コマンドを始める開始する原因となる状態を同定して固定化することも必要となる．強化学習では行動 $s_i$ と状態 $a_j$ の時系列で一連の動作が記述されるため，このような因果関係の同定は時系列から必要な要素だけ抽出して時系列を圧縮することに他ならない．

## 4. Q-Timer

線条体の入力に対する時間窓が長いことから, Q-Timer が考案された．Q-Timer とは状態行動対の訪問時に起動される

L step のタイマー (Q-Timer) に基づいて収集した L step 収益 ($R_L(s_t, a_t)$) を用いて Q 値を更新する非同期的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である．ワンステップの収益 (報酬 r) を用いるのではなく， L step 収益 ($R_L(s_t, a_t)$) を Q 値のバックアップに用いるという発想では適格度トレースを導入し学習アルゴリズム Sarsa(λ) が提案されているが, 後方観測という形態から TD 誤差を使用しなくてはならない．Q-Timer は TD 誤差を使わないために非同期にも Q 値を学習することができ，同期的なタスクでも Sarsa(λ) と同程度の性能を持っていることが示されている [太田 14].

## 5. 時系列圧縮抽出アルゴリズム Habit-Former 1.0

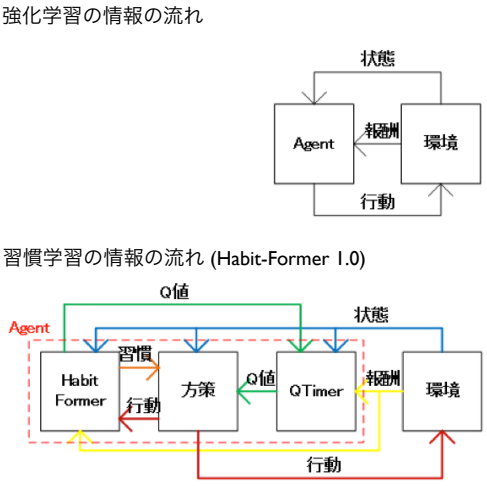


図 1: TD 学習と Habit-Former の情報の流れの違い

本研究では原因状態をスイッチとした行動系列の実行 (習慣学習) を獲得するアルゴリズムとして Habit Former 1.0 を考案した．Habit Former 1.0 は前述したように，プリミティブな行動を継続して行い続ける慣性的な行動として扱う．習慣は報酬の獲得をきっかけとして形成されていき, 習慣行動テーブルに保存されていく．非習慣行動中に習慣行動のスイッチ (開始条件) となっている状態を観測した場合, その習慣行動と一致する行動価値関数 Q 値を予測報酬としてエージェントに与える．最終的に観測された報酬を与えられる状態において，予測報酬と実際に与えられる報酬の差を各行動価値関数に分配する．このような報酬の出現は過去の生理実験の結果とも符合する [Schultz 95]．ここで行動価値関数 (Q 値) への報酬の分配は Q-Timer 1.0 [太田 14] を用いて行う．Q-Timer とは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的な行動価値関数の更新を行う強化学習モデルである．必要のない状態と行動を圧縮するため, 予測報酬を獲得した際に, その先の習慣行動と現在の行動が一致する場合, 慣性的な行動として習慣行動テーブルにスイッチとなる状態を前倒しした習慣行動を上書きしていく．ここで一致しなかった場合は新たにその中継地点となるスイッチ状態とその際に行った行動を記録する．習慣行動テーブルは推移先と行動, その状態で習慣がすでに形成されているかどうかを保持しており，形成されていれば推移先となる状態に到達するまで慣性的に行動し続ける．また，その際に仮想報酬としてその時点で観測される Q 値を Q-timer で蓄積される収益 $R_k(s_i, a_j)$ に蓄積される．習慣行動中にも訪問した状態行動対 $(s_i, a_j)$ の Q-timer はカウントは開始され，その際に通常，初期収益 $R_0(s_i, a_j) = 0$ から始まるのに対して，仮想報酬として当該の Q 値を与え，$R_0(s_i, a_j) = Q(s_i, a_j)$ から開始される．ただし，$\epsilon$-greedy 方策を用いて確率 $\epsilon$ でランダムな探索行動を併用する等，習慣行動中にそれから外れるような行動を取る事もある．習慣行動中に習慣から外れた場合，ただちにその時点での k step 収益 $R_k(s_i, a_j)$ を用いて対応する Q 値 $Q(s_i, a_j)$ を更新し，全ての Q-timer をゼロに戻す．その状態において既に習慣行動テーブルが作成されていた場合は $\epsilon$-greedy で習慣行動テーブルに記された習慣行動を行う．遅延報酬課題のような, マルコフ性の仮定できない場合には適格度トレースを導入したモデル (Sarsa(λ)) が提案されているが, Q-Timer では TD 誤差を使わずに Sarsa(λ) と同程度の性能を持っていることが示されている．

## 6. 正報酬の崖歩きシミュレーション

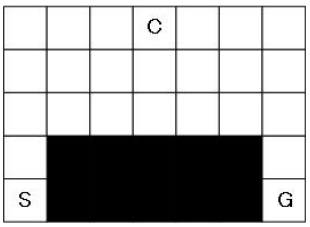


図 2: 条件付き崖歩き

崖歩き課題と条件付き崖歩き課題でシミュレーションを行った．本シミュレーションにおいてエージェントが取りうる行動は, $a_N$:上へ行く, $a_W$:左へ行く, $a_S$:下へ行く, $a_E$:右へ行くの 4 種類である．図 2 は問題環境を表しており，縦 5 マス, 横 7 マスから (黒く塗りつぶした) 崖領域の 10 マスを除いた 25 状態が存在する．初期状態 (S) から始まり, 崖歩き課題ではゴール (G) に着いた場合, 条件付き崖歩き課題では, 報酬条件状態 (C) を通ってからゴールに着いた場合にのみ正の報酬 (100/ ゴールまでに掛かった step 数) を与える．この課題で負の報酬を扱わないのは Habit-Former が正の報酬をきっかけに習慣行動を形成しはじめるためである．条件付き崖歩き課題はどの状態が報酬の条件なのか認識できないため, 通常の TD 学習では困難な非マルコフ課題に分類される．どちらも崖領域に落ちた場合，ペナルティとしてゴールまでかかった時間を 100 増加させる．シミュレーションは全 1,000 エピソードを 1,000 回行い, 平均を取った．また習慣の形成の開始は 500 エピソード以降とした．比較のために Habit Former 1.0 に加えて, Sarsa, Sarsa($\lambda = 0.9$), Q-Timer でシミュレーションを行った．各エージェントは学習率 $\alpha = 0.05$，割引率 $\gamma = 0.9$ のパラメータで学習を行う．行動の決定には $\epsilon$-greedy 方策, ランダム行動率 $\epsilon = 0.1$ を用いた．

### 6.1 結果

評価には, 獲得報酬と, どれだけ習慣に基づいた行動を取ったかという習慣行動使用率を用いる．習慣行動使用率の図 4, 6 は黒に近い程, その後の一貫した動作 (上下左右) を引き起こすスイッチとなっている事を表しており，形成された習慣がどの状態と結びついているかを意味する．図 4 から, 曲がる部分が濃くなっており, 間の行動はほとんど無視されているので習慣の形成が行われていることがわかる．また図 5 から, 条件付き崖歩き課題のような非マルコフ環境でも学習できることが示された．

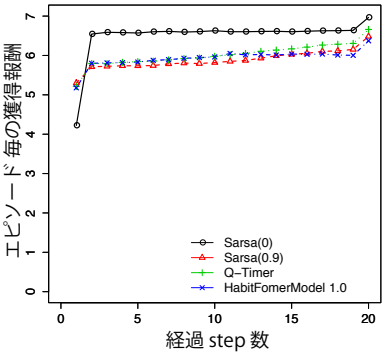


図 3: 崖歩きでの獲得報酬

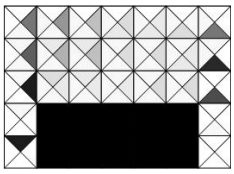

図 4: 崖歩きでの習慣行動使用率

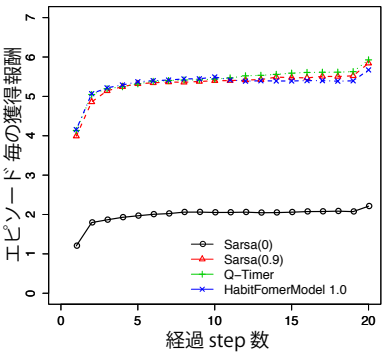


図 5: 条件付け崖歩きでの獲得報酬

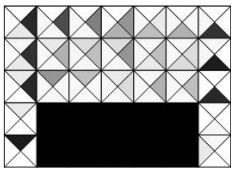

図 6: 条件付き崖歩きでの習慣行動使用率

## 7. 議論

図 4, 6 によって，本研究で考案した Habit-Former 1.0 が特定の状態と一貫した動作を結びつけられていることが確認された．これにより本研究の目的である，線条体の持続機構を前提として，プリミティブな行動を束ねて一連の動作とし，それを特定の状態と結びつけることで，ある状態信号に対して特定の動作が自動的に選択されるという，習慣学習をアルゴリズムとして実装可能なことを示すことができた．しかしながら，現時点で課題も多い．第一に一度形成された習慣が改善されないことが挙げられる．第二にどのような工学的応用が考えられるかという点である．以下ではそのに点について関連する議論を行う．

### 7.1 習慣の改善と探索との相性

Habit-Former 1.0 が習慣の改善を行えないことは，改善のための，その時点で最良な行動とは異なる行動を選択する探索の仕方が，現時点では乱数的で単純なものであるためだと考えられる．習慣の改善とは，探索によってよりよく報酬が得られる動作の発見とその習慣化によって行われると考えられる．しかしすでに習慣がある状態から，その習慣とは異なる行動をする事は難しい．さらにその既存の習慣とは食い違う動作を何度も行って固定化していかなければならない．前述した通り，本研究では習慣形成したのちも $\epsilon$-greedy 方策による確率的な探索行動が行われる．これは Q 値の学習には十分な探索が必要なためである．しかし，習慣による意思決定が Q 値に対する意思決定より優位である関係上，探索が行われたところで既存の習慣を破壊できない以上，いずれは再度，既存の習慣行動に合流してしまい，ダイナミックな探索が行えない．しかし，人間は習慣行動中においてはなんらかの切っ掛けが無い限りは，習慣行動を中断して探索行動がなされるとは限らない．強化学習においてこの切っ掛けとは外部，あるいは内部の状態が何らかの信号で変化することを意味すると考えらえる．そもそも生物は探索を行う際になんらかの傾向性を帯びていると思われる．このように Habit-Former 1.0 は生理的な背景を持つがゆえに，生物的な探索の仕方を取り入れる必要があると考えられる．

### 7.2 階層型強化学習との関係

本研究における習慣形成は複数の状態から形成される一つの課題環境を，一連の動作 (習慣) の自動的な選択という形で切り分けていると捉えることもできる．そのため，生理的知見を背景に持ちながら，階層型強化学習と関連があり，工学的にも応用が考えられる階層型強化学習ではサブゴールを定めることで環境を切り分け，サブゴールを超えると別の状態空間として認識する．このサブゴールの設定が困難であるとされるが，Habit-Former 1.0 は一連の動作の開始という “結果” を特定の状態を “原因” として扱うためサブゴール形成と関連深い性質を持つ．図 4, 6 の結果はそれを示しており，慣性的な動作が開始される原因として紐付けられている状態行動対 (灰色が濃くなっている箇所) が階層型強化学習におけるサブゴールになっていると考えらえる．部分観測マルコフ環境では，本来異なる状態を観測情報から混同して同一視してしまうことに問題がある．これに対して，習慣学習では “ある習慣 $H_k$ を行った” こと自体をメタ状態として認識することで空間の切り分けとサブゴール形成ができると考えられる．現時点ではまだ具体的なアルゴリズムの考案・実装には至っていないが，線条体の特性に由来する本研究は階層型強化学習に対する知見にもなり得ると考えられる．

## 参考文献

[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.

[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.

[Houk 95] Houk, J.C., Adams, J.L., Barto, A.G.: A model of how the basal ganglia generate and use neural signals that predict reinforcement, In Models of Information Processing in the Basal Ganglia, Houk, J.C., Davis, J.L., Beiser, D.G., Eds. MIT Press, pp. 215–232. (1995).

[太田 14] 太田 宏之, 甲野 佑, 高橋 達二: 線条体ニューロンの持続的発火と強化学習, JSAI 2014(2014 年度人工知能学会全国大会 (第 28 回)) 予稿集, 2N5-OS-03b-4. (2014).

[Schultz 95] Schultz, W., Romo, R., Ljungberg, T., Mirenowiczm, J., Hollerman, J.R., Dickinson, A.: Reward-related signals carried by dopamine neurons, In Models of Information Processing in the Basal Ganglia, Houk, J.C., Davis, J.L., Beiser, D.G., Eds. MIT Press, pp. 233–248. (1995).

[Smith 13] Smith, K.S., Graybeil, A.M.: A Dual Operator View of Habitual Behavior Reflecting Cortical and Striatal Dynamics, Neuron, Vol.79, No.2, pp. 361–374. (2013).

[Sutton 00] Sutton, R.S., Barto, A.G.: 強化学習, 森北出版, (三上, 皆川 訳) (2000).