マルチエージェント強化学習における割引率の社会適応的調節
—生物に見られる社会的階層構造の自己組織化と神経修飾物質—

Social adaptive modulation of temporal discount factor in Multi-agent reinforcement learning

-Self organizing social hierarchy and neuromodulator -

○ 矢野 史朗（東京大学） 青沼 仁志（北海道大学）
淺間 一 （東京大学）

Shiro YANO, The University of Tokyo, yano@robot.t.u-tokyo.ac.jp
Hitoshi AONUMA, Hokkaido University
Hajime ASAMA, The University of Tokyo

In the field of reinforcement learning, researchers have focused a role and dynamics of animal's neuromodulatory system, which is thought to correspond to meta-parameter of the learning system. Recently, the hypothesis is suggested and beginning to be verified that neuromodulator serotonin regulates the temporal discount factor. It is known that dominance hierarchy affects the amount of serotonin in animal society. It is also known that social hierarchy is generated from competitive behavior in a self-organizing manner. In this study, we propose self-organizing multi-agent reinforcement system which grows various temporal discount factors robust and fault-tolerant system. We show that this system has the capability adapting to environmental variability and robustness for vacancy or increasing of agents.

***Key Words*:** Reinforcement learning, Meta-parameter, Social interaction, Adaptation

1. 緒言

強化学習法は，環境から得る報酬の累積を最大化するように，試行錯誤的に方策を探索する学習法である [1]．累積報酬は式 (1) で表される．

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \qquad (1)$$

r_kはある時刻 k に得た報酬，γは割引率と呼ばれ，どれだけ先の報酬まで予測して評価に組み入れるかを決める減衰定数である．割引率は，以下で述べるように強化学習のメタパラメータの 1 つである．

強化学習の性能は，メタパラメータと呼ばれる学習率や割引率の設定により影響を受けることが知られている．計算論的神経科学の分野では，これらのメタパラメータが脳内の神経修飾物質に相当するという仮説がたてられ，現在その検証が進められている [2, 3]. 通常, 計算機を用いた強化学習では, これらメタパラメータは設計者の経験によって特定の定数に固定されるが，強化学習を行っていると考えられている生物の脳においては，このメタパラメータに相当する神経修飾物質は環境適応的に調整されている [4]. このメタパラメータの適応的調整を強化学習のメタ学習と呼び，アルゴリズムの構築が期待されている [5, 6].

メタ学習アルゴリズムの構築を行うにあたり，神経修飾物質がどのように環境適応的振舞いを示すかモデル化するという方法が考えられる．近年，計算論的神経科学ではメタ学習の構築にあたって，個体－環境間相互作用における神経修飾物質の適応的振舞いに注目しているが[9]，メタパラメータに対応する神経修飾物質は，個体－個体のような社会的相互作用によっても影響を受けることが知られている [7]. 現在マルチエージェント強化学習や群強化学習と呼ばれる分野では，複数のエージェントが報酬信号や行動方策に関する情報を交換しあいながら，行動方策を獲得するアルゴリズムの開発が盛んに行われているが[8]，上述したような，エージェントが社会的な相互作用を通してメタパラメータを調整するようなアルゴリズムに関しては，殆ど行われていない．

本研究では社会的相互作用と神経修飾物質の関係に関する生物学的知見に基づき，マルチエージェントメタ学習を構築する．また本稿では，本手法がエージェント数の増減に対して高いロバスト性・拡張性をもつことを示す．

より具体的には，

(i) 競争的な相互作用によって，社会に所属している個体ごとに異なる割引率が割り当てられる現象を用い（2.1～2.3 節）

(ii) 複数の割引率の中から信頼度の高い割引率を選び出す手法と組みあわせる（2.4 節）．

2.4 節に引用する研究はシングルプロセッサを前提としたアルゴリズムである．本研究ではこれをマルチプロセッサ・マルチエージェント用途へ拡張し，プロセッサ毎に異なる割引率を割り当て，信頼度の高いエージェントを選び出すアルゴリズムに変形する．本研究の特色として，上記(i)の自己組織化現象を利用していることで，各プロセッサに割引率を割り当てる際，初期値に依存しない設定が可能となっている．これによりプロセッサの交換や追加などに際して高いロバスト性・拡張性を持つシステムが構築可能である．このことに関しては，5 章で改めて検討する．

2. 先行研究

2.1 強化学習メタパラメータと神経修飾物質の仮説

強化学習メタパラメータには，TD 誤差δ，割引率γ，逆温度定数β，学習速度係数αがある．これらと脳機能，特に神経修飾物質を関係づける仮説は Schultz らによる大脳基底核におけるドーパミン系の研究が発端となり [5]，Doya によって提唱された [2]．本研究においてもこの Doya らの仮説に従い，マルチエージェントメタ学習のアルゴリズムを構築する．

上記仮説によると，哺乳類脳内では，TD 誤差δは大脳基底

核ドーパミン系，式(1)に現れる割引率γは縫線核セロトニン系，行動選択のランダムさを表す逆温度定数βは青斑核ノルアドレナリン系，学習速度係数αは前脳基底部アセチルコリン系に相当する [2]. 社会的相互作用との関係が良く研究されてきた神経修飾物質としては，セロトニン系が挙げられる[10, 11]．以下では，セロトニン系の社会適応的調整の生物学的知見を参考に，割引率のメタ学習を構築する．

2.2 セロトニン系と社会的順位

縫線核セロトニン系は，うつ病との関係[12]や強化学習の割引率に相当するという仮説などにより注目されてきた[11].

セロトニン系は様々な動物で，社会的な競争による敗北に対して反応することが知られている．しかしこの反応の仕方は脊椎動物，無脊椎動物などによって異なり，敗北によりセロトニン系が活性化される動物や[13]，抑制される動物など[14]，動物によって正反対の影響を受けることが知られる．

2.3 競争と社会的順位の自己組織化

集団の個体が相互に競争し合う場合にどのような現象が発現するかをモデル化した研究がある[15]．このモデルでは，個体ごとに 1 つずつ内部状態を持っており，i 番目の個体は h_i を持つ．このとき，モデルは以下のようになる．

個体 i と j は各時刻において確率 $P_{ij}(h_i, h_j)$で接触し，もし接触すれば確率 $Q_{ij}(h_i, h_j)$で勝敗を決定する．

$$Q_{ij}(h_i, h_j) = \frac{1}{1+\exp(-\eta_2(h_i - h_j))} \tag{2}$$

$$\frac{dh_i}{dt} = -\mu h_i + \frac{\rho}{N}\sum_j \left(\Delta_{dom} Q_{ij}(h_i, h_j) + \Delta_{sub}(1 - Q_{ij}(h_i, h_j))\right) \tag{3}$$

となり，これは以下の 3 つ

(i)　個体同士が遭遇し競争を開始する

(ii) 勝敗が決した後，各個体の内部状態に報酬/罰が加わる

(iii) 発生した内部状態の変化が，時間とともに忘却される

という一連の流れを数理的にモデル化している．この研究により，ある閾値以上の頻度で社会的競争を行うと，社会的順位が自己組織化することが示されている(Figure 1)[16].

以上のモデル説明を振り返ると，2.2 節で説明したセロトニン量の動態は，モデルの内部状態の動態と類似していることが分かる．そこで本研究では，セロトニンが本モデルの内部状態として振舞うと考える．

2.4 信頼度に基づく割引率の調整

2.3 節で述べたように，本研究では社会的相互作用によって，セロトニン量が Bonabeau モデルに従うと考えている．そのため，ある閾値以上の頻度で社会的競争を行った場合，Fig.1 に示したように，セロトニン量の個体差が自己組織化すると分かる．このセロトニンは 2.1 節で述べたように，強化学習の割引率というメタパラメータに対応するので，競争という社会的相互作用によって，様々な割引率が発生することになる．

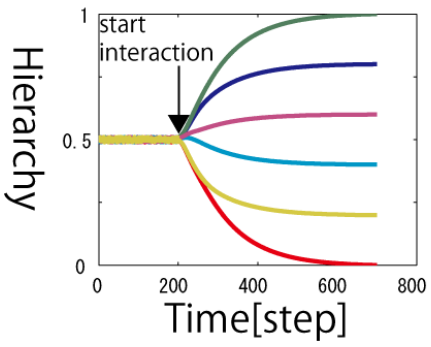

Fig.1　社会的順位の自己組織化（数値計算）
200 [step] から十分な頻度で相互作用を開始した．
各線は個体の社会的順位を表しており，200 [step] 以降，順位が自己組織化していることが分かる．ここでは，勝利/敗北時の内部状態変化量の絶対値を等しくした設定で計算を行った．縦軸の Hierarchy は内部状態を[0 1]に規格化した量．

尾川らは，強化学習のメタ学習の問題で，割引率の調整を扱った[17]. 学習の初期段階では内部モデルの構築が不十分であるため，将来を予測することが困難である．このアルゴリズムでは，学習進度に伴って割引率を調整する．

また，尾川らのアルゴリズムは，1 つのプロセッサの中で割引率を計算する形式となっている．

尾川らは，内部モデルをどれだけ信用できるか，つまり予測精度が高いか低いかを計量するために，信頼度 $\lambda_t(s_t)$という尺度を導入している．またその信頼度 $\lambda_t(s_t)$の更新には TD 誤差 δ_t を用いて，

$$\frac{1}{\lambda_{t+1}(s_t)} = \frac{1}{\lambda_t(s_t)} + \alpha_R {}^R\delta_t \tag{5}$$

$$\text{where } {}^R\delta_t = \delta_t^2 + \gamma_R \frac{1}{\lambda_t(s_{t+1})} - \frac{1}{\lambda_t(s_t)}$$

としている．ここで $\lambda_t(s_t)$は状態 s_t ごとに信頼度を定義する関数である．また，この信頼度を用いて割引率 γ_t の調整則を

$$\gamma_t = \gamma_0 \min(1, \sqrt{\lambda_t}) \tag{6}$$

としている．

3. 方法

尾川らのアルゴリズムはシングルプロセッサ内で，連続な割引率空間の中から，一点を選び出す形式となっている．これをマルチプロセッサのプロセッサ毎に割引率を割り当て，適切な信頼度を持つプロセッサを選び出すアルゴリズムにする場合，離散的な割引率空間の中から一点を選び出す形式となる．すなわち，信頼度の最も高い割引率のプロセッサを選ぶには，j 番目($j \in \mathbb{N}$)のプロセッサの割引率を γ_j とするとき，式(6)の γ_t を用いて，

$$\gamma = \text{argmin}(\gamma_t - \gamma_j) \tag{7}$$

とすることで拡張される．また γ_j は

$$\gamma_j = \tau + \frac{1-\tau}{2}\left(\frac{h_j}{\max(|\mathbf{h}|)} + 1\right) \quad (8)$$
$$\text{where } \forall h_j \in \mathbf{h}$$

とする．これにより， $\tau \le \forall\gamma_j \le 1$ となる．

3.1 実験プロトコル

尾川らと同様，本アルゴリズムを用いて Windy gridworld 問題を解く．以下の条件は文献[17]と等しい．

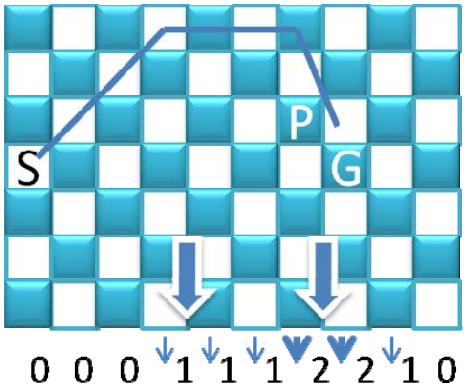

Fig.2　Windy gridworld 問題
文献[17]より改変して引用．線路は，最短経路の 1 つ．
S: スタート位置, G: ゴール位置, P: 信頼度の測定位置

Gridworld 問題では，エージェントはスタート位置 S からゴール位置 G まで移動することを学習する．GridWorld には Fig.2 の矢印にあるように下向きに風がふいており，テーブルの下にある数字分だけ移動させられる．報酬はゴール時に 1，その他のセルで 0，テーブルの外に出ると −1 とした．

強化学習のアルゴリズムは Actor-Critic 法を採用し，行動選択には Softmax 法を用いた．

$$\Delta V(s_t) = \alpha_V \delta_t,$$
$$\Delta p(s_t, a_t) = \alpha_p \delta_t,$$
$$\text{where } \delta_t = r_{t+1} + \gamma_t V(s_{t+1}) - V(s_t)$$
$$P(a) = \frac{\exp(\beta_p(s_t, a))}{\sum_{a'} \exp(\beta_p(s_t, a'))} \quad (9)$$

$\alpha_P = \alpha_V = \alpha_R = 0.1$，$\gamma_R = 0$，信頼度を測定する位置は Fig.2 の位置 P(7,4)とし，S(1,4), G(8,4)，開始時の個体数は 4，$\tau = 0.7$，全 3000 試行を行う．1 試行あたりの上限は 80 ステップまでとし，上回った場合は S 点に移動する．

特にプロセッサの故障や追加に対するロバスト性・拡張性を調べる為，1500 ステップ目と 6000 ステップ目でプロセッサを 1 つ不能にし，3500 ステップ目と 5500 ステップ目でプロセッサを追加する．

2.2 で述べたように，敗北時のセロトニン増減量は様々である．今回は勝利時と敗北時の内部状態の変化を $\Delta_{dom} = \Delta_{sub} = 1$ のように対称とした．

4.　結果

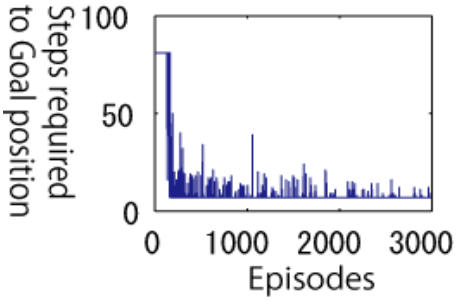

Fig. 3　ゴールまでに必要とする歩数の遷移
初期はゴールにたどり着けていない．後半はほぼ最短距離でゴールに辿り着いている．

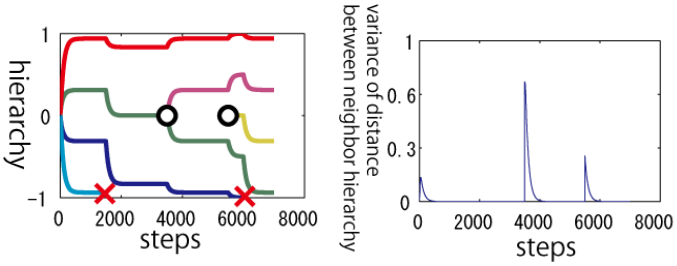

Fig. 4 左：途中でエージェントの故障・追加をした場合の順位変化
個体の故障・追加が発生すると，故障した場合は空いた空間を他の個体が埋め（例：1 つ目の赤クロス），追加された場合は，間隔を詰めていることが分かる（例：1 つ目の白丸）．
Fig.4 右：隣接した順位の個体間における順位差の分散
常に，隣接している順位の個体は，その隣接距離を一定に保とうとしていることが分かる．

5.　検討

Fig. 3 の結果は，尾川らのアルゴリズム［17］と比較して，本アルゴリズムがほぼ同性能であることを示している．

また Fig. 4 の結果は，プロセッサが故障し新しいプロセッサを補充しようとした際，どの割引率を担当するプロセッサが故障したか把握せずとも，自動的にその割引率の穴埋めを行ってくれることを示している．あるいは，新規にプロセッサを追加した際に，勝手に割引率を均等に調整することを示している．

本研究は Bonabeau モデルの簡易版である平均場近似を用いた計算法[16]を用いて，計算を行っている．平均場近似を用いた手法は，他の個体全部の内部状態の情報を常に取得できる状況を前提としている．しかしながらプロセッサ群が十分高頻度で相互作用しあい，また相互作用のネットワークが全結合となる状況であれば，平均場近似と同様の性能が発揮されると予想される．

通常の強化学習に加えメタ学習を追加することは，計算コストの問題があると考えられている[18]．計算を分散化するこ

とは，実環境下で活動する場合に有用となることが期待できる．

マルチエージェント強化学習の分野ではこれまで，報酬の共有・競争や[8]，行動則の共有など[19]の研究が行われてきており，今後，マルチエージェント強化学習におけるメタ学習も研究されると考えられる．多くの生物は社会相互作用を通して神経修飾物質を調整していることが知られている．今後の研究の発展にとって，生物社会模倣型のメタ学習という方向性が有効であろう．

6. 結言

本研究では，計算論的神経科学における強化学習の仮説に従い，特にこれまであまり行われて来なかった生物社会に対する修飾物質の応答に着目することで，新しいメタ学習のアルゴリズムを提案した．具体的には，セロトニン神経系と競争的社会相互作用に着目し，セロトニン系が割引率に対応することから，割引率のマルチエージェントメタ学習を提案した．

文　献

[1] Sutton, R. S. and Barto, A. G., "Reinforcement learning An Introduction," *MIT Press*, 1998.

[2] Doya, K., “Metalearning and neuromodulation,” *Neural Networks*, vol.15, pp.495-506, 2002.

[3] Daw, ND. and Doya, K., “The computational neurobiology of learning and reward,” *Curr. Opin. Neurobiol.*, vol. 16-2, pp. 199-204, 2006

[4] Luksys, G., Gerstner, W. & Sandi, C. “Stress, genotype and norepinephrine in the prediction of mouse behaviour using reinforcement learning,” *Nat. Neurosci*. vol. 12, pp. 1180-1186, 2009

[5] Schweighofer, N. and Doya, K., “Meta-learning in Reinforcement Learning,” *Neural Networks*, vol. 16, pp.5-9, 2003.

[6] Kobayashi, K., Mizoue, H., Kuremoto, T. and Obayashi, M., “A Meta-learning Method Based on Temporal Difference Error,” *Lect. Notes Comput. Sc.*,vol. 5863, pp. 530-537, 2009.

[7] Kravitz, E.A. and Huber, R.C., “Aggression in invertebrates,” *Curr. Opin. Neurobiol.* vol. 13, pp. 736–743, 2003.

[8] L. Busoniu, R. Babuska, B. De Schutter, “A Comprehensive Survey of Multi-Agent Reinforcement Learning,” *IEEE T. SYST. MAN. CY. C*, vol. 38-2, pp. 156–172, 2008.

[9] Doya, K., “Reinforcement learning: Computational theory and biological mechanisms,” *HFSP J.*, vol. 1(1), pp. 30-40, 2007.

[10] Dayan, P. and Huys, QJ. “Serotonin in affective control,” *Annu. Rev. Neurosci* , vol..32, pp. 95–126, 2009.

[11] Tanaka, SC., Doya, K., Okada, G., Ueda, K., Okamoto, Y., and Yamawaki, S., “Prediction of Immediate and Future Rewards Differentially Recruits Cortico-Basal Ganglia Loops,” *Nat. Neurosci.*, vol. 7 (8), pp. 887-893, 2004.

[12] Müller, C., and Jacobs, B. (eds), “Handbook of the Behavioral Neurobiology of Serotonin,” *Amsterdam: Academic Press/Elsevier.*, pp. 153-162, 2009.

[13] Blanchard, RJ, McKittrick, CR., Blanchard DC., “Animal models of social stress: effects on behavior and brain neurochemical systems,” *Physiol. Behav.*, vol. 73, pp. 261–271, 2001.

[14] Edwards, DH, Kravitz, EA. “Serotonin, social status and aggression,” *Curr Opin. Neurobiol.* vol. 7, pp. 812–819, 1997.

[15] Bonabeau, E., Theraulaz, G., and Deneubourg, J.-L., “Mathematical models of selforganizing hierarchies in animal societies,” *Bull. Math. Biol.* vol. 58, pp. 661–719, 1996.

[16] Lacasa, L. and Luque, B. “Statistical Mechanics and its Applications,” *Physica A* , vol. 366, pp. 472-484, 2006.

[17] Ogawa, N., Namiki, A., and Ishikawa, M., “Adjustment of discount rate using index for progress of learning,” *IEICE Technical Report*, vol. 129, pp.73-78, 2002.(in Japanese)

[18] Elfwing, S., “Embodied Evolution of Learning Ability,” *PhD thesis,KTH School of Computer Science and Communication,* pp. 22-23 , 2007.

[19] Iima, H, and Kuroe, Y., “Reinforcement Learning through Interaction among Multiple Agents,” *SICE-ICASE, 2006. International Joint Conference*, pp. 2457-2462, 2006