The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

# 大局基準値共有による社会的強化学習

Social reinforcement learning with shared global aspiration for satisficing

其田憲明 *1　神谷匠 *2　高橋達二 *1
Noriaki Sonota　Takumi Kamiya　Tatsuji Takahashi

*1東京電機大学理工学部
School of Science and Engineering, Tokyo Denki University

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

When humans learn, it is not just by individual trial-and-error, but the learning is accelerated by sharing information with others. There are social learning strategies such as imitating others' actions and emulating the high achievement of someone. As a model of social learning, sharing of state- and/or action-values are often implemented in reinforcement learning algorithms. However, sharing information of such huge amount is not realistic for a model of social learning of humans or animals. We propose an algorithm in which a mere "record" (achieved accumulated reward per episode) leads to efficient social learning. The algorithm is based on the model of satisficing integrated with different risk attitudes around the reference (aspiration level), and the conversion of the global aspiration onto each state.

## 1. はじめに

機械学習の分野の一つである強化学習では, 学習を行うエージェントが環境との相互作用によって得られた経験から行動価値を更新することで, 収益を最大化する最適な行動系列を学習することを目的とする.

一方で人間の学習は, ある目的水準を満たすことを目的とした場合に満足化原理 [Simon 56] と呼ばれる意思決定における損失回避の傾向がある. 満足化原理とは現状の収益が基準を満たさない場合には探索を行い, 基準を満たす行動を発見した場合にはその行動を選び続けることである. 満足化原理により, 人間は効率の良い探索を行うことができると考えられている.

この満足化原理を強化学習に応用したのが Risk-sensitive Satisficing (RS) である [高橋 16]. RS は最適な基準値を与えることで素早く最適な行動を学習し, 後悔の値を有限に抑えると証明されている [Tamatsukuri 18].

また, *"keeping up with the Joneses"*という慣用句が存在するように, 人間には自身を他者と社会的比較を行うことによって満足化の参照点が推移することが知られており, 今日ではインターネットの興隆などにより社会的比較の対象ははるか広範囲に達している [Manktelow 15].

強化学習における他者との情報共有は群強化学習 [飯間 06] のように行動価値に関連したものが多い. しかし行動価値の共有には状態行動対で情報を共有する必要があるため計算量が多いこと, 共有される情報次第では共有されたエージェントの探索傾向に偏りが生じることで準最適解に陥る可能性があることが考えられる. このような問題に対して満足化による強化学習を複数のエージェントで行い, 状態ごとに他者のより良い成績を自身の基準値として共有しつつ学習を行う手法が有効であることが示されている [其田 18]. しかし, 各状態ごとに行動価値を基準値として共有していたが, 現実には各状態ごとに基準となる成績を知ることは容易ではない. 一方で, 100m 走のタイムのような大局的な成績を知ることはあり, 大局的な情報であっても人間はより効果的に活用することができる [柄谷 85].

連絡先: 高橋達二, 東京電機大学理工学部, 350-0394 埼玉県比企郡鳩山町大字石坂, 049-296-1642, tatsujit@mail.dendai.ac.jp

本論文では大局的な成績から基準値の共有を行う社会的学習を検証し, その有効性を示すことを目的とする.

## 2. 強化学習と RS 価値関数

### 2.1 強化学習

強化学習とは学習を行うエージェントが環境との相互作用によって, 得られる報酬を最大化する行動系列の獲得を目標とする機械学習の分野の一つである. エージェントの行動決定手法を方策と呼び, 行動価値の推定手法を価値関数と呼ぶ. 強化学習の代表的な価値推定手法である Q-learning では, 時間 $t$ における状態を $s_t$, エージェントの方策に基づいて得られる行動を $a_t$ としたとき, 行動に対する環境からの作用として報酬 $r_t$, 次状態 $s_{t+1}$ を観測する. 行動価値 $Q(s_t, a_t)$ は学習率 $\alpha$, 割引率 $\gamma$ を用いることで 式 1 によって更新される.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big(r_t + \gamma \max Q(s_{t+1}) - Q(s_t, a_t)\Big) \tag{1}$$

エージェントは行動価値 $Q$ を利用して行動を決定する.

### 2.2 Risk-sensitive Satisficing

Risk-sensitive Satisficing (RS) は, 状態行動対 $(s_t, a_t)$ に対する試行量 $\tau(s_t, a_t)$ と行動価値 $Q(s_t, a_t)$, そして状態 $s_t$ に対して保持される満足化基準値 $\aleph(s_t)$ から, 式 2 によって RS 価値関数が定義される.

$$RS(s_t, a_t) = \tau(s_t, a_t)(Q(s_t, a_t) - \aleph(s_t)) \tag{2}$$

RS 方策は RS 価値関数を最大化する行動 $a_t$ を選択する方策である. また, 試行量 $\tau(s_i, a_i)$ は $\tau_{\mathrm{curr}}(s_t, a_t)$ と $\tau_{\mathrm{port}}(s_t, a_t)$ を用いて式 3 によって定義される. そして, $\tau_{\mathrm{curr}}(s_t, a_t)$ と $\tau_{\mathrm{port}}(s_t, a_t)$ は試行量割引率 $\gamma_\tau$, 試行量学習率 $\alpha_\tau$ を用いて式 4 と式 5 によって更新される.

$$\tau(s_t, a_t) = \tau_{\mathrm{curr}}(s_t, a_t) + \tau_{\mathrm{post}}(s_t, a_t) \tag{3}$$

$$\tau_{\mathrm{curr}}(s_t, a_t) \leftarrow \tau_{\mathrm{curr}}(s_t, a_t) + 1 \tag{4}$$

$$\tau_{\mathrm{post}}(s_t, a_t) \leftarrow \tau_{\mathrm{post}}(s_t, a_t) + \alpha_\tau\Big(\gamma_\tau \tau(s_{t+1}, a_{t+1}) - \tau_{\mathrm{post}}(s_t, a_t)\Big) \tag{5}$$

基準値 $\aleph$ に加え, 試行量 $\tau$ を用いることによって, 基準を満たしていない非満足状態においては楽観的探索を, 基準を満たしている満足状態においては悲観的活用を行う.

### 2.3 Global Reference Conversion

強化学習に拡張された RS は各状態に基準値 $\aleph(s_i)$ を持ち, 各状態の行動価値 $Q(s_i)$ に対して適切な基準値 $\aleph(s_i)$ を与えることで, 適切に学習できることが示されている [牛田 17]. しかし, エージェントはタスク全体としての大局的な基準値を知ることが出来たとしても, 全体目標を達成するための局所的な基準値は不明であることが多い. よって Global Reference Conversion (GRC) を用いることで, タスク全体の大局基準値 $\aleph_G$ から, 式 6 によって局所的な基準値 $\aleph(s_i)$ 変換を行う.

$$\begin{aligned} \delta_G &= \min(E_G - \aleph_G, 0) \\ \max_a Q(s_i, a) - \aleph(s_i) &= \zeta(s_i)\delta_G \\ \aleph(s_i) &= \max_a Q(s_i, a) - \zeta(s_i)\delta_G \end{aligned} \quad (6)$$

式中の $\zeta$ はスケーリングパラメータである. $E_G$ は大局観測期待値と呼ばれるものであり, エージェントが一定期間内に環境から得られた累計報酬 $E_{\text{tmp}}$ と $N_G$ を用いて式 7 で更新される.

$$E_G \leftarrow \frac{E_{\text{tmp}} + \gamma_G(N_G E_G)}{1 + \gamma_G N_G} \quad (7)$$

$$N_G \leftarrow 1 + \gamma_G N_G \quad (8)$$

パラメータ $\gamma_G$ は大局割引率を表し, $0.0 \leq \gamma_G \leq 1.0$ の範囲で定められる.

## 3. 満足化基準値共有による社会的学習

[其田 18] では同一設定のタスクを複数用意し, エージェントを 1 体ずつ配置して並列的に学習した. エージェント N 体からなるグループの n 番目のエージェントの状態 $s_i$ における最大行動価値を $Q_n^{best}(s_i)$ とした時, 式 9 によって グループ内で自律的に基準値 $\aleph(s_i)$ を更新した. その結果, 行動価値を直接共有するエージェントは準最適解に陥ったが, 基準値として共有するエージェントは準最適解に陥らずに学習することに成功した.

$$\aleph(s_i) \leftarrow \max_n Q_n^{best}(s_i), (\forall s) \quad (9)$$

しかし, この手法では状態 $s_i$ ごとに計算するため, 状態数に比例して情報共有に必要とする計算量が増加する問題が挙げられる. そこで, 本研究ではエージェント $n$ 体の観測した大局観測期待値 $E_G^n$ から式 10 のように大局基準値 $\aleph_G$ を定める.

$$\aleph_G \leftarrow \max_n E_G^n \quad (10)$$

この手法による, より少ない情報共有での学習の有用性を次の SuboptimaWorld タスクで評価した.

## 4. SuboptimaWorld

このタスクでは準最適解となるゴールが多数存在しており, エージェントは準最適解となるゴールを避けて, 最適解となる報酬を得られるゴールへの経路を学習することを目標とする.

### 4.1 シミュレーション設定

図 1 のように縦 9 マス, 横 9 マスの全 81 状態からなる格子空間上で報酬が得られる経路を学習する. 報酬が得られるゴールが 8 つ存在し, ゴールで得られる報酬は図 1 中のゴールの数字に対応してそれぞれ $1, 2, ..., 8$ と得られる. またゴールを終端状態とし, スタートからゴールにたどり着くまでを 1 エピソードとして 4000 エピソード行った.

提案手法である大局基準値共有を行うエージェント群を GRC グループとし, 比較対象として, 先行研究である 各状態の基準値 $\aleph(s_i)$ を共有する RS グループ, 最適基準を事前情報として保持している $\text{GRC}_{\text{opt}}$, そして強化学習における一般的な方策である $\epsilon$-greedy を用いた.

全ての手法において学習率 $\alpha = 0.1$, $\gamma = 0.9$ と設定した. 提案手法である GRC グループは大局基準値 $\aleph_G$ の初期値を一律 $\aleph_G = 0$ とし, $\text{GRC}_{\text{opt}}$ の大局基準値 $\aleph_G = 8$ とした. そして GRC グループ, $\text{GRC}_{\text{opt}}$ はそれぞれ $\zeta(s_i)$ を一律 1 に, $\alpha_\tau = 0.1$, $\gamma_\tau = 0.9$, $\gamma_G = 0.9$ し, エピソード単位の獲得報酬を $E_{\text{tmp}}$ とした. また, RS グループでの $\tau_\alpha$ と $\tau_\gamma$ は GRC グループと同様に設定した. そして, GRC グループと RS グループの基準値を共有するタイミングはどちらもグループに属する全てのエージェントが 1 エピソード終えた時点とし, 基準値を共有した後にエージェントは次のエピソードに移る. $\epsilon$-greedy は $\epsilon = 1.0$ から等速度で減少させ, 2000 エピソード時点で $\epsilon = 0$ となるように設定した.

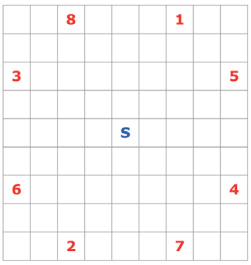


図 1: SuboptimaWorld 概要図

### 4.2 結果

提案手法である GRC エージェントを 4 体とした場合の 1000 回行った平均の結果を図 2 に示す. 図 2 から, 最高報酬にたどり着くのが早い順から, $\text{GRC}_{\text{opt}}$, GRC グループ, RS グループ, $\epsilon$-greedy となっており, $\epsilon$-greedy は 2000 エピソード経過時点ではわずかに最高報酬を下回る成績であることがわかる.

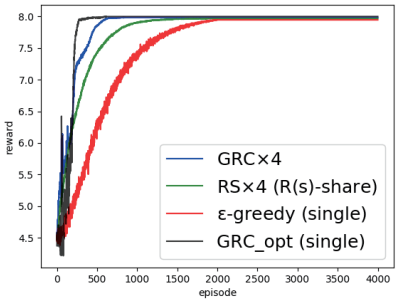


図 2: 獲得平均報酬の時間発展

そして, GRC エージェントを 1, 2, 3, 4 体とした 1000 回の平均の結果を図 3 に示す. 図 3 からエージェントが増えるごとに成績が良くなることがわかる.

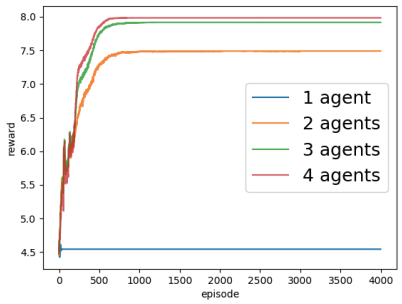


図 3: 獲得平均報酬の時間発展

そして, グループで具体的にどのように学習しているかを見るための 1 例として, 4 体グループで 1000 エピソードを 1 回のみ行った場合の大局観測期待値 $E_G$ を出力したグラフを図 4 に示す.

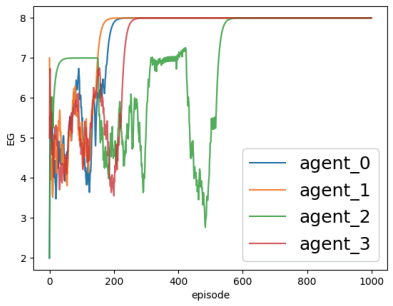


図 4: 大局観測期待値 $E_G$ の時間発展

報酬 7 で満足していたエージェントが他者がより高い報酬を獲得し始めた時点から再探索を行い, そして報酬 8 を得ることに成功しているのがわかる.

## 5. 考察

まずはじめに, 人数が増えるごとに GRC グループの成績が良くなることについて, 非満足状態のエージェントは共有された大局基準値 $\aleph_G$ 以上の報酬を得られる行動を学習するという性質から, 中には共有された大局基準値 $\aleph_G$ と等しい成績で満足する場合が存在することが考えられる. しかし, グループに属するエージェントが多いほど一度に観測される報酬のパターンが多くなることで 素早くより良い成績が得られるエージェントが出現する. その結果, グループの大局基準値 $\aleph_G$ が素早く上昇し, グループのエージェント数が多いほど共有された大局基準値 $\aleph_G$ と等しい成績に留まらず, より良い成績を発見することができると考えられる.

次に, 基準値の共有手法について, RS グループは行動価値から基準値を設定しているため, 価値更新手法の Q-learning の行動価値の更新式の学習率の影響から満足状態のエージェントの行動価値が収束するまでの間は, 非満足状態のエージェントの行動価値は満足状態のエージェントから共有される基準値以上の行動価値を満足状態のエージェントより少ない試行で獲得するのが困難である. 一方で GRC グループでは他者の行動価値を直接使わずに、エピソードで得られる報酬から計算される大局観測期待値 $E_G$ から基準値を更新しているため, 最適な行動の行動価値が他の行動価値よりも高くなっていれば最適な行動をすぐ取ることができる. したがって RS グループのように行動価値の収束を待つ必要があるために, 基準を満たす行動系列への収束が遅くなるということは起こりにくくなっている.

しかし, 準最適解で満足したエージェントのような, とある行動系列の行動価値が他の行動価値よりも高くなってしまった場合には, 探索によって他者から共有された基準値を満たす報酬が得られたとしても, すぐに基準を満たす行動をすることをせずに, 行動価値が高い今までの準最適解への行動を選びつつ, 徐々に基準を満たす行動へと移行していることが図 4 から推測できる. これは GRC が自身の最大行動価値から基準値を設定していることと, RS グループと同様に Q-learning の行動価値の更新式の学習率によって徐々に更新されることから生じるものであると考えられる. したがって, 価値更新手法の変更による, より素早い価値推定を行うことで改善が可能であると考えられる.

## 6. おわりに

大局基準値共有を用いた社会的学習により, より限られた情報共有で学習を有効に行うことに成功した. そして, グループのエージェント数が増えることで, 一定期間で探索する領域が広がり, より良い成績を素早く発見, 共有することで全体での成績が向上することが判明した. 今回提案した手法ではエピソード単位での獲得報酬を利用した情報を共有したため, 他のアルゴリズムともグループを作ることが可能であることが考えられる. そして, 今回では 1 エピソードごとに情報を共有していたところを, 情報を共有する間隔をより疎にすることによって, さらに学習中の情報共有に必要とする計算量を減少することが可能であると考えられる.

よって今後の課題として, GRC エージェントが他のアルゴリズムとグループを作った場合の挙動の変化の観測と, より少ない人数での成績の向上, より疎な間隔での情報共有による学習可能性の検証, そして今回のタスクではエージェントが確実に報酬を得られる設計であったため, より複雑なタスクでの適用手法を考案することが考えられる.

## 参考文献

[Manktelow 15] Ken Manktelow : Thinking and reasoning (2012) (邦訳: 思考と推論, 服部雅史, 山祐嗣 訳: 思考と推論　理性・判断・意思決定の心理学, 北大路書房 (2015), pp. 260-261

[Simon 56] Simon, H.A.: Rational choice and the structure of the environment, *Psychological Review*, 63(2), 129–138. (1956)

[Tamatsukuri 18] Akihiro Tamatsukuri, Tatsuji Takahashi: Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function. *arXiv preprint arXiv:1812.05795*, 2018.

[飯間 06] 飯間 等 & 黒江 康明: エージェント間の情報交換に基づく群強化学習法, 計測自動制御学会論文集, 42(11), 1224–1251. (2006)

[牛田 17] 牛田有哉, 甲野佑, 高橋達二: 生存を目的とする満足化強化学習, JSAI 2017, 4C2-2in2. (2017)

[其田 18] 其田憲明, 神谷匠, 甲野佑, 高橋達二: 満足化基準値共有を用いた社会的強化学習, JSAI2018 予稿集, 1N1-05. (2018)

[高橋 16] 高橋達二, 甲野佑, 浦上大輔: 認知的満足化 - 限定合理性の強化学習における効用, 人工知能学会論文誌, 31(6), 1–11. (2016)

[柄谷 85] 柄谷 行人: ブタに生れかわる話, 批評とポスト・モダン, pp. 257260. (1985)