4U-05

弱教示的強化学習における探索割合の自律調整

Autonomous adjustment of exploration ratio in weakly supervised reinforcement learning

佐鳥 玖仁朗 † 神谷 匠 † 高橋 達二 †
Kuniaki Satori Takumi Kamiya Tatsuji Takahashi

東京電機大学 †
Tokyo Denki University

1. 序論

機械学習の一種である強化学習では，エージェントは試行錯誤を通じて未知環境において収益を最大化する行動系列を学習する．環境に対する試行錯誤には探索と知識利用のトレードオフが存在し，エージェントは両者のバランスを調整しながら学習を行う必要があるが，実環境のような膨大な探索空間ではそのバランシングは困難である．対して人間は，満足化 [Simon 56] により現状が目標に対し非満足であれば探索を行い満足であれば知識利用を行うことで，膨大な探索空間であっても効果的に学習する．

満足化を実装した行動決定手法として，Risk-sensitive Satisficing(RS) が考案された．RS によるエージェントは，タスク達成目標を与えられたとき Global Reference Convertion(GRC)[牛田 17] と RS 価値関数 [高橋 16, 玉造 18] によって評価を非満足（不正解）と満足（正解）に変換し，弱教示的に学習する．しかし GRC を用いた学習は新奇探索性とスケールのパラメータ ζ の影響が強く働くため，多様な強化学習タスクに対して調整を要するがその調整は困難である．

そこで本研究ではハイパーパラメータ ζ を必要としない GRC ratio(GRCr) を提案し，実験を通して強化学習タスクへの広い適用可能性を示す．

2. Risk-sensitive Satisficing

RS による弱教示的強化学習を行うエージェントは基準に対し満足な収益の獲得を目的としており，収益と基準の差分である満足度を現方策の基準に対する弱教示的な信号とすることで探索性を調整し方策を改善する．

RS エージェントは，GRC により大局満足度と局所満足度を揃えるように各状態の基準を決定する．また，行動価値と基準の関数である RS 価値関数を最大化する行動を選択することで，満足時と非満足時で異なるリスク態度で振る舞う．現方策で満足な収益を得られるなら探索を打ち切りリスクを回避し，そうでないなら満足に近い行動の検証とリスク追求的な探索を均衡させて満足に至るまで学習する．

2.1 Global Reference Convertion

タスク達成目標として大局基準 $\aleph_G$ を定めた時，現方策で十分な収益を得ているのであればそれ以上の探索は不要である．収益が大局基準 $\aleph_G$ を下回る場合，その差分が小さいのであれば現方策の近傍に満足に至る方策が存在する可能性が高く，差分が大きいのであれば未だ満足方策との距離は遠いだろう．よって，この差分を用いて各状態の新奇探索性を調整することで，満足に向けて適当な探索が可能だと考えられる．

収益 E_G と大局基準 $\aleph_G$，状態基準 $\aleph(s_i)$ としたとき，大局満足度 δ_G と最大価値に対する状態満足度 $\delta_L(s_i)$ は次のように定義できる．

$$\delta_L(s_i) = \max Q(s_i) - \aleph(s_i) \quad (1)$$

$$\delta_G = \min(E_G - \aleph_G, 0) \quad (2)$$

両満足度の間には収益と価値のスケールの差が存在しており，そのスケールの吸収と探索性の制御のためのパラメータ $\zeta(s_i)$ を導入し，式 (3) のように両満足度を等価とみなすと，状態基準 $\aleph(s_i)$ は式 (4) で計算できる．

$$\delta_L(s_i) = \zeta(s_i)\delta_G \quad (3)$$

$$\aleph(s_i) = \max Q(s_i) - \zeta(s_i)\delta_G \quad (4)$$

状態基準 $\aleph(s_i)$ はエピソードが終了するたび更新する．

収益 E_G はタスクと達成目標により様々に扱う余地があるが，エピソードごとの収益 E_{tmp} と収益観測量 N_G，大局減衰率 γ_G を用いて次のように更新する．

$$E_G \leftarrow \frac{E_{\mathrm{tmp}} + \gamma_G(N_G E_G)}{1 + \gamma_G N_G} \quad (5)$$

$$N_G \leftarrow 1 + \gamma_G N_G \quad (6)$$

2.2 RS 価値関数

GRC により計算された状態基準 $\aleph(s_i)$ と信頼度 $\tau(s_i, a_j)$，行動価値 $Q(s_i, a_j)$ を用いて，RS 価値関数を次のように定義する．

$$RS(s_i, a_j) = \tau(s_i, a_j)\Big(Q(s_i, a_j) - \aleph(s_i)\Big) \quad (7)$$

$$\tau(s_i, a_j) = \tau_{\mathrm{curr}}(s_i, a_j) + \tau_{\mathrm{post}}(s_i, a_j) \quad (8)$$

また，信頼度 $\tau(s_i, a_j)$ は価値更新と合わせて以下のように更新される．

$$\tau_{\mathrm{curr}}(s_t, a_t) \leftarrow \tau_{\mathrm{curr}}(s_t, a_t) + 1 \quad (9)$$

$$\tau_{\mathrm{post}}(s_t, a_t) \leftarrow (1 - \alpha_\tau)\tau_{\mathrm{post}}(s_t, a_t) + \alpha_\tau\gamma_\tau(s_{t+1}, a_{t+1}) \quad (10)$$

このとき，γ_τ は信頼度の割引率を表し，α_τ は信頼度の学習率を表す． a_{t+1} は状態 s_{t+1} において選択する行動である．

RS 価値を最大化するような行動を選択する方策を RS 方策と呼び，本論では挙動方策として用いる．

3. GRC の問題点と GRC ratio の提案

GRC のパラメータ ζ には大きな問題が存在する．パラメータ ζ はタスクの複雑さや達成目標によって最適な値が変化し，また減衰率などが影響するため各状態ごとに異なる値を設定し適切なスケーリングが必要だが，それは困難である．よって現在は，経験的に全状態に共通の値が採用されることが多く，仮に最適な値がわかったとしてもスケーリングと探索性の 2 要因の分離は困難である．

よって，大局基準 $\aleph_G$ から状態基準 $\aleph(s_i)$ への変換の別アプローチとして，GRCr を提案する．

3.1 GRC ratio

GRC では満足度を $\aleph_G$ と E_G の差分 δ_G で表し，それを各状態の満足度に変換しようとしたためスケールの差が生じ ζ を必要とした．GRCr では満足度を $\aleph_G$ と E_G の割合 ρ_G とすることで，スケーリングを要せず各状態の満足度に変換する．

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

$$\rho_G = \max(\frac{\aleph_G}{E_G}, 1) \quad (11)$$

$$\aleph(s_i) = \max Q(s_i)\rho_G \quad (12)$$

割合による変換はスケーリングを要しない一方で，収益 E_G が小さい学習初期に探索過多となりやすい．正確な大局満足度の計算には探索がノイズとなるため，探索を抑制し正確な満足度を把握するための機構を導入する必要がある．

$$\sigma \leftarrow \begin{cases} 0 & \text{if } \sigma \geq \aleph_G N_G \\ \sigma + \aleph_G - E_{\text{tmp}} & \text{otherwise} \end{cases} \quad (13)$$

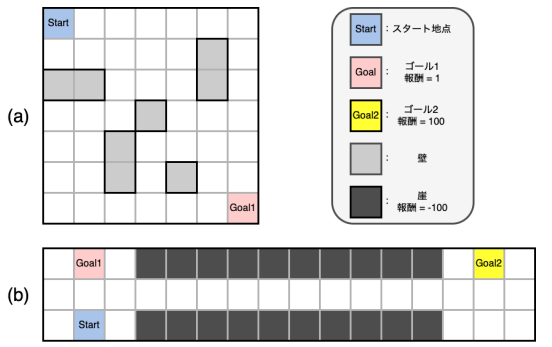
$$E_G \leftarrow \aleph_G \quad \text{if } \sigma \geq \aleph_G N_G \quad (14)$$

σ は累積非満足度を表し，その値が閾値 $\aleph_G N_G$ を超えた際に 0 へと初期化され，同時に大局収益 E_G を満足に足るものと見なし現方策による正確な大局満足度を評価する．閾値 $\aleph_G N_G$ は，大局減衰率 γ_G により収益観測量 N_G が定数に収束するため，同様に定数へ収束する．現方策が満足に至らない場合，大局収益 E_G の更新に伴い再度探索が行われる．

4. 実験

グリッドワールド拡張タスクである迷路タスクと一本道の崖渡タスクを用いて，GRC の ζ 依存性とそれに対する GRCr の適用可能性を示すためシミュレーションを行った．環境は図 1 と同様のものを用い，スタート S からゴール G もしくは崖にたどり着くまでを 1 エピソードとして 3000 エピソードを 1 シミュレーションとした．指標には，後悔を「最適行動系列収益と実行動系列収益の差」と定義し，その 1000 シミュレーション平均を用いた．

一般的な方策である ϵ-greedy によるエージェント (ϵ-greedy) と，GRC による RS エージェント (GRC)，GRCr による RS エージェント (GRCr) の 3 種のエージェントを用いて比較を行った．行動価値の更新は代表的な強化学習アルゴリズムの Q 学習で行い，学習率は $\alpha = 0.1$，割引率は $\gamma = 0.9$ とした．GRC と GRCr のパラメータは，大局割引率は $\gamma_G = 0.9$ とし，大局基準値 $\aleph_G$ はタスクごとの目標に従い設定した．また，スケーリングパラメータ $\zeta(s_i)$ は値を変化させ，ζ の値による後悔の変動を見た．ϵ-greedy は $\epsilon = 1$ から $1/n$ ずつ $\epsilon = 0$ まで減衰させ，減衰量にはタスクごとによい性能を示したエピソード数 n を用いた．



(a)：迷路タスク　(b)：一本道の崖渡タスク

図 1: 各タスクの設定

4.1 迷路タスク

迷路タスクでは，ゴールにたどり着く最短経路の学習を目的とした．大局基準 $\aleph_G$ は最適な方策のステップ平均収益 1/10 とし，同様にエピソード収益 E_{tmp} はステップ平均収益 $1/N_{\text{step}}$，後悔はその差分の累積とした．

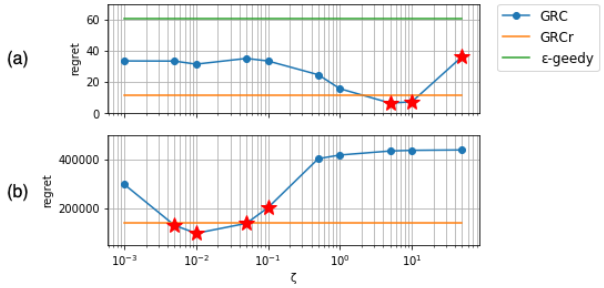
4.2 一本道の崖渡タスク

崖渡タスクでは，一本道を渡った先の最大報酬の獲得を目的とした．崖に落ちずに一本道を渡るにはランダムな探索では難しく，目標を持たない最適化では学習が困難となる．大局基準 $\aleph_G$ は最大収益である 100 とし，エピソード収益 E_{tmp} は単純な獲得報酬を用いて，後悔はその差分の累積とした．

4.3 結果

横軸を ζ，縦軸を後悔としたシミュレーションの結果を図 2 に示す．GRCr と ϵ-greedy は ζ を用いないため定数で示し，GRC は最適な方策を学習できた ζ を星で，それ以外を点で示す．一本道の崖渡タスクの ϵ-greedy は 3000 エピソードで最適な方策を獲得できなかったため，迷路タスクのもっとも性能が良かった減衰ステップ数 $n = 1500$ の結果のみを示す．

GRCr は，どちらのタスクでも後悔を抑えつつ最適な方策を獲得できている．また，GRC の最適な方策の獲得と後悔は，ζ の値が大きく影響しているのがわかる．



(a)：迷路タスク　(b)：一本道の崖渡タスク

図 2: 各 ζ の GRC と GRCr，ϵ-greedy の後悔の比較

4.4 考察

実験で行った 2 つのタスクは，複雑さや達成目標が異なったため ζ の最適な値が異なったと考えられる．最適な値付近の ζ より大きくても小さくても後悔が増加する理由として，最適な値の ζ に比べて大きいとき，満足するまでに探索が過剰に行われその過剰探索分の後悔が増加したと考えられる．また，最適な値の ζ に比べて小さいとき，探索が不足し満足する経路の学習が遅れ満足するまでに試行回数が増え後悔が増加したと考えられる．対して GRCr は満足度を割合として ζ を必要とせず，また，正確な満足度を正しく把握するための機構は，人間が行うような「練習による学習だけからでは目標の達成度合いが把握しにくいため，テストでどこまで学習できたのかを検証する」という学習の検証を行っている．そのため今まではタスクの複雑さや達成目標に左右され，ζ の値に頼っていた新奇探索性を自立的に調整することにより後悔が抑えられた学習ができたと考えられる．

5. 結論

本研究では GRC の問題点を提示し，それを解決する GRCr を提案した．また，実験を通して ζ を必要としない GRCr が自律的な探索調整による強化学習タスクへの広い適用可能性と GRC と同等の学習効率であることを示した．

今回の実験で GRC は最適な値の ζ を見つけ学習ができたが，より複雑なタスクになるほど最適な適値を見つけることが難しいと考えられる．さらに，深層強化学習のように 1 回のシミュレーションのコストが高く試行錯誤的に ζ を決められない場合もある．これらのことを踏まえると GRCr は，より複雑なタスクでの有用性が考えられるため，今後より複雑なタスクで GRC や他方策との比較検証したいと考えている．

参考文献

[Simon 56] Simon, H. A., "Rational choice and the structure of the environment", *Psychological Review*, Vol. 63, No. 2, pp. 129–138 (1956)

[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, Vol. 31, No. 6, pp. AI30-M-1–11 (2016)

[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 (2017)

[玉造 18] 玉造 晃弘, 高橋 達二, "認知的満足化価値関数の分析", JSAI 2018 (2018)