The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

満足化原理の強化学習全般への適用に向けて

Toward satisficing in general reinforcement learning

佐鳥 玖仁朗 *1 Kuniaki Satori　吉田 豊 *1 Yutaka Yoshida　山岸 健太 *1 Kenta Yamagishi　牛田 有哉 *2 Yuya Ushida　神谷 匠 *2 Takumi Kamiya　高橋 達二 *1 Tatsuji Takahashi

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

*2東京電機大学大学院 Graduate School of Tokyo Denki Univerity

As the scope of reinforcement learning broadens, optimization becomes less realistic, and bounded rationality that considers the limitations in agents gets more important. Satisficing, the principal model of bounded rationality, models how people and animals explore and exploit. However, there is no efficient algorithm that represents satisficing can be applied to reinforcement learning in general. We apply our satisficing model, reference satisficing (RS) value function, and the global reference conversion (GRC) technique to the broader reinforcement learning tasks than in previous studies. In the three tasks we deal with in this study, RS and GRC work well, while there are some open problems for general reinforcement learning tasks.

1. はじめに

強化学習では, エージェントの試行錯誤を通じて未知の環境において収益を最大化する行動系列を学習する. 試行錯誤における環境の探索と知識の活用には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が, 実環境のような連続状態空間を扱う場合に適切な探索と知識活用を行うことは非常に困難である. 一方で人間は, 満足化原理 [Simon 56] と呼ばれるように, 現状が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ように意思決定を行うことで, 探索空間の広大な未知の環境であっても目的とする基準を満たすような行動を可能としている. この満足原理を反映した意思決定手法として, 満足化価値関数 (reference satisficing: RS) が考案された. RS は多腕バンディット問題をはじめとした様々なタスクにおいて少数の探索で満足する行動系列を獲得でき, 基準値が適切であれば最適な行動系列を獲得できることがわかっている [高橋 16]. また, 状態が複数ある一般的な強化学習タスクでは各状態ごとに基準値が必要となるが, 未知の環境における各状態の基準値を推定することは非常に困難である. 一方で, タスク全体を通して得られる収益は既知であることも多く, タスク全体の期待収益に対する基準値を設定することで大局的な満足化度合いを定義しそれに対する各状態の基準値を動的に獲得する手法として大局基準変換法 (global reference convertion: GRC) が考案された [牛田 17a]. GRC を用いた RS(RS+GRC) は, 格子空間上に複数配置されたスイッチを決められた順番で押すことで報酬を獲得できるスイッチワールドタスクや運動制御課題で複雑なダイナミクスを有する大車輪タスクで, 一定以上の報酬を素早く獲得することに成功している [牛田 17b].

満足化原理は, 人間の意思決定・試行錯誤方策として重要視されながら, これまで強化学習全般に適用可能なアルゴリズムは存在していない. そこで本研究では, これまでに RS の性能と振る舞いを検証した諸タスクより一般的な強化学習の 3 タスクにおいて RS を用いる. 具体的には, 1 エピソード内で時間の経過によって報酬減衰して行くタスク, 確率的に報酬が変わるタスク, 負の報酬が与えられるタスクであり, それらのタスク対して GRC を適用することで, 様々な報酬関数に対する GRC の性質と GRC を用いた RS の適応可能性について議論する.

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

2. 満足化方策

人間の意思決定にはある基準値を定め, その値を超える行動を見つけるまで探索を行い, 発見した際は満足しその行動を選び続ける傾向が存在している. このような意思決定傾向は満足化と呼ばれ, 最良の行動を探索する最適化とは区別される. 満足化には, 複雑で最適化が困難な環境においても探索を打ち切る獲得報酬を基準として設けることで, 一定以上の報酬を獲得する行動系列を発見できる利点がある.

2.1 満足化価値関数 (RS)

RS は満足化に加え信頼性を考慮したアルゴリズムであり, 任意の状態行動対の行動価値関数を $Q(s_i, a_j)$, 任意の状態行動対とその後の信頼度を考慮した値を $\tau(s_i, a_j)$, 各状態に対する基準値を $R(s_i)$ として定義することで以下の式で算出される.

$$RS(s_i, a_j) = \tau(s_i, a_j)\Big(Q(s_i, a_j) - R(s_i)\Big) \quad (1)$$

$$\tau(s_i, a_j) = \tau_{\mathrm{curr}}(s_i, a_j) + \tau_{\mathrm{post}}(s_i, a_j) \quad (2)$$

各状態行動対の $\tau(s_i, a_j)$ はその対の訪問ごとに以下のように更新される.

$$\tau_{\mathrm{curr}}(s_t, a_t) \leftarrow \tau_{\mathrm{curr}}(s_t, a_t) + 1 \quad (3)$$

$$\tau_{\mathrm{post}}(s_t, a_t) \leftarrow \tau_{\mathrm{post}}(s_t, a_t) + \alpha_\tau\Big(\gamma_\tau(s_{t+1}, a_{t+1}) - \tau_{\mathrm{post}}(s_t, a_t)\Big) \quad (4)$$

このとき, γ_τ は未来信頼度割引率を表し, α_τ は信頼度学習率を表す. a_{t+1} は状態 s_{t+1} において選択する行動である. RS による評価値を最大化するような行動を選択する方策を満足化方策と呼ぶ.

3. 大局基準変換法 (GRC)

現実的なタスクにおいて, 各状態ごとの適切な基準値を推定することは困難である. 一方で, タスク全体を通して得られる収益は既知であることも多く, タスク全体の収益に対する適切な基準値の推定は可能である場合が多い. そこで, タスク全体を通しての基準値と現時点でのエージェントの達成度合いからタスク全体の満足度合いを求め, その値から各状態の基準値を求める手法である GRC が考案された. 以下に, GRC を用いた各状態の $R(s_i)$ の算出方法を示す.

まず, タスク全体の満足度合いを表すための達成度合いの更新式を定義する. 全体を通しての目標水準を大局満足化基準値 (global reference : R_G), それに対する大局観測期待値 (global expectation : E_G) と定義し, E_G はある期間 T_{tmp} ごとに消去, 再蓄積される一時的平均獲得報酬を用いて以下の式で更新される.

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

$$N_G \leftarrow 1 + \gamma_{N_G} \quad (6)$$

この E_G と報酬関数の設計に基づいてタスク設計者が与えた R_G との差 $R_G - E_G$ をタスク全体の満足化度合いとする.

各状態の満足化度合いは, 状態 s_i における最大の行動価値 Q を $\max Q(s_i)$ とした時, $\max Q(s_i) - R(s_i)$ で表される. E_G と行動価値 Q はスケールが異なるため, スケーリングパラメータ $\zeta(s_i)$ を導入し各状態に対してタスク全体の満足化度合いを適用する. 各状態の基準値 $R(s_i)$ は, 以下の式で算出する.

$$\delta_G = \min(E_G - R_G, 0) \quad (7)$$

$$\max Q(s_i) - R(s_i) = \zeta(s_i)\delta_G \quad (8)$$

$$R(s_i) = \max Q(s_i) - \zeta(s_i)\delta_G \quad (9)$$

RS+GRC は, スイッチワールドタスクや大車輪タスクで一定以上の報酬を素早く獲得することに成功している.

4. 報酬減衰グリッドワールド

報酬減衰グリッドワールドでは, 時間経過によって獲得可能な報酬が減衰する報酬関数に対して RS+GRC が適切な行動を獲得可能であるか検討する. 学習が最短経路の学習に対して行われたのではなく減衰の少ない報酬を得られる行動系列に対して行われたことを確認するため, 格子空間上に決められた位置のスタートと 4 つのゴールを配置した. エージェントは, 格子空間上の位置によって状態を識別し, 各状態で上下左右の隣接したマス目に進むために行動することができる.

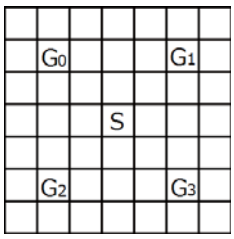

図 1: グリッドワールドタスクの設定

4.1 シミュレーションの設定

本タスクでは図 1 と同様の設定でシミュレーションを行った. 各ゴールの報酬は初期値が異なっており, エージェントがゴールへ到着すると初期値を到達ステップ数 t で割った $G_0 = 6/t, G_1 = 3/t, G_2 = 1/t, G_3 = 5/t$ が報酬として与えられる. また以降はマスの座標表現として, 格子空間を左上のマスを基準に横を x 座標, 縦軸を y 座標とてして座標 (x, y) と表現し, 実際にエージェントが通過したマスの系列を経路と呼ぶ. 格子空間は 7×7 の合計 49 マスで, エージェントは座標 $(3, 3)$ をスタートとして配置した. スタートから始まりゴールにたどり着くまでを 1 エピソードとして 1000 エピソードまで行い, 評価の指標として後悔を用いた. 強化学習タスクでは行動ではなく方策としての評価が必要であるため, 「最適経路を選び続けた場合の累計期待報酬と実際に選んだ経路の累計期待報酬の差」が後悔の定義となる.

アルゴリズムは代表的な強化学習アルゴリズムの Q 学習用いて, 学習率は $\alpha = 0.1$, 割引率は $\gamma = 0.9$ とした. エージェントの方策は, 検証方策である RS+GRC と, 比較方策として ϵ-greedy を用いた. 大局基準値 R_G は, 最短経路の 4 ステップで G_0 へ到達した際の収益である 1.5 を目指すため, その単位時間期待値として 0.375 に設定した. 大局割引率は $\gamma_G = 0.9$ とし, スケーリングパラメータ $\zeta(s_i)$ は経験的に良い成績を残した設定として全状態において一律に 1 とした. ϵ はエピソードごとに 1 から 1/600 ずつ減衰させ, 600 エピソード時点で $\epsilon = 0.0$ になるように設定した.

4.2 結果と考察

シミュレーションの結果として後悔の時間発展を図 2 に示す. ϵ-greedy は $\epsilon = 0$ になるエピソード数 600 以降で最適な経路を発見できずに後悔が増加し続ける結果となっている. ここで $\epsilon = 0$ になるまでのエピソード数を減らしても最適な経路を発見できる割合が減るのでさらに後悔が増え続ける傾きが大きくなる. 逆に $\epsilon = 0$ になるまでのエピソード数を増やしても最適な経路を発見できる割合が増えるので後悔の傾きが小さくなるが, 最適な経路を発見するまでの後悔が増えると考えられる. それに対し, 満足化アルゴリズム RS+GRC は ϵ-greedy より速く最適な経路を見つけ, 後悔も増えていない. この結果から RS+GRC は時間経過によって報酬減衰していくこのタスクに適応できたと言える. また ϵ-greedy より後悔が少なかったことよりこのタスクでは RS+GRC の有用性を示せた.

次にグリッドワールドタスクに GRC が適応できた理由は, ステップ数の増加によって獲得できる報酬が減って行くが, 1 エピソードで得られる最大報酬は決まっているので最大報酬から R_G が計算でき, E_G は R_G に収束し適応できた.

また, 時間の経過により報酬減衰するタスクでは, しないタスクと比べてゴールから得られる報酬がいつも最大とは限らないので, 最大報酬を与えるゴールでも辿り着くためのステップが多くかかったため得られる報酬が小さくなり, エージェントが他のゴールを選びやすくなってしまうような誤った学習をして局所解に陥りやすい. それに対して RS+GRC は, 最大報酬を学習していなければ図 3 のようにエピソード 500 あたりまでは E_G が R_G に達せず非満足状態となる (図 3 は 1 シミュレーションの E_G の時間発展を示したもの). そのため信頼度の影響により多く試行したものほど信頼度が低くなり, 試行回数が少ない行動の信頼度が高くなる. 非満足状態ではなるべく報酬が高いゴールへ行動しているが信頼度が低くなると探索を繰り返し E_G が R_G に収束するような最大報酬を見つけられると満足化した. そのため局所解に陥らず最適解を探索できたと考えられる.

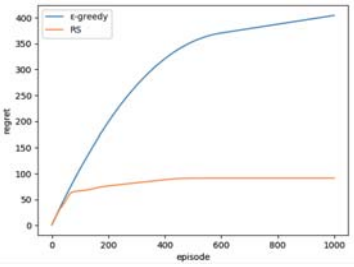

図 2: グリッドワールドの後悔の時間発展

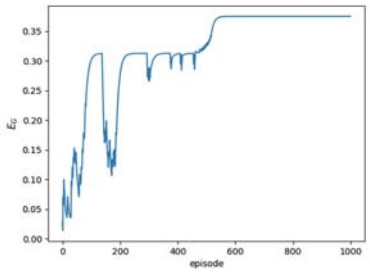

図 3: グリッドワールドの E_G の時間発展

5. ツリーバンディット

ツリーバンディットでは, 報酬が確率的に与えられる報酬関数に対して RS+GRC が適切な行動を獲得可能であるか検討する. 各状態では 4 本腕バンディットを行い, 選択した行動によって異なる状態へ遷移する. 層は全 2 層とし, $1+4$ の計 5 つの状態に対して最適な行動系列の獲得を目的とする.

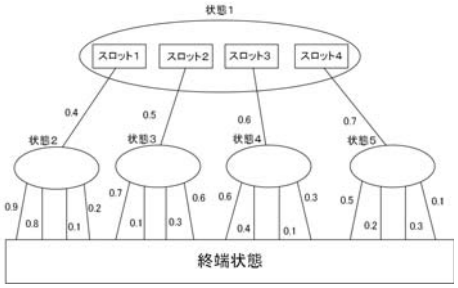

図 4: ツリーバンディットタスクの設定

5.1 シミュレーションの設定

報酬は報酬確率に従って, 0 または 1 とし, 報酬確率は図 4 と同様の値に設定した. 各層での行動選択を 1 ステップ, 終端状態である 2 層目の選択までを 1 エピソードとして 10000 エピソードまで行い, 評価の指標として後悔を用いた. アルゴリズムは代表的な強化学習アルゴリズムの Q 学習を用いて, 学習率は $\alpha = 1/n_j$, 割引率は $\gamma = 0.9$ とした. このとき n_j はその状態での行動の試行回数を意味する. エージェントの方策は, 検証方策である RS+GRC と, 比較方策として ϵ-greedy を用いた. 大局基準値 R_G は, 最適経路の期待累積収益 1.3 の単位時間期待値である 0.65 ではなく, 0.64 とした. 大局割引率は $\gamma_G = 1.0$ とし, スケーリングパラメータ $\zeta(s_i)$ は経験的に良い成績を残した設定として全状態において一律に 0.2 とした. ϵ はエピソードごとに 1.0 から 1/3000 ずつ減衰させ, 3000 エピソード時点で $\epsilon = 0.0$ になるように設定した.

5.2 結果と考察

シミュレーションの結果として後悔の時間発展を図 5 に示す. ϵ-greedy の結果は最終的にエピソード数が 3000 となった時点で最適な経路を発見できずに後悔が増加し続ける結果となっている. 一方, 満足化アルゴリズム RS+GRC は ϵ-greedy よりも後悔が少なく, 増加もしていない.

この結果より RS+GRC は確率的に報酬が変化する場合にも適用でき, ϵ-greedy よりも後悔が少なくこの確率的タスクにおいても有用性があるのではないかと考えられる.

ツリーバンディットに GRC が適応できた理由は, E_G と R_G の値を適切に設定できたからである. 報酬が確率的に与えられるタスクにおいては同じ腕を選んでも報酬が一定にもらえるわけではないので E_G は一定の値に収束することはなく R_G に近づきはするが R_G より大きくなったり小さくなったりと値が振れる. そのため最大報酬を取る腕を選んでいたとしても満足と探索を繰り返してしまう. ここで E_G の振り幅を小さくし, R_G の値の設定を E_G が振れる幅になるべく含まれていない値を設定する事により GRC が適応できる.

シミュレーションの設定では $\gamma_G = 1.0$ と $R_G = 0.64$ と設定していた. 理由は, γ_G とは大局割引率であり, 1 に近いほど E_G は昔の報酬も大きく考慮する値となり E_G の更新する変化量は小さくなる. また, 0 に近いほど E_G は直近の報酬を大きく考慮した値となり, 1 か 0 に近い値をとるので E_G の更新する変化量は大きくなる. 今回は E_G の更新する変化量を小さくしたいため 1 とした. その設定により E_G の値は全ての報酬の平均値となり探索した際の報酬も含まれるが, エピソード数が大きくなるたびに影響は小さくなり E_G は最大単位時間期待値に近づいていく.

次に R_G だが, 通常であれば期待収益の単位時間期待値である 0.65 に設定する. しかし, 報酬が確率的であり E_G が収束しないため. 最適値 0.65 に近くかつ次に良い値 0.6 で満足しない位置に設定することが望ましく, 0.64 とした.

この設定により GRC は報酬が確率的なタスクに適応できた. また, 報酬が確率的であっても E_G を最大単位時間報酬期待値へ収束させることができれば $R_G = 0.65$ に設定することでも今回のような結果が得られると考えられる.

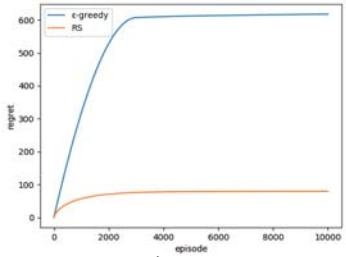

図 5: ツリーバンディットの後悔の時間発展

6. 崖歩き

崖歩きでは, 獲得報酬が負である報酬関数に対して RS+GRC が適切な行動を獲得可能であるか検討する. これまでの研究では負の報酬が与えられるタスクにおいて RS+GRC の動作の検証を行なって来なかった. GRC では E_G の更新にエピソードを通した累積獲得報酬を用いるため, 各状態ごとの基準 R に影響を与える. 崖歩きは強化学習におけるリスクを回避した最適な行動を獲得するタスクで, 格子空間上に崖を設置し崖マスに移動した際に大きな負の報酬を与える.

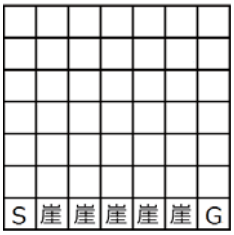

図 6: 崖歩きタスク

6.1 シミュレーションの設定

エージェントがゴールもしくは崖のあるマス目への行動を選択すると報酬が与えられ, エージェントはエピソード開始時の初期位置に戻る. 崖方向の報酬は大きな負の値として -100, ゴール方向への報酬は正の報酬として 1 を与えた.

格子空間は 7×7 の合計 49 マスで, エージェントの初期位置は座標 $(0, 6)$ に, ゴールの位置を座標 $(6, 6)$ に設定した. 崖の数は 5 マスとし, 図 6 のように配置した. 50 ステップを 1 エピソードとし, 50 エピソード行い, そのシミュレーションを 1000 回行なった結果を平均した. アルゴリズムは代表的な強化学習アルゴリズムの Q 学習を用いて, 学習率は $\alpha = 0.1$, 割引率は $\gamma = 0.9$ とした. エージェントの方策は, 検証方策である RS+GRC と, 比較方策として ϵ-greedy を用いた. 大局基準値 R_G は, 最適経路の期待累積収益 12.0 の単位時間期待値として 0.12 とした. 大局割引率は $\gamma_G = 0.9$ とし, スケーリングパラメータ $\zeta(s_i)$ は経験的に良い成績を残した設定として全状態において一律に 0.1 とした. ϵ はエピソードごとに 1.0 から 1/20 ずつ減衰させ, RS が最大報酬を獲得できる経路に収束する 20 エピソード時点で $\epsilon = 0.0$ になるように設定した.

6.2 結果と考察

シミュレーションの結果として後悔の時間発展を図 7 に示す. まず, 20 エピソードで ϵ が 0 になる ε -greedy 法の結果を

見る．この後悔は最初急激に増加し，その後一定の割合で増加し続けていることから，この時点で探索を打ち切ってしまうと，探索が足りないため最適な行動系列が発見できないことがわかる．一方で RS+GRC は ϵ-greedy 法よりも後悔を増加させずに早い段階で最適な行動系列を発見することができ，その有用性を示した．

RS+GRC が ϵ-greedy 法より高い成績を示した理由について，崖方向への行動は大きな負の報酬を与えられると，その行動の RS 値は一度の選択で急激に小さくなり再度選ばれることはなくなる．そのため崖方向への行動が選ばれなったことと，RS の信頼性を考慮した探索を行ったことが，結果後悔を $\epsilon-$greedy 法に比べ増やさなかったことに繋がったと考えられる．

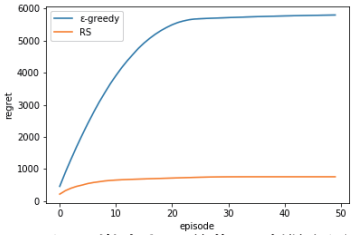

図 7: 崖歩きの後悔の時間発展

6.3 設定の変更

次に崖歩きタスクの設定を変更し，環境に状態が複数存在し，負の即時報酬が常に与えられるタスクでの挙動を確認する．そのために行動に対し常に移動コストとして負の報酬 -1 を与え続け，ゴールにたどりついたときに 0 の報酬を与えるように設定を変更したタスクでのシミュレーションを行う．また，このタスクではゴールにたどり着いた時にエピソードが終了するように設定を変更した．

格子空間や崖の配置は前タスク同様に図 6 のように設定をした．ゴールにたどり着くまでを 1 エピソードとし，300 エピソード行いそのシミュレーションを 1000 回行った結果を平均した．比較アルゴリズムは ϵ 減衰法の代わりに $\epsilon = 0$ の greedy 法を用いる．RS+GRC によるシミュレーションは，スケーリング変数 $\zeta(s_i)$ を $\zeta(s_i) = 1.0, 0.5, 0.1, 0.01$ の 4 つを用いて行なった．またこの際，$\zeta(s_i)$ は全状態に置いて一律に設定した．GRC に おける大局基準値 R_G には，1 エピソードにおいて最短の経路を通った場合における収益である -7 を目指すため，その単位時間期待値として $R_G = -0.875$ に設定した．大局割引率は $\gamma_G = 0.9$ とした．また，すべてのアルゴリズムにおいて学習率は $\alpha = 0.1$，割引率には $\gamma = 0.9$ を用いる．

6.4 結果と考察

シミュレーションの結果として後悔の時間発展を図 8 に示す．最初に greedy 法の結果をみると，後悔を序盤と中盤で伸ばすが，200 エピソード手前で最適な経路を発見することができている．これは，行動するたび常に一定の負の報酬が与えられるタスクにおいて，探索を十分な回数行った場合，最短経路の Q 値が最も小さくなりその経路を選ぶようになるためだと考えられる．次に $\zeta(s_i)$ を 1.0, 0.5, 0.1 に設定した RS の結果を見ると，最終的に最適な経路を発見できているが，他のアルゴリズムに比べ，後悔が大きくなっている．最後に $\zeta(s_i) = 0.01$ に設定した結果を見ると，最適経路を見つけることができているが，greedy 法と一致する結果となっている．この結果より，この設定のタスクでは RS+GRC は他のアルゴリズムに比べ，その有用性を示すことはできなかった．

次に，なぜ後悔が greedy 法と同じかそれ以上に蓄積されることに繋がったかを考える．RS+GRC においてスケーリング変数 $\zeta(s_i)$ はその値が適切なものより小さい場合，$Q(s_i)$ に差が出てきた際の $Q(s_i, a_j) - R(s_i)$ の最大値とそれ以外の大小関係は $\tau(s_i, a_j)$ による変動が起きにくくなるため，greedy 行動をしやすくなる傾向となる．反対に大きすぎる場合，$RS(s_i, a_j)$ の変動が起こりやすくなるので探索を行う傾向となる．今回のスケーリング変数の設定の一つである $\zeta(s_i) = 0.01$ は非常に小さい値であったため，greedy 法と結果が一致することになったと考えられる．また，RS の後悔が greedy 法を下回らなかった理由について，RS は E_G が R_G に近似し，満足化した際，以降 Q 値から greedy に行動決定を行うようになる．しかし，常に負の報酬を与え続けるタスクにおいて，最適な値に収束していない Q 値は行動を行うたびにその値を減少させていく．そのため Q 値が十分更新されていない場合，Q 値の大小関係が入れ替わり続けることから，最適な経路を選び続けるためには Q 値の減少によってその大小関係が変動しなくなるまで更新を行う必要がある．結果として RS は探索を行う分，greedy 法より後悔を蓄積することとなったと考えられる．

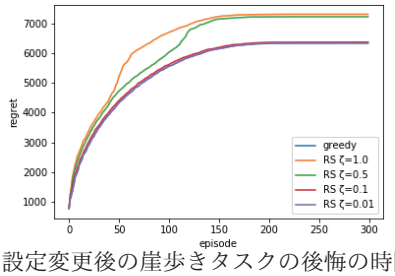

図 8: 設定変更後の崖歩きタスクの後悔の時間発展

7. 結論

本研究では GRC を用いた RS を先行研究で試されてこなかった，1 エピソード内で時間の経過によって報酬減衰して行くタスク，確率的に報酬が変わるタスク，負の報酬が与えられるタスクの 3 つの報酬関数をもつタスクに適用することでどのようなタスクに適応できるかの検証を行った．結果として，タスク内で正の報酬のみが与えられる場合では大局期待値 E_G が大局満足化基準値 R_G に収束またはツリーバンディットのように適切に超えるパラメータ設定をすることで最適な行動系列を取り続けることができ，適応できることがわかった．大きな負の報酬がタスク内にある場合では GRC を用いた RS は大きな負の報酬を避ける経路を少ない試行回数で学習し，その有用性を示したが，移動コストとして負の報酬を与える崖歩きの結果から，全状態の Q 値が行動のたびに減少するタスクの場合は RS は各状態の Q 値が十分に更新されるまで満足する行動を取り続けることはできないため，Q 値が減少を続けないような報酬関数の設計を考える必要があるといえる．

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)

[牛田 16] 牛田 有哉, 甲野 佑, 浦上 大輔, 高橋 達二: 探索割合を 自律調節する強化学習手法満足化基準の動的獲得, JSAI 2016 (2016 年度人工知能学会全国大会 (第 30 回)) (2016)

[牛田 17a] 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 年度人工知能学会全国大会 (第 31 回)) (2017)

[牛田 17b] 牛田 有哉, 甲野 佑, 高橋 達二:強化学習と満足化による素早い行動系列の獲得, 情報処理学会第 79 回全国大会講演論文集, 5M-04. (2017)

[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)