4C2-2in2

生存を目的とする満足化強化学習

Satisficing Reinforcement Learning for Survival

牛田 有哉 *1　甲野 佑 *2　高橋 達二 *2
Yuya Ushida　Yu Kohno　Tatsuji Takahashi

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

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

While the usual goal of reinforcement learning is to maximize the rewards, real agents like animals and humans tend to balance their survival costs and the rewards. For this more realistic goal, *satisficing* provides an efficient way. Some of the authors have proposed a new cognitive satisficing reinforcement learning algorithm called RS. In this study, we propose a technique for RS to allocate the global aspiration level to local states. We show its effectiveness in three tasks: a *switch world*, in which the agent needs to trip several switches by visiting the locations in a specific order to get a reward, a partially observable version, *blind switch world*, in which the agent does not know if it has tripped switches, and a *restless switch world*, in which the location of the switches periodically changes in an unobservable way.

1. はじめに

未知の環境において，環境との相互作用を通じて目的を達成するような行動系列を学習する枠組みに強化学習が存在する．しかし試行錯誤的に学習を行う強化学習では情報収集に時間を割く必要があり，タスクの複雑さが増すと現実的な時間内で最適化を行うことは非常に困難となる．一方で動物や人間にとっての最大の目的は (場当たり的ではあるが最低限以上に得し続ける)“生存” であり，1 日に消費する基礎代謝 (ライフコスト) に対して，そのコストを上回るような餌 (報酬) を獲得し続けることを優先することで，複雑な環境へ柔軟に適応しながら生活してると考えられる．そこで本研究では，学習の目的を，最適化 (報酬の最大化) から生存に必要な報酬の継続的獲得に緩和することで，現実時間内で有効な行動系列の獲得を達成する強化学習アルゴリズムを提案する．

ある目標水準を満たすことを目的とした場合，満足化と呼ばれる人間の意思決定傾向の側面を価値関数として表現した満足化価値関数 (reference satisficing : RS) が，多本腕バンディット問題において極めて高い成績を示している [高橋 16]．さらに，RS は強化学習課題全般への拡張も行われており，全状態に一律の目標水準 (基準値) をエージェントに与えた場合でも素早く良い行動系列を獲得できることが示されている [牛田 16]．しかし，生存という全体としての目標水準 (大局基準) から逆算して，それを達成するための各状態の基準値の適切な決定方法については言及されていなかった．そこで我々は大局基準を定義し，それを局所的な状態と対応付けて基準を割り振るような大局基準変換法 (global reference convertion : GRC) を考案し，GRC を適用した RS の有用性の検証を行う．

2. 強化学習と生存

未知の環境において動物が生存していくためには，自ら行動して適切な餌場を探す必要がある．このとき，動物自身は既に発見している餌場が最適な餌場であるかどうかを判断することができないため，他により良い餌場があるかどうかは実際に別の場所を探してみないとわからない．しかし，探している間は餌を獲得できなくなることから飢えのリスクが増加してしまうため，これ以上餌場の探索にコストを割くかどうかの判断が困難になる．

これは探索と利益追求のトレードオフと呼ばれ，強化学習における大きな課題の一つとされている．つまり良い行動系列を獲得するには，既にある情報の中で最良の行動を選択をする “利益追求” と，最良でない行動を試す “探索” を行う必要があるが，限られた時間内ではどちらも十分に行うことは困難な場合が多いため，どちらをどの程度行うかのバランシングが重要となる．

そこで我々は動物や人間の意思決定に習い，生存を目的とすることで，最適でなくても最低限の目標水準を素早く達成するように探索割合のバランシングを行う強化学習アルゴリズムの考案を目指す．

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. 満足化方策

任意の目標水準を満たすことを目的とする場合，人間の意思決定傾向の側面である満足化が有用であると考えられる [Simon 56]．人間は意思決定において，「ある基準値 (目標水準) を定め，基準値を超える価値をもつ選択肢 (行動) が見つかるまで探索を続け，発見したら探索を止めてその行動に満足する」という傾向を持つ．このような意思決定傾向を満足化と言い，最良の行動を追求し続ける最適化とは区別される．満足化は探索と利益追求の境目を基準値というパラメータによって切り替えることができるため，最適化が困難である場合においても，探索を打ち切る獲得報酬レベルを条件として決められるという利点が存在する．本研究では満足化の性質に加え，人間のリスク考慮の特性をモデル化した RS 価値関数 [高橋 16] を強化学習へ応用することで，生存という目標水準を満たすような強化学習の実現を目指す．

3.1 強化学習における満足化価値関数 (RS)

元々多本腕バンディット問題で扱われていた RS 価値関数の評価値はその選択で得られる報酬のサンプル平均，そのサンプリング回数に基づいた信頼度，そして基準値によって定義されていた [高橋 16]．それに対し強化学習全般では，一つの行動だけでなく行動系列に対して評価を行う必要があり，遅延報酬を考慮した価値の評価となるような形式にする必要がある．そのため，サンプル平均を任意の状態行動対からの未来の収益

予測である Q 値，その Q 値の信頼度として，任意の状態行動対とその後のサンプリング経験の蓄積である $\tau(s_i, a_j)$，そして各状態に対する基準値 $R(s_i)$ として定義し，以下のように評価値 $RS(s_i, a_j)$ が算出される．

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

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

状態行動対 (s_i, a_j) ごとの τ 値 (式 (2)) は，その対の訪問毎に式 (3) と (4) により更新される．

$$\tau_{\text{current}}(s_t, a_t) = \tau_{\text{current}}(s_t, a_t) + 1 \quad (3)$$

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

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_{up} は，方策毎に異なる，次に選択する行動である．本研究では，この価値関数に対して greedy 法を用いた選択を行うアルゴリズムとして使用する．

4. 大局基準変換法 (GRC)

強化学習に拡張された RS では各状態に対して基準値 $R(s_i)$ が存在する．このとき， $R(s_i)$ の比較対象となる Q 値は「ある状態において，その後得られると期待される収益」であり，その値は状態によって異なる．それに対し，先行研究では各状態ごとに適切な基準値を与えれば，素早く良い行動系列が獲得できることが示されている [牛田 16]．一方で生存を目的とすることを考えると，エージェントは生存に必要な餌の量という全体としての一つの目標水準を持っているが，そのために局所的にどのような目標を決めればよいかは不明である．つまり，生存においてはタスク全体としての基準値は存在するが，各状態での基準値は自律的に獲得する必要がある．しかし，従来の RS を用いた強化学習手法では，各状態ごとの適切な基準値が不明な場合は，全ての状態に対して一律の値に初期設定する手法を用いており，状態毎の Q 値のスケールの違いを考慮した基準値の設定方法が存在しなかった．そこで我々は，生存というタスク全体の満足度を表すため，達成度合いとして大局観測期待値 (global expectation : E_G) と，それに対応する大局満足化基準値 (global reference : R_G) を定義し，そこから各状態の基準値へ変換する手法である大局基準変換手法 (global reference convertion : GRC) を考案した．タスク全体に対する評価である大局観測期待値 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 (E_{tmp}: temporary expectation，ある期間中，期間内の試行回数で平均してどの程度報酬を獲得したか) を用いて以下の式で更新される．

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

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

ここで T_{tmp} は 1 エピソードのステップ数と仮定すると，一時的平均獲得報酬 E_{tmp} はあるエピソードでのステップ毎の平均獲得報酬を意味する．パラメータ γ_G は大局割引率を表し $0.0 \leq \gamma_G \leq 1.0$ をとる．

また動物の生存において，必要な餌の量はエージェントの内的欲求としてあらかじめ備わっているものだと考えられ，そのため本研究において大局基準値 R_G は所与とする．ここで，満足化度合いの定義を考えると，と大局的な満足化度合いは $E_G - R_G$ で表されるのに対し，各状態では状態 s_i における最大の Q 値を $\max Q(s_i)$ として $\max Q(s_i) - R(s_i)$ で表される．このとき，Q 値は状態ごとに異なる値であるため大局的な満足化度合いを各状態に反映させるためには，スケーリングの補正が必要となるため，各状態ごとにスケーリングパラメータ $\zeta(s_i)$ を導入し，$R(s_i)$ を式 (9) で計算する．この手法を大局基準値変換手法 (GRC) と呼ぶ．

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

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

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

5. スイッチワールドタスク

動物や人間が生存する際，1 日で消費する基礎代謝が決まっており，それを超えるような食物を摂取する行動系列の獲得を目指す．ここで，日によって大きな環境の変化がないと考えれば，目標を達成するような行動系列を見つけ，それを毎日行うようなループの獲得が重要となる．

そこで，本研究では格子空間上に番号のついた複数のスイッチが配置されたスイッチワールドタスクを用いて検証を行った．エージェントは，一度の行動で上下左右の隣のマス目に進むことができ，スイッチのあるマス目を通過すると，スイッチが押され，それに対応したランプが点灯する仕組みとなっている．スイッチは順番通りでないと押せない仕組みなっており，最後のスイッチが押されたタイミングで報酬が 1 で与えられ，すべてのスイッチがリセットされる．このとき，エージェントは現在自分がいるマス目の座標と，ランプの点灯状態を観測することができるため，エージェントの状態は座標に加えランプの点灯情報（どのスイッチが押されているか）で与えられるため，状態数は格子空間のマス目の数 × スイッチの数となる．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．

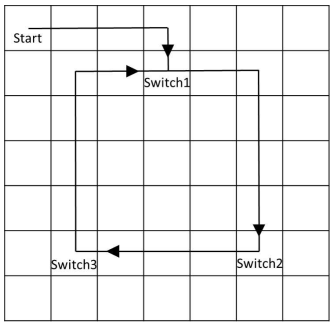

図 1: スイッチワールドタスクとその上での最適 = 最短経路

5.1 シミュレーションと結果

本シミュレーションでは図 1 と同様の設定でシミュレーションを行った．格子空間は 7×7 の合計 49 マス，スイッチの数は 3 つであり，エージェントの初期位置は座標 $(0, 0)$ に設定した．100 ステップを 1 エピソードとし，1,000 エピソード行い，そのシミュレーションを 1,000 回行った結果を平均した．比較アルゴリズムは，代表的な強化学習アルゴリズムとして Q 学習，満足する価値を持つ行動を見つけるまで行動決定方策に従う行動を取り，満足する行動を発見したらその行動を選択し続ける素朴満足化ポリシー (PS : Policy Satisficing) を用

いた．また，一般的な強化学習である価値の 1 ステップバックアップによる学習手法が，本タスクにおいて成否にどう関わるか調べるため， 1 ステップごとに観測した価値差分を過去の価値推定に反映するアルゴリズムである Sarsa (λ)[Sutton 00] も比較に用いる．Q 学習の行動決定方策には ϵ-greedy を使用し，ϵ はエピソードごとに 0.005 ずつ減衰させ 200 エピソード時点で $\epsilon = 0.0$ になるように設定した．Sarsa (λ) には反映する価値の重みパラメータである $\lambda = 0.9$ を設定した．Sarsa (λ) と PS の行動決定方策には $\epsilon = 0.1$ の ϵ-greedy を使用した．RS と PS の $R(s_i)$ 決定には GRC 手法を用い，GRC における大局基準値 R_G には，1 エピソードにおいて報酬が獲得できるような最短のループをし続けた場合における収益である 6.0 を目指すため，その単位時間期待値として $R_G = 0.06$ に設定した．大局割引率は $\gamma_G = 0.9$ とし，スケーリング変数は全状態において一律とし，経験的に良かった $\zeta(s_i) = 0.05$ に設定した．また，すべてのアルゴリズムにおいて，学習率 (Q 値の更新割合) は $\alpha = 0.1$，割引率には $\gamma = 0.9$ を用いる．

シミュレーションの結果としてエピソードごとに得られた報酬の時間発展を図 2 に示す．まず，ϵ-greedy の結果を見ていく．最終的な到達地点を見ると，ϵ の値が 0.0 になり greedy な行動のみを選択するようになる 200 エピソード以降，このタスクの最高収益である 6.0 にやや満たない結果となっている．つまり，この時点で探索を打ち切ってしまうと，探索が足りず最適な行動系列を発見することができないことで局所解に陥っていると考えられる．次に Sarsa (λ) の結果を見ると，立ち上がりは最も速いものの，1,000 エピソード時点での到達報酬では他のアルゴリズムに比べ低くなっている．また PS は到達が 6.0 に満たずに一定の値になっていることから，満足化できる行動系列を発見することができずに，ϵ-greedy による行動決定を行い続けていると考えられる．一方で RS は学習序盤から獲得報酬を伸ばし，200 エピソードの少し手前の時点で既に理論的にも最大の報酬を獲得することができている．このことから RS 大局基準としてタスク全体としての目標水準を与えるのみで，他のアルゴリズムと比較しても素早く良い行動系列を獲得していることがわかる．

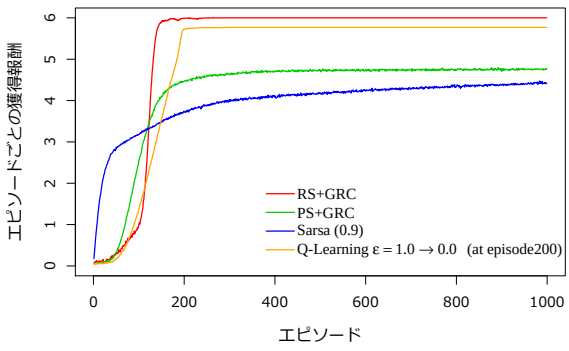

図 2: 獲得報酬の推移 (スイッチワールドタスク)

6. ブラインドスイッチワールドタスク

次にスイッチワールドタスクの設定を，エージェントがランプを観測することができない，すなわちスイッチを押したことを認識できないように変更したブラインドスイッチワールドタスクでシミュレーションを行う．これにより，エージェントはスイッチ状態の違いによる状態の分別ができなくなり，座標が同じであればすべて同じ状態として認識してしまうような状態の不完全知覚問題が発生する，部分観測マルコフ決定過程タスクとなる．

6.1 シミュレーションと結果

格子空間やスイッチの配置の設定は前タスク同様に図 1 のように設定をした．100 ステップを 1 エピソードとし，100,000 エピソード行い，そのシミュレーションを 1,000 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 $\epsilon = 0.1$ の ϵ-greedy を用い，それ以外は前タスクと同様のアルゴリズムを用いた．

シミュレーションの結果としてエピソードごとに得られた報酬の時間発展を図 3 に示す．まず，ϵ-greedy と PS の結果を見ると，どちらも全く報酬を獲得することができていないことが確認できる．それに対して Sarsa (λ) は，学習序盤から報酬を獲得することができており，その後も 1.0 前後の報酬の獲得をキープしている．このことから，ブラインドスイッチワールドタスクでは Sarsa (λ) の 1 ステップで過去の状態行動対の価値の更新を行うことできる特徴が報酬の獲得につながったと考えられ，タスクの成否には価値の更新タイミングが大きく影響することがわかる．一方で RS は学習序盤はほとんど報酬を獲得できていないものの，徐々に上昇してき最終的には Sarsa (λ) と同等の水準まで成績を伸ばすことができている．

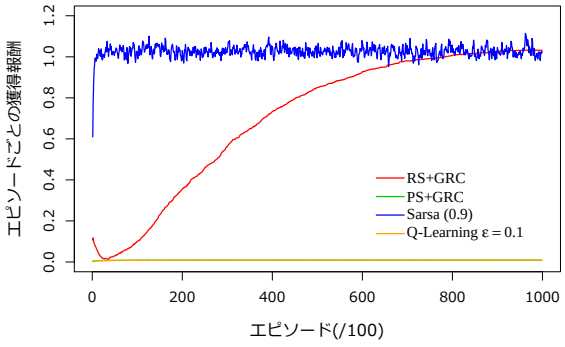

図 3: 獲得報酬の推移 (ブランドスイッチワールドタスク)

7. 非定常スイッチワールドタスク

より現実的な問題を考えると，大抵の日は同じ行動系列を行っていれば食糧を確保することができるが，環境は変化するものであり，突然食糧が得られなくなってしまうかもしれない．その場合，エージェントは獲得した行動系列を諦め，新たに一定以上の報酬の得られる行動系列を発見することが重要となる．そこで，一定のエピソード毎にエージェントが認識できない部分でスイッチの位置が入れ替わる (環境の変化が起こる) 非定常スイッチワールドタスクでシミュレーションを行う．

7.1 シミュレーションと結果

結果が解釈しやすいよう，格子空間は前タスクと同様に図 1 の格子空間を用いるが，スイッチの配置は異なる設定を行う．1 エピソードで得られる最大の合計報酬が変化しないようにし，かつ報酬を得るための最短ステップが 16 ステップで固定されるように，スイッチの位置を図 1 における最短経路となる正方形の経路の頂点として座標 $(1,1),(5,1),(1,5),(5,5)$ のいずれか 3 箇所に設置する．3 つのスイッチの位置は環境が変化するたびに，この 4 つの座標の中からランダムに配置される．しかし，これではスイッチの位置が変化しても，同じ正方形のループ経路を辿るだけで，最大ではなくとも報酬を獲得することは可能となってしまう．そこで，押すとすべてのスイッチがリセットされるような，リセットスイッチを新たに 1 つ

追加し，リセットスイッチを含めた 4 つのスイッチを先程の 4 箇所の座標にランダムに配置する．これにより，報酬を獲得するためには環境が変化するたびにリセットスイッチを避けるような新たな経路の発見が必要なタスクとなる．

100 ステップを 1 エピソードとし，1,000 エピソード毎にスイッチの位置を変化させる．これを 10,000 エピソードまで行い，そのシミュレーションを 1,000 回行った結果を平均した．比較アルゴリズムは，ブラインドスイッチワールドタスクで用いたアルゴリズムと同様のものを用い，ϵ-greedy の ϵ は $\epsilon = 0.1$ に設定し，GRC 手法におけるスケーリング変数 $\zeta(s_i)$ は経験的に成績の良かった $\zeta(s_i) = 1.0$ に設定した場合の結果を用いる．シミュレーションの結果としてエピソードごとに得られた報酬の時間発展を図 4 に示す．まず，ϵ-greedy の結果を見ていくと，環境が変化するたびに一度獲得報酬が下がった後，徐々に上昇していき，次の環境変化までに獲得報酬は 4.0 を超える水準に到達する結果となった．また PS を見ると，ϵ-greedy とほとんど同じ変化をしており，満足化する経路の発見に至らなかったために，ϵ-greedy による行動選択を行い続けていることがわかる．次に Sarsa (λ) の結果を見ると，ϵ-greedy と比較して環境が変化したタイミングからの報酬の上昇が速く，獲得報酬の到達地点もやや高い成績となっている．一方で RS は他アルゴリズムと比較して，環境が変化した直後の獲得報酬の上昇が速く，最終的に獲得している報酬も高いことから，良い経路を素早く再発見できていると言える．

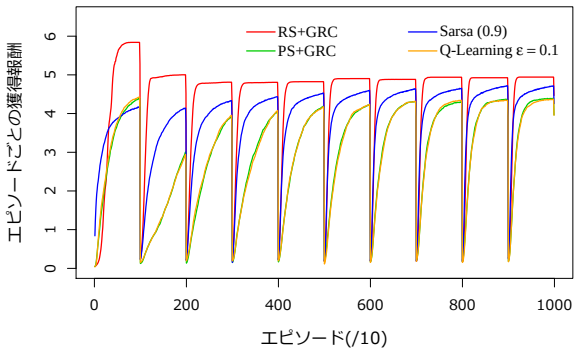

図 4: 獲得報酬の推移 (非定常スイッチワールドタスク)

8. 考察

ここでは，RS のどのような性質が高い成績につながったのか考察する．そのために，まずブラインドスイッチワールドタスクの結果に着目する．このタスクはスイッチが押されることに起因する状態の変化を観測できない部分観測マルコフ決定過程タスクであり，状態の不完全知覚問題が発生する．通常 TD 学習ではマルコフ性を前提にしたうえで価値の 1 ステップバックアップを繰り返すことを前提にしているため，状態の不完全知覚が発生する環境では正しい Q の更新が行われないため学習が困難となる．ただし，今回の設定では同じ観測の別状態に対して，同じ方向に移動する行動を選んでいれば良いため，報酬が得られる経路を見つけることは不可能ではない．そのため，スイッチ状態の変化による Q 値の更新の違いに関わらず，経路単位の学習をすることができれば報酬が得られると考えられる．まず，素早く報酬を獲得していた Sarsa (λ) について考える．Sarsa (λ) はあるステップでの価値の更新と同じタイミングで，過去の状態行動対の価値の更新に報酬情報を反映させるため，1 回のステップのみで経路単位で価値の更新を行っていると言える．そのため 1 ステップ分しか価値のバックアップをできない通常の Q 学習に比べ，経路としての学習を瞬時に行うことができ，それが報酬の獲得に繋がったと考えられる．しかし，Sarsa (λ) は 1 ステップごとに全状態行動対を更新するため，状態数が増えるほど 1 ステップにおける更新回数が増加する．一方で RS は各ステップにおいて 1 ステップのバックアップのみで，最終的には Sarsa (λ) と同様の成績を出している．これは，同じ満足化アルゴリズムである PS が獲得報酬を増やせていないことから，RS の信頼性を考慮した探索方法が学習の成功につながったと考えられる．RS の信頼性変数である $\tau(s_i, a_j)$ は，未来の行動の試行回数も考慮した行動系列としての信頼性を表し，非満足化時において $\tau(s_i, a_j)$ 値が小さいほど価値を高く設定するような価値付けを行うため，信頼性の低い試した回数が少ない経路を優先的に選択していく傾向をもつ．このように，RS は経路を考慮した探索ができていたことで，ランダム探索に依存する PS に比べて効率的な探索が行うことができ，徐々に獲得報酬が上昇したと考えられる．次に非定常スイッチワールドタスクの結果に着目すると，RS は環境の変化に対し素早く新たな経路を発見することができている．これは，RS が Q の大小関係だけでなく，大局期待値 E_G と大局基準値 R_G 値の差を考慮することで，得られる報酬の変化に対して敏感に探索度合いを変化させることができ，素早く再探索を行うことができたためであると考えられる．RS は満足化度合いである非負値の $E_G - R_G$ が大きいほど探索する傾向にあり，小さいほど Q 値に従った greedy 行動を多く行う傾向を持つ．そのため，環境が変化し報酬が得られなくなると E_G の値が減少し，$E_G - R_G$ の差が大きくなるため探索傾向が強くなる．さらに，信頼性の考慮による効率の良い経路の探索も加わり，環境の変化に対して素早く新たな経路を発見することができていたと考えられる．

9. 結論

本研究では，エージェントの生存を目的とした満足化強化学習の考案を目指した．そのために生存コストという一つの目標水準 (大局基準) を定め，それを各状態の基準に割り振る手法として大局基準変換法 (GRC) を考案し，満足化価値関数 RS への適用を行うことで，生存を目的とした満足化を強化学習へ実装した．その結果，GRC を適用した RS は一つ目標値として大局基準が与えられれば，素早く目的を達成するような行動系列の獲得が可能となり，強化学習における RS の有用性が示された．その要因として，RS の性質である目標の達成度合いを考慮した探索度合いの自律調節や，行動系列の信頼性を考慮した行動選択を行うことによる効率的な探索が学習の成功につながったことが示唆された．$\tau(s_i, a_j)$ の関数近似 (例えば疑似カウント) により深層強化学習にも応用可能である．

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. 1–11 (2016)

[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)

[Sutton 00] Sutton, R.S., Barto, A.G.,(三上貞芳　皆川邪章 共訳): 強化学習, 森北出版 (2000).