1M2-3

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－

Reinforcement Learning Method for Search Ratio Autonomous Adjustment: Online Update of Satisficing Aspiration Level

牛田 有哉 *1 Yuya Ushida　甲野 佑 *2 Yu Kohno　浦上 大輔 *3 Daisuke Uragami　高橋 達二 *1 Tatsuji Takahashi

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

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

*3日本大学生産工学部 School of Industrial Technology, Nihon University

In realistic reinforcement learning tasks, the abundance of possible actions and states make traditional optimization methods often inapplicable. We have proposed another, cognitive algorithm that implements satisficing behavior that explores for actions with values above the aspiration level. The algorithm is known to quickly optimize in bandit problems that deal with immediate rewards. In this study, we test it by a more general reinforcement learning task with delayed reward.

1. はじめに

環境と相互作用しながら行動系列を試行錯誤的に獲得する枠組みとして強化学習が存在する．強化学習は環境から与えられる数値化された報酬を手掛かりに環境に適応するための最適な行動の獲得を目的とするため，限られた時間内で学習を行うには新たな情報を得るための探索と，既に持っている知識から最良な行動を選ぶ知識利用を両立しなければならない．しかし探索と知識利用は一度に行うことはできないため学習は困難となり，複雑な環境になるほど最適な方策が得られるまでに時間がかかってしまう．しかし，そのような複雑な環境で生きる人間は情報収集とそのコストのバランスを上手くとりながら意思決定を行っていると考えられる。これは人間が満足化という最適化とは異なるルールで意思決定を行っているからであるとされる [Simon 56]．満足化は価値に対してある基準値を定め，そこを境目に探索と知識利用を切り替えることができ，満足化を応用した Reference Satisficing(RS) は単純な意思決定課題であるバンディット問題において基準値を適切に設定した場合に良い成績を示している [高橋 15]．しかし，一般的な強化学習では報酬が行動に対して遅れて返ってくるような遅延報酬の考慮が問題となるため最適行動の発見が困難になる．また，状態が複数の一般的な強化学習での最適な基準値は、バンディット問題での設定の方法が単純に適用できないため複数の状態を考慮して設定する必要がある．

そこで本研究では強化学習タスクにおける満足化の最適な基準値を設定し，RS が遅延報酬の問題に対して効率よく最適行動を見つけられるのかを検証する．

2. 強化学習とトレードオフ

多くの報酬を得るためには，エージェントは過去に試みた既にもっている情報の中から報酬の獲得につながる最良の行動を選択し続けなければならない (知識利用)．しかし，そのような行動を見つけるためには過去に試みたことのない行動も選択しなければならない (探索)．つまり，エージェントは報酬を得るために探索と知識利用の両方を行う必要がある．しかし，限られた時間内での学習では探索にコストを割けば知識利用を行う回数が減り報酬を得る機会を失ってしまうし，知識利用にコストを割けば他の良い行動を見つけられずに局所解に陥ってしまう．このように探索と知識利用は両立することが困難であり，この問題は探索と知識利用のトレードオフと呼ばれ強化学習の大きな問題の一つとなっている [Wickelgren 77]．環境が複雑になるほどこの問題は大きくなり，いかにして探索と知識利用の割合を調節して学習を効率よく行うかが課題となる．

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. バンディット問題における満足化

人間は意思決定において行動の価値にある基準値を定め，基準値を超えた価値を持つ行動が見つかるまで探索を続け，そのような行動が見つかったら探索を止め，その行動に執着するという傾向がある．このような意思決定傾向は満足化と呼ばれ最適化とは区別される [Simon 56]．強化学習は探索と知識利用のトレードオフの問題から，最適な方策を得るために多くの時間を必要とする．さらに適用する環境が複雑になれば現実的な時間内で最適な方策を見つけることは困難であり，実世界のような捉えきれない形で変化する環境では充分な探索が不可能であるため探索を打ち切る定義が難しい．それに対して満足化は，最適化が困難である場合においても基準値というパラメータによって探索の開始と停止の条件を明確に決められる利点が存在する．そこで本研究では満足化を価値関数として表した既存の満足化価値関数である RS に着目した．

3.1 RS

満足化価値関数 RS は価値 E_i と信頼度 n_i，基準値 R から以下の式で表される．この式は価値関数を表しているため ϵ-greedy などの方策と併用することができる利点がある．本研究では特に記述しない限り評価式の値に対して greedy な選択を行うものとする．

$$RS_i = n_i(E_i - R) \tag{1}$$

バンディット問題において，E_i は行動 a_i の報酬の標本平均，n_i はサンプルサイズ (試行回数) を表す．この $E_i < R$ となるような行動については，サンプルサイズが小さく信頼性の低い選択肢の価値を高くし，$E_i > R$ となるような行動について

は，サンプルサイズが大きく信頼性の高い選択肢の価値を高くする．このように設定することで探索の間はまだあまり試していない良く知らない行動を，知識利用の際はより信頼のできる行動を選択することができる．

3.2 バンディット問題における最適な基準値

RS において基準値 R の値をどのように設定するかという問題が存在する．適切な基準値はエージェントが適用する環境により異なると考えられるため，エージェント自身が観測できる情報から動的に設定することが理想的であると言える．しかし，そのためには目指すべき適切な基準値を知る必要がある．ここで，適切な基準値の定義が問題となる．満足化は最適化と異なる概念ではあるが基準値の設定によっては最適化と同義となることがわかっており，最適化が行える環境においてはそのような設定が適切な基準値であるといえる．バンディット問題では基準値 R が最適行動と次に良い行動の報酬確率の間に設定すれば基準値 R を超える行動は最適行動のみとなり最適化を行うこととなる．つまり最適基準は最適行動の報酬確率 P_1 と次に良い行動の報酬確率 P_2 を用いて以下の式で表される．

$$R_{\mathrm{opt}} = (P_1 + P_2)/2 \tag{2}$$

既存の研究ではバンディット問題において，この最適基準を用いることで RS が別のアルゴリズムに比べて素早く最適な行動を見つけることができることを示している [高橋 15].

4. 未来を考慮した意思決定の難しさ

最適基準を用いた RS はバンディット問題において高い成績を示した．しかし，より一般的な強化学習に適用した場合に同じく高い成績を出せるとは限らない．一般的な強化学習ではバンディット問題と異なり状態が複数存在する．そのため単純にその時の行動だけではなく，その後の行動系列を考慮しなくてはならない．それにより学習が困難になる理由として遅延報酬の発生が挙げられる．たとえ現在の行動で報酬が得られても，その後の行動で報酬が得られなければそれは良い行動と言えない．逆に現在の行動で報酬が得られなくても，その行動が後に大きな報酬につながる行動であるならばそれは良い行動と言える．このように最終的な収益を増やすためには目の前の報酬は小さくても後に大きな報酬につながる，急がば周れのような考え方が必要となってくる．つまり，強化学習では単純に行動の最適化をするのではなく未来を考慮した方策としてのの最適化が必要となる．そうなることで選択肢が行動のみのバンディット問題に比べ，一般的な強化学習では方策の選択肢が爆発的に多くなるため探索が非常に困難となる．本研究では遅延報酬の考慮という問題に対して最適基準を用いた RS がバンディット問題同様に素早く最適方策を見つけられるのかどうか検証を行い，強化学習における RS の有用性を検証する．

5. 強化学習への RS の拡張

RS を強化学習へ拡張するためには二つの問題が存在する．一つ目は基本的な強化学習手法である TD 学習では標本平均 E_i に対しては Q 値で置き換えることで適用できるが，信頼度であるサンプルサイズ n_i がそのまま使用できないことである．もう一つの問題は，最適基準が単純に行動の報酬確率で表せないことである．この二つの問題はそれぞれ以下のようにすることで解決される．

5.1 強化学習における信頼度

バンディット問題ではある行動における報酬の標本平均の信頼度として n_i で表すことができた．それに対し強化学習で行動価値関数として用いる Q 値はある方策 π に基づいて行動したときに得られる報酬の合計の期待値である．つまり強化学習における信頼度は，単純にその行動をした回数ではなく，方策 π の下でその行動をした回数を意味する．しかし方策 π は状態数が増えるにつれて増えていくため全ての方策に対する回数を数えるのは難しい．そのため強化学習における信頼度は Q 値の更新が未来の報酬を考慮するのと同じように現在の行動の信頼度に加え，その後の未来の行動系列の信頼度を考慮した信頼度変数として $\tau(s_i, a_j)$ で定義される．[甲野 15].

$$\tau(s_i, a_j) = \tau_{\mathrm{current}}(s_i, a_j) + \tau_{\mathrm{post}}(s_i, a_j) \tag{3}$$

$$\tau_{\mathrm{current}}(s_t, a_t) = \tau_{\mathrm{current}}(s_t, a_t) + 1 \tag{4}$$

$$\begin{aligned} \tau_{\mathrm{post}}(s_t, a_t) &= \tau_{\mathrm{post}}(s_t, a_t) \\ &+ \alpha\Big(\gamma_\tau \tau(s_{t+1}, a_{\mathrm{up}}) - \tau_{\mathrm{post}}(s_t, a_t)\Big) \end{aligned} \tag{5}$$

ここで，α は信頼度学習率であり，変化した行動系列の信頼度をどの程度反映させるかを意味し，γ_t は信頼度減衰率であり，その後の行動の信頼度をどの程度考慮するかを意味する．

5.2 強化学習における最適な基準値

バンディット問題と同様に，強化学習においても適した基準値は存在する．しかし，バンディット問題では行動に対して最適化を行ったのに対し，強化学習では方策に対して最適化を行わなければならない．そのため強化学習における最適基準は，最適な方策 π の下で獲得できる収益 $E^\pi = \sum_t r_t$ に対して一番高い値となる $E_{\mathrm{first}} = \max_{a,\pi} Q^\pi(s, a)$ と次に高い値となる E_{second} の間に設定すればよい．このことから RS で扱う満足化の最適基準を以下の式で与える．ただし，強化学習では状態ごとに最適な Q 値の値は異なるため適切な基準値も状態ごとに存在する．

$$R_{\mathrm{opt}}(s_i) = (E_{\mathrm{first}}(s_i) + E_{\mathrm{second}}(s_i))/2 \tag{6}$$

6. ツリーバンディットシミュレーション

本研究では強化学習における RS の評価を行うために，N 本腕バンディット問題をツリー構造状に拡張し，状態遷移を追加したタスクを用いた．エージェントは，まず通常の N 本腕バンディットを行い，報酬確率に従って 0 か 1 の報酬を得る．そして選択した腕に応じて状態遷移し次の N 本腕バンディットを行う．これを繰り返して終端状態まで行い，これを 1 エピソードとし報酬の累計が最大となるような経路をを見つけることを目的とする．本研究ではこのタスクをツリーバンディット問題と呼ぶ．ツリーバンディットをタスクとして用いた理由として，本研究では最適基準を設定した RS での検証を目的とするため，最適基準を計算がしやすいような，エピソード内に同じ状態を訪れることのない一方通行のタスクが望ましかったことがある．また，このタスクは強化学習の難しさである遅延報酬を考慮する必要があり，報酬が確率的であるためさらに最適経路を見つけるのが困難なタスクとなっている．図 1 を例として見ると，最適な経路は状態 1 で報酬確率 0.4 であるスロット 1 を選択し，遷移した先である状態 2 で報酬確率 0.9 のスロットを選択する経路である．しかし，この経路を見つけるためには状態 1 において報酬確率の低いスロットを選択しなければな

らない．つまり，目先の報酬のみに影響されずに一見報酬確率が低い経路をいかに探索するかがこのタスクの難しさとなる．さらに報酬が確率的であるため，より多くの探索が必要となり最適経路を見つけるのが困難なタスクであると言える．

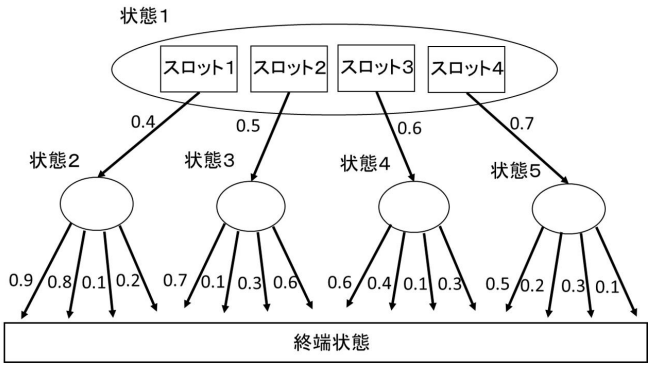

図 1: ツリーバンディット

6.1 設定

本シミュレーションは図 1 と同様の設定でシミュレーションを行った．1 つの状態につき 4 本腕のバンディット問題を行い，層の数が 2 層で状態数が 5 つのツリーバンディットである．報酬確率に関しても，図 1 と同様の値に固定した．報酬確率に固定値を扱った理由としては，本研究の目的である遅延報酬を考慮した学習が行えるかどうかを調べるために，毎エピソードで確実に遅延報酬の考慮が必要となる環境にしたかったためである．エージェントが終端状態にたどり着くまでを 1 エピソードとし 10,000 エピソード行こない，そのシミュレーションを 10,000 回行った結果を平均した．評価を行う指標としてはバンディット問題における後悔にあたる指標を用いた．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値報酬と実際に選んだ経路の累計期待値の差」が後悔の定義となる．比較に用いるアルゴリズムは代表的な強化学習アルゴリズムとして，方策オフ型 TD 学習である Q 学習と方策オン型 TD 学習である Sarsa 使用し，方策には，ランダム選択率 ϵ を 1.0 から一定の量を徐々に減らしていき 2,000 エピソードで 0 になるように設定した ϵ-greedy と，Q 値の更新量を利用して ϵ 調節するアルゴリズムで強化学習タスクにおいて優れた成績の出している適応性 ϵ-greedy(VDBE: Value-Difference Based Exploration)[Tokic 10]，最も単純な満足化アルゴリズムである素朴満足化ポリシー (PS: Policy Satisficing) を用いた．PS は信頼度の考慮がなく，単純に基準値を超える価値を持つ行動を見つけるまではランダム探索を行い，基準値を超える価値を持つ行動を見つけたらその行動を選択し続ける満足化アルゴリズムである．また，本研究では RS と PS における最適基準の設定を行うために正確な Q 値の更新が望ましかったことから，学習率 α に関しては $1/t$ で与え，割引率は $\gamma = 1.0$ とした．このとき，t はステップ数を意味する．RS と PS については最適基準 R_{opt} を与える．

7. 結果および考察

シミュレーションの結果として後悔の時間発展を，Q 学習を図 2 に，Sarsa を図 3 に示す．まず，ϵ-greedy と VDBE の結果から見ていく．ϵ-greedy は，Q 学習では良い経路を発見することができているものの，探索段階において探索割合を上手く調節できていないことから序盤に後悔が大きくなってしまっている．また，強化学習タスクにおいて一般に優れているとされる ϵ の自律調節を行う VDBE では，探索割合を上手く調節することで学習序盤での後悔は小さくなっているが，最終的に良い経路を発見できずに後悔が増加し続ける結果となった．これは，報酬が確率的であることや学習率 α を $1/t$ で与えたことにより，Q 値の更新量が小さくなってしまったために ϵ が早い段階でほぼ 0 になり間違った経路に収束してしまったからであると考えられる．つまり VDBE は α が一定以上の限られた環境でしか良い成績を出すことができないと言える．一方で満足化方策の 2 つを見てみると，Q 学習においてはどちらも後悔の値は最終的に一定となっていて最適経路を見つけられていることがわかる．Q 学習では，その時点で最も良い行動を選択しやすくなる．つまり，ある程度報酬の得られる経路を見つけてしまったら別の経路を探索しづらくなり最適経路を見つけるまでに時間がかかってしまう．さらにツリーバンディットシミュレーションにおいては最初の行動選択の際に高い報酬確

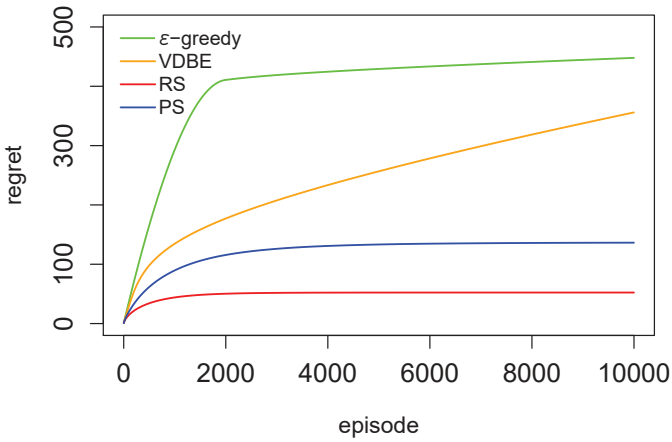

図 2: 後悔の時間発展 (Q 学習)

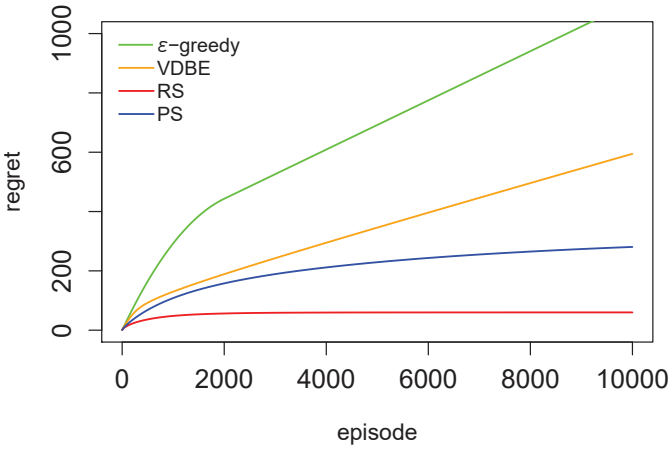

図 3: 後悔の時間発展 (Sarsa)

率の行動を選択してしまうと最適な経路は見つけられない設定を用いている．これにより先に見つける経路は間違った経路である場合が多く，再探索を行い最適経路を新たに発見しなくてはならないため，最適経路の獲得は困難であり時間を要するタスクになっている．それに対し，満足化アルゴリズムはその性質上，基準値を超えるような経路を発見するまで探索を行う．そのため最適基準を与えた PS や RS はある程度報酬が得られる経路を見つけたとしても，そこに執着することなく最適経路を目指して探索を行うことができる．また，PS と RS を比較した場合 RS の方が素早く最適経路を見つけられていることが確認できる．これは RS のもう一つの特性である信頼度の考慮が効率の良い探索につながったためであると考えられる．RS は基準値を下回る価値をもつ選択肢については信頼度である τ の値が小さいほど価値を高く設定する．そのためエージェントは試した回数が少ない経路に関して，「今まで試した結果高い報酬が得られなかったが，さらに試すことで基準値を超えているかもしれない」というような考えの下で探索を行っていく．この性質により基準値を超えるような経路を素早く発見できたのだと考えられる．さらに Q 学習と Sarsa の最終的な後悔の値を表 1 に示す．表 1 を見ると，ϵ-greedy，VDBE，PS では Q 学習に比べて Sarsa の後悔がそれぞれ約 2.47，1.67，2.06 倍となっており，かなり大きくなることが確認できる．しかし RS は Q 学習と比較して Sarsa は約 1.15 倍と他のアルゴリズムに比べて後悔が増加量は少ない．これは他のアルゴリズムがランダム探索を使用することで Sarsa では Q 値が最適な値に収束しないのに対して，RS は乱数を使用せずに信頼度考慮した選択をすることで適切な Q 値を素早く得ることができたからであると考えられる．現実のような環境でオンライン学習することを考えた場合，最適 Q 値の探索よりも方策として良いの経路を探索する方策オン型の学習が適していることもあり，RS の方策オンオフへの依存度が低いという特性は利点であると言える．

表 1: 学習終了時の後悔

	ϵ-greedy	VDBE	RS	PS
Sarsa	1106.378	594.605	59.572	280.717
Q 学習	447.8992	355.933	52.323	136.356
Sarsa/Q 学習	2.470	1.671	1.146	2.059

8. 満足化と基準値

今回のシミュレーションでは RS と PS に与える基準として予め設計者側が最適基準を知っていることが前提となっていた．しかし，本来は適切な基準値はタスクによって異なり，環境が複雑になるほど設計者が予め適切な基準値を知ることは困難な場合が多い．ゆえに基準値はエージェント自身が学習していく中で自律的に獲得していく必要がある．ただ，設計者が最適基準を知ることができるようなタスクにおいては，素早く最適経路を発見することができ，満足化は非常に有効であると言える．また，目的が基準値を満たす行動を見つけることであり，必ずしも最適化を目指すわけではない満足化という概念の観点からすれば，定めた基準値に対して基準値を超える行動を素早く見つけられるという点に関しては有益な結果である．これは最適化が困難な状況であった場合でも基準値を適切に獲得できれば未知の環境に素早く柔軟に適応することのできる可能性を示している．本研究では基準値を設計者が与えたが，適切な基準値が設定できれば高い成績を出せることを示せたことから，RS における課題点を基準値の動的設定方法という 1 点に落とし込むことができたと言える．また，基準値はエージェントの内的欲求を表していると解釈でき，設定方法については ϵ-greedy 法の ϵ などに比べると基準値パラメータ R は設計がしやすいという利点もある．

9. 結論

本研究では，人間の意思決定における特性を表した満足化価値関数である RS を一般的な強化学習タスクに応用し，遅延報酬や確率的な報酬の考慮が必要となるツリーバンディットタスクにおいてシミュレーションを行った．その結果，最適基準を与えた RS は素早く最適な経路を発見できることが確認でき，強化学習タスクにおける有用性を示した．これにより，適切な基準値の設定が行えれば素早く環境に適応できるアルゴリズムであると言え，問題を基準値の動的獲得の方法という点に絞ることができた．仮にエージェントが自ら適切な基準値を獲得できるようになれば，環境に応じて目標値を変更することのできる柔軟なアルゴリズムになると考えらる．このことから最適化が困難となる現実のような複雑な環境において，割くことのできる探索コストに合わせて目標を調節できる満足化は有効なアルゴリズムとなる可能性があると思われる．そのためにも今後はエージェントの基準値の動的獲得方法の考案が課題となる．

参考文献

[高橋 15] 高橋達二, 大用庫智, 甲野佑, 横須賀聡: 不確実性の下での満足化を通じた最適化, JSAI 2015 (2015 年度人工知能学会全国大会 (第 29 回)) 予稿集, 2D1-OS-12a-4in (2015).

[甲野 14] 甲野 佑, 高橋 達二 : 柔軟な意思決定機能のための認知特性の応用と検証, JSAI 2014(2014 年度人工知能学会全国大会 (第 28 回)) 予稿集, 2N5-OS-03b-2. (2014).

[Simon 56] Simon, H.A.: Rational choise and the structure of the enbironment, Psychological Review, 63, 261–273 (1956).

[篠原 07] 篠原修二, 田口亮, 桂田浩一, 新田恒雄: 因果性に基づく信念形成モデルと N 本腕バンディット問題への適用, 人工知能学会論文誌, 22(1), 58–68 (2007).

[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).

[甲野 15] 甲野佑，高橋達二: 満足化とその基準値の動的な更新による強化学習の促進，JSAI 2015(2015 年度人工知能学会全国大会 (第 29 回)). (2015)

[Wickelgren 77] Wickelgren, W.A.: Speed-accuracy tradeoff and information processing Dynamics, Acta Psychologica, 41, 67-85 (1977).

[Tokic 10] Tokic, M: Adaptive ε-greedy Exploration in Reinforcement Learning Based on Value Differences, KI'10 Proceedings of the 33rd annual German conference on Advances in artificial intelligence,203-210.