4S-01

非定常環境に適応する認知的満足化価値関数の提案

齋藤 建志 †　　高橋 達二 ‡
† 東京電機大学大学院　　‡ 東京電機大学

1 はじめに

人工知能分野で Alpha-Go[2] は囲碁において人間以上の成績を残して注目を集めたが、このモデルで用いられた技術は深層強化学習と呼ばれる強化学習の一種である。強化学習において探索空間が膨大だと最適な行動系列を現実的な時間で学習することが困難となる。そこで高橋らはハーバート・サイモンが提唱した限定合理性に着目し、基準を満たすことを目的とした戦略を満足化と呼び研究を行なっている [1]。特に、強化学習に適用可能な満足化のモデルを提案し、多腕バンディット問題において UCB1-tuned よりも早く学習できることが示された [1]。本研究では、高橋らが提案した満足化価値関数 RS をもとに非定常環境において有用なアルゴリズムを提案し、より実用的な満足化方策の実現を目指す。そして提案アルゴリズムの性能を評価するために非定常環境を想定した多腕バンディットシミュレーションを行い、提案したアルゴリズムの有用性を示す。

1.1 満足化価値関数 RS

高橋らは認知的な満足という意思決定特性を取り入れた満足化価値関数 (RS: Reference Satisficing) を以下のように定義した。

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

ここで、n_i は行動 a_i を試行した回数で、E_i は報酬平均、パイパーパラメータ R(基準値) である。そしてエージェントは常に最大の RS_i を持つ行動 a_i を選択する。RS は非満足状態 ($E_i < R$) であれば楽観的探索を行う。すなわち試行回数 n_i が少ない方が価値が高く過小評価しないようになっている。一方、満足時 ($R < E_i$) は悲観的利益追求を行う。これは試行回数 n_i が大きいほど価値を高く評価し、満足状態の確実性を試行回数で保証している。基準値 R が最大報酬確率 p_{first} とその次に高い報酬確率 p_{second} の間に設定されていれば、最適な行動獲得を行えるので $R_{\text{opt}} = \frac{p_{\text{first}}+p_{\text{second}}}{2}$ を最適切基準と呼び、これを用いた場合は RSopt と表記する。

Risk-Sensitive Satisficing in Unsteady Environments
†Kenshi Saito　‡Tatsuji Takahashi
†Graduate School of Tokyo Denki University
‡Tokyo Denki University

2 RSγ アルゴリズムの提案

RS アルゴリズムは最適切基準を設定することで定常環境下において理論的保証と高い性能をもつアルゴリズムである。しかし、RS 値は試行回数 n_i の増加によって ∞ もしくは $-\infty$ に発散してしまう。そのため、報酬確率が非定常な場合に満足できる腕への切り替えが遅くなる場合がある。さらに、満足化価値関数 RS が、人間や動物の満足するという性質をもとに提案されたことを考えると、それらと異なる点が見受けられる。特に信頼度として RS に用いられている試行回数 n_i は発散するが人間や動物の満足度合いは際限なく増加することは考えにくい。そこで過去の情報を忘却しつつ RS 値の発散を抑えるために、γ を導入して各ステップごとに全ての w_i, l_i を以下のように更新する。

$$w_i = \begin{cases} \gamma w_i + 1 & (a_i = a_{\text{select}} \wedge e) \\ \gamma w_i & (\text{otherwise}) \end{cases} \tag{2}$$

$$l_i = \begin{cases} \gamma l_i + 1 & (a_i = a_{\text{select}} \wedge \neg e) \\ \gamma l_i & (\text{otherwise}) \end{cases} \tag{3}$$

ここで、a_{select} はそのステップで選択した行動で e は行動 a_i をして報酬が獲得できたことを意味する。満足化価値関数 RS の更新において、(2)、(3) 式の処理を用いる価値関数を greedy に選択するアルゴリズムを RSγ アルゴリズムと呼ぶ。この時、全ての行動の集合を $\mathbf{A}$ とすると k ステップ目での全 w_i, l_i の和 $N_k := \sum_{i\in\mathbf{A}}(w_i + l_i)$ では、γ が条件 $-1<\gamma<1$ を満たすならば以下が成り立つ。

$$\lim_{k\to\infty} N_k = \frac{1}{1-\gamma} \tag{4}$$

したがって、-1 より大きく 1 未満の γ を用いて (2),(3) 式による更新を行うことで RS 値が ∞ もしくは $-\infty$ に発散することを防ぐことが可能である。さらに RSγ では (1) 式中の E_i を $E_i := w_i/(w_i + l_i)$ と定義することで、過去の獲得報酬情報を減衰させて過去情報に頼り過ぎないよう設定した。RSγ アルゴリズムで最適切基準を設定する場合は RSoptγ と表記する。しかし、最適切基準は事前情報を必要とするため計算が困難だといえる。そこで基準値を毎 step(5) 式のように更新し、動的に変化させる場合を RSγ と表記した。

$$R \leftarrow R + \alpha(E_i - R) \tag{5}$$

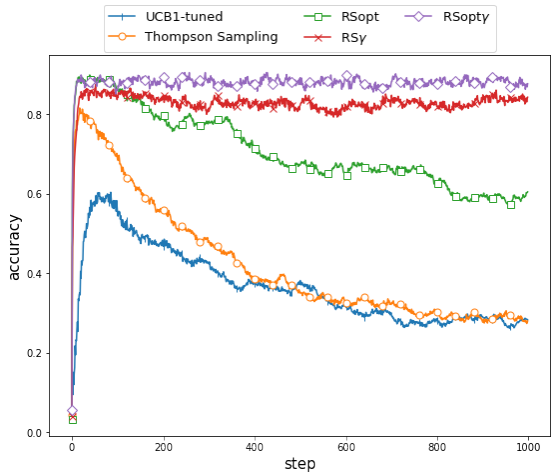

図 1: 確率的非定常 20 本腕バンディット問題正解率

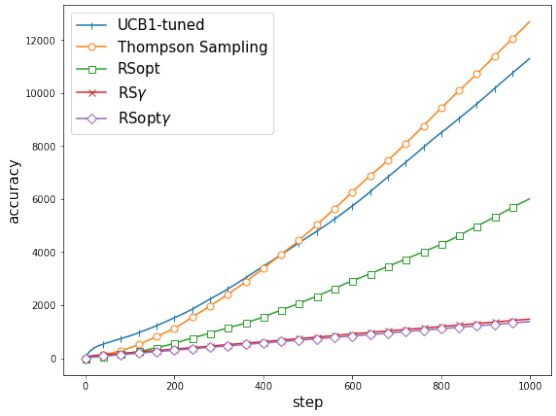

図 2: 確率的非定常 20 本腕バンディット問題 regret

3 確率的非定常シミュレーション

環境変化は特定の時間に発生するよりも確率的に発生すると捉えた方がより自然な場合がある。そこで、このような環境を再現するために一定の確率で環境変化が発生することを想定した多腕バンディットシミュレーションで UCB1-tuned[3]、Thompson Sampling アルゴリズム [4] との比較を行った。エージェントは各アルゴリズムにしたがって 100,000 回選択を行い、それを 1 シミュレーションとして 1,000 シミュレーションを行った際の平均値を指標として計算した。エージェントの行動は $|\mathbf{A}| = 20$ を想定し、各行動に対応する全ての報酬確率は毎 step ごとに確率 10^{-4} で一様分布から独立に再設定されるようにした。また RSγ アルゴリズムで用いるパイパーパラメータはそれぞれ、$\gamma = 0.999$、$\alpha = 5.0 \times 10^{-4}$、R の初期値 $R_0 = 1.0$ とした。

3.1 確率的非定常シミュレーション結果

Fig.1 に確率的非定常シミュレーションにおける正解率、Fig.2 には reget の変化を示した。Fig.1 をみると、UCB1-tuned,Thompson Sampling,RSopt はステップが進むにつれて正解率が低下している。一方で、RSγ と RSoptγ がステップが進んでもほぼ一定の正解率を保っていることがわかる。さらに、Fig.2 を見ると RSγ,RSoptγ は他のアルゴリズムよりも regret 増加が少なく、増加量がほぼ一定に抑えられている。これは、(2) 式と (3) 式による処理を加えて価値関数における過去の知識を圧縮して忘却のような働きをさせたため、突発的な環境変化に対してもエージェントが素早く対応できるようになった結果だと考えられる。

4 結論

確率的な環境変化を設定した非定常多腕バンディットシミュレーションの結果から、RSγ アルゴリズムは非定常環境を想定した場合で RS アルゴリズムや TS アルゴリズム、UCB1-tuned アルゴリズムよりも高い性能を有していることがわかった。また、漸進的に基準値を求める (5) 式を活用することで RSγ アルゴリズムでは事前情報を用いることなく非定常環境への適応を素早く行えることがわかった。したがって、RSγ アルゴリズムは非定常環境下で有用なアルゴリズムであることが確認できた。

参考文献

[1] 高橋 達二, 甲野 佑, 浦上 大輔, 認知的満足化, 人工知能学会論文誌, 2016, 31 巻, 6 号, p.AI30-M10

[2] David Silver et al. Mas- tering the game of Go with deep neural networks and tree search.Nature,Vol. 529, No. 7587, pp. 484-489, January 2016.

[3] Peter Auer et al. Finite-time analysis of the multi-armed bandit problem.Machine Learning, Vol. 47, No. 2, pp.235-256, May 2002

[4] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.Biometrika, Vol. 25, No. 3/4, pp. 285-294, 1933