The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

非定常環境における認知的満足化価値関数の適応性能

Adaptability of Cognitive Satisficing Algorithm in Nonstationary Environments

花安 勇人 *1 Yuto Hanayasu　齋藤 建志 *2 Kenshi Saito　吉井 佑輝 *1 Yuki Yoshii　甲野 佑 *1 Yu Kono　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 University

The environments where an agent performs trial-and-error learning is generally nonstationary because of unobservable information and various kinds of fluctuations. In order to make effective decisions in such an environment, the agent has to gradually or abruptly discard old information and put more weight on newer information, because some of the elements in the environment may have changed. As a result, there is a necessity of choosing a better option with smaller amount of information. We focus on the risk-sensitive satisficing (RS) algorithm which models the decision-making strategy of human beings and animals. We compare its performance in stationary and nonstationary bandit problems with other representative algorithms. We propose variants of RS combined with existing ideas for adaptation for nonstationary bandits such as meta-bandit and discounted update.

1. はじめに

近年，深層学習の目覚しい発展により，画像認識をはじめ現実での応用事例が増えてきている．深層学習は試行錯誤によって, 自律的に行動手順を学習するある種の万能性を有する強化学習 [Sutton 98] とも結びつき，従来困難だった人間レベルのビデオゲームのプレイを可能にした [Mnih 15]．しかしながら，強化学習の主要な問題である活用と探索のバランシングに関して，深層学習は直接的な解決策を与えていない．現在最適だと見積もられている行動を通して利益を追求する活用と，利益を捨ててでも真に最適な行動を探し出そうとする探索は同時に行うことができない．このような主体的な活用と探索が必要となる最も単純な課題の一つに多腕バンディット問題が存在する．多腕バンディット問題に対して探索と活用のバランシングを考慮し，適した試行錯誤を与える意思決定アルゴリズムは複数存在するが，基本的に環境が変化しない固定された定常環境を前提にしたものである．しかしながら, 往々にして現実の問題は非定常な環境である．そこで, 我々は実際に非定常な環境下で生きる人間の意思決定傾向を有する認知的満足化価値関数 (Risk Sensitive Satisficing Value Function: RS) を用いた選択アルゴリズム (RS アルゴリズム)[高橋 16] に着目した．本研究では非定常環境でのバンディット問題において RS アルゴリズムのバリエーションと既存のアルゴリズムを比較し，人間の意思決定傾向である満足化を利用することでの高い柔軟性を示す．

2. 多腕バンディット問題と非定常な環境

多腕バンディット問題ではエージェントが, 未知の報酬確率 $\{p_1, p_2, \ldots, p_k\}$ が割り当てられた行動 $\{a_1, a_2, \ldots, a_k\}$ の中から毎回一つ選択し，試行錯誤しながら得られる報酬を最大化することを目的とした最も単純な強化学習課題である．与えられる報酬が 1 か 0 のベルヌーイ試行からなるものをベルヌーイバンディットと呼び，本研究ではこれを扱う．現実には選んだ広告 (行動) がクリック (報酬) されるか否かという広告配信などの応用例と対応づけられる．多腕バンディット問題では最適な手段を知るために，現状は非最適な行動をあえて探索的に試行する必要がある．しかし前述の通り，高い累積報酬を得る (活用) ためにはどこかで探索を打ち切らなければならないという速さと正確さのトレードオフを端的に表した課題であり，探索と活用のバランシングが問題となる．バンディット問題において，このバランシングの良し悪しは，最適な行動を最初から知っていた場合に得られる報酬の累積の期待値との差分からなる regret で評価される．この regret が小さいほど，無駄なく報酬を最大化できたことを意味し，多くのバンディット問題のアルゴリズム (バンディットアルゴリズム) はこれの最小化を目指す．

$$\text{regret} = p^* N - \sum_{t=0}^{N} p_t \tag{1}$$

ここで変数 N は総試行回数，p^* は最も高い報酬確率であり，p_t は開始から t 回目に選択した選択肢の報酬確率である．Upper Confidence Bound (UCB) 系のアルゴリズム (UCB1-tuned[Wang 05], KL-UCB) や Thompson Sampling [Agrawal 12] など regret をなるべく小さくするバンディット問題に適したアルゴリズムが存在するが，これらは定常環境を前提としている (以下では UCB1-tuned を UCB1T，Thompson Sampling を TS とそれぞれ記載する)．しかし，現実的には対応すべき環境は非定常であり，既に最善の行動を見つけていても, 環境の変化によって探索を再開する必要がある場合を考慮しなければならない．これに対して複雑な準備をせず，迅速に対応するのは難しい．

3. RS アルゴリズム

人間の意思決定ではある基準を定め，基準を超える価値を持った選択肢を探索し続け，発見した場合は探索をやめて満足するという傾向がある．この意思決定における傾向を満足化と呼ぶ．満足化は，最も高い価値の選択肢を探索する最適化に比べ，基準の設定によって探索の打ち切りを早めることができるという利点がある．そのような基準の達成を目的に，観測された報酬期待値と思考回数から定義される認知的な価値関

連絡先: 高橋達二，東京電機大学理工学部，350-0394 埼玉県比企郡鳩山町大字石坂，049-296-1642, tatsujit@mail.dendai.ac.jp

数として，満足化価値関数 (Risk Sensitive Satisficing Value Function: RS) が考案されている [高橋 16].

$$RS_i = n_i\delta_i = n_i(E_i - \aleph) \tag{2}$$

ここで，n_i は行動 a_i を試行した回数, E_i は行動 a_i によって得た報酬の平均, $\aleph$ は満足化基準値である．そして, 常に最大の RS_i を持つ行動 a_i を選択する意思決定アルゴリズムを RS アルゴリズム (以下，断りがなければ RS はアルゴリズムのことを示す) と呼ぶ．RS は非満足状態 ($E_i < \aleph$) 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，満足状態 ($\aleph < E_i$) 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．また，$\aleph$ を最大の報酬確率 p_{first} とその次に大きい報酬確率 p_{second} の間に設定することで満足化は最適化となる．そのため，満足化基準値 $\aleph$ が以下のように設定された場合，最適切基準 $\aleph_{\text{opt}}$ と呼び，$\aleph_{\text{opt}}$ を用いた RS アルゴリズムを RS-OPT アルゴリズムと呼ぶ.

$$\aleph_{\text{opt}} = \frac{p_{\text{first}} + p_{\text{second}}}{2} \tag{3}$$

しかし，$\aleph_{\text{opt}}$ は報酬確率が既知のものとしているため活用は困難である．そこで，以下のように初期値 $\aleph_0$ から基準値を更新することで条件 (5) を目指す.

$$\aleph = \aleph + \alpha(E_i - \aleph) \tag{4}$$

$$P_{\text{first}} > \aleph > P_{\text{second}} \tag{5}$$

式 (4) によって求められる基準値 $\aleph$ は $a_{\text{select}} = a_{\text{first}}$ であるならば P_{second} を超えて条件式 (5) を満たすことができる．しかし， 式 (4) が条件式 (5) を満たす保証はないことに注意しなければならない.

4. 非定常環境に対処するアルゴリズム

環境が変異したという直接的な情報をエージェントが得られない場合，それはエージェントにとって非定常な環境であると言える．前述の UCB1T や TS は定常環境を前提として regret を最小化するアルゴリズムであり，一般に非定常環境には対応できない．そこで環境から得られる間接的な情報である報酬から環境の変化を察知したり，過去の情報の廃棄 (忘却) を用いて非定常環境に対応するアルゴリズムが考案されている.

4.1 メタバンディットアルゴリズム

非定常な多腕バンディット問題に対しては, 一般的にメタバンディットアルゴリズムが用いられる [Hartland 06]．メタバンディットは環境の変化の検出と，検出後の処理の組み合わせからなる．環境の変化は下記の Page-Hinkley 統計量 PH_T より, 最も高い価値をもつ選択肢を選び続けた結果, 得られた報酬の獲得割合の変化から検出する (PH_T が閾値 λ を超えると環境が変化したと判断する).

$$\overline{r_t} = \frac{1}{t}\sum_{l=1}^{t} r_l \tag{6}$$

$$m_T = \sum_{t=1}^{T}(r_t - \overline{r_t} + \delta) \tag{7}$$

$$M_T = \max\{m_t, t = 1 \ldots T\} \tag{8}$$

$$PH_T = M_T - m_T \tag{9}$$

$$\text{Change alarm} = \begin{cases} \text{true} & (PH_T > \lambda) \\ \text{false} & (\text{otherwise}) \end{cases} \tag{10}$$

そして変化が検出されると “従来の観測情報を記憶している旧エージェント” を保持したまま “観測情報を初期 step の状態に戻した新規エージェント” を生成する．また，新旧どちらのエージェントで意思決定を行うかを選ぶ上位エージェントを生成する．その後 L step の間，上位エージェントは旧エージェントと新規エージェントのどちらで意思決定を行うかを選び，選ばれたエージェントはこの環境に対して意思決定を行う．この L step の間をメタバンディット期間と呼ぶ．そしてメタバンディット期間終了後，旧エージェントと新規エージェントで高い報酬を得られていた方を残し，一方のエージェントと上位エージェントを破棄する．そして環境の変化の検出を再開する．メタバンディット期間に実際の環境に対して選択肢を試行して得られた報酬情報は新旧エージェントの両方に共有される．両者の違いはそれ以前の観測情報を持っているか否かである．また，上位エージェントには, 新旧どちらのエージェントが意思決定を担い，報酬を得ることができたかを記憶していく．本研究では UCB1T と TS, RS, RS-OPT アルゴリズムにメタバンディットアルゴリズムを適用して意思決定を行う (Meta UCB1T，Meta TS， Meta RS， Meta RS-OPT).

4.2 忘却率付き RS アルゴリズム

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，人間の満足化を目指す意思決定傾向から考案されたことから考えると，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, 人間の満足度合いが際限なく増加することは考えにくい．そこで，n_i の更新に忘却率 γ を導入して過去の情報を忘却しつつ, RS 値の発散を抑えるために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 [甲野 13].

$$w_i = \begin{cases} \gamma w_i + 1 & (a_i = a_{\text{select}} \wedge e) \\ \gamma w_i & (\text{otherwise}) \end{cases} \tag{11}$$

$$l_i = \begin{cases} \gamma l_i + 1 & (a_i = a_{\text{select}} \wedge \neg e) \\ \gamma l_i & (\text{otherwise}) \end{cases} \tag{12}$$

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 をして報酬を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，式 (11)，(12) で更新された信頼度 $n_i = w_i + l_1$ を用いた満足化価値関数に対して最大の選択肢を選ぶアルゴリズムを RSγ アルゴリズムと呼ぶ．また，同様に期待値 E_i に関しても忘却率を伴った更新 w_i, l_i によって更新された $E_i = w_i/(w_i + l_1)$ を用いて計算する．そして，RSγ との比較のためにこの忘却率を用いて減衰した w_i, l_i を用いてベータ分布のハイパーパラメータを更新する Thompson Sampling を新たに TSγ として定義する．この時，全ての行動の集合を $\mathbf{A}$ とすると k ステップ目での全体の和 N_k は以下のようになる.

$$N_k := \sum_{i \in \{i | a_i \in \mathbf{A}\}} (w_i + l_i) \tag{13}$$

$$\begin{aligned} N_{k+1} &= \gamma w_1 + \gamma l_1 + \gamma w_2 + \gamma l_2 + \cdots + \gamma w_{|\mathbf{A}|} + \gamma l_{|\mathbf{A}|} + 1 \\ &= \gamma N_k + 1 \end{aligned} \tag{14}$$

忘却率 γ が条件 $-1<\gamma<1$ を満たすならば以下が成り立つ．

$$\lim_{k\to\infty} N_k = \frac{1}{1-\gamma} \tag{15}$$

したがって，式 (11)，(12) 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．本研究では参考文献 [甲野 13] を参考に，忘却率 $\gamma = 0.999$ とした．

5. バンディットシミュレーション

以降, 式 (4) で算出される基準値を使用する満足化価値関数 RS を RS，RS を用いたメタバンディットアルゴリズムを Meta RS と表す．パラメータは $\lambda = 30$, $\delta = 0$, $\gamma = 0.999$, $\alpha = 0.001$, $\aleph_0 = 1.0$ と設定し, UCB1T, TS, RS, RS-OPT, TS γ, RS γ, RS-OPT γ, Meta UCB1T(L=500), Meta TS(L=30), Meta RS(L=30), Meta RS-OPT(L=30) を以下の多腕バンディットシミュレーションで比較する．性能の指標としては，最も高い行動の報酬確率と選んだ行動の真の報酬確率の差である regret を用いる．

5.1 定常環境でのシミュレーション

環境が変化しない多腕バンディット問題で各種アルゴリズムを比較する．選択アルゴリズムに従い, 選択を総 step 数として 100,000 steps 行い，その 1,000 シミュレーション回分を平均して regret を算出した．選択肢は 20 通りあり，全ての選択肢の真の報酬生起確率は，シミュレーション開始時に一様分布から独立に決定されて以降 100,000steps の間は変化しない．1,000 シミュレーションのあいだ毎回，全ての選択肢の真の報酬生起確率はサンプリングし直される．

5.1.1 定常環境でのシミュレーションの結果と考察

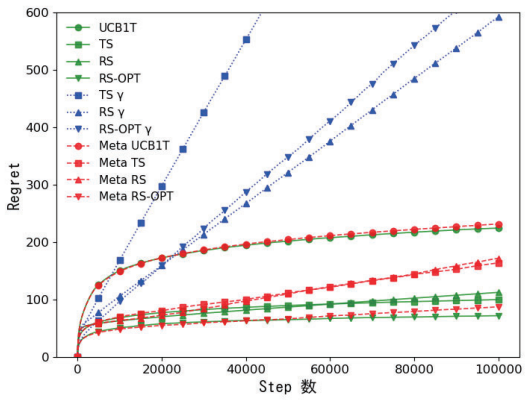

図 1: 定常環境 20 本腕バンディット問題での regret の推移

図 1 は定常環境での regret の推移を示している．忘却率 γ 用いていないアルゴリズムは regret が 250 以下に収まっているのに対し，忘却率 γ を用いたアルゴリズムは 500 以上まで伸びてしまっている． このことから忘却率 γ が定常環境での TS, RS, RS(OPT) の性能を低下させていることが分かる．一方で，Meta UCB1T, Meta TS, Meta RS, Meta RS-OPT は環境の変化の検出がなければ， それぞれ UCB1T, TS, RS, RS-OPT と同じ振る舞いをするが， 今回のシミュレーションでは環境の変化の誤検出を行ってしまい，regret が増えてしまっている．また, Chernoff Bound を背景に持ち，比較的優れている TS よりも最適切基準値を有する RS-OPT の方が優っている [Tamatsukuri 18]．Meta TS，Meta RS-OPT のシミュレーション結果も見ると，RS-OPT のこのような達成したい基準というただ一つのパラメータを与えるだけで優れた成績を示すのはメタバンディットアルゴリズムに用いた場合も変わらないことが示された．また, RS-OPT のシミュレーション結果より, RS の動的基準値更新を改善し, 今よりもう少し最適に近い基準を設定することができれば, 定常環境において TS を上回ることが期待できる．

5.2 非定常環境でのシミュレーション

環境が不定期的に変化する多腕バンディット問題で各種アルゴリズムを比較する．選択アルゴリズムに従い, 選択を総 step 数として 100,000 steps 行い，そのシミュレーション 1,000 回分を平均して regret を算出した．選択肢は 20 通りあり，全ての選択肢の真の報酬生起確率は，シミュレーション開始時に一様分布から独立に決定される．定常環境シミュレーションとの違いは，シミュレーション開始から，step 毎に 1/10000 の変異確率で各選択肢の真の報酬生起確率が一様分布から独立に再設定されることである．選択肢ごとの変異が独立で非同期なため, 各エージェントは選択肢の変化を直接的に観測することはできず，また，そのタイミングを学習することができない．

5.2.1 非定常環境でのシミュレーション結果と考察

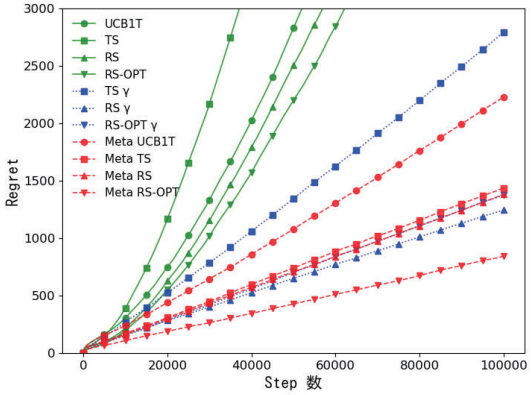

図 2: 非定常環境 20 本腕バンディット問題での regret の推移

図 2 は非定常環境での regret の推移を示している．UCB1T, TS, RS, RS-OPT は環境の変化に適応できず, regret が 6000 以上に跳ね上がってしまっている．また, Meta UCB1T と TS γ も regret が 2000 以上まで増えていて, 優秀とは言えない．一方で, RS γ, RS-OPT γ, Meta TS, Meta RS, Meta RS-OPT は regret が 1500 を下回っている．定常環境では, regret が多かった RS γ だったが, 非定常環境では, 最適切基準値を与えないアルゴリズムの中で最高の成績を残している．また，定常環境では, Meta RS は Meta TS より regret が多かったが, 非定常環境では, その関係が逆転している．さらに Meta RS-OPT と Meta TS の違いは定常環境と同じく，ただ一つの達成したい基準値パラメータ (最適切基準値) を与えるのみだが，非定常環境でも優れた成績を示せた．

5.3 定常環境と非定常環境での成績の総合考察

図 3，表 1 は各アルゴリズムの定常環境，非定常環境の 100,000 step 時点での regret の比較を示している．

図 3 では左下に近いほど，当該のアルゴリズムが定常，非定常環境の両方で優れていることを示す．Meta TS と Meta

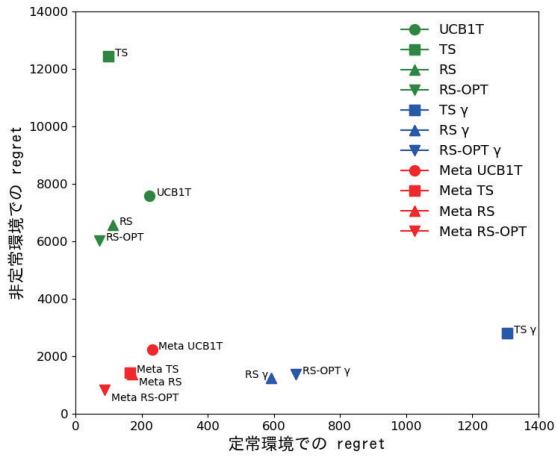

図 3: 100,000 step 時点での定常環境 (横軸)・非定常環境 (縦軸) regret 比較図

表 1: 100,000 step 時点での regret

	定常環境	非定常環境
UCB1T	224.6	7576.7
TS	100.0	12436.1
RS	112.5	6574.8
RS-OPT	71.9	6016.6
TS γ	1306.1	2793.5
RS γ	592.4	1246.0
RS-OPT γ	665.5	1375.7
Meta UCB1T	231.7	2228.9
Meta TS	164.2	1436.7
Meta RS	172.0	1381.7
Meta RS-OPT	87.6	842.7

RS がほぼ互角の水準であり，Meta RS-OPT がさらにそれに優れた性質を有している．これはメタバンディットアルゴリズムの定常環境のときは定常を考慮したアルゴリズムとして振る舞い， 非定常環境のときは非定常を考慮したアルゴリズムとして振る舞う柔軟性を示している．一方で, RS γ，および RS-OPT γ は常に忘却するという定常環境では不利なアルゴリズムではあるが，TS γ に対しては 定常環境，非定常環境での成績を比較的両立していると言える．これは人間の意思決定傾向を背景に持つ RS の柔軟性を示しており，メタバンディットアルゴリズムと比較して, 忘却率付き更新の容易さから，この点も RS の利点だと言える．

6. おわりに

以上の結果より，RS アルゴリズムが定常，非定常環境において有用であることが分かった．最適切基準値を与えないアルゴリズムの中で, 定常環境では TS ，非定常環境では RS γ が最も優秀な結果を残したが， Meta RS, Meta TS は一組のパラメータで定常，非定常問わず，優秀な結果を残した．現実の問題では環境が非定常であることが多く，環境の変化のタイミングや頻度を予測できない場合も多くあるため, この二つのアルゴリズムは現実の問題に対しても有用であると言える．また, 達成したい基準値パラメータ (最適切基準値) が与えられていれば, 定常・非定常問わず, 最も優秀な成績を残した Meta RS-OPT は最適切基準値をある程度予測できる場合には現実の問題にも有用であり, 動的に満足化基準値を設定する Meta RS や RS の潜在能力を示していると言える．

参考文献

[Agrawal 12] Agrawal, S., Navin Goyal, N.: Analysis of thompson sampling for the multi-armed bandit problem. In: *Proceedings of the 25th Annual Conference on Learning Theory* (COLT). (2012)

[Hartland 06] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag, Multi-armed bandit, dynamic environments and meta-bandits, In: *Advances in Neural Information Processing Systems(NIPS-2006) Workshop, Online Trading Exploration Exploitation*. (2006)

[Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Hassabis, D., et al.: Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533. (2015)

[Sutton 98] Sutton, R. and Barto, A.: *Reinforcement Learning: an Introduction*, MIT Press. (1998)

[Tamatsukuri 18] Tamatsukuri, A. and Takahashi, T: Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function, arXiv:1812.05795. (2018; manuscript under revision)

[Wang 05] S. Gelly, Y. Wang., R. Munos. and O. Teytaud.: Modification of UCT with Patterns in Monte-Carlo Go, *INRIA Technical Report*, No.6062. (2005)

[甲野 13] 甲野佑, 高橋達二, 価値推論ヒューリスティクスとしての規準学習と忘却, In: *Proceedings of 30 th Japanese Congnitive Science Society (JCSS)*, 74–79. (2013)

[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, 31(6), AI30-M_1–11. (2016)