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

認知的満足化価値関数の分析：保証付き満足化と有限 regret

Analysis of cognitive satisficing value function: Guaranteed satisficing and finite regret

玉造 晃弘 *1　　高橋 達二 *2
Akihiro Tamatsukuri　　Tatsuji Takahashi

*1東京電機大学大学院先端科学技術研究科
Graduate School of Advanced Science and Engineering, Tokyo Denki University

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

As the domains of reinforcement learning become more complicated and realistic, standard optimization algorithms may not work well. In this paper we introduce a simple mathematical model called *RS* (reference satisficing) that implements a satisficing strategy that look for actions with values above the aspiration level. We apply it to K-armed bandit problems. If there are actions with values above the aspiration level, we theoretically show that *RS* is guaranteed to find these actions. Also, if the aspiration level is set to an "optimal level" so that satisficing practically ends up optimizing, we prove that the regret (the expected loss) is upper bounded by a finite value. We confirm these results by simulations, and clarify the effectiveness of *RS* through comparison with other algorithms.

1. はじめに

強化学習の対象範囲が現実世界に拡大するに伴って探索空間は急激に膨張しており，真に最適な行動は現実的に不可能な場合が多い．そこで我々は，ある基準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り次第探索を止め，その行動で満足する，という満足化（satisficing）[Simon 56] の戦略に着目した．我々は，先行研究では価値関数のレベルで満足化を組み込んだ満足化価値関数 *RS* というモデルを提案し，シミュレーションにより経験的に有効性を確認した [高橋 16].

本論文で，我々はこの *RS* を最も簡単な強化学習問題である K 本腕バンディット問題に適用した場合についての理論的解析を試みる．具体的には最初に満足化の理論的保証を示す．つまり，十分な回数を試行すれば満足化基準を超える行動を安定して選択できるようになるという性質を理論的に示す．次に regret の有限性を示す．一般的に K 本腕バンディット問題のアルゴリズムの性能は regret と呼ばれる期待損失をどれだけ小さくできるかで表される．regret は試行回数に対して少なくとも対数オーダーであることが理論的に知られている [Lai 85]. 一方，*RS* では報酬分布の一部の情報を利用できると仮定すれば，満足化の結果が最適行動の選択（つまり最適化）を意味するようになり，regret は無限大に発散せず有限の値で抑えられる (上に有界になる) という注目すべき性質を示す．これらの性質については，シミュレーションによっても成立を確認する．また，*RS* と他のアルゴリズムとの比較も行う．

2. K 本腕バンディット問題

本論文では報酬確率にベルヌーイ分布を仮定した 2 値の K 本腕バンディット問題を扱う．エージェントには未知の報酬確率 $\{p_1, p_2, \ldots, p_K\}$ に従って，報酬 0 または 1 をもたらす，K 種類の行動 $\{a_1, a_2, \ldots, a_K\}$ があるとする．アルゴリズムの性能を評価する指標として期待損失を表す regret がある．最大の報酬確率をとる行動の添え字を i^* とする (i.e. $p_{i^*} = \max_i p_i$) と，n step 目 (n 回目の試行) の終了時点での regret は次のように定義される．

$$\mathrm{regret}(n) = \sum_{i=1}^{K} (p_{i^*} - p_i) E[n_i(n)]. \tag{1}$$

ここで $n_i(n)$ は n step 目の終了時点までの行動 a_i の選択回数である (step 数を明示しない場合は単に n_i と書く)．$E[\cdot]$ は期待値である．行動 a_i の基本的な価値付けは次で定義される報酬平均 E_i である．

$$E_i = n_i^1 / (n_i^1 + n_i^0). \tag{2}$$

ここで n_i^r は，行動 a_i を行って報酬 r を得た回数である．行動 a_i を選択した回数 n_i は $n_i = n_i^1 + n_i^0$ を満たし，$n = \sum_{i=1}^{K} n_i$ である．一般に価値関数が最も高くなる行動を選ぶ方法を greedy 法という．単に過去の知識を greedy に利用する「利益追求」だけでなく他の行動の価値も試すための「探索」も必要となる．

3. 満足化のモデル

3.1 素朴満足化ポリシー (PS)

満足化は最適化と違い，全ての行動を探索して最適な行動を決める必要がない点で探索のコストを下げることができる．これを強化学習のポリシーとして定式化すると，一つでも行動の報酬平均が基準 R を超えていれば greedy に知識利用を，そうでなければ (全ての行動の報酬平均が基準を下回っていれば) ランダムに選択して探索を，行うこととなる．これを素朴満足化ポリシー (PS: policy satisficing) と呼ぶ．

3.2 満足化価値関数 (RS)

基準との比較は行動 a_i の報酬平均 E_i と満足化の基準 R との差 $\delta_i = E_i - R$ が使われる．

正の δ_i が存在すれば，エージェントは満足してそのような a_i を選択し，そうでなければ不満足である．そこで基準満足

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石坂, 049-296-5416, tatsujit@mail.dendai.ac.jp

化価値関数 (RS: reference satisficing) を次のように定義する [高橋 16, Oyo 17].

$$RS_i = n_i\delta_i = n_i(E_i - R). \tag{3}$$

エージェントは最大の RS_i 値を持つ行動 a_i を選択するものとする．この形式は，次の 2 つの合理的な満足化行動を統合している．不満足の時は楽観的探索を行う．つまり全ての行動について $\delta_i < 0$ であれば，より選択回数 n_i が小さい行動がまだ試す価値のあるものとして優先される．満足であれば RS は悲観的知識利用を行う．δ_i を正とする行動が複数あれば，より確かなものを求めて選択回数 n_i が大きいものが優先される．

3.3 満足化基準の決め方

基準 R はエージェントの内的な必要性またはその環境についての知識に依存する．R の値が最適行動と次善行動の 2 つの間であれば，R を超える満足化は単純に最適化である．このように最適化は満足化の特別な場合を見なすことができる．この時の R を「最適切基準」と呼ぶ．報酬確率の最大値を p_1，2 番目の値を p_2 とすれば，もっとも単純には次のように設定すれば R は最適切基準となる：

$$R = (p_1 + p_2)/2. \tag{4}$$

なお，あらかじめ最適切基準を設定することは K 本腕バンディット問題の答えを最初から知っていること同じことではないか，という疑問が出てくるかもしれない．しかし，報酬確率の最大値と次に大きい値の間の 1 点の値が分かったとしても通常はどの行動が最適なのかその情報だけでは直ちには分からず，それを効率的に知る方法は決して自明ではない．

4. 理論的な解析

RS の基本的な性質について理論的な解析を行う．

4.1 満足化の理論的な保証

以降の証明では記述を明確にするため，step 数 s と選択行動 a_i を明示した記号を用いる．いずれも s step 目の終了時点での値を表すものとして，報酬平均は，$E(a_i, s) = n_i^1(s)/n_i(s)$，$RS$ の式は，$RS(a_i, s) = n_i(s) \cdot \big(E(a_i, s) - R\big)$ と書ける．

命題 1.（満足化の理論的な保証）行動 a_i の報酬確率を p_i とする（$i = 1, 2, \ldots, K$）．報酬確率が満足化基準 R 以上の行動の集合を A_U，小さい集合を A_L とする．つまり，$I_U = \{i \mid p_i \geq R\}$，$I_L = \{i \mid p_i < R\}$ として，$A_U = \{a_i \mid i \in I_U\}$，$A_L = \{a_i \mid i \in I_L\}$ とする．ただし，A_U は空集合でないと仮定する．このとき，RS では次が成り立つ．

> 「十分な回数を試行すれば必ず満足化基準 R より報酬確率の大きい行動の集合の中から行動を選択し，この状態は安定である.」

この内容は次のように定式化できる．$P(A)$ を事象 A が起きる確率として，

$$P\Big(\arg\max_{a_i} RS(a_i, s) \in A_U\Big) = 1 \ \ (s \to \infty). \tag{5}$$

Claim を二つ示したのちに，命題 1 を証明する．以降では $N_j = \Big\{s \;\Big|\; \arg\max_{a_i} RS(a_i, s) = a_j\Big\}$ とし，行動 a_j が選択される step の集合を表す．$\#N$ で集合 N の個数を表す．

Claim A.

$$\begin{aligned} &i \in I_L \text{のとき}, \\ &P\big(\#N_i = \infty \Leftrightarrow RS(a_i, s) \to -\infty \ \ (s \to \infty)\big) = 1. \end{aligned} \tag{6}$$

証明. (Claim A) $i \in I_L$ のとき，$RS(a_i, s) \to -\infty \ \ (s \to \infty)$ を仮定する．もし $\#N_i < \infty$ となれば，ある番号以上の s において $RS(a_i, s)$ が定数となり矛盾するため，$\#N_i = \infty$ が成立．

逆に $\#N_i = \infty$ のとき，大数の法則より任意の正の数 ϵ に対し，ある S があり，任意の S より大きい整数 $s \in N_i$ に対し，$P\big(|E(a_i, s) - p_i| < (R - p_i)/2\big) > 1 - \epsilon$ とできる．$|E(a_i, s) - p_i| < (R - p_i)/2$ ならば，$RS(a_i, s) = n_i(s) \cdot \big(E(a_i, s) - R\big) < n_i(s) \cdot \big(p_i + (R - p_i)/2 - R\big) = n_i(s) \cdot (p_i - R)/2 < 0$ となる．$s \to \infty$ とすれば，$n_i(s) \cdot (p_i - R)/2 \to -\infty$ となるため，$P\big(RS(a_i, s) \to -\infty \;\big|\; \#N_i = \infty\big) > 1 - \epsilon$. したがって，$\epsilon$ は任意より $P\big(RS(a_i, s) \to -\infty \;\big|\; \#N_i = \infty\big) = 1$. □

Claim B.

$$\text{ある } i \in I_U \text{について } P(\#N_i = \infty) = 1. \tag{7}$$

証明. (Claim B) 任意の $i \in I_U$ に対して $\#N_i < \infty$ であるとする．このとき，任意の $i \in I_U$ に対して，ある番号以上の s について $RS(a_i, s)$ は定数である．またこのとき，いずれかの $j \in I_L$ について $\#N_j = \infty$ となる．Claim A により，$P\big($ ある $j \in I_L$について $RS(a_j, s) \to -\infty \;\big|$ 任意の $i \in I_U$に対して$\#N_i < \infty\big) = 1$ となる．ところが，$RS(a_j, s) \to -\infty$ であることと，任意の $i \in I_U$ に対して $RS(a_i, s)$ がある番号以上で定数であることは互いに矛盾するため，$P\big($ ある $j \in I_L$について $RS(a_j, s) \to -\infty$, 任意の $i \in I_U$に対して$\#N_i < \infty\big) = 0$ でもある．したがって，確率の積の公式より P(任意の $i \in I_U$に対して$\#N_i < \infty) = 0$ でなければならない. □

証明.（命題 1）Claim B より，ある $k \in I_U$ について $\#N_k = \infty$ と仮定してよい．大数の法則より任意の正の数 ϵ に対し，ある S があり，任意の S より大きい整数 $s \in N_k$ に対し，$P\big(|E(a_k, s) - p_k| < (p_k - R)/2\big) > 1 - \epsilon$ とできる．$|E(a_k, s) - p_k| < (p_k - R)/2$ ならば，

$RS(a_k, s) = n_k(s) \cdot \big(E(a_k, s) - R\big) > n_k(s) \cdot \big(p_k + (R - p_k)/2 - R\big) = n_k(s) \cdot (p_k - R)/2 \geq 0$ となる．

よって，$P\big($十分大きい s について $RS(a_k, s) > 0\big) > 1 - \epsilon$. ϵ は任意であるから $P\big($十分大きい s について $RS(a_k, s) > 0\big) = 1$. ここで，もし，いずれかの $i \in I_L$ について，$\#N_i = \infty$ となれば，Claim A より $RS(a_i, s) \to -\infty$ と仮定してよい．ところがこれは，十分大きいすべての s について $RS(a_k, s) > 0$ となる条件と矛盾する．これは，任意の $i \in I_L$ に対して $P(\#N_i < \infty) = 1$ を意味する．したがって，

ある $k \in I_U$ について $P(\#N_k = \infty) = 1$ でかつ すべての $i \in I_L$ について $P(\#N_i < \infty) = 1$ が示せたから， 式 (5) が明らかに従う. □

4.2 regret の理論解析

基準 R を最適切な範囲に設定した RS では regret は有限な値で抑えられることを示す.

命題 2. RS の regret の有限性

全ての行動で報酬確率 p_i が最大となるものを p_1，報酬確率が 2 番目のものを p_2 とする．更に $R = (p_1 + p_2)/2$ となるように R を設定しておく（最適切基準）．

このとき，RS では次が成り立つ．

「$\mathrm{regret}(s) < f(s)$ となる step 数 s の単調増加な関数 $f(s)$ が存在して，$f(s) \to M$（$s \to \infty$）M: 定数 となる．すなわち，$\mathrm{regret}(s) < M$ である．」

以下の証明は RS の類似モデルである TOW（Tug-of-war）モデルを扱っている [Kim 15] の方法を手掛かりにした．ただし，TOW と比べても RS は一般の強化学習への適用が可能など RS の方が優れている点は多い．[Kim 15] では 2 本腕で報酬確率の分散が同じケースに限定して regret が有限であることを示している．つまり，$p_1 = 1 - p_2$ という非常に限られた問題のみ扱っており一般性がない．また，天下りなパラメータが与えられているなどの問題があった．ここでは一般化して，K 本腕でかつ等分散性を仮定しない証明を行う．

証明. （命題 2）$p_1 > p_2 > p_i$ $(i \neq 1, 2)$ とし，$RS(a_i, s) = n_i(s) \cdot \big(E(a_i, s) - R\big)$ $(i = 1, 2, \ldots, K)$ とする．$RS(a_i, s)$ の期待値 E と分散 V は $E[RS(a_i, s)] = n_i(s)\ (p_i - R)$, $V[RS(a_i, s)] = n_i(s)\sigma_i^2$ ただし $\sigma_i^2 = p_i(1 - p_i)$ ．また，

$$\begin{aligned} RS(a_i, s) &= n_i(s) \cdot \big(E(a_i, s) - R\big) \\ &= (X_{i,1} - R) + (X_{i,2} - R) + \cdots + (X_{i,n_i(s)} - R) \end{aligned} \tag{8}$$

と書けることに注意しておく．ただし，$X_{i,j} = 1 \text{ or } 0$ で行動 a_i が j 回目に選択された時の報酬を示す．

$\Delta RS_i(s) = RS(a_1, s) - RS(a_i, s)$ $(i \neq 1)$ と定義すると $E[\Delta RS_i(s)] = n_1(s)(p_1 - R) - n_i(s)(p_i - R) = \{(p_1 - p_i)/2\}(n_1(s) + n_i(s))\{(p_1 + p_i)/2 - R\}(n_1(s) - n_i(s))$. また $V[\Delta RS_i] = n_1(s)\sigma_1^2 + n_i(s)\sigma_i^2$. ここで $(p_1 + p_2)/2 = R$ より，$E[\Delta RS_i(s)] = \{(p_1 - p_i)/2\}(n_1(s) + n_i(s)) + \{(p_i - p_2)/2\}(n_1(s) - n_i(s))$. 命題 1 より step 数 s が十分に大きければ確率 1 で $n_1(s) \to s$ となることから，$E[\Delta RS_i(s)] = \{(p_1 - p_i)/2\}s + \{(p_i - p_2)/2\}s = \{(p_1 - p_i)/2\}s$，$V[\Delta RS_i(s)] \leq (n_1(s) + n_i(s))\sigma_i^2 \leq s\sigma_{1,i}^2$ ． ただし $\sigma_{1,i} = \max(\sigma_1, \sigma_i)$.

式 (8) より $\Delta RS_i(s)$ は中心極限定理により，期待値 $E[\Delta RS_i(s)]$，分散 $V[\Delta RS_i(s)]$ の正規分布に従う．$\Delta RS_i(s) < 0$ となる確率は，$Q(E[\Delta RS_i(s)]/\sqrt{V[\Delta RS_i(s)]})$ である．ここで $Q(x)$ は標準正規分布の裾確率を表す Q-関数である．即ち，$Q(x) = (1/\sqrt{2\pi}) \cdot \int_x^\infty \exp(-t^2/2)\, dt$ である．$(n + 1)$ step 目において行動 a_i を選択する確率 $P[s = n + 1, I = i]$ は，

$$\begin{aligned} P[s = n + 1, I = i] &= P[RS(a_j, n) < RS(a_i, n)\ \ (\forall j \neq i)] \\ &< P[\Delta RS_i(n) < 0] \qquad (9) \\ &= Q(\phi_i \sqrt{n}). \qquad (10) \end{aligned}$$

ここで $\phi_i = (p_1 - p_2)/(2\sigma_{1,i})$ と置いた．Chernoff 限界 $Q(x) \leq (1/2)\exp(-x^2/2)$ を用いて regret の上界を評価する．

$$\begin{aligned} E[n_i(n)] &= \sum_{t=0}^{n-1} P[s = t + 1, I = i] \\ &< \frac{1}{2} + \int_0^{n-1} \frac{1}{2} \exp\Big(-\frac{\phi_i^2}{2} t\Big)\, dt \\ &= \frac{1}{2} - \frac{1}{\phi_i^2}\left(\exp\Big(-\frac{\phi_i^2}{2}(n-1)\Big) - 1\right). \qquad (11) \end{aligned}$$

したがって regret は

$$\mathrm{regret}(n) < \sum_{i=1}^{K} (p_1 - p_i)\left\{\frac{1}{2} - \frac{1}{\phi_i^2}\left(\exp\Big(-\frac{\phi_i^2}{2}(n-1)\Big) - 1\right)\right\} \tag{12}$$

$$\to \sum_{i=1}^{K} (p_1 - p_i)\Big(\frac{1}{2} + \frac{1}{\phi_i^2}\Big) \quad (n \to \infty). \tag{13}$$

□

5. シミュレーションによる検証

証明した性質をシミュレーションで検証した．基準 R は式 (4) のように最適切に設定した．いずれも 1000 回シミュレーションして平均を求めた．パフォーマンスの指標としては，regret の他に，各 step において最適な行動を選んだ比率「正確さ」（accuracy）を用いる．つまり，t step 目の accuracy は，accuracy = （t step 目で報酬確率 p_1 である行動 a_1 を選択した回数）／シミュレーション回数（ここでは 1000 回）となる．最適切に設定した RS であれば報酬確率の差が小さい場合でもその差を検出できるか試すため 2 本腕の場合で報酬確率 $p_1 > p_2$ として $(p_1, p_2) = (0.51, 0.49), (0.501, 0.499)$ で調べた $(R = 0.5)$．結果は図 1 のとおり．regret の図の上部の点線は命題 2 で示した上界を示す．$(0.501, 0.499)$ のように差が 0.002 しかなくても 100 万 step 経過すればほぼ accuracy が 1 になることが分かる．また regret は命題 2 で求めた上界を超えていないことが確認できる．

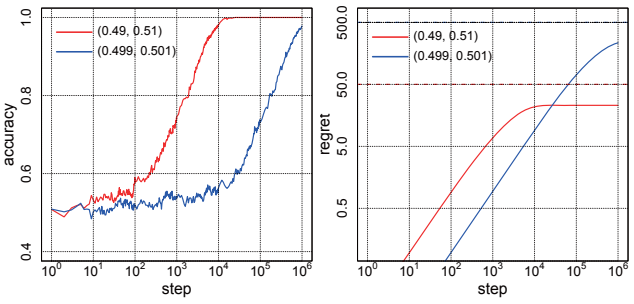

図 1: $K = 2$ (凡例の数値は報酬確率の値) の RS の accuracy (左) と regret (右) の時間発展

次に命題 1 と命題 2 は一般の K 本腕で示したので $K = 10$ でも成立を確認するためシミュレーションした． 今度は報酬確率は $[0, 1]$ の乱数で生成した．結果は図 2 のとおり．$K = 10$ でも accuracy が 1 に近づくことや，regret が命題 2 で示した上界 (式 (13) の値．図 2 の regret の上部にある点線) を超えていないことが分かる．なお，$K = 10$ の regret の上界が $K = 2$ に比べて実際の regret よりもかなり上にある．これは，命題 2 の証明中の式 (9) にあるように行動 a_i が選択される確率を報酬確率が最大の行動 a_1 とだけ比較することで評価しているため，行動の数が多くなるほど報酬確率が最大でない行動の選択確率が過大評価されるためであり，改善は可能であろう．

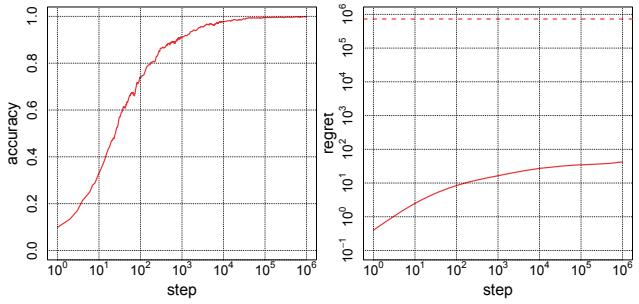

図 2: $K = 10$（報酬確率はシミュレーションごとに乱数で生成）の RS の accuracy（左）と regret（右）の時間発展

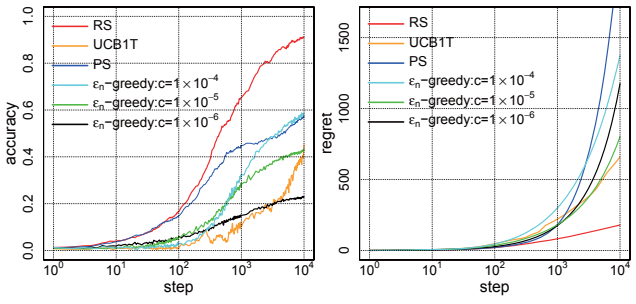

図 3: $K = 100$ の RS，UCB1T，PS，ϵ_n-greedy の比較 (左：accuracy，右：regret)

6. 他のアルゴリズムとの比較

6.1 UCB1-Tuned

UCB（upper confidence bound） [Auer 02] は，regret が理論限界である対数オーダーとなることが保証されている．UCB1 を性能向上させた UCB1-Tuned（UCB1T）と RS を比較する．

$$\mathrm{UCB1T}(a_i) = E_i + \sqrt{\frac{\ln n}{n_i} \min\{\frac{1}{4}, V_i(n_i)\}}\,. \qquad (14)$$

$V_i(n_i) = v_i + \sqrt{2 \ln n / n_i}$ で v_i は行動 a_i の報酬の分散である．

6.2 ϵ_n-greedy

RS と同じく報酬確率に関する情報を利用したアルゴリズムとしては ϵ_n-greedy があり，regret は対数オーダーとなる [Auer 02]．パラメータを $c > 0$ かつ $0 < d < 1$ とし，K 本腕において数列 $\epsilon_n \in (0, 1]$，$n = 1, 2, \ldots$ を次式で定義する．

$$\epsilon_n = \min\Big\{1, \frac{cK}{d^2 n}\Big\}. \qquad (15)$$

a_n が報酬平均最大の行動ならば，確率 $1 - \epsilon_n$ で a_n を選択し，確率 ϵ_n で無作為に選択する．p_1 を報酬確率の最大値とし，d は $\Delta_i = p_1 - p_i$ として $0 < d \leq \min_{i \neq 1} \Delta_i$ を満たす必要があり，設定は簡単ではない．

6.3 シミュレーションによる性能比較

UCB1T，PS，ϵ_n-greedy，RS の性能を $K = 100$ で比較した．性能の指標は accuracy と regret とし，報酬確率は $[0, 1]$ から乱数で選び，1,000 回シミュレーションして平均を求めた．ϵ_n-greedy のパラメータ c を決めるのが難しいが，10,000 step 時点での regret が最小になるのは試行錯誤の結果 $c = 1 \times 10^{-5}$ 付近と経験的に分かったため，$c = 1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}$ の結果を比較対象とした．d は p_2 を 2 番目に大きい報酬確率として $d = p_1 - p_2$ とし，PS と RS の基準 R は $(p_1 + p_2)/2$ とした．

結果は図 3 のとおり．accuracy は RS が最も速く 1 に近づく．regret については PS は R を超える報酬確率を持つ行動が見つからない限りランダムに行動を選択するため増加のスピードが速い．RS の regret は有限であり，UCB1T や ϵ_n-greedy は対数オーダーで発散するため，RS の方が低く抑えられている．最適切に基準 R を設定した RS は UCB1T，PS，ϵ_n-greedy よりパフォーマンスが良いことが分かる．ϵ_n-greedy と同じく報酬確率に関する情報が事前に必要とはいっても，パラメータ c のようなものなしに有限の regret を達成し，accuracy も良い RS は有用なアルゴリズムといえるだろう．

7. 結論

本論文では，満足化のモデルである RS について K 本腕バンディット問題に適用した際の理論的な分析を行った．基準を満たす行動があれば必ずそれを見つけ出すことや，満足化が最適化に一致する場合は有限の regret が成立することを理論的に示し，またシミュレーションでも成立を確認した．また，他の最適化アルゴリズムと比較し，RS の性能が優れていることを明らかにした．本論文では事前に設定している満足化基準の R をオンラインで推定するアルゴリズムを開発し，理論的な分析を行うことは今後の課題である．RS には本論文で挙げた以外にも多くの優れた点がある．例えば満足化を達成する速さは満足化基準を満たす行動の占める割合に依存し，行動数（つまり問題の規模）にはほぼ依存しないというスケーラビリティがある [Oyo 17]．また RS は特定の課題の形式に依存しない単純な価値関数であるので，一般化することでバンディット問題だけでなく他の強化学習の課題に適用可能である [高橋 16]．RS の更なる応用としては複数のエージェントに複数の基準 を与えて探索を行う並列化などが考えられる．これは最適化問題の判定問題への変換としての理論的な意味を持つ．

参考文献

[Auer 02] Auer, P., Cesa-bianchi, N., and Fischer, P.: Finite-time analysis of the multiarmed bandit problem, *Machine Learning*, Vol. 47, No. 2-3, pp. 235–256 (2002)

[Kim 15] Kim, S. J., Aono, M., and Nameda, E.: Efficient decision-making by volume-conserving physical object, *New Journal of Physics*, Vol. 17, No. 8, 083023 (2015)

[Lai 85] Lai, T. L. and Robbins, H.: Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics*, Vol. 6, No. 1, pp. 4–22 (1985)

[Oyo 17] Oyo, K. and Takahashi, T.: Optimization through satisficing with prospects, in *AIP Conference Proceedings*, Vol. 1863, 360013 (2017)

[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. AI30-M-1–11 (2016)