東京電機大学

博　士　論　文

満足化価値関数のバンディット問題における有効性：
満足化の保証と期待損失の有限性

Effectiveness of a Satisficing Value Function in the Multi-armed Bandit Problems: Guarantee of Satisficing and Finiteness of Regret

2020 年 3 月

東京電機大学大学院　先端科学技術研究科　情報学専攻

玉造晃弘

学籍番号 17UDJ01

目 次

図の目次

第 1 章

はじめに

1.1 背景

強化学習とは試行錯誤を通じて環境に適応する学習・制御の枠組みである．最近の強化学習を用いた技術の発展は目覚ましい．DQN (Deep Q-Network) を用いたコンピュータゲームの学習 [Mnih 15] や Alpha-Go による囲碁の学習 [Silver 16] などでは人間並みもしくは人間以上の成績を上げるほどまでになった．また，強化学習の適用対象がコンピュータ上のものだけでなく，現実のロボット [Muse 09] や無人飛行機 (UAV) [Zhao 18] の制御にまで広がっている．

このように強化学習の対象がより複雑なタスクや現実世界にまで広がると探索空間が急激に膨張するため，エージェントのセンサー，プロセッサーやアクチュエーターにおけるリソースの限界を考慮すると，従来のように真に最適な行動を見つけるまで試行を続ける最適化アルゴリズムだけでは上手く機能しない可能性がある．このような状況を動物や人間に対して表現した概念として Simon が提唱した限定合理性がある [Simon 57]．これは動物や人間は意思決定の際に全ての選択肢を比較し考慮するだけの認知的能力はなく，限られた能力しかない中で行動しなければならないというものである．限定合理的なエージェントは非合理的にふるまうように見えるが，制約や限界を考慮すると，エージェントの行動は合理的と理解できる．

限定合理性は近年大きな関心を集めている．限定合理性の更新された形態である計算合理性 [Lewis 14] は，神経科学 (脳)，認知科学 (心) および人工知能 (機械) の三分野を統合すると主張されている [Gershman 15]．また，そもそも認知の高度な機能と見なされてきた抽象化や階層性 [Tenenbaum 11] がむしろ計算論的な制約に由来するものという有力な議論も情報理論の枠組みを利用して展開されている [Genewein 15]．限定合理性の中で人間が意思決定を行うための方針としては満足化の原理（または満足化の戦略）が代表的である [Simon 55, Simon 56]．満足化の原理によれば人間は最適化を目指して行動

するのではなく，ある一定の満足できる基準を超える結果が得られるように行動し，満足できれば行動を終えるとされる．また，満足化の原理は動物の生存戦略とも見なせる．動物は 1 日の食物の摂取量を最大にするまで行動する必要はなく，生存に必要な量が得られれば十分であるし，食べきれない食物が腐るということもあるであろう．動物の生存戦略をモデル化しコンピュータの計算に応用したものは Biology-based algorithms (BBA) と呼ばれ，これまでも多くの研究例がある [Siddique 15]．しかし，今まで満足化の原理はあまり注目されてこなかった．そこで高橋 ([高橋 16, Oyo 17]) は価値関数のレベルで満足化の原理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その有効性を確認してきた．しかし，先行研究では *RS* の特徴について理論的に解明されていなかった．

本論文で，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増加することが理論的に知られている [Lai 85]．それゆえ試行を繰り返せば regret は無限に増加する．しかし，*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，満足化は最適化に一致し，その場合，regret は無限大に発散せず有限の値で抑えられる (つまり，上に有界になる) という注目すべき性質が成り立つ．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．これらの結果により *RS* の理論的な基盤や特徴が明らかになり，今後 *RS* のより広い範囲での活用が可能になるだろう．

1.2　本論文の構成

本論文の構成は次のとおりである．

「2　満足化のモデル」では最初に *K* 本腕バンディット問題を説明した後，*RS* の基本的な考え方と定式化を説明する．「3　理論的な解析」で本題である *RS* の理論的解析を行う．命題 1 で満足化基準を超える行動を選択できるという性質を証明する．命題 2 で満足化が最適化に一致する場合について regret の有限性を証明する．どちらの命題もでシミュレーションにより成立を確認する．「4　バンディット問題における性能比較」では他の最適化アルゴリズムや満足化アルゴリズムと定量的又は定性的に比較を行い，*RS* の特徴を明らかにする．「5　既存モデルとの関係」では，*RS* の起源ともいえる *S*0 モデルとの関係，命題 2 の証明で参考にした TOW モデルとの関係についてぞれぞれ考察する．

「6　満足化基準が変化する場合の命題の拡張」では命題 1 や命題 2 を拡張し，満足化基準が変動する場合にも同様の性質が成立することを理論的に示し，シミュレーションでも確認する．最後に「7 本研究の成果と今後の発展」で本研究で得られた成果と，その意義をまとめ，*RS* の更なる可能性を述べる．

第 2 章

満足化のモデル

以下では K 本腕バンディット問題の枠組みで，ポリシーと価値関数のそれぞれのレベルで満足化のモデルを導入する．

2.1 K 本腕バンディット問題

まず K 本腕バンディット問題の枠組みを述べる．本論文では最も広く用いられている設定である，報酬を確率変数とし，報酬が従う確率分布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．エージェントは 1 回の試行で 1 つの行動を選択する．その結果として，選択した行動に伴う報酬確率に従って，エージェントは確率的に報酬を得たり (報酬 1)，得なかったりする (報酬 0)．エージェントはこの試行を繰り返す．ここでの目標は累積報酬の最大化である．報酬確率の分布 $\{p_i\}$ に依らずに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．

$$\mathrm{regret}(n) = \sum_{i=1}^{K} (p_{i^*} - p_i) E[n_i(n)]. \tag{2.1}$$

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．regret は，「最初から最適行動を選んでいた場合の累積期待報酬に比べてどのくらい実際の選択行動の累積期待報酬が劣っているか」という期待損失を表す．regret の最小値はゼロであり，値が小さいほどアルゴリズムの性能がよいといえる．実際には step 数 n に対して少なくとも $\mathcal{O}(\log n)$ で増大することが理論的に知られている [Lai 85].

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．

$$E_i = n_i^1 / (n_i^1 + n_i^0). \tag{2.2}$$

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，$n = \sum_{i=1}^{K} n_i$ とする．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．報酬平均 E_i を行動 a_i の価値関数として greedy 法を用いると，たまたま試行の序盤で報酬平均が高くなった行動 a_i $(i \neq i^*)$ があれば，その行動をいつまでの選択し続けてしまう危険がある．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」も必要であり，利益追求と探索のトレードオフが K 本腕バンディット問題ではポイントとなる．

2.2　素朴満足化ポリシー (PS)

満足化 satisficing の標準的な定義は，

> 「基準 $\aleph$ 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，その行動で満足する」

というものである．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は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 $\aleph$ 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．

2.3　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 $\aleph$ との差

$$\delta_i = E_i - \aleph \tag{2.3}$$

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．しかし，これは単に定数 $(-\aleph)$ が切片としてついただけである．価値の大小のみで行動を決める greedy 法でこの価値を運用する限り，そのままでは単に価値 E に従うのと同じことになる．

そこで RS (risk-sensitive satisficing) 価値関数 を次のように定義する [高橋 16, Oyo 17].

$$RS_i = n_i\delta_i = n_i(E_i - \aleph). \tag{2.4}$$

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 n_i が小さい行動が優先される．すなわち，「実際は $P_i > \aleph$ となる a_i が存在するかもしれないが，a_i をこれまで試してみてたまたま外れが多かったため $E_i < \aleph$ なのかもしれず，もっと試せば $E_i > \aleph$ となるかもしれない」と考える．これは「不確実ならば楽観的に」という Cesa-Bianchi らによる強化学習における合理的な最適化理論のコンセプトであり [Bubeck 12]，UCB (後述の 4.1 参照) も同じ考え方である．

満足であれば RS は悲観的知識利用を行う．つまり，δ_i を正とする a_i が一つのみ存在すれば，エージェントはそれを選択し続ける．δ_i を正とする行動が複数あれば，n_i が大きいものが優先される．つまり，「選択した回数 n_i が大きいものほど，$E_i > \aleph$ という大小関係が信頼できる」という考え方に基づいている．

なお，[篠原 07] で因果推論で用いられていた指標 $S0$ ([Takahashi 11] では価値関数としても使用) は，RS で満足化基準 $\aleph = 0.5$ とした価値関数に等しいことが分かっている (5.2 を参照).

2.4 満足化基準の決め方

満足化基準 $\aleph$ は満足と不満足の境界であり，プロスペクト理論 [Tversky 81] における利得と損失の分岐点，または参照点に類似している．これ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．生態学的な例であればエージェントが動物，報酬の 1 と 0 が食物の在と不在を表し，行動はある餌場で食物を探すことであれば，だいたい二日に一度くらい食物を得たいとすれば $\aleph = 0.5$ 以上に設定すればよい．

$\aleph$ の値が最適行動と次善行動の 2 つの間であれば，$\aleph$ 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と見なすことができる．この時の $\aleph$ を「最適切基準」と呼ぶ．全ての行動で報酬確率 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 とすれば，もっとも単純には次のように設定すれば $\aleph$ は最適切基準となる：

$$\aleph = (p_1 + p_2)/2. \tag{2.5}$$

step 数 n に対して regret は少なくとも $\mathcal{O}(\log n)$ で増大することが理論的に示されているが [Lai 85]，これは報酬分布について何も情報を持ってないことが前提の結果である．この前提を緩和して上記のように RS に最適切基準を設定することを許せば，後述の命題 2 で示すように regret は有限の値で抑えられることが分かる．

なお，上記のように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 (どれが報酬確率最大の行動なのか) 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，最適と次善の間の 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．また，「4　バンディット問題における性能比較」では最適切基準の設定以外で報酬分布の一部の情報を利用する他のアルゴリズムとの比較も行う．

第 3 章

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．先ず，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ことを理論的に示す．次に，最適切基準が与えられ，満足化が最適化に一致する場合，regret が有限に抑えられることを理論的に示す．また，それらの性質をシミュレーションでも確認する．

3.1　満足化の理論的な保証

命題 1 の証明では記述を明確にするため，次のように 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．

報酬平均:

$$E(a_i, s) = \frac{n_i^1(s)}{n_i(s)}. \tag{3.1}$$

RS の式:

$$RS(a_i, s) = n_i(s) \cdot \big(E(a_i, s) - \aleph\big). \tag{3.2}$$

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 $\aleph$ より大きい行動の集合を A_U, $\aleph$ より小さい集合を A_L とする．つまり，$I_U = \{i \mid p_i > \aleph\}$，$I_L = \{i \mid p_i < \aleph\}$ 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 $\aleph$ より報酬確率が大きい行動の集合の中から行動を選択し，また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，

$$P\Big(\arg\max_{a_i} RS(a_i, s) \in A_U\Big) = 1 \ \ (s \to \infty). \tag{3.3}$$

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

Claim A.

$$\begin{aligned} &i \in I_L \text{のとき}, \\ &P\big(\#N_i = \infty \Leftrightarrow RS(a_i, s) \to -\infty \ (s \to \infty)\big) = 1. \end{aligned} \tag{3.4}$$

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，$\#N_i = \infty$ が成立する．

逆に $\#N_i = \infty$ のとき，大数の法則より任意の正の数 ϵ を取ると，ϵ に対応する S が存在し，S より大きい任意の整数 $s \in N_i$ に対し，$P\big(|E(a_i, s) - p_i| < (\aleph - p_i)/2\big) > 1 - \epsilon$ とできる．$|E(a_i, s) - p_i| < (\aleph - p_i)/2$ ならば，$E(a_i, s) < p_i + (\aleph - p_i)/2$ となり，これを RS の式に代入すると，

$$\begin{aligned} RS(a_i, s) &= n_i(s) \cdot \big(E(a_i, s) - \aleph\big) \\ &< n_i(s) \cdot \big(p_i + \frac{\aleph - p_i}{2} - \aleph\big) \\ &= n_i(s) \cdot \frac{p_i - \aleph}{2} < 0 \end{aligned} \tag{3.5}$$

となる．$s \to \infty$ とすれば，式 (3.5)$\to -\infty$ 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.

$$\text{ある } i \in I_U \text{について } P(\#N_i = \infty) = 1. \tag{3.6}$$

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，

$$\begin{aligned} &P\big(\text{ある } j \in I_L \text{について } RS(a_j, s) \to -\infty \;\big|\; \text{任意の } i \in I_U \\ &\quad \text{に対して} \#N_i < \infty\big) = 1 \end{aligned} \tag{3.7}$$

となる．ところが，次の (i) と (ii) はお互いに矛盾する：(i) $RS(a_j, s) \to -\infty$ であること，(ii) 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．よって，

$$\begin{aligned} &P\big(\text{ある } j \in I_L \text{について } RS(a_j, s) \to -\infty, \text{ 任意の } i \in I_U \\ &\quad \text{に対して} \#N_i < \infty\big) = 0 \end{aligned} \tag{3.8}$$

となる．したがって，

$$\begin{aligned}
&P\big(\text{ ある } j \in I_L \text{について } RS(a_j,s) \to -\infty, \text{ 任意の } i \in I_U \\
&\text{に対して} \#N_i < \infty\big) \\
&= P\big(\text{ ある } j \in I_L \text{について } RS(a_j,s) \to -\infty \,\big|\, \text{任意の } i \in I_U \\
&\text{に対して} \#N_i < \infty\big) P(\text{任意の } i \in I_U \text{に対して} \#N_i < \infty) \qquad (3.9)
\end{aligned}$$

より $P(\text{任意の } i \in I_U \text{に対して} \#N_i < \infty) = 0$ でなければならない． □

命題 1. (再掲)

$$P\Big(\arg\max_{a_i} RS(a_i,s) \in A_U\Big) = 1 \ \ (s \to \infty). \qquad (3.3)$$

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，$P\big(|E(a_k,s) - p_k| < (p_k - \aleph)/2\big) > 1 - \epsilon$ とできる．$|E(a_k,s) - p_k| < (p_k - \aleph)/2$ ならば，

$$\begin{aligned}
RS(a_k,s) &= n_k(s) \cdot \big(E(a_k,s) - \aleph\big) \\
&> n_k(s) \cdot \big(p_k + \frac{\aleph - p_k}{2} - \aleph\big) \\
&= n_k(s) \cdot \frac{p_k - \aleph}{2} > 0 \qquad (3.10)
\end{aligned}$$

となる．

よって，$P\big(\text{ 十分大きい } s \text{ について } RS(a_k,s) > 0\big) > 1 - \epsilon$．$\epsilon$ は任意であるから $P\big(\text{ 十分大きい } s \text{ について } RS(a_k,s) > 0\big) = 1$.

ここで，いずれかの $i \in I_L$ について，$\#N_i = \infty$ となるものと仮定する．Claim A より $RS(a_i,s) \to -\infty$ が成り立つとしてよい．一方，十分大きいすべての s について $RS(a_k,s) > 0$ となることから，$RS(a_i,s) \to -\infty$ より $\#N_i < \infty$ が導かれる．しかし，$\#N_i = \infty$ と $\#N_i < \infty$ はお互いに矛盾するので，最初の仮定は誤りである．よって，任意の $i \in I_L$ に対して $P(\#N_i < \infty) = 1$ が成立する．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，このことより $P\Big(\arg\max_{a_i} RS(a_i,s) \in A_U\Big) = 1 \ \ (s \to \infty)$ が明らかに従う． □

3.2　regret の理論解析

基準 $\aleph$ を最適切基準に設定した RS では regret は有限な値で抑えられることを示す．命題 1 の証明同様に step 数 s と選択行動 a_i を明示した記号を用いる．

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 $\aleph = (p_1 + p_2)/2$ となるように $\aleph$ 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．

「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，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) ダイナミクスのモデル (以下，TOW という) に対する論文 [Kim 15, Kim 16, Kim 18] を参照して考えたものである (RS と TOW の類似点と相違点は 5.3 を参照)．ただし，それらの論文では 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解析している．(報酬確率がベルヌーイ分布に従う 2 値のバンディット問題では，行動 a_i の分散を V_i とすると等分散であることは，$p_1 \neq p_2$ であれば，$V_1 = V_2 \Leftrightarrow p_1(1-p_1) = p_2(1-p_2) \Leftrightarrow p_1 + p_2 = 1$ となり，非常に強い仮定を置いた特殊なケースである．)

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で証明を行う．

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

$$\begin{aligned} RS(a_i, s) &= n_i(s) \cdot \big(E(a_i, s) - \aleph\big) \\ &= n_i^1(s) - n_i(s)\aleph \\ &= (X_{i,1} - \aleph) + (X_{i,2} - \aleph) + \cdots + (X_{i,n_i(s)} - \aleph) \end{aligned} \tag{3.11}$$

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

$$\begin{aligned} E[\Delta RS_i(s)] &= n_1(s)(p_1 - \aleph) - n_i(s)(p_i - \aleph) \\ &= \{(p_1 - p_i)/2\}(n_1(s) + n_i(s)) \\ &\quad + \{(p_1 + p_i)/2 - \aleph\}(n_1(s) - n_i(s)). \end{aligned} \tag{3.12}$$

$$V[\Delta RS_i(s)] = n_1(s)\sigma_1^2 + n_i(s)\sigma_i^2. \tag{3.13}$$

ここで $(p_1+p_2)/2=\aleph$ より

$$\begin{aligned} E[\Delta RS_i(s)] &= \{(p_1-p_i)/2\}(n_1(s)+n_i(s)) \\ &\quad +\{(p_i-p_2)/2\}(n_1(s)-n_i(s)). \end{aligned} \tag{3.14}$$

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，

$$\begin{aligned} E[\Delta RS_i(s)] &= \{(p_1-p_i)/2\}s+\{(p_i-p_2)/2\}s \\ &= \{(p_1-p_2)/2\}s. \end{aligned} \tag{3.15}$$

$$\begin{aligned} V[\Delta RS_i(s)] &\le (n_1(s)+n_i(s))\sigma_{1,i}^2 \le s\sigma_{1,i}^2, \\ &\text{ただし}\sigma_{1,i}=\max(\sigma_1,\sigma_i). \end{aligned} \tag{3.16}$$

となる．[*1]

式 (3.11) 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，

$$\begin{aligned} P[s=n+1,I=i] &\le P[RS(a_j,n)\le RS(a_i,n)\ \ (\forall j\neq i)] \\ &\le P[\Delta RS_i(n)\le 0] \\ &= Q\left(\frac{E[\Delta RS_i(n)]}{\sqrt{V[\Delta RS_i(n)]}}\right) \\ &\le Q\left(\frac{(p_1-p_2)\sqrt{n}}{2\sigma_{1,i}}\right) \\ &= Q(\phi_i\sqrt{n}). \end{aligned} \tag{3.17, 3.18}$$

ここで $\phi_i=(p_1-p_2)/(2\sigma_{1,i})$ と置いた．Chernoff 限界 $Q(x)\le(1/2)\exp(-x^2/2)$ を

[*1] 注：ここでは近似計算を行っているため，その誤差のために上界の値は正確なものでない可能性がある．いくつかのシミュレーションでは上界の値が実際に成立していることを確認した．例えば図 3.1 や図 3.2 を参照．

用いて regret の上界を評価する．

$$
\begin{aligned}
E[n_i(n)] &= \sum_{t=0}^{n-1} P[s = t+1, I = i] \\
&\leq \sum_{t=0}^{n-1} Q(\phi_i \sqrt{t}) \\
&\leq \frac{1}{2} + \sum_{t=1}^{n-1} \frac{1}{2} \exp\Big(-\frac{\phi_i^2}{2} t\Big) \\
&\leq \frac{1}{2} + \int_0^{n-1} \frac{1}{2} \exp\Big(-\frac{\phi_i^2}{2} t\Big)\, dt \\
&= \frac{1}{2} - \frac{1}{\phi_i^2}\left(\exp\Big(-\frac{\phi_i^2}{2}(n-1)\Big) - 1\right) \qquad (3.19) \\
&\to \frac{1}{2} + \frac{1}{\phi_i^2} \ \ (n \to \infty). \qquad (3.20)
\end{aligned}
$$

したがって regret は

$$
\begin{aligned}
\text{regret}(n) &= \sum_{i=1}^{K} (p_1 - p_i) E[n_i(n)] \\
&\leq \sum_{i=1}^{K} (p_1 - p_i)\left\{\frac{1}{2} - \frac{1}{\phi_i^2}\left(\exp\Big(-\frac{\phi_i^2}{2}(n-1)\Big) - 1\right)\right\} \qquad (3.21) \\
&\to \sum_{i=1}^{K} (p_1 - p_i)\Big(\frac{1}{2} + \frac{1}{\phi_i^2}\Big) \ \ (n \to \infty). \qquad (3.22)
\end{aligned}
$$

□

3.3　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 $\aleph$ は命題 2 と同じく報酬確率 p_i を $p_1 > p_2 > p_i \ \ (i \neq 1, 2)$ として $\aleph = (p_1 + p_2)/2$ と設定した．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 は次のようになる．

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) ／ 全シミュレーション回数 (ここでは 1000 回).

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

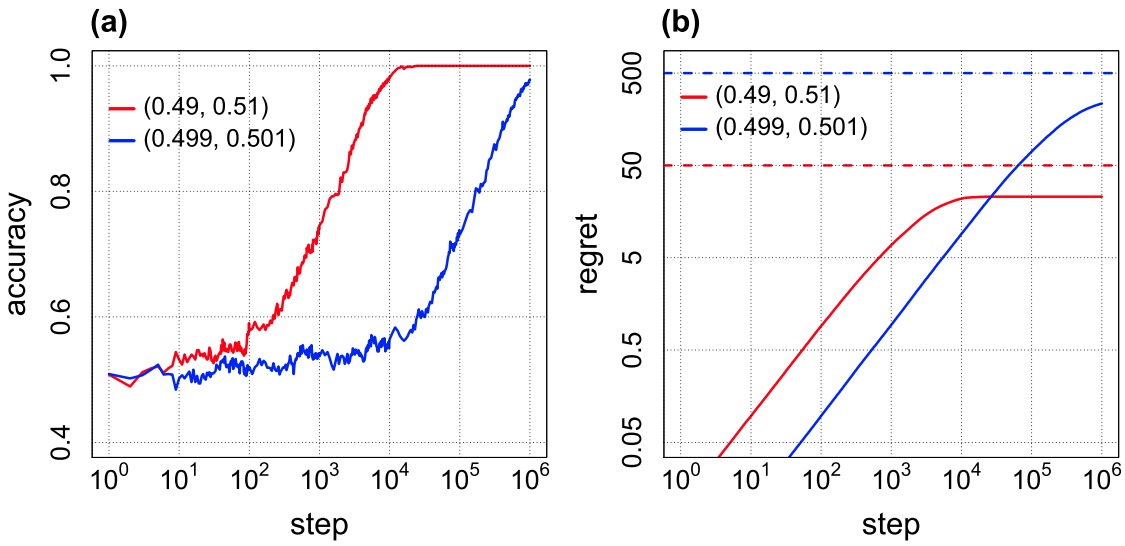

図 3.1　$K = 2$ の場合の RS のシミュレーション．報酬確率は $(0.51, 0.49)$ または $(0.501, 0.499)$ とした．**(a)** accuracy および **(b)** regret のプロット．**(b)** の上部の点線は命題 2 で計算された regret の上界を示す．

調べた．結果は図 3.1 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 (式 (3.22)) を超えていないことが確認できる．

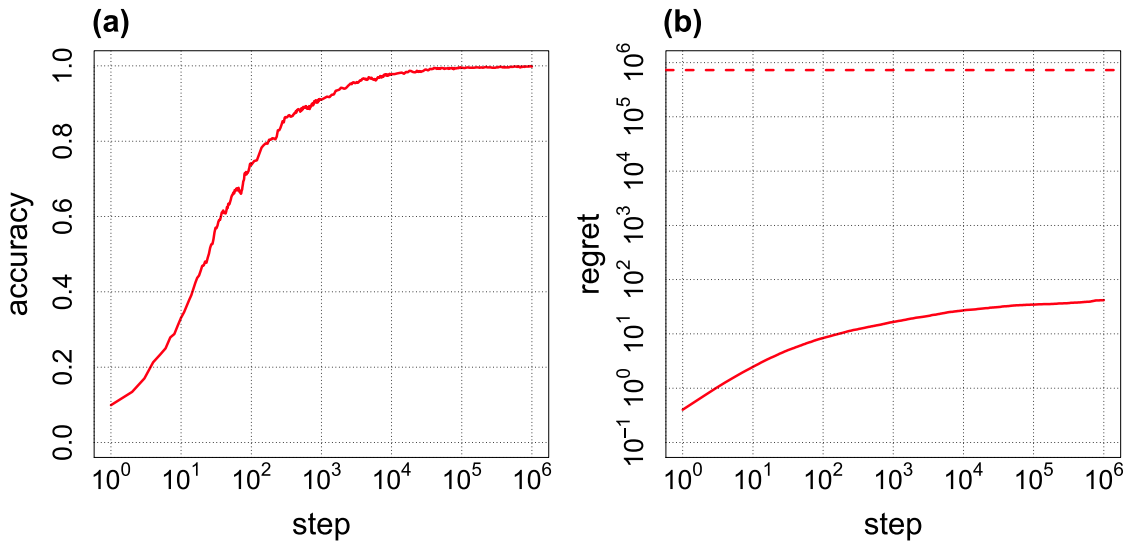

図 3.2　$K = 10$ の場合の RS のシミュレーション．それぞれのシミュレーションにおいて報酬確率は $[0, 1]$ からの一様乱数により生成されている．**(a)** accuracy および **(b)** regret のプロット．**(b)** の上部の点線は命題 2 により計算された regret の上界を示す．

次に命題 1 と命題 2 は K 本腕 $(K \geq 2)$ 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．報酬確率は $[0, 1]$ の乱数で生成した．結果は図 3.2 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 (3.22)) 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 (3.17) 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ことが原因の 1 つと考えられる．

第 4 章

バンディット問題における性能比較

ここではバンディット問題の有名なアルゴリズムである UCB1-Tuned や ϵ-greedy との比較を通じて *RS* の性能や特徴を明らかにする．

4.1 UCB1-Tuned

RS と同様に「まだあまり試していない行動 (不確実) のポテンシャルは高い (楽観)」という考え方に基づいたアルゴリズムとして UCB (upper confidence bd) [Auer 02] がある．UCB の regret は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．ここでは UCB1 のパフォーマンスを改良した UCB1-Tuned (以降 UCB1T という) を導入する．

$$\mathrm{UCB1T}(a_i) = E_i + \sqrt{\frac{\ln n}{n_i} \min\{\frac{1}{4}, V_i(n_i)\}}. \tag{4.1}$$

ここで $V_i(n_i) = v_i + \sqrt{2 \ln n / n_i}$ であり，v_i は行動 a_i の報酬の分散，$1/4$ は二項分布に従う確率変数の分散の上限である．このアルゴリズムでは UCB1T を各行動の価値とした greedy 法のもと，報酬平均である第一項 E_i が利益追求の傾向を，行動 a_i の試行につれて減少していく第二項が E の信頼性の低さを表現し，探索を担っているといえる．$n_i = 0$ だと第二項が計算できないので，最初の K 回は各行動を一度ずつ選択して値を有限にして用いる．

4.2 ϵ_n-greedy

RS ではバンディット問題において満足化が最適化となるためには，基準 $\aleph$ の設定に，最適と次善の行動の報酬確率という，通常は未知の報酬分布に関する情報が必要となり，チートであることは間違いない．ただし，適切な基準の範囲，即ち最適行動と次善行動の

報酬確率の間の区間という情報が利用可能な場合にそれをアルゴリズムが利用できること自体は優れたことであり，UCB のようなアルゴリズムではそれを行うのは難しい．

同じく区間の情報を利用する直観的なアルゴリズムとしては ϵ_n-greedy がある [Auer 02]．以下に示すように無作為選択の確率である ϵ_n を徐々に下げてアニーリングしていくことで，ϵ_n-greedy の regret は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．ϵ_n-greedy ではパラメータを $c > 0$ 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．

$$\epsilon_n = \min\left\{1, \frac{cK}{d^2 n}\right\}. \tag{4.2}$$

それぞれの $n = 1, 2, \ldots$ に対して，a_n を報酬平均が最大となる行動とすれば，確率 $1 - \epsilon_n$ で a_n を選択し，確率 ϵ_n で無作為に行動を選択する．ここで d は次を満たす必要がある．p_1 を報酬確率の最大値として

$$\Delta_i = p_1 - p_i \text{として } 0 < d \leq \min_{i \neq 1} \Delta_i. \tag{4.3}$$

上記のように d を決めるためには $\min \Delta_i$ を計算する必要があるが，これには最適行動の報酬確率 p_1 と次善行動の報酬確率 p_2 の差 $p_1 - p_2$ を事前に知っておく必要があり RS と同様にチートといえる．またパフォーマンスはパラメータ $c > 0$ の値に敏感で c の最適値を見つけることは難しい [Auer 02]．

一方，RS の基準 $\aleph$ を決めるためにはパラメータ c のようなものは不要で，最適行動と次善行動の報酬確率の和 $p_1 + p_2$ が分かればよい．$(p_1 + p_2)/2$ を $\aleph$ とすれば $\aleph$ は最適切になる．より一般的には最適行動と次善行動の報酬確率の区間あるいは区間内の一点の値が分かればよい．

4.3　シミュレーションによる性能比較

UCB1T，PS，ϵ_n-greedy，及び 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，1000 回シミュレーションして平均を求めた．4.2 で述べたように ϵ_n-greedy のパラメータ c を決めるのが難しいが，ここでは 10,000 step 時点での regret を目安とし，これ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て示す．d は $d = p_1 - p_2$ とした．RS と PS については，その満足化行動が最適化行動となり，その効率化について評価できるよう $\aleph = (p_1 + p_2)/2$ の最適切基準を与える．

結果は図 4.1 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 $\aleph$ 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

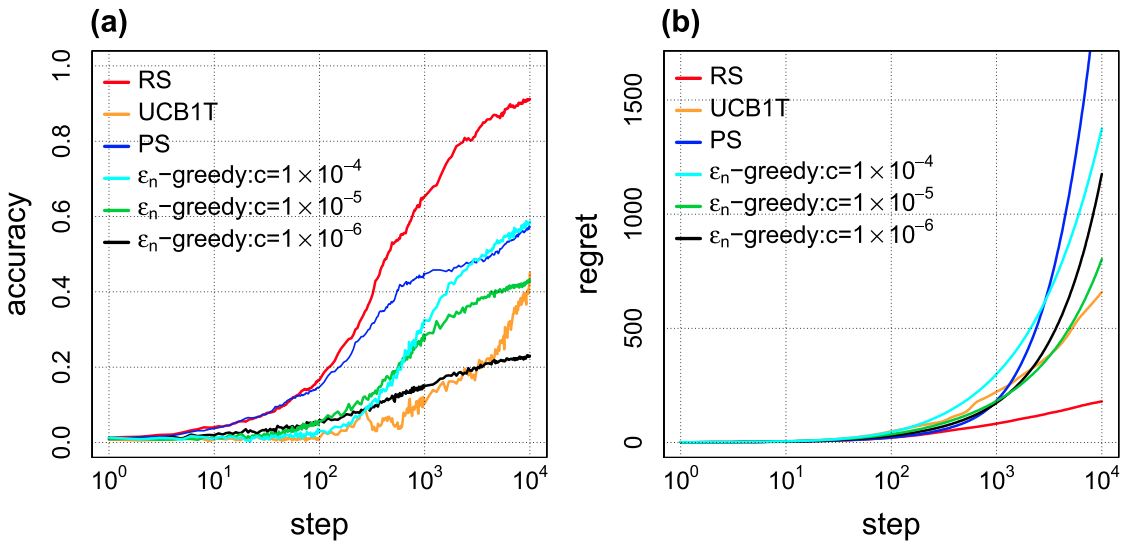

図 4.1　$K = 100$ の場合のシミュレーション．報酬確率は $[0, 1]$ からの一様乱数で生成．*RS*，UCB1T，*PS* および ϵ_n-greedy: **(a)** accuracy および **(b)** regret.

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 ℵ 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 と同じく最適行動と次善行動の報酬確率に関する事前情報が必要とはいっても，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．

4.4　価値関数の値の変化の期待値

最適切に基準 ℵ を設定した *RS* が他のアルゴリズムよりも性能が良い理由を定性的に考察する．$(n+1)$ step 目に i 番目の行動を選択時に，*RS* の値が n step 目からどのように変化するか考える．まず *RS* の式を再掲する．

$$RS(a_i, n) = n_i(n) \cdot (E(a_i, n) - \aleph) = n_i^1(n) - n_i(n)\aleph. \tag{4.4}$$

ここで $n_i^1(n)$ は行動 a_i の選択で 1 が出た回数であった．$(n+1)$ step 目では確率 p_i で

$$RS(a_i, n+1) = n_i^1(n) + 1 - (n_i(n) + 1)\aleph = RS(a_i, n) + 1 - \aleph \tag{4.5}$$

となり，確率 $(1 - p_i)$ で

$$RS(a_i, n+1) = n_i^1(n) - (n_i(n) + 1)\aleph = RS(a_i, n) - \aleph \tag{4.6}$$

と変化する．RS_i の変化 $\Delta RS(a_i, n) = RS(a_i, n+1) - RS(a_i, n)$ の期待値 $E[\Delta RS(a_i, n)]$ は，

$$E[\Delta RS(a_i, n)] = p_i(1-\aleph) + (1-p_i)(-\aleph) = p_i - \aleph \tag{4.7}$$

となる．よって

$$p_i > \aleph \text{であれば} E[\Delta RS(a_i, n)] > 0, \tag{4.8}$$
$$p_i < \aleph \text{であれば} E[\Delta RS(a_i, n)] < 0 \tag{4.9}$$

となり，これはどの step n でも変わらない．

このことから RS では最適切に $\aleph$ を設定していれば，最適でない行動では $E[\Delta RS(a_i, n)] < 0$ となり，その行動を選択し続ければ $RS(a_i, n)$ の値は減少し他の行動の RS の値は変化しないので局所解 (最適でない行動を選択し続けること) に陥ることなく，いつかは別の行動を選択する．最適な行動であれば $E[\Delta RS(a_i, n)] > 0$ となり，平均的にはそこにとどまるようになる．$E[\Delta RS(a_i, n)] = p_i - \aleph$ であるため最適でない行動を 1 回選択し RS の値が減少していく平均速度は報酬確率 p_i と基準 $\aleph$ の差が大きいほど大きい．つまり平均的に見れば報酬確率が小さい行動ほど一度その行動を選択したとしても早く見切りをつけて他の行動に移ることになる．

RS の特徴を明らかにするため，他の基本的なアルゴリズムについても同様の解析を行ってみよう．先ず単純に報酬平均を価値関数とする場合を考える．価値関数は $V(a_i, n) = n_i^1 / n_i$ なので 行動 a_i を選択する時，$E[\Delta Q(a_i, n)]$ は，

$$\begin{aligned} E[\Delta Q(a_i, n)] &= p_i\left(\frac{n_i^1 + 1}{n_i + 1} - \frac{n_i^1}{n_i}\right) + (1-p_i)\left(\frac{n_i^1}{n_i + 1} - \frac{n_i^1}{n_i}\right) \\ &= \frac{p_i - E_i}{n_i + 1} \end{aligned} \tag{4.10}$$

となる．選択しなかった行動の価値関数は変化しない．$E[\Delta Q(a_i, n)]$ は $p_i > E_i$ なら正，逆なら負であり，これは RS とは異なり，報酬確率 p_i の大きさに関わらず両方考えられる．行動 a_i の選択回数 n_i が大きければ $E[\Delta Q(a_i, n)] \approx 0$ となって $Q(a_i, n)$ はほとんど変化しなくなる．また，n_i が大きければ $E_i \approx p_i$ になる．このことは最適な行動を見つけられない可能性を示している．例えば行動が 2 つのケース ($p_1 > p_2$) で，最適な行動 a_1 の選択時にたまたま報酬があまり得られなかったため，$E_1 < p_2$，かつ $E_1 < E_2$ となった場合を考えてみよう．行動 a_2 の選択回数 n_2 が大きくなると E_2 の値は $E_2 \approx p_2$ でほとんど変わらなくなるが，$E_1 < p_2$ のため $E_1 < E_2$ の関係が固定されることになってしまう．そのため行動 a_2 の選択から抜けられなくなり，局所最適に陥いることになる．そこで ϵ_n-greedy では確率 ϵ_n でランダムに行動を選択することで局所最適を避けるようにしているが，いずれにせよ，RS のような方法で報酬確率が小さい行動ほど平均的に早く見切りをつけて他の行動に移る，とはいえない．

次に UCB 族の中で最も単純なアルゴリズム UCB1 について同じように $E[\Delta\mathrm{UCB1}(a_i, n)]$ を計算する．UCB1 の行動 a_i の step n での価値関数は

$$\mathrm{UCB1}(a_i, n) = E_i + \sqrt{\frac{2\ln n}{n_i}} \tag{4.11}$$

で表される．これより，行動 a_i を選択する時，

$$\begin{aligned}
E[\Delta\mathrm{UCB1}(a_i, n)] &= p_i\left\{\frac{n_i^1+1}{n_i+1} + \sqrt{\frac{2\ln(n+1)}{n_i+1}} - \left(\frac{n_i^1}{n_i} + \sqrt{\frac{2\ln n}{n_i}}\right)\right\} \\
&\quad + (1-p_i)\left\{\frac{n_i^1}{n_i+1} + \sqrt{\frac{2\ln(n+1)}{n_i+1}} - \left(\frac{n_i^1}{n_i} + \sqrt{\frac{2\ln n}{n_i}}\right)\right\} \\
&= \frac{p_i n_i - n_i^1}{(n_i+1)n_i} + \left(\sqrt{\frac{2\ln(n+1)}{n_i+1}} - \sqrt{\frac{2\ln n}{n_i}}\right) \\
&= \frac{p_i - E_i}{n_i+1} + \left(\sqrt{\frac{2\ln(n+1)}{n_i+1}} - \sqrt{\frac{2\ln n}{n_i}}\right)
\end{aligned} \tag{4.12}$$

となる．一方，この時，選択しない行動 a_j は，

$$\begin{aligned}
E[\Delta\mathrm{UCB1}(a_j, n)] &= E_j + \sqrt{\frac{2\ln(n+1)}{n_j}} - \left(E_j + \sqrt{\frac{2\ln n}{n_j}}\right) \\
&= \sqrt{\frac{2}{n_j}}\left(\sqrt{\ln(n+1)} - \sqrt{\ln n}\right)
\end{aligned} \tag{4.13}$$

となる．式 (4.12) の第 1 項は式 (4.10) と同じであり，第 2 項と第 3 項は行動 a_i が選択され続けるならばゼロになる．したがって，式 (4.12) だけでは前述のように局所最適に陥る可能性がある．しかし，UCB1 は選択しない行動 a_j も式 (4.13) のように価値関数が変化する．しかも，この選択しない行動の価値関数は式 (4.11) の第 2 項より，正の無限大に増加することが分かる．これにより，局所最適を回避している．

式 (4.12) で第 1 項は $p_i > E_i$ なら正，逆なら負であり，これは報酬確率 p_i の大きさに関わらず両方考えられる．第 2 項と第 3 項をまとめた括弧内は $n \geq 3$ であれば負になる (これは $f(x) = (\ln x)/x$ が $x > e\,(> 2)$ で単調減少であることから分かる)．よって $E[\Delta\mathrm{UCB1}(a_i, n)]$ は報酬確率の大きさに関わらず全体として正になることも負になることもある．したがって，*RS* のような方法で報酬確率が小さい行動ほど平均的に早く見切りをつけて他の行動に移る，とはいえない．

以上の解析を踏まえて本論文で導入した次の *RS* の式の意味を再考しよう．

$$RS_i = n_i\delta_i = n_i(E_i - \aleph). \tag{4.14}$$

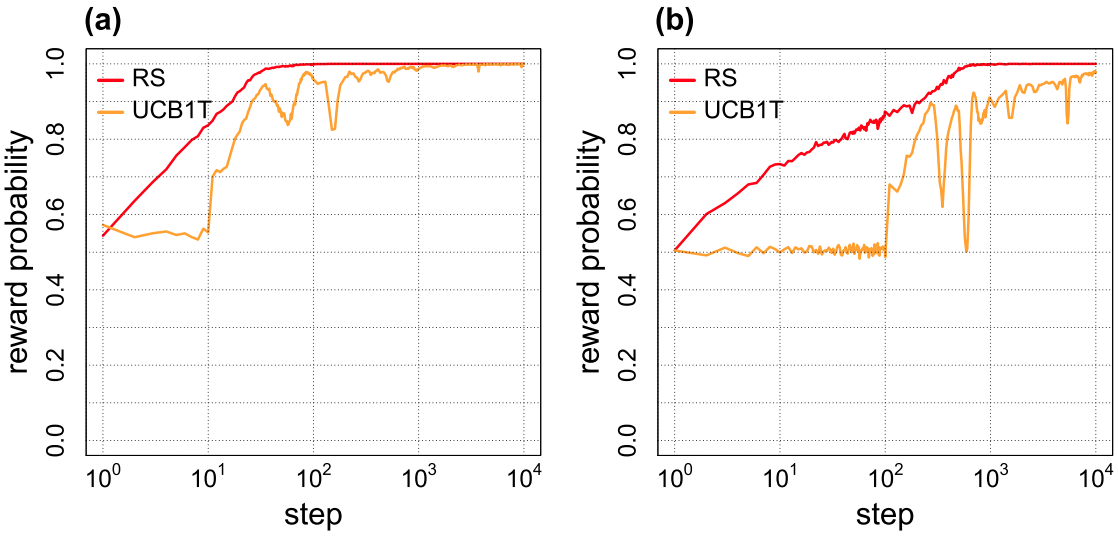

図 4.2　*RS* や UCB がどのような報酬確率の行動を平均的に選択するかを示したプロット．報酬確率は $[0,1]$ を K 等分した値．**(a)** $K = 10$ 及び **(b)** $K = 100$ のプロット．

もし δ_i だけを考慮するならば報酬平均 E_i に単に 定数 $(-\aleph)$ が付いただけであり，報酬平均を価値関数とする場合と挙動は変わらない．また，もし n_i だけを考慮するならば，$\aleph = 0$ とするのと同じで $n_i E_i = n_i^1$ となり，最初に報酬が得られた行動をずっと選択し続けることになってしまう．δ_i を作るのと，n_i を掛けることの両方があることによって，*RS* の値の変化の期待値 $E[\Delta RS(a_i, n)]$ が step 数 n によらない定数になる．そして，その定数は報酬確率 p_i から満足化基準 $\aleph$ を引いた値である．そのため基準を下回る報酬確率を持つ行動はふるい落とされることになる．また報酬確率が小さいほど平均的には早くふるい落とされるといえる．上で計算したように，UCB や ϵ_n-greedy にこのような性質はない．したがって，この性質が，最適切基準を利用した *RS* の性能が他の最適化アルゴリズムよりも優れている理由の 1 つと考えられる．

次に最適切基準を利用した *RS* がどのように行動を選択していくかをシミュレーションで確認する．行動数 K を $K = 10$ または $K = 100$ とし，報酬確率は $[0,1]$ を K 等分した値，つまり，$(0.1, 0.2, \ldots, 0.9, 1.0)$ または $(0.01, 0.02, \ldots, 0.99, 1.00)$ とした．満足化基準は $\aleph = 0.95$ または $\aleph = 0.995$ の最適切基準とした．1000 回シミュレーションして平均した結果は図 4.2 のとおりであった．縦軸は各 step において，どの報酬確率の行動が選択されたかの平均的な傾向を示す．*RS* は step 数が増加するにつれて選択する行動の報酬確率も概ね増加しており，報酬確率が最大の行動 (この場合は 1.0) を選択した後，そのまま選択し続けていることが分かる．(ただし，あくまで平均化された結果での傾向であ

り，1 回ごとのシミューションで必ずそういった選択をしているという意味ではない．) 一方，UCB は報酬確率が高い行動を選択した後であっても，そのまま選択し続けるのではなく，報酬確率が低い行動を選択してしまうことがあることが分かる．(なお，UCB は最初は K step 目までは全ての行動を 1 回づつ順番に選択していくので K step 目までは報酬確率の値はほぼ変わっていない．) *RS* は最適化された満足化基準を利用してスムーズに最適な行動を探索し，一度選択した最適な行動を平均的に選択し続ける傾向があることが分かった．

第 5 章

既存モデルとの関係

ここでは最初に既存の満足化モデルを紹介する．満足化の原理をモデル化した研究は多くなく，それらも *RS* とコンセプトが異なることを示す．次に *RS* がもう一つのモデル S0 の一般化として見なせることを示す．S0 は競合的な評価がある種のタスクでは高い性能のカギとなるという考えに基づいている．最後に，*RS* と命題 2 の証明で参照した TOW ダイナミクスモデルを比較する．*RS* の式を変形することで *RS* は TOW ダイナミクスモデルの 1 つと同じ数学的形式を持つことを示す．しかし，*RS* に組み込まれた risk-sensitive な満足化の考えが概念的な明晰さと一般化を可能にするものと考えられる．

5.1　既存の満足化モデル

ここでは既存の満足化モデルを紹介し，既存のモデルや *RS* との違いも簡単に述べる．まず，本研究の枠組みに最も近いのは，Bendor らのヒューリスティクスとしての満足化の政治学における分析 [Bendor 09] であり，ベルヌーイ報酬確率の場合の 2 本腕バンディット問題の分析がある．しかし，分析の枠組みが経済学に近いため，*PS* に似たポリシー的モデルの，極限における挙動を中心に分析している．そのモデルが *PS* と違うのは，不満足な場合には (必ずではなく) ある確率で行動を切り替える，という確率パラメータを与えている点である．

最近の，かつ包括的な研究としては [Reverdy 17] がある．これは最適性のあるバンディットアルゴリズムや PAC 学習などを参照しながら，満足化 satisfice をその語源である満足 satisfy と十分 suffice の二つに分解し，通常のバンディット問題を包括するような一般的な問題設定と (オーダーが) 最適なアルゴリズムを提案するものである．アルゴリズムは標準的な UCB [Auer 02] を作りかえたものであるため，本研究の提案する価値関数との違いも前述の UCB との違いと同様である．また問題としては報酬がガウシアンの場合に議論を限っている．彼 らの研究では regret の概念を拡張したうえで，確率

$(1-\delta)$ で満足化基準を超える行動を見つけることを目標としたアルゴリズムを開発しており，$\delta > 0$ であれば regret が有限になることを証明している．

しかし，彼らの研究は，あくまで regret の定義を変更した結果であることに注意が必要である．彼らの研究は一定の確率 $(1-\delta)$ で満足化基準を超えているかどうかで regret を計算しており，δ の確率で発生する regret をゼロとする定義が採用されている．実際，$\delta = 0$ とすれば，確実に満足化基準を超えるかどうかで regret を計算するため通常のバンディット問題と同じ枠組みになるが，この場合は彼らのアルゴリズムの regret は理論限界と同じく対数オーダーで増大し，有限にはならない．RS は regret の定義を変更せずに有限の regret を達成している．したがって本研究と彼らの研究とは目的や問題設定が全く異なる．

以上より，満足化の強化学習あるいは近接領域における先行研究で提案されたアルゴリズム等については，本研究とは目的と枠組みが違って比較ができないか，あるいは *PS* や UCB1 との比較で十分であるため，本研究では直接は扱わない．

5.2　指標 *S0* の一般化としての *RS* 価値関数

本論文で扱った *RS* は [篠原 07] が導入した *S0* モデルを特殊な形式で含む．この *S0* モデルは後に *RS* (rigidly symmetric) モデルとよばれ [中野 08]，その後価値関数として用いられた [Takahashi 11]．*S0* を 2 本腕バンディット問題に用いたところ，より複雑な *LS* とほぼ同様な性能が示されている [Takahashi 11]．満足化の観点からのこれらの挙動の解析は 2013 年に最初に公表された [大用 13, 大用 15]．2011 年に満足化の要求水準が変数とされた [大用 11]．続いて 2012 年に 2 つから任意の数の行動へのモデルの一般化が提案された [Kohno 12]．しかし，*LS* は *RS* よりはるかに複雑で，解析もかなり間接的であった．ここでは *RS* と *S0* の等価性とその条件を示す．なお，本内容は [高橋 16] の記載と同じ内容である．[*1]

2 本腕バンディット問題で行動 A と行動 B があるときに行動 $X \in \{A, B\}$ の選択が報酬を与えた回数を a^1_X，報酬を与えなかった回数を a^0_X とする．通常の報酬平均が $a^1_X/(a^1_X + a^0_X)$ となるところ，行動 A，B の価値を

$$S0(A) = \frac{a^1_A + a^0_B}{a^1_A + a^0_B + a^0_A + a^1_B}, \tag{5.1}$$

$$S0(B) = \frac{a^1_B + a^0_A}{a^1_B + a^0_A + a^0_B + a^1_A} \tag{5.2}$$

[*1] TOW モデルとの関係を考察するうえで *RS* の満足化基準が持つ認知的意味を明らかにしておく必要があったため再掲した．

と与える．これは，A で報酬を得ることと B で報酬を得ないことを同一視するもので，$S0(A) = 1 - S0(B)$ が成り立つ．分母が共通なので価値の比較は結局，不等式

$$a_A^1 + a_B^0 > a_B^1 + a_A^0 \tag{5.3}$$

ならば A をそうでなければ B を選択することになる．これを考慮すると 3 つ目の行動 C を考えると推移律が成り立つことが分かる．つまり A の B に対する価値関数 $S0(A)$ を $S_{AB}(A)$ と表せば，$S_{AB}(A) < S_{BA}(B)$，$S_{BC}(B) < S_{CB}(C)$ ならば $S_{AC}(A) < S_{CA}(C)$ となる．このことから比較可能な行動数は $K = 2$ とは限らないことが分かる．不等式 (5.3) は

$$a_A^1 - a_A^0 > a_B^1 - a_B^0 \tag{5.4}$$

と変形でき，本論文中の記号を用いると $a_X^1 = n_X E_X$，$a_X^0 = n_X(1 - E_X)$ が成り立つことから

$$n_A E_A - n_A(1 - E_A) > n_B E_B - n_B(1 - E_B) \tag{5.5}$$

$$\Leftrightarrow \quad n_A(2E_A - 1) > n_B(2E_B - 1) \tag{5.6}$$

$$\Leftrightarrow \quad n_A(E_A - 0.5) > n_B(E_B - 0.5) \tag{5.7}$$

となり，式 (5.7) の両辺は 式 (2.4)，(2.3) で $\aleph = 0.5$ と定数にした形式と一致する．推移性から任意の行動の組の価値が比較可能であるため，それぞれの行動について独立に式 (2.4) を計算し，その値が最大のものを選べば良いことになる．

このように RS は $S0$ を行動数 $K = 2$ から任意の $K \geq 2$ へと，また満足化基準として働く定数 0.5 を変数 $\aleph \in [0, 1]$ として一般化したものともいえる．

5.3 TOW と RS の類似点と相違点，RS の優位点

命題 2 の regret の証明で TOW ダイナミクスモデルに言及した [Kim 15, Kim 16, Kim 18](以下，簡単に TOW という)．ここで，RS と TOW を比較し，TOW に対する RS の相対的な利点について記述する．およそ 2010 年から TOW の研究 [Kim 10] が始まったが，TOW には様々なバリエーションが存在する．ここでは，RS に最も近い最近の論文 [Kim 16, Kim 18] で提案された TOW に注目する．行動 k を i 回目に選択して得られる報酬を X_k, i で表す．TOW では行動 k の価値関数 S_k は以下の形で表せることになる：

$$\begin{aligned} S_k &= X_{k,1} + X_{k,2} + \cdots + X_{k,n_k} - Kn_k \\ &= (X_{k,1} - K) + (X_{k,2} - K) + \cdots + (X_{k,n_k} - K), \end{aligned} \tag{5.8}$$

ここで K はあるパラメータである．

行動 k の選択回数を n_k，行動 k を選択して得られた報酬平均を E_k とする．この時，$E_k = \sum_{i=1}^{n_k} X_{k,i}/n_k$ となる．これまでは報酬の確率分布をベルヌーイ分布としていたが，TOW の論文 [Kim 16, Kim 18] に従い，ここではそれに限定しないものとする．

RS では行動 k の価値関数 RS_k は以下のように変形できる．

$$\begin{aligned} RS_k &= n_k(E_k - \aleph) \\ &= \sum_{i=1}^{n_k} X_{k,i} - n_i R \\ &= (X_{k,1} - \aleph) + (X_{k,2} - \aleph) + \cdots + (X_{k,n_k} - \aleph). \end{aligned} \tag{5.9}$$

式 (5.8) と式 (5.9) により，パラメータ K を満足化基準 $\aleph$ に対応させれば，TOW と式変形した *RS* は同じ形をしていることが分かる．このため TOW の regret の計算方法が *RS* でも利用でき，TOW と同じく上に有界であることが示せたのである．ただし，3.2 でも述べたように TOW の論文では 2 本腕で報酬確率の分散が同じケースに限定して分析をしている．

しかし，*RS* と TOW に相違点があり，*RS* は TOW に比べて優れている点を持っている．最も大きな違いは *RS* と TOW のバックグラウンドである．*RS* は人間の意思決定の仕方 (満足化) についてのモデルであり，5.2 で示したように因果推論での価値関数 *S*0 を一般化したモデルでもある．TOW は体積保存のような物理法則から導出されている．このような背景から両モデルはそれぞれ独立に考案されたものである．TOW と比較した *RS* の優れている点はその明晰さ，一般性にある．明晰さに関しては，*RS* は「得られた情報の信頼性」と「満足化の度合い」の積の形であり，パラメータ $\aleph$ は内在的な満足化基準としての意味を持つ．一方，*RS* の $\aleph$ に対応する TOW のパラメータ K の意味と解釈は必ずしも明確ではない．*RS* はこの 2 つの要素の概念を一般化することによって，K 本腕バンディット問題だけでなく，一般的な強化学習の設定にも適用可能である [高橋 16].

第 6 章

満足化基準が変化する場合の命題の拡張

6.1　拡張された命題とその証明

命題 1 や命題 2 では満足化基準 $\aleph$ は固定されていることが前提であった．しかし，基準 $\aleph$ が step ごとに変化する場合でも (確率的に変化する場合でさえ)，その変化が一定の範囲内であれば，証明を少し変更するだけで同様の命題が成立することが分かる．ここでは基準 $\aleph$ が変化する場合に成り立つよう命題 1 と命題 2 を拡張し、それらの証明を示す．またシミュレーションによる確認も行う．

命題 3 (命題 1 の基準 $\aleph$ を変数にしたバージョン)**.** 命題 1 において基準 $\aleph$ が時間的・確率的に変化しても集合 A_U，集合 A_L が変わらないとする．すなわち，変化する $\aleph$ の最小値，最大値を $\aleph_{min}$，$\aleph_{max}$ とし，A_L に属する報酬確率の最大値を p_l，A_U に属する報酬確率の最小値を p_u として，$p_l < \aleph_{min} \leq \aleph \leq \aleph_{max} < p_u$ とする．このとき元の命題 1 はそのまま成立する．

証明. Claim A の証明は大数の法則を使う箇所を次のように変えれば良い．$P\big(|E(a_i, s) - p_i| < (\aleph_{min} - p_i)/2\big) > 1 - \epsilon$ として $|E(a_i, s) - p_i| < (\aleph_{min} - p_i)/2$ ならば，

$$\begin{aligned} RS(a_i, s) &= n_i(s) \cdot \big(E(a_i, s) - \aleph\big) \\ &< n_i(s) \cdot \big(p_i + \frac{\aleph_{min} - p_i}{2} - \aleph_{min}\big) \\ &= n_i(s) \cdot \frac{p_i - \aleph_{min}}{2} < 0. \end{aligned} \tag{6.1}$$

以降は元の Claim A の証明と同じ．

Claim B の証明はそのままでよい．

命題 1 の証明は大数の法則を使う箇所を次のように変えれば良い．$P\big(|E(a_k,s)-p_k|<(p_k-\aleph_{max})/2\big)>1-\epsilon$ として $|E(a_k,s)-p_k|<(p_k-\aleph_{max})/2$ ならば，

$$\begin{aligned} RS(a_k,s) &= n_k(s)\cdot\big(E(a_k,s)-\aleph_{max}\big) \\ &> n_k(s)\cdot\big(p_k+\frac{\aleph_{max}-p_k}{2}-\aleph_{max}\big) \\ &= n_k(s)\cdot\frac{p_k-\aleph_{max}}{2}>0. \end{aligned} \tag{6.2}$$

以降は元の証明と同じである．□

命題 4 (命題 2 の基準 $\aleph$ を変数にしたバージョン)**.** 満足化基準 $\aleph$ が $p_2<\aleph<p_1$ を満たすように時間的・確率的に変化する場合，元の命題 2 と同じく regret は上に有界となる．

証明. 基準 $\aleph$ を $p_2<\aleph<p_1$ を満たす変数とし $\aleph$ の最大値，最小値を $\aleph_{max}$，$\aleph_{min}$ とする．$(p_2<\aleph_{min}\leq\aleph\leq\aleph_{max}<p1)$ $p_1>p_2>p_i(i\neq1,2)$ とし，$RS(a_i,s)=n_i(s)\cdot\big(E(a_i,s)-\aleph\big)$ $(i=1,2,\ldots,K)$ とする．ここで，$RS_{bd}(a_1,s)$ を

$$\begin{aligned} RS_{bd}(a_1,s) &= n_1(s)\cdot\big(E(a_1,s)-\aleph_{max}\big) \\ &\leq n_1(s)\cdot\big(E(a_1,s)-\aleph\big) \\ &= RS(a_1,s) \end{aligned} \tag{6.3}$$

と定義する．また，$i\neq1$ では $RS_{bd}(a_i,s)$ を

$$\begin{aligned} RS_{bd}(a_i,s) &= n_i(s)\cdot\big(E(a_i,s)-\aleph_{min}\big) \\ &\geq n_i(s)\cdot\big(E(ai,s)-\aleph\big) \\ &= RS(a_i,s) \end{aligned} \tag{6.4}$$

と定義する (bd は境界 (bound) の意味)．したがって $\aleph_i=\aleph_{max}$ $(i=1),\aleph_{min}$ $(i\geq2)$ とすれば $RS_{bd}(a_i,s)=n_i(s)\cdot\big(E(a_i,s)-\aleph_i\big)$ $(i=1,2,\ldots,K)$ となる．

$RS_{bd}(a_i,s)$ の期待値 E と分散 V は，$E[RS_{bd}(a_i,s)]=n_i(s)(p_i-\aleph_i)$，$V[RS_{bd}(a_i,s)]=n_i(s)\sigma_i^2$. ただし $\sigma_i^2=p_i(1-p_i)$ となる．ここで $\Delta RS_i(s)=RS(a_1,s)-RS(a_i,s)$ $(i\neq1)$，$\Delta RS_{bd,i}(s)=RS_{bd}(a_1,s)-RS_{bd}(a_i,s)$ $(i\neq1)$ と定義すれば，$RS(a_1,s)\geq RS_{bd}(a_1,s)$ かつ $RS(a_i,s)\leq RS_{bd}(a_i,s)$ $(i\neq1)$ より，$\Delta RS_i(s)\geq\Delta RS_{bd,i}(s)$ が成立することに注意しておく. さて，$\Delta RS_{bd,i}(s)$ の期待値 $E[\Delta RS_{bd,i}(s)]$ は，

$$\begin{aligned} E[\Delta RS_{bd,i}(s)] &= n_1(s)(p_1-\aleph_{max})-n_i(s)(p_i-\aleph_{min}) \\ &= n_1(s)(p_1-\aleph_{max})+n_i(s)(\aleph_{min}-p_i) \\ &\geq n_1(s)(p_1-\aleph_{max})+n_i(s)(\aleph_{min}-p_2)\ (\because i\neq1,\ p_2\geq p_i) \\ &\geq (n_1(s)+n_i(s))\min\big((p_1-\aleph_{max}),(\aleph_{min}-p_2)\big) \\ &(\because p_1-\aleph_{max}>0,\aleph_{min}-p_2>0) \end{aligned} \tag{6.5}$$

となる．[*1]

命題 3 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，$E[\Delta RS_{bd,i}(s)] \geq s \cdot \min\big((p_1 - \aleph_{max}), (\aleph_{min} - p_2)\big)$ となる．また，$\Delta RS_{bd,i}(s)$ の分散 $V[\Delta RS_{bd,i}(s)]$ は，$V[\Delta RS_{bd,i}(s)] \leq (n_1(s) + n_i(s))\sigma_{1,i}^2 \leq s\sigma_{1,i}^2$ となる．ただし $\sigma_{1,i} = \max(\sigma_1, \sigma_i)$ とする．

$\Delta RS_{bd,i}(s)$ は中心極限定理により，期待値 $E[\Delta RS_{bd,i}(s)]$，分散 $V[\Delta RS_{bd,i}(s)]$ の正規分布に従う．$\Delta RS_{bd,i}(s) \leq 0$ となる確率は，$Q(E[\Delta RS_{bd,i}(s)]/\sqrt{V[\Delta RS_{bd,i}(s)]})$ 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，

$$
\begin{aligned}
P[s = n+1, I = i] &\leq P[RS(a_j, n) \leq RS(a_i, n) \ \ (\forall j \neq i)] \\
&\leq P[\Delta RS_i(n) \leq 0] \\
&\leq P[\Delta RS_{bd,i}(n) \leq 0] \ \ (\because \Delta RS_i(s) \geq \Delta RS_{bd,i}(s)) \\
&= Q\left(\frac{E[\Delta RS_{bd,i}(n)]}{\sqrt{V[\Delta RS_{bd,i}(n)]}}\right) \\
&\leq Q\left(\frac{\sqrt{n}\min\big((p_1 - \aleph_{max}), (\aleph_{min} - p_2)\big)}{\sigma_{1,i}}\right) \\
&= Q(\phi_i\sqrt{n}). \qquad (6.6)
\end{aligned}
$$

ここで $\phi_i = \min\big((p_1 - \aleph_{max}), (\aleph_{min} - p_2)\big)/(\sigma_{1,i})$ と置いた．以降は，命題 1 と同じなので省略する．regret の上界は式 (3.22) の ϕ_i をこの ϕ_i に置き換えたものとなる． □

[*1] 注：命題 1 の証明同様，ここでは近似計算を行っているため，その誤差のために上界の値は正確なものでない可能性があるが，シミュレーションで確認できたように実際に上界として成立している．

6.2　シミュレーションによる検証

満足化基準が時間的に変動したり確率的に値が変わるケースについて，証明した性質が成り立つことをシミュレーションで確認する．2 本腕のケースで報酬確率は $p_1 = 0.6$ と $p_2 = 0.4$ とし，満足化基準 $\aleph$ が p_1 と p_2 の間を変動する 2 つのケースを考えた．1 つ目は基準 $\aleph$ が $[0.41, 0.59]$ の間を $\aleph = 0.5 + 0.09 * \sin(2 * \pi * \text{step}/100)$ の式に従って step ごとに周期的に変動するケースで，2 つ目は $[0.41, 0.59]$ の一様分布から毎 step で乱数を取って基準 $\aleph$ として設定するケースである．[*2] 比較のため満足化基準 $\aleph$ を中点の 0.5 に固定したケースも併せて示した．

1000 回のシミュレーションを平均した結果を図 6.1，図 6.2 に示す．どちらのケースも $\aleph$ が変動しても (a) accuracy が命題 3 のように 1 に収束すること，(b) regret は命題 4 で示した上界を超えていないこと，満足化基準を固定したケースの方が accuracy が早く 1 に収束し，regret は小さくなることが分かる．なお，図 6.1 の accuracy では振動しながら 1 に収束していく様子が見てとれるのは $\aleph$ が周期的に変動する影響と考えられる．

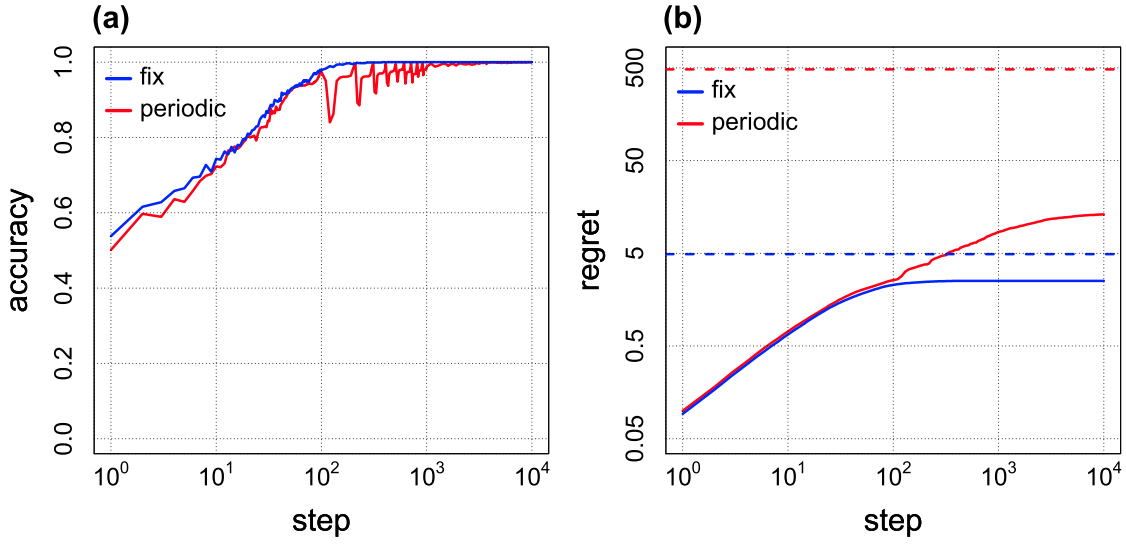

図 6.1　$K = 2$ の場合の RS のシミュレーション．報酬確率は $(0.4, 0.6)$．満足化基準 $\aleph$ は fix で 0.5 に固定，periodic で $[0.41, 0.59]$ の間を周期的に振動させた．$(\aleph = 0.5 + 0.09 * \sin(2 * \pi * \text{step}/100))$ **(a)** accuracy 及び **(b)** regret のプロット．**(b)** の上部の点線は命題 2 または命題 4 で計算された regret の上界を示す．

[*2] 注：$\aleph$ の変動範囲を $[0.4, 0.6]$ としなかったのは，命題 3 の証明で $p_l = \aleph_{min}$ であれば，$P\big(|E(a_i, s) - p_i| < (\aleph_{min} - p_i)/2\big) > 1 - \epsilon$ が成立しなくなるためである．$p_u = \aleph_{max}$ の時も同様に証明が成立しない．$\aleph$ が範囲に含まれないケース（$\aleph_{min} \leq p_l$ または $p_u \leq \aleph_{max}$）では 2 つの性質はいずれも成立しないことがシミュレーションで確認できる．

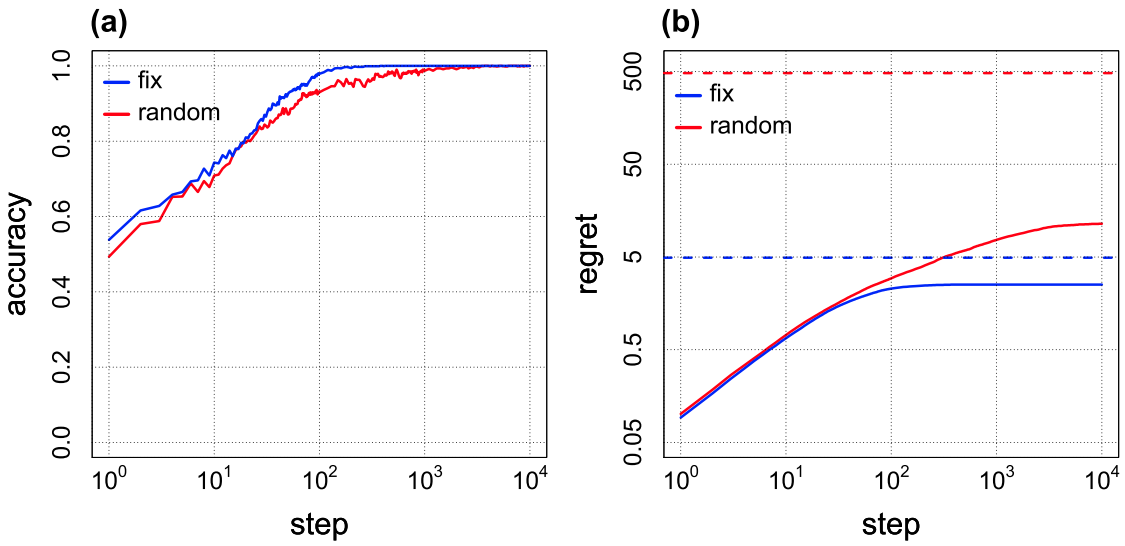

図 6.2　$K = 2$ の場合の *RS* のシミュレーション．報酬確率は $(0.4, 0.6)$．満足化基準 $\aleph$ は fix で 0.5 に固定，random で $[0.41, 0.59]$ の一様分布から step ごとに取って変化させた．**(a)** accuracy 及び **(b)** regret のプロット．**(b)** の上部の点線は命題 2 または命題 4 で計算された regret の上界を示す．

第 7 章

本研究の成果と今後の発展

本研究では，満足化の戦略を組み込んだ *RS* と呼ばれる単純な数理モデル [高橋 16, Oyo 17] について，最も基本的な強化学習課題である 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．以下では研究で得られた成果とその意義をまとめる．また，今後の研究の発展の可能性を述べる．

7.1　本研究の成果

本研究では *RS* をバンディット問題に適用した場合に，2 つの性質が理論的に成り立つことを示した．1 つ目は *RS* は満足化基準を満たす行動があれば，それを必ず見つけ出すことができるということである．2 つ目は，満足化基準を最適切基準に設定できる場合，regret は有限の値で抑えられるということである．1 つ目の性質は *RS* が満足化の戦略をモデル化したものである以上，必ず成立することが望まれる性質である．2 つ目の性質の成立については報酬確率についての事前情報が必要ではあるが，そのような情報が活用できる場合には *RS* は他の最適化アルゴリズムに比べれば非常に性能が良いことが分かった．

バンディット問題は現実世界ではインターネットにおける広告配信やレコメンデーション機能に応用されている．応用上，バンディット問題における選択可能な行動は配信される広告やレコメンドする商品の種類であり，報酬はユーザーのクリック数やサービスの購入数となり，サービス提供側からすれば報酬を最大にすることが目的となる．もし，蓄積された過去のデータから経験的に最適切基準を導出できる場合，*RS* を利用すれば他の最適化アルゴリズムよりも多くの報酬を得られることが理論的に保証されたことになる．

また，*RS* の性能が良い理由について価値関数の変化の期待値に着目して定性的に考察した．与えられた満足化基準と報酬確率の差を利用し，報酬確率が小さい行動ほど平均的に早く見切りをつけて他の行動に移ることが分かった．これは他のアルゴリズムにはない

RS に内在するメカニズムの一端を解明できたといえる．

更に前述した 2 つの性質を拡張し示した．満足化基準 ℵ が変動する場合であっても最大の報酬確率 p_1 と次に大きい報酬確率 p_2 の間の区間に ℵ が入っていれば，たとえ ℵ が確率的に変動する場合であっても，満足化の達成や有限の regret が成立することを理論的に示した．これまでの先行研究では *RS* の ℵ を最適切に設定する場合，p_1 と p_2 の中点とすることが多かったが，ℵ は中点でなくても良いと理論的に示せたことにまず意義がある．また逆に ℵ がどの区間に入っていれば regret が有限で抑えられるかをいえたことになる． 証明した拡張した命題はいずれも十分な step 数が経過した極限状態での性質であるため，有限の step 数での挙動は結果に影響を及ぼさない（ただし，regret の上界の値は命題 4 とは変わってくるが）．本研究では基本的に環境は時間的に変化しない定常状態を想定してきたが，たとえ非定常であっても環境変化の後に ℵ がその区間に入っているのであれば，regret は有限で収まることがいえたことになる．また，ℵ を自然環境などから決める場合，自然環境のゆらぎから ℵ 自体も変動する可能性があるが，その場合でも，その区間に ℵ が入っていればよいことが示せたことになる．

7.2　今後の発展

RS には本研究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．これは行動数の増加に伴い性能低下が起きる他の最適化アルゴリズムにはない優れた特徴である [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．実際，大車輪タスクに対して自律的かつ効率的な探索が可能であることが示されている [高橋 16]．

命題 2 や命題 4 では，満足化基準 ℵ を最適水準に設定できると仮定した．エージェントが常に最適切基準を得るとは限らないので，将来の研究方向はオンラインで最適切基準を学習できるアルゴリズムを開発することだろう．予備的な結果として，*RS* の性質を活用するアルゴリズムは Thompson サンプリング [Agrawal 12] と同等の性能を示した [甲野 18]．ただし，これまでのところ理論的には保証されていない．

今後は事前に設定している満足化基準の ℵ をオンラインで推定するアルゴリズムを開発し，本研究で行ったような理論的な分析を行うことが課題として残されている．今後，人間が実際に *RS* が表現するような満足化の原理に従って意思決定をしているかどうかの行動実験による検証も考えられる．特に満足化基準 ℵ を人間がどのように設定しているかが分かれば不確実な状況下での人間の意思決定の解明にもつながり，認知科学的にも意義深い研究となる．

今後考えられる RS の更なる応用としては並列化がある．例えば N 体のエージェントに基準 $\aleph_1, \aleph_2, \ldots, \aleph_N$ を与え，ある課題を並列に実行させる場合，$\aleph_i$ で成功して $\aleph_{i+1}$ で失敗するようなら，最適解が $[\aleph_i, \aleph_{i+1}]$ にありそうと分かる．既存の強化学習の手法のように乱数を使って行動を選択し，非常に多くの試行を繰り返して最適解を見つけるよりは，一定の時間的空間的な制限の下で基準を超えるかどうか試行し，可能であれば制限を少しずつ調整していく方法は，ある意味で人間的な学習方法と言え，人間が得意な課題に利用できる可能性がある．

第 8 章

謝辞

本論文は著者が東京電機大学大学院の先端科学技術研究科情報学専攻在籍中に得た研究成果をまとめたものである．社会人である著者の指導を引き受けて下さり，研究にあたって多大な労力と時間を割いて終始丁寧にご指導いただいた高橋達二准教授に心からの感謝を表する．

第 9 章

参考文献

[Agrawal 12] Agrawal, S. and Goyal, N.: Analysis of Thompson Sampling for the Multi-armed Bandit Problem, in *Proceedings of the 25th Annual Conference on Learning Theory*, pp. 39.1–39.26 (2012)

[Auer 02] Auer, P., Cesa-Bianchi, N., and Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem, *Machine Learning*, Vol. 47, No. 2, pp. 235–256 (2002)

[Bendor 09] Bendor, J. B., Kumar, S., and Siegel, D. A.: Satisficing: A 'Pretty Good' Heuristic, *The B.E. Journal of Theoretical Economics*, Vol. 9, No. 1 (2009)

[Bubeck 12] Bubeck, S. and Cesa-Bianchi, N.: Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems, *Foundations and Trends in Machine Learning*, Vol. 5, No. 1, pp. 1–122 (2012)

[Genewein 15] Genewein, T., Leibfried, F., Grau-Moya, J., and Braun, D. A.: Bounded Rationality, Abstraction, and Hierarchical Decision-Making: An Information-Theoretic Optimality Principle, *Frontiers in Robotics and AI*, Vol. 2, p. 27 (2015)

[Gershman 15] Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B.: Computational rationality: A converging paradigm for intelligence in brains, minds, and machines, *Science*, Vol. 349, No. 6245, pp. 273–278 (2015)

[Kim 10] Kim, S.-J., Aono, M., and Hara, M.: Tug-of-war model for the two-bandit problem: Nonlocally-correlated parallel exploration via resource conservation, *Biosystems*, Vol. 101, No. 1, pp. 29–36 (2010)

[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, p. 083023 (2015)

[Kim 16] Kim, S.-J., Tsuruoka, T., Hasegawa, T., Aono, M., Terabe, K., and Aono, M.: Decision maker based on atomic switches, *AIMS Materials Science*, Vol. 3, No. 1, pp. 245–259 (2016)

[Kim 18] Kim, S.-J. and Takahashi, T.: Performance in Multi-Armed Bandit Tasks in Relation to Ambiguity-Preference Within a Learning Algorithm, *Frontiers in Applied Mathematics and Statistics*, Vol. 4, p. 27 (2018)

[Kohno 12] Kohno, Y. and Takahashi, T.: Loosely Symmetric Reasoning to Cope with The Speed-Accuracy Trade-off, in *SCIS-ISIS 2012*, pp. 1166–1171 (2012)

[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)

[Lewis 14] Lewis, R. L., Howes, A., and Singh, S.: Computational Rationality: Linking Mechanism and Behavior Through Bounded Utility Maximization, *Topics in Cognitive Science*, Vol. 6, No. 2, pp. 279–311 (2014)

[Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015)

[Muse 09] Muse, D. and Wermter, S.: Actor-Critic Learning for Platform-Independent Robot Navigation, *Cognitive Computation*, Vol. 1, No. 3, pp. 203–220 (2009)

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

[Reverdy 17] Reverdy, P., Srivastava, V., and Leonard, N. E.: Satisficing in Multi-Armed Bandit Problems, *IEEE Transactions on Automatic Control*, Vol. 62, No. 8, pp. 3788–3803 (2017)

[Siddique 15] Siddique, N. and Adeli, H.: Nature Inspired Computing: An Overview and Some Future Directions, *Cognitive Computation*, Vol. 7, No. 6, pp. 706–714 (2015)

[Silver 16] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, van den G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, No. 7587, pp. 484–489 (2016)

[Simon 55] Simon, H. A.: A Behavioral Model of Rational Choice, *The Quarterly Journal of Economics*, Vol. 69, No. 1, pp. 99–118 (1955)

[Simon 56] Simon, H. A.: Rational choice and the structure of the environment., *Psychological Review*, Vol. 63, No. 2, pp. 129–138 (1956)

[Simon 57] Simon, H. A.: *Models of Man: Social and Rational*, John Wiley and Sons, Inc., New York (1957)

[Takahashi 11] Takahashi, T., Oyo, K., and Shinohara, S.: A Loosely Symmetric Model of Cognition, *Lecture Notes in Computer Science*, Vol. 5778, pp. 238–245 (2011)

[Tenenbaum 11] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D.: How to Grow a Mind: Statistics, Structure, and Abstraction, *Science*, Vol. 331, No. 6022, pp. 1279–1285 (2011)

[Tversky 81] Tversky, A. and Kahneman, D.: The framing of decisions and the psychology of choice, *Science*, Vol. 211, No. 4481, pp. 453–458 (1981)

[Zhao 18] Zhao, F., Zeng, Y., Wang, G., Bai, J., and Xu, B.: A Brain-Inspired Decision Making Model Based on Top-Down Biasing of Prefrontal Cortex to Basal Ganglia and Its Application in Autonomous UAV Explorations, *Cognitive Computation*, Vol. 10, No. 2, pp. 296–306 (2018)

[甲野 18] 甲野 佑, 高橋 達二：満足化を通じた最適な自律的探索, JSAI2018 (2018 年度人工知能学会全国大会 (第 32 回) 予稿集, pp. 1Z3–04 (2018)

[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)

[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, A Model of Belief Formation Based on Causality and Application to N-armed Bandit Problem, 人工知能学会論文誌, Vol. 22, No. 1, pp. 58–68 (2007)

[大用 11] 大用 庫智, 甲野 佑, 高橋 達二：非定常 N 本腕バンディット問題に対する人間の認知バイアスの適用, JSAI2011 (2011 年度人工知能学会全国大会 (第 25 回) 予稿集, pp. 1G1–2in (2011)

[大用 13] 大用 庫智, 高橋 達二：知識利用と探索のジレンマに対する因果的価値関数の適用とそのベイズ的分析, JSAI2013 (2013 年度人工知能学会全国大会 (第 27 回) 予稿集, pp. 1L4–OS–24b–4 (2013)

[大用 15] 大用 庫智, 市野 学, 高橋 達二：緩い対称性を持つ因果的価値関数の認知的妥当性と n 本腕バンディット問題におけるその有効性, 人工知能学会論文誌, Vol. 30, No. 2, pp. 403–416 (2015)

[中野 08] 中野 昌宏, 篠原 修二：対称性バイアスの必然性と可能性：無意識の思考をどうモデル化するか, 認知科学, Vol. 15, No. 3, pp. 428–441 (2008)