確率的満足化における最適な基準値の動的推定

Dynamic estimation of optimal aspiration level in Stochastic Risk-sensitive Satisificing

久米 淳 *1 Jun Kume　鈴木 裕毅 *1 Hiroki Suzuki　加藤 暦雄 *2 Toshikatsu Kato　甲野 祐 *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

Artificial intelligence technology has historically been developed by imitating certain aspects of neurophysiological and cognitive properties. In fact, although humans are apparently irrational, they are able to perform quick and congruent search under limited information. We believe that cognitive satisficing is involved in this quick search, and have developed an algorithm, Risk-sensitive Satisficing (RS), which can be applied to search in unknown environments such as the setting of reinforcement learning. Since RS is a deterministic search, it has difficulties in robustness to environmental noise and application to algorithms using probability distributions. To cope with these difficulties, Stochastic Risk-sensitive Satisficing (SRS), which expresses the search ratio inherent in RS as a probability distribution, was devised. However, it is debatable whether SRS retains the excellent characteristics that RS had in many cases. In this study, we examined the definition of congruence, which is one of the tasks of satisficing strategies, in short, the dynamic estimation of the optimal aspiration reward level can be performed in SRS for the bandit problem, and showed that it is possible to achieve both a quick search for congruent means and optimization.

1. はじめに

人工知能技術は歴史的に神経生理的・認知的性質のある側面の模倣により発展してきた．実際，人間の行動原理は，能力や情報収集の限界から限定された情報下で素早く探索を行い，合理的に判断する満足化が妥当であると考えられている [Simon 56]．この素早い探索に認知的満足化が関与していると考え，これらを数理化・分析を行い，強化学習のような未知環境への探索に応用可能なアルゴリズム Risk-sensitive Satisficing (RS) が考案された [高橋 16]．RS は一定の基準値 $\aleph$ (aleph; アレフ) が設定されており，それを超える選択肢を素早く発見するアルゴリズムである．しかし RS は決定論的な方策であり，強化学習一般で用いられる確率分布を利用したアルゴリズムへの応用に困難がある．そこで RS に内在する探索比率を確率分布として表現した確率化認知的満足化方策 Stochastic Risk-sensitive Satisficing (SRS) が考案された [加藤 21]．しかし SRS は RS が有していた優秀な特性を全て維持しているかは議論の余地がある．RS は Chernoff-Hoeffding bound[Lai 85] を用いることによってタスク中のオンライン情報のみから最適な基準値を動的に推定することが可能である [甲野 18]．

そこで本研究ではその性質が SRS においても適応できるのかを検証した．

2. バンディット問題

バンディット問題とは，報酬確率 p_i とそれに対応する行動の選択肢 a_i が存在し，毎回行動を選択することで得られる累積報酬の最大化を目的とした意思決定課題である．最大化のためには様々な行動を繰り返し選択することで最適な選択肢を探索する必要がある．しかし，探索を優先すると累積報酬の最大化が見込めない．また，探索を疎かにし，少ない情報のみで最良の行動を選択する活用を行うと最適な行動が選べず，局所解に陥ってしまう．この探索と活用のバランシングがバンディット問題の課題となる．

本研究では，報酬 r は任意の選択肢 a_i を試行すると報酬確率 p_i で $r = 1$，報酬確率 $1 - p_i$ で $r = 0$ が与えられるベルヌーイ試行からなるベルヌーイバンディットを扱う．

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，049-296-0394，tatsujit@mail.dendai.ac.jp

3. 満足化方策 RS

満足化方策 RS は式 (2)，式 (3) で表される．ここでの満足化基準値 $\aleph$ (以下，基準値とする) は満足化における基準として使われており，事前知識や経験をもとに設定することができる．V_i は経験期待値，n_i は試行量，N は総試行量を表しており，各種算出方法は 3.1 節にて後述する．また，各選択肢の試行量と総試行量の比率である試行量割合を ρ_i とする．

$$RS_i = \frac{n_i}{N}(V_i - \aleph) \quad (1)$$

$$= \rho_i(V_i - \aleph) \quad (2)$$

$$a^{\text{select}} = \arg\max_i(RS_i) \quad (3)$$

満足化方策 RS は式 (2) から算出された RS 値の中で一番大きい値を持つ行動を選択する．最大の RS 値が複数あった場合，その選択肢の中でランダムに選択する．

RS は全ての選択肢が基準値を下回る非満足状況 ($\max_i V_i < \aleph$) の時，試行量割合が低い選択肢の RS 値が高くなることで探索を促す．一方で，ある選択肢が基準値を上回る満足状況 ($\max_i V_i > \aleph$) の時は試行量割合が高い選択肢の RS 値を高くなることで活用を促す．このように設定することで探索の間は試行量の少ない行動の選択頻度を増やし，活用の間は試行量割合が高い，すなわち信頼性が高い行動を選択することができる．

3.1 経験期待値と試行量の更新

バンディット問題において，RS 値を算出するためには経験期待値 V と試行量 n が必要であり，式 (5)，式 (6) より更新

される．ここで α は学習率，K は選択肢数，N は総試行量を示しており，学習率 α は試行量 n が増加していくほど減少していく．

$$\alpha \leftarrow \frac{1}{1+n^{\text{select}}} \tag{4}$$

$$V^{\text{select}} \leftarrow (1-\alpha)V^{\text{select}} + \alpha r \tag{5}$$

$$n^{\text{select}} \leftarrow n^{\text{select}} + 1 \tag{6}$$

$$N = \sum_{i=1}^{K} n_i \tag{7}$$

ここで，n^{select} は選択された行動の試行量を表し，V^{select} は選択された行動の経験期待値を表している．

3.2 経験期待値と試行量の初期値

満足化方策 RS の定式において RS 値の算出過程でゼロ除算の発生を防ぐため，試行量 n_i にはごく微小な値として ϵ を，経験期待値 V_i には 0.5 を初期値として代入する．n_i を試行回数ではなく，試行量として表現しているのはこの初期値のためである．

$$n_i \leftarrow \epsilon \tag{8}$$

$$V_i \leftarrow 0.5 \tag{9}$$

4. 非満足状況

バンディット問題において，経験期待値 V_i が $\max_i V_i > \aleph$ を満たすとき，これを満足状況であると定義する．一方で，経験期待値 V_i が $\max_i V_i < \aleph$ を満たすとき，これを非満足状況であると定義する．満足状況と非満足状況のイメージ図はそれぞれ図 1，図 2 に示す．図 1，図 2 より，満足状況では一つ以上の選択肢が基準値を満たしており，非満足状況では全ての経験期待値が基準値より低いことが分かる．ここで，各選択肢の持つ経験期待値のうち，最大の経験期待値を V_G とし，V_G を持った選択肢の試行量割合を ρ_G とする．また，試行量割合 ρ_i のうち，ρ_G を除いた試行量割合を ρ_j とする．

非満足状況下では，V_G が基準値 $\aleph$ よりも低くなるため，式 (2) より，試行量割合 ρ_G が上昇するほど，その選択肢から算出される RS 値は低下していく．一方，ρ_G の上昇に伴い，ρ_G 以外の試行量割合 ρ_j は低下していくため，他の RS 値は上昇していく．RS 値の上昇率は試行量割合 ρ_i が増加していくほど減少していくため，総試行量 N が充分大きくなると RS 値は一意に定まる．これを RS 均衡と呼び，RS 均衡に陥った際の RS 値を RS 均衡値 $-Z$ と定義する [甲野 18]．また，RS 均衡のイメージ図は図 3 に示す．図 3 の左図は探索初期の状況を表し，非満足状況で探索を続けていくと，図 3 の右図のように RS 値は RS 均衡値 $-Z$ に均衡していく．

総試行量 N が充分大きいとき，試行量割合 ρ_i を式 (11) より RS 均衡値 $-Z$ を用いて逆算することができる．

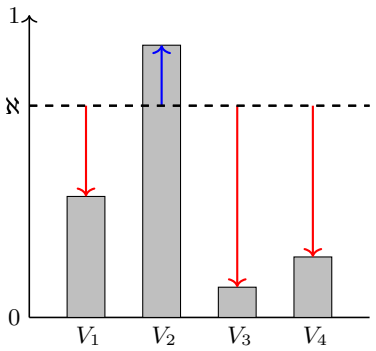

図 1: 満足状況

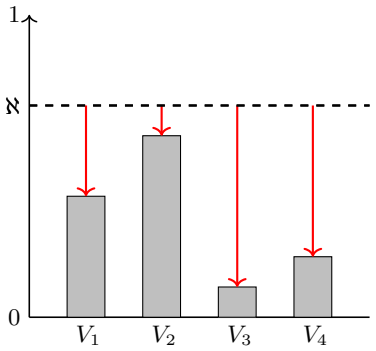

図 2: 非満足状況

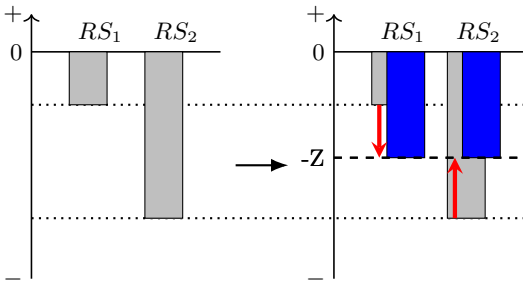

図 3: RS 均衡

$$RS_i = -Z \tag{10}$$

$$\rho_i = \frac{n_i}{N} = \frac{Z}{\aleph - V_i} \tag{11}$$

また RS 均衡値 $-Z$ は，試行量割合 ρ_i の総和が 1 になるこ

とから算出できる．

$$\begin{aligned}\sum_{i=1}^{K}\rho_i &= \sum_{i=1}^{K}\frac{n_i}{N}\\ &= \sum_{i=1}^{K}\frac{Z}{\aleph - V_i}\\ &= 1 \qquad (12)\end{aligned}$$

$$Z = \frac{1}{\sum_{i=1}^{K}\frac{1}{\aleph - V_i}} \qquad (13)$$

このように RS 均衡値 $-Z$ が式 (13) から算出できるため，試行量割合 ρ_i は基準値 $\aleph$ と経験期待値 V_i から定義できる．特に RS 均衡値 $-Z$ を用いて求められる試行量割合を ρ_i^z とし，これを理想試行量割合と呼ぶ．

4.1 非満足状況下における基準値の最適化

非満足状況下では，greedy な選択肢 a_G とそれ以外の選択肢 a_j の RS 値が RS 均衡によっていずれも等しくなることから式 (14) が成り立つ．ここで，式 (1) を代入することで式 (15) が得られ，$\aleph$ について解くと式 (16) が求まる．

$$RS_G = RS_j \qquad (14)$$

$$\frac{n_G}{N}(V_G - \aleph) = \frac{n_j}{N}(V_j - \aleph) \qquad (15)$$

$$\aleph = V_G\frac{1 - \frac{V_j}{V_G}\frac{n_j}{n_G}}{1 - \frac{n_j}{n_G}} \qquad (16)$$

式 (16) より，greedy な選択肢 a_G とそれ以外の任意の選択肢 a_j の理想的な選択比率 n_j^*/n_G^* の関係が分かると基準値 $\aleph$ を逆算することができる．そこで本項は greedy な選択肢 a_G とそれ以外の任意の選択肢 a_j の潜在的な選択比率 $\mu = n_j/n_G$ を最適化することを目的とし，その最適な選択比率 μ^* を Chernoff-Hoeffding bound[Lai 85] を用いて推定する．最適な潜在的選択比率 μ^* は式 (17) で表すことができる．ここで，ρ_G^* と ρ_j^* は最適な試行量割合を示す．

$$\begin{aligned}\mu_j^* &= \frac{n_j^*}{n_G^*} = \frac{n_j^*}{N}\frac{N}{n_G^*}\\ &= \frac{\rho_j^*}{\rho_G^*} \qquad (17)\end{aligned}$$

最適な試行量割合 ρ_j^* は現時点で greedy な選択肢 a_G が最良の選択肢ではなく $V_G^* \leq V_j^*$ となる確率に等しいと定義し，Chernoff-Hoeffding bound から式 (18) を定義した．

$$\begin{aligned}\rho_j^* &= \frac{n_j^*}{N}\\ &= \Pr(V_j^* \geq V_G^*)\\ &= \exp\bigl(-n_j \mathrm{D}_{\mathrm{KL}}(V_j||V_G)\bigr) \qquad (18)\end{aligned}$$

また，任意の選択肢が a_G の場合，ρ_G^* は 1 となる．

$$\begin{aligned}\rho_G^* &= \frac{n_G^*}{N}\\ &= \Pr(V_G^* \geq V_G^*)\\ &= \exp\bigl(-n_G \mathrm{D}_{\mathrm{KL}}(V_G||V_G)\bigr)\\ &= 1 \qquad (19)\end{aligned}$$

よって，任意の選択肢に対して Chernoff-Hoeffding bound より算出される最適な潜在的選択比率 $\mu_j^{\mathrm{CH}} \approx \mu_j^*$ は式 (20)，最適な基準値 $\aleph^{\mathrm{CH}}$ は式 (21) から定義される．

$$\begin{aligned}\mu_j^{\mathrm{CH}} &= \frac{n_j^*}{N}\frac{N}{n_G^*}\\ &= \exp\bigl(-n_j \mathrm{D}_{\mathrm{KL}}(V_j||V_G)\bigr) \qquad (20)\end{aligned}$$

$$\aleph^{\mathrm{CH}} = \max\left(V_G\frac{1 - \frac{V_j}{V_G}\mu_j^{\mathrm{CH}}}{1 - \mu_j^{\mathrm{CH}}}\right) \qquad (21)$$

算出された潜在的選択比率と基準値をそれぞれ最適潜在選択比率，非満足基準値と呼び，μ_j^{CH}，$\aleph^{\mathrm{CH}}$ とする．また，それらを用いた RS を RS-CH と呼ぶ．RS-CH は 2 本腕バンディットにおいて，Thompson Sampling[Agrawal 12, Thompson 33] と同等の成績が得られることが分かっている [甲野 18].

5. 確率的満足化方策 SRS

SRS は，RS 均衡値 $-Z$ から導出した理想試行量割合 ρ_i^z と現在の試行量割合 ρ_i における差分から確率分布を生成し，確率的に行動選択を行う方策である．選択確率であることから負の割合とゼロ除算の発生を防ぐ方法として調整パラメータ b, 及び ϵ を用いる．調整パラメータ b の更新式，SRS の定式，それらに従って導出される選択確率 π をそれぞれ式 (22)，式 (23)，式 (24) に示す．

$$b_i = \frac{n_i}{\rho_i} - N + \epsilon \qquad (22)$$

$$SRS_i = (N + b_{\max})\rho_i^z - n_i > 0 \qquad (23)$$

$$\pi_i = \frac{SRS_i}{SRS_1 + SRS_2 + \cdots + SRS_k} \qquad (24)$$

調整パラメータ b は，各選択肢について算出する．その中で最大の値を $b_{\max}$ として扱い，これを SRS 値の算出に用いる．

また，RS 均衡値 $-Z$ を用いることは，非満足状況であることが前提である．したがって，最大の経験期待値である V_G が $\aleph$ を超えた場合に対して一時的に式 (25) のような調整を行う．

$$V_i \leftarrow V_i - (V_G - \aleph) - \epsilon \qquad (25)$$

前章より SRS で用いられる基準値 $\aleph$ を非満足基準値 $\aleph^{\mathrm{CH}}$ とした方策，SRS-CH を提案する．

6. シミュレーション

6.1 実験概要

本研究の実験は 1 回の行動選択を 1 step とし，100,000 step の行動選択を行った．行動数は 2 であり，各々シード値の異なる 10,000 回のシミュレーションを行った．各選択肢の報酬確率は一様分布よりサンプリングされ，1 回のシミュレーションでは選択肢の報酬確率は固定される．検証に使用するアルゴリズムは RS-CH，SRS-CH，Thompson Sampling とした．評価指標は regret を用いた．regret とは，最適な選択肢を取り続けた場合と比べてどれくらいの差があるのかを示す値である．regret が低ければ低いほど報酬を最大化できていることを示す．regret の算出方法は以下の式 (26) である．

$$\mathrm{regret} = \sum_{t=1}^{N}(p_{\max} - p_t^{\mathrm{select}}) \qquad (26)$$

ここで，$p_{\max}$ は与えられた報酬確率で最も高い報酬確率であり，p_t^{select} は t 回目に選択した行動の報酬確率である．

6.2 実験結果

各アルゴリズムにおける regret の推移を表したものを図 4，最終的な regret の値を有効数字 4 桁に丸めたものを表 1 に示す．図 4，表 1 の結果から SRS-CH は，最終的な regret に注目すると，他の方策 RS-CH と Thompson Sampling と比較して同等の性能であることが分かった．

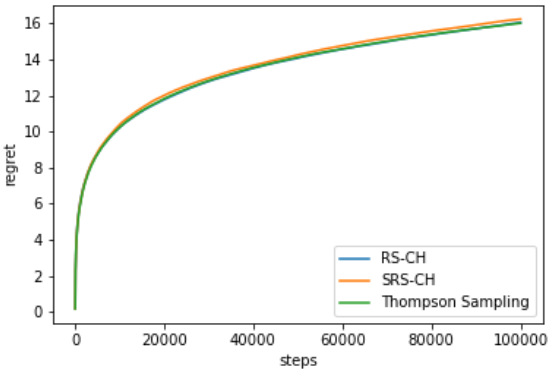

図 4: 2 本腕バンディットにおける regret の推移

表 1: 最終 regret の比較

方策	regret
RS-CH	16.05
SRS-CH	16.23
Thompson Sampling	16.01

7. 考察

図 4 の結果から，SRS-CH は既存方策 RS-CH と Thompson Sampling と同等の成績を得られた．仮説として，SRS-CH は確率論的な方策のため，必ずしも greedy な選択肢を選び続けるとは限らない．RS-CH と比較して regret がわずかに大きくなることを予想していた．しかし，予想に反して他方策に引けを取らない同等の成績が得られた．

本研究より 2 本腕の場合，SRS-CH は後半の方で決定論的な振る舞いをすると考えられる．そして，SRS には基準値 $\aleph$ の最適値を動的に算出する性質が保証されていることが分かった．これにより満足化と最適化の両方を連続的に行き来できるため，基準値を満たさない選択肢のみの場合でも最適化に切り替えることで，選択肢の中でより良い行動を選択することができる．

また，SRS-CH は 行動の選択分布を確率的にしたため，各選択肢の報酬確率が変化する非定常環境に対して RS-CH より素早く別の選択肢に切り替えられる柔軟性を有していると考えられる．

8. 結論

以上の結果より，確率分布を生成する方策 SRS において，基準値 $\aleph$ を動的更新することで本研究で提案した方策 SRS-CH は，既存方策 RS-CH，Thompson Sampling と同等の水準であり，非常に優れていることが分かった．先行研究では，基準値 $\aleph$ を事前知識や経験をもとに任意に設定していた．

本研究から，SRS に対して基準値 $\aleph$ に動的更新を適用することが可能であり，基準値を上回る選択肢の発見に優れていることが分かった．また，確率的な方策であるにもかかわらず決定的な振る舞いを示した．したがって，満足化と最適化の 2 つの性質を持っていることが言える．しかし，本研究では 2 本腕バンディットのみの検証であり，SRS-CH が多本腕バンディットに対しても決定的な振る舞いをするかは明らかでない．

今後の展望として，選択肢が多数ある多本腕バンディット問題や各選択肢の報酬確率が変化する非定常環境に適用して，SRS-CH の性質を明らかにしていくことが重要課題となる．

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).

[加藤 21] 加藤暦雄，甲野祐，高橋達二：満足化方策における非満足均衡を用いた確率的方策の検証，2021 年度 人工知能学会全国大会 (第 35 回) 論文集，(2021).

[甲野 13] 甲野祐，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 30th Japanese Cognitive Science Society (JCSS)*, 74–79, (2013).

[甲野 18] 甲野祐，高橋達二：満足化を通じた最適な自律探索，2018 年度 人工知能学会全国大会 (第 32 回) 論文集，(2018).

[Lai 85] Lai, T. L., Robbins, H.：Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics*, 6(1), 4–22, (1985).

[Simon 56] Simon, H.A.：Rational choice and the structure of the environment, *Psychological Review*, 63(2), 129–138, (1956).

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

[高橋 16] 高橋達二，甲野祐，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).

[玉造 18] 玉造晃弘，高橋達二：認知的満足化価値関数の分析：保証付き満足化と有限 regret，2018 年度 人工知能学会全国大会 (第 32 回) 論文集，(2018).

[Tamatsukuri 19] Tamatsukuri, A., Takahashi, T.:Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function, *Biosystems*, 180(June), 46–53, (2019).

[Thompson 33] Thompson, W. R.：On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, *Biometrika*, 25, 285–294, (1933).