正規分布報酬の Satisficing bandit における目標水準の最適調整

上野 誠弥 †，高橋 達二 †, 甲野 佑 †
Seiya Ueno, Tatsuji Takahashi, Yu Kono
† 東京電機大学
Tokyo Denki University
tatsujit@mail.dendai.ac.jp

概要

強化学習の応用例であるバンディット問題を解くアルゴリズムでは，事前分布など事前知識を与えることがある．その中でも人の試行錯誤の一種である満足化を再現した Risk-sensitive Satisficing (RS) では，最適な目標値を事前知識として与えた場合に高い性能が確認されていた．本研究では最適な目標値が未知の実数値になりうる場合でも，エージェントがオンライン情報から自律的に推定する方法を検討した．

キーワード：人工知能，機械学習，強化学習

1. はじめに

人間が行う試行錯誤の傾向として，一定の基準値を上回る行動を見つけるまで盛んに探索を行い，発見後利益追求に切り替える満足化 (Satisficing) と呼ばれる方策を行うと言われている (Simon, 1956)．実用上では最適化より目標達成を重視する満足化が重要なことがある．実際に強化学習の分野では Upper Confidence Bound (UCB) というバンディットアルゴリズムの提案者らにより，満足化を目的としたアルゴリズムが提案されている (Rouyer et al., 2024)．他にも強化学習全般に適用可能な満足化のアルゴリズムとして Risk-sensitive Satisficing (RS) が提案されている (高橋他, 2016)．本研究ではこの RS の拡張を行った．RS では希求水準という主観的な損益分岐基準を導入した価値関数を使用する．RS では 1 番目に大きい真の報酬と 2 番目に大きい真の報酬の平均値という最適な希求水準が与えられた場合に，後悔が定数オーダーで有限になることが証明されている (Tamatsukuri & Takahashi, 2019).

バンディットアルゴリズムではよくパラメータを用いることがあるが，パラメータが後悔の上界に影響を与えることもある (Bubeck & Cesa-Bianchi, 2012)．また Thompson Sampling (TS) でも適切な事前分布を与えることが重要だと示されている (Honda & Takemura, 2013)．RS における最適な希求水準は事前知識として直感的に与えやすい．ただしあまりに高い誤った事前知識（例えば明らかに無謀な営業目標）を与えられた際には，エージェント自らが無謀な希求水準を適切に調整する必要が出てくる．そこで本研究ではエージェントが自律的に最適な希求水準を導出する手法について検討する．先行研究において，オンライン情報のみから最適な後悔の度合いを実現する希求水準を算出する RS-CH を提案されている (甲野・高橋, 2018)．RS-CH では Chernoff-Hoeffding bound から選択割合に対する理想的な選択割合を推定し逆算することで，後悔の度合いを最小化する希求水準を算出することができる．

しかしこの希求水準の算出についての議論は，ベルヌーイ分布のような報酬が 0 or 1 の 2 値のみの課題に限定されているものでしか議論されてこられなかった．少なくとも報酬が正規分布に従うバンディット問題においては単純な漸化式のような形での希求水準の更新では後悔の度合いが大きくなってしまうことが示されている (小河他，2023)．そこで本研究では報酬が正規分布に従うバンディット問題でも RS-CH のような希求水準の算出が可能かどうか検討した．

2. 多腕バンディット問題

多腕バンディット問題は，強化学習における基本的な意思決定課題である．腕と呼ばれる複数の行動が存在し，各行動からは異なる確率分布に従う報酬値が得られる．そのような環境で，事前情報がない状態から試行錯誤を続けて累積報酬の最大化を目的とする．本研究では報酬が正規分布に従うより一般的バンディット問題について議論する．

2.1 正規分布に従う報酬のバンディット問題

バンディット問題は実用上，広告配信におけるクリック率のような値の範囲が限られた報酬分布ではなく，視聴時間のような任意の実数値の報酬分布を扱いたい場合も多い．また任意の実数値の報酬を与える環境でも使用できるアルゴリズムであれば，より複雑な環境の強化学習への一般化が容易になる．

2.2 評価指標

強化学習タスクでは，最適な行動を発見するためによく知らない行動の選択を促す探索が必要となる．その一方で累積報酬の最大化を行うには，探索を行わず現時点で最適だと考えられる行動を選択し続ける知識

利用もしなければならない．この探索と知識利用のトレードオフという課題にはいくつかの評価指標があり，本研究では式 (1) で示す Regret を評価指標として用いた．$E[r^*]$ は最も高い真の報酬期待値，$E[r_{\text{chosen}}]$ はエージェントが選択した行動の真の報酬期待値である．Regret は損失の期待値であり，最適な行動選択ができているほど小さな値となる．

$$\text{Regret} = \sum_{t=1}^{T} (E[r^*] - E[r_{\text{chosen}}]) \tag{1}$$

2.3 バンディットアルゴリズムの例

バンディット問題を解くアルゴリズムで代表的なものとして，Upper Confidence Bound (UCB) と Thompson Sampling (TS) などが挙げられる．

本研究で用いる UCB は正規分布報酬に対応した UCB1-Normal である (Auer et al., 2002)．n_i^t は時刻 t 時点の各行動の試行回数，N は総試行回数，$\bar{r}_i$ はエージェントが得た各行動の報酬の平均，q_i は各行動の報酬の 2 乗和である．式 (3) で定義される UCB スコアから greedy に時刻 t の行動 a_t を選択する．また UCB1-Normal では試行回数が $\lceil 8 \log n \rceil$ の腕があった場合にそれを選択するように定められている．

$$\text{Bonus}_i(t) = \sqrt{16 \cdot \frac{q_i^t - n_i^t \left(\bar{r}_i^t\right)^2}{n_i^t - 1} \cdot \frac{\ln(N_t - 1)}{n_i^t}} \tag{2}$$

$$\text{UCB-Normal}_i(t) = \bar{r}_i^t + \text{Bonus}_i(t) \tag{3}$$

$$i = \arg\max_i (\text{UCB1-Normal}_i(t)) \tag{4}$$

$$a_t = a_i \tag{5}$$

また TS は正規分布報酬に対応したものを使用する．特に今回は逆カイ 2 乗分布から分散をサンプリングし，その分散を用いて正規分布から報酬の予測値をサンプリングする手法を使用する．ν_i と s^2 はそれぞれ各選行動の自由度とスケールパラメータである．

$$\sigma_i^2 \sim \textit{Scale-inv-}\chi^2(\nu_i, s_i^2) \tag{6}$$

$$\text{TS}_i(t) \sim \mathcal{N}(\bar{r}_i^t, \frac{\sigma_i^2}{n_i^t + 1}) \tag{7}$$

$$i = \arg\max_i (\text{TS}_i(t)) \tag{8}$$

3. RS の定義

RS 価値関数の値は式 (9) で定義される．原則 RS ではこの評価値が最も高い行動を選択する．ここでの S は満足化における基準値のことであり，ここでは希求水準と呼ぶ．最適な希求水準は，1 番高い真の報酬期待値 $E[r^*]$ と 2 番目に高い真の報酬期待値 $E[r^{2\text{nd}}]$ を用いて式 (11) と定義される (Tamatsukuri & Takahashi, 2019)．常にその最適な希求水準を使用する手法を RS-opt と表記する．また式 (12) のような漸化式的に希求水準を更新する手法を RS-dyn と表記する (小河他，2023)．本研究では初期の希求水準を 0，学習率 $\beta = 0.001$ とする．

$$I_i^{\text{RS}} = \frac{n_i^t}{N_t}(\bar{r}_i^t - S) \tag{9}$$

$$i = \arg\max_i (I_i^{\text{RS}}) \tag{10}$$

$$S_{\text{opt}} := \frac{E[r^*] + E[r^{2\text{nd}}]}{2} \tag{11}$$

$$S_{\text{dyn}} \leftarrow S_{\text{dyn}} + \beta(\bar{r}_{\max}^t - S_{\text{dyn}}) \tag{12}$$

4. RS 平衡と目標値推定

RS の方策を実施するエージェントにとって全ての行動の報酬平均が希求水準を下回るときが存在し，このときを “未達成状況” と呼ぶ．この未達成状況では RS の価値関数は全ての行動に対して負の値の評価値を与えることになる．よって RS の定義により行動の試行割合が高いほどその行動の評価値は低い値となり，これまで選択されてこなかった行動が選択されやすくなる．その結果，未達成状況が継続した場合に全ての行動の RS の評価値はほぼ一定の値（$-\mathsf{E}$）となる．この現象を “RS 平衡”，その値 $-\mathsf{E}$ を “RS 平衡値” と呼ぶ．この平衡値から未達成状況では各行動の試行割合を式 (14) で逆算できる．

$$I_i^{\text{RS}} = \rho_i(\bar{r}_i - S) = -\mathsf{E} \tag{13}$$

$$\rho_i = \frac{n_i}{N} = \frac{\mathsf{E}}{S - \bar{r}_i} \tag{14}$$

また RS 平衡値は式 (15) のように各行動の試行割合の和が 1 になることから求められる．

$$\sum_{i=1}^{K} \rho_i = \sum_{i=1}^{K} \frac{\mathsf{E}}{S - \bar{r}_i} = 1$$

$$\mathsf{E} = \frac{1}{\sum_{i=1}^{K} \frac{1}{S - \bar{r}_i}} \tag{15}$$

4.1 未達成状況下での希求水準と RS-CH

未達成状況の RS 平衡の状態では希求水準も求めることができる．この状態では現時点で最適 (greedy) な行動 a_g とそれ以外の行動 a_j の RS の評価値は等しくなるため以下の式変形が行える．

$$I_g^{\text{RS}} = I_j^{\text{RS}}$$

$$\frac{n_g}{N}(\bar{r}_g - S) = \frac{n_j}{N}(\bar{r}_j - S)$$

$$S = \bar{r}_g \frac{1 - \frac{\bar{r}_j}{\bar{r}_g}\frac{n_j}{n_g}}{1 - \frac{n_j}{n_g}} \tag{16}$$

上記の式で論じられるのは，greedy な行動と任意の行動との関連性のみである．しかし greedy な行動と任意の行動の理想的な潜在選択比率 $\tau_j^* = n_j^*/n_g^*$ の情報が得られれば希求水準の逆算が可能になる．理想的

な潜在選択比率はそれぞれの腕の理想的な試行割合から式 (17) のように求められる．

$$\tau_j^* = \frac{n_j^*}{n_g^*} = \frac{n_j^*}{N}\frac{N}{n_g^*} = \frac{\rho_j^*}{\rho_g^*} \tag{17}$$

理想的な試行割合 ρ_j^* を “現時点で最適な行動の真の報酬期待値 $E[r_g]$ よりも任意の行動 j の真の報酬期待値 $E[r_j]$ の方が大きい（$E[r_g] \leqq E[r_j]$）可能性” と等しいとする．先行研究では ρ_j^* の推定に式 (18) で示される Chernoff-Hoeffding bound を用いて式 (19) と定義した．$p_j = E[r_j]$，$\epsilon > 0$ である．

$$P(\bar{r}_j \geqq p_j + \epsilon) \leqq \exp\left(-n_j D_{KL}(p_j + \epsilon || p_j)\right) \tag{18}$$

$$\rho_j^* := \exp\left(-n_j D_{KL}(\bar{r}_j || \bar{r}_g)\right) \tag{19}$$

任意の行動が greedy な行動 a_g の場合，確率 $P(E[r_g] \geqq E[r_g])$ は 1 となる．

$$\rho_g^* = \frac{n_g^*}{N} = \exp\left(-n_g D_{KL}(\bar{r}_g || \bar{r}_g)\right) = 1 \tag{20}$$

$\sum_{k=1}^{K} \rho_k^*$ は 1 を上回るが，この ρ_i^* は将来的に目指していきたい試行割合であるため問題はない．以上より理想的な潜在選択比率と希求水準は以下のように算出でき，これを使用した手法が RS-CH である．

$$\begin{aligned}\tau_j^{\mathrm{CH}} &= \frac{n_j^*}{N}\frac{N}{n_g^*} \\ &= \exp\left(-n_j D_{KL}(\bar{r}_j || \bar{r}_g)\right)\end{aligned} \tag{21}$$

$$S^{\mathrm{CH}} = \max\left(\bar{r}_g \frac{1 - \frac{\bar{r}_j}{\bar{r}_g}\tau_j^{\mathrm{CH}}}{1 - \tau_j^{\mathrm{CH}}}\right) \tag{22}$$

5. より一般的な報酬分布下での目標値推定

RS-CH での Chernoff-Hoeffding bound は確率変数がベルヌーイ分布に従う場合を想定しており，正規分布のように確率変数の範囲に制限のない場合には使用できず，本研究ではその代替案を検討する．

5.1 RS-t-test と RS-z-test

前章の議論では Chernoff-Hoeffding bound で理想的な選択割合の推定値である潜在選択比率を求めた．理想的な選択割合は “現時点で最適な行動の真の報酬期待値 $E[r_g]$ よりも任意の行動 j の真の報酬期待値 $E[r_j]$ の方が大きい（$E[r_g] \leqq E[r_j]$）可能性” と等しいとしていた．似たような判断を行う手法としては T 検定や Z 検定が挙げられる．そこで今回は片側検定での帰無仮説を $E[r_g] \leqq E[r_j]$，対立仮説を $E[r_g] > E[r_j]$ とみなしたときの p 値を用いる．

ここでの T 検定は 2 つの母集団の分散が異なる場合で有用なウェルチの T 検定を使用する．本来 Z 検定は母集団の分散が既知である場合を想定しており，標本分散を用いるのは不適当であるが大まかな推定値を求める用途で使用可能か合わせて検証する．Z 検定でもある程度の性能が確認できるのであれば，T 検定よりも自由度などの計算を省けるため利用の際の選択候補となりうる．任意の行動 j と greedy な行動との間の検定の p 値 p_{jg} と greedy な行動同士との検定の p 値 p_{gg} を用いると，理想的な潜在選択比率は式 (23) のようになる．

$$\begin{aligned}\tau_j^{\mathrm{test}} &= \frac{n_j^*}{N}\frac{N}{n_g^*} = \frac{p_{jg}}{p_{gg}} = \frac{p_{jg}}{0.5} \\ &= 2p_{jg}\end{aligned} \tag{23}$$

検定では帰無仮説を棄却する際に p 値と有意水準の比較を行うが，有意水準のような働きをするのはエージェントが実際に選択した a_g に対する a_j の試行割合 n_j/n_g である．もし p 値が n_j/n_g の値を上回ったとき，a_j の選択が不足していると判断すると考えられる．よってその有意水準は動的な値となる．

本研究ではこのように T 検定や Z 検定を用いることで報酬が正規分布に従う場合でも目標値推定を行う手法を RS-t-test，RS-z-test として提案する．検定に不偏分散を用いるため RS-t-test では始めに全ての腕を 2 回選択する．RS-z-test では始めにサンプルサイズを確保するため，全ての腕を 30 回選択する．

5.2 多腕バンディット問題への対応

RS-CH や RS-t-test，RS-z-test では任意の行動 a_j と greedy な行動の 2 つの間で理想的な潜在選択比率を使用しているため，そのままでは多腕バンディット問題には用いることはできない．

そのため多腕バンディット問題での RS-CH では実装上， 任意の行動 a_j の全てに対し greedy な行動 a_g とのペアでの比較を行う (甲野・高橋，2018)．具体的には行動 a_k と a_g の比較で得た希求水準 S_k を用いた RS の評価値の計算と比較を各行動で行っている．それらを行った結果，a_k の RS の評価値が a_g の RS の評価値を上回ったものが 1 つもなければ greedy な行動を，1 つあればその行動を選択する．もし a_g の RS の評価値を上回った行動が複数あった場合は，その中で理想的な潜在選択比率が高いものを選択し，潜在選択比率も同じであった場合には希求水準が高いものを選択する．希求水準も同じであった場合はそれらの行動の中からランダムに選択する．そのような手順を設定した RS-CH では報酬がベルヌーイ分布に従うバンディット問題で TS に次ぐ性能が確認されている．上記の手法を RS-t-test，RS-z-test でも行うことにする．

6. 実験と結果

報酬が正規分布に従うバンディット問題では，各行動の報酬期待値に任意の実数値を使用できるうえ，各

行動ごとに別の分散も指定できる．分散が大きいほど平均値の推定が難しくなるため，バンディット問題では報酬期待値が高い行動ほど分散も大きい環境の場合に最適な行動以外を誤って行動し続けてしまう状態が発生しやすくなると考えられる．また，行動の多さによる難易度への影響も計測する必要がある．そこで本研究では 2 種類の環境設定で実験を行う．実験 1 は 4 本腕 $(\mu, \sigma^2) = \{(1.0, 3.0^2),\ (0.8, 2.0^2),\ (0.5, 1.5^2),\ (0.0, 0.3^2)\}$，実験 2 は 10 本腕 $(\mu, \sigma^2) = \{(2.0, 3.0^2),\ (1.8, 2.7^2),\ (1.6, 2.4^2),\ (1.4, 2.1^2),\ (1.2, 1.8^2),\ (1.0, 1.5^2),\ (0.8, 1.2^2),\ (0.6, 0.9^2),\ (0.4, 0.6^2),\ (0.2, 0.3^2)\}$ である．今回の実験では 1 回の行動選択を 1 ステップ，100 万ステップを 1 シミュレーションとして 1 万回シミュレーション行い，その Regret の平均を評価とした．

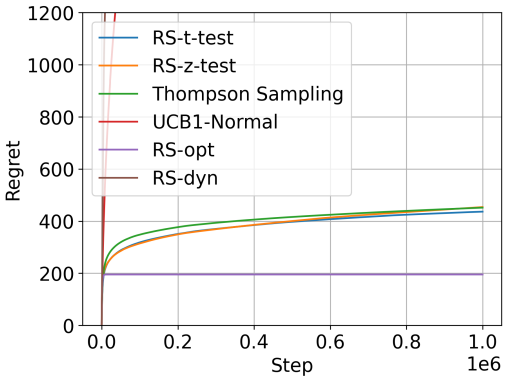

図 1　実験 1 における各手法の Regret 推移（4 本腕）

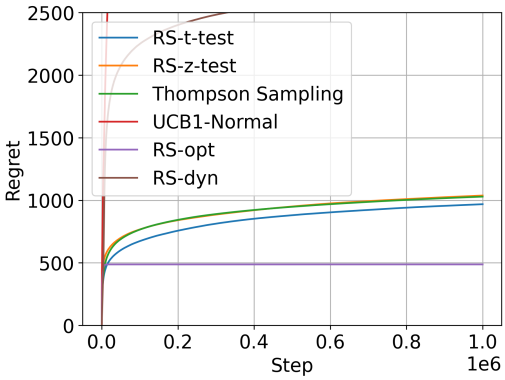

図 2　実験 2 における各手法の Regret 推移（10 本腕）

図 1 と図 2 より RS-t-test と RS-z-test は 2 つのいずれの環境でも TS と概ね同じ水準の性能が確認できた．しかし 4 本腕と 10 本腕を比較した際に TS と RS-z-test の Regret の差が縮まっているため，腕の本数が多い場合では TS がより有利になると考えられる．

7. 考察

既存の RS-CH と提案手法である RS-t-test と RS-z-test の違いは潜在選択比率の推定方法を Chernoff-Hoeffding bound によるものから検定における p 値によるものに変更したことである．報酬分布が特定の分布に従っていることがわかれば，その特定の分布にあった検定を実施することで潜在選択比率の推定が行える可能性が十分に考えられる．

7.1 推定された希求水準は最適なのか

図 1 と図 2 のグラフを見る限り RS-t-test と RS-z-test では TS 並みに最適化を行えている．しかしこれらの方策の希求水準は RS-opt の最適な希求水準とは一致しない．式 (11) と式 (16) を比較すると，RS-t-test と RS-z-test での動的な希求水準はより厳しい目標値であることがわかる．推定された希求水準は，検定の p 値から予測された潜在選択比率に近づける作用を持つと考えられる．本研究では実装の容易さから検定の p 値を用いたが，検定は計算コストがかかる作業であるため，検定の頻度を下げつつ最適な希求水準を推定する手法の検討を行いたい．また，検定の p 値を用いる手法が適切かどうかの検証を行う必要がある．

8. 結論

人間の認知傾向に由来する価値関数 RS では目標となる基準値を上回る行動を見つける能力に長けている．その一方で人間もある程度行うであろう適切な基準の形成においては報酬がベルヌーイ分布に従う場合でしか試されていなかった．本研究では任意の実数値報酬のタスクにおいて，潜在的探索比率の推定に検定における p 値を用いることが可能であることを示した．これにより報酬が正規分布に従うバンディット問題でオンライン情報のみを用いて Regret の最適化が行えるようになった．

文　献

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Junya Honda and Akimichi Takemura. Optimality of thompson sampling for gaussian bandits depends on priors, 2013.

Chloé Rouyer, Ronald Ortner, and Peter Auer. Understanding the gaps in satisficing bandits. In *Seventeenth European Workshop on Reinforcement Learning*, 2024.

Herbert A Simon. Rational choice and the structure of the environment. *Psychological review*, 63(2):129, 1956.

Akihiro Tamatsukuri and Tatsuji Takahashi. Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function. *Biosystems*, 180:46–53, 2019.

小河将真 ・ 有村柊一 ・ 高橋達二 ・ 甲野佑. Gaussian 分布の報酬への自然強化学習の拡張. In **人工知能学会全国大会論文集 第** *37* **回** *(2023)*, pages 2Q4OS27b04–2Q4OS27b04. 一般社団法人 人工知能学会, 2023.

甲野佑・高橋達二. 満足化を通じた最適な自律的探索. In **人工知能学会全国大会論文集 第** *32* **回** *(2018)*, pages 1Z304–1Z304. 一般社団法人 人工知能学会, 2018.

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, 2016.