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

満足化を通じた最適な自律的探索

Autonomous optimal exploration through satisficing

甲野 佑 *1　　高橋達二 *2
Yu Kono　　Tatsuji Takahashi

*1株式会社ディー・エヌ・エー　　*2東京電機大学
DeNA Co., Ltd.　　Tokyo Denki University

As deep layered neural networks enables reinforcement learning in huge action-state spaces, the exploration–exploitation tradeoff becomes more serious. Several heuristics have been proposed to deal with the tradeoff, utilizing noises. The probabilistic methods have difficulty in parameter tuning, and they amplify the problem of huge dispersion in performance of deep reinforcement learning algorithms. We propose a deterministic action selection algorithm based on a cognitive satisficing value function (RS) inspired by how humans explore under uncertainty. We define a method to enable optimal (minimal) exploration, utilizing the relationship between the aspiration level and the potential exploration distribution. The resulting algorithm exhibits an optimal performance in multi-armed bandit problems, and it opens the possibility for a new class of reinforcement learning algorithms.

1. はじめに

人間は一定の基準値 R を超える行動を探索し，発見後は探索を停止する，満足化と呼ばれる意思決定・試行錯誤の方策を実施するとされる．効率的な満足化に基づく試行錯誤学習を行うエージェントやロボットが開発できれば，R を利用者やエージェント自身が与えられる場合には，工学的にも有用である．また，満足化アルゴリズムを人間や動物の行動の分析に用いることで，行動科学的な知見を得ることもできるであろう．しかしながら，多腕バンディット問題のような単純な問題を除いては，満足化を効率的に行うアルゴリズムは知られていなかった．これに関し，その傾向を多腕バンディット問題において実現する認知的な満足化価値関数 RS が提案されている [高橋 16]．RS は R 以上の価値を持つ選択肢を速やかに見つけ出し, また R が最適，すなわち最適な行動のみが満足な行動であるように設定すれば，満足化の結果として最適化を非常に効率的に行い, 有限 regret を実現することが証明されている [玉造 18]. R の設定に関しては，まず R が入るべき範囲がわかる場合は色々ありうる．一日に必要な食物の量などのエージェントの内的な需要を R とすれば良い場合や，タスク・環境の情報から最適な R が計算できる場合がある．また，同様のタスクを解いている他エージェントから有益な R をもらえる場合もある．例えば陸上競技で誰かが 100m 走で 10 秒を切ると，他の選手も相次いで後続する，といった場合には，当初の新記録が R となっていると考えられる．満足化については，そのスケーラビリティ，すなわち行動数が爆発しても有効に働くことも分かっている [Oyo 17]．これは現実的な強化学習タスクを扱うには重要な性質である．しかし現在, 静的なバンディット問題を扱う際に RS は二つの課題を有している．

1. タスク中のオンライン情報のみからの最適基準値の算出法
2. 満足できる選択肢の発見までの探索配分の最適性

前者は事前情報がない状態での動的な基準値の算出の最適化, 後者は満足化における後悔の度合いの最適化についてを意味する．本研究では Chernoff–Hoeffding bound [Lai 85] からこの二点に関する RS の改善案を提示する．

連絡先: 甲野　佑, 株式会社ディー・エヌ・エー, 150–8510, 東京都渋谷区渋谷 2–11–1 渋谷ヒカリエ, yu.kohno.02@gmail.com

2. 多腕バンディット問題

多腕バンディット問題とは確率的に報酬が得られる複数の行動の選択肢 $a_i \in A$ が存在する環境で, 事前知識のない状態から単位時間毎に一種の行動を選んで発生する報酬値を観測する試行錯誤を続け, 報酬を最大化することを目的とした最もシンプルな意思決定課題である．また割引率を無視すれば，K 種の選択肢を有する初期状態を持ち, かつ 1 回の行動毎に終端状態へとたどり着く強化学習のエピソードタスクとしても捉える事もできる．本研究では報酬 r は必ず任意の選択肢 a_i を試行すると確率 p_i で $r = 1$, 確率 $1 - p_i$ で $r = 0$ が与えられる二値バンディットを扱う．

2.1 経験期待値の更新方法

バンディット問題における RS の算出に必要な数値は試行量 $n(a_i)$ と経験期待値 $V(a_i)$ であり, 以下のように更新される．ここで α は学習率, K は選択肢の数で, a^{select} はそのステップで選択された選択肢である．学習率 α は試行量 $n(a_i)$ に応じて減少していく．また総試行量を N と表記する．

$$\alpha \leftarrow \frac{1}{1 + n(a^{\text{select}})} \tag{1}$$

$$V(a^{\text{select}}) \leftarrow (1-\alpha)V(a^{\text{select}}) + \alpha r \tag{2}$$

$$n(a^{\text{select}}) \leftarrow n(a^{\text{select}}) + 1 \tag{3}$$

$$N = \sum_{k=1}^{K} n(a_k) \tag{4}$$

2.2 経験期待値の初期値

その初期値は以下のようにしている．試行量 $n(a_i)$ にはごく微小な ϵ を, RS 式でゼロ除算を防ぐために初期値として代入する．変数 $n(a_i)$ を試行量と呼ぶのはこの初期値のため試行回数そのものではないためである．

$$n(a_i) \leftarrow \epsilon \tag{5}$$

$$V(a_i) \leftarrow 0.5 \tag{6}$$

3. RS の定義

満足化価値関数 RS は式 (7) で定義される．ここで変数 R は満足化における基準値で, 既存研究では事前知識として与えられたり, なんらかの経験的な手法で動的に算出されていた．

$$RS(a_i) = \frac{n(a_i)}{N}\big(V(a_i) - R\big) \tag{7}$$

$$a^{\text{select}} \leftarrow \arg\max_{a_k}\big(RS(a_k)\big) \tag{8}$$

本研究ではその時点で RS の評価値が最も高い (RS–greedy) 選択肢を試行するエージェントを便宜上 RS エージェントと呼ぶ.

4. 非満足状況

エージェントにとって, 多腕バンディット問題における K 個の選択肢すべての経験期待値 $V(a_i)$ が $V(a_i) < R$ であると観測されている環境を “非満足状況” であると定義する. 非満足状況では現時点で経験期待値 $V(a_i)$ が最も高い (すなわち greedy な) 選択肢 a_G より基準値 R は高くなる ($R \geq V(a_G)$) ため, RS の定義では試行量とその総和の比率 (試行量割合) $\rho(a_i) = n(a_i)/N$ が高ければ高いほど RS 価値関数が算出する評価値は低くなっていく. ある選択肢 a_i が選択されるほどそれ以外の選択肢 a_j に対する試行量割合 $\rho(a_j) = n(a_j)/N$ は減少するため, 非満足状況ではある選択肢を試行するほど, それ以外の選択肢の RS 評価値は高くなる. その際の RS 評価値の上昇率は試行割合の変化量に応じて低くなるため, 総試行量 N が大きくなればなるほど, 非満足状況における RS 評価値の変化は小さくなる. しかし任意の選択肢の試行に対して, 不満足状況では RS は減少しかしないため, いずれ全ての選択肢に対してほぼ一定 (均衡値 $-Z$) になる. これを “非満足下の RS 均衡” と呼び, 前述の均衡値 $-Z$ を RS 均衡値と呼ぶ. それに伴い総試行量 N が充分大きい際の RS エージェントの試行割合 $n(a_i)/N$ を逆算的に算出できる. この均衡はある一定の割合に試行割合の配分を矯正する働きを有しており, 非満足状況下では常にその時点で RS の評価値が最も高い (RS–greedy) 選択肢を選び続けると, なるべく試行量 $n(a_i)$ が式 (10) に近づくように選択が行われていく.

$$RS(a_i) = -Z \tag{9}$$

$$\rho(a_i) = \frac{n(a_i)}{N} = \frac{Z}{R - V(a_i)} \tag{10}$$

非満足状況における RS 均衡値 ($-Z$) は, 試行量割合 $\rho(a_i)$ の総和が 1 になることから算出できる.

$$\sum_{k=1}^{K}\rho(a_i) = \sum_{k=1}^{K}\frac{n(a_i)}{N} \tag{11}$$

$$= \sum_{k=1}^{K}\frac{Z}{R - V(a_i)} \tag{12}$$

$$= 1 \tag{13}$$

$$Z = \frac{1}{\sum_{k=1}^{K}\frac{1}{R - V(a_k)}} \tag{14}$$

また RS 均衡値 $-Z$ が式 (14) から得られるため, 試行割合 $n(a_i)/N$ は基準値 R と経験期待値 $V(a_i)$ から定義できる. これによって RS エージェントの挙動が予想でき, 目的とする試行量割合 $\rho(a_i)$ から基準値 R を算出することができる.

4.1 非満足状況下で regret を最適化する基準値

前述の議論から非満足状況の時に RS–greedy に試行し続けると, greedy な選択肢 a_G の RS 評価値とそれ以外の選択肢 a_j の RS による評価値は “非満足下の RS 均衡” によっていずれ等しくなる.

$$RS(a_G) = RS(a_j) \tag{15}$$

$$n(a_G)\big(V(a_G) - R\big) = n(a_j)\big(V(a_j) - R\big) \tag{16}$$

$$n(a_j) = n(a_G)\frac{\big(R - V(a_G)\big)}{\big(R - V(a_j)\big)} \tag{17}$$

$$R = V(a_G)\frac{1 - \frac{V(a_j)}{V(a_G)}\frac{n(a_j)}{n(a_G)}}{1 - \frac{n(a_j)}{n(a_G)}} \tag{18}$$

ここで $V(a_j)/V(a_G)$ はその時点での greedy な選択肢の定義と報酬が非負である定義上, 必ず 1 以下になる.

$$\frac{V(a_j)}{V(a_G)} \leq 1 \tag{19}$$

式 (17) から, 現時点で我々が RS において対象とできるのは, 現時点で greedy な選択肢 a_G 任意の選択肢 a_j の関係のみである. しかしゆえに非満足化状況では理想的な $n^*(a_j)/n^*(a_G)$ の関係がわかれば, それを逆算的に実現する基準値 R を定義できるということがわかる. そこで本項では greedy な選択肢とそれ以外の任意の選択肢 a_j の潜在的な選択比率 $\mu = n(a_j)/n(a_G)$ を最適化することを目的とし, その最適な比率 μ^* を Chernoff-Hoeffding bound [Lai 85] から推定する. 各選択肢の現時点からの試行回数の目的値 $n^*(a_i)$ で表すとき, 任意の選択肢 a_j に対する最適な潜在的選択比率 μ_j^* は $n^*(a_j)/n^*(a_G)$ で表される. ここで ρ_k^* は最適な試行割合を意味する.

$$\mu_j^* = \frac{n^*(a_j)}{n^*(a_G)} = \frac{n^*(a_j)}{N}\frac{N}{n^*(a_G)} = \frac{\rho_j^*}{\rho_G^*} \tag{20}$$

あくまでも μ は比率であるため, $n(a_i)$ の絶対的大きさに依存しない. ここで現時点での総試行回数で $n^*(a_i)$ を割ると, 現時点で要求されるべき試行割合 ρ_i^* になる. 本研究では要求されるべき試行割合 ρ_i^* を現時点で greedy な選択肢の真の期待値 $V^*(a_G)$ が任意の選択肢の真の期待値 $V^*(a_i) = p_i$ 以上である (つまり実は a_G が 最良の選択肢ではなく $V^*(a_G) \leq V^*(a_i)$ である) 可能性と等しいと定義し, Chernoff–Hoeffding bound を用いて以下の式 (21) から式 (22) を定義した.

$$\Pr\big(V(a_j) \geq p_i + \epsilon\big) \leq \exp\big(-n(a_j)D_{KL}(p_i + \epsilon || p_i)\big) \tag{21}$$

$$\begin{aligned}\rho_i^* &= \frac{n^*(a_i)}{N} \\ &= \Pr\big(V^*(a_i) \geq V^*(a_G)\big) \\ &= \exp\big(-n(a_i)D_{KL}(V(a_i)||V(a_G))\big)\end{aligned} \tag{22}$$

ここで任意の選択肢 a_i が greedy な選択肢の a_G だった場合, $V^*(a_i) = V^*(a_G)$ であるので確率 $\Pr\big(V^*(a_G) \geq V^*(a_G)\big)$ は 1 になる.

$$\begin{aligned}\rho_G^* &= \frac{n^*(a_G)}{N} \\ &= \Pr\big(V^*(a_G) \geq V^*(a_G)\big) \\ &= \exp\big(-n(a_G)D_{KL}(V(a_G)||V(a_G))\big) \\ &= 1\end{aligned} \tag{23}$$

ここで ρ_k^* はあくまでも今後目指すべき試行割合であり, 総和が $\sum\rho_k^* = 1$ を満たさなければならないわけではない. よって任意の選択肢に対して Chernoff–Hoeffding bound から導かれる近似的に最適な潜在選択比率の推定値 $\mu_j^{\text{CH}} \simeq \mu_j^*$ は以下のように定義される.

$$\mu_j^{\mathrm{CH}} = \frac{n^*(a_j)}{N}\frac{N}{n^*(a_G)} = \exp\big(-n(a_j)D_{KL}(V(a_j)||V(a_G))\big) \quad (24)$$

$$R^{\mathrm{CH}} = \max\left(V(a_G)\frac{1-\frac{V(a_j)}{V(a_G)}\mu_j^{\mathrm{CH}}}{1-\mu_j^{\mathrm{CH}}}\right) \quad (25)$$

また近似的な最適潜在選択比率 μ_j^{CH} を用いて算出された基準値 R を非満足基準値 R^{CH} とし, それを用いた RS 価値関数を RS–CH と呼ぶ.

5. RS–CH の多腕対応

前述した段階での RS–CH はある選択肢 a_j と greedy 選択肢との間の Chernoff–Hoeffding Theorem から導かれる理想的な比率 μ_j^{CH} に基づくのみで, 選択肢が二つしかない 2 本腕バンディット問題のみにしか使用できない. それ以上に後述する理論的な問題があるために現行の RS の形式では 3 本腕以上のバンディット問題では最適にはならないと思われる.

5.1 3 本腕バンディット問題以上での非最適性

総試行回数 N が十分大きい場合の a_i に対する理想的な試行割合は式 (26) になるため, ゆえに greedy ではない任意の二つの選択肢 a_i と a_j の選択比率は式 (27) になる.

$$\rho^*(a_i) = \exp\big(-n(a_i)D_{KL}(V(a_i)||V(a_G))\big) \quad (26)$$

$$\frac{\rho^*(a_i)}{\rho^*(a_j)} = \frac{\exp\big(-n(a_i)D_{KL}(V(a_i)||V(a_G))\big)}{\exp\big(-n(a_j)D_{KL}(V(a_j)||V(a_G))\big)} \quad (27)$$

RS 価値関数に対して greedy な選択をし続けた最終的な試行割合 ρ^{RS} (RS 目的試行割合) は式 (28) であるため, greedy ではない任意の二つの選択肢 a_i と a_j の間の選択比率は式 (29) のようになる.

$$\rho^{\mathrm{RS}}(a_i) = \frac{Z}{R-V(a_i)} \quad (28)$$

$$\frac{\rho_i^{\mathrm{RS}}}{\rho_j^{\mathrm{RS}}} = \frac{R-V(a_j)}{R-V(a_i)} \quad (29)$$

全ての選択肢に対する RS 目的試行割合の間の選択比率 (式 (29)) が $N\to\infty$ の時に式 (27) を満たす基準値 R が存在しない限り, RS はいかなる基準値 R の定義を持とうと 3 本腕以上のバンディット問題では最適ではないと言える. 従来の RS エージェントでは基準値 R は全ての選択肢に対する RS 評価値に対して共通であるため, 式 (27) と一致はしない.

ただしこれは非満足化状況 ($R \geq V(a_G)$) での満足化に限定した場合であり, 満足化状況 ($R \leq V(a_G)$) を取りうる基準値 R の定義である場合には適用されない.

5.2 3 本腕以上のバンディット問題への対応

従来の RS 価値関数のように全ての選択肢に対して等しい基準値 R を用いる場合, 3 本腕以上のバンディット問題での後悔の最小化は不可能である. そのため本研究では実装上, 現時点で経験期待値が最も高い greedy な選択肢 a_G とそれ以外の選択肢 a_j を一つずつ比較する手法を提案する. 具体的には $V(a_G)$, $n(a_G)$, $V(a_k)$, $n(a_k)$ の四つの変数を用いて 2 本腕バンディットと同じく式 (25) 及び RS 価値関数を計算する. その上で全ての非 greedy 選択肢 a_j と比較して greedy 選択肢 a_G の RS 評価値が上回る場合は a_G を選択する. 非 greedy 選択肢 a_k の RS 価値関数が a_G を上回る選択肢があった場合, その選択肢を選ぶ. そして greedy 選択肢 a_G の RS 値を上回る非 greedy 選択肢が複数あった場合, その中で a_G との潜在的選択比率 μ_j^{CH} が最も高い選択肢, もし最も高い潜在的選択肢 $a^{\max} = \arg\max \mu_j^{\mathrm{CH}}$ が複数あった場合は R^{CH} が最も高い選択肢を選択する (それも同値だった場合はその選択肢の中でランダムに選択する).

6. 実験と結果

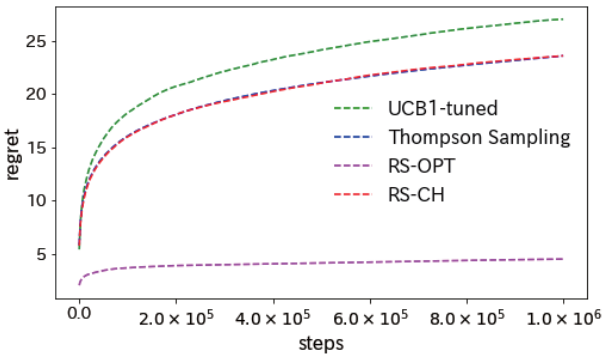

図 1: 2 本腕の結果

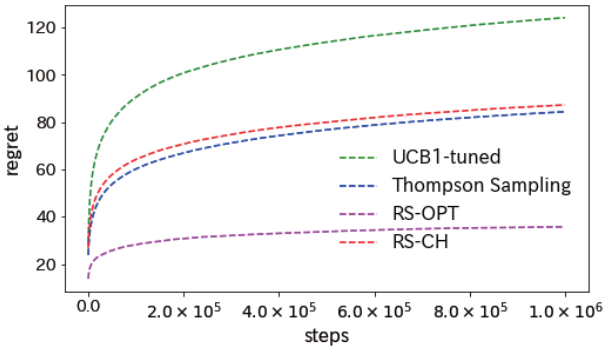

図 2: 8 本腕の結果

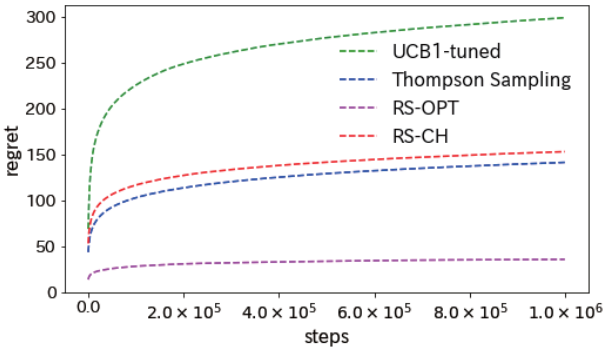

図 3: 20 本腕の結果

実験は各々シードの異なる 10,000 回のシミュレーションを平均で評価した. 選択肢数は 20 個で, それぞれ選択肢 a_k の真の報酬出現確率 p_k はシミュレーション毎に一様分布からサンプリングされ, 1 回のシミュレーション中は固定される. 1 回の選択を 1 step として 1,000,000 step 選択を行なった. 比較する選択アルゴリズムには UCB1–tuned [Wang 05] と Thompson Sampling [Agrawal 12] を用いた. また理想的な基準値 R^{OPT} [高橋 16, Oyo 17, 玉造 18] が既知の場合にそれを用いた RS–OPT も併記する. 理想的な基準値 R^{OPT} は選択肢 a_k を真の報酬出現確率 p_k に対して同値を排除し, 高い順に降順に並べた選択肢 q_m に対して, 以下の式で定義される.

$$R^{\mathrm{OPT}} = \frac{q_1 + q_2}{2} \tag{30}$$

評価は各プロット点の step 数を T と, ある step t において選ばれた選択肢 a_t^{select} の真の確率 p_t^{select} から定義される regret (後悔の度合い) によって比較する.

$$\text{regret} : \sum_{t=0}^{T} \left(q_1 - p_t^{\mathrm{select}}\right) \tag{31}$$

選択指数はそれぞれ $K = 2, 8, 20$ の場合をそれぞれ図 1, 2, 3 にしめす. 本研究で提案した基準値を用いた価値関数 RS–CH により 2 本腕では Thompson Sampling と同水準の成績を得られた. しかし選択肢の数が増えるほどに徐々に regret が劣っていく結果になった. 同時に RS–CH が UCB1–tuned に対して優っている度合いも増えていった.

7. 今後の展望

RS–CH と従来の RS の違いは基準値 R の定義のみであるため, いままで考案されてきた従来の基準値 R と定義式を組み合わせるなどして, 非定常環境などにも用いることが出来る可能性がある. それ以外にも R^{CH} の更新に使う変数をステップ数でずらすだけで非定常環境に強くなる可能性はあると考えらえる. RS–CH は 2 値の報酬でのバンディット問題における KL 情報量での定義なため, 報酬の分布族がわかればベルヌーイバンディット問題以外にも用いることが可能だと思われる.

7.1 非満足状況化での最適基準値と理想基準値との関係

RS–CH は 2 本腕バンディット問題における非満足化状態 ($R \geq V(a_G)$) における RS 評価値の均衡を利用して, Chernoff–Hoeffding bound から導かれる regret の下限以上になりつつ $O(\log N)$ になるよう調整したアルゴリズムである. それにより損失の上限を抑えるという基本的な考え方は Upper Confidence bound (UCB) 系アルゴリズムの導出や Thompson Sampling の最低性の証明にも用いられている. これらの既存最適化アルゴリズムは “greedy な選択肢 a_G の期待値よりその他の選択肢 a_i の期待値が低いと確信するための探索回数を稼ぐ” ことを目的としていると言える.

それに対して満足化方策は RS–greedy な選択肢 a_G^{RS} の期待値が理想的な基準値 R^{OPT} より高い事を確信する事を目的としている. 理想基準値 R^{OPT} であれば, それを上回る真の期待値を持つ選択肢が一つしかないので, greedy に選択し続けても基準値 R^{OPT} より高い選択肢は最適な選択肢である. さらに RS 満足化は前述したように非満足化状態の際に RS 評価値の均衡が起こり, 現環境で観測された経験期待値と基準値 R との差に応じて探索具合を調整する性質がある. 基準値 R が高ければ高いほど試行割合が均等な配分になる.

最適化では step 毎に 1 対 $(K-1)$ の比較検証 (greedy 選択肢とそれ以外の選択肢を一つずつ比べなければならない) が必要になり, さらにその比較は regret 下限の制約を持つため, 非常に検証コストが大きい. それに対して満足化は, 飽くまでも常に基準値と選択肢の比較であり, 一つでもそれ以上の選択肢を見つければ良いため, 1 対 $M \leq (K-1)$ の検証の比較検証で済む. すなわち満足化は最適化と検証対象の違うために非常に検証コストが低い. これは満足化方策は基準値 R の値によっては選択収束 ($N \to \infty$ の時に $n(a_G) \simeq N$ になること) ができず, 最適化方策はいずれ必ず選択収束ができると言い換えることも出来るため一長一短である.

7.2 満足化の柔軟性としての展望

非満足化状態から満足化を目指す満足化過程を前提とする場合, 満足化は最適化方策と同じ問題を抱えてしまう. さらに前述の問題から 3 本腕以上のバンディット問題では後悔の度合いに対するタイトな最適化にはなり得ない. ゆえに理想基準値 R^{OPT} の動的な推定コストが RS における本研究の提案本手法 RS–CH のコストより低いものであれば本手法はその存在意義を失う. ゆえに今後は満足状況をも扱える満足化としての, 理想基準値の推定を検証すべきだと考えられる.

しかしバンディット問題に対して, 事前情報なしで基準値 R を最適化できることを示したのは本研究が初となる. これは満足化の枠組みの中で選択的に最適化も可能であることを示している. コストや, 与えられた目標値など, 事前に与えられた基準値とした R^{PRE} と比較したり二重の満足化を行うことで, 満足化を緩めて事前の目的を優先する事もできる.

8. 結論

満足化価値関数 RS は, 与えられた, あるいは動的に推定した基準値以上の選択肢を発見に優れた性質を有する. しかし, 満足が不可能な基準値を与えられた際の挙動や, その基準にどう対処すべきかは議論されてこなかった. 本研究の理論的な貢献はこの非満足化状況を定式化し, 非満足化時にそこから目指すべき探索配分, 潜在的探索比率 μ を導出したことにある. この潜在的探索比率 μ に Chernoff–Hoeffding bound を用いることで, Thompson sampling のような乱数や, KL–UCB のような数値計算を用いず, 2 値報酬のバンディット問題において regret をタイトに最適化できることを示した. この意味で, 決定論的な楽観的探索の形式の提案を行ったとも言える. 潜在的探索比率の導出により, RS の定量的な探索配分を明らかになったため, 状態空間を広げた一般的な強化学習課題に拡張した際の探索配分の分析, 改善に貢献できると考えられる.

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).

[Kohno 17] Kohno, Y., Takahashi, T.: A cognitive satisficing strategy for bandit problems, *International Journal of Parallel, Emergent and Distributed Systems*, 32, 2, 232–242. (2017)

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

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

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

[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). doi:10.1527/tjsai.AI30-M

[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)