2D1-OS-12a-4in

不確実性の下での満足化を通じた最適化

Optimization through Satisficing under Uncertainty

高橋 達二 *1 Takahashi, Tatsuji　甲野 佑 *2 Kohno, Yu　大用 庫智 *3 Oyo, Kuratomo　横須賀 聡 *4 Yokosuka, Satoshi

*1*4東京電機大学 理工学部
School of Science and Engineering, Tokyo Denki University

*2*3東京電機大学大学院 先端科学技術研究科
Graduate School of Advanced Science and Technology, Tokyo Denki University

We introduce a value function that implements the risk attitudes characteristic of human cognition. The function, RS (reference satisficing value function), enables an efficient satisficing in N-armed bandit problems when operated under the greedy policy, and when the reference level for satisficing is appropriate, it leads to effective optimization.

1. はじめに

強化学習は，環境自体のメカニズムや観測の不完全性に由来する不確実性を有する状況において，試行錯誤を通じて有効な行動系列を獲得する機械学習の分野である．その適用範囲がロボットや脳科学へと拡がることで，求められる技術は変化してきている．これまでの，いかなる状況においても極限で収束・最適化する手法の考案から，対応できる環境をある程度制限し，またエージェントの制限（計算能力の限界やアクチュエーターの損耗）を考慮に入れた上でより現実的な時間・試行回数でそれなりに有効な行動を獲得でき，かつ様々な状況に柔軟に対応できる，合理性なヒューリスティクスの必要性が高まってきた．

環境やエージェントの制約を込みにした合理性，あるいは非合理性ということで言えば，サイモンの提唱した限定合理性が重要である．限定合理性を持つ戦略としては，満足化 (satisficing) が代表的である．しかしながら，これまで強化学習の分野での満足化戦略の研究は盛んではなかった．理由としては，後述するように，従来の満足化戦略の形式化が素朴なものであったことが考えられる．また，満足化の概念は通常は完全な合理性の下での最適化に対する，限定された合理性の下での満足化としての，消極的な意味合いしか与えられてこなかった．

本研究では，ある前提（環境あるいは目的の限定）の下で満足化が即ち最適化を意味することを確認した上で，N 本腕バンディット問題の枠組みの中で，新しい満足化のモデルを提案する．従来の満足化が強化学習のポリシー（状況から行動を導く確率的関数）のレベルで形式化されてきたのに対し，我々のモデルは価値関数（状態行動対の価値付け）のレベルに存する．そのことで，定義と振る舞いが単純となり，性質の分析が容易になるだけでなく，パフォーマンスの点においても優れたものになることを示す．また，我々の満足化価値関数が，人間と同種の信頼性・リスク評価を価値に組み込んでいることも議論する．

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

2. N 本腕バンディット問題

不確実性の下，エージェントが環境との相互作用を通じて情報をその場で獲得し，またそれを活用して意志決定を繰り返していく枠組みは強化学習と呼ばれる．強化学習の中でも最も基本的な問題として，N 本腕バンディット問題がある．この問題においては，とりうる行動が N 種類ある $(a_1, a_2, ..., a_N)$．それぞれの行動は，それを行う結果として確率的に，報酬をもたらしたり（これを報酬 1 とする），もたらさなかったりする（報酬 0）．報酬確率は各行動 a_i について異なったものが与えられており，これを P_i とする．ただし，各確率はエージェントにとっては未知であり，行動の選択とその結果得られる報酬の対の情報を集めることで推定していくしかない．ゴールは獲得報酬の累積の最大化である．

そのための基本的な価値づけとして，行動 a_i についてその報酬の期待値を価値 E_i とし，

$$E_i = a_i^1/(a_i^1 + a_i^0) \tag{1}$$

と定義する．ここで a_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 $n_i = a_i^1 + a_i^0$ を満たし，さらに $n = \Sigma n_i$ とする．また，各行動 a_i について，それを期待値 E_i と報酬確率 P_i で降順にソートしたものをそれぞれ s_i, o_i と表記することとする．初回から最後までの o_1 の選択がベストの累積獲得報酬の期待をもたらす．

2.1 greedy 法

この問題では，それまでに得られた報酬の情報を考慮して，一度に一つの行動を選択し，それにより報酬というフィードバックを受け，さらに行動の選択を行っていく．行動選択のアルゴリズム（ポリシー）として最も基本的なものは greedy 法であり，これは常に最も価値の高い (greedy な) 行動 s_1 を実行するものである．

2.2 知識利用と探索のジレンマと，速さと正確さのトレードオフ

この状況で累積獲得報酬の最適化を目指す場合，単にこれまでの知識を利用して greedy に行動するだけでは局所解に陥ってしまう．すなわち，N 個の中からたまたま序盤に選択した行動 a_i の期待値がその他よりも高ければ，その行動を選択し続けてしまう．このような場合，行動 a_i が実際に最適な行動

であるということ，すなわち $a_i = s_1$ が $a_i = o_1$ も満たすということが重要である．この検証には，a_i 以外についての試行，探索が必要となる．探索は，ある決まった確率 ϵ におけるランダムな行動の選択でも良い．この場合確率 $1-\epsilon$ で知識利用を行う．このようなポリシーを ϵ-greedy 法と言う．また，ルーレット選択のように，その期待値の単調関数となる確率での行動のランダムな選択としても良い．この種のポリシーで代表的なものは Gibbs 分布を用いた softmax 行動選択法である．

このように，知識利用=greedy 行動，探索=非 greedy 行動，という通念に従うと，両カテゴリーの相互排他性により，知識利用と探索は定義上両立できないというジレンマが成立してしまう．これは，知識利用を優先すると，パフォーマンスの立ち上がりは速いが後での伸びが悪く，探索を優先すると最終的な伸びは良いが序盤のパフォーマンスが悪い，という，速さと正確さのトレードオフを導く．これは強化学習の根本的な問題の一つであり，遅延報酬の扱いの難しさにも深い関連がある．

2.3 不確実性には楽観性で (UCB)

強化学習における合理的な最適化理論としては，Cesa-Bianchi らによる「不確実ならば楽観的に optimism in face of uncertainty」というものがある [Bubeck 12]．たとえば行動 a_i, a_j について期待値が同じ $E_i = E_j$ であっても，試行回数が異なり，$n_i \ll n_j$ であれば，真の価値といえる P_i より P_j の方がはるかに現在の期待値に近いと考えるべきである．このように，これまでに獲得した報酬情報の信頼性を考慮し，「まだあまり試してないもの (不確実な選択肢) のポテンシャルは高い (楽観)」という評価を繰り込む．

この考えに基づくバンディット問題の標準的アルゴリズムは UCB (upper confidence bound) [Auer 02] と呼ばれ，モンテカルロ木探索に応用され囲碁 AI を飛躍的に強化したのもこの手法である．UCB は単なる価値関数でありながら，十分な選択回数が許されれば高い成績を示し，「後悔」(後述) の上界を保証している．ここでは UCB1 のパフォーマンスを改良した UCB1-Tuned (UCB1T) を導入する．

$$\mathrm{UCB1T}(a_i) = E_i + \sqrt{\frac{\ln n}{n_i} \min\{\frac{1}{4}, V_i(n_i)\}} \qquad (2)$$

ここで $V_i(s) = v_i + \sqrt{(2\ln n)/s}$ であり，v_i は腕 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界である．このアルゴリズムでは，UCB1T を各行動の価値とするが，E が知識利用を，E の信頼性の低さを表現し，試行につれて減衰していく第二項が探索を担っている．$n_i = 0$ のとき $\mathrm{UCB1T}(a_i)$ が発散すると考え，greedy 法では最初の N 回は各行動を一度ずつ選択する（それにより値が有限に落ちる）ことにして用いる．

3. 満足化のモデル

N 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．

3.1 素朴満足化ポリシー (PS)

満足化 satisficing の標準的な定義は，

> 基準 R 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．これを強化学習のポリシーとして定式化すると，一つでも行動の期待値が基準 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．つまり，R を超える期待値をもつ行動が存在すれば $\epsilon = 0$，そうでなければ $\epsilon = 1$，という探索確率 ϵ の設定をする ϵ-greedy 法である．これを素朴満足化ポリシー (PS: policy satisficing) と呼ぶ．

3.2 満足化と価値づけ

満足する satisfice というのは，ある基準よりも優れた行動を発見し，それを選択するということである．基準との比較は行動 a_i の期待値と基準 R との差を

$$\delta_i = E_i - R \qquad (3)$$

とした価値で行うことができるが，これは $E_i = \delta_i + R$ が成立するため，たんに定数 R 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 にしたがうのと同じ事になる．

3.3 価値関数による満足化

素朴満足化では δ_i が正となる行動 a_i が見つかり次第探索を止めてそれを選択し続ける．δ_i が正となる行動が複数あれど，s_1 を選ぶ．このようにすると，$s_1 = o_1$ でない場合には，s_1 の選択という局所解から抜け出るのが難しい．E の値，あるいは s の順序の情報の信頼性の考慮が必要である．

3.3.1 価値の信頼性・リスクの考慮

「信頼性の表現」の問題に関し，期待値や条件付き確率にサンプルサイズも付加情報としてつけるのが一つのやり方であり，ベイズの立場からは，初期の事前分布が，それまでに得た情報によってどのくらい尖ってくるか，として表現できる．ここで，探索と知識利用において，そのそれぞれの方針の趣旨に従ったかたちで行動 a_i について，期待値 E_i の値のみでなく，その試行回数・サンプルサイズ n_i を考慮することで，E_i の信頼性を単純・直接的に扱う．

3.3.2 楽観的探索

全ての行動の価値が基準を下回る際に行われる探索においては，とにかく基準を超えるものを探す満足化の観点から言えば，「もしかしたら基準を超える行動が存在するが，その行動について，これまで試してきてたまたま外れが多かったため期待値が基準を超えておらず，もっと試してみたら基準を超えていることが分かるかも知れない」と考えることが有効であろう．そうすると，サンプルサイズの小さい行動の価値を上げるような補正を行うことになる．これは UCB1 と同様である．

3.3.3 悲観的知識利用

基準を上回る価値を持つ行動が一つは存在する知識利用においては，期待値が基準を上回る行動を選ぶとして，そのような行動が複数ある場合にどの行動を選ぶかについて，やはり基準との関係において信頼性を考慮することが有効である．つまり，「ある行動の価値が基準を超えていることはどれだけ確かか」を評価し，その価値が最も基準を超えていそうな行動を選択する．すなわち，E_i が高いだけでなく，それが信頼できる，つまり n_i が大きいような行動 a_i を選択しやすいように価値付けを行うことで，たまたま乱数の出具合により一時的に基準を超えている行動 ($P_j < R$ となるような a_j) については，$\delta_j < 0$ となるようなふるい落としを行うこととともなる．

3.3.4 満足化価値関数 (RS)

以上をまとめると，基準を下回る行動については，サンプルサイズが小さいものを高く価値付けし，他方基準を上回る行動

については，サンプルサイズが大きいものを高く価値付けすることとなる．ここで δ_i を基本とすると，基準を下回る行動の価値は負，基準を上回る行動は正の値を持つので，それらの負や正の値にサンプルサイズを掛けてやれば，上の価値付けは場合分けなしに，単なる価値関数

$$\mathrm{RS}_i = n_i \delta_i \tag{4}$$

として実現されることになる．これは $R = 0.5$ の場合には篠原が RS (rigidly symmetric) と呼んだ価値関数に等しい [篠原 07]．これにちなんで式 (4) を基準満足化価値関数 (RS: reference satisficing) として提案する．この論文では今後，この価値関数の性質と機能について述べていく．その際，特に断りが無ければ，この価値関数を，それを greedy 法で運用したアルゴリズムとも区別しない．

3.4 満足化基準の値の決め方

PS や RS の運用においてまだ定まっていないのが基準 R の値である．満足化の従来の研究ではこの基準は aspiration level と呼ばれるものに相当する．より生態学的な例で考えれば，バンディット問題の枠組みで行動する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，ステップ（一回の試行）は一日という時間単位を意味するとしよう．その動物が，大体二日に一度くらいは食物を摂取する必要があるとすれば，基準は $R = 0.5$ 以上に，しかし高望みしすぎない程度に，設定すれば良いことになる．満足化が成功することを前提とすれば，そうすることによって，平均的には二日に一度ほど食物が見つかるような餌場を探すこととなる．

このような基準がどのように設定されるのかはいくつかの変数に依る．生存や再生産のために必要な食物の平均的な摂取量，あるいは一試行のコストは代表的な例であろう．可能な試行回数 n によっても基準は変化しうる．すなわち，n が小さければ，高いレベルを基準とするよりも，そこそこのレベルで満足するべきかもしれないし，n が充分に大きければ，基準を少しずつ上げていき，最適化を目指すべきかもしれない．あるいはこれまでの知識や経験から，目指すべき基準がじりじりと上下してきているのかもしれない．このような知識や経験は事前分布として表現することもできる．

3.4.1 生物が要求される最適化の多次元性

他にもバンディット問題内だけでなくその外にあるファクターも効いてくると考えられる．一般には，エージェントが解くタスクは唯一つではない．動物にとって基本的なタスクとしては採餌行動とメーティングがあるが，一方のために他方を犠牲にする（より低い基準でよしとする），ということはあるだろう．より現代的にも，他のことに思い悩むことなくゲームに熱中していれば，そのゲームの中ではプレイヤーは自らのゴールを達成すべく，ほとんどのリソースを投入できる．しかし，ゲームをしていても，生存のための最低限の必要はあり，食事や水分補給は行うわけだが，ゲームに熱中しているときの食事や飲み物の質は，熱中していないときに比べて，そこそこの質でよしとされると考えられる．

3.4.2 満足化が最適化となる場合：最適切基準

満足化は最適化と対比されるように導入された概念ではあるが，一般にそう思われているのと異なり，対立するわけではなく，ある条件が満たされれば満足化を行うことは最適化でもありうる．二値バンディット問題の範囲では，最適化は最も報酬確率の高い最適行動を選択する（し続ける）ことであるが，もしも満足化の基準 R が，最適行動 $a_i = o_1$ と次善の行動 $a_j = o_2$ の報酬確率の間に設定されているとすれば，R を満足する行動は最適行動の選択となる．これを「最適切基準」$R_{\mathrm{ap}} = (P_i + P_j)/2$ と呼ぶ．

4. シミュレーションと結果

N 本腕バンディット問題での満足化モデルの性能を検証する．結果は全て，1000 回のシミュレーションの平均であり，各シミュレーションでの試行回数 (step) は 1,000,000 である．全行動の報酬確率は毎回のシミュレーションで，$[0, 1]$ の一様乱数とする．パフォーマンスの指標としては，各 step において最適な行動を選んだ比率の「正確さ」(accuracy) と，各 step において，「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」，という期待損失の「後悔」(regret) を用いる．RS と PS については $R = R_{\mathrm{ap}}$ の最適切基準を与える．

$N = 2$ の場合の RS の結果は [Oyo 13] の LS という価値関数による結果とほぼ全く同様である．$N = 10, 100, 1000$ の結果を図 1 に示す．

まず UCB1T と PS の結果から見ていく．UCB1T は前述のメカニズムにより，$N = 10$ の場合の 10^1 step までの正確さは 10% となり，その後順調に上昇する (図 1 左上)．後悔は理論的に保証されている通り，小さく抑えられている．他方，素朴な満足化＝最適化を行う PS は初期からなめらかに正確さを上げるが，R を超える報酬確率を持つ行動（実際には最適行動一つしか存在しない）が見つからない限りランダムに行動を選択することから，後悔の成長が速い．$N = 10$ では微妙なところであるが，$N = 100, 1000$ の結果からは，UCB1T と PS の結果に，弱い速さと正確さのトレードオフが見てとれる．それに対して RS は，UCB1T と PS に比べると，全

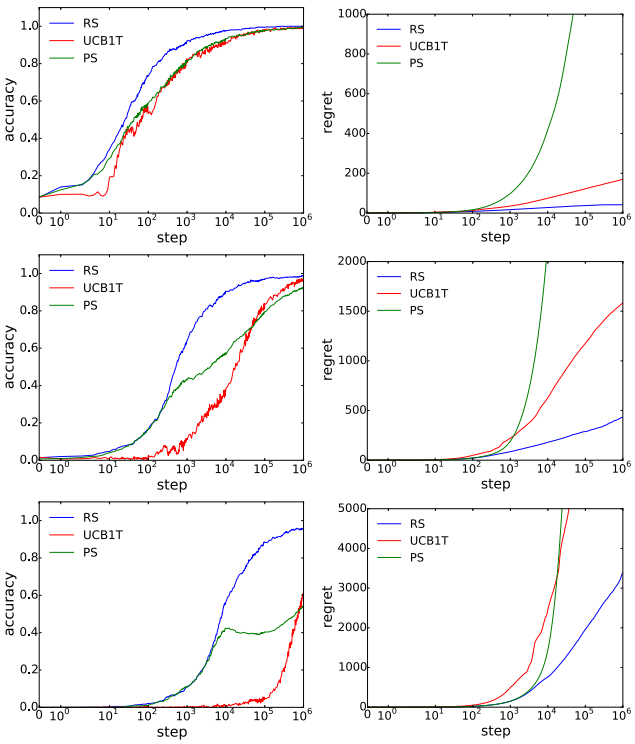

図 1: 指標の時間発展．左列が正確さ，右列が後悔．上行，中行，下行がそれぞれ $N = 10, 100, 1000$ である．

てのステップについて他の二つのアルゴリズムよりも正確さが高く後悔も小さい.，速さと正確さを両立していると言える．

5. 議論

腕が N=10, 100, 1000 と 10 倍になっていくにつれて，たとえば正解率 50%に達するのは RS と UCB1T でそれぞれ 10 の 1.5, 2.75, 4 乗と 10 の 1.8, 4.2, 5.8 乗 step 付近と，N に対する steps 数のオーダーが異なるようである．この点は今後の分析を要する．

5.1 今後の理論的課題

RS による満足化の達成に関する理論的な保証が必要である．最適行動 $a_i = o_1$ の期待値 E_i，次善の行動 $a_j = 0_2$ の期待値 E_j 、そして基準 R の三項の大小関係の 6 つの組み合わせを 6 状態とする確率オートマトンとしての分析が中心的になる．遷移確率は報酬情報の関数として求められ，どの状態からでも 状態 $E_j < R < E_i$ に遷移し，基準 R が期待値に対しても適切となること，またこの状態が安定であると示せば良い．

また， RS による満足化の効率性を定量的に扱えることが望ましい．UCB はそれが発表された論文 ([Auer 02]) のタイトルにもあるように，極限ではなく，有限の step におけるパフォーマンスを保証した点で画期的であった．RS の価値の比較は $n_i/n_j < \delta_j/\delta_i$ のような単純かつ直感的な比率で行うため，この点も容易であることが期待できる．

5.2 内的な満足化

満足化の基準は， 3.4 で議論したように aspiration level として，環境から推定すべきものではなくむしろエージェントに備わっているもの，とすれば，その満足化が効率的であるだけで十分である．この点はロボットや生物が無際限な実世界において自己維持を賭けた強化学習を行う場合に有効である．

5.3 動的な基準の設定

R をオンラインで更新していくことも可能である．これに関しては，報酬値や期待値，分散を用い，また方策 on/off で，様々な更新方法を考えることが出来る．また， UCB のように信頼性を考慮して基準をアニーリングしていくこともできる．現在それなりにうまいやり方はいくつか存在するが，理論的な保証は今後与える必要がある．

5.4 「最適切な基準」はどの程度のチートなのか

シミュレーションでは RS と PS が最適切基準を持つ，つまりその満足化行動は成功すれば最適化行動となるという設定をとった．その基準設定には，最適と次善の行動の報酬確率という，通常は未知の環境情報が必要となり，チートであるのは間違いない．つまり，基準 R というパラメータをいかにして設定するか，あるいは満足化エージェントがそれをいかにして自ら試行錯誤を通じて獲得するか，という問題が存在する．

ただし，適切な基準という情報が利用可能な場合にそれをアルゴリズムが活用できること自体は優れたことである．UCB を含めて他のアルゴリズムにはそれができない．また，パラメータ R の設定の難しさということであれば，他に同様に直感的な ϵ-greedy の ϵ と比べることができる．どのくらいの ϵ の値が最適であるのかについては，問題を固定した上での試行錯誤によるチューニングが多くの場合に必要であろう．また，実際には ϵ は固定でなく，序盤は高く，広く探索し，徐々に下げてアニーリングしていかなければ，長期的な最適行動への収束は望めない．このアニーリングにどのような単調減少関数を用いるのかは，それなりに難しい問題である．この意味で，パラメータ R は扱いやすいものである．

5.5 人間のリスク態度との関係

RS の期待値と満足化基準との差にサンプル数をかけた価値づけは，人間のリスクの扱い方に類似している．満足化基準 R を損益分岐点のような，得失の境になる水準と考える．このとき， RS 的な価値付けは，リスク追求とリスク回避を， R を軸として対称に表現することとなり (反射効果)，R がどこに置くかがフレーミングの効果を決定する．これはアジア病気問題における「400 人の死者」という表現と「200 人の生存者」という表現が，600 人の死者を見込み・基準とした場合にどちらも同じく，そのうちの 200 人が助かるということを意味するとしても，表現の違いによってリスクの扱いと選好が逆転することと似ている [Tversky 81]．同様に，同じ当たり確率，たとえば 0.6 という期待値にしても,「基準である 0.4 よりも 0.2 高い当たり確率」と考えるか,「基準である 0.8 よりも 0.2 低い当たり確率」と考えるかで，人間の考え方も変わってくるのかも知れない．いずれ，このような意味で， RS は人間の認知の特性を実装した価値関数を言えるし，また逆に，RS のパフォーマンスの高さや分析から，人間がなぜそのようにリスクを扱っているのかの解明も可能となるかもしれない．今後この点はバンディット問題などの実験で検証したい．

6. 結論

本論文では，人間のリスク態度を組み込んだ，満足化戦略の価値関数のレベルでのモデルを導入し，その基本的なパフォーマンスを調べた．今後は詳細な理論的分析と，満足化基準パラメータのオンラインでの更新方法の開発を行う．また，ベルヌーイ的報酬に限らない，強化学習一般への適用も必要となる．満足化はその基準が「最適切」に設定されているという条件の下で最適化を意味することになり，その意味で矛盾しない．他方で，満足化と最適化のプロセスは性質が異なっていることが予想され，それが今回の結果のスケーラビリティに繋がっているのかもしれない．いずれ，強化学習の適用範囲が拡がり，脳やロボットの強化学習を現実的なオーダーで行うためには，あるいは動物の生存やロボットの機能維持という課題に対しては，満足化戦略は有効な一方策となりうると思われる．

参考文献

[Auer 02] Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47:235–256 (2002).

[Bubeck 12] Bubeck, S., Cesa-Bianchi, N.: Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5:1–122 (2012).

[Oyo 13] Oyo, K., Takahashi, T.: A cognitively inspired heuristic for two-armed bandit problems: The loosely symmetric (LS) model. *Procedia Computer Science*, 24:194–204 (2013).

[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, 人工知能学会論文誌, 22(1):58–68 (2007).

[Tversky 81] Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458 (1981).