2L1-3

N 本腕バンディット問題における素朴満足化と満足化基準の更新法

Satisficing and update of its reference value in *n*-armed bandit problems

後藤田 大地[*1] Daichi Gotoda
大用 庫智[*2] Kuratomo Oyo
高橋 達二[*1] Tatsuji Takahasi

[*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 Tokyo Denki University

Satisficing, a heuristics that represents the theory of bounded rationality, is optimization when the aspiration level is set somewhere between the value of the optimal and second optimal actions. This fact leads to another possibility to more efficient optimization in action planning under uncertainty, known as n-armed bandit problems or reinforcement learning in general, as far as we can show that satisficing can be executed at an optimal speed. We propose a way to satisfice with online update of aspiration level and compare the result with one of the standard algorithms introduced with the idea of "optimism in face of uncertainty."

1. はじめに

現在，バンディット問題はインターネット上の広告事業事業(e.g, ABテスト)などに用いられるようになってきている．特に重要なのは限られた情報の中で，獲得した情報を如何に効率よく活用し報酬を獲得することである．ここで問題になるのは，速さと正確さのトレードオフから導かれる探索と知識利用のジレンマである．近年，不確実性下の行動選択であるバンディット問題の解法として，「受容可能な基準を満たす選択肢を見つけるまで探索する」という人間の満足化[Simon 56]を行う緩い対称性(LS)モデルの有用性が示されてきている[大用 15]．先行研究では報酬確率を開示した状態での満足化調整による成績の変化を静的に行っていたが，通常のバンディット問題で高成績を出すためには常に満足化のパラメータを動的に変化させ，最適な基準を保つ必要がある．そこで本研究では，満足化のパラメータ R を動的に変化させ，効率的な満足化を行うことによって，既存アルゴリズム(UCB[Auer 02])よりも高い成績を目指す．

2. バンディット問題

バンディット問題は不確実な環境下で探索と知識利用を行い，報酬の最大化を目的としている．この問題は既存の情報を活用し最適な選択を行うという側面と，現状では結果を生み出さないが新しい結果を導くために探索を行うという 2 つの側面を合わせ持っている．バンディット問題では報酬確率の不明な複数のスロットマシン(腕)に対して 1 度に 1 回の試行を行う．その報酬確率に従い報酬の有無(0 または 1)が決定する．この問題は探索と知識利用のジレンマと速さと正確さのトレードオフを表現しており，機械学習における基本的な問題とみなされている[Sutton 98]．

3. Greedy 法

Greedy 法は貪欲法とも呼ばれる強化学習の基本的なポリシーである．このポリシーは，2 本腕の場合，腕 A と腕 B の価値を条件付き確率，M(1|A)，M(1|B)として計算する．その後腕を選択する際は，この価値が高い方の腕を選択する．ただし全ての腕の価値が等しい場合は腕の選択はランダムに行われる．Greedy 法において，探索は主観的な価値が低い方の腕の選択となり，知識利用は主観的な価値が高い方の選択となる．探索と知識利用はどちらも重要だが両立できないため，探索と知識利用のジレンマが発生する．

4. UCB

現在，バンディット問題の標準的なアルゴリズムとしてUCB(Upper Confidence Bound)が良く知られている[Auer 02]．UCB は推定した報酬確率に信頼区間を加えるという楽観主義的なアルゴリズムであり，評価値を本来よりも高く見積もる．このアルゴリズムは試行回数が十分でない腕に対し探索を行い，試行回数が多くなったときに知識利用を行う．そのため長期的に見ると成績は良くなるが，最適解に到達するまで試行回数が必要になることが欠点となっている．UCB は最初に全ての腕を選択し，その後価値関数$\mathrm{UCB1(j)} = \bar{X}_j + \sqrt{2\ln n / n_j}$ を用いて Greedy 法を行う．$\bar{X}_j$ は腕 j の評価値，n_jの選択回数，n は全ての腕に対する選択回数を意味している．

4.1 UCB1-Tuned

今回のシミュレーションで用いるのは，UCB1 に改良を加えたUCB1-Tuned(UCB1T)を用いる．UCB1T の価値は次のような式で定義される．

$$\mathrm{UCB1T(j)} = \bar{X}_j + \sqrt{\frac{\ln n}{n_j}\min\{\frac{1}{4}, V_j(n_j)\}}$$

ここでの，$V_j(s)$は$v_j + \sqrt{(2\ln n / s)}$を表しており，$v_j$は腕j の報酬に対する分散である．本研究では，UCB1よりも UCB1T の方が高成績を示したため，UCB1T の結果のみ示す．

5. LS

5.1 対称性と相互排他性

LS は，人間の因果的直観の特性を保ちながら効率的に良い意思決定(バンディット問題)を行うモデルとして発見された[篠

連絡先:高橋達二,東京電機大学,350-0394 埼玉県比企郡鳩山町石坂,049-296-5416,tatsujit[at]mail.dendai.ac.jp

表 1:2 本腕バンディット問題の 2×2 分割表

		報酬	
		1	0
行動	A	a	b
	B	c	d

原 07]. 対称性は, この「もしpならばq」という条件文に対し, その逆である「もしqならばp」を推論するバイアスである. 同様に, 相互排他性はその裏である, 「もし$\bar{p}$なら$\bar{q}$」($\bar{X}$はXの否定を意味している)を推論するバイアスである. このようなバイアスは日常生活でもしばしばみられる. 対称性は, 幼児が未知の物体から未知の文字の関係性を学んだ際, その反対である文字から物体を特に親などから教えられることなくそれを理解していることなどである(実際のりんごと音声のりんごの関係性を理解するなど). 相互排他性は, 実際の 2 つの物体と 1 つの物体名を知っている状況で, 新しい物体名を与えられた際, その物体名と 2 つのうちの 1 つの物体名の関係性を推論することである(物体:りんごとみかんと音声リンゴを理解している場合に, 新しい音声:ミカンとみかんの関係性を推論するなど).

5.2 緩い対称性モデル

対称性と相互排他性バイアスを実装した LS モデルは結果qと原因候補pの 2×2 分割表を用い, 共変動情報から推定される因果関係の強さを検討する[大用 15]. 表 1 は 2 本腕バンディット問題での腕の選択と獲得した報酬の組み合わせを表している. aは腕Aを選択し報酬1を獲得した回数, bは腕Bを選択し報酬0を獲得した回数, cは腕Bを選択し報酬1を獲得した回数, dは腕Bを選択し報酬0を獲得した回数となっている. LS ではpからqへの関係性の強さを表現する条件付き確率と同様の形式を持つ確率的な関数である. また LS は対称性と相互排他性バイアスが常に完全に満たされる訳ではなく, 状況に応じて調整される方が自然であるという着想から, 条件付き確率に含まれていない項, (3), (4)を含む形で定義された[篠原 07].

$$LS(q|p) = \frac{a + \gamma_1}{a + \gamma_1 + b + \gamma_2} \quad (1).$$

$$LS(q|\bar{p}) = \frac{c + \gamma_1}{c + \gamma_1 + d + \gamma_2} \quad (2).$$

$$\gamma_1 = bd/(b+d) \quad (3).$$

$$\gamma_2 = ac/(a+c) \quad (4).$$

5.3 LS と満足化

本研究では, 満足化基準を用いて拡張した LS を用いる. この式に含まれている R は満足化基準を意味しており, R, $\bar{R}$ は $[0, 1] \ni R$, $\bar{R} = 1 - R$ を満たす. 満足化基準 R を含んだ LS の式は以下のように定義される.

$$LS(q|p) = \frac{2\bar{R}a + 2R\gamma_1}{2\bar{R}(a + \gamma_2) + 2R(\gamma_1 + b)} \quad (5).$$

$$LS(q|\bar{p}) = \frac{2\bar{R}c + 2R\gamma_1}{2\bar{R}(c + \gamma_2) + 2R(\gamma_1 + d)} \quad (6).$$

6. シミュレーションと設定

このシミュレーションでは, 2 本腕と N 本腕バンディット問題を用いる. シミュレーションを行う際, 全てのシミュレーションは 10 万回実施し, その平均を結果としている.

6.1 2 本腕バンディット問題

この 2 本腕バンディット問題は, 腕A, 腕Bにそれぞれに対応する報酬確率(P_A, P_B)によって定義される. ここでのP_iは腕i の報酬確率であり, 腕iを選択することによって, P_iの確率で報酬 1, $1 - P_i$の確率で報酬 0 を得ることが出来る. 1 回の試行で選択する事が出来る腕は 1 本に制限している. また, 腕を選択して報酬(1 か 0)を得る回数を本論中では Step としている.

本論文ではバンディット問題で最も基本的な指標である後悔を示す. 後悔は期待損失のことであり, 全ての試行において最適な腕を選択したときと, 実際し選択した腕によって発生した報酬の期待値の差である.

このシミュレーションでは 2 本腕バンディット問題における腕A, 腕Bの報酬確率の設定を 2 設定に分けて行う. 1 つ目は, 報酬確率P_A, P_Bに対し, [0.0, 1.0] 区間で一様乱数を生成する. これを全範囲確率と呼ぶ. 2 つ目は, 報酬確率P_A, P_Bを [0.5, 1.0] と [0.0, 0.5] の区間でそれぞれ一様乱数を生成し各腕に設定する. これを単高確率とする.

この 2 つのそれぞれの設定で UCB1T と適切に満足化基準が調節された LS, 満足化基準を動的に変化させる LS を比較した. 適切に満足化基準が調節された LS は最適な満足化を行った際の結果を用いており, $R = \min(P_A, P_B) + |P_A - P_B| \times \tau$ を用いている. τ は[大用 15]の結果より 0.6 に設定した. 新しい手法(動的に変化させる LS)は 2 種類ある. 1 つ目は満足度 R を現在の推定報酬確率の中間値に設定するパターン(Setting1)である. この設定の場合, やや緩やかではあるが動的に最適な満足化が可能であり, 比較的早く効率的な満足化を実現することが可能であると考えた. R の変化式は $R = (valueA + valueB)/2.0$である. この変化式は一定回数ごとに更新を行い, 更新頻度の変化による後悔の変動も調べた. またここでの valueA, valueBとは真の報酬確率ではなく, 現在の Step 数での推定報酬確率である. 2 つ目は, 楽観的な満足化基準の抑制(CI)である. この手法は信頼区間(Confidence Interval)を用いており, 略称はその頭文字を取り CI としている. 問題設定に対して基準を楽観的に高く設定し, 探索を増加させ, 得られた情報に応じ楽観的に見積もった満足化基準を下げる事によって適切な満足化を図る手法である. R の変化式は以下の通りである.

$$R = P(1|X) + c\sqrt{P(1|X) \times (1.0 - P(1|X))/n} \quad (7).$$

ここでの右辺の第二項は信頼区間の計算を行っており, 上位の腕の価値を信頼区間の範囲で変動させることにより, 満足化基準を楽観的に見積もる事が出来る. また, cは任意で設定する変数となっており, 今回のシミュレーションでは一貫して $c = 0.5$にしている.

6.2 N 本腕バンディット問題

これまでは, 2 本腕の場合を扱ったが, 現実ではそれ以上の選択肢があり, 多くの不確実情報が存在している. そのため次は, N 本腕バンディット問題を扱う. N 本腕バンディット問題では, 各腕の報酬確率(P_1, P_2, …, P_N)を一様乱数で[0.0, 1.0] 区間の実数値を生成し, 設定した. 腕の本数 N は 10, 100, 1000 とした. シミュレーションは 2 本腕バンディットとは異なり, UCB1T, LS に加えて R の変化パターンである CI の結果のみを示す Setting3 の方法では, 全範囲確率においては探索に掛ける手数が多くなり, 成績がとても悪くなってしまうためである.

N 本腕バンディット問題において LS の性質を見るためにトーナメント形式の評価 LS を用いる. この手法は, N 本腕に対してその中からランダムに 2 本の腕を選択し組を生成していく. その

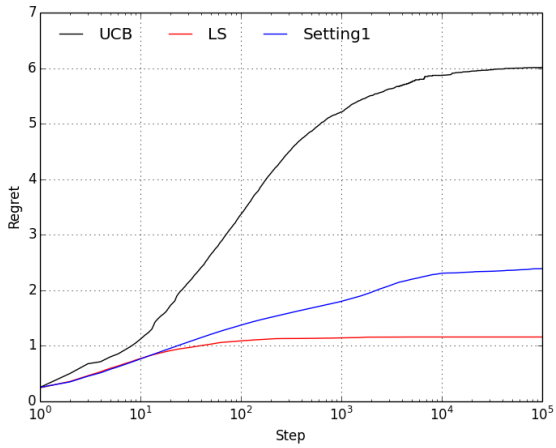

図 1:2 本腕バンディット問題における Setting1 と LS(τ=0.6)と UCB1T の後悔

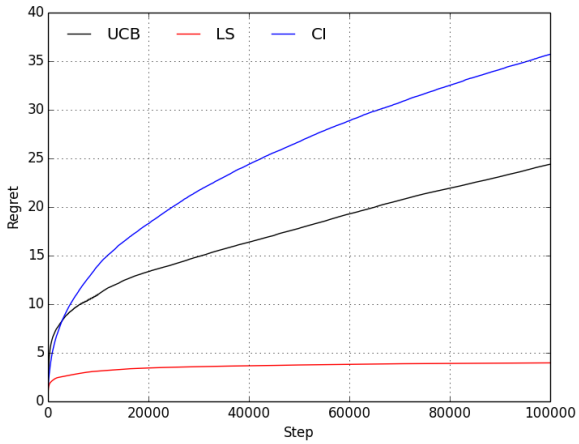

図 2:2 本腕バンディット問題における CI と LS(τ=0.6)と UCB1T の後悔

後その組毎に LS による評価を行い評価の高い腕同士でまたランダムに組を生成する. 以上のような動作を繰り返し行い, 腕が 1 本になるまで続けていく. また, 腕の本数が奇数になった場合は, その腕を決勝戦までシード扱いとし, 決勝戦まで残った腕と組みを作り LS による評価を行うようになっている.

7. 結果

7.1 2 本腕バンディット問題

単高確率環境下における, 考案した 1 つ目の手法(Setting1)と, LS(τ =0.6), UCB1T の後悔を図 1 に示す. その結果, Setting1 は常に最適な満足化を行う LS と UCB1T の間に位置し, Setting3 はやや LS よりに後悔が収束した. 推定報酬確率を利用したシンプルな更新を用いたが, 良い結果となった.

全範囲確率環境下における, 2 つ目の手法(CI)をと LS(τ =0.6), UCB1T の後悔に図 2 に示す. その結果, 1 万回 step までの結果は, UCB1T と LS はほぼ同じ値を示していた. 今回は, 10 万 step を行ったため, step が増えたことにより最終的な後悔は UCB1T と差がついてしまった.

7.2 N 本腕バンディット問題

N 本腕バンディット問題では, R の変化パターンとして CI の結果と最も理想的な満足化を行った結果を示す. UCB1T と CI の後悔を比較した結果を図 3 に示す. なお N 本腕バンディット

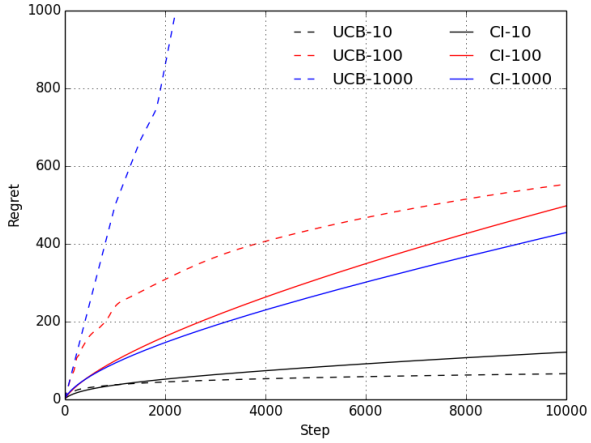

図 3:N 本腕バンディット問題における CI と UCB1T の後悔

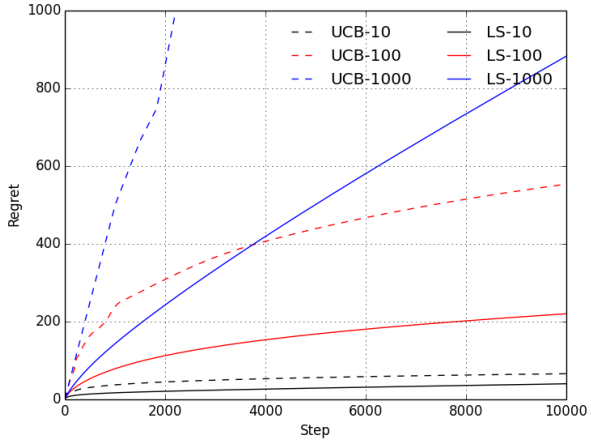

図 4:N 本腕バンディット問題における UCB1T と LS(τ=0.6)の後悔

のグラフの UCB1T-10 などの表記におけるハイフン以降の数字は腕の本数を示している. 前項の 2 本腕バンディット問題においては CI と UCB1T の後悔に差があったが, 今回 N 本腕バンディット問題を用い, 腕の本数を増加させる毎に双方の後悔の差は逆転し, 腕が増加するごとに CI の後悔が UCB1T より減り, 優秀になっていった.

次に, 理想的な満足化を行った LS と UCB1T の比較を行う. LS と UCB1T の後悔の比較を図 4 に示す. 理論上最適な満足化を行うと腕の本数に限らず, 早期に最適な満足化を行うため, UCB1T よりはるかに良い成績を示している(図 4)

8. 総合議論

まず 2 本腕と N 本腕バンディット問題に対し, LS の満足化パラメータ R を動的に変化させた結果を要約する. 本研究では Setting1 と CI という R の新しい変化式を提案し, 2 本腕バンディット問題と N 本腕バンディット問題においてシミュレーションを行った. 2 本腕バンディット問題では単高確率環境下で Setting1 を, 全範囲確率環境で CI のシミュレーションを試みた. Setting1 では最適な満足化を行った LS には劣るものの UCB1T より高い成績を示した(図 1). CI では Setting1 の環境よりも厳しい環境下でシミュレーションを行ったが UCB1T より高い成績を示した(図 2). N 本腕バンディット問題ではバンディットの腕の本数を 10, 100, 1000 に設定, CI のみシミュレーションを行った. 10 本腕では UCB1T, LS にも劣る成績を示した(図 3, 図

4). しかし, 100 本腕では LS にはまだ劣るが UCB1T よりも高い成績を示し, 1000 本腕では LS よりも高い成績を示した.

Setting1 が 2 本腕バンディット問題において高い成績を示したのは, 報酬確率の設定と R の変化式に関係があると考えられる. Setting1 での R の変化式は$R = (valueA + valueB)/2.0$である. 最適な満足化は早期に報酬確率が 1 番高い腕と 2 番目に高い腕の間に R を設定することであるため, Setting1 のような推定報酬確率の平均を取る式を用いる場合, 報酬確率のばらつきが小さいほど早い段階で R を最適化できると考えられる. そのため今回は, ばらつきが小さくなり Setting1 が高い成績を示したと考えられる.

CI が 2 本腕バンディット問題において余りよい成績を示せないのは情報のサンプル数と R の変化式に信頼区間を用いていることだと考えられる. CI においては, R を決定する際, 推定した報酬確率上位 1 位の腕を基準とし, それに対し信頼区間の計算を行い毎回その範囲内で R を変動させている. 2 本腕の場合 R を決定する際に利用できる情報は 2 つだけであり, この場合 2 つの腕の報酬確率の差が大きいと信頼区間の範囲に入るまでに手数を必要とし後悔が大きくなる. そのため良い成績を出せないと考えられる.

しかし, 信頼区間を用いている事が N 本腕バンディット問題で UCB1T より良い成績を示すことに繋がっているとも考えられる. N 本腕バンディット問題では腕の数が増加することによって報酬確率が一様にばらついた場合でも腕同士の報酬確率間隔は小さくなる. そのため報酬確率が 2 番目に高い腕が信頼区間に収まるまでの手数が減り, 早い段階で最適な満足化が行えたのではないかと考えられる.

今回は多腕での最適な満足化には成功したが, 2 本腕のような少ない情報化での満足化は行えなかった, 今後より動的な満足化パラメータの変化を良くするには, 報酬確率の差が大きい場合の最適な R がどの程度なのかを記憶し, そこから平均的な R, 人間の妥協点のようなものを設定できれば少ない情報下でも最適な満足化が行え, どんな環境でもそれなりの成績を出すことが期待できるのではないかと考える.

9. まとめ

今回のシミュレーションによって, より現実に近い複数の腕を持つ状態での試行において効率的な満足化を行うことが出来た. Setting1 を行ったような限られた範囲での試行という条件での結果においても, 人間が日常生活の中で意図的に選択肢を絞り, 限られた選択肢の中から推定を行うことがある事を考えると, 妥当な結果が得られたと考える. CI においては, 事前結果では 2 本腕且つ, 少ない試行回数の結果のみを出していたため, 腕を増やした場合の結果が見えなかったが, 今回のシミュレーションによって選択肢が多いほど信頼区間を用いた CI の満足化効率が高まることが判明した. 普段の生活では無数の選択肢から選択しなければいけない状況が多々ある. それはバンディット問題で言う N 本腕の状態と重なる. そのような状況である程度効率的な満足化をする事が出来たという事は, 一歩人間の満足感の変化に近づけたのではと考えられる. 今回信頼区間を用いただけでこのような結果を得られたことにより, 今後はこれに更なるパラメータを加えることで, 現状より効率的な満足化を行えることが期待できる.

参考文献

[Auer 02] Auer, P., Cesa-Bianchi, N., and Fischer, P.: Finite-time analysis of the multiarmed bandit problem, *Machine learning*, 47, 23–256 (2002).

[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).

[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, 30(2), 403--416 (2015).

[Sutton 98] Sutton, R. S., Barto, A. G.: *Reinforcement Learning: An Introduction*, MIT Press, Cambridge (1998).

[Sidman 94] Sidman, M.: *Equivalence relations and behavior: A research story*, Boston, M.A., Authors Cooperative (1994).

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