1G1-2in

非定常 N 本腕バンディット問題に対する人間の認知バイアスの適用

Application of Human Cognitive Bias to Nonstationary N-armed Bandit Problems

大用　庫智[*1]　甲野　佑[*1]　高橋　達二[*2]
Kuratomo Oyo　Yu Kohno　Tatsuji Takahashi

[*1]東京電機大学大学院　[*2]東京電機大学
Graduate School of Tokyo Denki University　Tokyo Denki University

N-armed bandit problems are fundamental tasks in machine learning, statistics and economics, where agents act under uncertainty tackling the dilemma between exploitation and exploration. Previous studies have shown that a value function implementing human cognitive biases (Shinohara ' s loosely symmetric (LS) model) performs very well as a heuristics in 2-armed bandit problems. However, the general application of LS has not been given yet. In this study, we analyze and generalize it to enable decision making among three or more options and adaptation to nonstationary environment. As the result, it is suggested that LS has several general properties of judgment under risk as the prospect theory describes.

1. 概要

N 本腕バンディット問題は強化学習において最も基本的な課題として知られており [Sutton 00]，機械学習の分野だけでなく経済学など幅広い分野で研究されてきた．企業の経済活動もある種のバンディット問題といえる [有賀 04, 中野 08]．バンディット問題の難しさは，より良い結果を得るためには結果に結びつくとは限らない探索行動をとらざるを得ないという探索と報酬のジレンマで表される．そのようなジレンマ最も端的に表現するバンディット問題をより良く解く手段として幾つもの数学的モデルが考案されており，その中でも損失の上限が保証された $UCB1$ が有用なアルゴリズムであるとされる [Auer 02]．

本研究では人間が持つとされる認知バイアスに着目し，二本腕バンディット問題と因果帰納[*1]のパフォーマンスを両立する緩い対称性 (*loosely symmetric*; LS) モデル [篠原 07] を用いて複雑なバンディット問題への適応性を論じている．

LS は認知的に導出されたモデル [Takahashi 10] である．LS は因果帰納のメタ分析において高い相関 ($\bar{r} = 0.96$) を示しており [大用 10]，かつ二本腕バンディット問題に対して高い適応性が示されている [篠原 07]．これらの結果から，LS は幾つかの人間の認知傾向を備えていると言える．しかし LS の理論的な分析や一般化，複雑な環境での分析は未だ不十分であった．そこで本研究では LS の分析と一般化を行い，またバンディット問題をより現実的な問題とするために，報酬確率が定常的でなく変化する環境での適応を検証した．さらに試行に対するコストが存在する選択肢を用いて分析を行った．これらの一般化された現実的な環境と N 本腕の設定で，LS と $UCB1$ を比較し分析を行っている．

2. N 本腕バンディット問題

バンディット問題は動物の採餌行動と対応付ける事で自然環境内に存在する問題として理解し易い．バンディット問題における選択肢は訪問可能な餌場 (手段) に対応し，それはバンディット問題において“ 腕 ”と呼ばれる．餌場に訪れた結果として，設定された報酬確率に従い餌の獲得の有無 (報酬 0 または 1) が決定される．未知の環境に放り込まれた動物は試行を行わずに餌場の情報を得る術を持たない．そのため，どのように探索を行い，どの段階で蓄積された情報をどの段階でどのくらい活用するかのバランシングが問題となる．その解決には方策と価値関数を工夫して適切に評価を行う事が必要となる [Sutton 00]．

本研究では複数の餌場の報酬確率が 0.5 以上である富んだ環境を高確率環境，0.5 未満の貧しい環境を低確率環境と定義する．また 0.5 以上の報酬確率を持つ餌場が一つしか無い環境を単高確率環境と定義し，餌場環境を報酬確率の (0.5 を基準とした) 高低から分類する．この問題に対し，完全に合理的な存在であれば利潤を最大化させる様に行動すると考えられる．しかし行動経済学では人間や動物はしばしば非合理的な行動をしてしまう事が知られており，その誤りは非論理的な認知バイアスとして表す事が出来る．

2.1 損失の上限が保証されたアルゴリズム

最終的には確実に正確な選択肢の判断が行えるアルゴリズムとして $UCB1$ が知られている [Auer 02]．このアルゴリズムは初めに選択可能な選択肢を全て一回づつ選択し，その後，各 $UCB1(X_j) = \bar{X}_j + \sqrt{\frac{2\log n}{T(n_j)}}$ という評価式から算出される値が最も高い腕を選択し適応していく．ここで $\bar{X}_j$=$\bar{X}_{j,T_j(n)} = \frac{1}{T_j(n)}\sum_{\gamma=1}^{T_j(n)} x_{j,\gamma}$，$n$ は選択回数，T_j(n) は腕 j の選択の回数，$x_{j,\gamma}$ は γ 時点での腕 j の報酬である．更に $UCB1(X_j)$ を改良した $UCB1'(X_j)$[*2]が存在し，より高い成績を持つとされる．

$$V_j(s) = \left(\frac{1}{s}\sum_{\gamma=1}^{s} x_{j,\gamma}^2\right) - \bar{X}_{j,s}^2 + \sqrt{\frac{2\log n}{s}}, \qquad (1)$$

$$UCB1'(X_j) = \bar{X}_j + \sqrt{\frac{2\log n}{T_j(n)} min\{\frac{1}{4}, V_j(T_j(n))\}}. \qquad (2)$$

式 1～ 2 から分かる様に $UCB1'$ は数学・統計学的なアルゴリズムであるとされる．

連絡先: 大用庫智 e-mail:kuratomo.oyo[at]gmail.com

*1　因果帰納とは，原因 C と結果 E の二事象間の因果関係を C と E の共起情報から帰納的に推論する事である．共起情報は原因 C と結果 E の在，不在の組合せからなる二事象の発生情報 (CE, $C\bar{E}$,$\bar{C}E$, $\bar{C}\bar{E}$) であり，表 1 のように頻度 a, b, c, d で表す事が出来る．

*2　[Auer 02] ではこのアルゴリズムは $UCB1$-*tuned* と呼ばれている．

3. 認知バイアスの有用性

バンディット問題において「餌場 A ならば餌を獲得しやすい」という情報が与えられた時「餌を獲得するならば餌場 A が良い」と推論する傾向を対称性バイアスと言い, また「餌場 A でないならば餌を獲得しにくい」を導く傾向を相互排他性バイアスと言う．これらの連想は与えられた命題に対して逆と裏の関係であり, 論理的に正しい推論ではない．しかし, この誤った推論が因果関係の帰納に有用であるとされる [篠原 07, 中野 08]. 因果帰納における対称性は逆向きの推論が真であると判断出来る場合に,「餌場 A ならば餌を獲得」の因果強度を強める役割を担っている．バンディット問題でも同様に, 報酬から餌場に対する逆向きの確率を参照することで餌場と報酬の結びつきを強めている．しかしそれが過ぎれば判断を誤らせる原因にもなりえる．相互排他性は餌場 A における餌の不獲得から餌場 A 以外への探索を促すが, このバイアスが効き過ぎれば効果的な判断を阻害してしまう．このように人間は常に対称性や相互排他性バイアスを働かせているとは考え難く, 状況によってバイアスの強さを柔軟に調整していると考えるのが妥当である．

4. 人間の因果帰納の傾向を含む価値関数:LS

バンディット問題において試行によって得られた情報は，原因と結果からなる共起情報として 2 × 2 の分割表 (表 1) で表現でき，LS は共起情報である頻度 a, b, c, d の関数として定義される．

表 1: 餌場の選択と餌獲得に関する共起情報の分割表表現.

	獲得	喪失
餌場 A	a	b
餌場 B	c	d

a: 餌場 A での餌獲得回数
b: 餌場 A での餌喪失回数
c: 餌場 B での餌獲得回数
d: 餌場 B での餌喪失回数

LS は柔軟に調整される対称性バイアスと相互排他性バイアスを持つモデルとして考案された．他にも図と地の分離，地の不変性という特性を持つ [Takahashi 10].

$$LS(\text{獲得} \mid \text{餌場 } A) = \frac{a + \frac{b}{b+d}d}{a + \frac{b}{b+d}d + b + \frac{a}{a+c}c}. \quad (3)$$

$$LS(\text{獲得} \mid \text{餌場 } B) = \frac{c + \frac{b}{b+d}d}{c + \frac{b}{b+d}d + d + \frac{a}{a+c}c}. \quad (4)$$

ここで式 3 の $ac/(a+c)$ は対称性と $bd/(b+d)$ は相互排他性とある程度関係があるが, 本研究では前者をネガティブ項，後者をポジティブ項と呼ぶ．ネガティブ項は高確率と単高確率環境，ポジティブ項は低確率と単高確率環境で影響が大きくなる．それは高確率環境では問題設定から $a > b, c > d$ となり, 低確率環境では $a < b, c < d$ となる傾向が強いためである．

地の不変性はここで, 式 3 と 4 のポジティブ項とネガティブ項が等しいということであり, 結果計算が $UCB1'$ よりもはるかに簡便となる．ある選択肢に対して着目すると, それ以外の選択肢を視覚における”地”と判断して結果に対して中立な選択肢と判断する．これを図と地の分離と言う．この性質から餌場 A を選択し続けると LS(獲得 | 餌場 A) $\simeq P$(獲得 | 餌場 A), LS(獲得 | 餌場 B)$\simeq$ 0.5 に収束する．LS は二種のバイアス項の関係から視覚における図と地に例えられる視点を導き，その視点から良く環境を捉えてバンディット問題に適応しているといえる．

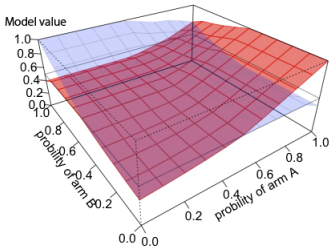

図 1: LS(獲得 |A)(赤) と LS(獲得 |B)(青) の各平均値.

4.1 LS エージェント:greedy 法と基準率

二本腕バンディット問題で LS エージェントを実装する際には方策として $greedy$ 法を用いる．$greedy$ 法では行動価値を 2 × 2 の分割表から LS で計算し，報酬獲得という結果に対する因果強度が最大になる餌場 (原因) を選択する．即ち LS(獲得 | 餌場 A) と LS(獲得 | 餌場 B) における評価値が大きい餌場を常に選択していく．定常的な二本腕バンディット問題における LS(獲得 | 餌場 A) と LS(獲得 | 餌場 B) の各平均値 (1000 回選択後,1000 回平均) を図 1 に示す．ただし共起情報の初期値を 1 として LS の振る舞いを観測する。LS(獲得 | 餌場 A) の値が赤，LS(獲得 | 餌場 B) の値が薄い青，z 軸は各値である．図 1 から LS の評価値は報酬確率 0.5 を境に値の取り方が変化している事が分かる．LS の行動選択は低確率環境では探索を多く行い，高確率環境ではある特定の選択肢に執着する傾向がある．また，単高確率環境では瞬時に正しい選択をしており，たった一つの価値関数でありながら環境に対し三種の方策状態を持つと解釈出来る．

4.2 一般化

これまでに, 二本腕バンディット問題における LS の振る舞いが論じられている [篠原 07, Takahashi 11]. しかし現実の環境では報酬を得るための餌場は一般に二つ以上存在する．そこで LS を N 本腕バンディット問題で運用するため, トーナメント形式の評価方法で選択を行う．この方法では表 1 が構成出来る様に多数の餌場から二つずつの組をランダムに生成する．その組毎に LS による評価を行い低い餌場を選択肢から除いていく．これを餌場が残り一カ所になるまで繰り返し行う．餌場の数が奇数であれば，ランダムに選択した選択肢をシードとして,「決勝戦」で評価する．評価の結果は, 評価のための組の生成方法や順序にほぼ全く依存しない．

5. シミュレーション

以下に LS を用いたバンディット問題に対するシミュレーションを行う．ただし本研究ではバンディット問題の解決を目的としていない事に注意されたい．人間の因果推論と相関が最も高いモデルをバンディット問題で運用した場合，数学的なモデルとどのような差異が生まれ，それがどのような意味を持つのかを知る事が最大の目的である．

5.1 問題設定と基本的な指標

N 本腕バンディット問題は N 個の餌場 (選択肢) とそれに対応する確率 (P_A,P_B,...,P_N) によって定義される．ここで P_X は，餌場 X の報酬確率を意味している．餌場 X は P_X の確率で 1 を，$(1 - P_X)$ の確率で 0 を返す．餌場を例にとっている事から解るように 1 step につき餌場の選択は 1 度に制限されている．P_X の設定確率は低確率，単高確率，高確率，全範囲の 4 種類の環境設定から選ばれる．低確率環境は [0,0.5] 区間の一様乱数 6 個の平均，高確率環境では [0.5,1] 区間の一様乱

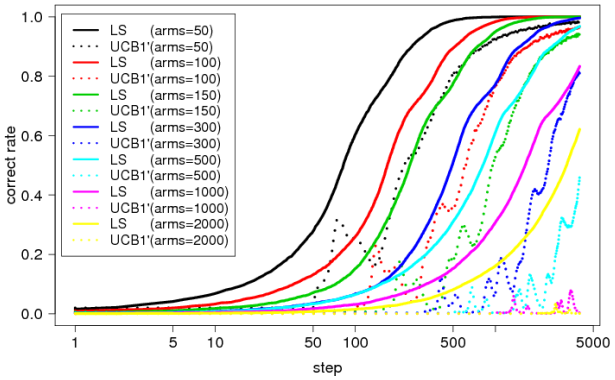

図 2: 餌場の数が 50,...,2000 本の問題正解率時間発展.

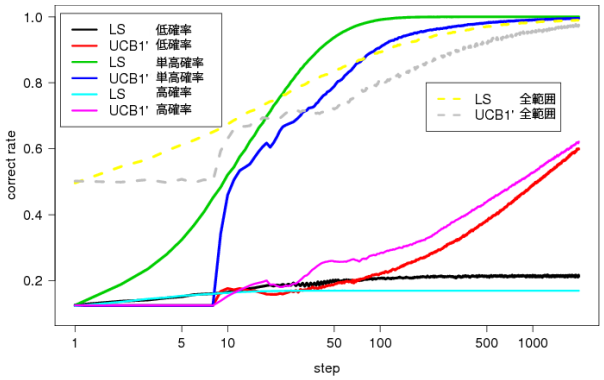

図 3: 3 種類の環境と全範囲環境で報酬確率が 50%以上の腕を正解としたときの正解率時間発展.

数 6 個の平均で設定される．単高確率環境では一つの選択肢を高確率環境，それ以外の選択肢は低確率環境で設定される．全範囲環境は [0,1] 区間の一様乱数 6 個の平均で設定される．

指標は正解率と探索率を用いる．正解率は各 N ステップ毎に $\arg\max_X(P_X)$ となる腕を選択した割合であり，探索率は N ステップまでに腕を変更した割合を意味する．この指標は偶然の影響を極力排するため，断りが無い限り各報酬確率の組毎に N ステップ行うシミュレーションを 1 万回行った平均で表されている．また本シミュレーションでは以下の 3 つの条件でエージェントを運用する:1. モデル値が一つでも計算出来ない場合はランダムに腕を選択, 2. 初期値として a, b, c, d に 1 を代入,3. モデルの評価値が最も高い腕が複数存在する場合は，それらの中からランダムに選択する．

5.2 定常 N 本腕バンディット問題

無数の選択肢の中に一カ所のみ餌が豊富な餌場が存在する単高確率環境を設定する．餌場の数は {50, 100, ..., 2000}．餌が豊富な餌場 RX の P_{RX} は，高確率環境から選ばれ，貧弱な餌場 FX の P_{FX} は低確率環境から選ばれる．選択回数 ($step$) が 4000 回のシミュレーションを 1 万回繰り返し行った LS と $UCB1'$ の正解率を図 2 に示す．同様に選択肢の本数を 8 本とし，4 種類の環境での正解率を図 3 に示す．ただし全範囲環境での正解率指標は $P_X > 0.5$ となればカウントする．図 3 に示した三つの環境での探索率を表 2 に示す．二本腕でも図 2,3 と表 2 と同様な結果の傾向性を持つ事は既に確認されている．

表 2: 50, 200, 2000 step における LS と $UCB1'$ の探索率.

環境	低確率		単高確率		高確率	
step	LS	UCB1′	LS	UCB1′	LS	UCB1′
50	.5125	.5151	.1731	.2299	.0192	.1999
200	.4574	.4441	.0460	.1151	.0047	.1682
2000	.4161	.1861	.0046	.0218	5e-04	.0772

5.3 非定常環境への問題：選択肢の環境変化

前節では定常的なバンディット問題を扱った．しかし現実においては意思決定を行うエージェントとの相互作用や外部からの影響で環境に変化が生じていくものと考えるべきである．そこである時点で餌場環境が不連続に変化する三種類の設定 (低確率，単高確率，高確率環境) で二本腕バンディット問題を行う．ここでは全ての設定で餌場確率 (P_A, P_B) の変化は 3000 ステップ目で発生する事とする．変化により低確率環境では $min(\{P_A, P_B\})$ に 0.5 を加算し，高確率環境では $max(\{P_A, P_B\})$ に 0.5 を減算する．また単高確率環境では P_A と P_B の入替を行う．これにより LS が報酬確率の大小関係の逆転に対し如何に適応するか観測する．LS と $UCB1'$ の低確率，単高確率，高確率環境変化の結果をそれぞれ 破線，実線，

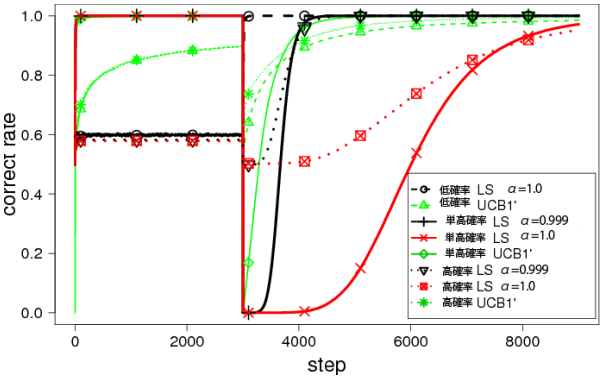

図 4: 報酬確率が変化する環境における LS と $UCB1'$，共起情報に割引率を含む LS の正解率時間発展.

点線として図 4 に示す．また人間や動物が過去の記憶の影響を減衰させゆくと考えられるため，強化学習で多く用いられている割引率 α を LS に適用する．適用方法は，2×2 分割表の各セルを更新する際に過去の情報を α で乗算した値と現時点で得られた情報を足し合わせて更新する．$UCB1'$ には方策の特徴から適用しない．割引率 α を 0.999 とした結果は凡例に α=0.999 として図 4 に示す．ただし低確率環境での LS と割引率を共起情報に含む LS の成績が同一であったため一方のみを示す.

5.4 非定常環境への問題：結果の環境変化

これまでの問題設定では餌場へ行くコストを無視して論じて来た．しかし，現実には行動は必ずそのコストを伴う．餌場であれば到達するためのカロリー消費がそれに相当するだろう．バンディット問題をより現実的な問題として扱うため選択肢 (餌場) にコストを定義した一般的な餌場を用いてシミュレーションを行う．一般的な餌場 X を選択すると P_X の確率で R_X，$(1 - P_X)$ の確率で F_X が得られる．この一般的な餌場を選択するためにはコスト CT_X が必要である．ここで R_X, F_X, CT_X は実数の変数である．ただし $CT_X \geq 0$ とする． この餌場を選択した時の利潤は R_A - F_A - CT_A である．また利潤の期待値 $E(PF(X))$ は $P_X R_X - (1 - P_X)F_X - CT_X$ である．正解率の指標を利潤の期待値が最も高い餌場を選択した割合と変更する．ここで共起情報は 4. 節の定義を拡張し利潤の分割表として扱う．つまり，$a(c)$ は正の利潤，$b(d)$ は負の利潤である．各セルの更新方法は一般的な餌場 X を引いた結果，正の利潤であれば a に，負の利潤であれば b に利潤値を加算する.

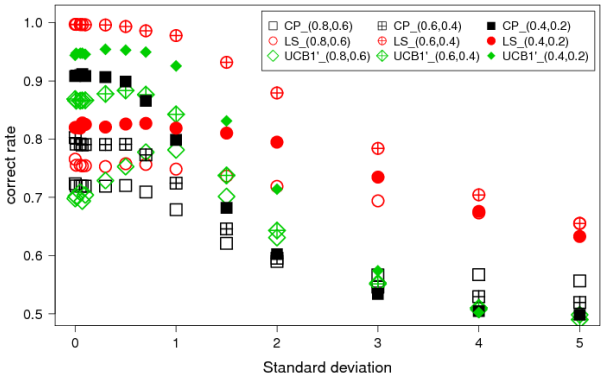

図 5: $N(1, sd)$ に対する標準偏差毎の 200 step の正解率.

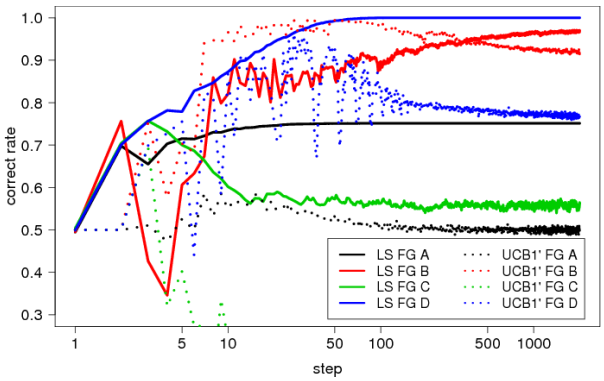

図 6: コストが異なる各餌場環境 (FG) の正解率.

また，現実において試行を行う際には必ず偶然の影響を受ける．そこで統計学の観点から偶然の程度を標準偏差として LS と $UCB1'$ の偶然性許容能力をシミュレーションする．一般的な餌場設定は $R_X = N(1, sd) = CT_X$, $F_X = 0$ である．二つの餌場のパラメータ sd(標準偏差) を [0,0.01,...,3,5] 区間 (14 種) とした際の条件付き確率と LS, $UCB1'$ の 200 ステップでの正解率を図 5 示す．報酬確率の組 (P_A,P_B) に対応するプロットは記号 (丸, 四角, 菱形) が (0.8, 0.6)，記号内に'+' を描いたものが (0.6, 0.4)，記号塗り潰しが (0.4, 0.2) である．コストや利潤の期待値 ($E(PF(Xn))$) が異なる環境 X の設定を (CT_{X1}, CT_{X2}, $E(PF(X1))$, $E(PF(X2))$) の組として A から D を設定する．環境 A は (3, 4, 0.5, 1.6), 環境 B は (2, 4.5, -0.5, -0.9), 環境 C は (2, 3.9, -0.5, -0.3), 環境 D は (2, 3, -0.5, 0.6) である．各環境の $P_{X1}, P_{X2}, R_{X1}, R_{X2}, F_{X1,X2}$ はおのおの 0.5, 0.9, 3, 4, 0 とし，その正解率を図 6 に示す． 50 ステップでの LS の各環境における探索率 (A, B, C, D) は (.01, .20, .11, .03)，2000 ステップでは (.0, .04, .1, .0) であった．

6. 考察

以上のシミュレーション結果から $UCB1'$ に対して LS はバンディット問題の目的を達成するための成績は劣ってしまう場合がある事が解る．しかし LS は無数の原因候補から真の原因の抽出が可能であり，真の原因に相当する報酬確率が 0.5 以上の選択肢を効率的に選ぶ傾向性がある．また価値関数のみで環境毎に方策を自律的に変化させ，環境変化に対してもいち早く適応する事が可能である．さらに，報酬 R とコスト CT が正規分布から生成される乱数として定義される餌場においても高い成績を示した．さらに LS は利潤が正の選択肢を発見すると選択肢を変更せず，利潤が負の選択肢のみであれば幅広く探索する傾向がある．ただし利潤が負の環境でも，一方の損失が相対的に大きければ正解の選択肢を選択する事が可能である．

7. 結語

本研究では非定常的なバンディット問題とコストを含んだ選択肢でバンディット問題をより現実的な問題として検証を行った．また LS の行動分析を回数分割表，利潤分割表を用いて行った．回数分割表のセル更新は選択肢を選択すると $a = a+1$ か $b = b+1$ であり，LS の行動選択は報酬確率 0.5 で変化していた．利潤分割表が回数分割表と同一なセル更新と選択基準を持つためには，利潤の期待値から $R = 2, CT = 1, F = 0$ の一般的な餌場と決定できる (利潤 0 になるのは報酬確率 0.5)[*3]．即ち LS はセル更新式と行動基準から意味的に二種の餌場を同一視して行動している．これらは LS の行動変化が利潤の期待値 0 において変化する事を意味し，利潤の期待値が正の環境ではリスク回避の挙動をとる．利潤の期待値が負の環境ではリスク追及の挙動をとるが，相対的に損失が大きい選択肢がある場合にはホットストーブエフェクトのような効果がある事が分かる．このように，LS はプロスペクト理論が扱うような参照点依存性と感応度逓減性，損失回避性などを持つ人間的なモデルである可能性が得られた．また，この利得基準値は共起情報に定数を乗算する事で歪めることができ LS が不得意な部分でも正解率 100%に到達する事が出来る．LS は利得基準値よりどの程度変化しているかを見ながら判断している．また LS が分数式であるため，初期では値の変動が大きく後半ではあまり変化しない．今後，これらの性質をさらに分析し，餌場を市場として捉えることで，様々な基準を持つエージェントによる人工市場などのシミュレーションを行っていく．

参考文献

[有賀 04] 有賀 裕二: 進化経済学の数理入門, 共立出版

[Auer 02] Auer, P. Cesa-Bianchi, N. Fischer, P.: Finite-time analysis of the multi-armed bandit problem, *Machine Learning*, 47, 235-256 (2002)

[Hattori 07] Hattori, M. Oaksford, M.: Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science*, 31(5), 765-814, (2007)

[中野 08] 中野昌宏, 篠原修二.: 対称性バイアスの必然性と可能性：無意識の思考をどうモデル化するか, 認知科学, Vol.15, No.3, pp.428-441 (2008)

[大用 10] 大用庫智, 高橋達二.: 因果推論と意思決定を結ぶ緩い対称モデル, 日本認知科学会第 27 回大会発表論文集, 799-800. (JCSS2010)

[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, Vol.22, No.1, pp.58-68 (2007)

[Sutton 00] Sutton, R. S. Barto, A. G.:強化学習, 森北出版, (翻訳) 三上 貞芳, 皆川 雅章

[Takahashi 10] Takahashi, T. Nakano, M. and Shinohara, S.: Cognitive Symmetry:Illogical but Rational Biases, *Symmetry:Culture and Science* 21, 1-3, 275-294. (2010)

[Takahashi 11] Takahashi, T. Oyo, K. and Shinohara, S.: A Loosely Symmetric Model of Cognition, *LNCS*, 5778, 238-245,(2011)

*3 二本腕において報酬確率全組合せ (10,201 種) で試行した結果各指標の誤差が殆どなかった.