1L4-OS-24b-4

知識利用と探索のジレンマに対する因果的価値関数の適用とそのベイズ的分析

An Application of Causal Value Function for the Exploration–Exploitation Dilemma and its Bayesian Analysis

大用 庫智[*1] Kuratomo Oyo
高橋 達二[*2] Tatsuji Takahashi

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

Value function, LS model, that express the causal intuition of human balances exploitation and exploration and achieves high performance in the most basic bandit problems in reinforcement learning for modeling the learning under uncertain environment. We perform Bayesian analysis to LS, discuss about the meaning of the model, show that the optimality can be obtained by a parameter called intuitive standard of value and clarify a mechanism for balancing exploration and exploitation.

1. はじめに

バンディット問題は、当たり確率不明の複数のスロットマシンの中から一度に一つ選択し、獲得当たり数の最大化を目的とした強化学習の最も基本的な問題として知られている[Sutton 98]。この問題は、一見単純かつ簡単に思えるが、より良い結果を得るために結果に結びつくとは限らない情報収集(探索)とその探索中に得た情報から報酬を得ようとする報酬獲得(知識利用)の両者が相反するために、知識利用と探索のジレンマとして知られる難問を抱えている。ただ、他の強化学習の問題と比較するとバンディット問題は単純であるために、知識利用と探索のスイッチングまたはバランシングを行うモデル研究が盛んに行われている[e.g. Auer 02]。バンディット問題の応用範囲は広く、たとえば脳科学では、逐次的意思決定を行う人間の脳が選択肢に対して相対的に評価を行うことが明らかにされており[e.g. Daw 06]、ゲーム理論では囲碁 AI の着手選択アルゴリズム(モンテカルロ木探索)[Kocsis 06]などが幅広く研究されている。

本研究では対称性と相互排他性と呼ばれる認知バイアス(人間の認知に観られる規範的な合理性に従わない偏り)を自律的に柔軟に調節するという観点から考案された緩い対称性(LS)モデル[篠原 07]をバンディット問題に適用する。対称性と相互排他性バイアスは、ほぼ人間固有のバイアスとして知られている。我々が対称性バイアスと呼ぶものは、条件文$p \to q$から$q \to p$を意識下で想起する推論であり、行動分析学の刺激透等価性の対称性、条件文研究では双条件解釈を確率的に拡張した等確率仮説[服部 08]である。$p \to q$(例えば約束や脅し、命令など)から$\bar{p} \to \bar{q}$を意識下で誘導する誘導推論[Geis 71]や発達心理学の幼児の語彙獲得の相互排他性などを単に相互排他性バイアスと呼ぶ。特に対称性については、服部が思考心理学における広範な存在を示し、確率論的観点から等確率性仮説として定式化し、それに着目して高次の認知モデルを提唱している。たとえば因果帰納とは事象間の共変動情報から因果関係を帰納的に形成する推論であるが、これに関して服部は DFH モデルを提唱し[Hattori 07]、その有効性を証明している[服部 08]。

本論文で扱う LS は人間の因果帰納の傾向を良く記述する(122 種類の刺激を含む因果帰納メタアナリシス[Hattori 07]で LS は人間の因果的直感との相関が 0.96 である[Oyo 12])。LS はモンテカルロ木探索[西村 12]や人間の認知バイアスを実装した音楽生成システム[大村 12]、強化学習の Giant-Swing Robot のコントロール[Uragami 11]等で幅広く利用されている。LS の定義の必然性については、高橋らによってゲシュタルト心理学における「図と地」に例えた分析がある[Takahashi 11a, 11b]。

本研究では、ベイズ理論を用いて、人間の因果的な直感を含む LS を理論化する。そして、LS を価値関数としてバンディット問題に適用し、直感的なパラメータにより性能が向上する事を示し、知識利用と探索の両立の機能について議論を行なう。

2. 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ず、いつまで試行を行なえるかさえも知り得ない。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(またはスイッチング)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 98]。

2.1 標準的なバンディット問題のアルゴリズム

バンディット問題の現在最も標準的なアルゴリズムは UCB1 である[Auer 02]。このアルゴリズムは十分な選択回数が許されれば高い成績を示し、後悔の上限が対数的に保証されている。その性質のために、囲碁の着手選択部分のモンテカルロ木探索に UCB1 を応用した UCT アルゴリズムが広く利用されている[Kocsis 06]。UCB1–Tuned は UCB1 を更に改良したモデルであり、餌場 j の価値は次の様に定義される[Auer 02]。

$$\mathrm{UCB1\text{-}Tuned}(j) = \bar{X}_j + \sqrt{\frac{\ln n}{n_j}\min\left\{\frac{1}{4}, V_j(n_j)\right\}},$$
$$V_j(s) = \left(\frac{1}{s}\sum_{\tau=1}^{s} X_{j,\tau}^2\right) - \bar{X}_{j,s}^2 + \sqrt{\frac{2\ln n}{s}}, \quad (1).$$

ここで、$\bar{X}_j$は選択された餌場 j の報酬の期待値である。n_jと s は餌場 j を選択した回数である。nは全体の選択回数である。$X_{j,\tau}^2$は τ 回選時の選択した餌場 j の獲得報酬である。このモデルは、最初に全ての餌場を一度ずつ選択する。その後、UCB1–Tuned の値が最も高い餌場を選択しながら問題に適合する。UCB1–Tuned は、サンプル数を考慮する($\bar{X}_j$でない)項によって、$\bar{X}_j$が高い餌場への試行(知識利用)を停止し、$\bar{X}_i$が低い餌場への試行(探索)を促すかどうかを決定(バランシング)する。

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

3. 人間の因果的価値関数:LS

LS は以下の 2×2 の分割表上の共変動情報a, b, c, d から信念の計算を行なう因果推論のモデルとして篠原によって定義された[篠原 07]。

表 1：2×2 の分割表と共変動情報

	試行の結果	
	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

表 1 から、認知バイアスとは無相関であり、「餌場 A ならば獲得」の最も基本的な指標、条件付き確率（CP）は$P(\text{獲得}|\text{餌場 A}) = a/(a+b)$として計算できる。CP の各評価値は各餌場の情報のみから評価（絶対評価）される。対称性と相互排他性バイアスを常に完全に満足する rigidly symmetric (RS)はRS$(\text{獲得}|\text{餌場 A}) = (a+d)/(a+b+c+d)$として定義される。RS の評価方法は CP とは異なり、例えば、餌場 A の喪失情報（b）から餌場 B の価値を高める（餌場 A の価値は下がる）といったシーソーゲームのような評価（相対評価）を行う。

人間の認知はバイアスと無関係または完全にバイアスを満足するとは考えにくいというアイディアから、LS は人間に根本的な認知のバイアスとして考えられる対称性と相互排他性を緩く柔軟に満たすモデルとして以下の様に提案された。

$$LS(\text{獲得}|\text{餌場 A}) = \frac{a + biasB}{a + b + biasA + biasB}, \quad (2).$$

$$LS(\text{獲得}|\text{餌場 B}) = \frac{a + biasB}{c + d + biasA + biasB}, \quad (3).$$

LS は CP に無い共通項を用いて CP と RS の間でバイアスを調整し、RS の様な相対評価を行う。biasA は獲得情報 a と c、biasB は喪失情報 b と d の調和平均の 1/2 の値（$biasA = ac/(a+c), biasB = bd/(b+d)$）である。バンディット問題での LS は式(2)と(3)の値が大きい方の餌場を選択する greedy 法で運用され、餌場を選択した結果は表 1 に蓄積される。

LS はバイアスを調節するだけでなくゲシュタルト心理学における図と地に例えられる形で認知的に正当化されている[Takahashi 11b]。バイアスを調整する項（biasA と biasB）は全ての餌場（選択肢）に対して共通である。この性質はたとえ図（選択肢）が変化しても地（基準）が不変である地の不変性と呼ばれている。この性質は計算量削減に役立つ。LS は共通の項を含むが各選択肢について評価することができる。例えば、一つの餌場に焦点を当て続けると、LS は他の餌場を視覚における地（結果に対して中立）と判断する（0.5 を割り当てる）。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。

3.1 経験ベイズ法と LS

前節の様に、LS は高橋らを中心に十分な認知的な意味を持つモデルとして分析されている。それに対して、ここでは経験ベイズ法の枠組みを用いて LS に理論的な背景を与える。

経験ベイズ法は、一般的に知られているベイズ推定とは事前分布を過去の経験情報から形成するという点で異なる（データ全体の平均や分散などから事前分布を形成する）。その方法は、例えば、各野球選手の 45 打席目の打率データを用いて、以下の事後分布（正規分布）の期待値（ベイズ推定値）から本来の各選手の能力を推定するために使われている[Casella 09]。

$$\bar{\theta} = \left(\frac{(\rho-3)\sigma^2}{\sum(X_j - \bar{X})^2}\right) \times \bar{X} + \left(1 - \frac{(\rho-3)\sigma^2}{\sum(X_j - \bar{X})^2}\right) \times X_j \quad (4).$$

ここで ρ と$X_j, \bar{X}, \sigma$ は、それぞれ、選手数、選手 j の打率, 打率の平均、正規分布のパラメータを意味する。式(4)から、推定対象の選手の打率は、45 打席の打率の全体的な傾向とその選手の打率の加重平均であり、個の推定を行なうために全体の傾向を加味するスタインの縮減推定である。その他に売買時の意思決定で事前と事後分布にベータ分布を用いた例などがある[Casella 09]。この推定法は、誤差の少ない推定や全体の傾向を加味する事で個々の比較を滑らかに行なえるという利点がある。経験ベイズ法は事前分布に（相場観、巨視的、基準の様な）全体傾向として意味を与える。

扱う問題が二項分布であるため、LS の導出には事前と事後分布がベータ分布になることが知られているベータ–二項モデルを用いる。このモデルを用いる事で事後分布∝事前分布×尤度関数の計算が簡単になり、事前分布$g(\theta|\alpha, \beta)$に対して、事後分布は以下の様に$g(\theta|a + \alpha, b + \beta)$ となる。

$$g(\theta|a + \alpha, b + \beta) = \frac{g(\theta_i|\alpha,\beta) f(a|(a+b), \theta_i)}{\int_0^1 g(\theta_i|\alpha,\beta) f(a|(a+b), \theta_i) d\theta} \quad (5).$$

ベータ分布$g(\theta|\alpha, \beta)$は以下の通りである。

$$g(\theta|\alpha, \beta) = \left(\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}\right)/Be(\alpha, \beta) \quad (6).$$

ここで、α, β はベータ分布の形状を決定するパラメータであり、$Be(\cdot)$はベータ関数を意味する。ベータ分布の期待値は$\alpha/(\alpha+\beta)$である。事後分布の期待値$\bar{\theta}$に対して$\alpha = bd/(b+d)$ と$\beta = ac/(a+c)$とすると以下の様に LS が導かれる。

$$\bar{\theta} = \frac{M}{M+n_j}\mu + \frac{n_j}{M+n_j}\left(\frac{a}{n_j}\right) = W1 \times \mu + W2 \times \left(\frac{a}{n_j}\right) \quad (7).$$

$$= W1 \times \left(\frac{bd}{b+d} / \left(\frac{ac}{a+c} + \frac{bd}{b+d}\right)\right) + W2 \times P(\text{獲得}|\text{餌場 A}) \quad (8).$$

$$= LS(\text{獲得}|\text{餌場 A}) \quad (9).$$

ここでMとμ , n_j は、それぞれ、$\alpha + \beta$と事前分布（ベータ分布）の期待値、腕 j の選択回数である。$W1$ と $W2$ は、それぞれ[0, 0.5]と[0.5, 1]の実数であり、$W1 + W2 = 1$である。同様に$LS(\text{獲得}|\text{餌場 B})$は$W'1 \times \mu + W'2 \times (c/(c+d))$となる。LS は異なる餌場に対して共通の事前分布が含まれる。LS は CP（図の候補）が餌獲得と喪失の全体的な情報（地）にどの程度引きずられるかを加重平均で決めている。

LS の評価値と事前は、事前分布を$g(\theta|2R\alpha, 2\bar{R}\beta)$、事後分布を$g(\theta|2\bar{R}a + 2R\alpha, 2Rb + 2\bar{R}\beta)$とするパラメータ R を用いることで歪めることができる。分布を左右に歪めるための R は $R + \bar{R} = 1$であり、[0, 1]の実数である。$R > \bar{R}$（$R < \bar{R}$）の場合、事前分布は右（左）方向に歪み、事後分布は左（右）方向に歪む傾向がある。$R = \bar{R}$の場合は元々の LS と同じである。この論文ではこの R を満足化基準と呼ぶ。この R は標準で 0.5 であり、損得 0（=報酬確率 0.5 の期待値）に対応する[大用 11]。

バンディット問題での LS の行動の極限を考える。例えば、餌場 A に着目し続ける場合、バイアス項は$\lim_{P(\text{餌場 A})\to 1} biasA \approx c$, $\lim_{P(\text{餌場 A})\to 1} biasB \approx d$となる。これより$W'1$と$W'2$が等しくなり、$\mu$が$d/(c+d)$となるため、$LS(\text{獲得}|\text{餌場 B}) \approx 0.5$となる。$LS(\text{獲得}|\text{餌場 A})$ はバイアスの重み$W1$が 0 に近づくため$P(\text{獲得}|\text{餌場 A})$となり、LS は正確な判断を行なえる。

4. シミュレーション

本研究では二本腕バンディット問題おいて、LS と UCB1–Tuned を比較する。

4.1 基本的な設定と指標

2 本腕バンディット問題は 2 個の餌場(選択肢)とそれに対応する確率（P_A, P_B）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)

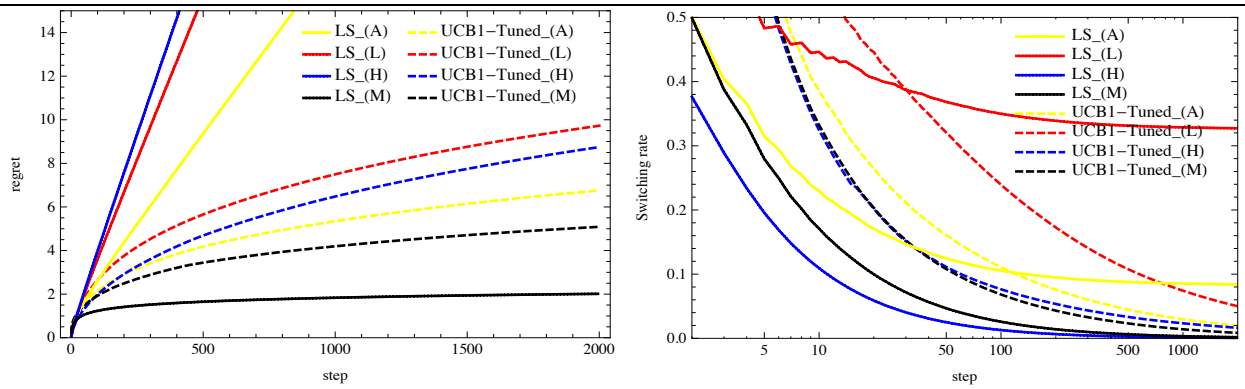

図 1：報酬確率の範囲毎の LS と UCB1-Tuned の後悔（図左）と切り替え率（図右）

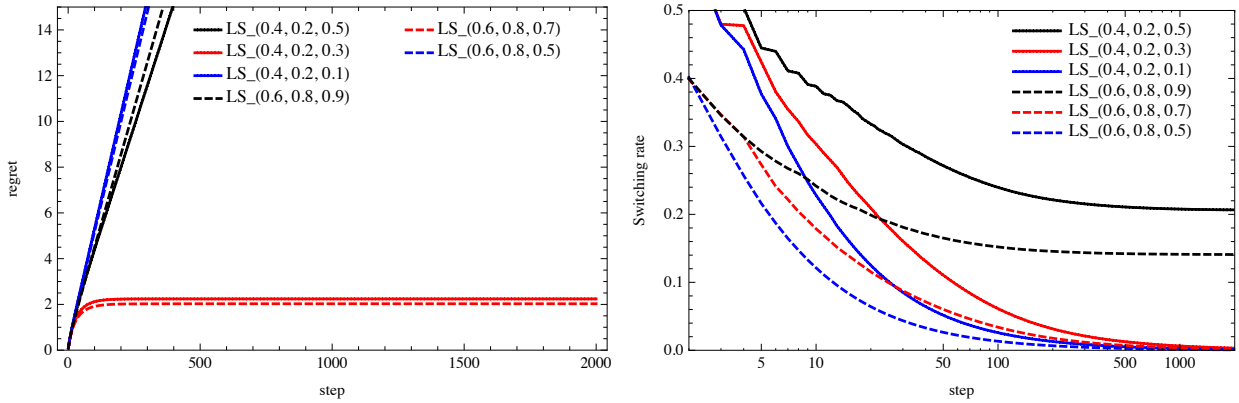

図 2：報酬確率設定の満足化基準毎の LS の後悔（図左）と切り替え率（図右）。括弧内の数値は LS_(P_A, P_B, R)

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 回選択（step）につき餌場の選択は 1 度に制限されている。

指標は後悔と切り替え率を採用した。切り替え率は n step までに腕を変更した割合である。後悔は最適な腕の期待値から選択した腕の期待値の差を累積した値である。偶然性の影響を極力排除するために、これらの指標は 10 万回の平均である。

本論文でのシミュレーションは以下の二つの条件に従った。1. 選択肢の価値が等しい場合はランダムに選択する。2. 初期値としてa, b, c, d に 1 を代入する。

4.2 異なる報酬確率範囲のパフォーマンス

ここでは 4 種類の報酬確率の範囲毎に結果を示す。報酬確率の範囲は、範囲(A)：P_Aと P_B が[0, 1]の実数、(H)：P_AとP_Bが[0.5, 1]の実数、(M)：P_Aが[0.5, 1]の実数とP_Bが[0, 0.5]の実数、(L)：P_AとP_Bが[0.0, 0.5]の実数として、その範囲から一様乱数として生成した。この論文では上記の範囲を全範囲、高確率、単高確率、低確率環境とそれぞれ呼ぶ。各範囲において、step は 2000 までとした。**図 1** に LS と UCB1–Tuned の結果を示す。

全範囲と高確率、低確率環境では、LS は UCB1–Tuned に劣ってしまうが、単高確率においては UCB1–Tuned よりも高い性能を示した。単高確率よりも高確率と低確率の UCB1–Tuned の性能が低いのは、$P_A - P_B$の差が大きい問題において最適な腕を発見するまでに多くの step が必要であるためである。

高確率と単高確率環境では LS の切り替え率が低いが、低確率環境では切り替え率が極端に高くなるという特徴がある。つまり、LS は満足化基準よりも高いと判断した餌場に執着し、それよりも低ければ満足に至る餌場を発見するために探索する。

4.3 満足化基準の変化

前節の結果から LS の振る舞いが満足化基準 R に従って決定している様に見える。例えば、$P_A, P_B < R$ならば餌場を多く探索し、$P_A, P_B > R$ならば一方の餌場に執着し、$P_A < R < P_B$か$P_B < R < P_A$ならば最適な餌場を選択する。そこで、LS のRの変化に伴う振る舞いを観測するために、報酬確率の設定(P_A, P_B)は(0.6, 0.8)と(0.4, 0.2)とした。(0.4, 0.2)の場合のRは 0.5 と 0.3, 0.1 とした。(0.6, 0.8)の場合のRは 0.9 と 0.7, 0.5 とした。LS の満足化基準を変化させた結果を図 2 に示す。

環境に対してRが高い LS は切り替え率が高い(R = 0.5 と R = 0.9 の黒い実線と破線)。またRが低い LS は切り替え率が最も低い(R = 0.1 と R = 0.7 の青い実線と破線)。中間的なRの LS の成績は元々の基準(R = 0.5)の成績よりも非常に良くなる(R = 0.3 と R = 0.7 の赤い実線と破線)。同様な設定を用いても UBC1-Tuned の切り替え率は、LS の様な傾向は見られず、特に(0.6, 0.8)では UBC1-Tuned の性能は逆に悪くなり、このパラメータの趣旨に反するため、その結果は図 2 に掲載していない。

4.4 中間的な基準の扱いやすさ

前節では問題に対して中間的な満足化基準であれば、LS の性能が向上する事が分かった。そこで、ここでは、中間的な基準に設定された LS の性能を測るために、報酬確率は 4.1 節の全範囲環境とした。Rは $\min(P_A, P_B)+|P_A - P_B|\times\gamma$とした。$\gamma$ の値は[0, 1]の実数であり、シミュレーションでは0.1, 02, …, 0.9 の 9 種類を調査した。LS と設定毎に基準を変更した LS、UCB1–Tuned の結果を図 3 に示す。

全範囲環境において、基準を変化させた LS は UCB1–Tuned よりも非常に高い成績を示した。また、γ =0.1, …, 0.9の 10 万ステップ後の後悔は、それぞれ 9.9, 6.2, 4.7, 4.1 ,3.8, 3.7, 3.9, 4.3, 5.7であった。最も成績が良いのはγ = 0.6 であり、ある程度の報酬確率の間であれば、それほど差がないことが分かる。即ち、パラメータ推定は直感的かつ簡単である。

5. 議論

以上のシミュレーション結果から LS は UCB1–Tuned に対して成績が劣る部分があるが、直感的なパラメータ R を用いることで、二本腕バンディット問題の全範囲で高い成績を示せること

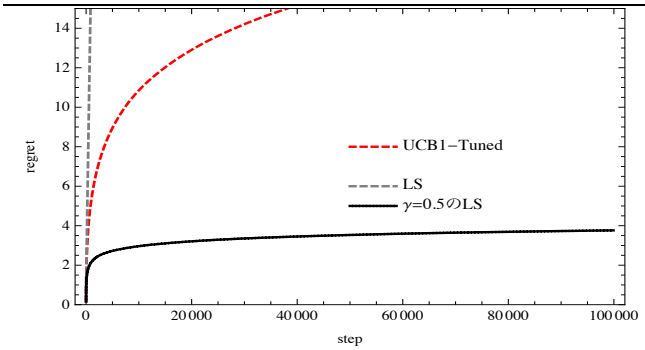

図 3： LS と UCB1-Tuned の後悔

が分かる。バンディット問題の多くのモデルは確率的に探索と知識利用を分けている。例えば、ϵの確率で腕をランダムに選択（探索）し、1-ϵの確率で最も価値の高いと思われる腕を選択（知識利用）するϵ-greedy 法や各腕の価値を比率にしてバランシングを行なう SoftMax 法などがある。LS はこれらのモデルよりも性能が良いことは確かめられており、上記のモデルとは異なり単に greedy 法で運用されている。LS は餌獲得情報（a）から餌場 A の価値を上げて（または図と見立てる）、餌場 B の価値（または地）を下げ、餌喪失情報（d）から餌場 B の価値を下げて餌場 A の価値を上げる。この様な相対評価の中には、全体傾向として意味を持つ事前分布が二つの餌場の価値の間を仲立ちする。LS が正確に餌場の価値を推定する場合には、全体傾向の重みは$W1 \approx 0$（$W'1 \approx 0$）となり、個の餌場の評価（条件付き確率）の重みは$W2 \approx 1$（$W'2 \approx 1$）となり、個の餌場の評価が全体傾向（地）の影響から分離される（図地分離）。LS が餌場を中立的（餌獲得に対して無相関）と判断する場合には、$W1 + W2 = 0.5$となり個の餌場が全体傾向と同一化される（W2 は 0.5 未満にならないため、LS の全ての振る舞いは、P(餌場A) ≈ 1の場合は餌場 A を図とし、P(餌場 A) ≈ 0.5の場合には図が曖昧になり、P(餌場 A) ≈ 0においては餌場 B を図として、視覚における図と地に例えられる）。LS はその基準を歪める事で、各共変動情報の価値を歪めて認識することができる。基準が低（高）ければポジティブ情報（ネガティブ情報）をより価値の高いものと認識する。これにより、例えば、LS は低確率環境を単高確率環境と認識することで振る舞いが変化する。LS は事前分布により仲立ちされる相対評価で、知識利用と同時に行なわれる探索により効率的にジレンマを克服しようとする。

満足化基準に満たない環境では、満足に至る餌場を探すために高めの切り替え率を保ち、満足化基準を満たす餌場には執着する。この様な性質は、その基準を下げれば勝率が低い着手でも満足するため勝ちこだわらない囲碁 AI などの柔軟な AI 構築に役立つと考えられる。

6. 結語

本論文では経験ベイズ法の枠組みを用いて LS にベイズ的分析を施し、満足化基準を変更するパラメータ R の拡張を施した。LS がベータ分布の期待値である事と条件付き確率（図）が全体傾向（地）にどの程度引きずられるかを加重平均で表現可能なことを示した。そして、LS をバンディット問題に適用させた結果、相対評価と価値基準の変化という二つの簡単な方法で高成績を示せることが分かった。

心理学や脳科学、行動経済学の研究から、人間が相対評価を行なうこと[Daw 06; Tversky 74]や一定の基準で満足に至る選択肢を探す満足化基準[Simon 56]、選択肢の価値に大きな影響を与える参照点[Kahneman 79]等の存在は知られており、それらの性質から認知バイアスの説明が行なわれいてる。LS はこれらの人間特性をヒューリスティックスとして捉えていると考えられる。LS は基準が適切であれば速くかつ正確に餌場を選択できる。基準が高ければ損を回避するため探索を幅広く行なう様なリスク追及傾向があり、基準が低ければ損失を被るよりも利益を生まれる餌場に執着するリスク回避傾向があるが、自然環境内での動物を考えれば、これらは生き残る上では合理的な振る舞いと考えられる。行動経済学や認知心理学で発見されている認知バイアスは、相対評価と基準変化に起因するものである可能性がある。これらは対称性と相互排他性が大きく影響を与えていると考えられ、二つのバイアスを中心とした諸バイアスの解明も今後行なって行きたい。また、満足化基準 R を静的に決定したが、今後は、動的な R の導入を目指す。

参考文献

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

[Casella 09] Casella., G. An Introduction to Empirical Bayes Data Analysis The American Statistician, 39(2), 83-87, 2009.

[Daw 06] Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., Dolan, R. J., 2006. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095), 876–879, 2006.

[Geis 71] Geis, M. L., & Zwicky, A. M. On invited inferences. *Linguistic Inquiry*, 2(4), 561–566, 1971.

[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.

[服部 08] 服部雅史, 推論と判断の等確率性仮説:思考の対称性とその適応的意味,認知科学,15(3), 408-427, 2008.

[Kocsis 06] Kocsis, L., Szepesvári, C., 2006. Bandit Based Monte-Carlo Planning. In J. Fürnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), *Machine Learning: ECML 2006* (Vol. 4212, pp. 282–293).

[Kahneman 79] Kahneman, D., Tversky, A., Prospect theory: An analysis of decision under risk, Econometrica, 47, 263-291, 1979.

[西村 12] 西村友伸, 大用庫智, 高橋達二, 可変参照型緩対称性推論のモンテカルロ木探索での効果 *The 17th Game Programming Workshop*. 2012.

[大村 12] 大村英史, 柴山拓朗, 高橋達二, 澁谷智志, 古川聖, 岡ノ谷一夫. 人間の因果推論による認知バイアスに基づいたメロディ生成システム. 知能と情報, 24(5), 954-966, 2012.

[Oyo 12] Oyo, K., Takahashi, T., Loosely symmetric heuristics as the basis for biases and the empirical Bayes methods, *CogSci2012*, 2012.

[大用 11] 大用 庫智, 甲野 佑, 高橋 達二, 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, JSAI 2011, 1G1-2in, 2011.

[Sutton 98] Sutton, R. S., Barto, A. G., 1998. Reinforcement Learning: An Introduction. *MIT Press*, Cambridge, MA.Sidman, M. (1994). Equivalence relations and behavior: A research story. Boston, M.A.: Authors Cooperative.

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

[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.

[Takahashi 11a] Takahashi, T., Oyo, K., Shinohara, S., A Loosely Symmetric Model of Cognition, In: *LNCS Springer Proceedings of the 10th European Conference on Artificial Life (ECAL 2009)*, Springer, 5778, 234–241, 2011a.

[Takahashi 11b] Takahashi, T., Nakano, M., and Shinohara, S., Cognitive Symmetry: Illogical but Rational Biases, Symmetry, *Culture and Science,* 21, 1-3, 275–294, 2011b.

[Tversky 74] Tversky, A., Kahneman, D., Judgment under Uncertainty: Heuristics and Biases, Science, 185(4157), 124-1131, 1974.

[Uragami 11] Uragami, D., Takahashi, T., Alsubeheen, H., Sekiguchi, A., Matsuo, Y., The Efficacy of Symmetric Cognitive Biases in Robotic Motion Learning. *Proceedings of the 2011 IEEE International Conference on Mechatronics and Automation*, 410–415, 2011.