2N5-OS-03b-2

柔軟な意思決定機能のための認知特性の応用と検証

Applying cognitive properties for flexible decision making and the analysis

甲野 佑[*1] 高橋 達二[*2]
Yu Kohno Tatsuji Takahashi

[*1]東京電機大学大学院 先端科学技術研究科 [*2]東京電機大学 理工学部
Graduate School of Tokyo Denki University Tokyo Denki University

We have shown that efficient adaptation to uncertain environments can be realized by three irrational cognitive properties: satisficing, risk attitudes, and comparative valuation. These properties, the most major biases in our cognition, have been extensively studied in artificial intelligence, cognitive psychology, and behavioral economics in isolation. The three properties, combined together, form a kind of suspension in value judgment. The suspension enables efficient valuation of actions in reinforcement learning where agents need to effectively balance exploration (search for novel information) and exploitation (local optimization with old information) under uncertainty. This study proposes Extended LS (LSX), a more general and simpler version of the loosely symmetric model (LS) that implements the properties. LSX is simpler in the sense the three properties of LS can be analyzed into individual terms, while in LS the properties are fused into a single term. This enables separate examination of the properties. Only when all the three are combined, superior performance in reinforcement learning is realized.

1. はじめに

我々は人間の評価能力に習って，未知あるいは不確実な環境での適応的評価手法，評価関数を考案する事を目的としている．未知の環境においてある選択に対してより多くの報酬を獲得するためには，探索的試行を行い，何らかの評価手法を用いて取りうる選択肢の価値を評価しなければならない．正確に価値を評価しようとする場合，何度も探索的試行を行い，より多くの知識を獲得する必要がある．しかし，探索してばかりでは高い報酬獲得を疎外するため，多くの報酬を得るためには探索的試行はある程度に収めなければならず，そのバランシングが困難である．これを探索と知識利用のジレンマと呼び，意思決定における速さと正確さにはトレードオフの関係がある事を表している．従来より，ϵ-$greedy$，$softmax$，$UCB1$ 等の手法により，数学的，統計学的に探索と知識利用を上手く使い分けるような仕組みが考案されて来た．しかし人間や動物は複雑な統計学的背景を持たずとも未知な環境に対するトレードオフに対応する事ができる．逆に言えば数学的において非規範である人間の認知特性が，未知な環境における限定下の合理性を有する可能性がある．ここでいう人間の評価手法における認知特性とは，相対評価，信頼性考慮，満足化の 3 種を指す．

$$LS(E|A) = \frac{P(A,E) + S_p}{P(A,E) + S_p + P(C,\bar{E}) + S_n} \quad (1)$$

$$\text{Positive bias}: S_p = P(\bar{E})P(A|\bar{E})P(\bar{A}|\bar{E}) \quad (2)$$

$$\text{Negative bias}: S_n = P(E)P(A|E)P(\bar{A}|E) \quad (3)$$

本研究では人間の評価感覚と一致する数理モデルとして，篠原氏が確信度形成のモデルとして考案した Loosely Symmetric model(以下 LS，式 1)[篠原 07] に着目する．LS は任意の原因事象 A と結果事象 E の生起不生起によって記述され，従来は 2 要因間の因果帰納課題，2 つの選択肢に対する意思決定課題（2 本腕バンディット問題）に対して良い成績を持つ認知的な確率モデルだとされていた．その後 LS は高橋により，視覚における着目対象と周辺視野との類似性（地の不変性）が指摘されている [Takahashi 10]．我々は LS モデルを主観的な確率の評価関数として解釈し，複数の選択肢に対する一般化を行った ($NormalizedLS$)．また，評価基準値をパラメータ化したモデルを開発し，動的な学習をする事で成績が飛躍的に上昇する事を示した (LS-VR)[Kohno 12]．更に筆者と高橋は LS の評価性能が 3 つの認知特性との類似に起因すると述べている．筆者では，報酬となる事象の種類を生起不生起の二種のみでなく任意の数に一般化し，また．前述の 3 つの認知特性を分離した記述を可能とする $EXtended\ LS$(以下 LSX) を考案した．LSX は本研究では LSX と 3 つの認知特性との関係を述べ，更にその特性を任意に除外し，組み合わせる事で 3 つの性質が意思決定課題においてどのような影響を及ぼすか推定する．

連絡先: 東京電機大学 理工学部
〒 350-0394 埼玉県比企郡鳩山町大字石坂
E-mail: yu.kohno.02@gmail.com

2. 人間の評価手法における 3 つの特性

人間や動物は複雑な数学的，統計学的知識を持たなくても，速さと正確さのトレードオフに対応する能力を有している．それが未知の環境での行動選択を迫られた際に，適応的，経験的に獲得して行く能力なのか，高度な認知能力を有する生物に先天的に備わった能力なのかはここでは触れない．少なくとも生物が自然環境に対して進化的に獲得した能力である事は間違いない．そのような意思決定における特性は複数存在するが，後に示す LS との関連が深い 3 つの特性を挙げる．これらは Hattori[Hattori 07] や Tenenbaum[Tenenbaum 11] の主張にも関連があり，諸々の認知特性の中でも特に原始的な特性だと考えられる．

2.1 相対評価

人間は手段 A_1 を試行した際に，報酬 E が得られなかった際，その他の手段である A_2 に対する評価が上昇する．逆に，A_1 で報酬が得られた際，A_2 に対する評価を下げる傾向がある．このように，一つの手段に対する試行結果が関係の無い (正確には関係あるかどうか解らない) 他全ての手段の評価に影響する事は，規範的な論理学から導出されない．しかし，人間はよくこのような評価をしてしまう [Tversky 74].

このように選択可能な手段の間に相対的な関係を想定して評価する形式は相対評価と呼ばれ，ある手段が上手くいけば其れに執着し，上手くいかなければ他の手段を試すよう促す効果を生む．これは正に “報酬の最大化” と “探索” を毎時の個別的な試行からバランシングしているに等しい．

2.2 信頼性考慮

信頼性考慮とは，評価の期待値のみでなく，サンプル数による信頼性を評価値に考慮する事である [Kahneman 84]．相対評価の具体的な形式として，信頼性を考慮する性質が考えられる．人間は確率的に等しい期待値と観測される選択肢にに対して，サンプル数の相対的な比率で評価が異なる場合がある．また，サンプル数によって評価値の順位が逆転する事もある．サンプル数の大きさは，その選択肢の客観的に観測した期待値がどれだけ信用できるかを表している．統計的知識を持たなくても，相対的な比率を参照することで評価に異なりを与える事が出来る点で優れている．

2.3 規準充足化

人間は評価を連続値ではなく “良い” と “悪い” 等，緩く二値化する性質がある．二値に分別するには基準値 (Reference value) が必要となり，規準と個々の選択肢との間の相対評価によって評価値の二値化が行われる [Simon 56]．また，その規準そのものも全体の評価値の分布や経験等から形成される．この評価値を二値化する性質によって，“良い” を見つけるまで探索するという満足化の性質を有する．更に評価値の二値化と信頼性考慮を組み合わせる事で，“良い” 評価が多い場合と “悪い” 評価が多い場合によってリスク忌避とリスク追及という真逆の傾向にわかれるという反射効果が表される．この時，二値化の規準が反射効果の参照点となる．

3. 意思決定課題 -N 本腕バンディット問題-

本研究では工学的な有用性を示す指標として N 本腕バンディット問題を例に，何も情報の無い状態から，トレードオフを抱える課題，環境に対し主体的に情報を獲得して行く際の不確実な知識の扱い方や値付けを論じる．ここでの不確実な知識とは観測が不十分で，正しいか否か断定出来ない曖昧な知識を意味する．これは強化学習課題における初期において学習を促進するためにどのような方策や価値観数を用いるかの問題に対応する [Sutton 00]．N 本腕バンディット問題とは目的となる報酬を確率的に得る事の出来る幾つかの手段 (腕)A_i から最適な手段を探索し，得られる報酬 E を最大化させる事を目的とする問題である．表 1 はバンディット問題で扱われる変数に対する確率的な表現である．

表 1: 事象 A，E 間の完全結合分布

	E	$\bar{E}$
A_1	$P(A_1, E)$	$P(A_1, \bar{E})$
A_2	$P(A_2, E)$	$P(A_2, \bar{E})$
$\vdots$	$\vdots$	$\vdots$
A_n	$P(A_n, E)$	$P(A_n, \bar{E})$

生き物が効率的に生きるためには，度々このようなバンディット問題的な課題に直面する．この課題の難しさは探索と収穫のジレンマという単語で表される．高い報酬を得るためにはどこかで探索を辞めるべきである．しかし探索しなければ高い報酬を得る事はできない．N 本腕バンディット問題はこのような知識の獲得とその利用からなる普遍的な “早さ” と “正確さ” のトレードオフを端的に表す事が出来る．

3.1 選択収束状態

本論文では議論を簡略化するために，いずれかの選択肢の選択された割合がほぼ 100 %になる状態を “選択収束状態” と定義する ($P(A_i) \approx 1.0$)．言い換えると，ある選択肢に執着して他の選択肢を相対的に殆ど選択していない状態を意味する．その執着している選択肢が真に最も期待値の高い正解の選択肢である場合，期待損失の上昇が止まり，上限が決定する．期待値が最も高い訳では無い誤った選択肢に執着してしまっている場合，その状態から抜け出せなければ期待損失が上昇し続ける．

4. EXtended Loosely symmetric model

本研究では LS と三つの認知特性の関係を明らかにするため，一般化を施した LS(式 1) として LSX を定義する．LSX は LS-VR[Kohno 12] と同様に，複数選択肢への一般化と，規準価値 (Reference) の動的なパラメータ化が行われている．更に LSX は排中立を満たすため，ある種の確率モデルとしてより規範的な性質を有するモデルであると言える (式 8)．変数 R はある種の基準点であり，以下の漸化式により，選択した選択肢の報酬獲得の標本平均（サンプリングされたあたり確率）から漸進的に学習する．

$$A_H = \arg\max_{A_k} P(A_k), A_L = \arg\min_{A_k} P(A_k) \quad (4)$$

$$S_E = \frac{P(E|A_H)P(E|A_L)}{(P(E|A_H) + P(E|A_L)} \quad (5)$$

$$S_{\bar{E}} = \frac{P(\bar{E}|A_H)P(\bar{E}|A_L)}{P(\bar{E}|A_H) + P(\bar{E}|A_L)} \quad (6)$$

$$S_{sum} = S_E + S_{\bar{E}} \quad (7)$$

$$LSX(E|A_i) = \frac{(P(E|A_i) + 2RS_{sum} - S_E)}{(P(A_i) + S_{sum}} \quad (8)$$

$$LSX(E|A_i) + LSX(\bar{E}|A_i) = 1.0 \quad (9)$$

$$R_0 = 0.5 \quad (10)$$

$$R_{t+1} = \alpha\ R_t + (1 - \alpha)P(E|A_{chose}) \quad (11)$$

ここで更に信頼性考慮と関連する重みを式 12，規準充足化に関する項を式 13，相対評価に関する項を式 14 と定義する事により，LSX は式 15 として三つの特性に分離した式として整理される．

$$\text{RC wight}: \omega_i = S_{sum}/(P(A_i) + S_{sum}) \quad (12)$$

$$\text{RS 差分}: \sigma_i = R - P(E|A_i) \quad (13)$$

$$\text{RE 差分}: \eta_i = \frac{S_E}{S_{sum}} - P(E|A_i) \quad (14)$$

$$LSX(E|A_i) = P(E|A_i) + \omega_i(2\sigma_i - \eta_i) \quad (15)$$

4.1 RC weight

RC wight (式 12) とは信頼性考慮 (Reliability Consideration) に関係する重み係数であると解釈できる．この重みの役割は後ろの RS 差分項と RE 差分項の強さを，抽象選択肢の試行回数に応じて修飾する事で信頼性を評価値に考慮する事である．上述の式では，重みの値域は $0 < \omega_i < 1/2$ になり，着目選択肢 A_i の試行割合 $P(A_i)$ が増える程に減少する．選択収束状態である $P(A_H) \to 1.0$ のとき $\omega_H = 0$ になり，後ろの二項の影響がなくなる．同様に $P(A_L) \to 0.0$ のとき $\omega_L = 1/2$ になり，後ろの二項が最大になる．

4.2 RS 差分

RS 差分 (式 13) は規準充足化 (Reference Satisficing) に関係する項であると解釈できる．選択収束状態において満足化に寄与するための項．しかしこの項自体がやっているのは参照点 (Reference value) へと近似する反射効果であり，“中庸化”と呼ぶべき物である．即ち，$\omega_i = 1/2$ となる選択肢を 参照点 に近似する事によって，参照点を越える評価値を持つ選択肢が無い時は探索し，逆に参照点を越える評価値があればその選択肢に執着するという規準充足化の振る舞いを間接的に表現している．

$$\lim_{P(A_H)\to 1.0} LSX(E|A_H) = P(E|A_H) \quad (16)$$

$$\lim_{P(A_L)\to 0.0} LSX(E|A_L) = R \quad (17)$$

4.3 RE 差分

相対評価 (Relative Estimation) に関する RE 差分の項 (式 14) は抽象的な期待値 S_E/S_sum と任意の選択肢の観測報酬確率を差分する事で，相対評価的な性質を評価値に与えているのだと考えられる．係数が負であるため，この項は中庸化に抗う項であり，2 本腕のときは選択収束状態において RE 差分の値は 0 になるので基本的には不要な項だと考えられる．

しかし N 本腕時には．選択収束状態において殆ど選択されない選択率ほぼ 0% の選択肢は RC wight と RS 差分項の影響でリファレンス値に収束する．そこに RE 差分項は抽象期待値との差分だけ値を上昇させる．つまり選択収束状態においては選択率ほぼ 0 %の中でも最も高い選択肢が選択され易くなる．

5. 実験 1 -3 つ特性の相互作用-

LSX を通して三つの特性の相互関係を理解するため，三つの特性を司る変数の値をそれぞれ以下の様に固定した場合の組み合わせで N 本腕バンディット問題を用いたシミュレーションを行う．

$$\text{dummy RC wight} : \omega_{dummy} = 1/4 \quad (18)$$

$$\text{dummy RS 差分} : \sigma_{dummy} = 0.0 \quad (19)$$

$$\text{dummy RE 差分} : \eta_{dummy} = 0.0 \quad (20)$$

表 2: LSX から構築可能な 3 つの認知特性の組み合わせ

	RC wight	RS differenc	RE 差分
CP	dummy	dummy	dummy
LSX-S	dummy	○	dummy
LSX-E	dummy	dummy	○
LSX-CS	○	○	dummy
LSX-CE	○	dummy	○
LSX	○	○	○

ここで RC wight のみが dummy でない式を考えても RS 差分，RS 差分の二項の値が 0 である以上 CP と等価になるため省略する．シミュレーション毎に選択肢の真の報酬確率は毎回一様乱数から設定し直している．エージェントは事前情報を持たない状態から 1step 一回の選択を行う．1,000 step を一回のシミュレーションとして，それを 10,000 回行い，得られる結果である正解率 (真の確率が最も高い選択率を選択できた確率) を平均して算出した．本シミュレーションの目的はあくまでも三つの特性の理解のために行うのだが，成績的指標として UCB1 の優れた改良モデルである UCB1-tuned との比較も行った [Wang 05].

$$P(E|A_i) + \sqrt{\frac{\ln n}{n_i} min\left(1/4, V_i(n_i)\right)} \quad (21)$$

$$V_i(s) = \left(\frac{1}{s}\sum_{k=1}^{s} r_{k,i}\right) - P(E|A_i) + \sqrt{\frac{\ln n}{s}}$$

5.1 実験結果

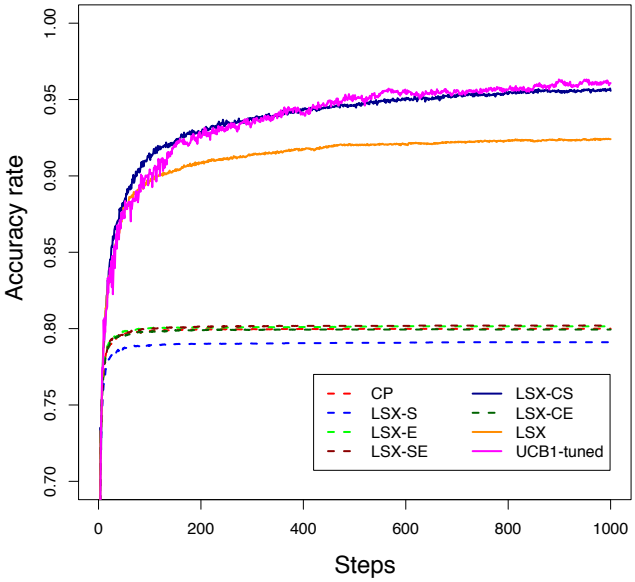

図 1: 正解率：選択肢数 2

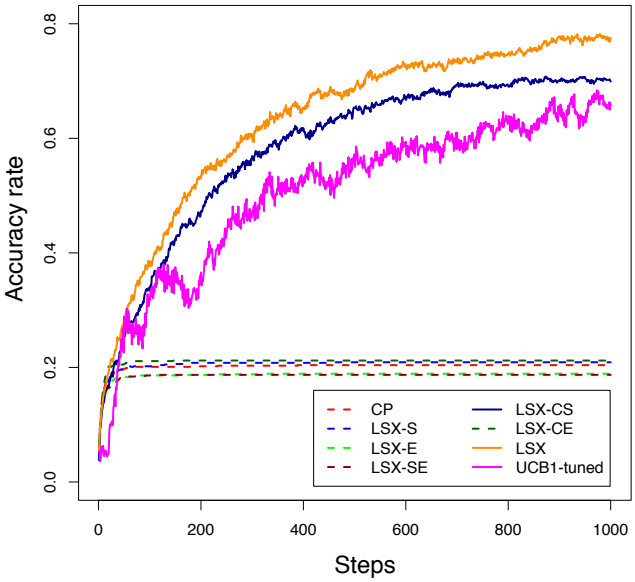

図 2: 正解率：選択肢数 20

本シミュレーションでは選択肢が 2 つの場合と，20 つの場合でそれぞれ行った．結果を図 1 に示す．これらの図は横軸が step の推移を表し，縦軸が真の報酬確率が最も高い選択肢（正解の選択肢）を選択できた割合をシミュレーション回数の平均によって示している．これらの結果から，少なくとも RC weight と RS 差分を併せ持たなければ，良い成績を得る事が出来ない事が解る．また，2 本腕バンディット問題では RE 差

分を持つ事で正解率の向上を疎外しているように見える．しかし，より選択肢が多い 20 本腕バンディット問題（図 2）では，RE 差分を持つ事でより高い成績を示す事が解る．この結果から，RE 差分は抽象化に抗う事によって，規準充足化の齎す探索の終了を抑制し，探索行動を誘発しているのではないかと考えられる．

6. 実験 2 -選択収束状態からの脱出-

前節のシミュレーション結果から得られた RE 差分が探索を誘発しているという仮説を検証するため，予め選択収束が起こっているような状況からの 20 本腕バンディット問題のシミュレーションを行った．実験 1 との違いは，既に 100,000 回選択をおえ，その際に観測された各選択肢の期待値が 95 %信頼区間に収まる最低の値になっている状態を引き継ぐ事にある．その際，真の正解確率は最も高い訳では無い選択肢（比正解の選択肢）の観測された期待値のみ，95 %信頼区間を越えないように，正解の選択肢に対して観測された期待値を上回るよう設定する．其れ以外の実験設定はは全て実験 1 と同様にした．

6.1 実験結果

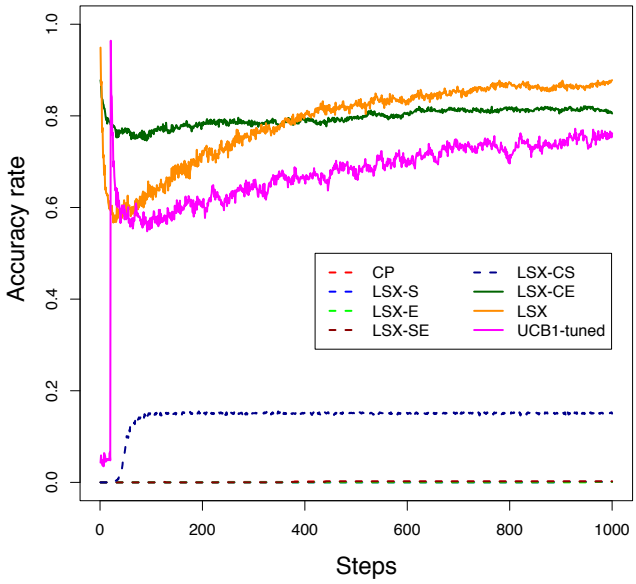

図 3: 正解率：誤った偏情報からの回復の時間推移

本シミュレーションでは選択肢数 20 の場合で行った．結果を図 3 に示す．これらの図は実験 1 と同様に，横軸が step の推移を表し，縦軸が真の報酬確率が最も高い選択肢（正解の選択肢）を選択できた割合をシミュレーション回数の平均によって示している．仮説通り，RE 差分を有する *LSX-CE* が高い成績を有している．しかし実験 2 でも実験 1 と同様に最も高い成績を有するのは三つの特性を併せ持つ *LSX* であった．

7. 総合考察

単純な反射効果のみなら RC weight と RS 差分のみで表現でき，2 本腕バンディット問題であるなら RE 差分は必要ない事がわかる． RE 差分は選択肢が多い場合に機能し，理想収束時の利益追求条件に合致するまで探索を続行させる性質を持つ．基本的には *LSX* はサンプル数（知識量）に応じて RC weight を変え，RS 差分を修飾する事で評価値をリファレンス値周辺に値を収束させる．それによって選択肢の評価値をドングリの背比べ状態にし，順序の逆転を発生させ易くする．そして，RE 差分によって選択収束状態と探索状態とのシフトを制御しているものだと考えられる．本研究では *LSX* というモデルを用い，三つの特性の関係を端的に示した．これは飽くまでも *LSX* が人間の特性を有し，かつ N 本腕バンディット問題で高い成績を有する事を示したに過ぎない．しかし，3 つの特性が合わさる事でより高いパフォーマンスを発揮する示唆でもあり，今後の認知的な工学研究に寄与する物である．

参考文献

[Wickelgren 77] W.A. Wickelgren, “Speed-accuracy tradeoff and information processing Dynamics,” *Acta Psychologica* 41, pp. 67–85, 1977.

[Tenenbaum 11] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, “How to Grow a Mind: Statistics, Structure, and Abstraction,” *Science*, vol. 331, no. 6022, pp. 1279–1285, 2011.

[Hattori 07] M. Hattori and M. Oaksford, “Adaptive noninterventional heuristics for covariation detection in causal induction: Model comparison and rational analysis,” *Cognitive Science*, 31, 5, pp. 765–814, 2007.

[篠原 07] 篠原修二, 田口亮, 桂田浩一, 新田恒雄 (2007), “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.

[Takahashi 10] T. Takahashi, M. Nakano and S. Shinohara, “Cognitive symmetry: Illogical but rational biases,” *Symmetry: Culture and Science*, 21, 1–3, pp. 275–294, 2010.

[Takahashi 11] T. Takahashi, K. Oyo and S. Shinohara, “A Loosely Symmetric Model of Cognition,” *Lecture Notes in Computer Science*, No. 5778, Springer, pp. 234–241, 2011.

[Kahneman 79] D. Kahneman and A. Tversky, “Prospect Theory: An Analysis of Decision under Risk,” *Econometrica*, 47(2), pp. 263–292, 1979.

[Kohno 12] Kohno, Y., Takahashi, T. (2012), “Loosely Symmetric Reasoning to Cope with The Speed-Accuracy Trade-off”, *SCIS-ISIS 2012*, Kobe Convention Center (Kobe Portopia Hotel), pp.1166–1171.

[Sutton 00] Sutton, R. S., Barto, A. G. (2000), “強化学習”, **森北出版**, (三上, 皆川 訳).

[Tversky 74] Tversky, A., Kahneman, D. (1974). “Judgment under uncertainty: Heuristics and biases”. *Science 185 (4157)*, 1124–1131.

[Kahneman 84] Kahneman, D.; Tversky, A. (1984). “Choices, values and frames”. *American Psychologist 39 (4)*, 341–350.

[Simon 56] Simon, H. A. (1956) “Rational choice and the structure of the environment”, *Psychological Review, 63*, 261–273.

[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.