

満足化とその基準の動的な更新による強化学習の促進

Improvement of reinforcement learning with satisficing and online update of its reference value

甲野 佑 *1
Yu Kohno

高橋 達二 *2
Tatsuji Takahashi

*1 東京電機大学大学院

Graduate School of Tokyo Denki University

*2 東京電機大学理工学部

School of Science and Technology, Tokyo Denki University

As the application range of reinforcement learning becomes broader, improvement of the learning algorithms in a way different from the past progresses is needed. The algorithms need to take into account both the cognitive, sensory, and motor limitations in agents and the structure of the environment, as discussed in the theory bounded rationality. We propose a cognitively inspired, effective satisficing algorithm for reinforcement learning that updates online the reference (aspiration level) of satisficing.

1. はじめに

事前知識が無い状態で目的を達成するためには試行錯誤して情報を収集する事が重要となる。環境に対する試行錯誤と得られる報酬から合目的的行動を獲得する枠組みである強化学習では、情報収集のための行動選択を“探索”と定義している。他方、得られた情報から最も目的達成に近づけると見積もられる行動を選択する事を“利益追求”と呼ぶ [Sutton 00]。合目的的行動の獲得に際する大きな問題の一つが、この探索行動と利益追求行動のバランスである。

本研究では、前述の問題に対して柔軟に合目的的行動を発見する人間の認知特性に習ったアルゴリズムの考案を目的としている。我々は人間の因果関係の強さの推定量と高い相関を持つ信念モデル Loosely Symmetric model (以下 LS) [篠原 07] に着目して、人間の意思決定における認知特性である満足化に関する拡張を行った EXtended LS (以下 LSX) を考案した [甲野 14]。同時に価値関数に LSX を用いた学習アルゴリズムが最も単純な強化学習課題の一種である N 本腕バンディット問題において良い成績を有する事を示した。しかしながら、LSX アルゴリズムは報酬の生起確率のみに対応しており、実数値域の報酬や価値には対応していなかった。そこで本研究では、より広い強化学習課題に応用する際の LSX の問題点を考察し、それを改善した Real scaLize Loosely Symmetric model (以下 RLLS) を考案した。また複雑な強化学習課題における RLLS アルゴリズムの振る舞いを通して LS 系モデルの根幹的な性質 (満足化) がどのような効果を生んでいるか考察した。

2. 強化学習とトレードオフ

基本的な強化学習アルゴリズムの一種である TD 学習 (Q 学習, Sarsa 等) は、獲得した報酬を Q 値と呼ばれる任意の状態においてある行動を取る事 (状態行動対) に対する価値として格納する。TD 学習では Q 値と選択方策を参照して次に取るべき行動を選択し、それが現時点での Q 値に照らして合理的であった場合は利益追求行動、そうでない場合を探索行動と定義される。最も目的に沿った行動系列を発見するためには、なるべく多くの報酬に関する情報を探索する必要がある。しかしながら探索する程に報酬を直接的に得る機会は失

われて行く。このように探索と利益追求は両立できないため、報酬獲得の速さ (Speed) と高い報酬を得るための Q 値の正確さ (Accuracy) にはトレードオフの関係がある (Speed and Accuracy Trade-off) [Wickelgren 77]。実際に選択される行動は方策アルゴリズムによって決定される。例えばシンプルな方策アルゴリズムである ϵ -greedy では、確率 ϵ でランダムな行動を行い、残りの $(1 - \epsilon)$ 確率で利益追求行動を行うという乱数を用いた選択を行う。このように、トレードオフ に対処して良い探索と利益追求のバランスを行えるか否かは行動選択の方策アルゴリズムが担う問題となる。そのシンプルさから ϵ -greedy は扱い易いが、 ϵ が一定である場合はいつまでも確率的に探索が発生する上に、その減衰はパラメータの経験や解析による設計が学習課題毎に個別に必要となる。より効率的な探索を目的とした Boltzman 分布を応用した softmax 方策等も存在するが、 ϵ が減衰する ϵ -greedy より更に複雑なパラメータ設計が必要となる。そこで我々は人間の柔軟な意思決定に習う事で、柔軟に探索と利益追求を配分し、かつ扱い易い方策アルゴリズムを作れないかと考えた。

3. 満足化方策と基準値

人間の意思決定方策の特徴として、必ずしも最適な行動を目指すわけではない点が挙げられる。大抵の人間はある行動系列がある基準を満たす成果を得られた時、その行動系列に執着してあまり探索をしなくなる。このような傾向は満足化と呼ばれ [Simon 56]、最適化とは区別される。厳密には満足化は強化学習の目的である報酬の最大化とは異なる。しかしトレードオフを考慮する場合、満足化には探索を止める条件を明確に規定できるという利点が存在する。だが満足化という枠組みには、基準値を超える行動系列を如何にして“効率的に”探索して発見するかという議論は含まれていない。また、成績に直結する基準値をどのように獲得するかについても規定していない。そこで我々は既存のモデルの中から、ある性質から定義される価値関数に対して利益追求するのみで満足化という目的を与えられる、信念の強さのモデルである LS に着目し、満足化方策の実装形式の一案として LSX という拡張モデルを考案した [甲野 14]。

3.1 LSX

EXtended Loosely Symmetric model (LSX) は客観的な価値を歪めて表現する価値関数である。その最も重要な評価の性

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@mail.dendai.ac.jp

質は、各選択肢の試行比率に応じて観測された情報から曖昧な評価に近づける事にある。価値関数 LSX の評価値は以下の式により、互いに独立な選択肢 ($a_i \in A$) と、その選択肢を試行した際に観測された目的事象 (報酬, $e \in \{e, \bar{e}\}$) の発生割合 (\bar{X}_{a_i}), またその試行回数 (n_{a_i}) によって定義される。

$$a_H = \arg \max_{a_k} (n_{a_k}), a_L = \arg \min_{a_k} (n_{a_k}) \quad (1)$$

$$V_e = \frac{n_{a_H} \bar{X}_{a_H} n_{a_L} \bar{X}_{a_L}}{n_{a_H} \bar{X}_{a_H} + n_{a_L} \bar{X}_{a_L}} \quad (2)$$

$$V_{\bar{e}} = \frac{n_{a_H} (1 - \bar{X}_{a_H}) n_{a_L} (1 - \bar{X}_{a_L})}{n_{a_H} (1 - \bar{X}_{a_H}) + n_{a_L} (1 - \bar{X}_{a_L})} \quad (3)$$

$$n_V = V_e + V_{\bar{e}} \quad (4)$$

$$X_V = \frac{V_e}{n_V} \quad (5)$$

$$\omega_{n_i} = \frac{n_i}{n_i + n_V} \quad (6)$$

$$LSX(e; a_i) = \omega_{n_i} \bar{X}_{a_i} + (1 - \omega_{n_i})(2R - X_V) \quad (7)$$

$$a_{Select} = \arg \max_{a_k} (LSX(e; a_k)) \quad (8)$$

行動選択は LSX の評価値 (式 7) において最大となる行動選択肢が選ばれる (式 8) ため、評価値がまったく等しい場合以外には乱数を用いず、エージェント側としては決定論的に行動が選択される。また、前述した通り、行動 a_i の全試行回数に対する試行割合 T_{a_i} が 1.0 に近づいた時、即ち試行全体において既にほぼ探索の行動をしておらず、ある行動選択肢に執着している際には LSX の評価値は客観的な評価と一致する (式 11)。逆に試行割合 T_{a_i} が 0.0 に近づいた時、即ち相対的にほとんど選択されていない場合は最も曖昧な評価値である基準値 R に収束する (12)。LSX のこのような性質のうち前者を鮮鋭化、後者を背景化と呼ぶ。

$$T_{a_i} = \frac{n_{a_i}}{\sum_{a_k \in A} n_{a_k}} \quad (9)$$

$$0.5 < \omega_{n_i} \leq 1.0 \quad (10)$$

$$\lim_{T_{a_H} \rightarrow 1.0} LSX(e; a_H) = X_{a_H} \quad (11)$$

$$\lim_{T_{a_L} \rightarrow 0.0} LSX(e; a_L) = R \quad (12)$$

LSX の評価値は以上の性質から、ある選択肢に執着すると他の選択肢が背景化して基準値 R に近づいていく。そのため執着した選択肢が基準値 R より低い場合、背景化した他の選択肢の方が評価値が高くなる。基準値 R を越える選択肢を発見しない限り、執着するほど評価値の逆転が発生するため、結果的に探索が促進され、基準値 R を越える選択肢を発見した時に探索を打ち切る事ができる。これが価値関数 LSX が最大となる選択肢を選択し続けるのみで、乱数を用いずに探索と利益追求を柔軟に配分できる理由である。また、基準値 R が環境の中にある選択肢が有する最も高い報酬獲得割合 P_{First} と、二番目に高い報酬獲得割合 P_{Second} の間であれば、非常に速く正確に最も良い選択肢を見つけ出す事が出来る [Kohno 14]。

$$P_{Second} < R < P_{First} \quad (13)$$

3.2 価値関数としての LSX の問題

LSX には強化学習に応用する際の大きな問題が二つ存在する。第一は良い基準値 R を如何にして獲得するかであり、第二の問題は客観的な価値である標本平均 \bar{X} を Q 値のような実数値の範囲にすると背景化等の性質が失われること等が挙げられる。本研究ではまず第二の問題について議論する。LSX の評価値は経験的に得られた客観的な価値 \bar{X} と基準値 R 、仮想的な価値 X_V との重み付け平均によって算出される [甲野 14]。性質が失われる理由は、この際に用いられる重み ω_{n_i} が、価値 \bar{X} が正負に股がった実数値を扱う際に重みとしての性質を破綻させるためである。具体的には V_e と $V_{\bar{e}}$ の分母に価値 \bar{X} が含まれる事が原因である。

$$n'_V = \frac{n_{a_H} n_{a_L}}{n_{a_H} + n_{a_L}} \quad (14)$$

$$w_{V_e} = \frac{n_{a_H} \bar{X}_{a_H}}{n_{a_H} \bar{X}_{a_H} + n_{a_L} \bar{X}_{a_L}} \quad (15)$$

$$w_{V_{\bar{e}}} = \frac{n_{a_H} (1 - \bar{X}_{a_H})}{n_{a_H} (1 - \bar{X}_{a_H}) + n_{a_L} (1 - \bar{X}_{a_L})} \quad (16)$$

$$V_e = n_V (w_{V_e} \bar{X}_{a_L} + (1 - w_{V_e}) \bar{X}_{a_H}) \quad (17)$$

$$V_{\bar{e}} = n_V (w_{V_{\bar{e}}} (1 - \bar{X}_{a_L}) + (1 - w_{V_{\bar{e}}}) (1 - \bar{X}_{a_H})) \quad (18)$$

3.3 RLLS

我々は前述の推定に基づき、 w_{V_e} 、 $w_{V_{\bar{e}}}$ を等しく w'_V に差し替える事で重み ω'_{n_i} から価値 \bar{X} を排除した。これにより価値 \bar{X} が正負にまたがった実数値の範囲を取った場合にも鮮鋭化 (式 26) と背景化 (式 27) の性質を保つことが出来るようになった。我々はこの評価式を Real scaLize Loosely Symmetric model (RLLS) と名付けた。

$$w'_V = \frac{n_{a_H}}{n_{a_H} + n_{a_L}} \quad (19)$$

$$V'_e = n_V (w'_V \bar{X}_{a_L} + (1 - w'_V) \bar{X}_{a_H}) \quad (20)$$

$$V'_{\bar{e}} = n_V (w'_V (1 - \bar{X}_{a_L}) + (1 - w'_V) (1 - \bar{X}_{a_H})) \quad (21)$$

$$X'_V = \frac{V'_e}{n'_V} \quad (22)$$

$$\omega'_{n_i} = \frac{n_i}{n_i + n'_V} \quad (23)$$

$$RLLS(e; a_i) = \omega'_{n_i} \bar{X}_{a_i} + (1 - \omega'_{n_i})(2R - X'_V) \quad (24)$$

$$0.5 < \omega'_{n_i} \leq 1.0 \quad (25)$$

$$\lim_{T_{a_H} \rightarrow 1.0} RLLS(e; a_H) = X_{a_H} \quad (26)$$

$$\lim_{T_{a_L} \rightarrow 0.0} RLLS(e; a_L) = R \quad (27)$$

4. 強化学習における RLLS 方策

RLLS により価値 \bar{X} を Q 値のような実数値に置き換える事は可能となった。しかしながら、TD 学習には試行割合に相当する概念が存在しないため、その導入が必要となる。そこで我々はある状態 s_i において a_j を試行した強さとして $\tau(s_i, a_j)$ という量を定義した (表 1)。

Q 値の更新は従来通り方策 on 型 TD 学習を用いるか方策 off 型 TD 学習をベースとするかによって更新法が異なる。強化学習アルゴリズムにおける RLLS は式 24 において価値 \bar{X}

表 1: 状態 s_i における Q 値と τ 値

	Q	τ
a_1	$Q(s_i, a_1)$	$\tau(s_i, a_1)$
a_2	$Q(s_i, a_2)$	$\tau(s_i, a_2)$
\vdots	\vdots	\vdots
a_n	$Q(s_i, a_n)$	$\tau(s_i, a_n)$

を Q 値に、試行の強さ n を τ 値に置き換える事で定義できる (式 32) . ここで基準値 R_i は状態 s_i 毎に個別の値を持つ .

$$a_H = \arg \max_{a_k} (\tau(s_i, a_k)), a_L = \arg \min_{a_k} (\tau(s_i, a_k)) \quad (28)$$

$$Q_u = \frac{\tau(s_i, a_H)Q(s_i, a_L) + \tau(s_i, a_L)Q(s_i, a_H)}{\tau(s_i, a_H) + \tau(s_i, a_L)} \quad (29)$$

$$\tau_u = \frac{\tau(s_i, a_H)\tau(s_i, a_L)}{\tau(s_i, a_H) + \tau(s_i, a_L)} \quad (30)$$

$$\omega_{ij} = \frac{\tau(s_i, a_j)}{\tau(s_i, a_j) + \tau_u} \quad (31)$$

$$RLLS(s_i, a_j) = \omega_{ij}Q(s_i, a_j) + (1 - \omega_{ij})(2R_i - Q_u) \quad (32)$$

強化学習における RLLS 方策も LSX と同様に RLLS 価値関数が最も高い行動を選択する (式 33) . この選択は前述の通り鮮鋭化と背景化の性質も有しているため、乱数を用いずに満足化方策として機能する .

$$a_{Select} = \arg \max_{a_k} (RLLS(s_i, a_k)) \quad (33)$$

$$T_{ij} = \frac{\tau(s_i, a_j)}{\sum_{a_k \in A_{s_i}} \tau(s_i, a_k)} \quad (34)$$

$$0.5 < \omega_{ij} \leq 1.0 \quad (35)$$

$$\lim_{T_{iH} \rightarrow 1.0} RLLS(s_i, a_H) = Q(s_i, a_H) \quad (36)$$

$$\lim_{T_{iL} \rightarrow 0.0} RLLS(s_i, a_L) = R \quad (37)$$

4.1 τ 値の定義と更新手法

本研究で導入した τ 値は現状態 s_i での行動 a_j を行った回数を $\tau_{current}(s_i, a_j)$ と、その後の状態行動系列を試行した強さを $\tau_{post}(s_i, a_j)$ の和によって決定される .

$$\tau(s_i, a_j) = \tau_{current}(s_i, a_j) + \tau_{post}(s_i, a_j) \quad (38)$$

$$\tau_{current}(s_t, a_t) = \tau_{current}(s_t, a_t) + 1 \quad (39)$$

$$\tau_{post}(s_t, a_t) = \tau_{post}(s_t, a_t) + \alpha(\gamma_\tau \tau(s_{t+1}, a_{up}) - \tau_{post}(s_t, a_t)) \quad (40)$$

また、 $\tau_{current}$ 値と τ_{post} の更新はそれぞれ、式 39, 40 によって行われる . 更新に用いる a_{up} は方策 on 型の TD 学習アルゴリズムであれば実際に方策に従い選択された行動 a_{t+1} を用い、方策 off 型であれば $\arg \max_{a_k} Q(s_{t+1}, a_k)$ となる . 試行の強さと呼ぶ理由は、パラメータ γ_τ ($0 \leq \gamma_\tau \leq 1$) によってその後の系列の試行回数を割り引いて扱うためである . 即ち試行割引率 γ_τ は選択肢 a_j を選んだ後の試行回数をどの程度考慮するかを意味している . パラメータ α は学習率であり Q 値の更新と同様に扱う .

5. 大車輪シミュレーション

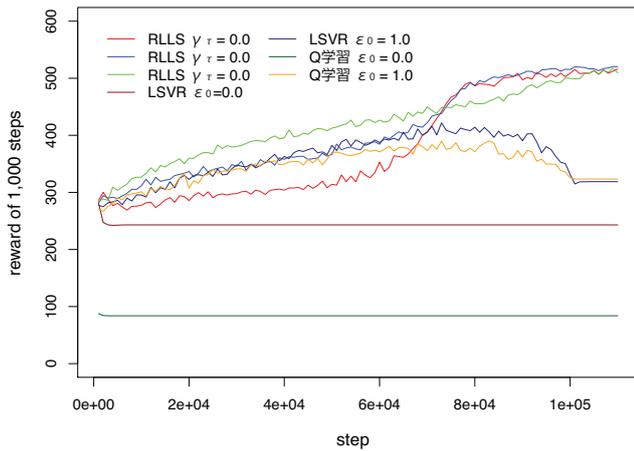
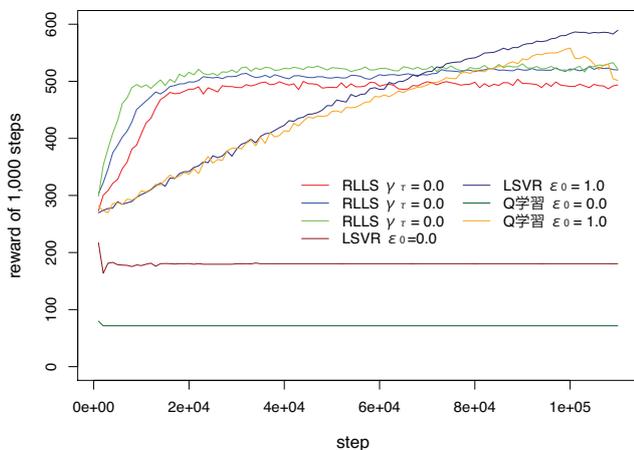
RLLS 方策を用いた学習アルゴリズムの性質を検証するため、本研究では複雑なダイナミクスを有する運動課題である大車輪強化学習課題を用いる . 大車輪課題を用いた理由は、RLLS とは異なる価値関数 LS の強化学習の応用案 (LS-Q アルゴリズム [浦上 13], LS-VR-Q アルゴリズム [高橋 13]) で検証されている課題である事も考慮している . LS-Q アルゴリズムやその発展系である LS-VR-Q アルゴリズムは、C-table と呼ばれる頻度テーブルを各状態毎に定義して、その上で LS 価値関数を計算して行動を選択していた . そのため鮮鋭化や背景化が Q 値に対して行われるわけではなく、構造が複雑化された事で本来の満足化方策とは異なる性質を持つ . また、LS-Q 系アルゴリズムは行動選択方策に ϵ -greedy を用いなければならず、そのみでは探索が充分に行えずに学習を促進する事が出来なかった . ゆえに満足化の性質をそのまま保ち、乱数を用いずに学習を行う事が出来る RLLS は元々の LS の性質を引き継いだまま一般化した学習アルゴリズムであると言える .

5.1 設定

学習課題に用いるシミュレーションの物理環境は過去の応用案と同様にした [浦上 13] . 腰のみを任意に動かす事が出来る大車輪ロボットが鉄棒に繋がれて静止した状態から学習は始まり、ロボットが可能な 3 つの行動、腰を“曲げる”、“延ばす”、“動かさない”、から逐次的に選択して、大車輪運動を獲得する事を目的とした課題である . エージェントが認識できる状態数は上半身の角度を 24、上半身と下半身のなす角度を 5、上半身の角速度を 7 に等分割した 840 種である . 一回の行動選択と状態の変化を 1 step として、それを 110,000 step 行い、そのシミュレーションを 50 回行った結果を平均した . また状態は 1,000 step 毎に初期状態に強制的に戻される . 報酬は初期状態であるロボットが垂直に下に向いている状態を角度 $\theta = 0$ として、step 毎に $r = \theta/\pi$ が与えられる . 比較に用いるエージェントは高橋の研究で良い成績を有していた LS-VR-Q アルゴリズム ($R = 0.8$ の場合) と、最も一般的な学習アルゴリズムとして Q 学習を用いる . 上述のアルゴリズムの行動選択には ϵ -greedy を用い、 ϵ は 1.0 から始まり、等間隔で徐々に減衰して 100,000 step の時点で 0.0 になるように設定する . 本研究で提案する RLLS アルゴリズムは ϵ -greedy を必要としないため、最初から $\epsilon = 0.0$ に設定する . その代わりに、全ての状態 s_i が持つ R_i 値は全て経験的な $R_i = 4.5$ に固定し、試行割引率には $\gamma_\tau = \{0.0, 0.5, 1.0\}$ を用いてそれぞれ比較した . また、乱数を用いない RLLS アルゴリズムとの比較のために、最初から $\epsilon = 0.0$ に設定した LS-VR-Q Learning と Q 学習とも比較する . 各アルゴリズムの割引率には $\gamma = 0.9$ を用いた .

5.2 結果及び考察

シミュレーションの結果として、縦軸は初期状態に戻されるまでの 1,000 step 毎の報酬の総和の推移を図 1 (学習率 $\alpha = 0.1$) と図 2 (学習率 $\alpha = 0.9$) に示す . 試行割引率 γ_τ はそれ以降に出現する行動系列を試行した強さを、どの程度その行動に対する試行の強さに反映するかを意味する影響度であると解釈できる . シミュレーション結果には ϵ -greedy を用いて学習初期の探索を促さなければ学習が行えない事が示されている . それに対して RLLS は $\epsilon = 0$ というランダムな探索を全く行わない学習でも学習が行っていた . また学習率 $\alpha = 0.1$ の場合、試行割引率 γ_τ が低い時には、ある段階での獲得報酬の急激な上昇が見られるが、 γ_τ が高くなる毎に、徐々に報酬が上昇していく傾向が見られる . 学習率 $\alpha = 0.9$ の場合にも試行割引率 γ_τ に対する基本的傾向は変わっていないが、全体的な学習速

図 1: 獲得報酬の推移: 学習率 $\alpha = 0.1$ の場合図 2: 獲得報酬の推移: 学習率 $\alpha = 0.9$ の場合

度が圧倒的に早まっている。しかし最終的には LS-VR に劣っている。これは決め打ちした RLLS の基準値 R がさほど良い基準でなかったためであると考えられる。しかしながら大車輪課題は多次元の物理量を離散化して状態認識しているのでも、そもそも学習率が高い方が行動を学習し易い。他方、RLLS は学習率 α や ϵ -greedy によるランダム探索等のパラメータに依存せず、決めうちの基準値 R でも一定以上の学習を行える点で、従来の C-table を導入して LS を実装したアルゴリズムより汎用性が高いと解釈できる。

6. 結論

本研究では、柔軟に探索と利益追求のバランスを行う意思決定における人間の特性を強化学習に応用する事を目的としていた。人間の意思決定傾向、満足化を表現できる価値関数である LS 及びその発展モデルである LSX は、共変動情報を背景としているために、実数スケールの価値関数への応用が出来ていなかった。我々は LSX の式に対する考察を基に、満足化傾向を直接的に残した実数スケールへの拡張モデルである RLLS を考案した。基準値を如何にして自然に獲得するかという、満足化の方策としての中核的な問題を残したままではあるが、ある程度正しい基準値を付与する事が出来れば、強化学習全般でも満足化方策が有効である事を示す事が出来た。また、強化学習における試行の強さの伝搬に関する手法を考案

し、それが学習の速さに関する事も明らかになった。満足化における基準値はある種のエネルギーコストに対する目標値と見なす事が出来る。即ち、満足化が実装可能になる事により、より動物的なエージェントとして強化学習エージェントを自身の消費カロリーに釣り合う行動を見つけて、生き続ける事を目的としたエージェントとして定義する事が可能になったと言える。更に満足しているという均衡した状態から環境を学習し、目標値(基準値)を動的に獲得できるアルゴリズムが開発されれば、より高度な知的活動エージェントとしての強化学習エージェントの発展を望む事が出来ると考えられる。

謝辞

本研究は特別研究員奨励費 26・10453 の助成を受けたものです。

参考文献

- [Kahneman 79] Kahneman D. and Tversky, A.: Prospect Theory: An Analysis of Decision under Risk, *Econometrica*, 47(2), 263–292 (1979).
- [Kahneman 84] Kahneman, D., Tversky, A.: Choices, values and frames, *American Psychologist*, 39(4), 341–350 (1984).
- [Kohno 12] Kohno, Y. and Takahashi, T.: Loosely Symmetric Reasoning to Cope with The Speed-Accuracy Trade-off, *SCIS-ISIS 2012*, 1166–1171 (2012).
- [甲野 14] 甲野佑, 高橋達二: 柔軟な意思決定機能のための認知特性の応用と検証, JSAI 2014(2014 年度人工知能学会全国大会 (第 29 回)) 予稿集, 2N5-OS-03b-2 (2014).
- [Kohno 14] Kohno, Y. and Takahashi, T.: A Satisficing Strategy with Variable Reference in the Multi-armed Bandit Problems, *ICNAAM 2014*, (2014).
- [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).
- [Simon 56] Simon, H.A.: Rational choice and the structure of the environment, *Psychological Review*, 63, 261–273 (1956).
- [Sutton 00] Sutton, R.S. and Barto, A.G.: 強化学習, 森北出版, (三上, 皆川 訳) (2000).
- [Takahashi 11] Takahashi, T., Oyo, K. and Shinohara, S.: A Loosely Symmetric Model of Cognition, *Lecture Notes in Computer Science*, 5778, Springer, 234–241 (2011).
- [高橋 13] 高橋優太, 甲野佑, 高橋達二: 認知的な強化学習モデルに対する基準学習の応用と考察, JSAI 2013(2013 年度人工知能学会全国大会 (第 28 回)) 予稿集, 1L3-OS-24a-4in (2013).
- [浦上 13] 浦上大輔, “対称性推論と運動学習の分節化: LS モデルを応用した Q 学習による大車輪ロボットの実現,” JSAI 2013(2013 年度人工知能学会全国大会 (第 27 回)) 予稿集, 1L3-OS-24a-5, (2013).
- [Wickelgren 77] Wickelgren, W.A.: Speed-accuracy trade-off and information processing Dynamics, *Acta Psychologica*, 41, 67–85 (1977).