2L1-4in

限定合理性に触発された強化学習法によるロボット運動学習

Robot motion control inspired by boundedly rational heuristics

水戸 亜友美 *1 Ayumi Mito　牛田 有哉 *1 Yuya Ushida　朝倉 勇護 *1 Yugo Asakura　甲野 佑 *2 Yu kohno　横須賀 聡 *1 Satoshi Yokosuka　浦上 大輔 *3 Daisuke Uragami

高橋 達二 *1 Tatsuji Takahashi

*1東京電機大学理工学部 School of Science and Technology, Tokyo Denki University

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

*3日本大学生産工学部 School of Industrial Technology, Nihon University

The rationality of an agent is bounded in various ways, because of imperfect observation, finite computational and memory resources, and incomplete control exerted in the environment. Under the bounded rationality, instead of optimize, one would satisfice with a certain level of aspiration. We model satisficing, the heursitics central to bounded rationality, in the framework of reinforcement learning and test the performance with the task of robot motion learning, which is a coarse grained dynamical system control problem.

1. はじめに

不確実性下での人間の選択傾向に触発された意思決定アルゴリズムが，複雑なダイナミクスを持つロボットの運動制御の強化学習課題において良い成績を示している [Uragami 11]．人間には完全に探索されていない環境下で意思決定を行う際に，ある基準を満たす選択肢が存在すれば，その選択肢に執着する満足化という傾向がある．これは人間が持つ限定合理性によって引き起こされるとされる [Simon 56]．この傾向を有するとされる Loosely Symmetric model （LS）を通じて，強化学習に対して LS-Q Learning[浦上 13] やその発展系である LS-VR-Q Learning[高橋 13] という形に応用されている．しかし，LS-Q Learning 及び LS-VR-Q Learning は従来強化学習に使われる Q 値とは別に C-table と呼ばれる特殊な記憶構造を持ち，その複雑な構造から満足化がどのような形で学習に影響しているのかが曖昧になっていた．本研究では C-table 上で定義されるより単純に満足化を表す RS-Q Learning や RS-Q Learning と比較し易い形式を持ちながら LSVR-Q Learning と同等の性質と成績を持つ LSX-Q Learning の比較を通じて，C-table を用いた満足化方策の性質について整理した．

2. 強化学習

強化学習とは環境に対する試行錯誤によって目的を達成する行動系列を学習する機械学習の一種である [Sutton 00]．エージェントは環境の状態を観測でき，ある状態において，ある行動を行うことがどれだけ良いのかを評価する報酬の累積である行動価値関数を参照して，各状態への適切な行動を学習していく．エージェントは行動価値関数と，それをどのように扱って選択を行うかを決める方策を用いて，実際に取る行動を決定する．エージェントはより多くの報酬を獲得するため，過去の行動の中から価値関数上で良い行動を優先的に選ばなくてはならない（知識利用）．しかしながら良い行動を発見するためには，現在最適だと思われる行動以外を選択する探索による情報収集も必要になる．探索と知識利用は一度に両立できないため，両者の割合をどのように調整すればよいかというトレードオフの問題が強化学習には存在する．

連絡先: 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

3. 満足化方策

人間は意思決定において，選択肢の評価をある基準値に対して “満たす” と “満たさない” に離散化する傾向を持ち，複数の選択肢から基準を満たす選択肢を探す事という目的を持つ事で素早く選択を行うことができる．これは人間の持つ限定合理性から引き起こされ [Simon 56]，このような選択傾向を満足化方策と呼ぶ．満足化方策はその性質上，満たすべき基準値 R というパラメータを併せ持つ．基準とは人間において “基準を満たす=良い選択肢” と “基準を満たさない=悪い選択肢” を分ける値で，環境に対する振る舞いが変化する境界を意味する．本研究では実際に実装されている満足化方策である LS-Q-Learning [浦上 13] の価値関数部をより単純な満足化傾向を持つ Reference Satisfacing model（以下 RS）や LS の発展モデルであり任意の基準値に変更可能にした EXtended Loosely Symmetric model （以下 LSX）を用いて様々な比較を行う．

3.1 RS policy

RS は元々 Rigid symmetric model と呼ばれる人間の認知バイアスを考慮した二事象間の繋がりの強さを表す価値関数である [篠原 07]．人間は「p ならば q」が真であるとき「q ならば p」も真であると捉えてしまう傾向（対称性バイアス）と「p ならば q」が真であるとき，「q でないなら p でない」も真であると捉えてしまう傾向（相互排他性バイアス）を持ち，RS はこれらを完全に満たす．近年，RS を基準値 R に対して拡張する事で，価値関数として意思決定に用いる事で，満足化する選択方策を表す事が出来る事が明らかになった [高橋 15]．

3.2 LSX policy

RS を基に人間が推定する因果関係の強さと相関の高い Loosely Symmetric model [篠原 07] が考案されている．LS

は観測された共起頻度から原因事象から結果事象への関係の強さを表す信念の強さのモデルである．また，RS と同様に満足化方策としての性質を持つ．LS は基準値 R が 0.5 に固定され，変えることができなかったが，Loosely Symmetric model with Variable Reference（LS-VR）[Kohno 12] や LS-VR をより形式的にした EXtended Loosely Symmetric model（以下 LSX）[甲野 14] に拡張される事により，任意の値に変更できるようになった．

4. C-table-Q Learning

前述した RS や LSX は原因と結果の二事象間の共起頻度として定義されるため，そのままでは強化学習に応用できなかった．そこで浦上は C-table（表 1）と呼ばれる，自身の行動に対する評価を頻度的に保存した情報を記憶し，その上で RS や LSX を定義する手法を考案した [浦上 13]．このベースとなる概念的学習アルゴリズムを本研究では C-table-Q Learning と呼ぶ．ここで $A = \{a_1, a_2, \cdots, a_n\}$ は行動を意味する．また，Greedy は学習された状態行動対の収益を意味する Q 値が，その時点で最も高い行動を選択した事を意味し，同様に Non-Greedy は Q 値が最も高いわけではない行動を選択した事を意味する．即ち，例えばその時点の Q 値において最も高いと評価されている a_0 が選択された場合，頻度 g_0 が +1 される．

表 1: 頻度記憶 “C-table”

	Greedy	Non-greedy
a_0	g_0	f_0
a_1	g_1	f_1
$\vdots$	$\vdots$	$\vdots$
a_n	g_n	f_n

C-table 上で RS，LSX の値は以下の式で定義され，これらの評価値が最も高いものを選択するという学習形式を RS-Q Learning，LSX-Q Learning と呼ぶ．ここで R は基準値を意味し，g_{max} はそれまでに最も多く試行された行動，g_{min} は最も試行されていない行動に対する Greedy な頻度を意味する．同じく f_{max}，f_{min} は最も試行された，試行されていない行動の Non-greedy な頻度を意味する．

$$RS(Greedy|a_i) = (g_i + f_i)(\frac{g_i}{g_i + f_i} - R) \tag{1}$$

$$S_f = \frac{f_{max} f_{min}}{f_{max} + f_{min}} \tag{2}$$

$$S_g = \frac{g_{max} g_{min}}{g_{max} + g_{min}} \tag{3}$$

$$S_{max} = S_f + S_g \tag{4}$$

$$LSX(greedy|a_i) = \frac{g_i + 2S_{sum}R - S_g}{g_i + f_i + S_{sum}} \tag{5}$$

C-table-Q Learnig の最大の特徴は従来の強化学習で用いられる行動価値関数 Q 値を意思決定に直接使わない点である．Q 値である $Q(s_i, a_j)$ とは任意の状態 s_i にて行動 a_j を行う事を意味する状態行動対 (s_i, a_j) がその後に得るであろう収益を意味する．C-table-Q Learnig では Q 値は自身の選択がその時点での Q 値において Greedy な選択，即ち獲得情報内で合理的な選択であったかを評価するのみに使われる．言わば C-table-Q Learnig では C-table から意思決定するエージェントと，意思決定エージェントの選択を現在の Q 値によって評価する行動評価エージェントの階層構造が存在する事になる．意思決定エージェントは直接的な形で本来満足化が対象とする Q 値を参照できないため，LS-Q Learnig や LS-VR-Q Learnig がどのような指標を対象として満足化を行っているのかが不明確だった．また満足化方策においては，基準値を試行錯誤から動的に獲得できれば，より良い学習が自動的に行う事が出来る．よって C-table-Q Learnig における満足化の定義が明らかになれば，それに伴って基準値の動的な更新法の構築に寄与できると考えられる．

4.1 C-table-Q Learning と N 本腕バンディット問題

本研究では，C-table-Q Learnig 上での満足化の性質を明らかにする事を目的としている．C-table-Q Learning において意思決定エージェントは実際の Q 値を観測できないため，C-table の情報のみを参照して行動を選択しているため，報酬を得てからその行動を評価する．我々は本研究において，この枠組みが N 本腕バンディット問題として捉えられると考えた．N 本腕バンディット問題とは，スロットマシンに例えられる，試行する事で確率的に報酬が得られる未知の選択肢が複数あるとき，選択と試行を通して報酬の最大化を目指す最も単純な強化学習課題の一種である [Sutton 00]．C-table-Q Learnig では，行動評価エージェントによる Greedy であったという評価がスロットマシンから得られる報酬に相当する．そのためどの行動が Greedy であるかは行動評価エージェントが保持する Q 値の更新による大小関係の変化に伴って変化する．すなわち C-table-Q Learnig 上において，意思決定エージェントは行動評価エージェントが提供するスロットマシン環境に対して非定常 N 本腕バンディット問題を状態毎に行っているに等しい．だとすれば Q 値の大小関係がなるべく変化しない事が行動 a_i の価値に相当する．その価値に対する満足化の基準値 R とは，Q 値の大小関係の変化に対して “安定して” Greedy であり続ける割合を意味する事になる．しかし意思決定エージェントの選択する行動に変化があれば，Q 値の大小関係にも影響が出るため，Q 値の大小関係の変化は課題環境のみに依存せず，意思決定エージェントと行動評価エージェントと課題環境の相互作用が重要となる．そこで本研究では以上の仮説を検証するために，Q 値の大小関係の変化の割合とエージェントの選択の変化に着目したシミュレーションを行った．

5. 大車輪運動課題シミュレーション

本研究では大車輪運動課題というロボットの運動制御課題を用いてシミュレーションを行う．大車輪運動課題は鉄棒に接続されたロボットに鉄棒を一回転させる大車輪を学習させる課題で，強化学習で扱われる課題の中でも複雑な物理ダイナミクスを有する問題として知られている [Sutton 00]．本研究の大車輪運動課題は過去に行われた LS 系の C-table-Q Learning と同様に，ロボットは腰のアクチュエータのみ稼働させることができる [浦上 13, Uragami 14, 高橋 13] ため，エージェントが取りうる行動は，“a_0：関節を曲げる”，“a_1：関節を伸ばす”，“a_2：動かない” の 3 種になる．それぞれの行動について価値関数と比較して “Greedy”，“Non-greedy” かによって C-table(表 1) を更新し，その値から各ポリシーを用いて行動を選択する．状態は “エージェントの位置”，“下半身の角度”，“エージェントの角速度” で決まり，報酬は垂直に静止した角

度を $\theta = 0$ としてロボット先端の成す角度 θ から $r = \theta/\pi$ によって 0〜1 の範囲で与えられる．状態には上半身の角度は 2π，下半身の角度は $0〜5\pi/6$，上半身の角速度は $-3\pi[rad/s]$ 〜$3\pi[rad/s]$ の状態量を用い，それぞれ任意の数に等分割に離散化されて認識される．満足化と C-table 上での評価値の関係を明らかにするため，学習エージェントには RS-Q Learning と LSX-Q Learning を用いて比較した．

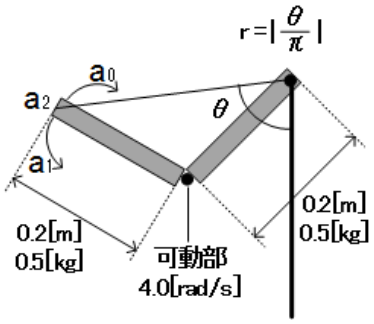

図 1: ロボットの構造

5.1 設定

エージェントの初期状態は，鉄棒に大して垂直にまっすぐな姿勢のまま，ぶら下がって静止した状態である．エージェントは各方策に基づいた行動選択を行い，それを 1 step として 1,000 step 毎に初期状態に戻る．シミュレーションでは 80,000 step 行い，1,000 step 毎の獲得報酬の合計を記録していき，その時間発展を学習曲線とした．行動の選択には ϵ の確率でランダムな選択を行う ϵ-greedy を用いた．学習開始時の $\epsilon = 1.0$ とし，1,000 step 毎に 0.05 減らすことで 20,000 step にはランダムな行動を完全にしなくなる．本シミュレーションではこの ϵ が減衰していく 20,000 step までを学習フェイズと呼ぶ．

状態は上半身の角度の状態数を 6，下半身の角度の状態数を 3，上半身の角速度の状態数を 3 に等分割し，状態数を 54 に離散化してエージェントに認識させた．それぞれの方策において学習率 α は $\alpha = 0.9$ [浦上 13] とし，基準値 R は RS-Q Learning は 0.1〜0.9 から 0.1 ずつ，LSX-Q Learning は 0.2, 0.5, 0.8 についてシミュレーションを行い比較した．

5.2 結果

図 2，図 3 は RS-Q Learning，LSX-Q Learning の学習曲線である．横軸は [step/1,000]，縦軸は 1,000 step 毎の獲得報酬の合計を表している．RS-Q Learning と LSX-Q Learning を比較すると，RS-Q Learning は基準値 R が 0.6〜0.7 以上において一度上がったところから徐々に落ちていく様子がみられる．しかしながら LSX-Q Learning は基準値 $R = 0.8$ の時でも高い獲得報酬を保っていた．

また前述した C-table-Q Leraning の性質を検証するため．更新によって Q 値の大小関係が変化しない割合を安定率，エージェントが選択する行動が前回と変更される割合であるスイッチ率として，獲得報酬が或る段階から落ち始めた 基準値 $R = 0.8$，そうでない基準値 $R = 0.5$ における RS-Q Learning，LSX-Q Learning の安定率とスイッチ率を比較した．安定率，スイッチ率はほとんどの状態において基準値 $R = 0.5, 0.8$ 共に RS-Q Learning と LSX-Q Learning の間に大きな差は認められなかった．しかし基準値 R が 0.8 の場合は RS-Q Learning と LSX-Q Learning の状態 24 において大きな差がみられた．

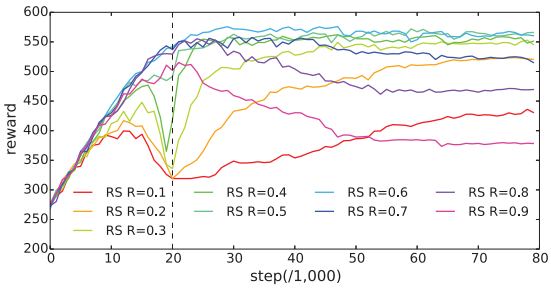

図 2: 基準ごとの RS の学習曲線

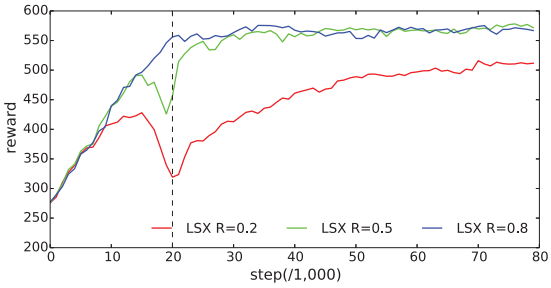

図 3: 基準ごとの LSX の学習曲線

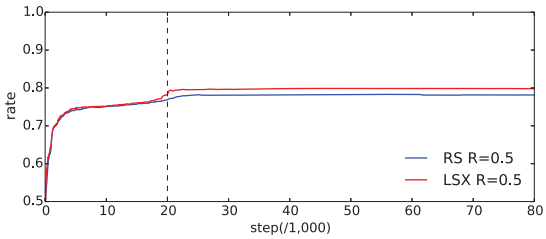

図 4: 基準値 R が 0.5 における RS と LSX の安定率

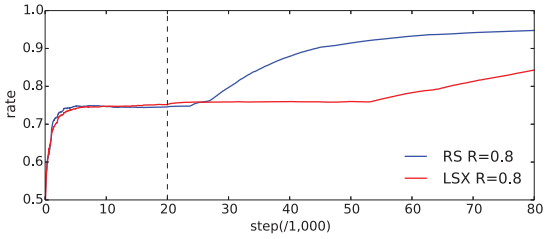

図 5: 基準値 R が 0.8 における RS と LSX の安定率

5.3 考察

高い基準値を伴う RS-Q Learning においてのみ見られた，一度は高い報酬を得られたのにも関わらず，その後下がっていくのは学習過程として不自然であると考えられる．LSX-Q Learning ではその過程はみられなかった．基準値 $R = 0.8$ における安定率の推移である図 5 を見ると，学習フェイズが終

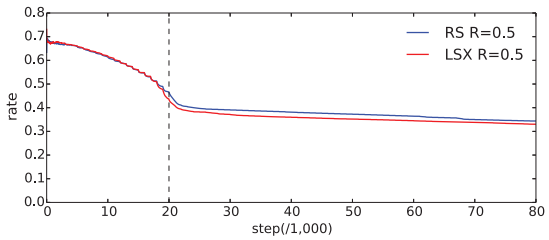

図 6: 基準値 R が 0.5 における RS と LSX のスイッチ率

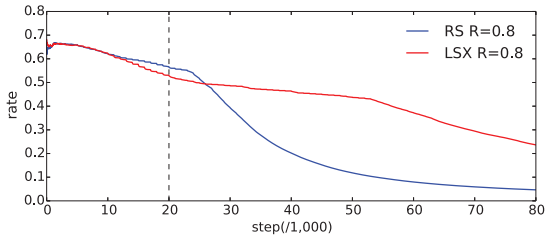

図 7: 基準値 R が 0.8 における RS と LSX のスイッチ率

了してしばらくしてから安定率の急激な上昇がみられた．仮説通りなら RS-Q Learning において基準値 $R = 0.8$ とは，かなり高い水準で Q 値の大小関係の変化が起こらない安定している選択肢を見つけない限り，探索を続けてしまう．しかし連続量である状態が粗く離散化されている本課題では，安定率を求め過ぎる事が必ずしも正しい行動の獲得に直結するとは限らない．更に言えば，強化学習においては一つの状態において行動選択を誤るだけで，行動系列全体に乱れが生じる．そのため RS-Q Learning では安定率を求め過ぎるがあまり，何らかの理由で過度な探索をしなくて良い状態でも探索を行い過ぎてしまい，負のループに陥っているのではないかと考えら得る．これが基準値 $R = 0.5$ では獲得報酬の現象がみられず，基準値 R が高い時にそれが見られた理由である．しかし，基準値が低い場合では RS-Q Learning の学習はうまく促進されていないため，大車輪運動課題において RS-Q Learning の基準値は高過ぎても低過ぎても適さない事がわかった．他方，LSX はその形式から N 本腕バンディット問題において非定常課題に強く，また定常課題でもある程度効率的に探索を抑えることができるとされている [甲野 14]．本研究のシミュレーション結果でも RS-Q Learning との比較において，基準値が高い場合でも成績が落ちず，また低い場合でも成績が優っている事から，強化学習においても LSX-Q Learning は基準値 R への依存度が少ない事がわかる．厳密に最適な基準値 R を求める手法はまだ確立されていない．そのため，ある程度いい加減な基準値でも良い成績を示す LSX-Q Learning の性質は満足化方策においても重要だと考えられる．

6. 結論

本研究のシミュレーション結果から，目的であった C-table-Q Learning における基準値 R と成績の関係についてある程度，明らかになった．すなわち Q 値の大小関係の変化の起き難さ (安定率) と基準値 R の関係に，行動選択に対する探索と知識利用の配分が依存するという点である．また，一度良い成績を得たにもかかわらず探索し過ぎると，安定率が下がってしまい，探索の促進と，それに伴う安定率の更なる低下を招いてしまい，悪い意味で安定した行動系列に陥るまで負のループを続けてしまう事がわかった．それに対して LSX-Q Learning は行動選択が基準値に依存しすぎないため，強化学習においても厳密に基準を定めなくてもある程度良い行動系列を獲得できる事がわかった．しかしながら満足化方策において良い基準を動的に獲得できる事が望ましい事には変わらない．本研究において C-table-Q Learning 上での満足化方策における基準値と安定率の関係が明らかになった事は，動的な基準値の更新法の構築に大きな寄与となったのではないかと考えられる．

参考文献

[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 28 回)) 予稿集, 2N5-OS-03b-2. (2014).

[Simon 56] Simon, H.A.: Rational choise and the structure of the enbironment, *Psychological Review*, 63, 261–273 (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, 22(1), 58–68 (2007).

[Sutton 00] Sutton, R.S., Barto, A.G.,(三上貞芳　皆川邪章 共訳): 強化学習, 森北出版 (2000).

[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 27 回)) 予稿集, 1L3-OS-24a-4in (2013).

[高橋 15] 高橋達二, 大用庫智, 甲野佑, 横須賀聡, 不確実性の下での満足化を通じた最適化, JSAI 2015 (2015 年度人工知能学会全国大会 (第 29 回)) 予稿集, 2D1-OS-12a-4in (2015).

[Uragami 11] Uragami, D., Takahashi, T., Alsubeheen., H., Sekiguchi, A., and Matsuo, Y.: The efficacy of symmetric cognitive biases in robotic motion learning, *Proceeding of the 2011 IEEE International Conference on Mechatronics and Automation*, August 7-10, Beijing, China, 410-415 (2011).

[浦上 13] 浦上大輔, 高橋達二, アルスビヒーン・ヒシャム, アルアルワン・アリー, 関口 暁宜, 松尾 芳樹: 対称性推論と運動学習の分節化, 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).

[Uragami 14] Uragami, D., Takahashi, T., Matsuo, Y.: Cognitively inspired reinforcement learning architecture and its application to giant-swing motion control, *BioSystem*, 116, 1–9 (2014).