5M-04

強化学習と満足化による素早い行動系列の獲得

牛田　有哉 *1　　甲野　佑 *2　　高橋　達二 *2

*1東京電機大学大学院　　*2東京電機大学理工学部

1. はじめに

環境との相互作用から行動系列を学習する強化学習には，環境情報のサンプリングを経るがゆえに試行錯誤の時間が膨大になる問題がある．それに対し最も単純な強化学習タスクの一種，多本腕バンディット問題において人間のリスク態度と意思決定傾向のある側面である満足化を組み合わせた満足化価値関数 (reference satisficing, 以下 RS) が少ないサンプリング下で高い成績を示している．RS は “任意の達成度合い” を目標値 (満足化基準値，あるいは単に基準値) として与えた場合に特に速さを発揮し，この基準値を動的に得る手法も存在する．RS は強化学習課題全般にも拡張されているが，基準値を複雑な状態表現にどのように適応すべきかは整理されておらず，経験的に良いとされる数値を基準値として学習に用いることが多かった．本研究では報酬関数に基づいた一つの値を基準値として設定し，それを適切に複数の状態表現に変換する方法，大局基準値変換法 global reference convertion (GRC) を考案した．以下では経験的な基準値の設計法と比較した GRC の性質について大車輪の運動制御課題を通して議論する．

2. 満足化方策

満足化とは人間の持つ「ある基準値を超える価値の選択肢 (行動) が見つかるまで探索を続け，発見したら探索を止める」という意思決定傾向であり，最良の行動を追求し続ける最適化とは区別される．本研究では RS 価値関数 [高橋 16] を用いて選択ポリシーに満足化傾向を反映する．RS は基準値を境界にして探索とその打ち切りを自律的に切り替えることができ，柔軟に一度打ち切った探索を再開することもできる．

強化学習以前に多本腕バンディット問題で扱われていた RS 価値関数の評価値はその選択で得られる報酬のサンプル平均，そのサンプリング回数に基づいた平均の信頼度，そして基準値によって算出された．強化学習では前述したサンプル平均を任意の状態行動対 (状態 s_i で行動 a_j) からの未来の収益予測である Q 値，そしてその Q 値の信頼度として，任意の状態行動対とその後のサンプリング経験の蓄積である $\tau(s_i, a_j)$，そして各状態に対する基準値 $R(s_i)$ として定義し，以下のように RS 価値関数の評価値 $RS(s_i, a_j)$ を計算する．

$$RS(s_i, a_j) = \tau(s_i, a_j)\big(Q(s_i, a_j) - R(s_i)\big) \quad (1)$$

$$\tau(s_i, a_j) = \tau_{cnt}(s_i, a_j) + \tau_{post}(s_i, a_j) \quad (2)$$

$$\tau_{cnt}(s_t, a_t) \leftarrow \tau_{cnt}(s_t, a_t) + 1 \quad (3)$$

$$\tau_{post}(s_t, a_t) \leftarrow \tau_{post}(s_t, a_t) + \alpha_\tau\Big(\gamma_\tau \tau(s_{t+1}, a_{up}) - \tau_{post}(s_t, a_t)\Big) \quad (4)$$

Quicker action sequence acquisition by reinforcement learning with cognitive satisficing, Yuya Ushida: Graduate School of Tokyo Denki University, Tatsuji Takahashi and Yu Kohno: School of Science and Engineering, Tokyo Denki University.

このとき，γ_τ は未来信頼度割引率を表し，α_τ は信頼度学習率を表す．本研究では，この価値関数に対して greedy 法を用いた選択を行うアルゴリズムとして使用する．

2.1 従来の基準値 $R(s_i)$ の決定法と問題点

バンディット問題だけでなく，強化学習においても RS は理想的な基準値が与えられれば最良の結果が速く正確に得られることが示されている [牛田 16]．理想的な基準値とは潜在的に達成できる収益期待値を意味し，達成不可能な収益目標でさえなければ RS は速く正確にそれを達成する行動系列を獲得する．しかし強化学習において収益とは，「ある状態からスタートして得られる報酬の (割引率を込みにした) 合計」であり，その値は状態毎によってかなり異なる．従来の多くの場合で，強化学習の基準値 $R(s_i)$ は状態毎の収益見込みのスケールの差を無視して学習タスクの提示者が経験的に与えていた．これは状態毎の収益の見込みを計算する手法と，その源となるタスク全体の達成度を表す概念が存在しなかったためであり，状態全体にタスクの満足化度合いを反映して探索の促進・打ち切りの判断を一貫させることができない．そこで我々はタスク全体の満足度を表すため，達成度合いとして大局観測期待値 (global expectation : GE) と，それに対応する大局満足化基準値 (global reference : GR) と，それによって導かれる満足化度合いを各状態基準への変換手法を考案した．

3. 基準値決定手法：GRC

タスク全体に対する評価である大局観測期待値 GE は一時的試行回数 (TN: temporary number) と一時的期待値 (TE: temporary expectation) を用いて 1,000 回ごとに以下の式によって更新される．また一般的な強化学習においてはエージェントが感じる報酬関数はタスクの提示者によって提供されるため，本研究において大局基準値 GR は動的ではなく報酬関数の設計に基づいて与えられるものとする．ここでパラメータ γ_G は大局割引率を表し $0.0 \leqq \gamma_G \leqq 1.0$ をとる．

$$GE \leftarrow \frac{\frac{TE}{TN} + \gamma_G(GN \times GE)}{1 + \gamma_G GN} \quad (5)$$

$$GN \leftarrow 1 + \gamma_G GN \quad (6)$$

大局的な満足化の度合いは $GE - GR$ で表されるのに対し，各状態での満足化度合いは状態 s_i における最大の Q 値を $\max Q(s_i)$ として $\max Q(s_i) - R(s_i)$ で表される．しかし，前述した通り各状態における $\max Q(s_i)$ はそれぞれ数値が異なるため，満足化度合い ($GE - GR$) を各状態にどのように対応づけるかスケーリングする必要があるため，我々はその差異を補正する変数 $\zeta(s_i)$ を導入し，以下の式で各状態の基準値 $R(s_i)$ を定義する．この GR から $R(s_i)$ への変換手法を大局基準値変換手法 global reference convertion (GRC) と名付けた．正確にバンディット問題で用いられる RS と定量的に対応づけるためには単純な補正ではなく状態毎の特性を反映する必

要があり，その定義や学習法にはさらなる議論が必要となる．しかし本研究ではバンディット問題における RS と正確な対応を取る必要が薄く，従来の経験手法との定性的な性質を比較を行うため，後述するようにスケーリングパラメータ $\zeta(s_i)$ を全ての状態に対して等しくした．

$$\zeta(GR - GE) = R(s_i) - \max Q(s_i) \quad (7)$$

$$\delta_G = \max(GR - GE, 0) \quad (8)$$

$$R(s_i) = \max Q(s_i) + \zeta\delta_G \quad (9)$$

4. 大車輪シミュレーションと結果

GRC と従来の基準値の与え方を比較するため，状態の離散化に起因する不完全知覚問題により，低い水準の報酬しか得られない停滞ループに陥りやすい [坂井 10]，複雑なダイナミクスを有する大車輪運動制御課題のシミュレーションを先行研究と同様の設定 [Uragami 14] で行った．状態は上半身の角度を 24，上半身と下半身のなす角度を 5，上半身の角速度を 7 に等分割した合計 840 種，取りうる行動は腰の関節を “曲げる”，“伸ばす”，“動かさない” の 3 種であり，鉄棒を中心として垂直と大車輪ロボットの先端のなす角度 θ に対して報酬 $r = |\theta/\pi|$ が与えられる．エピソード毎に大車輪ロボットは垂直状態に戻され，シミュレーションは 1 エピソードを 1,000 ステップ，200 エピソードを 50 回行い，以下に示す結果はその平均である．比較に用いたアルゴリズムは，Q 学習，RS とは異なる満足化傾向を有する LS-Q アルゴリズム [Uragami 14]，報酬の分散を用いて探索率を自律調節する VDBE アルゴリズム [Tokic 11] を用いた．LS-Q と Q 学習には行動選択に ϵ-greedy を用い，ϵ は 0.0 に固定したものと，1.0 から始まり，徐々に減衰し 100 エピソードの時点で 0.0 になるように設定したものを用いた．GRC における大局基準にはエージェントが一定のスピードで鉄棒を回転した際の期待報酬となる $GR = 0.5$ に設定し，大局割引率には $\gamma_G = 0.9$，パラメータ $\zeta(s_i)$ は全状態において一律で $\zeta(s_i) = 10.0$ に設定した．また，従来の経験的な基準値には先行研究に習い，成績の良かった全ての状態に対して $R(s_i) = 5.0$ に固定した場合を用いる．

シミュレーションの結果として，単位ステップあたりのエピソード総報酬の時間発展とエピソード毎の greedy な行動を選択した割合を図 1 に示す．獲得報酬の推移 (左図) から，既存手法ではランダム探索率 ϵ による強制的な探索が必要であるのに対し，RS はランダム探索を用いることなく最終的に一定以上の報酬を獲得できている．従来の R 固定 RS と GRC を用いた RS の結果を比較すると，R 固定では徐々に上昇する傾向にあるが，GRC では早く上昇する代わりに一度減少し，その後再び上昇する様子が見られる．また高いほど最大の Q 値を持つ選択肢を選択し，低いほど探索した事を意味する greedy 選択率の推移 (右図) から，GRC では報酬が一時的に減少していた 50 から 100 エピソード付近において同じく探索傾向が強くなっており，GRC では獲得の減少と探索傾向の上昇が連動していると言える．また GRC におけるパラメータ $\zeta(s_i)$ の違いによる影響を考察するため，異なる $\zeta(s_i)$ を用いた場合との比較を図 2 に示す．最終的な獲得報酬は $\zeta(s_i) = 5.0$ が若干低いもののほぼ同じ水準である一方で，学習途中における探索度合いでは $\zeta(s_i)$ の違いに対して差異がみられた．

5. 総合考察

各状態での Q 値のスケールは異なることから本来各状態における達成目標は異なるはずであり，従来の経験的に良さそうな基準値すべての状態に対して一定に固定する方法は望ましくない．本研究は新たに考案した GRC を用いた RS，すなわち各状態の基準値を固定することない，打倒で汎用性の高い満足化形式でも，より直感的な大局基準値で一定以上の報酬を獲得できることを示した．

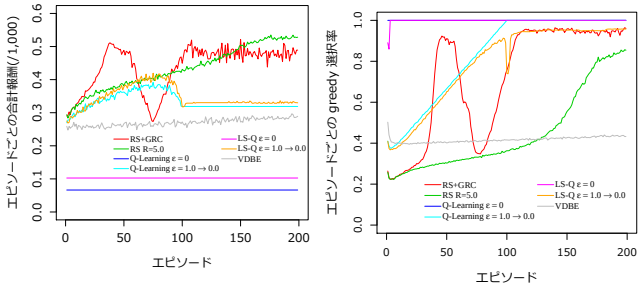

図 1: 報酬の推移と greedy 率の推移

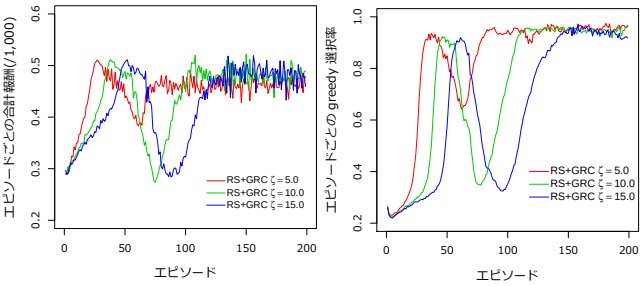

図 2: $\zeta(s_i)$ の違いによる報酬の推移と greedy 率の推移比較

新たに導入したパラメータ $\zeta(s_i)$ は大局的な満足化度合いを各状態の反映度を意味し，その大小は非満足化時の探索度合と連動する．図 2 の結果から最終的な報酬は全て $GR = 0.5$ 近くの報酬を獲得することができており，大局的な満足化に成功していることがわかる．学習途中を比較すると，$\zeta(s_i)$ が大きいほど立ち上がりが遅く，再探索の期間が長いことがわかった．これは再探索の際に，大きすぎる $\zeta(s_i)$ が設定されていたため過度に探索が促されたためであると考えられる．また，$\zeta(s_i)$ が小さい場合に過度な再探索は起きていないが，本来もっと探索が必要なためか良い方策の発見が遅れて，最終的な報酬が他に比べ少し低くなったと考えられる．今後は探索時のロスを減らすために，適切な $\zeta(s_i)$ を決定する手法を考案する．

参考文献

[Simon 56] Simon, H.A.: Rational choice and the structure of the environment, *Psychological Review*, Vol. 63, No. 2, pp. 129–138 (1956)

[Tokic 11] Tokic, M.: Value-difference based exploration: Adaptive control between epsilon-greedy and softmax, in *Proceedings of KI2011*, pp. 335–346 (2011)

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

[坂井 10] 坂井直樹, 川辺直人, 原正之, 豊田希, 薮田哲郎: 強化学習を用いたスポーツロボットの大車輪運動の獲得とその行動形態の考察, 計測自動制御学会論文集, Vol. 46, No. 3, pp. 178–187 (2010)

[牛田 16] 牛田有哉, 甲野佑, 浦上大輔, 高橋達二: 探索割合を自律調節する強化学習手法－満足化基準の動的獲得－, JSAI 2016 (2016 年度人工知能学会全国大会 (第 30 回)) (2016)

[高橋 16] 高橋達二, 甲野佑, 浦上大輔 : 認知的満足化 - 限定合理性の強化学習における効用, 人工知能学会論文誌, Vol. 31, No. 6, pp. 1–11 (2016)