The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

深層満足化強化学習に向けて

Toward Deep Satisficing Reinforcement Learning

佐鳥玖仁朗 *1 Kuniaki Satori　吉田豊 *1 Yutaka Yoshida　神谷匠 *2 Takumi Kamiya　高橋達二 *1 Tatsuji Takahashi

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

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

For dealing with continuous state spaces, DQN and other algorithms have been proposed in reinforcement learning (RL). However, it is hard for DQN to explore efficiently, as it depends on random search strategies such as epsilon-greedy. Humans are known to effectively search and learn through "satisficing" instead of optimizing. Although the risk-sensitive satisificing (RS) algorithm enables satisficing in RL, it depends on the count of visiting each state, which poses a problem for continuous spaces. We propose a method for solving this problem by pseudocount and hash+auto encoder methods that enables intrinsically motivated exploration. Through two experiments, we show that RS combined with the two methods enables deep satisficing RL that searches and learns efficiently in continuous spaces.

1. はじめに

強化学習では，エージェントの試行錯誤を通じて未知の環境において収益を最大化する行動系列を学習することを目的とする．エージェントの試行錯誤には，探索と知識利用の間にトレードオフが存在し，エージェントは両者のバランスを調整しながら学習を行う必要がある．

近年，連続状態空間を扱う方法として深層学習と強化学習を組み合わせた深層強化学習の Deep Q-Network(DQN)[Mnih 2013] がある．DQN は，Atari2600 のゲーム画面を入力とし得られたスコアを報酬とすることで人間レベルのパフォーマンスを達成したアルゴリズムである．しかし，現状では行動を決定する意思決定方法として ϵ-greedy のような単純なルールに依存しているため，膨大な探索空間を扱う場合に探索と知識利用のバランス調整を適切に行えているとは言い難い．

人間は，莫大な探索空間上で意思決定を行う場合に最適化ではなく満足化により意思決定を行う．満足化による意思決定では，現状が目的とする基準を満たしていなければ探索を行い，満たしていれば知識利用を行うように意思決定を行うことで，探索空間の広大な未知の環境であっても基準を満たすような行動を可能としている [Simon 1956].

この満足化を反映した意思決定手法として，Risk-sensitive Satisficing (RS) が考案された． RS は多腕バンディット問題をはじめとした様々なタスクにおいて基準値が適切であれば少数の探索で，最適な行動系列を学習できることがわかっている [高橋 2016][牛田 2017]．深層強化学習においても，RS を適用することで莫大な探索空間上での探索を改善し，基準を満たすような行動系列の短時間での学習が期待される．しかし，RS の深層強化学習への適用する場合，RS がカウントベース手法であり，意思決定に経験した状態行動対の試行数を利用することが問題となる．状態数が極めて多いタスクでは完全に同じ状態に訪れることが少なく，多くの状態行動対が新奇の観測となってしまう．

連絡先: 高橋達二， 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， 049-296-1642, tatsujit@mail.dendai.ac.jp

そこで本研究では，連続状態空間で状態の不確実性から内発的動機付けを行うために状態をカウントする手法として考案されている擬似カウント (pseudo count)[Bellmare 2016] とハッシュ関数+AutoEncorder(AE)[Tang 2017] によって抽象化された状態を用いることで，RS の深層強化学習への適用を図る．

2. Risk-sensitive Satisficing (RS)

本章では，満足化を反映した意思決定手法である RS について説明する． 強化学習において RS を用いる場合，エージェントは RS 価値関数と大局基準変換法 (Global Reference Conversion: GRC) を用いることで意思決定を行う．

2.1 RS 価値関数

RS 価値関数は，方策の信頼度 $\tau(s_i, a_j)$ と行動価値関数 $Q(s_i, a_j)$，各状態の基準値 $\aleph(s_i)$ から式 (1) のように定義される．信頼度 τ はある方策に従った際の行動価値 Q がどの程度信頼できるかの指標であり，状態 s_i において行動 a_j を試行した回数 $\tau_{\mathrm{curr}}(s_i, a_j)$ と，未来信頼度 $\tau_{\mathrm{post}}(s_i, a_j)$ の和として定義される (式 (2)).

$$RS(s_i, a_j) = \tau(s_i, a_j)\Big(Q(s_i, a_j) - \aleph(s_i)\Big) \quad (1)$$

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

$\tau_{\mathrm{curr}}(s_i, a_j)$ と $\tau_{\mathrm{post}}(s_i, a_j)$ は，実際の試行によって式 (3)，式 (4) のように更新される．このとき，状態 s' は状態行動対 (s_i, a_j) により遷移した状態で，行動 a' は状態 s' において選択する行動である．また，γ_τ は未来信頼度割引率，α_τ は信頼度学習率を表す．

$$\tau_{\mathrm{curr}}(s_i, a_j) \leftarrow \tau_{\mathrm{curr}}(s_i, a_j) + 1 \quad (3)$$

$$\tau_{\mathrm{post}}(s_i, a_j) \leftarrow \tau_{\mathrm{post}}(s_i, a_j) + \alpha_\tau\Big(\gamma_\tau \tau(s', a') - \tau_{\mathrm{post}}(s_i, a_j)\Big) \quad (4)$$

RS 価値関数による評価値 RS を最大化する行動を選択する方策を，RS 方策と呼ぶ．

2.2 大局基準変換法 (GRC)

現実的なタスクにおいて，各状態ごとの適切な基準値を推定することは困難である．一方で，タスク全体を通して得られる収益は既知であることも多く，タスク全体の収益に対する適切な基準値の推定は可能である場合が多い．そこで，タスクの収益に対する大局基準値 $\aleph_G$ と現在の方策による収益の大局観測期待値 E_G からタスク全体の満足度合いを求め，その値から各状態の基準値 $\aleph(s_i)$ を求める手法である GRC が考案された [牛田 2017]．

E_G は現在の方策による収益を E_{tmp} としたとき，式 (5) のように更新される．

$$E_G \leftarrow \frac{E_{\mathrm{tmp}} + \gamma_G (N_G E_G)}{1 + \gamma_G N_G} \tag{5}$$

$$N_G \leftarrow 1 + \gamma_{N_G} \tag{6}$$

E_G と $\aleph_G$ の差 δ_G を現在の方策の満足化度合いとしたとき，各状態の基準値 $\aleph(s_i)$ は式 (9) のように計算される．$\zeta(s_i)$ はスケーリングパラメータで，各状態と大局のスケールの違いを吸収するために用いられる．

$$\delta_G = \min(E_G - \aleph_G, 0) \tag{7}$$

$$\max Q(s_i) - \aleph(s_i) = \zeta(s_i)\delta_G \tag{8}$$

$$\aleph(s_i) = \max Q(s_i) - \zeta(s_i)\delta_G \tag{9}$$

2.3 RS の深層強化学習へ適用への問題点

RS を深層強化学習で用いる場合，深層強化学習で扱うようなタスクの状態数の多さが問題となる．観測した状態を信頼度 τ にそのまま用いた場合，その状態へ再訪することが少ないため，適切な評価値 RS の比較が行えない．

そこで，状態の不確実性から擬似的にカウントを行う擬似カウントや，画像上の物体間の位置等を考慮した AENO による状態表現をハッシュ関数を用いることで，抽象的な状態のカウントを行う．それにより信頼度 τ を計算することで，深層強化学習への RS 適用の問題解決を図る．

3. 擬似カウントによる RS の深層強化学習への適用

本章では，疑似カウントによる状態のカウント方法を説明した上で，その信頼度への拡張方法を提案する．

3.1 擬似カウント (pseudo count)

擬似カウントとは状態が極めて多い場合でも状態の経験回数をカウントできるように，状態の不確実性から算出した擬似的な状態の経験回数である．擬似カウントでは，状態が要素の集合で表現されているとみなし，状態の各要素が過去に出現した比率を計算しその比率を全要素について掛け合わせることで，その状態の出現確率を近似する．状態 x の出現確率 $\rho_n(x)$ と新たに状態 x が観測された後の出現確率 $\rho'_n(x)$ は，記録された状態数 n と記録された状態内の x の出現回数 $N_n(x)$ から次のように定義できる．

$$\rho_n(x) = \frac{N_n(x)}{n} \tag{10}$$

$$\rho'_n(x) = \frac{N_n(x)+1}{n+1} \tag{11}$$

n と $N_n(x)$ を未知数 $\widehat{n}$ と $\widehat{N}_n(x)$ とした時，$\widehat{N}_n(x)$ は式 (12) のように求められる．

$$\widehat{N}_n(x) = \frac{\rho_n(x)\,(1-\rho'_n(x))}{\rho'_n(x) - \rho_n(x)} \tag{12}$$

疑似カウントでは，この $\widehat{N}_n(x)$ を擬似的な状態 x の出現回数として使用する．

3.2 擬似カウントによる信頼度の表現

疑似カウントを用いて τ を表現するにあたり，状態ではなく状態行動対の経験回数を擬似的に表現する必要がある．状態 s の要素を s_n とし，行動 a を選択したときの各要素の出現回数を $N_{\tau_{\mathrm{curr}}}(s_n, a)$，未来出現回数を $N_{\tau_{\mathrm{post}}}(s_n, a)$ と表すと，τ_{curr} と τ_{post} は式 (15) のように定義される．ここで以下の式の τ_x には τ_{curr} もしくは τ_{post} が対応する．

$$\rho_{\tau_x}(s,a) = \prod_{k=1}^{n} \frac{N_{\tau_x}(s_k,a)}{\sum_{l=1}^{n} N_{\tau_x}(s_l,a)} \tag{13}$$

$$\rho'_{\tau_x}(s,a) = \prod_{k=1}^{n} \frac{N_{\tau_x}(s_k,a)+1}{\sum_{l=1}^{n} N_{\tau_x}(s_l,a)+1} \tag{14}$$

$$\tau_x(s,a) = \frac{\rho_{\tau_x}(s,a)\,(1-\rho'_{\tau_x}(s,a))}{\rho'_{\tau_x}(s,a) - \rho_{\tau_x}(s,a)} \tag{15}$$

$$\tau(s,a) = \tau_{\mathrm{curr}}(s,a) + \tau_{\mathrm{post}}(s,a) \tag{16}$$

行動 a を選択したときの各要素の出現回数 $N_{\tau_{\mathrm{curr}}}(s_n, a)$ と未来出現回数 $N_{\tau_{\mathrm{post}}}(s_n, a)$ は，以下のように更新される．

$$N_{\tau_{\mathrm{curr}}}(s_n,a) \leftarrow N_{\tau_{\mathrm{curr}}}(s_n,a) + 1 \tag{17}$$

$$N_{\tau_{\mathrm{post}}}(s_n,a) \leftarrow N_{\tau_{\mathrm{post}}}(s_n,a) + \alpha_\tau(\gamma_\tau \tau(s',a') - \tau_{\mathrm{post}}(s,a)) \tag{18}$$

以上の手法により信頼度 τ を擬似カウントで表現することで，RS を深層強化学習へ適用した．

4. ハッシュ関数と AE による RS の深層強化学習への適用

本章では，ハッシュ関数と AE による状態表現と，その状態表現を用いた信頼度の拡張方法を提案する．

4.1 SimHash と AE による状態表現

先行研究において，直接カウントを行うことが困難な連続状態の環境で状態カウントを行う手法として，SimHash と AE を用いた手法が提案されている [Tang 2017]．SimHash はハッシュ関数の一種で，類似値を類似ハッシュ値にハッシュ化する性質を持つため，この性質を利用することで類似状態を抽象化して同一状態として扱うことが可能である．また，Atari2600 など，画像を直接的に状態入力へ用いる環境の場合，非常に複雑であるため状態の抽象化は困難である．そのため，AE を用いることで状態を次元削減し，その削減された状態表現をハッシュ関数を通して抽象化することで，連続状態のカウントを行う．

4.2 SimHash と AE による信頼度の表現

式 (3), (4) では直接 s_t を用いていたが，SimHash と AE を用いた信頼度は SimHash によって離散化された k 次元のベクトル $g(s_t)$ を用いる．

また，画像を状態として用いる場合，SimHash への入力は，s_t を入力とした AE の隠れ層 h からの出力となる．

本実験の AE は CNN を用い，バイナリ層の活性化関数は sigmoid 関数，それ以外の層の活性化関数は Relu を用いた．具体的な構成，各層のノード数を図 1 に示す．

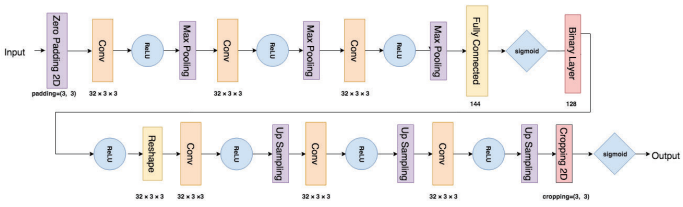


図 1: AE の構成

また，損失関数は式 (19) のように定義した．

$$Loss = \frac{1}{N}\sum_{n=1}^{N}[\log p(s_n) - \sum_{i=1}^{D}\min\left\{(1-b_i(s_n))^2, b_i(s_n)^2\right\}] \quad (19)$$

損失関数は AE の入力と出力の誤差を表す第 1 項と，バイナリ層における出力を 0 か 1 に偏るようにコストをかける正則化項から構成される．AE はエージェントの学習中にそれまでに観測した状態によって動的に更新を行う．実験を通して AE の更新は 3 ステップに 1 回, SimHash の出力のハッシュ値は 16 次元のベクトルとした．

以上の SimHash と AE を用いた手法により信頼度 τ を表現することで，RS を深層強化学習へ適用した．

5. MountainCar

既存手法と提案手法の比較のため，状態が極めて多くかつ報酬が疎である MountainCar タスクで実験を行った．

5.1 環境設定

環境は OpenAI Gym の MountainCar タスクを用いた．MountainCar はスタート位置の左右に二つの山があり，エージェントは台車を操作する．台車がゴールにたどり着くためには，一度ゴールと反対側の山を途中まで登り反動をつける必要がある．今回は報酬を疎にするため，ゴールに到達した時のみ報酬 1 を得るように MountainCar の設定を変更した．MountainCar では速度と位置情報により状態が表現されており， それらを等分割して離散化することで信頼度に使用した．また他手法との比較のため，離散化の必要のないハッシュ関数による信頼度も状態を等分割して離散化した場合と連続値の場合を別々に用いた．状態の分割数を変化させ，各手法による信頼度の表現の性能比較を行った．

Q 値の推定には DQN を使用し，近似するニューラルネットワークのノードは入力層 2 個（連続値），隠れ層は 128 個が 3 層，出力層 3 個とし，活性化関数は Relu，損失関数は HuberLoss，学習率 lr =5e-4，割引率 $\gamma = 0.99$ とした．リプレイメモリサイズは 50,000，バッチサイズは 32，ターゲットネットワークの更新頻度は 500 ステップとした．RS は全ての分割数設定と信頼度表現手法で 200 ステップ以内のゴールを目的として, $\aleph_G = 1/200$，E_{tmp} はエピソードの平均報酬として 報酬/ステップ数，$\zeta = 1$，$\gamma_G = 0.9$，$\alpha_\tau = 0.1$，$\gamma_\tau = 0.9$ とした．

5.2 結果

図 2 に各手法の分割数ごとの目的達成までの累計ステップ数の 10 シミュレーションの平均の結果を示す．縦軸は累計ステップ数，横軸は分割数とし，エピソード終了条件はゴールに到達したときのみとした．また, タスクのクリア条件を移動平均 200 ステップ以内の到達を良い行動系列の条件とし, 移動平均はエピソードごとのステップ数の 100 エピソード分の移動平均を用いた．分割数が増えるごとに状態数が増えているため，経験カウントの累計ステップ数が大幅に増加したのに対して，擬似カウントとハッシュ関数は少ない累積ステップ数でクリア条件を達成できている．このことから, 状態が極めて多い場合では経験カウントに比べて提案手法の擬似カウントとハッシュ関数で信頼度を表現した方法が有用であることが示せている．

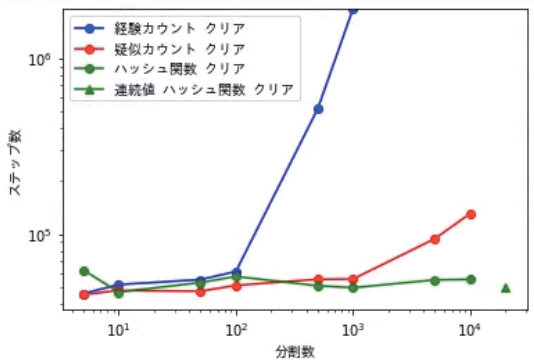


図 2: 擬似カウントと経験カウントを用いた RS の比較

6. SuboptimaWorld

SuboptimaWorld は準最適解が複数存在するタスクである．MountainCar の実験で状態の要素から信頼度を計算したが，SuboptimaWorld では画像から信頼度を計算し RS の学習が行えるか検証した．また，DQN で通常使われる ϵ-greedy と内発的動機付けの手法である MBIE-EB との比較も行った．

6.1 環境設定

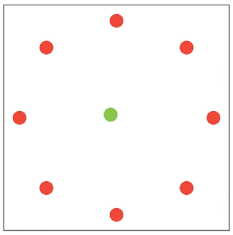
実験環境は連続状態を扱い，ゴールが複数あるタスクを用いた．このタスクは状態の要素として x 軸，y 軸の位置が $[0, 1]$ の連続状態で与えられる．図の赤い点のゴールは 8 つ存在し，報酬は $[1.0, 0.8, 0.8, 0.6, 0.6, 0.4, 0.4, 0.4]$ の 8 つの値がランダムで配置される．スタートの位置は中心付近でランダムに決められ，エージェントの行動は上下左右とその斜めの計 8 つの行動が存在する．各手法による信頼度の比較のため，Q 値の近似は簡易的に x 座標と y 座標を入力とした NN を用いた．

DQN の設定は，学習率 lr=1e-4，近似するニューラルネットワークのノードは入力層 2 個（連続値），隠れ層は 128 個が 3 層，出力層 8 個とし，他の設定は MountainCar の実験と同じものを使用した．RS は最大報酬の獲得を目的として $\aleph_G$=1，E_{tmp} は 1 エピソードの獲得報酬，ζ は擬似カウントを用いた RS は 0.01，ハッシュ関数+AE を用いた RS は 0.5 で，$\gamma_G = 0.9$，$\alpha_\tau = 0.1$，$\gamma_\tau = 0.9$ とした．

RS の信頼度は画像から計算するため，図 3 の画像をグレースケールに変換し，ピクセルを 42 × 42 へ平均値によるダウンサンプリングを行なった．擬似カウントの信頼度の計算では，各ピクセルのグレースケール値を 8 分割 (3bit) に変換し，

各ピクセルを状態の要素とし，各ピクセルの出現確率の積を状態の出現確率とした．

比較アルゴリズムとして，DQN で通常使われる ϵ-greedy と，擬似カウントとハッシュ関数+AE の論文で使われた内発的動機付け手法である MBIE-EB を擬似報酬として用いた ϵ-greedy を使用した．ϵ-greedy は初期値を $\epsilon = 1$ とし，定量的に減少させ 20000 ステップ以降は $\epsilon = 0.1$ で固定した．MBIE-EB のパラメータである β は擬似カウントを用いた場合では 0.01，ハッシュ関数+AE を用いた場合では 0.001 とした．MBIE-EB の報酬ボーナスは経験再生が行われる際に計算し報酬に加算した．



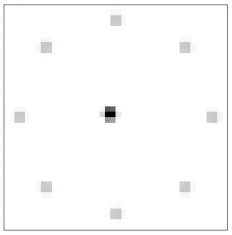


図 3: SuboptimaWorld　図 4: 加工した画像 (42 × 42)

6.2 結果

図 5 に，200000 ステップを 1 シミュレーションとして行い，10 シミュレーションを平均した結果を示す．縦軸は報酬で横軸はステップ数である．ϵ-greedy は局所解に陥った．それに対して，提案手法の二つの RS と擬似カウントを用いた MBIE-EB は，最大報酬 1 への行動系列を学習することができていることがわかる．

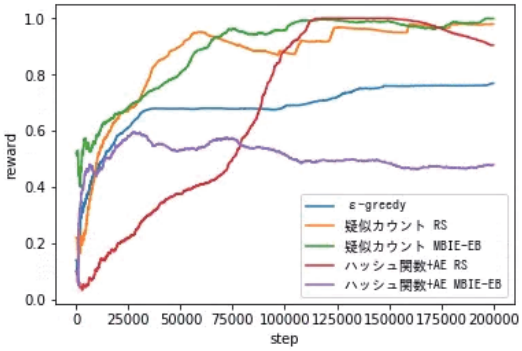


図 5: 提案手法 RS と比較アルゴリズムの獲得報酬

7. 考察

MountainCar の実験では，経験カウントを用いた RS がタスクをクリアするまでにかかった累計ステップ数が要素の分割数が増えるごとに一気に増加している．状態数が増えたことで観測の必要数が増え学習速度が低下した結果，安定してゴールにたどり着けなかったためだと考えられる．対して擬似カウントを用いた RS は，状態が抽象化された結果，新奇な状態が少なくなり，分割数が増えても対応できたと考えられる．ハッシュ関数を用いた RS は，SimHash を用いることで類似状態がある程度同一ハッシュ値にまとめられるため，状態を分割した場合と連続値を扱った場合の離散化の度合いが類似しタスククリアのステップ数が一定になったと考えられる．

SuboptimaWorld の実験では，提案手法 2 つを含む 3 つの手法が局所解に陥らずに学習できている．提案手法の 2 つの RS は，目標の報酬 1 のゴールの行動系列を見つけるまで満足できないため探索を続けたことで効率的な探索が行えたためだと考えらえる．ハッシュ関数+AE を用いた RS は，擬似カウントを用いた RS に比べて最大報酬 1 の行動系列の学習に時間がかかっている．AE は状態表現の学習を行う必要があるため，安定した状態の抽象化が行われるまでカウント数が安定しないためと考えられる．擬似カウントを用いた MBIE-EB は，不確実性に与えられるボーナス報酬により未探索状態を進んで探索し，ほぼ全ての報酬を観測できたことにより最終的に高い報酬を獲得する行動系列を学習することができたと考えられる．

カウントを用いた手法のうち，ハッシュ関数+AE を用いた MBIE-EB のみ局所解に陥っている．ハッシュ関数+AE を用いた RS と同様，カウント数が不安定であり，シミュレーション内でボーナスの加算が適切に行われなかったためと考えられる．疑似カウントによる手法は RS と MBIE-EB の両者とも安定した学習ができているが，ハッシュ関数+AE による手法は MBIE-EB のみ学習が進まなかった．RS と MBIE-EB に差がでた理由として，MBIE-EB ではボーナス報酬が直接 Q 値に影響を与えるため AE の学習初期の影響が強く，RS は Q 値に直接影響を及ぼさないため状態の不確実性を素早く利用できたと考えられる．

8. おわりに

本研究では RS を深層強化学習へ適用する方法を提案した．状態が極めて多い場合でも状態のカウントが可能な擬似カウントとハッシュ関数+AE により，RS の信頼度を表現し深層強化学習へ適用できた．今後の課題として，さらに擬似カウントとハッシュ関数+AE のカウント表現の性質の差を明らかにする必要がある．また RS の信頼度の表現方法は，より複雑な Atari2600 タスクへの適用や，今回の手法以外の状態の不確実性を測り定量化する方である RND などに関しても比較検証する必要がある．

参考文献

[Bellmare 2016] Bellemare, M.G. Srinivasan, S. Ostrovski, G. Schaul, T. Saxton, T. Munos, R.: Unifying Count-Based Exploration and Intrinsic Motivation, NIPS2016. (2016)

[Mnih 2013] Mnih, V. Kavukcuoglu, K. Silver, D. Graves, A. Antonoglou, I. Wierstra, D. Riedmiller, M.: Playing Atari with Deep Reinforcement Learning, *Nature*, 518(7540), 529–533. (2015)

[Simon 1956] Simon, H.A.: Rational choice and the structure of the environment, *Psychological Review*, 63(2), 129138. (1956)

[Tang 2017] Tang, H. Houthooft, R. Foote, D. Stooke, A. Chen, X. Duan, Y. Schulman, J. Turck, F.D. Abbeel, P: #Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning, NIPS2017. (2017)

[牛田 2017] 牛田 有哉, 甲野 佑, 高橋 達二: 生存を目的とする満足化強化学習, JSAI2017. (2017)

[高橋 2016] 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)