1M2-4

成功確率と収益を組み合わせた行動価値に基づく強化学習

Reinforcement Learning Based on Action Values Combined with Success Probability and Profit

堀江 直人 *1 Naoto Horie　松井 藤五郎 *2 Tohgoroh Matsui　森山 甲一 *1 Koichi Moriyama　武藤 敦子 *1 Atsuko Mutoh　犬塚 信博 *1 Nobuhiro Inuzuka

*1名古屋工業大学大学院工学研究科情報工学専攻
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

*2中部大学生命健康科学部臨床工学科
Department of Clinical Engineering, College of Life and Health Sciences, Chubu University

Success probability based reinforcement learning is a method that calculates the expected value of the success probability of each state action pairs, removes actions with low success probability and selects a safe action from the remaining actions. However, this method is not suitable in the environment where risk is not avoidable, i.e., no action has sufficient success probability. So we propose a method based on the combination of the success probability and conventional action values and investigate the behavior selection when risk is not avoidable.

1. はじめに

強化学習 (RL) は，試行錯誤に基づいて行動を学習する機械学習の手法であり，ロボットの行動学習などに利用されている．ロボットの行動学習においては，実際のロボットで試行錯誤をすることは困難であるため，一般的には，コンピュータ・シミュレーションによって反復学習を行い，学習した結果が実際のロボットに適用される．しかしながら，シミュレーションによる反復学習では，ロボットの転倒など，シミュレーションでは問題のない結果をもたらす行動であっても，実際のロボットに適用した際には問題となる行動を学習してしまうことがある．このような背景から，強化学習において安全な行動を学習させることを目的とした安全な強化学習 (Safe RL) の研究が進められている [Garcia 15].

安全な強化学習の手法の 1 つに，得られる利益を低下させてでも破滅的状況に陥ることを回避する，リスクに敏感な強化学習がある [Geibel 05][Mihatsch 02]．リスクに敏感な強化学習の一つに，竹山らが提案した，成功確率に基づく強化学習 [竹山 15] がある．成功確率に基づく強化学習は，成功確率の低い行動をリスクと考えて，各状態行動対の成功確率の期待値を求め，成功確率の低い行動を行動選択から取り除き，安全な行動選択を行う手法である．しかし，リスクが回避不可能な場合，すべての行動が取り除かれるため，行動選択を獲得することができないという問題がある．

そこで，本論文では成功確率と従来の行動価値を組み合わせた新しい価値に基づく手法を提案する．また，提案手法が，リスクが回避可能な場合に加え，リスクが回避不可能な場合でも適切な行動を学習することを示す.

2. 強化学習

強化学習 [Sutton 00] とは，目標を達成するためにエージェントに相互作用から行動を学習させる手法である．強化学習における相互作用とは，離散的な時間ステップ $t = 0, 1, 2, 3, \ldots$ において，エージェントが，環境から現在の状態 s_t を受け取り，状態 s_t で選択できる行動 a_t を選択・実行し，エージェントの行動に対して環境から新しい状態 s_{t+1} と報酬 r_t を受け取るサイクルのことである．

連絡先: 堀江直人，名古屋工業大学，愛知県名古屋市昭和区御器所町，n.horie.026@nitech.jp

割引報酬の和を収益といい，収益 R_t は，

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots$$

で表される．γ は割引率といい，将来得られる報酬の重みを表す値で，0 以上 1 以下の値をとる．

エージェントには各時間ステップにおいて各状態で選択可能な行動を選択する確率を表す方策 π_t を持つ．$\pi_t(s, a)$ は時刻 t で $s_t = s$ ならば $a_t = a$ となる確率である．方策の例として，ϵ グリーディ方策が挙げられる．ϵ グリーディ方策は，確率 ϵ でその状態で選択可能なすべての行動からランダムに行動を選択し，確率 $1 - \epsilon$ で収益の期待値が最大化されるような行動を選択する．これにより，まだ十分な回数訪問されていない状態行動を訪問することができるようになるため，最適な方策を作ることができるようになる．

価値関数は，強化学習において，エージェントがいる状態の好ましさを表す関数で，収益の期待値で表される．方策 π の下で状態 s において行動 a を実行した場合の収益の期待値を

$$Q^\pi(s, a) = E_\pi\{R_t | s = s_t, a_t = a\}$$

で定義し，これを行動価値関数 (Q 値) と呼ぶ.

全ての方策の中で最大の Q 値を最適行動価値関数 Q^* という．Q^* は全ての状態と行動に対して

$$Q^*(s, a) = \max_\pi Q^\pi(s, a)$$

で定義される．Q^* を求める手法として，Sarsa や Q 学習が存在する．

3. 成功確率に基づく強化学習

竹山らが提案した成功確率に基づく強化学習 [竹山 15] は，各状態 s において成功する確率が最も高い行動を学習する手法である．状態 s において行動 a を選択したときの成功確率 $P(s, a)$ は

$$P(s, a) = P(r_{t+1} = 1 | s_t = s, a_t = a)$$

で表される．時刻 t での成功確率を $P(r_{t+1})$ と表記するとき，時刻 t 以降に失敗しない確率 (成功確率) $\mathcal{P}_t$ は

$$\mathcal{P}_t = P(r_{t+1})P(r_{t+2})^{\gamma}P(r_{t+3})^{\gamma^2}\ldots = \prod_{k=0}^{\infty} P(r_{t+k+1})^{\gamma^k}$$

で表される．上式の対数をとることで，成功確率の対数を再帰的に定義できる．

$$\begin{aligned}\log \mathcal{P}_t &= \log \prod_{k=0}^{\infty} P(r_{t+k+1})^{\gamma^k} \\ &= \log P(r_{t+1}) + \gamma \sum_{k=0}^{\infty} \gamma^k \log \mathcal{P}_{t+1}\end{aligned}$$

方策 π の下での状態 s における行動 a の行動価値関数 $\mathrm{Pr}Q^{\pi}(s,a)$ は次のように表される．

$$\begin{aligned}\mathrm{Pr}Q^{\pi}(s,a) &= E_{\pi}[\log \prod_{k=0}^{\infty} P(r_{t+k+1})^{\gamma^k} | s_t = s, a_t = a] \\ &= \sum_{s' \in S} \mathbb{P}^a_{ss'} \left\{ \log P(s,a) + \gamma V^{\pi}(s') \right\}\end{aligned}$$

ここで，$\mathbb{P}^a_{ss'}$ は状態 s で行動 a を選択した場合に状態 s' に遷移する確率を表す．

行動価値関数 $\mathrm{Pr}Q(s_t,a_t)$ の更新式を

$$\begin{aligned}\mathrm{Pr}Q(s_t,a_t) \leftarrow \mathrm{Pr}Q(s_t,a_t) + \alpha(&\log P(s_t,a_t) \\ &+ \gamma \mathrm{Pr}Q(s_{t+1},a_{t+1}) - \mathrm{Pr}Q(s_t,a_t))\end{aligned}$$

で表す．$P(s_t,a_t)$ は，状態 s で行動 a を選択したときの成功率であり，以下の式で表す．

$$P(s,a) = \frac{\sum_{k=0}^{t} I_{ss_k} I_{aa_k} r\,'_{k+1} + 1}{\sum_{k=0}^{t} I_{ss_k} I_{aa_k} + 1}$$

I_{xy} は $x = y$ の時に 1，そうでないときに 0 をとる一致関数である．$r\,'$は行動が成功した場合 1，失敗した場合に 0 となる値である．成功確率が 0 のとき，$\log P(s_t,a_t)$ が $-\infty$ に発散するが，すべての行動に対して訪問回数 (分子) と成功回数 (分母) を 1 に初期化することで，発散を回避する．

竹山らは，PrSarsa を使い $\mathrm{Pr}Q(s,a)$ を求め，その値が一定以下の行動を行動選択から取り除き，残った行動の中から Sarsa の行動価値関数に基づいて行動を選択することによって，危険な行動を避けるような行動を学習させた．

4. 提案手法

3 章で述べたように，竹山らが提案した手法は，PrSarsa と Sarsa を組み合わせることで成功確率の低い行動を行動選択から取り除き，残った行動の中から Sarsa の価値関数に基づいて行動選択を行うため，安全な行動を学習することができる．しかし，すべての行動に失敗する可能性が存在する環境では，最適な行動を学習できない．そこで，PrQ を使用した新しい価値関数を定義し，危険を回避できない環境での行動学習について検討する．

本手法では，PrSarsa の価値関数 PrQ と，従来の強化学習の価値関数 Q を組み合わせたものを，価値関数 EQ (Expected Q) として定義する．

$$Q(s,a) = E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k (r_{t+k+1})\right]$$

$$\mathrm{Pr}Q(s,a) = E_{\pi}\left[\log \prod_{k=0}^{\infty} (P_{t+k+1})^{\gamma^k}\right]$$

$$EQ(s,a) = \begin{cases} e^{\mathrm{Pr}Q(s,a)} \times Q(s,a) & (Q(s,a) \geq 0) \\ (1 - e^{\mathrm{Pr}Q(s,a)}) \times Q(s,a) & (Q(s,a) < 0) \end{cases} \quad (1)$$

上式の $EQ(s,a)$ では，ネイピア数 e を $\mathrm{Pr}Q(s,a)$ 乗している．イェンゼンの不等式から，$e^{\mathrm{Pr}Q^{\pi}(s,a)} \times Q(s,a) \geq E_{\pi}[\prod_{k=0}^{\infty} P(r_{t+k+1})^{\gamma^k} | s_t = s, a_t = a] \times Q(s,a)$ となり，$EQ(s,a)$ は価値関数 $Q(s,a)$ の期待値以上の値となる．これにより，従来の Q 値に成功確率に基づいた重みをつけることができ，成功確率が低い行動の価値を小さくしつつも，PrSarsa と Sarsa を組み合わせた手法とは異なり，成功確率が低い行動でも行動選択から除外しないようにすることができる．

$Q(s,a) < 0$ のときに $(1 - e^{\mathrm{Pr}Q(s,a)}) \times Q(s,a)$ としているのは，期待値が負になるような環境で，成功確率の低い状態行動対の価値の絶対値が 0 に近づくのを防ぐためである．

図 1 に，$EQ(s,a)$ のアルゴリズムを示す．$N(s,a)$ は訪問回数，$M(s,a)$ は成功回数を表す．$N(s,a)$ はその状態行動対を訪問するたびに 1 加算され，$M(s,a)$ はその状態行動対が成功した場合 $(r' = 1)$ に 1 加算される．

5. 実験および結果

5.1 実験 1: 安全経路が存在する格子世界

図 2 に示す格子世界を用意し，リスクのある環境でのエージェントの行動について調査する．格子世界には，終端状態として，G, F が存在する．灰色のマスは移動不能なマスとして扱う．エージェントは移動すると，報酬と，成功または失敗の情報を得る．ここでは，F に到達した場合を失敗とし，それ以外を成功とする．G と F の隣に数字が書かれているマスが存在するとき，そのマスから G に移動しようとした場合に，マスに書かれている値の確率で，F のマスに移動する（つまり，失敗する）．今回の実験では，0.5 の確率で F のマスに移動するようにした．エージェントは格子空間上の G と F と灰色のマス以外のいずれかのマスを初期状態とし，行動を繰り返して終端状態に到達するまでを 1 エピソードとする．

報酬は，$(0,1)$ では 150，$(6,1)$ では 100，$(3,4)$ では 200，$(1,2)$ では 50，$(4,3)$ では 0 の報酬を得る．それ以外のマスはすべて 0 の報酬を得る．PrSarsa では，F のマスに到達した場合は 0，それ以外のマスに到達した場合は 1 の報酬を得る．$(6,1)$ 以外の G のマスは 0.5 の確率で異なった報酬を得るので，最終的にどの終端状態も得られる収益の期待値は 100 に収束する．

竹山らの手法（以下 PrSarsa）は，現在の状態 s において $\mathrm{Pr}Q(s,a) \leq -0.1$ となるすべての行動を行動選択から除き，それ以外の行動の中から Sarsa の行動価値関数によって導かれる方策を用いて行動選択を行う．すべての行動が取り除かれ，選択可能な行動がない場合は，ランダムに行動することにした．行動の方策を ϵ グリーディ法，1 試行を 100,000 エピソード，1 エピソードを 1,000 ステップとして，実験を 20 回試行し，結果を調べた．パラメータは $\epsilon = 0.2, \alpha = 0.01, \gamma = 0.9$ とし，乱数アルゴリズムはメルセンヌツイスタ法を使用した．乱数に使用するシードは実験ごとに変更している．

1. $Q(s,a), PrQ(s,a), EQ(s,a)$を任意の値で初期化
2. すべてのs,aに対し，$N(s,a)=1, M(s,a)=1$
3. 各エピソードに対して繰り返し:
    1. sを初期化
    2. Qに導かれる方策（行動選択確率）にしたがってsでの行動aを選択
    3. エピソードの各ステップに対して繰り返し:
        1. 行動aを選択し，報酬rと成功失敗判定r'と次状態s'を観測
        2. $Q(s,a)$に導かれる方策に従ってs'での行動a'を選択
        3. $N(s,a) \leftarrow N(s,a)+1$
        4. $M(s,a) \leftarrow M(s,a)+r'$（$r'$は成功した場合1、そうでない場合0）
        5. $Q(s,a) \leftarrow Q(s,a)+\alpha[r+\gamma Q(s',a')-Q(s,a)]$
        6. $PrQ(s,a) \leftarrow PrQ(s,a)+\alpha[\log\left(\frac{M(s,a)}{N(s,a)}\right)+\gamma PrQ(s',a')-PrQ(s,a)]$
        7. $Q(s,a) \geq 0$の場合
            1. $EQ(s,a) \leftarrow e^{PrQ(s,a)} \times Q(s,a)$
        8. $Q(s,a) < 0$の場合
            1. $EQ(s,a) \leftarrow (1-e^{PrQ(s,a)}) \times Q(s,a)$
        9. s', a'をそれぞれs,aに代入
        10. sが終端状態なら終了

図 1: EQ のアルゴリズム

図に実行結果を示す．図 3，図 4，図 5，図 6 はそれぞれの手法で学習した経路である．図中の矢印は，移動方向を示し，×は PrSarsa により取り除かれた行動が存在するマスである．図 3 は PrSarsa，図 4 は提案手法の結果である．図 5 は Sarsa，図 6 は Q 学習である．PrSarsa と提案手法では，右に進む，確実に成功することができる経路を学習した．Sarsa では，右に進む経路が学習されたが，提案手法とは異なり，失敗を回避する経路を学習できなかった．Q 学習では，すべての経路が均等に選択され，Sarsa と同様に失敗を回避する経路は学習できなかった．

実験 1 の結果から，提案手法，PrSarsa は安全な経路が存在する場合，安全な経路を学習することが分かった．しかし，この実験のみでは PrSarsa との違いを説明するには不十分である．よって，実験 2 では，安全な経路を取り除いた環境で，実験を行った．

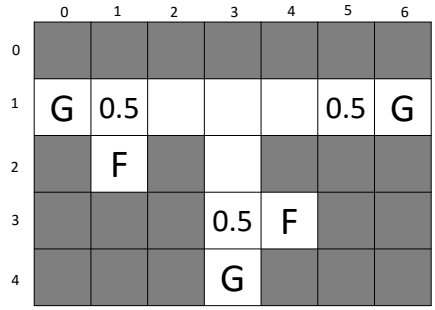

図 2: 実験 1 の格子世界

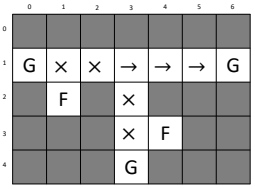

図 3: 実験 1 の PrSarsa の結果

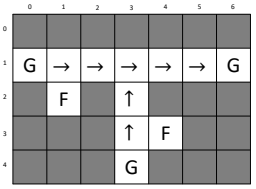

図 4: 実験 1 の提案手法の結果

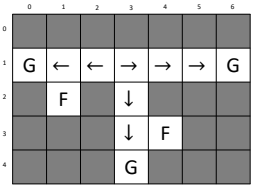

図 5: 実験 1 の Sarsa の結果

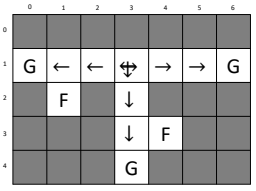

図 6: 実験 1 の Q 学習の結果

5.2 実験 2: 安全経路が存在しない格子世界

実験 2 では，安全な経路を取り除いた場合の学習経路を確認するための実験を行った．竹山らの手法と，提案手法について，リスクを回避できない環境での行動選択の学習について実験を行った．

実験に使用した環境を図 7 に示す．実験 1 との違いは，リスクを回避できる経路が存在しないという点であり，G のマスに到達した場合は 100，F のマスに到達した場合は 0 の報酬を得るよう変更した．各パラメータ，乱数アルゴリズムは実験 1 に使用したものと同様である．

図に実行結果を示す．図 8，図 9 はそれぞれの手法で学習した経路である．図 8 は PrSarsa，図 9 は提案手法の結果であ

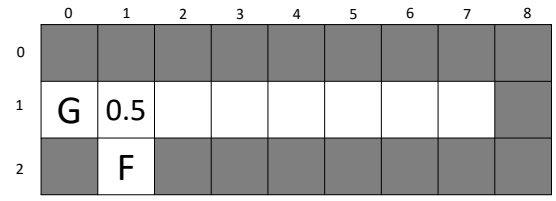

図 7: 実験 2 の格子世界

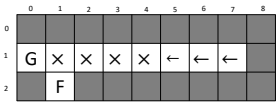

図 8: 実験 2 の PrSarsa の結果

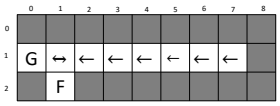

図 9: 実験 2 の提案手法の結果

る．PrSarsa では，図 8 に示すように，$(1,3)$ より左側のマスに移動するような行動は $\Pr Q(s,a) \leq -0.1$ となっており，行動選択から取り除かれ，$(1,3)$ より左のマスでは立ち往生する．よって，$(1,4)$ と $(1,5)$ を往復するような行動を学習した．一方提案手法では，図 9 に示すように，実験 20 回中 14 回目標状態に向かうような行動を学習したが，6 回 $(2,1)$ で右に進むような経路を学習した．

実験 2 では，実験 1 で説明することができなかった，提案手法と PrSarsa の違いを説明するために，安全経路を取り除いた．結果から，提案手法はリスクを回避できない場合でも，ある程度目標状態に進む経路を学習することができることがわかった．

6. まとめ

成功確率に基づく強化学習手法では，価値関数を成功確率に基づいて定義し，成功確率が小さい行動を選択しないようにすることで安全な経路を学習する．しかし，成功確率が小さい行動を選択しないため，リスクを回避できない環境で，目標状態までの経路を学習させたい場合には適さない．

そこで，本論文では，成功確率に基づく強化学習手法を拡張し，成功確率に基づく収益の期待値を用いた行動価値関数 EQ を提案した．EQ はネイピア数を $\Pr Q(s,a)$ 乗することで連続的な成功確率に基づく Q 値の期待値を求めることができる．これにより，従来の Sarsa や Q 学習ではリスクのある経路を学習する問題において，EQ を用いることでリスクを回避する経路を学習することができ，また，リスクを回避することができない環境で目標地点に到達する経路を学習させたい場合でも，ある程度目標状態に到達する行動を学習させることができた．

本研究の発展として，災害救助ロボットの行動選択に使用することが挙げられる．従来の Sarsa や Q 学習が危険な経路を学習する環境で，安全な経路を学習できる．また，竹山らの手法で正しく学習できない，リスクを回避することができない環境で経路を学習させることができるため，このような環境でも安全な経路を学習させることが期待できる．

今回の研究では，格子世界のような単純な環境において提案手法が動作することを確かめることができたが，複雑な環境では実験していない．そこで，より複雑な環境を設定し，その環境で動作することを確かめることと，時々目標状態に到達できない場合があったので，確実に目標状態に到達する経路を学習させるために，本論文で提案した EQ の式 1 を改善することが今後の課題である．

参考文献

[Garcia 15] Javier Garcia and Fernando Fernandez: A Comprehensive Survey on Safe Reinforcement Learning, *Journal of Machine Learning Research*, Vol. 16, pp. 1437–1480 (2015)

[Geibel 05] Peter Geibel and Fritz Wysotzki: Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Articial Intelligence Research*, Vol. 24, pp. 81–108 (2005)

[Mihatsch 02] Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, Vol. 49, No. 2-3, pp. 267–290 (2002)

[Sutton 00] Richard S. Sutton and Andrew G.Barto 著, 三上貞芳, 皆川雅章共訳: 強化学習. 森北出版 (2000)

[竹山 15] 竹山大貴, 加納政芳, 松井藤五郎, 中村剛士: 成功確率に基づく強化学習によるロボットの危険回避行動の獲得, 知能と情報, Vol. 27, No. 6, pp. 877–884 (2015)