強化学習を用いた腐敗性を有する在庫問題の最適化について

On an optimization method for perishable inventory problem using reinforcement learning

髙橋勇人 [1]，星野満博 [2]
[1] 秋田県立大学大学院システム科学技術研究科
[2] 秋田県立大学システム科学技術学部
Hayato Takahashi and Mitsuhiro Hoshino
[1]Graduate School of Systems Science and Technology,
Akita Prefectural University
[2]Faculty of Systems Science and Technology,
Akita Prefectural University

1　腐敗性を有する在庫管理モデル

本研究では，腐敗性を有する定期不定量発注方式の在庫管理問題について，マルコフ決定過程 (Markov Decision Process; MDP) と強化学習 (Reinforcement Learning) という二つの観点から定式化と数値実験を行う．比較のために，どちらも同じ問題となるように定式化することを考慮する．MDP は状態や行動など問題のサイズが大きくなると最適解の計算が困難になる．これを回避するための代替手法として強化学習が有効かどうか検討するため，学習中の状態価値関数などについて考察する．

腐敗性在庫とは，食品や医療品など消費期限，有効期限があるようなものを指す．腐敗性在庫には保管期間に限りがあり，それを過ぎた製品は廃棄しなければならない．そのため，保管期間を超えるような期間で続けて在庫管理問題について分析することは難しい．

MDP は不確実性を伴う逐次決定に関するモデルで，在庫管理における需要などの不確実性や定期または不定期な発注を表現できる [1][2]．MDP の関連分野である強化学習は機械学習の一種で，在庫管理問題についても強化学習を利用した研究 [3] が行われている．

本研究では，在庫管理モデル [4] に廃棄処理を追加し，固定期間，確率的需要，定期不定量発注方式の腐敗性を有する在庫管理モデルを扱う．1 期間の在庫管理は図 1 に示す流れで行われ，各期間の始めの定期発注量を決定する問題を考える．

この在庫管理モデルは以下を仮定する．

(a) 発注量の決定は固定された各期間の始めに行われる．
(b) 需要は固定された各期間の終わりに満たされる．
(c) リードタイムは 0 期間とする (発注は直ちに到着する)．

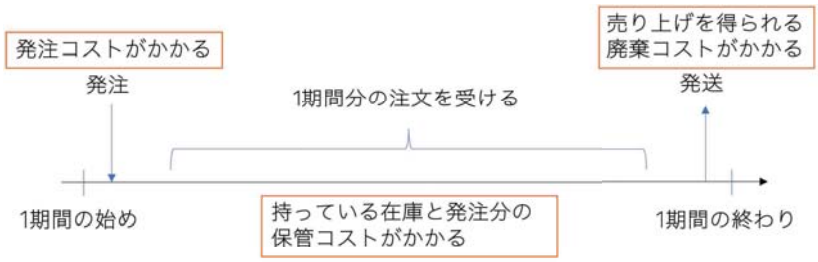

図 1　1 期間の在庫管理の流れ

(d) 需要が在庫を超える場合，超過需要は失われる．
(e) 収益，費用，需要分布は定常とする．
(f) 製品は単位でのみ販売する．
(g) 保管期間 τ を過ぎた製品は廃棄する．$(\tau \geq 2, \tau \in \mathbb{N})$
(h) 在庫管理は先入れ先出しによって行う．
(i) 倉庫の容量は M 単位とする．$(M \in \mathbb{N})$

2　マルコフ決定過程と強化学習

MDP の要素と MDP を用いた強化学習の枠組み，使用する学習アルゴリズムについて述べる．本研究では，強化学習の枠組みに用いられる数理モデルとして，MDP を用いる場合について考える．

有限状態有限行動 MDP は要素の組 $(\mathcal{S}, \mathcal{A}, p, r)$ によって特徴付けられる．

(i) 有限状態集合 $\mathcal{S} = \{s^1, s^2, \ldots, s^N\}$
(ii) 有限行動集合 $\mathcal{A} = \{a^1, a^2, \ldots, a^M\}$
(iii) 推移関数 $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$
(iv) 報酬関数 $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

意思決定者は各時刻のある状態で実行可能な行動を選択し，それに応じた確率的状態推移と報酬獲得が行われる．このような過程において，目的関数を最適化する方策を導出する．

強化学習の MDP による枠組みは，図 2 のように表現される．エージェントの行動による環境への働きかけで状態と報酬を観測するという環境とエージェント間の相互作用を繰り返し，目的関数を最適化する方策を求める．

強化学習で扱う価値関数は以下の二つである．ここで，割引率 $\gamma \in [0, 1]$，$\pi \in \Pi_{\mathrm{S}}$(定常方策) とし，S_t, A_t, R_t はそれぞれ時刻 t における状態，行動，報酬の確率変数を表す．

図 2　環境とエージェント間の相互作用

(i) 状態価値関数

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}R_{t+1} \mid S_0 = s\right], \forall s \in \mathcal{S} \tag{1}$$

(ii) 状態行動価値関数

$$Q^{\pi}(s,a) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}R_{t+1} \mid S_0 = s, A_0 = a\right], \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{2}$$

本研究では，強化学習手法のうち，TD 学習 (Temporal Difference Learning) と呼ばれる手法を用いる．Q-Learning と SARSA は TD 学習に分類される手法で，この二つの手法で学習を行う．使用するアルゴリズムを以下に示す．Q-Learning と SARSA では，学習式が異なる．

Tabular TD(0) のアルゴリズム

1 行動価値 $Q(s,a)$ を初期化する．
2 各エピソードに対して以下を繰り返す．
　2.1 初期状態 s を決定する．
　2.2 初期状態 s での行動 a を決定する．
　2.3 エピソード中の各ステップに対して以下を繰り返す．
　　2.3.1 行動 a を実行し，報酬 r と次の状態 s' を観測する．
　　2.3.2 次の状態 s' での行動 a' を決定する．
　　2.3.3 学習式に従って行動価値 $Q(s,a)$ を更新する．
　　[Q-Learning の場合]
　　$Q^{\mathrm{new}}(s,a) = (1-\alpha)Q(s,a) + \alpha(r(s,a,s') + \gamma \max_{a' \in \mathcal{A}}\{Q(s',a')\})$
　　[SARSA の場合]
　　$Q^{\mathrm{new}}(s,a) = (1-\alpha)Q(s,a) + \alpha(r(s,a,s') + \gamma Q(s',a'))$
　　2.3.4 $s \leftarrow s', a \leftarrow a'$ とする．
　　2.3.5 状態 s がエピソードの終端状態なら繰り返しを終了する．

3　腐敗性を有する在庫管理モデルの定式化

腐敗性を有する在庫管理モデルについて，MDP と強化学習によって定式化する．それぞれを，マルコフ決定問題としての定式化，強化学習問題としての定式化とする．マルコフ決定問題としての定式化では，期限の異なる在庫をそれぞれ別々に扱うことで確定的な在庫の腐敗を表現する．強化学習問題としての定式化では，MDP 環境を用いること，時刻を状態の一つとし，有限期間を無限期間として扱えるようにすることによって，強化学習問題がマルコフ決定問題と同等な問題となるように定式化を行う．

3.1　マルコフ決定問題としての定式化

(i) 有限期間

$$T = \{1, 2, ..., N\},\ N < \infty \tag{3}$$

(ii) 有限状態集合

$$\mathcal{S} = \left\{ (s_0, s_1, \ldots, s_{\tau-2}) \mid 0 \leq \sum_{k=0}^{\tau-2} s_k \leq M,\ s_0, s_1, \ldots, s_{\tau-2} \in \mathbb{N} \right\} \tag{4}$$

(iii) 有限行動集合

$$\mathcal{A} = \{0, 1, 2, \ldots, M\} \tag{5}$$

$$\mathcal{A}_{(s_0, s_1, \ldots, s_{\tau-2})} = \left\{ 0, 1, 2, \ldots, M - \sum_{k=0}^{\tau-2} s_k \right\} \subset \mathcal{A} \tag{6}$$

(iv) 決定規則，方策

$$\delta_t : \mathcal{S} \to \mathcal{A}, t = 1, 2, \ldots, N-1 \tag{7}$$

$$\pi = (\delta_1, \delta_2, \ldots, \delta_{N-1}) \tag{8}$$

(v) 需要分布 (D, d: それぞれ需要の確率変数と実現値)

$$p_d = P\{D = d\},\ d = 0, 1, 2, \ldots \tag{9}$$

(vi) 推移確率 ($\tau = 2$)

$$p(s_0' \mid s_0, a) = \begin{cases} 0 & \text{if } s_0 + a < s_0' \le M \text{ or } s_0' > a \\ \displaystyle\sum_{d=0}^{s_0} p_d & \text{if } 0 < s_0' \le s_0 + a \le M, s_0' = a \\ p_{s_0+a-s_0'} & \text{if } 0 < s_0' \le s_0 + a \le M, s_0' \ne a \\ \displaystyle\sum_{d=s_0+a}^{\infty} p_d & \text{if } s_0 + a \le M, s_0' = 0, a > 0 \\ 1 & \text{if } s_0 + a \le M, s_0' = 0, a = 0 \end{cases} \tag{10}$$

(vii) 推移確率 ($\tau \ge 3,\ m = \sum_{k=0}^{\tau-2} s_k + a,\ m' = m - \sum_{k=0}^{\tau-2} s_k'$)

$$\begin{aligned} & p((s_0', s_1', \ldots, s_{\tau-2}') \mid (s_0, s_1, \ldots, s_{\tau-2}), a) \\ & = \begin{cases} 0 & \text{if } \sum_{k=0}^{\tau-2} s_k + a < \sum_{k=0}^{\tau-2} s_k' \le M, \\ & \quad \text{or } (s_{\tau-3}' > s_{\tau-2} \text{ for some } \tau) \text{ or } s_{\tau-2}' > a \\ & \quad \text{or } \left(\sum_{k=0}^{\tau-4} s_k' > 0, s_{\tau-2} > s_{\tau-3}' \text{ for some } \tau\right) \\ & \quad \text{or } \left(\sum_{k=0}^{\tau-3} s_k' > 0, a > s_{\tau-2}'\right) \\ \displaystyle\sum_{d=0}^{s_0} p_d & \text{if } 0 < \sum_{k=0}^{\tau-2} s_k' \le \sum_{k=0}^{\tau-2} s_k + a \le M, \\ & \quad s_0' = s_1, \ldots, s_{\tau-3}' = s_{\tau-2}, s_{\tau-2}' = a \\ p_{m'} & \text{if } 0 < \sum_{k=0}^{\tau-2} s_k' \le \sum_{k=0}^{\tau-2} s_k + a \le M, \\ & \quad s_0' \ne s_1 \text{ or } \ldots \text{ or } s_{\tau-3}' \ne s_{\tau-2} \text{ or } s_{\tau-2}' \ne a \\ \displaystyle\sum_{d=m}^{\infty} p_d & \text{if } \sum_{k=0}^{\tau-2} s_k + a \le M, \sum_{k=0}^{\tau-2} s_k' = 0, \sum_{k=1}^{\tau-2} s_k + a > 0 \\ 1 & \text{if } \sum_{k=0}^{\tau-2} s_k + a \le M, \sum_{k=0}^{\tau-2} s_k' = 0, \sum_{k=1}^{\tau-2} s_k + a = 0 \end{cases} \end{aligned} \tag{11}$$

(viii) 期待収益 (b: 単位当たり収益)

$$F(m) = \sum_{d=0}^{m-1} bdp_d + bm \sum_{d=m}^{\infty} p_d \tag{12}$$

(ix) 発注費 (c_1: 固定発注費，c_2: 単位当たり変動発注費)

$$O(a) = \begin{cases} c_1 + c_2 a & \text{if } a > 0 \\ 0 & \text{if } a = 0 \end{cases} \tag{13}$$

(x) 在庫保管費 (c_3: 単位当たり在庫保管費)

$$H(m) = c_3 m \tag{14}$$

(xi) 期待廃棄費 (c_4: 単位当たり廃棄費)

$$W(s_0) = \sum_{d=0}^{s_0-1} c_4 \max\{s_0 - d, 0\} p_d + c_4 s_0 \sum_{d=s_0}^{\infty} p_d \tag{15}$$

(xii) 報酬関数

$$r_t((s_0, s_1, \ldots, s_{\tau-2}), a) = F(m) - O(a) - H(m) - W(s_0),\ t = 1, 2, \ldots, N-1 \quad (16)$$

(xiii) 末期在庫価値

$$r_N(s_0, s_1, \ldots, s_{\tau-2}) = g(s_0, s_1, \ldots, s_{\tau-2}),\ t = N \quad (17)$$

目的関数を

$$v_N^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{N-1} r_t(S_t, A_t) + r_N(S_N) \,|\, S_0 = s\right], s \in \mathcal{S} \quad (18)$$

とする．ここで，$v_N^{\pi}(s)$ を初期状態 s で方策 π に従ったときの期待報酬和，Π_{MD} を確定的マルコフ方策集合とし，S_t, A_t はそれぞれ時刻 t における状態，行動の確率変数を表す．また，最適化問題を

$$\begin{array}{ll} \text{maximize} & v_N^{\pi}(s_0, s_1, \ldots, s_{\tau-2}) \\ \text{subject to} & \pi \in \Pi_{\mathrm{MD}} \end{array}$$

とする．

3.2 強化学習問題としての定式化

(i) 有限状態集合

$$\mathcal{S} = \left\{ (t, s_0, s_1, \ldots, s_{\tau-2}) \,|\, t \in \{1, 2, ..., N\}, N < \infty, \right. \\ \left. 0 \le \sum_{k=0}^{\tau-2} s_k \le M, s_0, s_1, \ldots, s_{\tau-2} \in \mathbb{N} \right\} \quad (19)$$

(ii) 有限行動集合

$$\mathcal{A} = \{0, 1, 2, \ldots, M\} \quad (20)$$

$$\mathcal{A}_{(s_0, s_1, \ldots, s_{\tau-2})} = \left\{ 0, 1, 2, \ldots, M - \sum_{k=0}^{\tau-2} s_k \right\} \subset \mathcal{A} \quad (21)$$

(iii) 推移確率

TD 学習では，推移確率は用いずに次の状態をシミュレーションによって観測する．

(iv) 収益 (b: 単位当たり収益，d: 需要，$m = \sum_{k=0}^{\tau-2} s_k + a$)

$$F(m) = b \min\{m, d\} \quad (22)$$

(v) 発注費 (c_1: 固定発注費，c_2: 単位当たり変動発注費)

$$O(a) = \begin{cases} c_1 + c_2 a & \text{if } a > 0 \\ 0 & \text{if } a = 0 \end{cases} \tag{23}$$

(vi) 在庫保管費 (c_3: 単位当たり在庫保管費)

$$H(m) = c_3 m \tag{24}$$

(vii) 廃棄費 (c_4: 単位当たり廃棄費，d: 需要)

$$W(s_0) = c_4 \max\{s_0 - d, 0\} \tag{25}$$

(viii) 報酬関数

TD 学習では，需要はシミュレーションによって観測される．この報酬以外に推移できない行動を選んだ場合にペナルティを設定する．

$$r((t, s_0, s_1, \ldots, s_{\tau-2}), a) = F(m) - O(a) - H(m) - W(s_0) \tag{26}$$

目的関数を

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} r_t(S_t, A_t) \,|\, S_0 = s \right], \forall s \in \mathcal{S} \tag{27}$$

とする．ここで，$V^\pi(s)$ を初期状態 s で方策 π に従ったときの期待報酬和，Π_{S} を定常方策集合とし，S_t, A_t はそれぞれ時刻 t における状態，行動，報酬の確率変数を表す．また，最適化問題を

$$\begin{array}{ll} \text{maximize} & V^\pi(t, s_0, s_1, \ldots, s_{\tau-2}) \\ \text{subject to} & \pi \in \Pi_{\mathrm{S}} \end{array}$$

とする．

4　数値実験による TD 学習の実装の比較

数値実験を行い，強化学習問題がマルコフ決定問題と同様の結果が得られるか検証を行う．また，異なる実装方法やパラメータについて，学習中の状態価値，状態行動価値からどの場合が適切か比較する．

ここでは，以下のパラメータを使用する．

(1) 保管期間 $\tau = 3$
(2) 倉庫容量 $M = 10$
(3) 有限期間 $N = 10$

(4) 単位当たり収益 $b = 10$
(5) 固定発注費 $c_1 = 4$
(6) 単位当たり変動発注費 $c_2 = 2$
(7) 単位当たり在庫保管費 $c_3 = 2$
(8) 単位当たり廃棄費 $c_4 = 2$
(9) 推移不可能な行動に対するペナルティ: -1000
(10) 末期在庫価値 $g(s_0, s_1) = 0$
(11) 需要は平均 $\lambda = 1$ のポワソン分布に従う.

学習率 $\alpha = 0.01$ の Q-Learning で，100000 エピソード学習した．行動選択には ϵ-Greedy を用い，$\epsilon = 0.01$ とした．行動選択の方策として用いる ϵ-Greedy は，$[0, 1]$ の一様乱数を発生させ，ϵ 以上なら価値が最大の行動を選ぶ (利用)，ϵ 未満ならランダムに行動を選ぶ (探索) という方策である．行動価値に関して Greedy な行動方策だと局所解に陥る可能性があるため，一定の確率 ϵ でランダムに行動を選び，他の解を探索する．

得られた方策と価値の一部を表 1 に示す．このとき MDP で求められた最適方策は表 2 のようになっている．100000 エピソード時点では，方策は一致しているが，まだ最適な価値にはなっていないことが確認できる．

表 1　Q-Learning で得られる方策と期待報酬和

状態	行動	価値 (期待報酬和)
(1, 0, 0)	2	1.199
(1, 0, 1)	0	4.378
(1, 0, 2)	0	5.076
(1, 0, 3)	0	4.551
(1, 0, 4)	0	0.805
(1, 0, 5)	0	-4.570
(1, 0, 6)	0	-10.494
(1, 0, 7)	0	-13.815
(1, 0, 8)	0	-20.736
(1, 0, 9)	0	-25.877
(1, 0, 10)	0	-33.609

表 2　MDP から得られる最適方策と期待報酬和

$t = 1$ の状態	行動	価値 (期待報酬和)
(0, 0)	2	4.651
(0, 1)	0	9.014
(0, 2)	0	10.066
(0, 3)	0	8.054
(0, 4)	0	4.064
(0, 5)	0	-1.043
(0, 6)	0	-6.695
(0, 7)	0	-12.579
(0, 8)	0	-18.544
(0, 9)	0	-24.533
(0, 10)	0	-30.531

実装を 2 パターンに分け，学習中の価値の値を比較する．実装 A では，行動を $\mathcal{A}$ から選択し，推移不可能な行動にペナルティを課す．実装 B では，行動を $\mathcal{A}_{(s_0, s_1, \ldots, s_{\tau-2})}$ から選択し，推移不可能な行動は選ばれないためペナルティは無い．

実装 A の Q-Learning ($\epsilon = 0.01$)，SARSA ($\epsilon = 0.01$) で，1000000 エピソード学習し

た結果を図 3 に示す．左図は学習中の 1 エピソードごとの状態行動価値の最大更新量を表し，右図は状態 (1,0,0) における学習中の 1 エピソードごとの状態価値を表す．図 4，図 5，図 6 も同様で，それぞれ実装 A の Q-Learning ($\epsilon = 1$)，SARSA ($\epsilon = 0.01$)，実装 B の Q-Learning ($\epsilon = 0.01$)，SARSA ($\epsilon = 0.01$)，実装 B の Q-Learning ($\epsilon = 1$)，SARSA ($\epsilon = 0.01$) のときの結果を示す．

実装 A の場合，SARSA では学習の途中で価値が減少に転じる．価値が負の値まで減少していることから，設定したペナルティが適切に機能していない可能性が考えられる．実装 B の場合，実装 A で見られた SARSA での価値の減少がなく，最適価値 4.651 に近い値を取っている．実装 A，実装 B に共通して Q-Learning は $\epsilon = 1$ の場合に収束すべき最適価値を超えて価値が増加している．$\epsilon = 1$ よりも $\epsilon = 0.01$ のほうがこのモデルに適していると考えられる．

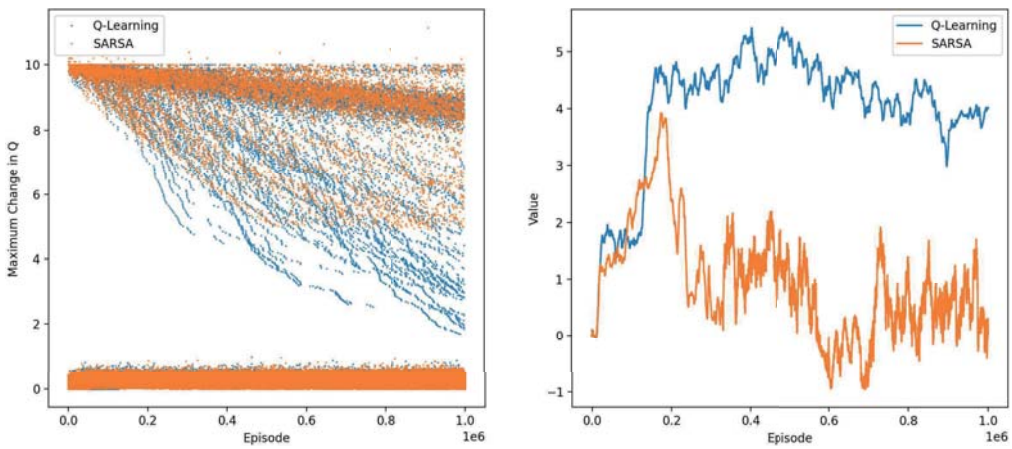

図 3　実装 A，Q-Learning ($\epsilon = 0.01$)，SARSA ($\epsilon = 0.01$)：状態行動価値の最大更新量 (左)，状態 (1,0,0) の状態価値 (右)

5　結論

腐敗性在庫の最適化問題について，MDP と強化学習という二つの観点から定式化と数値実験を行い，状態行動価値などから TD 学習における各実装の比較をした．学習で得られる状態価値は MDP の最適な状態価値に近づくものの，収束には至っていないことがわかった．学習式や学習中の行動選択方策のパラメータなどの条件によっては，最適価値の値を超える，学習の途中で価値が減少に転じるなど最適価値の値から遠ざかる場合がある．そのため，学習中の行動選択などの実装に問題点があることによって適切な学習が行われていない可能性がある．

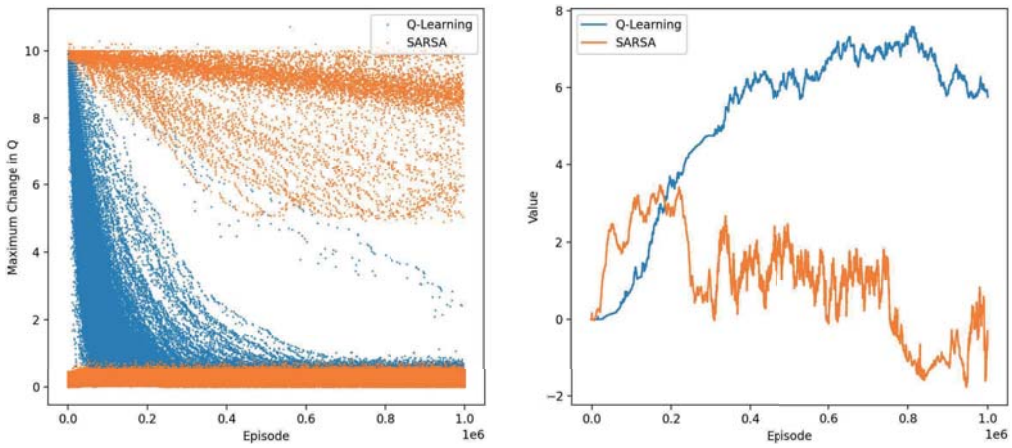

図 4　実装 A，Q-Learning ($\epsilon = 1$)，SARSA ($\epsilon = 0.01$): 状態行動価値の最大更新量 (左)，状態 (1,0,0) の状態価値 (右)

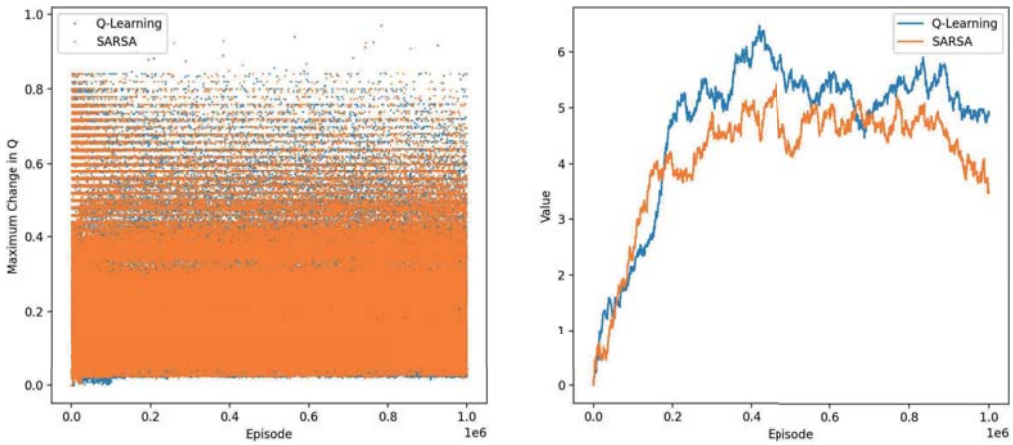

図 5　実装 B，Q-Learning ($\epsilon = 0.01$)，SARSA ($\epsilon = 0.01$): 状態行動価値の最大更新量 (左)，状態 (1,0,0) の状態価値 (右)

課題として，適切な学習が行われなかった詳細な原因を検証すること，計算時間の比較のために学習終了条件について検討することが挙げられる．

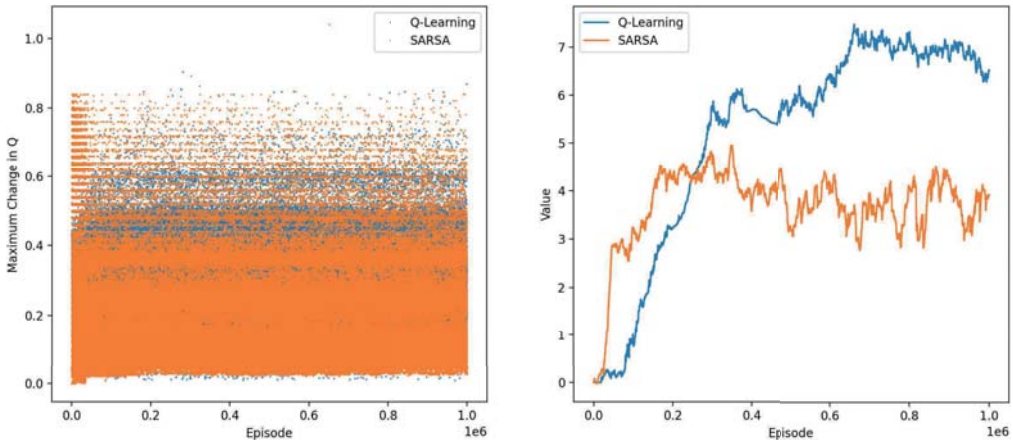

図 6　実装 B，Q-Learning ($\epsilon = 1$)，SARSA ($\epsilon = 0.01$): 状態行動価値の最大更新量 (左)，状態 (1,0,0) の状態価値 (右)

参考文献

[1] Yasuo Adachi, Toyokazu Nose, and Sennosuke Kuriyama. Optimal inventory control policy subject to different selling prices of perishable commodities. *International Journal of Production Economics*, Vol. 60, pp. 389–394, 1999.

[2] Guilherme O Ferreira, Edilson F Arruda, and Lino G Marujo. Inventory management of perishable items in long-term humanitarian operations using markov decision processes. *International Journal of Disaster Risk Reduction*, Vol. 31, pp. 460–469, 2018.

[3] Ahmet Kara and Ibrahim Dogan. Reinforcement learning approaches for specifying ordering policies of perishable inventory systems. *Expert Systems with Applications*, Vol. 91, pp. 150–158, 2018.

[4] Martin L Puterman. *Markov Decision Processes Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.

[5] Douglas J White. *Markov Decision Processes*. Wiley, 1993.

[6] Frederick S Hillier and Gerald J Lieberman. *INTRODUCTION TO OPERATIONS RESEARCH EIGHTH EDITION*. McGraw-Hill Publishing Company, 2005.

[7] Richard S Sutton and Andrew G Barto. *Reinforcement Learning, second edition: An Introduction*. Bradford Books, 2018.

[8] Csaba Szepesvari. *Algorithms for Reinforcement Learning (Synthesis Lectures on*

Artificial Intelligence and Machine Learning). Morgan and Claypool Publishers, 2010.

[9] 牧野貴樹. これからの強化学習. 森北出版, 2016.

[10] 曽我部東馬. 強化学習アルゴリズム入門: 「平均」からはじめる基礎と応用. オーム社, 2019.