二人零和マルコフゲームにおける オフ方策評価のためのQ学習

阿部 拳之[1,a)]　金子 雄祐[1,b)]

概要：オフ方策評価は，ある方策から取得した履歴データを使用してオフラインで新しい方策を評価する問題である．本研究では，二人零和マルコフゲームにおけるオフ方策評価のために，新しい Q 学習アルゴリズムである *Best Response (BR) Q-learning* を提案する．BR Q-learning は，二人零和マルコフゲームにおける履歴データを用いて，与えられた戦略に対する最適反応戦略の状態行動価値関数を推定する．本論文では，BR Q-learning によって更新される状態行動価値関数が，最適反応戦略の状態行動価値関数へと確率 1 で収束することを証明する. さらに，BR Q-learning を用いることで，与えられた戦略プロファイルの exploitability を推定する手法を提案し，推定された exploitability が, 真の exploitability に確率 1 で収束することを示す. また，実験によって BR Q-learning の有効性を確認する．

Q-Learning for Off-Policy Evaluation in Two-Player Zero-Sum Markov Games

Kenshi Abe[1,a)]　Yusuke Kaneko[1,b)]

Abstract: Off-policy evaluation (OPE) is the problem of evaluating new policies using historical data obtained from a different policy. In this study, we propose a novel Q-learning algorithm, called *Best Response (BR) Q-learning*, for OPE in two-player zero-sum Markov games. BR Q-learning estimates the state-action value of the best response to the given strategy. We prove that BR Q-learning converges the state-value of the best response with probability one. Further, we propose the novel off-policy estimator for exploitability using BR Q-learning. Then, we show that the estimated exploitability converges to the true exploitability with probability one. Finally, we demonstrate the effectiveness and performance of BR Q-learning through experiments.

1. はじめに

オフ方策評価は，ある方策から取得した履歴データを使用してオフラインで新しい方策を評価する問題である．オンラインで方策を評価することは医療 [11] や教育 [10] などにおいてはリスクが伴うため，オフ方策評価は近年盛んに研究が行われている [2], [6], [7], [9], [13], [14]．これらのオフ方策評価に関する研究のほとんどは，プレイヤが一人しか存在しない環境におけるオフ方策評価に焦点を当てており，単一の方策の期待報酬和を推定するための推定量を提案している．

阿部ら [1] は，プレイヤが複数人存在するマルチエージェント環境におけるオフ方策評価手法を提案した．この手法では，二人零和マルコフゲームにおける戦略プロファイルを履歴データから評価することを可能にしている．具体的には，期待報酬和を推定する代わりに，二人零和ゲームにおいて戦略プロファイルがナッシュ均衡にどれだけ近いかを判断するための指標である exploitability を推定することで戦略プロファイルを評価する. この手法では, まず与えられた戦略プロファイルの期待報酬和を推定する推定量を構築する．次に，その期待報酬和の推定量によって推定された期待報酬和を最大化する戦略を特定の方策クラスから選択することで, 最適反応戦略を推定する．しかし，方策クラスが複雑な場合，最適反応戦略の推定は時間・空間計算量が大きくなることが想定される．そのため, 阿部らの手

[1] 株式会社サイバーエージェント
CyberAgent, Inc.
[a)] abe_kenshi@cyberagent.co.jp
[b)] kaneko_yusuke@cyberagent.co.jp

法では，結果として exploitability の推定に大きな計算コストがかかることが考えられる．

そこで，本研究では，二人零和マルコフゲームにおける履歴データを用いて最適反応戦略を効率的に計算するための新しい Q 学習 [15] アルゴリズムである *Best Response (BR) Q-learning* を提案する．BR Q-learning は，評価対象の戦略と履歴データをサンプルした戦略の不一致を補正するために，importance weight を用いる．本論文では，BR Q-learning によって更新される状態行動価値関数が，最適反応戦略の状態行動価値関数へと確率 1 で収束することを証明する. さらに，BR Q-learning を用いることで，与えられた戦略プロファイルの exploitability を推定する手法を提案し，推定された exploitability が, 真の exploitability に確率 1 で収束することを示す. また，実験によって BR Q-learning の有効性を確認する．

2. 問題設定

2.1 二人零和マルコフゲーム

二人零和マルコフゲームは、$\langle \mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, P_T, P_R, \gamma \rangle$ で定義されるゲームである．二人零和マルコフゲームは、以下の要素から構成されるゲームである．

- $\mathcal{S}$：有限の状態空間
- $\mathcal{A}_i$：プレイヤー i ($\in \{1,2\}$) の有限の行動空間
- $P_T : S \times \mathcal{A}_1 \times \mathcal{A}_2 \times S \to [0,1]$：状態遷移関数
- $P_R : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times \mathbb{R} \to [0,1]$：報酬分布関数
- $\gamma \in [0,1)$：割引率

ここで，$R : S \times \mathcal{A}_1 \times \mathcal{A}_2 \to \mathbb{R}$ を期待報酬関数として定義する．また，時刻 $t = 1,2,\cdots$ に対して，$r_{1,t} = r_t \sim P_R(s_t, a_t^1, a_t^2)$ を行動 a_t^1 と a_t^2 が状態 s_t で取られた場合のプレイヤー 1 の報酬とし，$r_{2,t} = -r_t$ をプレイヤー 2 の報酬とする．$\pi_i : S \times A_i \to [0,1]$ をプレイヤー $i \in \{1,2\}$ の定常なマルコフ方策とし，$\pi = (\pi_1, \pi_2)$ を戦略プロファイルもしくは**方策プロファイル**と呼ぶ. さらに，Σ_i をプレイヤー i の方策の集合として定義する．

ある方策プロファイル (π_1, π_2) が与えられたとき, 状態 s における各プレイヤーの状態価値関数は次のように書ける：

$$V_1^{\pi_1,\pi_2}(s) = \mathbb{E}_{\pi_1,\pi_2}[\sum_{t=1}^{\infty} \gamma^t r_t | s], \; V_2^{\pi_1,\pi_2}(s) = -V_1^{\pi_1,\pi_2}(s).$$

簡単のため，$\mathbb{E}_{a_i \sim \pi_i(\cdot|s), s' \sim P_T(\cdot|s,a_{-i},a_i)}[\cdot|s,a]$ を $\mathbb{E}_{\pi_i}[\cdot]$ と表記すると，行動 a^1，a^2 が状態 s で取られた際の状態行動価値関数は，次のように書ける：

$$Q_1^{\pi_1,\pi_2}(s,a) = \mathbb{E}_{\pi_2}[R(s,a,a_2) + \gamma V_1^{\pi_1,\pi_2}(s')],$$
$$Q_2^{\pi_1,\pi_2}(s,a) = \mathbb{E}_{\pi_1}[-R(s,a_1,a) + \gamma V_2^{\pi_1,\pi_2}(s')].$$

最適反応 $\mathcal{B}_i(\pi_{-i}) = \arg\max_{\pi_i' \in \Sigma_i} \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} V_i^{\pi_i', \pi_{-i}}(s)$ は，π_{-i} に対して最適なプレイヤー i の方策のことである．ただし，π_{-i} は i 以外の方策とする．最後に，**最適反応価値**を $\delta_i(\pi_{-i}) = \max_{\pi_i' \in \Sigma_i} \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} V_i^{\pi_i', \pi_{-i}}(s)$ で定義する．

Algorithm 1 Off-Policy Exploitability Estimation Method using BR Q-learning

Initialize Q value $Q_{1,1}, Q_{2,1}$.
for $t = 1, \cdots, T$ **do**
  Observe state s_t.
  Select actions a_t^1, a_t^2 according to π_1^b, π_2^b.
  Receive reward r_t.
  $y_{1,t} = \frac{\pi_2^e(a_t^2|s_t)}{\pi_2^b(a_t^2|s_t)} (r_t + \gamma \max_a Q_{1,t}(s_{t+1}, a))$
  $Q_{1,t+1}(s_t, a_t^1) = Q_{1,t}(s_t, a_t^1) + \alpha_{1,t}(s_t, a_t^1)\left(y_{1,t} - Q_{1,t}(s_t, a_t^1)\right)$
  $y_{2,t} = \frac{\pi_1^e(a_t^1|s_t)}{\pi_1^b(a_t^1|s_t)} (-r_t + \gamma \max_a Q_{2,t}(s_{t+1}, a))$
  $Q_{2,t+1}(s_t, a_t^2) = Q_{2,t}(s_t, a_t^2) + \alpha_{2,t}(s_t, a_t^2)\left(y_{2,t} - Q_{2,t}(s_t, a_t^2)\right)$
end for
return $\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \max_{a \in \mathcal{A}_1} Q_{1,T+1}(s,a) + \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \max_{a \in \mathcal{A}_2} Q_{2,T+1}(s,a)$

2.2 二人零和マルコフゲームにおけるオフ方策評価

本研究では, 文献 [4] と同様に, 評価対象の方策プロファイルの推定値の更新のために新しいサンプル $(s_t, a_t^1, a_t^2, r_t, s_{t+1})$ が異なる方策プロファイルによって継続的に得られる設定におけるオフ方策評価問題を対象とする．本研究では，これらのサンプルを**履歴データ**と呼び，履歴データは時刻に依存しない方策プロファイル $\pi^b = (\pi_1^b, \pi_2^b)$ からサンプルされると仮定する．また，この方策プロファイルを**行動方策プロファイル**と呼ぶ.

二人零和マルコフゲームにおけるオフ方策評価問題では，行動方策プロファイルによって得られた履歴データから，与えられた**評価方策プロファイル** $\pi^e = (\pi_1^e, \pi_2^e)$ の評価を行うことを目的とする．本研究では，文献 [1] に従い，評価方策プロファイルの期待報酬和を推定する代わりに，以下で定義される **exploitability** を推定する：

$$v^{\exp}(\pi_1, \pi_2) = \max_{\pi_1' \in \Sigma_1} \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} V_1^{\pi_1', \pi_2}(s) + \max_{\pi_2' \in \Sigma_2} \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} V_2^{\pi_1, \pi_2'}(s),$$
$$= \delta_1(\pi_2) + \delta_2(\pi_1).$$

$v^{\exp}(\pi_1, \pi_2)$ を計算するには，最適反応価値 $\delta_1(\pi_2), \delta_2(\pi_1)$ をそれぞれ計算する必要がある．

3. Best Response Q-Learning

本章では，方策 π_{-i}^e に対する最適反応価値 $\delta_i(\pi_{-i}^e)$ を推定するための Q 学習アルゴリズムである、**BR Q-learning** を提案する．BR Q-learning は，最適反応に対する状態行動価値 $Q_i^{\mathcal{B}_i(\pi_{-i}^e), \pi_{-i}^e}(s,a)$ を近似することで，$\delta_i(\pi_{-i}^e)$ を推定する．

履歴データから最適反応を計算するには，行動方策 π_{-i}^b が評価方策 π_{-i}^e と異なる，共変量シフト [3] 下での学習を行うことが重要になる．この π_{-i}^b と π_{-i}^e の不一致を補正するために，BR Q-learning は importance weighting $\rho_t^{-i} = \frac{\pi_{-i}^e(a_t^{-i}|s_t)}{\pi_{-i}^b(a_t^{-i}|s_t)}$ を用いて状態行動価値の更新を行う．具体的には，BR Q-

learning は π_{-i} に対する状態行動価値 $Q_{i,t}(s_t, a_t^1)$ を次のように更新する．

$$y_{i,t} = \rho_t^{-i}\left(r_{i,t} + \gamma \max_a Q_{i,t}(s_{t+1}, a)\right),$$
$$Q_{i,t+1}(s_t, a_t^i) = Q_{i,t}(s_t, a_t^i) + \alpha_{i,t}(s_t, a_t^i)\left(y_{i,t} - Q_{i,t}(s_t, a_t^i)\right). \quad (1)$$

ただし，$\alpha_{i,t}(s,a) \in [0,1)$ は時刻 t における，状態 s と行動 a に対する学習率である．Algorithm 1 に，BR Q-learning を用いた exploitability 推定手法を示す．

以降では，BR Q-learning，および exploitability 推定手法の収束性を証明する．まず，BR Q-learning によって更新される状態行動価値関数が最適な状態行動価値関数へと収束保証を与える．

定理 1. 以下の条件を仮定する：
(1) 状態空間と行動空間が有限，
(2) $\forall (s,a) \in \mathcal{S}\times\mathcal{A}_i$, $\sum_t \alpha_{i,t}(s,a) = \infty$, $\sum_t \alpha_{i,t}^2(s,a) < \infty$, .
(3) $\forall\, t$, $|r_t| \le R_{\max}, 0 \le \rho_t^{-i} \le \eta$.
このとき，すべての $(s,a) \in \mathcal{S}\times\mathcal{A}_i$ に対して，*BR Q-learning* によって更新される状態行動価値関数 $Q_{i,t}(s,a)$ は最適な状態行動関数 $Q_i^{\mathcal{B}_i(\pi^e_{-i}),\pi^e_{-i}}(s,a)$ に確率 *1* で収束する．

Proof. まず，確率過程の収束に関する以下の定理 [5] を導入する．

定理 2. 確率過程 $\Delta_{t+1}(s) = (1-\alpha_t(s))\Delta_t(s) + \beta_t(s)F_t(s)$ が以下の仮定を満たすとき，確率 *1* で *0* へと収束する*:*
(1) 状態空間 $\mathcal{S}$ が有限，
(2) $\forall s \in \mathcal{S}$, $\sum_t \alpha_t(s) = \infty$, $\sum_t \alpha_t^2(s) < \infty$, $\sum_t \beta_t(s) = \infty$, $\sum_t \beta_t^2(s) < \infty$.
(3) $\forall\, s \in \mathcal{S}$, $\mathbb{E}[\beta_t(s)|\chi_t] \le \mathbb{E}[\alpha_t(s)|\chi_t]$.
(4) $\gamma \in (0,1)$ に対して，$\|\mathbb{E}[F_t(s)|\chi_t]\|_W \le \gamma\|\Delta_t\|_W$.
(5) ある定数 C が存在して，$\mathbb{V}[F_t(s)|\chi_t] \le C(1+\|\Delta_t\|_W)^2$.
ここで，$\chi_t = \{\Delta_t, \Delta_{t-1}, \cdots, F_{t-1}, \cdots, \alpha_{t-1}, \cdots, \beta_{t-1}, \cdots\}$ を時刻 t までの履歴とする．また，$\|\cdot\|_W$ はある重み W に対する重み付き最大値ノルム（*weighed maximum norm*）とする．

簡単のため $Q_i^*(s_t, a_t^i) = Q_i^{\pi_i^*,\pi^e_{-i}}(s_t, a_t^i)$ と表記すると，式 (1) は以下のように書き表せる：

$$Q_{i,t+1}(s_t, a_t^i) = (1-\alpha_t(s_t, a_t^i))Q_{i,t}(s_t, a_t^i) + \alpha_t(s_t, a_t^i)\rho_t^{-i}\left(r_t + \gamma \max_a Q_{i,t}(s_{t+1}, a)\right).$$

この式の両辺から $Q_i^*(s_t, a_t^i)$ を引き，さらに $\Delta_{i,t}(s_t, a_t^i) = Q_{i,t}(s_t, a_t^i) - Q_i^*(s_t, a_t^i)$ と定義すると，以下の式を得る：

$$\Delta_{i,t+1}(s_t, a_t^i) = (1-\alpha_t(s_t, a_t^i))\Delta_{i,t}(s_t, a_t^i) + \alpha_t(s_t, a_t^i)\left(\rho_t^{-i}\left(r_t + \gamma \max_a Q_{i,t}(s_{t+1}, a)\right) - Q_i^*(s_t, a_t^i)\right).$$

ここで，

$$F_t(s, a^i) = \rho_t^{-i}\left(r_t + \gamma \max_a Q_{i,t}(s', a^i)\right) - Q_i^*(s, a^i),$$

と定義することで，BR Q-learning の更新式 (1) を，定理 2 において $\beta_t(s,a^i) = \alpha_t(s,a^i)$ とした場合の確率過程 $\Delta_{i,t}$ として書くことができる．したがって，以降では，BR Q-learning による状態行動価値関数の更新が定理 2 の仮定 (4) および (5) を満たすことを示す．

まず，仮定 (4) である $\|\mathbb{E}[F_t(s,a^i)|\chi_t]\|_W \le \gamma\|\Delta_{i,t}\|_W$ を示す．$\|\mathbb{E}[F_t(s,a^i)|\chi_t]\|_W$ に関して，以下のように展開できる：

$$\begin{aligned}
&\mathbb{E}[F_t(s,a^i)|\chi_t] \\
=&\mathbb{E}_{\pi^b_{-i}}\left[\frac{\pi^e_{-i}(a^{-i}|s)}{\pi^b_{-i}(a^{-i}|s)}\left(r_t + \gamma \max_a Q_{i,t}(s',a)\right) - Q_i^*(s,a^i)|s,a^i\right] \\
=&\mathbb{E}_{\pi^e_{-i}}\left[r_t + \gamma \max_a Q_{i,t}(s',a)|s,a^i\right] - Q_i^*(s,a^i) \\
=&\mathbb{E}_{\pi^e_{-i}}\left[r_t + \gamma \max_a Q_{i,t}(s',a) - R(s,a^i,a^{-i}) + \gamma \max_a Q_i^*(s',a)|s,a^i\right] \\
=&\gamma\mathbb{E}_{\pi^e_{-i}}\left[\max_a Q_{i,t}(s',a) - \max_a Q_i^*(s',a)|s,a^i\right].
\end{aligned}$$

したがって，以下の式を得る．

$$\begin{aligned}
&\max_{a^i}\left|\mathbb{E}[F_t(s,a^i)|\chi_t]\right| \\
&= \gamma \max_{a^i}\left|\mathbb{E}_{\pi^e_{-i}}\left[\max_a Q_{i,t}(s',a) - \max_a Q_i^*(s',a)|s,a^i\right]\right| \\
&\le \gamma \max_{a^i}\mathbb{E}_{\pi^e_{-i}}\left[\max_a\left|Q_{i,t}(s',a) - Q_i^*(s',a)\right||s,a^i\right] \\
&\le \gamma \max_{s,a^i}\left|Q_{i,t}(s,a^i) - Q_i^*(s,a^i)\right| = \gamma\left|\Delta_{i,t}\right|_\infty.
\end{aligned}$$

つぎに，$\mathbb{V}[F_t(s,a^i)|\chi_t] \le C(1+\|\Delta_{i,t}\|_W)^2$ を示す．$\rho^{-i} = \frac{\pi^e_{-i}(a^{-i}|s)}{\pi^b_{-i}(a^{-i}|s)}$，$\mathbb{E}_e[\cdot] = \mathbb{E}_{\pi^e_{-i}}[\cdot|s,a^i]$ と置くと，以下が成り立つ：

$$\begin{aligned}
&\mathbb{V}[F_t(s,a^i)|\chi_t] \\
\le&\mathbb{E}\left[\left(\rho^{-i}\left(r_t + \gamma \max_a Q_{i,t}(s',a)\right) - Q_i^*(s,a^i)\right)^2|s,a^i\right] \\
=&\mathbb{E}_e\Big[\Big(\rho^{-i}r_t - \mathbb{E}_e[r_t] \\
&\quad + \gamma\left(\rho^{-i}\max_a Q_{i,t}(s',a) - \mathbb{E}_e\left[\max_a Q_i^*(s',a)\right]\right)\Big)^2\Big] \\
\le&\mathbb{E}_e\left[\left(\rho^{-i}r_t - \mathbb{E}_e[r_t]\right)^2\right] \\
&+\gamma^2\mathbb{E}_e\left[\left(\rho^{-i}\max_a Q_{i,t}(s',a) - \mathbb{E}_e\left[\max_a Q_i^*(s',a)\right]\right)^2\right] \\
&+2\gamma\max_{s,a^{-i}}\left|\rho^{-i}r_t - \mathbb{E}_e[r_t]\right| \\
&\quad\cdot\left(\max_{s',a^{-i}}\left|\rho^{-i}\max_a Q_{i,t}(s',a) - \mathbb{E}_e\left[\max_a Q_i^*(s',a)\right]\right|\right) \\
\le&4\eta^2R_{\max}^2 + \gamma^2\mathbb{E}_e\left[\left(\rho^{-i}\max_a Q_{i,t}(s',a) - \mathbb{E}_e\left[\max_a Q_i^*(s',a)\right]\right)^2\right] \\
&+4\gamma\eta R_{\max}\left(\max_{s',a^{-i}}\left|\rho^{-i}\max_a Q_{i,t}(s',a) - \rho^{-i}\max_a Q_i^*(s',a)\right|\right) \\
&+4\gamma\eta R_{\max}\left(\max_{s',a^{-i}}\left|\mathbb{E}_e\left[\max_a Q_i^*(s',a)\right] - \rho^{-i}\max_a Q_i^*(s',a)\right|\right).
\end{aligned}$$

ここで，最後の不等式の導出は仮定 (3) を用いた．さらに，仮定 (3) より：

$$
\begin{aligned}
&\mathbb{V}[F_t(s,a^i)|\chi_t] \\
&\leq \frac{8\eta^2 R_{\max}^2}{1-\gamma} + \gamma^2 \mathbb{E}_e\left[\left(\rho^{-i}\max_a Q_{i,t}(s',a) - \mathbb{E}_e\left[\max_a Q_i^*(s',a)\right]\right)^2\right] \\
&+4\gamma\eta R_{\max}\left(\max_{s',a^{-i}}\left|\rho^{-i}\max_a Q_{i,t}(s',a) - \rho^{-i}\max_a Q_i^*(s',a)\right|\right) \\
&\leq \frac{8\eta^2 R_{\max}^2}{1-\gamma} + \gamma^2 \mathbb{E}_e\left[\left(\rho^{-i}\max_a Q_{i,t}(s',a) - \mathbb{E}_e\left[\max_a Q_i^*(s',a)\right]\right)^2\right] \\
&+4\gamma\eta^2 R_{\max}\left(\max_{s',a}\left|Q_{i,t}(s',a) - Q_i^*(s',a)\right|\right),
\end{aligned}
$$

が成り立つ．ここで，$\mathbb{V}_e[\cdot] = \mathbb{V}_{\pi^e_{-i}}[\cdot|s,a^i]$ とすると，同様にして，以下が成り立つ：

$$
\begin{aligned}
&\gamma^2 \mathbb{E}_e\left[\left(\rho^{-i}\max_a Q_{i,t}(s',a) - \mathbb{E}_e\left[\max_a Q_i^*(s',a)\right]\right)^2\right] \\
&=\gamma^2 \mathbb{E}_e\left[\left(\rho^{-i}\max_a Q_{i,t}(s',a) - \rho^{-i}\max_a Q_i^*(s',a)\right.\right. \\
&\quad \left.\left. -\mathbb{E}_e\left[\max_a Q_i^*(s',a)\right] + \rho^{-i}\max_a Q_i^*(s',a)\right)^2\right] \\
&\leq \gamma^2\eta^2\left(\max_{s',a}\left|\Delta_{i,t}(s',a)\right|\right)^2 + \gamma^2 \mathbb{V}_e\left[\rho^{-i}\max_a Q_i^*(s',a)\right] \\
&+2\gamma^2\eta\left(\max_{s',a}\left|\Delta_{i,t}(s',a)\right|\right) \\
&\quad \cdot\left(\max_{s,a^{-i},s'}\left|\mathbb{E}_e\left[\max_a Q_i^*(s',a)\right] - \rho^{-i}\max_a Q_i^*(s',a)\right|\right).
\end{aligned}
$$

したがって，再び仮定 (3) を用いると，以下が導ける：

$$
\begin{aligned}
&\gamma^2 \mathbb{E}\left[\left(\rho^{-i}\max_a Q_{i,t}(s',a) - \mathbb{E}_e\left[\max_a Q_i^*(s',a)\right]\right)^2\right] \\
&\leq \left(\frac{\gamma\eta R_{\max}}{1-\gamma}\right)^2 + \frac{2\eta R_{\max}}{1-\gamma}\max_{s',a}|\Delta_{i,t}(s',a)| + \gamma^2\eta^2\left(\max_{s',a}|\Delta_{i,t}(s',a)|\right)^2.
\end{aligned}
$$

したがって，ある定数 C が存在して

$$
\mathbb{V}[F_t(s,a^i)|\chi_t] \leq C\left(1 + \max_{s',a}|\Delta_{i,t}(s',a)|\right)^2,
$$

が成り立つことが導ける．よって，定理 2 から，$Q_{i,t}(s,a)$ は確率 1 で $Q_i^*(s,a)$ に収束する． □

定理 1 に基づいて，Algorithm 1 によって推定された exploitability が，真の exploitability $v^{\exp}(\pi_1,\pi_2)$ に収束することが導ける．

系 1. 以下の条件を仮定する：
(1) 状態空間と行動空間が有限，
(2) $\forall\,(s,a)\in\mathcal{S}\times\mathcal{A}_i$, $\sum_t \alpha_{i,t}(s,a) = \infty$, $\sum_t \alpha_{i,t}^2(s,a) < \infty$, .
(3) $\forall\, t$, $|r_t| \leq R_{\max}, 0 \leq \rho_t^{-i} \leq \eta$
次の式で定義される *exploitability* の推定値 $\hat{v}^{\exp}(\pi_1^e,\pi_2^e)$ は，確率 *1* で $v^{\exp}(\pi_1,\pi_2)$ に収束する：

$$
\hat{v}^{\exp}(\pi_1^e,\pi_2^e) = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\max_{a\in\mathcal{A}_1} Q_{1,t}^{\pi_2^e}(s,a) + \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\max_{a\in\mathcal{A}_2} Q_{2,t}^{\pi_1^e}(s,a).
$$

4. 実験

本章では，文献 [1] にて用いられたベンチマーク問題である repeated biased rock-paper-scissors (RBRPS) とマルコフサッカー [8] による実験を通して，BR Q-learning の性能を確認する．

4.1 ベンチマーク問題

RBRPS は，バイアス付きじゃんけんゲーム [12] を複数回繰り返し行うシンプルな二人零和ゲームである．本研究では，ゲームを一回のみ行う，つまり繰り返しなしの RBRPS を RBRPS1 と呼び，二回繰り返す RBRPS を RBRPS2 と呼ぶ．図 1 (a) に RBRPS2 における利得行列と状態遷移グラフを示す．時刻 $t = 1$ では，利得行列は従来のじゃんけんゲームと同様で利得に偏りを持たない．時刻 $t = 1$ の結果を基に，次の状態と利得行列が決定される．RBRPS2 は全部で 5 個の状態を持ち，各状態が図 1 (a) における各利得行列にそれぞれ対応している．

マルコフサッカーは，図 1 (b) に示すような，4×5 のグリッド上で行われる一対一のサッカーゲームである．図中における A と B はそれぞれプレイヤ 1 と 2 を表しており，丸印はボールを表している．各ターン t ごとに，各プレイヤは隣接するセルのいずれかに進むか，その場に留まるかを選択することができる．プレイヤ 1 と 2 の行動は各ターンごとにランダムな順番で実行される．いずれかのプレイヤが他のプレイヤがすでに存在するセルに移動しようとした場合，ボールの所有権はセルにすでにいたプレイヤに移動する．このとき，両プレイヤの位置は変化しない．ボールを持っているプレイヤがゴール（プレイヤ 1 はセル 10 または 15 の右，プレイヤ 2 はセル 6 または 11 の左）に到達した場合，ゲームが終了する．このとき，ゴールに到達したプレイヤは報酬 +1 を受け取り，相手プレイヤは報酬 −1 を受け取る．プレイヤとボールの位置は図 1 (b) に示す位置関係で初期化される．

4.2 実験手順

全ての実験において，まず Minimax-Q learning[8] を用いて近似的な最適な方策プロファイル π_d を学習する．続いて，π_d を基にして行動方策プロファイル π^b と評価方策プロファイル π^e を構築する．その後，π^b から履歴データのサンプリングを行い，それを基に BR Q-learning の学習，すなわち π^e の exploitability の推定を行う．

4.3 比較手法

Algorithm 1 において importance weighting を用いない，すなわち $y_{i,t} = r_{i,t} + \gamma\max_a Q_{i,t}^{\pi^e_{-i}}(s_{t+1},a)$ とした手法との比較を行う．本実験では，この手法を Behavior Q-learning と呼ぶ．また，Behavior Q-learning において評価方策プロファイ

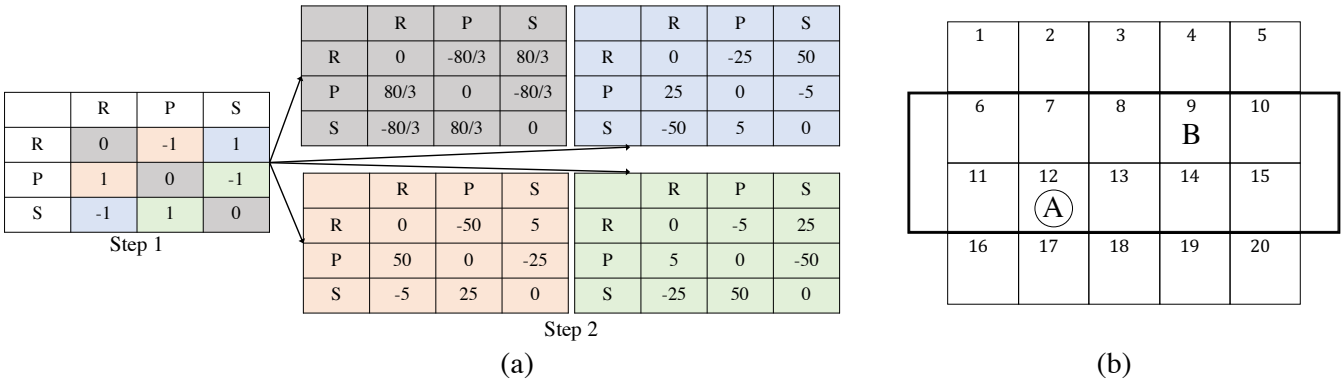

図 1 (a) Repeated biased rock-paper-scissors における利得行列と状態遷移グラフ．最初の時刻であいこだった場合，利得行列は灰色で示した行列に遷移する．どちらかのプレイヤがグー/パー/チョキで勝利した場合，利得行列はそれぞれ青/赤/緑の行列へと遷移する．(b) マルコフサッカーにおける初期配置．

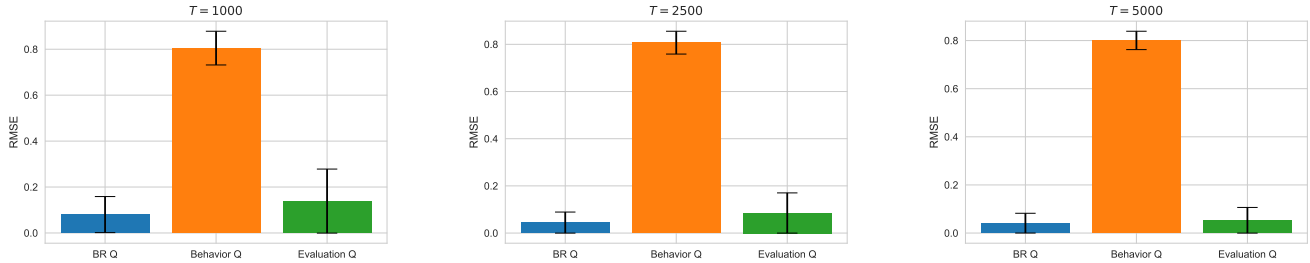

図 2 RBRPS1 における exploitability の推定誤差（RMSE）．エラーバーは標準誤差を表している．

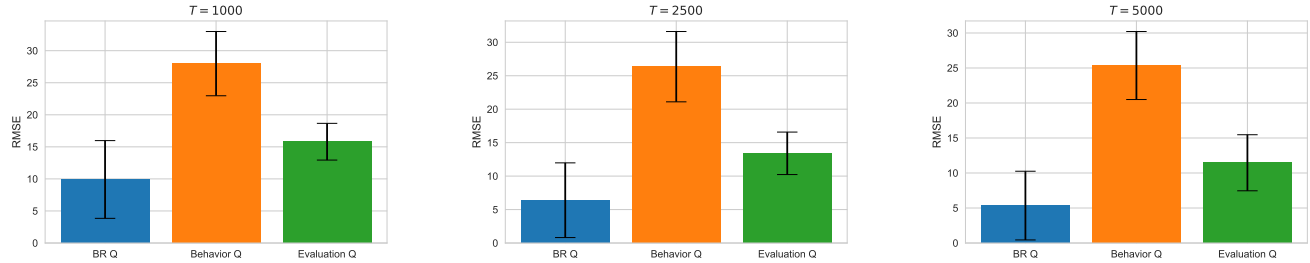

図 3 RBRPS2 における exploitability の推定誤差（RMSE）．エラーバーは標準誤差を表している．

ル π^e からの履歴データを用いて学習する手法を Evaluation Q-learning と呼ぶ．Evaluation Q-learning は評価方策プロファイル π^e からのサンプリングを行うため，BR Q-learning や Behavior Q-learning よりも有利な設定であることに注意されたい．

4.4 実験結果

4.4.1 Repeated biased rock-paper-scissors

RBRPS1 と RBRPS2 における実験では，行動方策プロファイルを $\pi_1^b = 0.8\pi_1^d + 0.2\pi^r$ と $\pi_2^b = 0.4\pi_2^d + 0.6\pi^p$ に，評価方策プロファイルを $\pi_1^e = 0.2\pi_1^d + 0.8\pi^r$ と $\pi_2^e = 0.2\pi_2^d + 0.8\pi^p$ に設定した．ここで，π^r は確率 1 でグーを出す方策を，π^p は確率 1 でパーを出す方策を表している．履歴データのサンプル数 T は $T \in \{1000, 2500, 5000\}$ とし，それぞれの設定に対して 100 試行ずつ実験を行った．

図 2，3 に，各手法が推定した exploitability と真の exploitability $v^{\exp}(\pi_1^e, \pi_2^e)$ の root-mean-squared error (RMSE) を示す．各図においてエラーバーは標準誤差を表している．BR Q-learning は Behavior Q-learning と比較して大幅に低い推定誤差を達成していることがわかる．また，評価方策プロファイル π^e からの履歴データを用いる Evaluation Q-learning よりもわずかに低い推定誤差を達成していることも確認できる．これは，π^e が π^b と比較して決定的な意思決定を行うために，Evaluation Q-learning が特定の状態

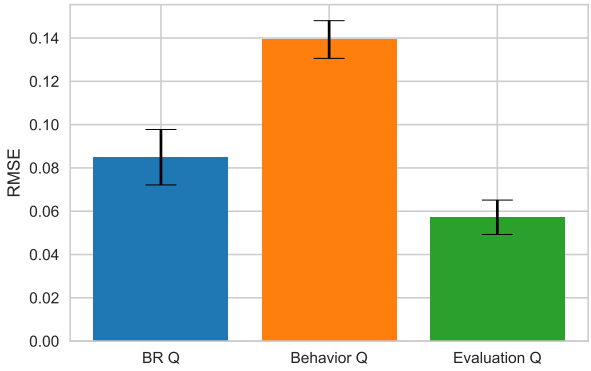

図 4 マルコフサッカーにおける exploitability の推定誤差（RMSE）．エラーバーは標準誤差を表している．

行動に対して十分な学習が行えなかったことが要因であると考えられる．

4.4.2 マルコフサッカー

マルコフサッカーにおける実験では，行動方策プロファイルを $\pi_1^b = 0.2\pi_1^d + 0.8\pi^u$ と $\pi_2^b = 0.5\pi_2^d + 0.5\pi^u$ に，評価方策プロファイルを $\pi_1^e = 0.4\pi_1^d + 0.6\pi^r$ と $\pi_2^e = 0.6\pi_2^d + 0.4\pi^u$ に設定した．なお，π^u は一様ランダムに意思決定を行う方策を表している．履歴データのサンプル数 T は $T = 1{,}000{,}000$ とし，それぞれの手法に対して 10 試行ずつ実験を行った．マルコフサッカーでは真の exploitability を計算することが困難である．そこで，π_e^1 と π_e^2 それぞれの方策を固定した環境に対して Q-learning を適用することで近似的に exploitability を計算し，その値との誤差を計測する．

図 4 に，各手法の exploitability の推定誤差を示す．RBRPS の結果と同様に，BR Q-laerning は Behavior Q-learning よりも少ない推定誤差を達成していることが確認できる．

5. おわりに

本研究では，二人零和マルコフゲームにおけるオフ方策評価のための Q-learning アルゴリズムである BR Q-laerning を提案した．また，BR Q-learning によって更新される状態行動価値関数は、最適反応の状態行動価値関数に確率 1 で収束することを証明した．さらに，BR Q-learning を用いて推定された exploitability は，真の exploitability に確率 1 で収束することを示した．また，実験によって BR Q-learning の有効性を確認した．今後の研究として，BR Q-learning や exploitability の推定値の漸近的な収束レートを示すなどの方針がある．

参考文献

[1] Kenshi Abe and Yusuke Kaneko. Off-policy exploitability-evaluation and equilibrium-learning in two-player zero-sum markov games. *arXiv preprint arXiv:2007.02141*, 2020.

[2] Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

[3] Hirotaka Hachiya, Masashi Sugiyama, and Naonori Ueda. Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. *Neurocomputing*, 80:93–101, 2012.

[4] Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *ICML*, pages 1372–1383, 2017.

[5] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *NeurIPS*, pages 703–710, 1994.

[6] Nathan Kallus and Masatoshi Uehara. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *NeurIPS*, pages 3320–3329, 2019.

[7] Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

[8] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *ICML*, pages 157–163. 1994.

[9] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *NeurIPS*, pages 5356–5366, 2018.

[10] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, pages 1077–1084, 2014.

[11] Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

[12] Mohammad Shafiei Nathan Sturtevant Jonathan Schaeffer, N Shafiei, et al. Comparing uct versus cfr in simultaneous games. In *IJCAI Workshop on General Game Playing*.

[13] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

[14] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *ICML*, pages 2139–2148, 2016.

[15] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.