Off-Policy Exploitability-Evaluation in Two-Player Zero-Sum Markov Games

Kenshi Abe
CyberAgent, Inc.
Shibuya, Tokyo
abe_kenshi@cyberagent.co.jp

Yusuke Kaneko
CyberAgent, Inc.
Shibuya, Tokyo
kaneko_yusuke@cyberagent.co.jp

ABSTRACT

Off-policy evaluation (OPE) is the problem of evaluating new policies using historical data obtained from a different policy. In the recent OPE context, most studies have focused on single-player cases, and not on multi-player cases. In this study, we propose OPE estimators constructed by the doubly robust and double reinforcement learning estimators in two-player zero-sum Markov games. The proposed estimators project exploitability that is often used as a metric for determining how close a policy profile (i.e., a tuple of policies) is to a Nash equilibrium in two-player zero-sum games. We prove the exploitability estimation error bounds for the proposed estimators. We then propose the methods to find the best candidate policy profile by selecting the policy profile that minimizes the estimated exploitability from a given policy profile class. We prove the regret bounds of the policy profiles selected by our methods. Finally, we demonstrate the effectiveness and performance of the proposed estimators through experiments.

KEYWORDS

Off-Policy Evaluation; Markov Games; Causal Inference; Reinforcement Learning

ACM Reference Format:
Kenshi Abe and Yusuke Kaneko. 2021. Off-Policy Exploitability-Evaluation in Two-Player Zero-Sum Markov Games. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 10 pages.

1 INTRODUCTION

Off-policy evaluation (OPE) is the problem of evaluating new policies using historical data obtained from a different policy. Because online policy evaluation and learning are usually expensive or risky in various applications of reinforcement learning (RL), such as medicine [36] and education [33], OPE is attracting considerable interest [2, 22, 26, 30, 50, 51, 60]. In the recent OPE context, most studies have focused on single-player cases rather than multi-player cases.

Multi-Agent Reinforcement Learning (MARL) is a generalization of single-agent RL for multi-agent environments. It is widely applicable to situations where there are multi-agent interactions, such as security games, auctions, and negotiations. In recent years, MARL has achieved many successes in the games Go [46, 47] and poker [6, 7]. MARL is a field with potential real-world applications, such as automated driving [44].

In this study, we propose OPE estimators in two-player zero-sum Markov games (TZMGs), which is one of the problems dealt with in MARL. In general, existing OPE estimators in RL estimate the discounted value of a new policy. However, in multi-agent environments, estimating the discounted value is ineffective when the policy of the other player is unknown. Unlike these estimators, for OPE in MARL, our OPE estimators evaluate a strategy profile by estimating exploitability, which is a metric for determining how close a strategy profile is to a Nash equilibrium in TZMGs. The proposed exploitability estimators are constructed by the doubly robust (DR) [18] and double reinforcement learning (DRL) [23] value estimators. We prove that the proposed exploitability estimators are $\sqrt{n}$-consistent estimators for the true exploitability.

We also propose the methods to find the best candidate strategy profile from a given strategy profile class. The proposed methods select the strategy profile that minimizes the exploitability projected by our exploitability estimators. Then, we prove that we can consistently select the true lowest-exploitability policy profile using the proposed methods.

To demonstrate the effectiveness of our exploitability estimators, we compare our estimators to the estimators based on the following representative value estimators: importance sampling (IS), marginalized importance sampling (MIS), direct method (DM) value estimators. The results show that the exploitability estimators based on the DR and DRL value estimators generally outperform the other estimator-based methods. To the best of our knowledge, this is the first proposed estimators for exploitability for OPE in TZMGs.

2 PRELIMINARY

2.1 Two-Player Zero-Sum Markov Game

A TZMG is defined as a tuple $\langle \mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, T, P_I, P_T, P_R, \gamma \rangle$, where $\mathcal{S}$ represents a finite state space; $\mathcal{A}_i$ represents an action space for player $i \in \{1, 2\}$; T represents a horizon; $P_I : \mathcal{S} \to [0, 1]$ represents an initial state distribution; $P_T : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{S} \to [0, 1]$ represents a transition probability function; $P_R : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times \mathbb{R} \to [0, 1]$ represents a reward distribution; and $\gamma \in [0, 1]$ represents a discount factor. We define $R : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2$ as a mean reward function of P_R. For $t = 1, \cdots, T$, we define $r_t \sim P_R(s_t, a_t^1, a_t^2)$ as a player 1's reward for taking actions a_t^1 and a_t^2 at state s_t, and define $-r_t$ as a player 2's reward. Let $\pi_{i,t} : \mathcal{S} \times \mathcal{A}_i \to [0, 1]$ be a Markov policy for player i at step $t \leq T$, and let $\pi_i = (\pi_{i,t})_{t \leq T}$. We define $\pi = (\pi_1, \pi_2)$ as a strategy profile or a *policy profile*. The T-step discounted value of the policy profile (π_1, π_2) for each player is

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

represented as follows:

$$v_1(\pi_1,\pi_2)=\mathbb{E}_{\pi_1,\pi_2}[\sum_{t=1}^{T}\gamma^{t-1}r_t],\ v_2(\pi_1,\pi_2)=-v_1(\pi_1,\pi_2).$$

We further define the state value function of state s_t at step t ($1\leq t\leq T$) as follows:

$$V_{1,t}(s_t)=\mathbb{E}_{\pi_1,\pi_2}[\sum_{k=t}^{T}\gamma^{k-t}r_k|s_t],\ V_{2,t}(s_t)=-V_{1,t}(s_t).$$

Based on the state value function, we define the state-action value function of taking actions a^1_t and a^2_t at state s_t as follows:

$$Q_{1,t}(s_t,a^1_t,a^2_t)=R(s_t,a^1_t,a^2_t)+\mathbb{E}_{P_T}[\gamma V_{1,t+1}(s_{t+1})|s_t,a^1_t,a^2_t],$$
$$Q_{2,t}(s_t,a^1_t,a^2_t)=-Q_{1,t}(s_t,a^1_t,a^2_t).$$

For a given policy profile π, we recursively define the marginal state-action distribution $p^\pi_t(s_t,a^1_t,a^2_t)$ at step t as follows:

$$p^\pi_t(s_t,a^1_t,a^2_t)=\pi_{1,t}(a^1_t|s_t)\pi_{2,t}(a^2_t|s_t)$$
$$\cdot\sum_{s_{t-1}\in\mathcal{S}}\sum_{a^1_{t-1}\in\mathcal{A}_1}\sum_{a^2_{t-1}\in\mathcal{A}_2}P_T(s_t|s_{t-1},a^1_{t-1},a^2_{t-1})p^\pi_{t-1}(s_{t-1},a^1_{t-1},a^2_{t-1}),$$

where $p^\pi_1(s_1,a^1_1,a^2_1)=\pi_{1,1}(a^1_1|s_1)\pi_{2,1}(a^2_1|s_1)P_I(s_1)$.

2.2 Nash Equilibrium and Exploitability

A common solution concept for two-player zero-sum games is a Nash equilibrium [37, 45], where no player cannot improve by deviating from their specified strategy. In TZMGs, a Nash equilibrium $\pi^\star=(\pi^\star_1,\pi^\star_2)$ ensures the following condition:

$$\forall\pi_1\in\Omega_1,\ \forall\pi_2\in\Omega_2,\ v_1(\pi^\star_1,\pi_2)\geq v_1(\pi^\star_1,\pi^\star_2)\geq v_1(\pi_1,\pi^\star_2),\quad(1)$$

where Ω_1 and Ω_2 are the *whole policy sets*, i.e., the sets of all possible Markov policies for players 1 and 2, respectively. The best response is a policy for player i that is optimal against π_{-i}, where π_{-i} is a policy for a player other than i. Here, we introduce the value known as *exploitability* [19, 32, 52], which is a metric for measuring how close a policy profile π is to a Nash equilibrium $\pi^\star=(\pi^\star_1,\pi^\star_2)$ in two-player zero-sum games. Formally, the exploitability of π_1,π_2 is represented as follows:

$$\begin{aligned}v^{\exp}(\pi_1,\pi_2)&=\max_{\pi'_2\in\Omega_2}v_2(\pi_1,\pi'_2)-v_1(\pi_1,\pi_2)\\&\quad+\max_{\pi'_1\in\Omega_1}v_1(\pi'_1,\pi_2)-v_2(\pi_1,\pi_2)\\&=\max_{\pi'_1\in\Omega_1}v_1(\pi'_1,\pi_2)+\max_{\pi'_2\in\Omega_2}v_2(\pi_1,\pi'_2).\end{aligned}$$

Note that in two-player zero-sum games, we can rewrite the exploitability as $v^{\exp}(\pi_1,\pi_2)=v_1(\pi^\star_1,\pi^\star_2)-\min_{\pi'_2\in\Omega_2}v_1(\pi_1,\pi'_2)+v_2(\pi^\star_1,\pi^\star_2)-\min_{\pi'_1\in\Omega_1}v_2(\pi'_1,\pi_2)$. From the definition, a Nash equilibrium $\pi^\star$ has the lowest exploitability of 0.

3 OFF-POLICY EVALUATION IN TWO-PLAYER ZERO-SUM MARKOV GAMES

In this study, we assume that we can observe the *historical data*

$$\mathcal{D}=\{(s_{i,1},a^1_{i,1},a^2_{i,1},r_{i,1},\cdots,s_{i,T},a^1_{i,T},a^2_{i,T},r_{i,T},s_{i,T+1})\}^n_{i=1},$$

where $n\in\mathbb{N}$ denotes the number of sampled trajectories. The data is sampled using a fixed policy profile $\pi^b=(\pi^b_1,\pi^b_2)$. We refer to this policy profile as a *behavior policy profile*. The distribution of $\mathcal{D}$ is then defined as follows:

$$P_I(s_1)\prod_{t=1}^{T}\pi^b_{1,t}(a^1_t|s_t)\pi^b_{2,t}(a^2_t|s_t)P_R(r_t|s_t,a^1_t,a^2_t)P_T(s_{t+1}|s_t,a^1_t,a^2_t).$$

In most of the studies related to OPE, the goal is to estimate the discounted value of a given *target policy* from the historical data. However, this goal is not appropriate for multi-agent environments. This is because, in general, in TZMGs, the policy of the opponent player is unknown, and one may play a game against a different policy than the target policy. In this case, the discounted value of the target policy depends critically on the opponent player's policy. Therefore, when the opponent policy is unknown, it is not worth estimating the discounted value against a specific policy. In this study, for OPE in TZMGs, we estimate the exploitability of a given *target policy profile* $\pi^e=(\pi^e_1,\pi^e_2)$ from the historical data instead of estimating the discounted value. In other words, we estimate the value against the worst opponent policy for each player.

In this study, we assume that we are constrained to consider each player's policies within pre-defined policy classes $\Pi_1\subset\Omega_1$ and $\Pi_2\subset\Omega_2$. In this case, if the best responses $\arg\max_{\pi'_1\in\Omega_1}v_1(\pi'_1,\pi^e_2)$ and $\arg\max_{\pi'_2\in\Omega_2}v_2(\pi^e_1,\pi'_2)$ are not included in Π_1 and Π_2, we cannot calculate the true exploitability $v^{\exp}(\pi^e_1,\pi^e_2)$. Therefore, instead of calculating $v^{\exp}(\pi^e_1,\pi^e_2)$, our exploitability estimators project the following value:

$$v^{\exp}_\Pi(\pi^e_1,\pi^e_2)=\max_{\pi'_1\in\Pi_1}v_1(\pi'_1,\pi^e_2)+\max_{\pi'_2\in\Pi_2}v_2(\pi^e_1,\pi'_2),$$

where $\Pi=\Pi_1\times\Pi_2$ is a policy profile class. Note that our exploitability estimators project the exploitability from the historical data, without the structure information P_I, P_T, P_R, and R.

3.1 Notation

For simplicity, we abbreviate terms like $V_{1,t}(s_t)$ as $V_{1,t}$. For a policy profile π, we define the following variables (note that each variable implicitly depends on π):

- $\eta_k=\frac{\pi_{1,k}(a^1_k|s_k)\pi_{2,k}(a^2_k|s_k)}{\pi^b_{1,k}(a^1_k|s_k)\pi^b_{2,k}(a^2_k|s_k)}$: the density ratio;
- $\rho_t=\prod_{k=1}^{t}\eta_k$: the cumulative density ratio;
- $\mu_t=\frac{p^\pi_t(s_t,a^1_t,a^2_t)}{p^{\pi^b}_t(s_t,a^1_t,a^2_t)}$: the marginal density ratio;
- $\hat{\pi}^b_i$: the estimators of π^b_i;
- $\hat{Q}_{1,t}$: the estimators of $Q_{1,t}$;
- $\hat{\rho}_t=\prod_{k=1}^{t}\frac{\pi_{1,k}(a^1_k|s_k)\pi_{2,k}(a^2_k|s_k)}{\hat{\pi}^b_{1,k}(a^1_k|s_k)\hat{\pi}^b_{2,k}(a^2_k|s_k)}$: the estimator of ρ_t.

Besides, we use the notation $\mathbb{E}_{\mathcal{D}}[f(X)]=\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}f(x)$ as an empirical average over $\mathcal{D}$, and we use $\mathbb{V}[\cdot]$ as a variance.

In the proofs presented in this study, we make the following assumptions regarding the overlapping of the policies and bounds of rewards and estimators, which are standard in the existing OPE literature [21, 23, 60]:

Assumption 1. $0\leq\eta_t\leq C$, $|r_t|\leq R_{\max}$ *for all* $1\leq t\leq T$.

Assumption 2. $0\leq\hat{\rho}_t\leq C^t$, $0\leq\hat{\mu}_t\leq C^t$, $0\leq|\hat{Q}_{1,t}|\leq(T+1-t)R_{\max}$ *for all* $1\leq t\leq T$.

4 OFF-POLICY VALUE ESTIMATORS

In this study, we construct the exploitability estimators using DR and DRL value estimators [18, 23], which are the efficient estimators for the discounted value $v_i(\pi_1, \pi_2)$. Therefore, in this section, we discuss the off-policy value evaluation and propose DR and DRL estimators for the discounted value in TZMGs. To distinguish these estimators from the exploitability estimators, we refer to them as *value estimators.*

4.1 Efficiency Bound in Two-Player Zero-Sum Markov Games

First, we discuss the (semiparametric) efficiency bound, which is the lower bound of the asymptotic mean squared error of OPE, among regular $\sqrt{n}$-consistent estimators. Following the general literature [53], we discuss the efficiency bound of the discounted value in TZMGs. An efficiency bound is defined for estimators under several conjectured models of the data generating process. If the conjectured model is parametric, the efficiency bound is equal to the Cramér-Rao lower bound. Even if the conjectured model is non-parametric or semi-parametric, we can still define a corresponding Cramér-Rao lower bound. Here, we introduce the following theorem from [23].

Theorem 1 (Efficiency bound in TZMGs). *The efficiency bound of $v_1(\pi_1, \pi_2)$ in TZMGs is*

$$\Upsilon_{\mathrm{EB}} = \mathbb{V}[V_{1,1}] + \sum_{t=1}^{T} \mathbb{E}[\gamma^{2(t-1)} \mu_t^2 \mathbb{V}[r_t + \gamma V_{1,t+1} | s_t, a_t^1, a_t^2]],$$

where $V_{1,T+1} = 0$.

4.2 Efficient Off-Policy Value Estimators

In this section, we propose the DR and DRL value estimators in TZMGs and their asymptotic properties.

Double Robust Estimator: We extend the DR value estimator for Markov decision processes (MDPs) proposed by [18] to apply to TZMGs. For the theoretical guarantees, we consider the *cross-fitting* version of the DR value estimator. We split the historical data into K evenly-sized folds. Next, for each fold k, we construct estimators $\hat{\rho}_t^{-k}$ and $\hat{Q}_{1,t}^{-k}$ based on all the data except fold k. We define the DR value estimator as follows:

$$\hat{v}_1^{\mathrm{DR}}(\pi_1, \pi_2) = \mathbb{E}_{\mathcal{D}}\left[\sum_{t=1}^{T} \gamma^{t-1} \left(\hat{\rho}_t^{-k(i)} \left(r_t - \hat{Q}_{1,t}^{-k(i)}\right) + \hat{\rho}_{t-1}^{-k(i)} \hat{V}_t^{-k(i)}\right)\right],$$
$$\hat{v}_2^{\mathrm{DR}}(\pi_1, \pi_2) = -v_1^{\mathrm{DR}}(\pi_1, \pi_2),$$

where $\hat{V}_t^{-k(i)} = \mathbb{E}_\pi[\hat{Q}_{1,t}^{-k(i)} | s_t]$[1] and $k(i)$ denotes the fold that contains the i-th data point. By extending the proof of Theorem 6 in [23] to the case of TZMG, we can easily show the asymptotic property of the DR value estimator.

Theorem 2 (Asymptotic property of the DR value estimator). *Suppose* $1 \leq t \leq T, 1 \leq k \leq K, \|\hat{Q}_{1,t}^{-k} - Q_{1,t}\|_2 = o_p(n^{-\alpha_1}), \|\hat{\rho}_t^{-k} - \rho_t\|_2 = o_p(n^{-\alpha_2})$*, where* $\alpha_1 > 0, \alpha_2 > 0$*, and* $\alpha_1 + \alpha_2 \geq 1/2$*. Then,*

$$\sqrt{n}(\hat{v}_1^{\mathrm{DR}}(\pi_1, \pi_2) - v_1(\pi_1, \pi_2)) \xrightarrow{d} \mathcal{N}(0, \Upsilon^{\mathrm{DR}}),$$

$$\sqrt{n}(\hat{v}_2^{\mathrm{DR}}(\pi_1, \pi_2) - v_2(\pi_1, \pi_2)) \xrightarrow{d} \mathcal{N}(0, \Upsilon^{\mathrm{DR}}),$$

where

$$\Upsilon^{\mathrm{DR}} = \mathbb{V}[V_{1,1}] + \sum_{t=1}^{T} \mathbb{E}[\gamma^{2(t-1)} \rho_t^2 \mathbb{V}[r_t + \gamma V_{1,t+1} | \{s_k, a_k^1, a_k^2\}_{k=1}^t]],$$

and $V_{1,T+1} = 0$.

The proof of this theorem is shown in the full version of the paper [1]. As in [18, 23], we can easily show that Υ^{DR} is the semiparametric efficiency bound under games where the current state s_t uniquely determines a trajectory.

Double Reinforcement Learning Estimator: In addition to the DR value estimator, we extend a DRL value estimator with cross-fitting for MDPs proposed by [23] to one for TZMGs. The DRL value estimator is defined as follows:

$$\hat{v}_1^{\mathrm{DRL}}(\pi_1, \pi_2) = \mathbb{E}_{\mathcal{D}}\left[\sum_{t=1}^{T} \gamma^{t-1} \left(\hat{\mu}_t^{-k(i)} \left(r_t - \hat{Q}_{1,t}^{-k(i)}\right) + \hat{\mu}_{t-1}^{-k(i)} \hat{V}_{1,t}^{-k(i)}\right)\right],$$
$$\hat{v}_2^{\mathrm{DRL}}(\pi_1, \pi_2) = -\hat{v}_1^{\mathrm{DRL}}(\pi_1, \pi_2).$$

By extending the proof of Theorem 13 in [23] to the TZMG case, we can again show the asymptotic property of the DRL value estimator.

Theorem 3 (Efficiency of the DRL value estimator). *Suppose* $1 \leq t \leq T, 1 \leq k \leq K, \|\hat{Q}_{1,t}^{-k} - Q_{1,t}\|_2 = o_p(n^{-\alpha_1}), \|\hat{\mu}_t^{-k} - \mu_t\|_2 = o_p(n^{-\alpha_2})$*, where* $\alpha_1 > 0, \alpha_2 > 0$*, and* $\alpha_1 + \alpha_2 \geq 1/2$*. Then,*

$$\sqrt{n}(\hat{v}_1^{\mathrm{DRL}}(\pi_1, \pi_2) - v_1(\pi_1, \pi_2)) \xrightarrow{d} \mathcal{N}(0, \Upsilon_{\mathrm{EB}}),$$

$$\sqrt{n}(\hat{v}_2^{\mathrm{DRL}}(\pi_1, \pi_2) - v_2(\pi_1, \pi_2)) \xrightarrow{d} \mathcal{N}(0, \Upsilon_{\mathrm{EB}}),$$

where Υ_{EB} *is an efficiency bound in Theorem 1.*

According to this result, the DRL value estimator is efficient under mild assumptions, whereas the IS, MIS, DM, and DR estimators may be inefficient.

4.3 Other Candidates of Value Estimators

In this study, we compare our exploitability estimators to the estimators constructed by the IS, MIS, and DM value estimators. This section summarizes these value estimators.

Importance Sampling Estimator: An IS estimator is represented as follows:

$$\hat{v}_1^{\mathrm{IS}}(\pi_1, \pi_2) = \mathbb{E}_{\mathcal{D}}\left[\sum_{t=1}^{T} \gamma^{t-1} \hat{\rho}_t r_t\right], \hat{v}_2^{\mathrm{IS}}(\pi_1, \pi_2) = -\hat{v}_1^{\mathrm{IS}}(\pi_1, \pi_2).$$

When the behavior policy profile is known, i.e., $\hat{\rho}_t = \rho_t$, the IS estimator is an unbiased and consistent estimator of $v_1(\pi_1, \pi_2)$ and $v_2(\pi_1, \pi_2)$. However, in general, the variance of the IS estimator grows exponentially with respect to horizon T [18].

Marginalized Importance Sampling Estimator: A MIS estimator is represented as follows:

$$\hat{v}_1^{\mathrm{MIS}}(\pi_1, \pi_2) = \mathbb{E}_{\mathcal{D}}\left[\sum_{t=1}^{T} \gamma^{t-1} \hat{\mu}_t r_t\right], \hat{v}_2^{\mathrm{MIS}}(\pi_1, \pi_2) = -\hat{v}_1^{\mathrm{MIS}}(\pi_1, \pi_2).$$

The MIS estimator can be regarded as one of the IS-type estimators. Although the MIS estimator addresses the curse of horizon by exploiting the Markov decision process (MDP) structure, it is still inefficient [23, 55].

[1] $\mathbb{E}_\pi[\hat{Q}_{1,t}^{-k(i)} | s_t]$ is the expected value taken only over $a^1 \sim \pi_{1,t}(a^1|s_t)$ and $a^2 \sim \pi_{2,t}(a^2|s_t)$.

Algorithm 1 Off-Policy Exploitability Estimator with $\hat{v}_i^{\mathrm{DR}}$

```
Input: Historical data $\mathcal{D}$
Input: A target policy profile $\pi^e = (\pi_1^e, \pi_2^e)$
Input: A policy classes $\Pi_1$ and $\Pi_2$
  Take a $K$-fold random partition $(I_k)_{k=1}^K$ of observation indices $\{1, \cdots, n\}$ such that the size of each fold $I_k$ is $n/K$.
  Let $\mathcal{D}_k = \{\mathcal{D}^{(i)} | i \in I_k\}, \mathcal{D}_{-k} = \{\mathcal{D}^{(i)} | i \notin I_k\}$
  Construct value estimators
    $v_1^k(\pi_1, \pi_2) = \mathbb{E}_{\mathcal{D}_k}\left[\sum_{t=1}^{T} \gamma^{t-1}\left(\hat{\rho}_t^{-k}\left(r_t - \hat{Q}_{1,t}^{-k}\right) + \hat{\rho}_{t-1}^{-k}\hat{V}_t^{-k}\right)\right]$,
    $v_2^k(\pi_1, \pi_2) = \mathbb{E}_{\mathcal{D}_k}\left[\sum_{t=1}^{T} \gamma^{t-1}\left(\hat{\rho}_t^{-k}\left(-r_t + \hat{Q}_{1,t}^{-k}\right) - \hat{\rho}_{t-1}^{-k}\hat{V}_t^{-k}\right)\right]$,
  where $\hat{Q}_{1,t}^{-k}$ and $\hat{\rho}_t^{-k}$ are the estimators of $Q_{1,t}$ and $\rho_t$, rerspectively, constructed using $\mathcal{D}_{-k}$.
Output: $\max_{\pi_1 \in \Pi_1} \frac{1}{K}\sum_{k=1}^K v_1^k(\pi_1, \pi_2^e) + \max_{\pi_2 \in \Pi_2} \frac{1}{K}\sum_{k=1}^K v_2^k(\pi_1^e, \pi_2)$
```

Direct Method Estimator: A DM estimator is represented as follows:

$$\hat{v}_1^{\mathrm{DM}}(\pi_1, \pi_2) = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\pi}[\hat{Q}_{1,1}(s_1, a_1^1, a_1^2)|s_1]\right],$$
$$\hat{v}_2^{\mathrm{DM}}(\pi_1, \pi_2) = -\hat{v}_1^{\mathrm{DM}}(\pi_1, \pi_2).$$

The DM estimator is not consistent if $\hat{Q}_{1,1}$ is not consistent, and it is not unbiased if $\hat{Q}_{1,1}$ is not correct.

5 OFF-POLICY EXPLOITABILITY ESTIMATORS

For OPE in TZMGs, we propose the following exploitability estimators constructed by the DR and DRL value estimators, respectively:

$$\hat{v}_{\mathrm{DR}}^{\exp}(\pi_1^e, \pi_2^e) = \max_{\pi_1 \in \Pi_1} \hat{v}_1^{\mathrm{DR}}(\pi_1, \pi_2^e) + \max_{\pi_2 \in \Pi_2} \hat{v}_2^{\mathrm{DR}}(\pi_1^e, \pi_2), \tag{2}$$
$$\hat{v}_{\mathrm{DRL}}^{\exp}(\pi_1^e, \pi_2^e) = \max_{\pi_1 \in \Pi_1} \hat{v}_1^{\mathrm{DRL}}(\pi_1, \pi_2^e) + \max_{\pi_2 \in \Pi_2} \hat{v}_2^{\mathrm{DRL}}(\pi_1^e, \pi_2). \tag{3}$$

Similarly, we define $\hat{v}_{\mathrm{IS}}^{\exp}$, $\hat{v}_{\mathrm{MIS}}^{\exp}$, and $\hat{v}_{\mathrm{DM}}^{\exp}$ as the exploitability estimators based on the IS, MIS, and DM value estimators, respectively. We present the pseudocode of the proposed estimator with $\hat{v}_i^{\mathrm{DR}}$ in Algorithm 1. The procedure of the exploitability estimator with $\hat{v}_i^{\mathrm{DRL}}$ is the same as Algorithm 1 except that $\hat{\rho}_t$ is replaced with $\hat{\mu}_t$.

In this section, we demonstrate the exploitability estimation error bounds of $\hat{v}_{\mathrm{DR}}^{\exp}(\pi_1^e, \pi_2^e)$ and $\hat{v}_{\mathrm{DRL}}^{\exp}(\pi_1^e, \pi_2^e)$. To obtain theoretical implications, we define the ϵ-Hamming covering number $N_H(\epsilon, \Pi)$ under the Hamming distance $H_n(\pi^a, \pi^b) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(\{\bigvee_{t=1}^T \pi_{1,t}^a(s_{i,t}) \neq \pi_{1,t}^b(s_{i,t})\} \vee \{\bigvee_{t=1}^T \pi_{2,t}^a(s_{i,t}) \neq \pi_{2,t}^b(s_{i,t})\})$ and its entropy integral $\kappa(\Pi) = \int_0^\infty \sqrt{\log N_H(\epsilon^2, \Pi)}$. In the proofs of the remaining theorems, we make the following assumptions on the covering number $N_H(\epsilon, \Pi)$:

Assumption 3. *For any* $0 < \epsilon < 1, N_H(\epsilon, \Pi) \leq D_1 \exp(D_2(\frac{1}{\epsilon})^\omega)$ *for some constants* $D_1, D_2 > 0, 0 \leq \omega < 0.5$.

Assumption 3 is precisely the same as the assumption in the proof of [25, 60], and this is not strong assumption [60]. Furthermore, to establish uniform error bounds on $\hat{Q}_{1,t}$ and $\hat{\mu}_t$, in the remaining theorems, we assume that $\hat{Q}_{1,t}$ and $\hat{\mu}_t$ are computed using the estimated TZMG model $\hat{R}$, $\hat{P}_T$, $\hat{p}_t^{\pi^b}$. Under similar consistency assumptions as in [25, 60], the estimation error bounds of $\hat{v}_{\mathrm{DR}}^{\exp}$ and $\hat{v}_{\mathrm{DRL}}^{\exp}$ are then obtained as follows:

Theorem 4 (Estimation error bound of $\hat{v}_{\mathrm{DR}}^{\exp}(\pi_1^e, \pi_2^e)$). *Let us define* $\hat{\pi}_l^{b,-k}(a_l^1, a_l^2|s_l) = \hat{\pi}_{1,l}^{b,-k}(a_l^1|s_l)\hat{\pi}_{2,l}^{b,-k}(a_l^2|s_l)$ *and* $\pi_l^b(a_l^1, a_l^2|s_l) = \pi_{1,l}^b(a_l^1|s_l)\pi_{2,l}^b(a_l^2|s_l)$. *Assume Assumptions 1, 2, 3, (4a)* $1 \leq t \leq T$ *and* $t \leq t' \leq T$,

$$\mathbb{E}\left[\left(\hat{R}^{-k}(s_{t'}, a_{t'}^1, a_{t'}^2)\prod_{l=t}^{t'-1}\hat{P}_T^{-k}(s_{l+1}|s_l, a_l^1, a_l^2) - R(s_{t'}, a_{t'}^1, a_{t'}^2)\prod_{l=t}^{t'-1}P_T(s_{l+1}|s_l, a_l^1, a_l^2)\right)^2\right] = o(n^{-2\alpha_1}),$$

and (4b) $1 \leq t \leq T$,

$$\mathbb{E}\left[\left(\prod_{l=1}^{t}\frac{1}{\hat{\pi}_l^{b,-k}(a_l^1, a_l^2|s_l)} - \prod_{l=1}^{t}\frac{1}{\pi_l^b(a_l^1, a_l^2|s_l)}\right)^2\right] = o(n^{-2\alpha_2}),$$

where $\alpha_1 > 0, \alpha_2 > 0$, *and* $\alpha_1 + \alpha_2 \geq 1/2$. *Then, for any* $\delta > 0$, *there exists* $C > 0, N_\delta > 0$, *such that with probability at least* $1 - 2\delta$ *and for all* $n \geq N_\delta$:

$$|v_\Pi^{\exp}(\pi_1^e, \pi_2^e) - \hat{v}_{\mathrm{DR}}^{\exp}(\pi_1^e, \pi_2^e)| \leq C\left(\kappa(\Pi) + \sqrt{\log(1/\delta)}\right)\sqrt{\Upsilon_{\mathrm{DR}}^*/n},$$

where $\Upsilon_{\mathrm{DR}}^* = \sup_{\pi \in \Pi}\mathbb{E}\left[\left(\sum_{t=1}^T \gamma^{t-1}\left(\rho_t(r_t - Q_{1,t}) + \rho_{t-1}V_{1,t}\right)\right)^2\right]$.

Theorem 5 (Estimation error bound of $\hat{v}_{\mathrm{DRL}}^{\exp}(\pi_1^e, \pi_2^e)$). *Assume Assumptions 1, 2, 3, (4a), and (5a)* $1 \leq t \leq T$,

$$\mathbb{E}\left[\left(\frac{\prod_{t'=1}^t \hat{P}_T^{-k}(s_{t'}|s_{t'-1}, a_{t'-1}^1, a_{t'-1}^2)}{\hat{p}_{b,t}^{-k}(s_t, a_t^1, a_t^2)} - \frac{\prod_{t'=1}^t P_T(s_{t'}|s_{t'-1}, a_{t'-1}^1, a_{t'-1}^2)}{p_{b,t}(s_t, a_t^1, a_t^2)}\right)^2\right] = o(n^{-2\alpha_2}),$$

where $\alpha_1 > 0, \alpha_2 > 0$, *and* $\alpha_1 + \alpha_2 \geq 1/2$. *Then, for any* $\delta > 0$, *there exists* $C > 0, N_\delta > 0$, *such that with probability at least* $1 - 2\delta$ *and for all* $n \geq N_\delta$:

$$|v_\Pi^{\exp}(\pi_1^e, \pi_2^e) - \hat{v}_{\mathrm{DRL}}^{\exp}(\pi_1^e, \pi_2^e)| \leq C\left(\kappa(\Pi) + \sqrt{\log(1/\delta)}\right)\sqrt{\Upsilon_{\mathrm{DRL}}^*/n},$$

where $\Upsilon_{\mathrm{DRL}}^* = \sup_{\pi \in \Pi}\mathbb{E}\left[\left(\sum_{t=1}^T \gamma^{t-1}\left(\mu_t(r_t - Q_{1,t}) + \mu_{t-1}V_{1,t}\right)\right)^2\right]$.

Theorems 4 and 5 mean that $\hat{v}_{\mathrm{DR}}^{\exp}$ and $\hat{v}_{\mathrm{DRL}}^{\exp}$ are $\sqrt{n}$-consistent estimators for the true exploitability defined among Π. In particular, when $\Pi = \Omega_1 \times \Omega_2$, the error between the estimated exploitability and the true exploitability $v^{\exp}(\pi_1^e, \pi_2^e)$ converges to 0 at a rate $O_p(\frac{1}{\sqrt{n}})$. Because $\Upsilon_{\mathrm{DR}}^* = \sup_{\pi \in \Pi}(\Upsilon_{\mathrm{DR}} + v_1^2(\pi_1, \pi_2))$ and $\Upsilon_{\mathrm{DRL}}^* = \sup_{\pi_1, \pi_2 \in \Pi}(\Upsilon_{\mathrm{EB}} + v_1^2(\pi_1, \pi_2))$, it is necessary to use the value estimator with a small (asymptotic) variance to reduce the exploitability estimation error. That is, the exploitability estimation error would be small using the value estimator with a small asymptotic variance. Therefore, from Theorems 2 and 3, using the efficient value estimator $\hat{v}^{\mathrm{exp}}_{\mathrm{DRL}}$ would result in a small estimation error. Note that we do not assume that the behavior policy profile is known in Theorems 4 and 5. We sketch the proof of Theorem 4. The proof of Theorem 5 is almost the same as Theorem 4.

Proof sketch of Theorem 4. First, we define the DR value estimator with oracles $Q_{1,t}$ and ρ_t as follows:

$$v^{\mathrm{DR}}_1(\pi^e_1,\pi^e_2)=\mathbb{E}_{\mathcal{D}}\left[\sum_{t=1}^{T}\gamma^{t-1}\left(\rho_t\left(r_t-Q_{1,t}\right)+\rho_{t-1}V_t\right)\right],$$
$$v^{\mathrm{DR}}_2(\pi^e_1,\pi^e_2)=-v^{\mathrm{DR}}_1(\pi^e_1,\pi^e_2).$$

Besides, we define the value difference between two policy profiles π^α and π^β in Π as follows:

$$\Delta(\pi^\alpha,\pi^\beta)=v_1(\pi^\alpha_1,\pi^\alpha_2)-v_1(\pi^\beta_1,\pi^\beta_2),$$
$$\hat{\Delta}(\pi^\alpha,\pi^\beta)=\hat{v}^{\mathrm{DR}}_1(\pi^\alpha_1,\pi^\alpha_2)-\hat{v}^{\mathrm{DR}}_1(\pi^\beta_1,\pi^\beta_2),$$
$$\tilde{\Delta}(\pi^\alpha,\pi^\beta)=v^{\mathrm{DR}}_1(\pi^\alpha_1,\pi^\alpha_2)-v^{\mathrm{DR}}_1(\pi^\beta_1,\pi^\beta_2).$$

We mainly show the uniform concentration of these difference functions following the proof of [60].

Uniform concentration of the difference of influence functions: First, we prove that the influence difference function $\tilde{\Delta}(\cdot,\cdot)$ concentrates uniformly around its mean $\Delta(\cdot,\cdot)$:

Lemma 1. *Under Assumptions 1 and 3, for any $\delta>0$, with probability at least $1-2\delta$,*

$$\sup_{\pi^\alpha,\pi^\beta\in\Pi}\left|\tilde{\Delta}(\pi^\alpha,\pi^\beta)-\Delta(\pi^\alpha,\pi^\beta)\right|$$
$$\leq O\left(\left(\kappa(\Pi)+\sqrt{\log\frac{1}{\delta}}\right)\sqrt{\frac{\Upsilon^*_{\mathrm{DR}}}{n}}\right)+o(\frac{1}{\sqrt{n}}).$$

The proof of Lemma 1 is the extension of the concentration result in [60] to the TZMG setting. The proof of this lemma is shown in the full version of the paper [1].

Uniform concentration of the estimated value difference function: Next, we prove that with high probability, the estimated value difference function $\hat{\Delta}(\cdot,\cdot)$ concentrates around $\tilde{\Delta}(\cdot,\cdot)$ uniformly at a rate $o_p(\frac{1}{\sqrt{n}})$:

Lemma 2. *Under Assumptions 1, 2, 3, (4a)-(4b):*

$$\sup_{\pi_\alpha,\pi_\beta\in\Pi}\left|\hat{\Delta}(\pi^\alpha,\pi^\beta)-\tilde{\Delta}(\pi^\alpha,\pi^\beta)\right|=o_p(\frac{1}{\sqrt{n}}).$$

The proof of this lemma is shown in the full version of the paper [1]. Here, we have:

$$\sup_{\pi^\alpha,\pi^\beta\in\Pi}\left|\hat{\Delta}(\pi^\alpha,\pi^\beta)-\Delta(\pi^\alpha,\pi^\beta)\right|$$
$$=\sup_{\pi^\alpha,\pi^\beta\in\Pi}\left|\hat{\Delta}(\pi^\alpha,\pi^\beta)-\tilde{\Delta}(\pi^\alpha,\pi^\beta)-\Delta(\pi^\alpha,\pi^\beta)+\tilde{\Delta}(\pi^\alpha,\pi^\beta)\right|$$
$$\leq\sup_{\pi^\alpha,\pi^\beta\in\Pi}\left|\hat{\Delta}(\pi^\alpha,\pi^\beta)-\tilde{\Delta}(\pi^\alpha,\pi^\beta)\right|$$
$$+\sup_{\pi^\alpha,\pi^\beta\in\Pi}\left|\tilde{\Delta}(\pi^\alpha,\pi^\beta)-\Delta(\pi^\alpha,\pi^\beta)\right|.$$

Therefore, combining Lemmas 1 and 2, we can show the uniform concentration of $\hat{\Delta}(\cdot,\cdot)$ on $\Delta(\cdot,\cdot)$:

Lemma 3. *Assume Assumptions 1, 2, 3, (4a)-(4b). Then, for any $\delta>0$, there exists $C>0, N_\delta>0$, such that with probability at least $1-2\delta$ and for all $n\geq N_\delta$:*

$$\sup_{\pi^\alpha,\pi^\beta\in\Pi}\left|\hat{\Delta}(\pi^\alpha,\pi^\beta)-\Delta(\pi^\alpha,\pi^\beta)\right|\leq C\left(\kappa(\Pi)+\sqrt{\log(1/\delta)}\right)\sqrt{\frac{\Upsilon^*_{\mathrm{DR}}}{n}}.$$

Estimation error bound of the exploitability estimator: Next, we define the best response and the estimated best response as follows:

$$\pi^\dagger_1=\arg\max_{\pi_1\in\Pi_1}v_1(\pi_1,\pi^e_2),\ \pi^\dagger_2=\arg\max_{\pi_2\in\Pi_2}v_2(\pi^e_1,\pi_2),$$
$$\hat{\pi}^\dagger_1=\arg\max_{\pi_1\in\Pi_1}\hat{v}^{\mathrm{DR}}_1(\pi_1,\pi^e_2),\ \hat{\pi}^\dagger_2=\arg\max_{\pi_2\in\Pi_2}\hat{v}^{\mathrm{DR}}_2(\pi^e_1,\pi_2).$$

Then, by some algebra, we have:

$$v^{\mathrm{exp}}_\Pi(\pi^e_1,\pi^e_2)-\hat{v}^{\mathrm{exp}}_{\mathrm{DR}}(\pi^e_1,\pi^e_2)$$
$$\leq 3\sup_{\pi^\alpha\in\Pi,\pi^\beta\in\Pi}|\Delta((\pi^\alpha_1,\pi^\alpha_2),(\pi^\beta_1,\pi^\beta_2))-\hat{\Delta}((\pi^\alpha_1,\pi^\alpha_2),(\pi^\beta_1,\pi^\beta_2))|,$$

and

$$v^{\mathrm{exp}}_\Pi(\pi^e_1,\pi^e_2)-\hat{v}^{\mathrm{exp}}_{\mathrm{DR}}(\pi^e_1,\pi^e_2)$$
$$\geq -3\sup_{\pi^\alpha\in\Pi,\pi^\beta\in\Pi}|\Delta((\pi^\alpha_1,\pi^\alpha_2),(\pi^\beta_1,\pi^\beta_2))-\hat{\Delta}((\pi^\alpha_1,\pi^\alpha_2),(\pi^\beta_1,\pi^\beta_2))|.$$

Therefore, we have:

$$|v^{\mathrm{exp}}_\Pi(\pi^e_1,\pi^e_2)-\hat{v}^{\mathrm{exp}}_{\mathrm{DR}}(\pi^e_1,\pi^e_2)|$$
$$\leq 3\sup_{\pi^\alpha\in\Pi,\pi^\beta\in\Pi}|\Delta((\pi^\alpha_1,\pi^\alpha_2),(\pi^\beta_1,\pi^\beta_2))-\hat{\Delta}((\pi^\alpha_1,\pi^\alpha_2),(\pi^\beta_1,\pi^\beta_2))|.$$

Then, from Lemma 3 and this equation, the statement is concluded.

6 BEST EVALUATION POLICY PROFILE SELECTION

In this section, we consider the problem of selecting the best candidate policy profile from a given policy profile class, one of the most practical applications of OPE. For given historical data $\mathcal{D}$, our goal is to select the best policy profile with the lowest exploitability from the candidate policy profile class Π, i.e.,

$$(\pi^*_1,\pi^*_2)=\arg\min_{\pi_1,\pi_2\in\Pi_1\times\Pi_2}v^{\mathrm{exp}}_\Pi(\pi_1,\pi_2).$$

According to Equation (1), when $\Pi_1=\Omega_1$ and $\Pi_2=\Omega_2$, the policy profile (π^*_1,π^*_2) is a Nash equilibrium.

To this end, we propose methods based on the exploitability estimators proposed in the previous section. Based on the exploitability estimator $\hat{v}^{\mathrm{exp}}_{\mathrm{DR}}$, we select the policy profile that minimizes the estimated exploitability as follows:

$$(\hat{\pi}^{\mathrm{DR}}_1,\hat{\pi}^{\mathrm{DR}}_2)=\arg\min_{\pi_1,\pi_2\in\Pi_1\times\Pi_2}\hat{v}^{\mathrm{exp}}_{\mathrm{DR}}(\pi_1,\pi_2).$$

From the definition of $\hat{v}^{\mathrm{exp}}_{\mathrm{DR}}$, we can rewrite the $\hat{\pi}^{\mathrm{DR}}_1$ and $\hat{\pi}^{\mathrm{DR}}_2$, respectively, as follows:

$$\hat{\pi}^{\mathrm{DR}}_1=\arg\max_{\pi_1\in\Pi_1}\min_{\pi_2\in\Pi_2}\hat{v}^{\mathrm{DR}}_1(\pi_1,\pi_2), \tag{4}$$

$$\hat{\pi}^{\mathrm{DR}}_2=\arg\max_{\pi_2\in\Pi_2}\min_{\pi_1\in\Pi_1}\hat{v}^{\mathrm{DR}}_2(\pi_1,\pi_2). \tag{5}$$

Algorithm 2 Off-Policy Best Evaluation Policy Profile Selection with $\hat{v}^{\mathrm{exp}}_{\mathrm{DR}}$

```
Input: Historical data $\mathcal{D}$
Input: A policy classes $\Pi_1$ and $\Pi_2$
Take a $K$-fold random partition $(I_k)_{k=1}^{K}$ of observation indices $\{1, \cdots, n\}$ such that the size of each fold $I_k$ is $n/K$.
Let $\mathcal{D}_k = \{\mathcal{D}^{(i)} | i \in I_k\}, \mathcal{D}_{-k} = \{\mathcal{D}^{(i)} | i \notin I_k\}$.
Construct value estimators
    $v_1^k(\pi_1, \pi_2) = \mathbb{E}_{\mathcal{D}_k}\left[\sum_{t=1}^{T} \gamma^{t-1}\left(\hat{\rho}_t^{-k}\left(r_t - \hat{Q}_{1,t}^{-k}\right) + \hat{\rho}_{t-1}^{-k}\hat{V}_t^{-k}\right)\right]$,
    $v_2^k(\pi_1, \pi_2) = \mathbb{E}_{\mathcal{D}_k}\left[\sum_{t=1}^{T} \gamma^{t-1}\left(\hat{\rho}_t^{-k}\left(-r_t + \hat{Q}_{1,t}^{-k}\right) - \hat{\rho}_{t-1}^{-k}\hat{V}_t^{-k}\right)\right]$,
    where $\hat{Q}_{1,t}^{-k}$ and $\hat{\rho}_t^{-k}$ are the estimators of $Q_{1,t}$ and $\rho_t$, respectively, constructed using $\mathcal{D}_{-k}$.
Obtain $\hat{\pi}_1$ and $\hat{\pi}_2$ by solving the following optimization problem:
    $\hat{\pi}_1 = \arg\max_{\pi_1 \in \Pi_1} \min_{\pi_2 \in \Pi_2} \frac{1}{K}\sum_{k=1}^{K} v_1^k(\pi_1, \pi_2)$,
    $\hat{\pi}_2 = \arg\max_{\pi_2 \in \Pi_2} \min_{\pi_1 \in \Pi_1} \frac{1}{K}\sum_{k=1}^{K} v_2^k(\pi_1, \pi_2)$.
Output: $(\hat{\pi}_1, \hat{\pi}_2)$
```

Similarly, we define $\hat{\pi}^{\mathrm{IS}}$, $\hat{\pi}^{\mathrm{MIS}}$, $\hat{\pi}^{\mathrm{DM}}$, and $\hat{\pi}^{\mathrm{DRL}}$ as the estimators based on $\hat{v}^{\mathrm{exp}}_{\mathrm{IS}}$, $\hat{v}^{\mathrm{exp}}_{\mathrm{MIS}}$, $\hat{v}^{\mathrm{exp}}_{\mathrm{DM}}$, and $\hat{v}^{\mathrm{exp}}_{\mathrm{DRL}}$, respectively. We describe the pseudocode of the proposed method with $\hat{v}^{\mathrm{exp}}_{\mathrm{DR}}$ in Algorithm 2. The procedure of the proposed method with $\hat{v}^{\mathrm{exp}}_{\mathrm{DRL}}$ is the same as Algorithm 2 except that $\hat{\rho}_t$ is replaced with $\hat{\mu}_t$.

We can derive the exploitability bounds of $\hat{\pi}^{\mathrm{DR}}$ and $\hat{\pi}^{\mathrm{DRL}}$ similarly as in the proofs of Theorems 4 and 5.

Theorem 6 (Exploitability bound of $\hat{\pi}^{\mathrm{DR}}$). *Assume Assumptions 1, 2, 3, (4a)-(4b). Then, for any $\delta > 0$, there exists $C > 0, N_\delta > 0$, such that with probability at least $1 - 2\delta$ and for all $n \geq N_\delta$:*

$$v^{\mathrm{exp}}(\hat{\pi}_1^{\mathrm{DR}}, \hat{\pi}_2^{\mathrm{DR}}) - v^{\mathrm{exp}}(\pi_1^*, \pi_2^*) \leq C\left(\kappa(\Pi) + \sqrt{\log(1/\delta)}\right)\sqrt{\frac{\Upsilon^*_{\mathrm{DR}}}{n}}.$$

Theorem 7 (Exploitability bound of $\hat{\pi}^{\mathrm{DRL}}$). *Assume Assumptions 1, 2, 3, (4a), and (5a). Then, for any $\delta > 0$, there exists $C > 0, N_\delta > 0$, such that with probability at least $1 - 2\delta$ and for all $n \geq N_\delta$:*

$$v^{\mathrm{exp}}(\hat{\pi}_1^{\mathrm{DRL}}, \hat{\pi}_2^{\mathrm{DRL}}) - v^{\mathrm{exp}}(\pi_1^*, \pi_2^*) \leq C\left(\kappa(\Pi) + \sqrt{\log(1/\delta)}\right)\sqrt{\frac{\Upsilon^*_{\mathrm{DRL}}}{n}}.$$

These theorems mean that we can consistently select the true lowest-exploitability policy profile π^* using the proposed methods. According to the minimax theorem, if $\Pi_1 = \Omega_1$ and $\Pi_2 = \Omega_2$, then $v^{\mathrm{exp}}(\pi_1^*, \pi_2^*) = 0$. Therefore, in this case, the exploitability of the selected policy profile converges asymptotically to 0. This means that the selected policy profile converges asymptotically to a Nash equilibrium when $\Pi_1 = \Omega_1$ and $\Pi_2 = \Omega_2$. We sketch the proof of Theorem 6. The proof of Theorem 7 is almost the same as Theorem 6.

Proof sketch of Theorem 6. Let define:

$$\mathcal{B}_i(\pi_{-i}) = \arg\max_{\pi_i' \in \Omega_i} v_i(\pi_i', \pi_{-i}),\ \hat{\mathcal{B}}_i(\pi_{-i}) = \arg\max_{\pi_i \in \Pi_i} \hat{v}_i^{\mathrm{DR}}(\pi_i, \pi_{-i}).$$

Besides, for simplicity, we write $\hat{\pi}_i^{\mathrm{DR}}$ as $\hat{\pi}_i$ and $\hat{v}_i^{\mathrm{DR}}$ as $\hat{v}_i$. From the definitions of π_i^* and $\hat{\pi}_i$, we have:

$$\begin{aligned}
&v_1(\hat{\mathcal{B}}_1(\pi_2^*), \pi_2^*) \leq v_1(\mathcal{B}_1(\pi_2^*), \pi_2^*),\\
&v_1(\pi_1^*, \mathcal{B}_2(\pi_1^*)) \leq v_1(\pi_1^*, \hat{\mathcal{B}}_2(\pi_1^*)),\\
&\hat{v}_1(\mathcal{B}_1(\hat{\pi}_2), \hat{\pi}_2) \leq \hat{v}_1(\hat{\mathcal{B}}_1(\hat{\pi}_2), \hat{\pi}_2) \leq \hat{v}_1(\hat{\mathcal{B}}_1(\pi_2^*), \pi_2^*),\\
&\hat{v}_1(\hat{\pi}_1, \mathcal{B}_2(\hat{\pi}_1)) \geq \hat{v}_1(\hat{\pi}_1, \hat{\mathcal{B}}_2(\hat{\pi}_1)) \geq \hat{v}_1(\pi_1^*, \hat{\mathcal{B}}_2(\pi_1^*)).
\end{aligned}$$

Therefore, the exploitability bound of $\hat{\pi}$ is:

$$\begin{aligned}
&v^{\mathrm{exp}}(\hat{\pi}_1, \hat{\pi}_2) - v^{\mathrm{exp}}(\pi_1^*, \pi_2^*)\\
&= \Delta((\mathcal{B}_1(\hat{\pi}_2), \hat{\pi}_2), (\hat{\pi}_1, \mathcal{B}_2(\hat{\pi}_1))) - \hat{\Delta}((\mathcal{B}_1(\hat{\pi}_2), \hat{\pi}_2), (\hat{\pi}_1, \mathcal{B}_2(\hat{\pi}_1)))\\
&- \Delta((\mathcal{B}_1(\pi_2^*), \pi_2^*), (\pi_1^*, \mathcal{B}_2(\pi_1^*))) + \hat{\Delta}((\mathcal{B}_1(\pi_2^*), \pi_2^*), (\pi_1^*, \mathcal{B}_2(\pi_1^*)))\\
&+\hat{v}_1(\mathcal{B}_1(\hat{\pi}_2), \hat{\pi}_2) - \hat{v}_1(\hat{\pi}_1, \mathcal{B}_2(\hat{\pi}_1)) - \hat{v}_1(\mathcal{B}_1(\pi_2^*), \pi_2^*) + \hat{v}_1(\pi_1^*, \mathcal{B}_2(\pi_1^*))\\
&\leq \Delta((\mathcal{B}_1(\hat{\pi}_2), \hat{\pi}_2), (\hat{\pi}_1, \mathcal{B}_2(\hat{\pi}_1))) - \hat{\Delta}((\mathcal{B}_1(\hat{\pi}_2), \hat{\pi}_2), (\hat{\pi}_1, \mathcal{B}_2(\hat{\pi}_1)))\\
&- \Delta((\mathcal{B}_1(\pi_2^*), \pi_2^*), (\pi_1^*, \mathcal{B}_2(\pi_1^*))) + \hat{\Delta}((\mathcal{B}_1(\pi_2^*), \pi_2^*), (\pi_1^*, \mathcal{B}_2(\pi_1^*)))\\
&+ \hat{\Delta}((\hat{\mathcal{B}}_1(\pi_2^*), \pi_2^*), (\mathcal{B}_1(\pi_2^*), \pi_2^*)) - \Delta((\hat{\mathcal{B}}_1(\pi_2^*), \pi_2^*), (\mathcal{B}_1(\pi_2^*), \pi_2^*))\\
&- \hat{\Delta}((\pi_1^*, \hat{\mathcal{B}}_2(\pi_1^*)), (\pi_1^*, \mathcal{B}_2(\pi_1^*))) + \Delta((\pi_1^*, \hat{\mathcal{B}}_2(\pi_1^*)), (\pi_1^*, \mathcal{B}_2(\pi_1^*)))\\
&\leq 4 \sup_{\pi^\alpha \in \Pi, \pi^\beta \in \Pi} |\Delta((\pi_1^\alpha, \pi_2^\alpha), (\pi_1^\beta, \pi_2^\beta)) - \hat{\Delta}((\pi_1^\alpha, \pi_2^\alpha), (\pi_1^\beta, \pi_2^\beta))|.
\end{aligned}$$

Then, from Lemma 3 and this equation, the statement is concluded.

Remark 1. *In general, when the policy classes Π_1 and Π_2 are complicated, it is not easy to solve the optimization problems in Equations (2)-(5). However, when we use the specific policy class such as the kernel functions, we can estimate the exploitability efficiently by a similar procedure as in [25]. Furthermore, if we use tabular policies and estimated TZMG model $\hat{R}$, $\hat{P}_T$, and $\hat{p}_t^{\pi^b}$, we can solve the optimization problems via linear programming.*

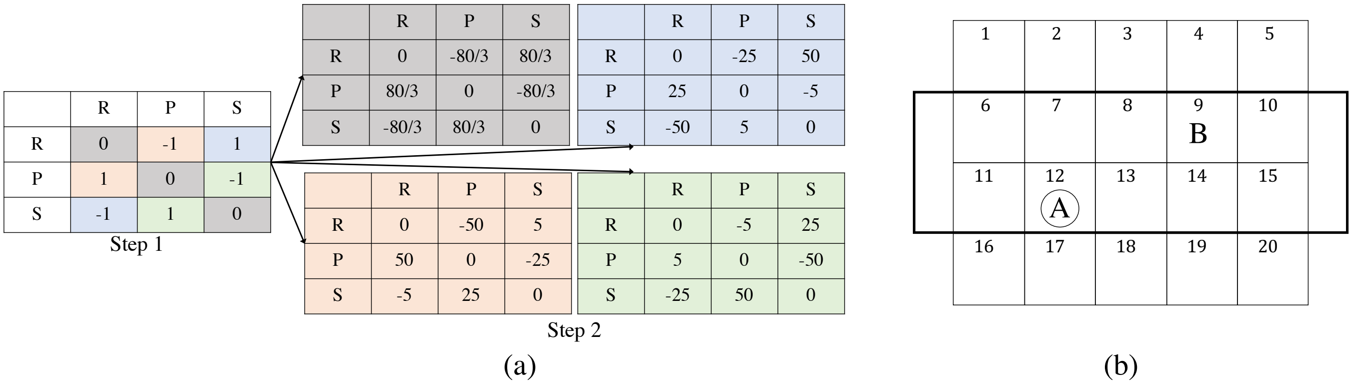
7 EXPERIMENTS

We conduct experiments to analyze and evaluate the proposed exploitability estimators and the policy profile selection methods. We conduct our experiments in two environments: repeated biased rock-paper-scissors (RBRPS) and Markov soccer [28].

In all the experiments, we first prepare a near optimal policy profile π_d using Minimax-Q learning [28], after which we construct the behavior and target policy profiles using π_d. We use an off-policy temporal difference learning [49] to construct a Q-function model, and we use a histogram estimator for μ, as in Section 5.2 in [23]. In our experiments, we assume that the behavior policy profile is known and fixed.

7.1 Environments

RBRPS is a simple TZMG where two players play an one-shot biased rock-paper-scissors game [42] multiple times. We refer to a game that is repeated once as RBRPS1 and a game that is repeated



	R	P	S
R	0	-1	1
P	1	0	-1
S	-1	1	0

Step 1

	R	P	S
R	0	-80/3	80/3
P	80/3	0	-80/3
S	-80/3	80/3	0

	R	P	S
R	0	-25	50
P	25	0	-5
S	-50	5	0

	R	P	S
R	0	-50	5
P	50	0	-25
S	-5	25	0

	R	P	S
R	0	-5	25
P	5	0	-50
S	-25	50	0

Step 2

Figure 1: (a) Payoff matrices and a state transition graph in repeated biased rock-paper-scissors. When the result at the first step is a draw, the payoff matrix at the second step will be the gray one. When either player wins by rock/paper/scissors, the payoff matrix at the next step will be the blue/red/green one. (b) An initial board in Markov soccer.

two times as RBRPS2. Note that RBRPS1 is precisely the same as the conventional rock-paper-scissors game. Figure 1 (a) shows the payoff matrices and the state transition graph of RBRPS2. In the first step, the payoff matrix is the same as in the conventional rock-paper-scissors game. Depending on the result of the one-shot game, the next state and the payoff matrix transition. There are five states in RBRPS2, and each state corresponds to each payoff matrix.

Markov soccer is a 1 vs 1 soccer game on a 4×5 grid , as shown in Figure 1 (b). A and B denote players 1 and 2, respectively, and the circle in the figure represents the ball. In each turn, each player can move to one of the neighboring cells or stay in place, and the actions of the two players are executed in random order. When a player tries to move to the cell occupied by the other player, the ball's possession goes to the stationary player, and the positions of both players remain unchanged. When the player with the ball reaches the goal (right of cell 10 or 15 for A, left of cell 6 or 11 for B), the game is over. At this time, the player receives a reward of +1, and the opponent receives a reward of -1. The player's positions and the ball's possession are initialized as shown in Figure 1 (b).

7.2 Exploitability Evaluation

In the first experiment, we compare the performance of $\hat{v}^{\text{exp}}_{\text{IS}}$, $\hat{v}^{\text{exp}}_{\text{MIS}}$, $\hat{v}^{\text{exp}}_{\text{DM}}$, $\hat{v}^{\text{exp}}_{\text{DR}}$, and $\hat{v}^{\text{exp}}_{\text{DRL}}$ in RBRPS1 and RBRPS2. We define the behavior policy profile as $\pi^b_1 = 0.7\pi^d_1 + 0.3\pi^r$ and $\pi^b_2 = 0.7\pi^d_2 + 0.3\pi^p$, where π^r is a deterministic policy that always chooses rock, and π^p is one that always chooses paper. Similarly, we define the target policy profile as $\pi^e_1 = 0.9\pi^d_1 + 0.1\pi^r$ and $\pi^e_2 = 0.5\pi^d_2 + 0.5\pi^p$. We define the policy classes as $\Pi_1 = \Omega_1, \Pi_2 = \Omega_2$. We conduct 100 trials using varying historical data sizes.

Tables 1 and 2 show the root-mean-squared error (RMSE) of each exploitability estimator in RBRPS1 and RBRPS2, where bold font indicates the best estimator in each case. For further details on the results, see the full version of the paper [1]. We find that $\hat{v}^{\text{exp}}_{\text{DR}}$ and $\hat{v}^{\text{exp}}_{\text{DRL}}$ generally outperform the other estimators. Note that $\hat{v}^{\text{exp}}_{\text{DRL}}$ has no advantage over $\hat{v}^{\text{exp}}_{\text{DR}}$ because the current state s_t uniquely determines a trajectory. Because the exploitability evaluation requires estimating best response value using historical data, the estimation error of the discounted value must be small. Therefore, $\hat{v}^{\text{exp}}_{\text{DR}}$ and $\hat{v}^{\text{exp}}_{\text{DRL}}$, with a small estimation error of the discounted value, would perform better than the other estimators.

Table 1: Off-policy exploitability evaluation in RBRPS1: RMSE.

N	$\hat{v}^{\text{exp}}_{\text{IS}}$	$\hat{v}^{\text{exp}}_{\text{MIS}}$	$\hat{v}^{\text{exp}}_{\text{DM}}$	$\hat{v}^{\text{exp}}_{\text{DR}}$	$\hat{v}^{\text{exp}}_{\text{DRL}}$
250	0.085	0.232	4.8×10^{-3}	$\mathbf{3.6 \times 10^{-3}}$	4.5×10^{-3}
500	0.065	0.230	6.9×10^{-5}	$\mathbf{3.6 \times 10^{-5}}$	6.1×10^{-5}
1000	0.044	0.226	2.9×10^{-9}	$\mathbf{1.1 \times 10^{-9}}$	2.5×10^{-9}

Table 2: Off-policy exploitability evaluation in RBRPS2: RMSE.

N	$\hat{v}^{\text{exp}}_{\text{IS}}$	$\hat{v}^{\text{exp}}_{\text{MIS}}$	$\hat{v}^{\text{exp}}_{\text{DM}}$	$\hat{v}^{\text{exp}}_{\text{DR}}$	$\hat{v}^{\text{exp}}_{\text{DRL}}$
250	36.6	11.3	7.07	8.98	**6.52**
500	21.7	11.2	6.04	6.10	**5.56**
1000	15.5	11.1	4.87	**4.33**	4.39

7.3 Best Evaluation Policy Profile Selection

In the second experiment, we analyze the performance of our policy profile selectors in RBRPS1, RBRPS2, and Markov soccer. We compare the five policy profiles $\hat{\pi}^{\text{IS}}$, $\hat{\pi}^{\text{MIS}}$, $\hat{\pi}^{\text{DM}}$, $\hat{\pi}^{\text{DR}}$, and $\hat{\pi}^{\text{DRL}}$, which are selected by each policy profile selector.

In the experiments on RBRPS1 and RBRPS2, we define the behavior policy profile as $\pi^b_1 = 0.5\pi^d_1 + 0.5\pi^r$ and $\pi^b_2 = 0.5\pi^d_2 + 0.5\pi^p$. We define the candidate policy classes as $\Pi_1 = \Omega_1, \Pi_2 = \Omega_2$ in RBRPS1, and set them to $\Pi_1 = \{\{\alpha_1(s)\pi^d_1(s) + (1 - \alpha_1(s))\pi^r(s)\}_{s \in \mathcal{S}} | 0 \leq \alpha_1(s) \leq 1\}$ and $\Pi_2 = \{\{\alpha_2(s)\pi^d_2(s) + (1 - \alpha_2(s))\pi^p(s)\}_{s \in \mathcal{S}} | 0 \leq \alpha_2(s) \leq 1\}$ in RBRPS2. Note that the number of policy parameters is reduced to simplify minimax optimization in RBRPS2. We conduct ten trials in each experiment with a historical data size of 250.

Table 3: Best evaluation policy profile selection in RBRPS: Exploitability (and standard errors).

	π^b	$\hat{\pi}^{\mathrm{IS}}$	$\hat{\pi}^{\mathrm{MIS}}$	$\hat{\pi}^{\mathrm{DM}}$	$\hat{\pi}^{\mathrm{DR}}$	$\hat{\pi}^{\mathrm{DRL}}$
RBRPS1	1.00	0.236(0.04)	0.738(0.05)	0.058(0.01)	**0.036(0.01)**	0.054(0.01)
RBRPS2	39.6	29.2(5.12)	37.4(4.33)	22.5(2.49)	20.5(0.66)	**19.4(0.45)**

Table 4: Best evaluation policy profile selection in Markov soccer: Win rate ×100 (and standard errors).

		Player 2					
		π_2^b	$\hat{\pi}_2^{\mathrm{IS}}$	$\hat{\pi}_2^{\mathrm{MIS}}$	$\hat{\pi}_2^{\mathrm{DM}}$	$\hat{\pi}_2^{\mathrm{DR}}$	$\hat{\pi}_2^{\mathrm{DRL}}$
Player 1	π_1^b	48.9(0.52)	31.7(9.5)	54.2(10.7)	18.2(3.4)	22.6(3.6)	**15.6(0.9)**
	$\hat{\pi}_1^{\mathrm{IS}}$	81.2(3.0)	54.9(7.9)	74.9(8.0)	46.8(6.0)	53.5(5.3)	**44.7(4.7)**
	$\hat{\pi}_1^{\mathrm{MIS}}$	88.1(1.6)	65.5(6.2)	79.7(6.4)	57.8(3.7)	63.2(5.0)	**55.5(3.0)**
	$\hat{\pi}_1^{\mathrm{DM}}$	88.8(3.1)	65.5(6.7)	81.3(6.2)	58.3(6.0)	67.0(4.5)	**56.7(4.9)**
	$\hat{\pi}_1^{\mathrm{DR}}$	89.0(3.0)	**70.0(5.5)**	82.0(5.6)	60.8(5.8)	66.2(6.0)	**57.5(4.1)**
	$\hat{\pi}_1^{\mathrm{DRL}}$	**92.2(1.5)**	69.8(5.9)	**82.5(5.8)**	**63.6(4.5)**	**71.0(5.1)**	**62.4(3.2)**

Table 3 shows the exploitability of each selected policy profile in RBRPS1 and RBRPS2. We find that all selected policies are better than the behavior policy profile. Again, bold font indicates the best policy profile in each case. Notably, $\hat{\pi}^{\mathrm{DR}}$ and $\hat{\pi}^{\mathrm{DRL}}$ outperform the policy profiles obtained by the other estimators.

In the Markov soccer experiment, we define the behavior policy profile as $\pi_1^b = 0.3\pi_1^d + 0.7\pi^u$ and $\pi_2^b = 0.5\pi_2^d + 0.5\pi^u$, where π^u is a uniform random policy. We set the candidate policy classes to $\Pi_1 = \{\alpha_1\pi_1^d + (1-\alpha_1)\pi^u | 0 \leq \alpha_1 \leq 1\}$ and $\Pi_2 = \{\alpha_2\pi_2^d + (1-\alpha_2)\pi^u | 0 \leq \alpha_1 \leq 1\}$. As before, we conduct ten trials in each experiment with a historical data size of 250. Because it is difficult to calculate the exploitability accurately in Markov soccer accurately, we compare the selected policy's winning rates against other policies. Here, we approximate the winning rate using the rate of reaching the goal in $10,000$ games. Note that player 1 has an advantage over player 2 because the possession of the ball always goes to player 1 at the initial state.

Table 4 shows the winning rates of each selected policy in Markov soccer. In this table, we show the winning rate of player 1. The winning rates of $\hat{\pi}_1^{\mathrm{DRL}}$ and $\hat{\pi}_2^{\mathrm{DRL}}$ are generally higher than those of the other policies. Unlike the results in RBRPS, the policy profile selected using $\hat{v}_{\mathrm{DRL}}^{\mathrm{exp}}$ is more robust and better than that obtained using $\hat{v}_{\mathrm{DR}}^{\mathrm{exp}}$. These results suggest that we can select the policy profile the lowest exploitability when using $\hat{v}_{\mathrm{DRL}}^{\mathrm{exp}}$.

8 RELATED WORK

In the context of OPE, there are many previous studies focusing on the theoretical properties of the value estimators, such as the IS [16], MIS [55], DR [9, 13, 14, 18, 31, 41, 51], and DRL [21, 23] estimators. In particular, the DRL estimator has the crucial advantage of using Markov properties to avoid the curse of horizon. The main difference between these studies and our study is that we propose exploitability estimators for OPE in MARL.

There are some studies on inverse MARL that assume the situation where the historical data is obtained in multi-agent environments [27, 38, 40, 54, 56, 59]. These studies differ from ours in that they aim to restore the reward function from the historical data. In contrast, our study uses the historical data to estimate the exploitability of a given policy profile.

MARL in Markov games has been studied extensively in the literature [3, 8, 17, 28, 29, 34, 57]. Most existing studies on MARL focus on online policy learning. In contrast, our study focuses on offline policy evaluation.

As with policy learning in Markov games, there is a large body of literature on policy learning in extensive-form games [12, 15, 35, 43, 48, 62]. These studies focus on developing efficient method for computing Nash equilibria in extensive-form games, such as counterfactual regret minimization [62]. On the other hand, we focus on policy evaluation in Markov games. Various works have investigated policy evaluation in extensive-form games [4, 5, 10, 11, 20, 61]. While these studies have focused on online strategy evaluation with known structure, our study focuses on offline estimating exploitability without structural information.

There are several studies on the offline policy selection in bandit problems or RL [2, 24–26, 50, 60]. Unlike these studies, we propose the policy selection methods in multi-agent settings. Various studies on batch MARL [39, 58] also have considered the off-policy data setting. The most significant difference between these studies and our study is that our study's main objective is to develop OPE estimators in MARL. Furthermore, we consider the situation where candidate policies belong to a restricted policy class. This has advantages in practical situations where only specific policies can be implemented.

9 CONCLUSION

In this study, we proposed estimators for TZMGs. The proposed estimators project the exploitability of a target policy profile from historical data. We proved the exploitability estimation error bounds for the proposed estimators. Besides, we proposed the methods for selecting the best policy profile from a given policy profile class based on our exploitability estimators. We proved the exploitability bounds of the policy profiles selected by the proposed methods. In future studies, we will explore the application of our exploitability estimators in more general settings, such as large extensive-form games.

REFERENCES

[1] Kenshi Abe and Yusuke Kaneko. 2020. Off-Policy Exploitability-Evaluation in Two-Player Zero-Sum Markov Games. *arXiv preprint arXiv:2007.02141* (2020).

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

[3] Yu Bai and Chi Jin. 2020. Provable Self-Play Algorithms for Competitive Reinforcement Learning. *arXiv preprint arXiv:2002.04017* (2020).

[4] Nolan Bard, Michael Johanson, Neil Burch, and Michael Bowling. 2013. Online implicit agent modelling. In *AAMAS*. 255–262.

[5] Michael Bowling, Michael Johanson, Neil Burch, and Duane Szafron. 2008. Strategy evaluation in extensive games with importance sampling. In *ICML*. 72–79.

[6] Noam Brown and Tuomas Sandholm. 2019. Superhuman AI for multiplayer poker. *Science* 365, 6456 (2019), 885–890.

[7] Noam Brown, Tuomas Sandholm, and Strategic Machine. 2017. Libratus: The Superhuman AI for No-Limit Poker.. In *IJCAI*. 5226–5228.

[8] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.

[9] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (2018), C1–C68.

[10] Joshua Davidson, Christopher Archibald, and Michael Bowling. 2013. Baseline: practical control variates for agent evaluation in zero-sum domains.. In *AAMAS*. 1005–1012.

[11] Trevor Davis, Neil Burch, and Michael Bowling. 2014. Using response functions to measure strategy strength. In *AAAI*. 630–636.

[12] Trevor Davis, Martin Schmid, and Michael Bowling. 2020. Low-Variance and Zero-Variance Baselines for Extensive-Form Games. In *ICML*. 2392–2401.

[13] Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. 2014. Doubly robust policy evaluation and optimization. *Statist. Sci.* 29, 4 (2014), 485–511.

[14] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. 2018. More robust doubly robust off-policy evaluation. In *ICML*. 1447–1456.

[15] Richard G Gibson, Marc Lanctot, Neil Burch, Duane Szafron, and Michael Bowling. 2012. Generalized Sampling and Variance in Counterfactual Regret Minimization.. In *AAAI*. 1355–1361.

[16] Keisuke Hirano, Guido W Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 4 (2003), 1161–1189.

[17] Junling Hu and Michael P Wellman. 2003. Nash Q-learning for general-sum stochastic games. *Journal of machine learning research* 4, Nov (2003), 1039–1069.

[18] Nan Jiang and Lihong Li. 2016. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. In *ICML*. 652–661.

[19] Michael Johanson, Nolan Bard, Neil Burch, and Michael Bowling. 2012. Finding optimal abstract strategies in extensive-form games. In *AAAI*. 1371–1379.

[20] Michael Johanson and Michael Bowling. 2009. Data biased robust counter strategies. In *AISTATS*. 264–271.

[21] Nathan Kallus and Masatoshi Uehara. 2019. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850* (2019).

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

[23] Nathan Kallus and Masatoshi Uehara. 2020. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research* 21, 167 (2020), 1–63.

[24] Nathan Kallus and Masatoshi Uehara. 2020. Statistically efficient off-policy policy gradients. In *ICML*. 5089–5100.

[25] Masahiro Kato, Masatoshi Uehara, and Shota Yasui. 2020. Off-Policy Evaluation and Learning for External Validity under a Covariate Shift. *arXiv preprint arXiv:2002.11642* (2020).

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

[27] Xiaomin Lin, Peter A Beling, and Randy Cogill. 2017. Multiagent inverse reinforcement learning for two-person zero-sum games. *IEEE Transactions on Games* 10, 1 (2017), 56–68.

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

[29] Michael L Littman and Csaba Szepesvári. 1996. A generalized reinforcement-learning model: Convergence and applications. In *ICML*. 310–318.

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

[31] Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. 2018. Representation balancing mdps for off-policy policy evaluation. In *NuerIPS*. 2644–2653.

[32] Edward Lockhart, Marc Lanctot, Julien Pérolat, Jean-Baptiste Lespiau, Dustin Morrill, Finbarr Timbers, and Karl Tuyls. 2019. Computing approximate equilibria in sequential adversarial games by exploitability descent. *arXiv preprint arXiv:1903.05614* (2019).

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

[34] Carlos Martin and Tuomas Sandholm. 2020. Efficient exploration of zero-sum stochastic games. *arXiv preprint arXiv:2002.10524* (2020).

[35] Peter McCracken and Michael Bowling. 2004. Safe strategies for agent modelling in games. In *AAAI Fall Symposium on Artificial Multi-agent Learning*. 103–110.

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

[37] John Nash. 1951. Non-cooperative games. *Annals of mathematics* (1951), 286–295.

[38] Sriraam Natarajan, Gautam Kunapuli, Kshitij Judah, Prasad Tadepalli, Kristian Kersting, and Jude Shavlik. 2010. Multi-agent inverse reinforcement learning. In *ICMLA*. 395–400.

[39] Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. 2017. Learning nash equilibrium for general-sum markov games from batch data. In *Artificial Intelligence and Statistics*. 232–241.

[40] Tummalapalli Sudhamsh Reddy, Vamsikrishna Gopikrishna, Gergely Zaruba, and Manfred Huber. 2012. Inverse reinforcement learning for decentralized non-cooperative multiagent systems. In *SMC*. 1930–1935.

[41] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89, 427 (1994), 846–866.

[42] Mohammad Shafiei Nathan Sturtevant Jonathan Schaeffer, N Shafiei, et al. 2009. Comparing UCT versus CFR in simultaneous games. In *IJCAI Workshop on General Game Playing*.

[43] Martin Schmid, Neil Burch, Marc Lanctot, Matej Moravcik, Rudolf Kadlec, and Michael Bowling. 2019. Variance reduction in monte carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines. In *AAAI*. 2157–2164.

[44] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295* (2016).

[45] Lloyd S Shapley. 1953. Stochastic games. *Proceedings of the national academy of sciences* 39, 10 (1953), 1095–1100.

[46] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484.

[47] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *Nature* 550, 7676 (2017), 354–359.

[48] Finnegan Southey, Bret Hoehn, and Robert C Holte. 2009. Effective short-term opponent exploitation in simplified poker. *Machine Learning* 74, 2 (2009), 159–189.

[49] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

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

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

[52] Finbarr Timbers, Edward Lockhart, Martin Schmid, Marc Lanctot, and Michael Bowling. 2020. Approximate exploitability: Learning a best response in large games. *arXiv preprint arXiv:2004.09677* (2020).

[53] Anastasios Tsiatis. 2007. *Semiparametric theory and missing data*. Springer Science & Business Media.

[54] Xingyu Wang and Diego Klabjan. 2018. Competitive multi-agent inverse reinforcement learning with sub-optimal demonstrations. *arXiv preprint arXiv:1801.02124* (2018).

[55] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. 2019. Towards Optimal Off-Policy Evaluation for Reinforcement Learning with Marginalized Importance Sampling. In *NeurIPS*. 9665–9675.

[56] Lantao Yu, Jiaming Song, and Stefano Ermon. 2019. Multi-agent adversarial inverse reinforcement learning. *arXiv preprint arXiv:1907.13220* (2019).

[57] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2019. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635* (2019).

[58] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. 2018. Finite-Sample Analysis For Decentralized Batch Multi-Agent Reinforcement Learning With Networked Agents. *arXiv preprint arXiv:1812.02783* (2018).

[59] Xiangyuan Zhang, Kaiqing Zhang, Erik Miehling, and Tamer Basar. 2019. Non-cooperative inverse reinforcement learning. In *NeurIPS*. 9487–9497.

[60] Zhengyuan Zhou, Susan Athey, and Stefan Wager. 2018. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778* (2018).

[61] Martin Zinkevich, Michael Bowling, Nolan Bard, Morgan Kan, and Darse Billings. 2006. Optimal unbiased estimators for evaluating agent performance. In *AAAI*. 573–579.

[62] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2008. Regret minimization in games with incomplete information. In *NeurIPS*. 1729–1736.