On Faking a Nash Equilibrium

Young Wu
University of Wisconsin-Madison
yw@cs.wisc.edu

Jeremy McMahan
University of Wisconsin-Madison
jmcmahan@wisc.edu

Xiaojin Zhu
University of Wisconsin-Madison
jerryzhu@cs.wisc.edu

Qiaomin Xie
University of Wisconsin-Madison
qiaomin.xie@wisc.edu

Abstract

We characterize offline data poisoning attacks on Multi-Agent Reinforcement Learning (MARL), where an attacker may change a data set in an attempt to install a (potentially fictitious) unique Markov-perfect Nash equilibrium. We propose the unique Nash set, namely the set of games, specified by their Q functions, with a specific joint policy being the unique Nash equilibrium. The unique Nash set is central to poisoning attacks because the attack is successful if and only if data poisoning pushes all plausible games inside it. The unique Nash set generalizes the reward polytope commonly used in inverse reinforcement learning to MARL. For zero-sum Markov games, both the inverse Nash set and the set of plausible games induced by data are polytopes in the Q function space. We exhibit a linear program to efficiently compute the optimal poisoning attack. Our work sheds light on the structure of data poisoning attacks on offline MARL, a necessary step before one can design more robust MARL algorithms.

1 Introduction

Data poisoning attacks are well-known in supervised learning (intentionally forcing the learner to train a wrong classifier) and reinforcement learning (wrong policy) [1, 5, 6, 10, 11, 9, 13, 17, 8, 12, 15, 16]. Can data poisoning attacks be a threat to Markov Games, too? This paper answers this question in the affirmative: Under mild conditions, an attacker can force two game-playing agents to adopt any fictitious Nash Equilibrium (NE), which does not need to be a true NE of the original Markov Game. Furthermore, the attacker can achieve this goal while minimizing its attack cost, which we define below. Obviously, such power poses a threat to the security of Multi-Agent Reinforcement Learning (MARL).

Formally, we study two-player zero-sum offline MARL. Let D be a dataset $\{(s^{(k)}, \mathbf{a}^{(k)}, r^{(k)})\}_{k=1}^{K}$ with K tuples of state s, joint action $\mathbf{a} = (a_1, a_2)$, rewards $(r, -r)$. The attacker's target NE is an arbitrary pure strategy pair $\pi^\dagger := (\pi_1^\dagger, \pi_2^\dagger)$. The attacker can poison D into another dataset $D^\dagger$ by paying cost $C(D, D^\dagger)$. Two MARL agents then receive $D^\dagger$ instead of D. The attacker wants to ensure that the agents learn the target NE $\pi^\dagger$ while minimizing C.

This problem is not well studied in the literature. Naive approaches – such as modifying all the actions in the dataset to those specified by the target policy $(\pi_1^\dagger, \pi_2^\dagger)$ – might not achieve the goal for MARL learners who assign penalties due to the lack of data coverage. Modifying all the rewards in the dataset that coincides with the target policy to the reward upper bound might be feasible, but would not be optimal in terms of attack cost C. Results on data poisoning against single-agent reinforcement learning also cannot be directly applied to the multi-agent case. In particular, there

Preprint. Under review.

are no optimal policies in MARL, and equilibrium policies are computed instead. There could be multiple equilibria that are significantly different, and as a result, installing a target policy as the unique equilibrium is difficult.

Adversarial attacks on MARL have been studied in [7, 3, 4], but we are only aware of one previous work [14] on offline reward poisoning against MARL. Nonetheless, they made the strong assumption that the learners compute the Dominant Strategy Markov Perfect Equilibrium (DSMPE). In contrast, we assume a weaker solution concept, Markov Perfect Equilibrium (MPE). Our general attack framework also accommodates other forms of data poisoning.

Our framework can be summarized by the mnemonic "ToM moves to the UN". (i) UN stands for the Unique Nash set, which is the set of Q functions that make the target $\pi^\dagger$ the unique NE. Uniqueness is crucial for the attacker to ensure that MARL agents choose the target NE with certainty, and not breaking ties arbitrarily among multiple NEs. (ii) ToM stands for the attacker's Theory of Mind of the MARL agents, namely the plausible set of Q functions that the attacker believes the agents will entertain upon receiving the poisoned dataset $D^\dagger$. (iii) The attack is successful if, by controlling $D^\dagger$, the attacker can move the Tom set inside the UN set. A successful attack with the smallest cost $C(D, D^\dagger)$ is optimal.

Summary of Contributions:

- We show that the set of zero-sum Markov games for which a deterministic policy is the unique MPE is equivalent to the set of games for which the policy is a strict MPE, and can be characterized by a polytope in the Q function space.
- We describe a class of MARL learners that compute equilibrium policies based on games within confidence regions around a point estimate of the Q function of the Markov game. With appropriate parameters, an attack on these learners would work on most of the model-based and model-free offline MARL learners proposed in the literature.
- We convert a version of the reward poisoning problem to a linear program that can be solved efficiently, and we provide an attack that is always feasible as long as the sizes of the attacker's confidence regions are sufficiently small.
- We provide a unified framework for offline data poisoning attacks on MARL agents. Our results highlight a security threat to multi-agent reinforcement learning agents, a necessary step before one can design novel MARL algorithms robust to adversarial attacks.

2 Faking a Nash Equilibrium

2.1 The Unique Nash Set (UN) of a Normal-form Game

We present the main components of our approach with a normal-form game, in particular, a two-player zero-sum game is a tuple $(\mathcal{A}, R)$, where $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ is the joint action space and $R : \mathcal{A} \to [-b, b]$ is the mean reward function. We use $b = \infty$ in the case of unbounded rewards. Given $\mathcal{A}$, we denote the set of reward functions by $\mathcal{R} = \{R : \mathcal{A} \to \mathbb{R}\}$.

A pure strategy profile $\pi = (\pi_1, \pi_2)$ is a pair of actions, where $\pi_i \in \mathcal{A}_i$ specifies the action for agent $i \in \{1, 2\}$. We focus on pure strategies, but we allow mixed strategies in which case we use the notation $\pi_i(a_i)$ to represent the probability of i using the action $a_i \in \mathcal{A}_i$, and R computes the expected reward $R(\pi) \coloneqq \sum_{a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2} \pi_1(a_1)\, \pi_2(a_2)\, R((a_1, a_2))$.

Definition 1 (Nash Equilibrium)**.** A Nash equilibrium (NE) of a normal-form game $(\mathcal{A}, R)$ is a mixed strategy profile π that satisfies,

$$R((\pi_1, a_2)) = R(\pi) = R((a_1, \pi_2)), \forall\, a_1 : \pi_1(a_1) > 0, a_2 : \pi_2(a_2) > 0,$$
$$R((\pi_1, a_2)) \leq R(\pi) \leq R((a_1, \pi_2)), \forall\, a_1 : \pi_2(a_1) = 0, a_2 : \pi_2(a_1) = 0,$$

in particular, for a pure strategy profile π, it is a Nash equilibrium if,

$$R((\pi_1, a_2)) \leq R(\pi) \leq R((a_1, \pi_2)), \forall\, a_1 \neq \pi_1, a_2 \neq \pi_2. \tag{1}$$

We define $\mathcal{N}(R) \coloneqq \{\pi : \pi \text{ is an NE of } (\mathcal{A}, R)\}$ to be the set of all Nash equilibria of a normal-form game $(\mathcal{A}, R)$.

Now, we define the inverse image of $\mathcal{N}$ from a single pure strategy profile π back to the space of reward functions to be the unique Nash set.

Definition 2 (Unique Nash)**.** The unique Nash set of a pure strategy profile π is the set of reward functions R such that $(\mathcal{A}, R)$ has a unique Nash equilibrium π,

$$\mathcal{U}(\pi) \coloneqq \mathcal{N}^{-1}(\{\pi\}) = \{R \in \mathcal{R} : \mathcal{N}(R) = \{\pi\}\}. \tag{2}$$

To characterize $\mathcal{U}(\pi)$, we note that for normal-form games, a pure strategy profile π is the unique Nash equilibrium of a game if and only if it is a strict Nash equilibrium, which is defined as a policy π that satisfies (1) with strict inequalities.

Proposition 1 (Unique Nash Polytope)**.** *For any pure strategy profile π,*

$$\begin{aligned}\mathcal{U}(\pi) &= \{R \in \mathcal{R} : \pi \textit{ is a strict NE of } (\mathcal{A}, R)\} \\ &= \{R \in \mathcal{R} : R((\pi_1, a_2)) < R(\pi) < R((a_1, \pi_2)), \forall\, a_1 \neq \pi_1, a_2 \neq \pi_2\}.\end{aligned} \tag{3}$$

Here, the uniqueness is among all Nash equilibria including mixed-strategy Nash equilibria. The proof of the equivalence between (2) and (3) is in the appendix. We restrict our attention to pure-strategy equilibria and defer the discussion of mixed strategy profiles to the last section.

To avoid working with strict inequalities, we define a closed subset of $\mathcal{U}(\pi)$ of reward functions that lead to strict Nash equilibria with an ι reward gap, which means all strict inequalities in (3) are satisfied with a gap of at least ι, for some $\iota > 0$.

Definition 3 (Iota Strict Unique Nash)**.** For $\iota > 0$, the ι strict unique Nash set of a pure strategy profile π is,

$$\underline{\mathcal{U}}(\pi; \iota) \coloneqq \{R \in \mathcal{R} : R((\pi_1, a_2)) + \iota \leq R(\pi) \leq R((a_1, \pi_2)) - \iota, \forall\, a_1 \neq \pi_1, a_2 \neq \pi_2\}. \tag{4}$$

For every pure strategy profile π and $\iota > 0$, we have $\underline{\mathcal{U}}(\pi; \iota) \subset \mathcal{U}(\pi)$, and the set is a polytope in $\mathcal{R}$.

2.2 The Attacker's Theory of Mind (ToM) for Offline Normal-form Game Learners

We provide a model of the attacker's theory of mind of the victim. We assume that the victims compute the Nash equilibria based on the reward functions estimated from a dataset $D \in \mathcal{D}$, where $\mathcal{D}$ is the set of possible datasets with K episodes in the form $\left\{\left(\mathbf{a}^{(k)}, r^{(k)}\right)\right\}_{k=1}^{K}$, with $\mathbf{a}^{(k)} \in \mathcal{A}$ and $r^{(k)} \in [-b, b]$ for every $k \in [K]$.

Definition 4 (Theory of Mind)**.** Given a dataset $D \in \mathcal{D}$, the theory-of-mind set $\mathcal{T}(D) \subseteq \mathcal{R}$ is the set of plausible reward functions that the victims estimate based on D to compute their equilibria. In particular, if the victims learn an action profile π, then $\pi \in \bigcup_{R \in \mathcal{T}(D)} \mathcal{N}(R)$.

The theory-of-mind sets can be arbitrary and could be difficult to work with. We define an outer approximation the set that is a hypercube in $\mathcal{R}$.

Definition 5 (Outer Approximation of Theory of Mind)**.** An outer approximation of $\mathcal{T}(D)$ is a set denoted by $\overline{\mathcal{T}}(D)$ that satisfies $\mathcal{T}(D) \subseteq \overline{\mathcal{T}}(D)$ for every $D \in \mathcal{D}$, and can be written in the form,

$$\overline{\mathcal{T}}(D) = \left\{R \in \mathcal{R} : \left|R(\mathbf{a}) - \hat{R}(\mathbf{a})\right| \leq \rho^{(R)}(\mathbf{a}), \forall\, \mathbf{a} \in \mathcal{A}\right\}, \tag{5}$$

for some point estimate $\hat{R}$ and radius $\rho^{(R)}$.
We call $\overline{\mathcal{T}}(D)$ a linear outer approximation if $\hat{R}$ is linear in $\left\{r^{(k)}\right\}_{k=1}^{K.}$

We present a few examples of the theory-of-mind sets as follows.

Example 1 (Theory of Mind for Maximum Likelihood Victims)**.** Given a dataset $D \in \mathcal{D}$, if the attacker believes the victims are maximum likelihood learners, then $\mathcal{T}(D)$ is a singleton R^{MLE}, where, for every $\mathbf{a} \in \mathcal{A}$,

$$R^{\text{MLE}}(\mathbf{a}) \coloneqq \begin{cases} \dfrac{1}{N(\mathbf{a})} \displaystyle\sum_{k=1}^{K} r^{(k)} \mathbb{1}_{\left\{\mathbf{a}^{(k)} = \mathbf{a}\right\}} & \text{if } N(\mathbf{a}) > 0 \\ 0 & \text{if } N(\mathbf{a}) = 0 \end{cases}; N(\mathbf{a}) \coloneqq \sum_{k=1}^{K} \mathbb{1}_{\left\{\mathbf{a}^{(k)} = \mathbf{a}\right\}}. \tag{6}$$

The smallest outer approximation $\overline{\mathcal{T}}(D)$ can be specified using $\hat{R} = R^{\text{MLE}}$ and $\rho^{(R)} = 0$, and $\overline{\mathcal{T}}$ is linear since (6) is linear in $\left\{r^{(k)}\right\}_{k=1}^{K}$.

Example 2 (Theory of Mind for Pessimistic Optimistic Victims)**.** Given a dataset $D \in \mathcal{D}$, if the attacker believes the victims are learners that use pessimism and optimism by adding and subtracting bonus terms and estimating one or two games, as in [2], then $\mathcal{T}(D)$ may contain two reward functions $\underline{R}$ and $\overline{R}$, where for every $\mathbf{a} \in \mathcal{A}$,

$$\underline{R}(\mathbf{a}) \coloneqq R^{\text{MLE}}(\mathbf{a}) - \beta(\mathbf{a}); \overline{R}(\mathbf{a}) \coloneqq R^{\text{MLE}}(\mathbf{a}) + \beta(\mathbf{a}), \tag{7}$$

with $\beta(\mathbf{a}) = \dfrac{c}{\sqrt{N(\mathbf{a})}}$ being the bonus term, for some constant c.

The smallest outer approximation $\overline{\mathcal{T}}(D)$ can be specified using $\hat{R} = R^{\text{MLE}}$ and $\rho^{(R)}(\mathbf{a}) = \beta(\mathbf{a})$ for every $\mathbf{a} \in \mathcal{A}$, and $\overline{\mathcal{T}}$ is linear since (6) and (7) are both linear in $\left\{r^{(k)}\right\}_{k=1}^{K}$.

Example 3 (Theory of Mind for Data Splitting Victims)**.** Given a dataset $D \in \mathcal{D}$, if the attacker believes the victims use maximum likelihood estimates on a subsample of the D, similar to the data-splitting procedure in [2], then $\overline{\mathcal{T}}(D)$ could be viewed as a high-probability set of rewards that the victims are estimating and $\rho^{(R)}$ would be half of the confidence interval width for the mean of the subsample around the mean of the complete dataset R^{MLE}.

2.3 The Cheapest Way to Move ToM into UN for Normal-form Games

The goal of the attacker is to install a specific action profile as the unique Nash equilibrium of the game learned by the victim while minimally modifying the training data. We consider a general attacker's cost as a function $C : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^{+}$ where $C\left(D, D^{\dagger}\right)$ is the cost of modifying the dataset from D to $D^{\dagger}$. Given the original data set $D \in \mathcal{D}$, the attacker's attack modality $\mathcal{D}(D)$ is the set of datasets the attacker is allowed to modify the original dataset to. For the reward poisoning problem, where $\mathcal{D}^{(R)}(D)$ is all possible datasets in which only rewards are modified from $r^{(k)}$ to $r^{\dagger,(k)}$, we consider the following cost function.

Example 4 (L_1 Cost Function)**.** For reward poisoning problems, we define the L_1 cost of modifying the dataset from $D = \left\{\left(\mathbf{a}^{(k)}, r^{(k)}\right)\right\}_{k=1}^{K}$ to $D^{\dagger} = \left\{\left(\mathbf{a}^{(k)}, r^{\dagger,(k)}\right)\right\}_{k=1}^{K}$ by $C^{(1)}\left(D, D^{\dagger}\right) \coloneqq \sum_{k=1}^{K}\left|r^{(k)} - r^{\dagger,(k)}\right|$.

Now, given the original dataset D and the attacker's target action profile $\pi^{\dagger}$, we formally state the attacker's problem as finding the cheapest way to move $\mathcal{T}(D)$ into $\mathcal{U}\left(\pi^{\dagger}\right)$.

Definition 6 (Attacker's Problem)**.** The attacker's problem with the target action profile $\pi^{\dagger}$ is,

$$\begin{aligned} &\inf_{D^{\dagger} \in \mathcal{D}(D)} C\left(D, D^{\dagger}\right) \\ &\quad s.t.\ \mathcal{T}\left(D^{\dagger}\right) \subseteq \mathcal{U}\left(\pi^{\dagger}\right). \end{aligned} \tag{8}$$

In general, (8) cannot be solved efficiently, but for reward poisoning problems with L_1 cost objective, we can relax the attacker's problem using ι strict unique Nash sets, which is a polytope described by (4), and a linear outer approximation of the theory-of-mind set, a hypercube described by (5), which can be converted into a linear program and solved efficiently. We state this observation as the following proposition and depict the relationship between the sets in Figure 1.

Proposition 2 (Reward Poisoning Linear Program)**.** *Given $\iota > 0$ and a linear $\overline{\mathcal{T}}$, the following problem is a relaxation of the attacker's reward poisoning problem and can be converted into a linear program,*

$$\begin{aligned} &\min_{D^{\dagger} \in \mathcal{D}^{(R)}(D)} C^{(1)}\left(D, D^{\dagger}\right) \\ &\quad s.t.\ \overline{\mathcal{T}}\left(D^{\dagger}\right) \subseteq \underline{\mathcal{U}}\left(\pi^{\dagger}; \iota\right). \end{aligned} \tag{9}$$

In Figure 1, given a dataset D, the general attacker's problem (8) of moving $\mathcal{T}(D)$ (light green) to $\mathcal{T}\left(D^{\dagger}\right)$ (light red) such that it is inside $\mathcal{U}\left(\pi^{\dagger}\right)$ (light blue) while minimizing the distance from D

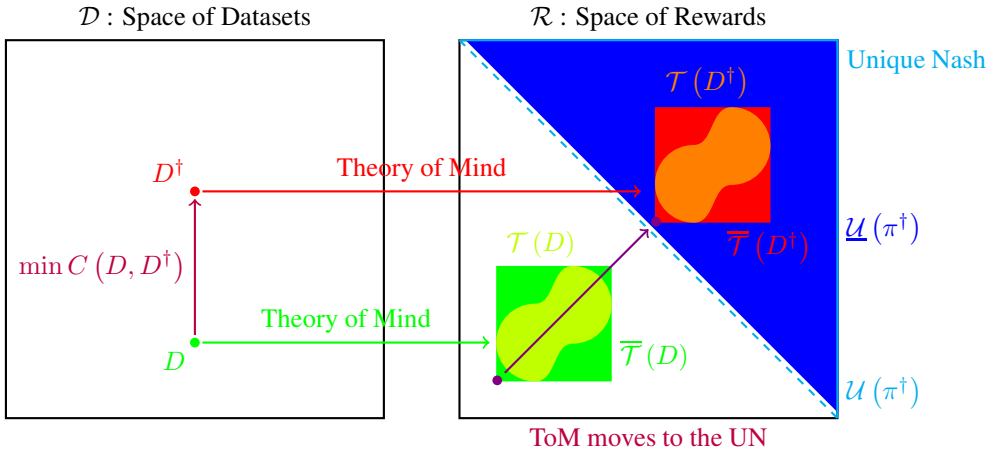

Figure 1: Attacker's Problem

to $D^\dagger$ is often intractable. We construct a relaxed problem (9) of moving $\overline{\mathcal{T}}(D)$ (green) to $\overline{\mathcal{T}}(D^\dagger)$ (red) such that it is inside $\underline{\mathcal{U}}(\pi^\dagger)$ (blue), in which all sets are polytopes and thus can be converted to a linear program for linear costs and linear theory-of-mind mappings.

In the appendix, we provide the complete linear program and show that the solution of (9) is feasible for (8). The optimality of the linear program solution depends on how close the outer approximation of the theory-of-mind set is, and in the case when the theory-of-mind set is already a hypercube, the infimum in (8) can be achieved by taking the limit as $\iota \to 0$. The following is an example illustrating the conversion of (9) into a linear program.

Example 5 (Maximum Likelihood Centered Linear Program)**.** In the case $\hat{R} = R^{\text{MLE}}$ in the theory-of-mind set, (9) is given by,

$$\min_{r^\dagger \in [-b,b]^K} \sum_{k=1}^{K} \left| r^{(k)} - r^{\dagger,(k)} \right| \tag{10}$$
$$s.t.\ R^{\text{MLE}} \text{ is a linear function of } r^\dagger \text{ satisfying (6)}$$
$$\overline{R} \text{ and } \underline{R} \text{ are upper and lower bounds of } \overline{\mathcal{T}}\left(r^\dagger; R^{\text{MLE}}\right) \text{ satisfying (5)}$$
$$\left(\overline{R}, \underline{R}\right) \text{ is in } \underline{\mathcal{U}}\left(\pi^\dagger\right) \text{ satisfying (4)}$$

Since $\overline{\mathcal{T}}\left(r^\dagger; R^{\text{MLE}}\right)$ is a hypercube and $\underline{\mathcal{U}}\left(\pi^\dagger\right)$ is a polytope, the fact that the corners of the hypercube are inside the unique Nash set if and only if every element in the hypercube is in the unique Nash set implies that the constraint in (9) is satisfied. Technically, we only require one corner of the hypercube to be inside the unique Nash polytope, as shown in Figure 1, and we leave the details to the proof of Proposition 2 in the appendix. Then, because the objective and all of the constraints in (10) are linear in $r^\dagger, \overline{R}, \underline{R}$ and R^{MLE}, this problem is a linear program.

3 Faking a Markov Perfect Equilibrium

3.1 The Unique Nash Set (UN) of a Markov Game

We now consider the attacker's problem for Markov games. A finite-horizon two-player zero-sum Markov game G is a tuple $(\mathcal{S}, \mathcal{A}, P, R, H)$, where $\mathcal{S}$ is the finite state space; $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ is the joint action space; $P = \{P_h : \mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}\}_{h=1}^H$ is the transition function with the initial state distribution $P_0 \in \Delta\mathcal{S}$; and $R = \{R_h : \mathcal{S} \times \mathcal{A} \to [-b, b]\}_{h=1}^H$ is the mean reward function; and H is the finite time horizon.

A deterministic Markovian policy $\pi = (\pi_1, \pi_2)$ is a pair of policies, where $\pi_i = \{\pi_{i,h} : \mathcal{S} \to \mathcal{A}_i\}_{h=1}^H$ for $i \in \{1, 2\}$, and $\pi_{i,h}(s)$ specifies the action used in period h and state s. Again, we focus on deterministic policies, but we allow stochastic policies in which case we use

the notation $\pi_i = \{\pi_{i,h} : \mathcal{S} \to \Delta\mathcal{A}_i\}_{h=1}^H$ for $i \in \{1,2\}$, and $\pi_{i,h}(s)(a_i)$ represent the probability of i using the action $a_i \in \mathcal{A}_i$ in period h state s.

The Q function is defined as, for every $h \in [H], s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$,

$$Q_h(s,\mathbf{a}) = R_h(s,\mathbf{a}) + \sum_{s' \in \mathcal{S}} P_h(s'|s,\mathbf{a}) \max_{\pi_1 \in \Delta\mathcal{A}_1} \min_{\pi_2 \in \Delta\mathcal{A}_2} Q_{h+1}(s',\pi), \tag{11}$$

with the convention $Q_{H+1}(s,\mathbf{a}) = 0$, and in the case π is stochastic, we write,

$$Q_h(s,\pi_h(s)) = \sum_{a_1 \in \mathcal{A}_1} \sum_{a_2 \in \mathcal{A}_2} \pi_{1,h}(s)(a_1)\,\pi_{2,h}(s)(a_2)\,Q_h(s,(a_1,a_2)).$$

Given $\mathcal{S}, \mathcal{A}, H$, we denote the set of Q functions by $\mathcal{Q} = \left\{\{Q_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}_{h=1}^H\right\}$. Technically, $\mathcal{Q}$ is not the set of proper Q functions of Markov games since both the reward functions and the transition functions do not have to be proper, and given $Q \in \mathcal{Q}$, we may not be able to construct a Markov game that induces Q. This choice is made to accommodate both model-based and model-free victims who may or may not estimate the rewards and transitions explicitly from the dataset.

A stage game of a Markov game G in period $h \in [H]$, state $s \in \mathcal{S}$ under policy π is a normal form game $(\mathcal{A}, Q_h(s))$, where $\mathcal{A}$ is the joint action space of G; and $Q_h(s)$ is the mean reward function, meaning the reward from action profile $\mathbf{a} \in \mathcal{A}$ is $Q_h(s,\mathbf{a})$. We define Markov perfect equilibria as policies in which the action profile used in every stage game is a Nash equilibrium.

Definition 7 (Markov Perfect Equilibrium)**.** A Markov perfect equilibrium (MPE) policy π is a policy such that $\pi_h(s)$ is a Nash equilibrium in the stage game $(\mathcal{A}, Q_h(s))$.
We define the set of all Markov perfect equilibria policies of a Markov game that induces $Q \in \mathcal{Q}$ by $\mathcal{M}(Q) = \{\pi : \pi \text{ is an MPE of a Markov game with Q function } Q\}$.

We note that Nash equilibria for Markov games can also be defined by converting the Markov game into a single normal-form game, but we only consider Markov perfect equilibria since Nash equilibria that are not Markov perfect require coordination and commitment to policies in stage games that are not visited along equilibrium paths, which is not realistic in the multi-agent reinforcement learning setting.

We define the unique Nash set for Markov games as follows.

Definition 8 (Unique Nash)**.** The unique Nash set of a deterministic Markovian policy π for a Markov game G is the set of Q functions such that π is the unique Markov perfect equilibrium under policy π,

$$\mathcal{U}(\pi) \coloneqq \mathcal{M}^{-1}(\{\pi\}) = \{Q \in \mathcal{Q} : \mathcal{M}(Q) = \{\pi\}\}. \tag{12}$$

Next, we extend the characterization of the unique Nash set for normal-form games to the Markov game setting.

Theorem 1 (Unique Nash Polytope)**.** *For any deterministic policy π,*

$$\begin{aligned}\mathcal{U}(\pi) &= \{Q \in \mathcal{Q} : \pi_h(s) \text{ is a strict NE of } (\mathcal{A}, Q_h(s)), \forall\, h \in [H], s \in \mathcal{S}\} \\ &= \left\{Q \in \mathcal{Q} : \begin{cases} Q_h(s,(\pi_{1,h}(s),a_2)) < Q_h(s,\pi(s)) < Q_h(s,(a_1,\pi_{2,h}(s))), \\ \quad \forall\, a_1 \neq \pi_{1,h}(s), a_2 \neq \pi_{2,h}(s), h \in [H], s \in \mathcal{S} \end{cases}\right\},\end{aligned} \tag{13}$$

We show the equivalence between (12) and (13) in the proof of Theorem 1 in the appendix. To avoid working with strict inequalities in (13), we again define the ι strict version of the unique Nash polytope.

Definition 9 (Iota Strict Unique Nash)**.** For $\iota > 0$, the ι strict unique Nash set of a deterministic policy π is,

$$\underline{\mathcal{U}}(\pi;\iota) \coloneqq \left\{Q \in \mathcal{Q} : \begin{cases} Q_h(s,(\pi_{1,h}(s),a_2)) + \iota \leq Q_h(s,\pi(s)), \\ Q_h(s,\pi(s)) \leq Q_h(s,(a_1,\pi_{2,h}(s))) - \iota, \\ \quad \forall\, a_1 \neq \pi_{1,h}(s), a_2 \neq \pi_{2,h}(s), h \in [H], s \in \mathcal{S} \end{cases}\right\}. \tag{14}$$

For every deterministic policy π and $\iota > 0$, we have $\underline{\mathcal{U}}(\pi;\iota) \subset \mathcal{U}(\pi)$, and the set is a polytope in $\mathcal{Q}$.

3.2 The Attacker's Theory of Mind (ToM) for Offline Multi-Agent Reinforcement Learners

Similar to the theory-of-mind set for normal-form game learners, we define the set for Markov game learners in the $\mathcal{Q}$ space. Here, $\mathcal{D}$ is the set of datasets with K episodes in the form $\left\{\left\{\left(s_h^{(k)}, \mathbf{a}_h^{(k)}, r_h^{(k)}\right)\right\}_{h=1}^{H}\right\}_{k=1}^{K}$ with $s_h^{(k)} \in \mathcal{S}, \mathbf{a}_h^{(k)} \in \mathcal{A}$ and $r_h^{(k)} \in [-b, b]$ for every $k \in [K]$, and the victims compute the Markov perfect equilibria based on the Q functions estimated from such datasets.

Definition 10 (Theory of Mind)**.** Given a dataset $D \in \mathcal{D}$, the theory-of-mind set $\mathcal{T}(D) \subseteq \mathcal{Q}$ is the set of Q functions that the victims estimate based on D to compute their equilibria. In particular, if the victims learn a policy π, then $\pi \in \bigcup_{Q \in \mathcal{T}(D)} \mathcal{M}(Q)$.

Example 6 (Theory of Mind for Maximum Likelihood Victims)**.** To extend Example 1 in the Markov game setting, we define R^{MLE} the same way and P^{MLE} as follows,

$$R_h^{\text{MLE}}(s, \mathbf{a}) \coloneqq \begin{cases} \frac{1}{N_h(s, \mathbf{a})} \sum_{k=1}^{K} r_h^{(k)} \mathbb{1}_{\left\{s_h^{(k)} = s, \mathbf{a}_h^{(k)} = \mathbf{a}\right\}} & \text{if } N_h(s, \mathbf{a}) > 0 \\ 0 & \text{if } N_h(s, \mathbf{a}) = 0 \end{cases}, \tag{15}$$

$$N_h(s, \mathbf{a}) \coloneqq \sum_{k=1}^{K} \mathbb{1}_{\left\{s_h^{(k)} = s, \mathbf{a}_h^{(k)} = \mathbf{a}\right\}},$$

$$P_h^{\text{MLE}}(s' | s, \mathbf{a}) \coloneqq \begin{cases} \frac{\sum_{k=1}^{K} r_h^{(k)} \mathbb{1}_{\left\{s_{h+1}^{(k)} = s', s_h^{(k)} = s, \mathbf{a}_h^{(k)} = \mathbf{a}\right\}}}{N_h(s, \mathbf{a})} & \text{if } N_h(s, \mathbf{a}) > 0, \\ \frac{1}{|\mathcal{S}|} & \text{if } N_h(s, \mathbf{a}) = 0 \end{cases} \tag{16}$$

$$P_0^{\text{MLE}}(s) \coloneqq \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{\left\{s_1^{(k)} = s\right\}}.$$

We can construct Q^{MLE} based on R^{MLE} and P^{MLE} according to (11), and since all Nash equilibria have the same value for zero-sum games, Q^{MLE} is unique for every Markov perfect equilibrium of the Markov game with rewards R^{MLE} and transitions P^{MLE}. Then we have that $\mathcal{T}(D)$ is a singleton Q^{MLE}.

Example 7 (Theory of Mind for Confidence Bound Victims)**.** Given a dataset $D \in \mathcal{D}$, if the attacker believes the victims estimate the Markov game by estimating the rewards and transitions within some confidence region around some point estimates such as the maximum likelihood estimates, as described in [14], then $\mathcal{T}(D)$ would be a polytope with Q functions induced by the Markov games $(\mathcal{S}, \mathcal{A}, P, R, H)$ with P and R satisfying, for every $h \in [H], s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$,

$$R_h(s, \mathbf{a}) \in \mathcal{C}_h^{(R)}(s, \mathbf{a}) \coloneqq \left\{R \in \mathbb{R} : \left|R - \hat{R}_h(s, \mathbf{a})\right| \leq \rho_h^{(R)}(s, \mathbf{a})\right\}, \tag{17}$$

$$P_h(s, \mathbf{a}) \in \mathcal{C}_h^{(P)}(s, \mathbf{a}) \coloneqq \left\{P \in \Delta\mathcal{S} : \left\|P - \hat{P}_h(s, \mathbf{a})\right\|_1 \leq \rho_h^{(P)}(s, \mathbf{a})\right\}, \tag{18}$$

for some point estimates $\hat{P}, \hat{R}$, and radii $\rho^{(R)}$ and $\rho^{(P)}$. We note that $\mathcal{T}(D)$ is a polytope in $\mathcal{Q}$, but it has an exponential number of vertices. We can construct a tight hypercube around this polytope and call it the outer approximation of $\mathcal{T}(D)$. It contains all the Q functions in the following set, for every $h \in [H], s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$,

$$Q_h(s, \mathbf{a}) \in \left[\underline{Q}_h(s, \mathbf{a}), \overline{Q}_h(s, \mathbf{a})\right], \tag{19}$$

$$\underline{Q}_h(s, \mathbf{a}) \coloneqq \min_{R \in \mathcal{C}_h^{(R)}(s, \mathbf{a})} R + \min_{P \in \mathcal{C}_h^{(P)}(s, \mathbf{a})} \sum_{s' \in \mathcal{S}} P(s') \max_{\pi_1 \in \Delta\mathcal{A}_1} \min_{\pi_2 \in \Delta\mathcal{A}_2} \underline{Q}_{h+1}(s', \pi),$$

$$\overline{Q}_h(s, \mathbf{a}) \coloneqq \max_{R \in \mathcal{C}_h^{(R)}(s, \mathbf{a})} R + \max_{P \in \mathcal{C}_h^{(P)}(s, \mathbf{a})} \sum_{s' \in \mathcal{S}} P(s') \max_{\pi_1 \in \Delta\mathcal{A}_1} \min_{\pi_2 \in \Delta\mathcal{A}_2} \overline{Q}_{h+1}(s', \pi).$$

We omit Example 2 and Example 3 for Markov games since the constructions are identical, except it is done for every stage game. As described in Example 7, we formally define the outer approximation of the theory-of-mind set for Markov games as follows.

Definition 11 (Outer Approximation of Theory of Mind)**.** An outer approximation of $\mathcal{T}(D)$ is a set denoted by $\overline{\mathcal{T}}(D)$ that satisfies $\mathcal{T}(D) \subseteq \overline{\mathcal{T}}(D)$ for every $D \in \mathcal{D}$, and can be written in the form,

$$\overline{\mathcal{T}}(D) = \left\{ Q \in \mathcal{Q} : \left| Q_h(s, \mathbf{a}) - \hat{Q}_h(s, \mathbf{a}) \right| \leq \rho_h^{(Q)}(s, \mathbf{a}), \forall\, \mathbf{a} \in \mathcal{A}, h \in [H], s \in \mathcal{S} \right\}, \tag{20}$$

for some point estimate $\hat{Q}$ and radius $\rho^{(Q)}$.

We call $\overline{\mathcal{T}}(D)$ a linear outer approximation if $\hat{Q}$ is linear in $\left\{ \left\{ r_h^{(k)} \right\}_{h=1}^{H} \right\}_{k=1}^{K}$.

3.3 The Cheapest Way to Move ToM into UN for Markov Games

In this subsection, we restate the attacker's problem for multi-agent reinforcement learners.

Definition 12 (Attacker's Problem)**.** The attacker's problem with target policy $\pi^\dagger$ is,

$$\begin{aligned} &\inf_{D^\dagger \in \mathcal{D}(D)} C\left(D, D^\dagger\right) \\ &\quad s.t.\ \mathcal{T}\left(D^\dagger\right) \subseteq \mathcal{U}\left(\pi^\dagger\right). \end{aligned} \tag{21}$$

For reward poisoning problems, we consider the following L_1 cost.

Example 8 (L_1 Cost Function)**.** For reward poisoning problem, where $\mathcal{D}^{(R)}(D)$ is all possible datasets in the form $D^\dagger = \left\{ \left\{ \left(s_h^{(k)}, \mathbf{a}_h^{(k)}, r_h^{\dagger,(k)} \right) \right\}_{h=1}^{H} \right\}_{k=1}^{K}$ that are modified from $D = \left\{ \left\{ \left(s_h^{(k)}, \mathbf{a}_h^{(k)}, r_h^{(k)} \right) \right\}_{h=1}^{H} \right\}_{k=1}^{K}$, we define the L_1 cost by $C^{(1)}\left(D, D^\dagger\right) = \sum_{k=1}^{K} \sum_{h=1}^{H} \left| r_h^{(k)} - r_h^{\dagger,(k)} \right|$.

We use the same ι strictness relaxation of the unique Nash set and the linear outer approximation of the theory-of-mind set to convert (21) into a linear program, which can be solved efficiently. We state this observation as the following theorem.

Theorem 2 (Reward Poisoning Linear Program)**.** *Given $\iota > 0$ and a linear $\overline{\mathcal{T}}$, the following problem is a relaxation of the attacker's reward poisoning problem and can be converted into a linear program,*

$$\begin{aligned} &\min_{D^\dagger \in \mathcal{D}^{(R)}(D)} C^{(1)}\left(D, D^\dagger\right) \\ &\quad s.t.\ \overline{\mathcal{T}}\left(D^\dagger\right) \subseteq \underline{\mathcal{U}}\left(\pi^\dagger; \iota\right). \end{aligned} \tag{22}$$

Example 9 (Maximum Likelihood Centered Linear Program)**.** In the case $\hat{R} = R^{\text{MLE}}$ and $\hat{P} = P^{\text{MLE}}$, and we construct $\overline{\mathcal{T}}(D)$ as described in Example 7, (22) can be converted into a linear program even without explicitly constructing the $\overline{\mathcal{T}}(D)$ set. We provide an intuition here and the formal construction in the proof of Theorem 2,

$$\begin{aligned} &\min_{r^\dagger \in [-b,b]^K} \sum_{k=1}^{K} \sum_{h=1}^{H} \left| r_h^{(k)} - r_h^{\dagger,(k)} \right| \\ &\quad s.t.\ R^{\text{MLE}} \text{ is a linear function of } r^\dagger \text{ satisfying (15)} \\ &\qquad P^{\text{MLE}} \text{ is independent of } r^\dagger \text{ satisfying (16)} \\ &\qquad Q^{\text{MLE}} \text{ is a linear function of } R^{\text{MLE}} \text{ thus } r^\dagger \text{ satisfying (11)} \\ &\qquad \overline{Q} \text{ and } \underline{Q} \text{ are upper and lower bounds of } \overline{\mathcal{T}}\left(r^\dagger; Q^{\text{MLE}}\right) \text{ satisfying (19)} \\ &\qquad \left(\overline{Q}, \underline{Q}\right) \text{ is in } \underline{\mathcal{U}}\left(\pi^\dagger\right) \text{ satisfying (14)} \end{aligned} \tag{23}$$

Similar to Example 5, we move the hypercube $\overline{\mathcal{T}}\left(r^\dagger; Q^{\text{MLE}}\right)$ into the polytope $\underline{\mathcal{U}}\left(\pi^\dagger\right)$ by moving one of the corners into the polytope. Note that if $\overline{Q}$ and $\underline{Q}$ are not constructed directly as linear

functions of $r^\dagger$, and are computed by (19), then these constraints are not linear in $r^\dagger$. We avoid this problem by using the dual linear program of (19). We present the details in the appendix in the proof of Theorem 2. All other constraints are linear in $r^\dagger$, and as a result, (23) is a linear program.

In the end, we present a sufficient but not necessary condition for the feasibility of (22) and (21). This condition applies directly to normal-form games with $H = 1$.

Theorem 3 (Reward Poisoning Linear Program Feasibility)**.** *For $\iota > 0$, $\mathcal{T}(D)$ with $\hat{Q} = Q^{MLE}$, and $N_h(s, \boldsymbol{a}) > 0$ for every $h \in [H], s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}$ where either $a_1 = \pi^\dagger_{1,h}(s)$ or $a_2 = \pi^\dagger_{2,h}(s)$, the attacker's reward poisoning problem is feasible if for every $h \in [H], s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}$,*

$$\rho_h^{(Q)}(s, \boldsymbol{a}) \leq \frac{b - \iota}{4H}. \tag{24}$$

$\mathcal{A}_1 \setminus \mathcal{A}_2$	$1^\dagger$	2	3
$1^\dagger$	0	$-b$	$-b$
2	b	no change	no change
3	b	no change	no change

Table 1: A Feasible Attack

$$r_h^{\dagger,(k)} = \begin{cases} 0 & \text{if } \mathbf{a}_h^{(k)} = \pi^\dagger\left(s_h^{(k)}\right) \\ -b & \text{if } a_{1,h}^{(k)} = \pi_{1,h}^\dagger\left(s_h^{(k)}\right) \\ b & \text{if } a_{2,h}^{(k)} = \pi_{2,h}^\dagger\left(s_h^{(k)}\right) \\ r_h^{(k)} & \text{otherwise} \end{cases} \tag{25}$$

To construct a feasible attack under (24), we use the poisoned rewards in (25). An example where each agent has three actions and the target action profile being action $(1, 1)$ is shown in Table 1. With this $r^\dagger$, the maximum likelihood estimate of the game has a unique Nash equilibrium $\pi_h^\dagger(s)$ with a value of 0 in every stage (h, s). Furthermore, if either the radius of rewards or the radius of Q functions for the theory-of-mind set is less than $\frac{b-\iota}{4H}$, we can show inductively that $\pi_h^\dagger(s)$ remains the unique Nash equilibrium in every stage (h, s), thus showing that every Q function in the theory-of-mind set is also in the unique Nash set, which means the attack is feasible. The complete proof is in the appendix.

4 Discussions

We discuss a few extensions.

- Faking a Unique Mixed Strategy Nash Equilibrium: due to the sensitivity of mixing probabilities from small perturbations of the reward function, as long as the theory-of-mind set has non-zero volume, it is impossible to install a mixed strategy profile (or stochastic policy for Markov games) as the unique equilibrium in general. However, this could be possible when the theory-of-mind set is a singleton. To characterize the unique Nash set for a mixed strategy profile, we need to extend Proposition 1 to include an additional invertibility condition on the reward function, but it is difficult to convert this condition into a linear constraint. We leave the technical details for future work.
- Faking an Optimal Policy for Single-Agent Reinforcement Learners: to attack a single-agent Markov decision process, we observe that a policy π is the unique optimal policy if and only if π is deterministic and is the strict optimal policy. As a result, the unique optimal policy set is also a polytope and can be viewed as a special case of the unique Nash set for a one-player game. In the case of reward poisoning, the attacker's problem can be formulated as a linear program similar to (22).
- Faking a Unique Coarse Correlated Equilibrium in Every Stage Game: for two-player zero-sum Markov games, π is the unique Markov Perfect Coarse Correlated Equilibrium if and only if π is the unique Markov Perfect Equilibrium. Therefore, the results in the previous section apply directly.
- Faking a Unique Markov Perfect Dominant Strategy Equilibria for General-Sum Games: for n-player general-sum Markov games, if π is a deterministic policy and it is a Markov Perfect Strict Dominant Strategy Equilibrium, then π is the unique Markov Perfect Equilibrium. The attacker's formulation in [14] can be viewed as a special case of our results when Nash equilibria are replaced by dominant strategy equilibria.

References

[1] Kiarash Banihashem, Adish Singla, Jiarui Gan, and Goran Radanovic. Admissible policy teaching through reward design. *arXiv preprint arXiv:2201.02185*, 2022.

[2] Qiwen Cui and Simon S Du. When is offline two-player zero-sum markov game solvable? *arXiv preprint arXiv:2201.03522*, 2022.

[3] Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. *arXiv preprint arXiv:1905.10615*, 2019.

[4] Wenbo Guo, Xian Wu, Sui Huang, and Xinyu Xing. Adversarial policy learning in two-player competitive games. In *International Conference on Machine Learning*, pages 3910–3919. PMLR, 2021.

[5] Yunhan Huang and Quanyan Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *International Conference on Decision and Game Theory for Security*, pages 217–237. Springer, 2019.

[6] Guanlin Liu and Lifeng Lai. Provably efficient black-box action poisoning attacks against reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Yuzhe Ma, Young Wu, and Xiaojin Zhu. Game redesign in no-regret game playing. *arXiv preprint arXiv:2110.11763*, 2021.

[8] Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. Policy poisoning in batch reinforcement learning and control. *Advances in Neural Information Processing Systems*, 32:14570–14580, 2019.

[9] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *International Conference on Machine Learning*, pages 7974–7984. PMLR, 2020.

[10] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching in reinforcement learning via environment poisoning attacks. *Journal of Machine Learning Research*, 22(210):1–45, 2021.

[11] Amin Rakhsha, Xuezhou Zhang, Xiaojin Zhu, and Adish Singla. Reward poisoning in reinforcement learning: Attacks against unknown learners in unknown environments. *arXiv preprint arXiv:2102.08492*, 2021.

[12] Anshuka Rangi, Haifeng Xu, Long Tran-Thanh, and Massimo Franceschetti. Understanding the limits of poisoning attacks in episodic reinforcement learning. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3394–3400. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[13] Yanchao Sun, Da Huo, and Furong Huang. Vulnerability-aware poisoning mechanism for online rl with unknown dynamics. *arXiv preprint arXiv:2009.00774*, 2020.

[14] Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie. Reward poisoning attacks on offline multi-agent reinforcement learning. In *The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[15] Haoqi Zhang and David C Parkes. Value-based policy teaching with active indirect elicitation. In *AAAI*, volume 8, pages 208–214, 2008.

[16] Haoqi Zhang, David C Parkes, and Yiling Chen. Policy teaching through reward function learning. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 295–304, 2009.

[17] Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. Adaptive reward-poisoning attacks against reinforcement learning. In *International Conference on Machine Learning*, pages 11225–11234. PMLR, 2020.

5 Supplementary Material

5.1 Proof of Proposition 1 and Theorem 1

We show that for zero-sum games, strict MPEs are MPEs and they are unique. We use the following definition of MPE and strict MPE for zero-sum games rewritten in terms of Q functions. Proposition 1 is a special case of Theorem 1 with $H = |\mathcal{S}| = 1$.

Definition 13. (Markov Perfect Equilibrium for Zero-sum Games)$\pi^\dagger$ is a MPE if for each $h \in [H], s \in \mathcal{S}$,

$$Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \geq Q_h^{\pi^\dagger}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right), \forall\, a_1 \neq \pi_{1,h}^\dagger(s), \tag{26}$$

$$Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \leq Q_h^{\pi^\dagger}\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right), \forall\, a_2 \neq \pi_{2,h}^\dagger(s). \tag{27}$$

Definition 14. (Strict Markov Perfect Equilibrium for Zero-sum Games)$\pi^\dagger$ is a strict MPE if for each $h \in [H], s \in \mathcal{S}$,

$$Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) > Q_h^{\pi^\dagger}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right), \forall\, a_1 \neq \pi_{1,h}^\dagger(s), \tag{28}$$

$$Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) < Q_h^{\pi^\dagger}\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right), \forall\, a_2 \neq \pi_{2,h}^\dagger(s). \tag{29}$$

Proof. Fix a period $h \in [H]$, and assume in periods $h+1, h+2, ..., H$, $\pi^\dagger$ is the unique NE in every state $s \in \mathcal{S}$. This is vacuously true in period H.
First, $\pi_h^\dagger(s)$ is a NE since (28) implies (26) and (29) implies (27).
Now, for a contradiction, assume $\left(a_1', a_2'\right) \neq \pi_h^\dagger(s)$ is another NE in the stage game in period h in some state $s \in \mathcal{S}$, then,

$$Q_h^{\pi^\dagger}\left(s, \left(a_1', a_2'\right)\right) \geq Q_h^{\pi^\dagger}\left(s, \left(\pi_{1,h}^\dagger(s), a_2'\right)\right), \tag{30}$$

$$Q_h^{\pi^\dagger}\left(s, \left(a_1', a_2'\right)\right) \leq Q_h^{\pi^\dagger}\left(s, \left(a_1', \pi_{2,h}^\dagger(s)\right)\right). \tag{31}$$

From the strict MPE conditions,

$$Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \overset{(28)}{>} Q_h^{\pi^\dagger}\left(s, \left(\pi_{1,h}^\dagger(s), a_2'\right)\right), \tag{32}$$

$$Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \overset{(29)}{<} Q_h^{\pi^\dagger}\left(s, \left(a_1', \pi_{2,h}^\dagger(s)\right)\right). \tag{33}$$

Combine the above inequalities, we get,

$$Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \overset{(30),\,(32)}{>} Q_h^{\pi^\dagger}\left(s, \left(a_1', a_2'\right)\right), \tag{34}$$

$$Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \overset{(31),\,(33)}{<} Q_h^{\pi^\dagger}\left(s, \left(a_1', a_2'\right)\right), \tag{35}$$

which is a contradiction.
Therefore, $\pi^\dagger$ is the unique NE in period h, state s. Since h and s are arbitrary, $\pi^\dagger$ is the unique MPE.

□

5.2 Proof of Proposition 2 and Theorem 2

We first write out the complete optimization problem for (23) in Example 9, then we show that the optimization is a relaxation by showing for any $Q^{\pi^\dagger} \in \left[\underline{Q}^{\pi^\dagger}, \overline{Q}^{\pi^\dagger}\right]$ elememtwise, $\pi^\dagger$ is a strict MPE, and as a result Theorem 1 implies its uniqueness. The proof that the problem can be converted into a linear program is similar to LP conversions in [14]. We do not write out the complete LP, and instead we show that each constraint can be converted into a linear constraint. Theorem 2 is a

special case of (23) with given $\underline{Q}^{\pi^\dagger}$ and $\overline{Q}^{\pi^\dagger}$ that are not derived from the rewards and transitions, and Proposition 2 is a special case of Theorem 2 when $H = |\mathcal{S}| = 1$.

$$\min_{r^\dagger \in [0,1]^{HK}} \sum_{k=1}^{K} \sum_{h=1}^{H} \left| r_h^{\dagger,(k)} - r_h^{(k)} \right|$$

$$\text{subject to } R_h(s,\mathbf{a}) = \frac{\sum_{k=1}^{K} \sum_{h=1}^{H} r_h^{(k)} \mathbb{1}_{\left\{s_h^{(k)}=s, \mathbf{a}_h^{(k)}=\mathbf{a}\right\}}}{\max\left\{N_h(s,\mathbf{a}), 1\right\}}, \tag{36}$$

$$\forall\, h \in [H], s \in \mathcal{S}, \mathbf{a} \in \mathcal{A},$$

$$\underline{Q}_h^{\pi^\dagger}(s,\mathbf{a}) = \min_{R \in \mathcal{C}_{i,h}^{(R)}(s,\mathbf{a})} R + \min_{P \in \mathcal{C}_h^{(P)}(s,\mathbf{a})} \sum_{s' \in \mathcal{S}} P(s')\, \underline{Q}_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right), \tag{37}$$

$$\forall\, h \in [H], s \in \mathcal{S}, \mathbf{a} \in \mathcal{A},$$

$$\overline{Q}_h^{\pi^\dagger}(s,\mathbf{a}) = \max_{R \in \mathcal{C}_{i,h}^{(R)}(s,\mathbf{a})} R + \max_{P \in \mathcal{C}_h^{(P)}(s,\mathbf{a})} \sum_{s' \in \mathcal{S}} P(s')\, \overline{Q}_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right), \tag{38}$$

$$\forall\, h \in [H], s \in \mathcal{S}, \mathbf{a} \in \mathcal{A},$$

$$\underline{Q}_{H+1}^{\pi^\dagger}(s,\mathbf{a}) = \overline{Q}_{H+1}^{\pi^\dagger}(s,\mathbf{a}) = 0, \qquad \forall\, s \in \mathcal{S}, \mathbf{a} \in \mathcal{A},$$

$$\underline{Q}_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \geq \overline{Q}_h^{\pi^\dagger}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) + \iota, \tag{39}$$

$$\forall\, h \in [H], s \in \mathcal{S}, a_1 \neq \pi_{1,h}^\dagger(s),$$

$$\overline{Q}_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \leq \underline{Q}_h^{\pi^\dagger}\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right) - \iota, \tag{40}$$

$$\forall\, h \in [H], s \in \mathcal{S}, a_2 \neq \pi_{2,h}^\dagger(s).$$

Since we evaluate the $\underline{Q}$ and $\overline{Q}$ functions on the policy $\pi^\dagger$, we add superscript $\pi^\dagger$ on $\underline{Q}$ and $\overline{Q}$ inside the optimization for clarity.

Proof. Take any $R \in \mathcal{C}^{(R)}$ and $P \in \mathcal{C}^{(P)}$, due to the definition of $\underline{Q}^{\pi^\dagger}$ and $\overline{Q}^{\pi^\dagger}$, which are replicated in (37) and (38), we know that, for each $h \in [H], s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$,

$$\underline{Q}_h^{\pi^\dagger}(s,\mathbf{a}) \leq Q_h^{\pi^\dagger}(s,\mathbf{a}) \leq \overline{Q}_h^{\pi^\dagger}(s,\mathbf{a}). \tag{41}$$

Fix period $h \in [H]$, and assume in periods $h+1, h+2, ..., H$, $\pi^\dagger$ is the Nash equilibrium in every state $s \in \mathcal{S}$. This is vacuously true in period H.
For a fixed $s \in \mathcal{S}$, for any $a_1 \neq \pi_{1,h}^\dagger(s)$,

$$\begin{aligned} Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) &\overset{(41)}{\geq} \underline{Q}_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \\ &\overset{(39)}{\geq} \overline{Q}_h^{\pi^\dagger}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) + \iota \\ &\overset{(41)}{\geq} Q_h^{\pi^\dagger}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) + \iota, \end{aligned} \tag{42}$$

and for any $a_2 \neq \pi_{2,h}^\dagger(s)$,

$$\begin{aligned} Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) &\overset{(41)}{\leq} \overline{Q}_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) \\ &\overset{(40)}{\leq} \underline{Q}_h^{\pi^\dagger}\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right) - \iota \\ &\overset{(41)}{\geq} Q_h^{\pi^\dagger}\left(s, \left(\pi_{2,h}^\dagger(s), a_2\right)\right) - \iota, \end{aligned} \tag{43}$$

(42) and (43) imply that $\pi_h^\dagger(s)$ is the Nash equilibrium in period h state s.
Therefore, $Q^{\pi^\dagger} \in \underline{\mathcal{U}}\left(\pi^\dagger; \iota\right)$, and by Theorem 1, $\pi^\dagger$ is the unique MPE.
Now, to show that the problem can be converted into an LP, we note that (36), (39) and (40) are linear constraints. We only have to convert (37) and (38) into linear constraints, in particular, we convert the following linear program, for some $h \in [H], s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$,

$$\begin{aligned}
\min_{P} & \sum_{s' \in \mathcal{S}} P(s')\, \underline{Q}_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right) \\
\text{subject to } & P(s') \leq \hat{P}_h(s'|s,\mathbf{a}) + \rho_h^{(P)}(s,\mathbf{a}), \forall\, s' \in \mathcal{S}, \\
& P(s') \geq \hat{P}_h(s'|s,\mathbf{a}) - \rho^{(P)}(s,\mathbf{a}), \forall\, s' \in \mathcal{S}, \\
& \sum_{s' \in \mathcal{S}} P(s') = 1, \\
& P(s') \geq 0, \forall\, s' \in \mathcal{S},
\end{aligned}$$

into its dual problem,

$$\begin{aligned}
\max_{\underline{u} \in \mathbb{R}^{\mathcal{S}}, \underline{v} \in \mathbb{R}^{\mathcal{S}}, \underline{w} \in \mathbb{R}} & \sum_{s' \in \mathcal{S}} \hat{P}_h(s'|s,\mathbf{a})\left(\underline{u}_{s'} - \underline{v}_{s'}\right) + \rho_h^{(P)}(s,\mathbf{a})\left(\underline{u}_{s'} + \underline{v}_{s'}\right) + \underline{w} \\
\text{subject to } & \underline{u}_{s'} - \underline{v}_{s'} + \underline{w} \geq -\underline{Q}_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right), \forall\, s' \in \mathcal{S}, \\
& \underline{u}_{s'} \geq 0, \underline{v}_{s'} \geq 0, \forall\, s' \in \mathcal{S}.
\end{aligned}$$

Therefore, (37) can be rewritten as the following linear constraints,

$$\begin{aligned}
\underline{Q}_h^{\pi^\dagger}(s,\mathbf{a}) = & R_h(s,\mathbf{a}) - \rho_h^{(R)}(s,\mathbf{a}) \\
& + \sum_{s' \in \mathcal{S}} \hat{P}_h(s'|s,\mathbf{a})\left(\underline{u}_{s'} - \underline{v}_{s'}\right) + \rho_h^{(P)}(s,\mathbf{a})\left(\underline{u}_{s'} + \underline{v}_{s'}\right) + \underline{w}, \\
\underline{u}_{s'} - \underline{v}_{s'} + \underline{w} \geq & -\underline{Q}_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right), \forall\, s' \in \mathcal{S}, \\
\underline{u}_{s'} \geq 0, & \underline{v}_{s'} \geq 0, \forall\, s' \in \mathcal{S}.
\end{aligned}$$

The similar dual problem can be written out for the $\overline{Q}$ to replace (38),

$$\begin{aligned}
\overline{Q}_h^{\pi^\dagger}(s,\mathbf{a}) = & R_h(s,\mathbf{a}) + \rho_h^{(R)}(s,\mathbf{a}) \\
& + \sum_{s' \in \mathcal{S}} \hat{P}_h(s'|s,\mathbf{a})\left(\overline{u}_{s'} - \overline{v}_{s'}\right) + \rho_h^{(P)}(s,\mathbf{a})\left(\overline{u}_{s'} + \overline{v}_{s'}\right) + \overline{w}, \\
\overline{u}_{s'} - \overline{v}_{s'} + \overline{w} \geq & \overline{Q}_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right), \forall\, s' \in \mathcal{S}, \\
\overline{u}_{s'} \geq 0, & \overline{v}_{s'} \geq 0, \forall\, s' \in \mathcal{S}.
\end{aligned}$$

The linearization of the other $\underline{Q}$ and $\overline{Q}$ constraints are similar.

□

5.3 Proof of Theorem 3

Again, we write the proof for (23) in Example 9, and Theorem 3 is a special case with given $\underline{Q}^{\pi^\dagger}$ and $\overline{Q}^{\pi^\dagger}$ that are not derived from the rewards and transitions. In particular, setting $\rho^{(Q)} = \rho^{(R)}$ and $\rho^{(P)} = 0$ would like to the result stated in Theorem 3. We first provide the intuition behind the proofs. The proof is at the end of this subsection.

Suppose the target action profile is $(1,1)$ in some state s in period h, we show that the target action profile $(1,1)$ is the unique NE for any $Q_h(s,\cdot) \in \left[\underline{Q}_h(s,\cdot), \overline{Q}_h(s,\cdot)\right]$ under the following attack,

$$r_h^{\dagger,(k)} = \begin{cases} -b & \text{if } a_{1,h}^{(k)} \neq \pi_{1,h}^{\dagger}\left(s_h^{(k)}\right), a_{2,h}^{(k)} = \pi_{2,h}^{\dagger}\left(s_h^{(k)}\right) \\ 0 & \text{if } a_{1,h}^{(k)} = \pi_{1,h}^{\dagger}\left(s_h^{(k)}\right), a_{2,h}^{(k)} = \pi_{2,h}^{\dagger}\left(s_h^{(k)}\right) \\ b & \text{if } a_{1,h}^{(k)} = \pi_{1,h}^{\dagger}\left(s_h^{(k)}\right), a_{2,h}^{(k)} \neq \pi_{2,h}^{\dagger}\left(s_h^{(k)}\right) \\ r_h^{(k)} & \text{otherwise} \end{cases}. \tag{44}$$

To simplify the notations, we define the bounds on the cumulative Q value in period $h+1, h+2, ..., H$ as,

$$\underline{S}_h = \sum_{h'=h+1}^{H} \min_{s' \in \mathcal{S}} \underline{Q}_{h'}\left(s', \pi_{h'}^{\dagger}(s')\right)$$

$$\overline{S}_h = \sum_{h'=h+1}^{H} \max_{s' \in \mathcal{S}} \overline{Q}_{h'}\left(s', \pi_{h'}^{\dagger}(s')\right)$$

$\underline{Q}_h(s)$ is lower bounded by,

$\mathcal{A}_1 \setminus \mathcal{A}_2$	1	2	...	$\lvert\mathcal{A}_2\rvert$
1	$0 - \rho_h^{(R)}(s,(1,1)) + \underline{S}_h$	$b - \rho_h^{(R)}(s,(1,2)) + \underline{S}_h$	...	$b - \rho_h^{(R)}(s,(1,\lvert\mathcal{A}_2\rvert)) + \underline{S}_h$
2	$-b - \rho_h^{(R)}(s,(2,1)) + \underline{S}_h$	?	...	?
...	...	...	...	...
$\lvert\mathcal{A}_1\rvert$	$-b - \rho_h^{(R)}(s,(\lvert\mathcal{A}_1\rvert,1)) + \underline{S}_h$	?	...	?

$\overline{Q}_h(s)$ is upper bounded by,

$\mathcal{A}_1 \setminus \mathcal{A}_2$	1	2	...	$\lvert\mathcal{A}_2\rvert$
1	$0 + \rho_h^{(R)}(s,(1,1)) + \overline{S}_h$	$b + \rho_h^{(R)}(s,(1,2)) + \overline{S}_h$	...	$b + \rho_h^{(R)}(s,(1,\lvert\mathcal{A}_2\rvert)) + \overline{S}_h$
2	$-b + \rho_h^{(R)}(s,(2,1)) + \overline{S}_h$	?	...	?
...	...	...	...	...
$\lvert\mathcal{A}_1\rvert$	$-b + \rho_h^{(R)}(s,(\lvert\mathcal{A}_1\rvert,1)) + \overline{S}_h$	?	...	?

For $(1,1)$ to be the ι strict, thus unique, Nash equilibrium for all $Q \in \left[\underline{Q}, \overline{Q}\right]$, sufficient conditions are, for $a_1 \neq 1$ and $a_2 \neq 1$,

$$-\rho_h^{(R)}(s,(1,1)) + \underline{S}_h - \frac{\iota}{2} \geq -\frac{b}{2H}(H-h+1) \geq -b + \rho_h^{(R)}(s,(a_1,1)) + \overline{S}_h + \frac{\iota}{2},$$

$$\rho_h^{(R)}(s,(1,1)) + \overline{S}_h + \frac{\iota}{2} \leq \frac{b}{2H}(H-h+1) \leq b - \rho_h^{(R)}(s,(1,a_2)) + \underline{S}_h - \frac{\iota}{2},$$

which would be true in period 1 if the following is satisfied for $\mathbf{a}$ such that either $a_1 = \pi_{1,h}^{\dagger}(s)$ or $a_2 = \pi_{2,h}^{\dagger}(s)$,

$$\rho_h^{(R)}(s,\mathbf{a}) \leq \frac{b-\iota}{4H} \leq \frac{b}{2H} - \frac{\iota}{2,}$$

which in turn implies,

$$\underline{S}_h \geq -\frac{b}{2H}(H-h+1) + \frac{b}{4H},$$

$$\overline{S}_h \leq \frac{b}{2H}(H-h+1) - \frac{b}{4H}.$$

We provide the formal proof below.

Proof. We assume is satisfied, meaning, for each $h \in [H], s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$,

$$\rho_h^{(R)}(s, \mathbf{a}) \leq \frac{b-\iota}{4H} \leq \frac{b}{2H} - \frac{\iota}{2}. \tag{45}$$

In addition, take $R \in \mathcal{C}^{(R)}$, based on (44), we can compute $\hat{R}$ using (36), and for each $h \in [H], s \in \mathcal{S}$,

$$-\rho_h^{(R)}\left(s, \pi_h^\dagger(s)\right) \leq R_h\left(s, \pi_h^\dagger(s)\right) \leq \rho_h^{(R)}\left(s, \pi_h^\dagger(s)\right), \tag{46}$$

$$\begin{aligned} -b - \rho_h^{(R)}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) &\leq R_h\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) \\ &\leq -b + \rho_h^{(R)}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right), \end{aligned} \tag{47}$$

$$\begin{aligned} b - \rho_h^{(R)}\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right) &\leq R_h\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right) \\ &\leq b + \rho_h^{(R)}\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right). \end{aligned} \tag{48}$$

We proceed by induction. In period H, for $a_1 \neq \pi_{1,H}^\dagger(s)$,

$$\begin{aligned} Q_H^{\pi^\dagger}\left(s, \pi_H^\dagger(s)\right) - \frac{\iota}{2} &= R_H\left(s, \pi_H^\dagger(s)\right) - \frac{\iota}{2} \\ &\overset{(46)}{\geq} -\rho_h^{(R)}\left(s, \pi_H^\dagger(s)\right) - \frac{\iota}{2} \\ &\overset{(45)}{\geq} -\frac{b}{2H} \\ &\geq -b + \frac{b}{2H} \\ &\overset{(45)}{\geq} -b + \rho_h^{(R)}\left(s, \left(a_1, \pi_{2,H}^\dagger(s)\right)\right) + \frac{\iota}{2} \\ &\overset{(47)}{\geq} R_H\left(s, \left(a_1, \pi_{2,H}^\dagger(s)\right)\right) + \frac{\iota}{2} \\ &= Q_h^{\pi^\dagger}\left(s, \left(a_1, \pi_{2,H}^\dagger(s)\right)\right) + \frac{\iota}{2}, \end{aligned} \tag{49}$$

and for $a_2 \neq \pi_{2,H}^\dagger(s)$,

$$\begin{aligned} Q_H^{\pi^\dagger}\left(s, \pi_H^\dagger(s)\right) + \frac{\iota}{2} &= R_H\left(s, \pi_H^\dagger(s)\right) + \frac{\iota}{2} \\ &\overset{(46)}{\leq} \rho_h^{(R)}\left(s, \pi_H^\dagger(s)\right) + \frac{\iota}{2} \\ &\overset{(45)}{\leq} \frac{b}{2H} \\ &\leq b - \frac{b}{2H} \\ &\overset{(45)}{\leq} b - \rho_h^{(R)}\left(s, \left(a_1, \pi_{2,H}^\dagger(s)\right)\right) - \frac{\iota}{2} \\ &\overset{(47)}{\leq} R_H\left(s, \left(\pi_{1,H}^\dagger(s), a_2\right)\right) - \frac{\iota}{2} \\ &= Q_H^{\pi^\dagger}\left(s, \left(\pi_{1,H}^\dagger(s), a_2\right)\right) - \frac{\iota}{2}. \end{aligned} \tag{50}$$

Now, fix a period $h < H$, we assume in periods $h' \in \{h+1, h+2, ..., H\}$, in every state $s \in \mathcal{S}$, $\pi^\dagger$ is the Nash equilibrium, and,

$$-\frac{b}{2}(H - h' + 1) \leq Q_{h'}^{\pi^\dagger}\left(s, \pi_{h'}^\dagger(s)\right) \leq \frac{b}{2}(H - h' + 1). \tag{51}$$

This is true in period H due to (49) and (50).

Now in period h, for a fixed $s \in \mathcal{S}$, for any $a_1 \neq \pi_{1,h}^\dagger(s)$,

$$Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) - \frac{\iota}{2}$$

$$
\begin{aligned}
&= R_h\left(s, \pi_h^\dagger(s)\right) + \sum_{s' \in \mathcal{S}} P_h\left(s' | s, \pi_h^\dagger(s)\right) Q_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right) - \frac{\iota}{2} \\
&\geq R_h\left(s, \pi_h^\dagger(s)\right) + \min_{s' \in \mathcal{S}} Q_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right) - \frac{\iota}{2} \\
&\overset{(51)}{\geq} R_h\left(s, \pi_h^\dagger(s)\right) - \frac{b}{2}(H-h) - \frac{\iota}{2} \\
&\overset{(46)}{\geq} -\rho_h^{(R)}\left(s, \pi_h^\dagger(s)\right) - \frac{b}{2H}(H-h) - \frac{\iota}{2} \\
&\overset{(45)}{\geq} -\frac{b}{2H} - \frac{b}{2H}(H-h) \\
&\geq -\frac{b}{2H}(H-h+1) \\
&\geq -b + \frac{b}{2H} + \frac{b}{2H}(H-h) \\
&\overset{(45)}{\geq} -b + \rho_h^{(R)}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) + \frac{b}{2H}(H-h) + \frac{\iota}{2} \\
&\overset{(47)}{\geq} R_h\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) + \frac{b}{2H}(H-h) + \frac{\iota}{2} \\
&\overset{(51)}{\geq} R_h\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) + \max_{s' \in \mathcal{S}} Q_{h+1}^{\pi^\dagger}\left(s', \left(a_1, \pi_{h+1}^\dagger(s')\right)\right) + \frac{\iota}{2} \\
&\geq R_h\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) + \sum_{s' \in \mathcal{S}} P_h\left(s' | s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) Q_{h+1}^{\pi^\dagger}\left(s', \left(a_1, \pi_{h+1}^\dagger(s')\right)\right) + \frac{\iota}{2} \\
&= Q_h^{\pi^\dagger}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) + \frac{\iota}{2},
\end{aligned} \tag{52}
$$

and for $a_2 \neq \pi_{2,h}^\dagger(s)$,

$$
\begin{aligned}
&Q_h^{\pi^\dagger}\left(s, \pi_h^\dagger(s)\right) + \frac{\iota}{2} \\
&= R_h\left(s, \pi_h^\dagger(s)\right) + \sum_{s' \in \mathcal{S}} P_h\left(s' | s, \pi_h^\dagger(s)\right) Q_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right) + \frac{\iota}{2} \\
&\leq R_h\left(s, \pi_h^\dagger(s)\right) + \max_{s' \in \mathcal{S}} Q_{h+1}^{\pi^\dagger}\left(s', \pi_{h+1}^\dagger(s')\right) + \frac{\iota}{2} \\
&\overset{(51)}{\leq} R_h\left(s, \pi_h^\dagger(s)\right) + \frac{b}{2H}(H-h) + \frac{\iota}{2} \\
&\overset{(46)}{\leq} \rho_h^{(R)}\left(s, \pi_h^\dagger(s)\right) + \frac{b}{2H}(H-H) + \frac{\iota}{2} \\
&\overset{(45)}{\leq} \frac{b}{2H} + \frac{b}{2H}(H-h) \\
&= \frac{b}{2H}(H-h+1) \\
&\leq b - \frac{b}{2H} - \frac{b}{2H}(H-h) \\
&\overset{(45)}{\leq} b + \rho_h^{(R)}\left(s, \left(a_1, \pi_{2,h}^\dagger(s)\right)\right) - \frac{b}{2H}(H-h) - \frac{\iota}{2} \\
&\overset{(47)}{\leq} R_h\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right) - \frac{b}{2H}(H-h) - \frac{\iota}{2} \\
&\overset{(51)}{\leq} R_h\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right) + \min_{s' \in \mathcal{S}} Q_{h+1}^{\pi^\dagger}\left(s', \left(\pi_{1,h}^\dagger(s), a_2\right)\right) - \frac{\iota}{2} \\
&\leq R_h\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right) + \sum_{s' \in \mathcal{S}} P_h\left(s' | s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right) Q_{h+1}^{\pi^\dagger}\left(s', \left(\pi_{1,h}^\dagger(s), a_2\right)\right) - \frac{\iota}{2} \\
&= Q_h^{\pi^\dagger}\left(s, \left(\pi_{1,h}^\dagger(s), a_2\right)\right) - \frac{\iota}{2}.
\end{aligned} \tag{53}
$$

Therefore, $\pi^\dagger$ is the Nash equilibrium in period h state s, and (52) and (53) are consistent (51). By induction, $\pi^\dagger$ is a strict, thus unique, Nash equilibrium in every stage game, making $\pi^\dagger$ the unique MPE.

□