

Soft-Bellman Equilibrium in Affine Markov Games: Forward Solutions and Inverse Learning

Shenghui Chen, Yue Yu, David Fridovich-Keil, and Ufuk Topcu

Abstract—Markov games model interactions among multiple players in a stochastic, dynamic environment. Each player in a Markov game maximizes its expected total discounted reward, which depends upon the policies of the other players. We formulate a class of Markov games, termed *affine Markov games*, where an affine reward function couples the players' actions. We introduce a novel solution concept, the *soft-Bellman equilibrium*, where each player is boundedly rational and chooses a soft-Bellman policy rather than a purely rational policy as in the well-known Nash equilibrium concept. We provide conditions for the existence and uniqueness of the soft-Bellman equilibrium and propose a nonlinear least-squares algorithm to compute such an equilibrium in the *forward problem*. We then solve the *inverse game problem* of inferring the players' reward parameters from observed state-action trajectories via a projected-gradient algorithm. Experiments in a predator-prey OpenAI Gym environment show that the reward parameters inferred by the proposed algorithm outperform those inferred by a baseline algorithm: they reduce the Kullback-Leibler divergence between the equilibrium policies and observed policies by at least two orders of magnitude.

I. INTRODUCTION

Markov games model the interaction of multiple decision makers in stochastic and dynamic environments [1]. In a Markov game, each player's transition and reward depend on the policies of the other players, and each player aims to find an optimal policy that maximizes its expected discounted total reward.

The concept of Nash equilibrium, which refers to a set of policies where no player can benefit by unilaterally changing their policy [1], overlooks the reality that players are often boundedly rational. For example, humans have limited cognitive capacity and are subject to biases and heuristics that can affect their decision-making. As a result, the outcomes of games played by humans may not always align with the predictions from the Nash equilibrium concept. Recent efforts attempt to address this limitation by accounting for players' bounded rationality in games with specific structures, including matrix games [2], fully cooperative games [3–5], and two-player games [6, 7]. Another recent work tackles the same limitation in dynamic games with continuous state and action spaces [8]. However, to our best knowledge, no work has addressed this limitation in *general-sum, multi-player* Markov games with *discrete* state and action spaces yet.

We propose the *soft-Bellman equilibrium* as a new solution concept to capture the dynamics of boundedly rational

players in *affine Markov games*. Affine Markov games are a class of Markov games where each player has independent dynamics and an affine reward function couples the players' actions. In a soft-Bellman equilibrium, each player chooses a policy that maximizes the expected reward with causal entropy regularization while satisfying independent transition dynamics. We provide conditions for the existence and uniqueness of the soft-Bellman equilibrium.

We study the *forward problem* of computing a soft-Bellman equilibrium in a given affine Markov game. We propose a least-squares-based algorithm to solve this problem by minimizing the residuals of the soft-Bellman equilibrium conditions.

We then turn to the *inverse game problem* of inferring the players' reward parameters that best explain observed interactions. We propose an iterative algorithm that leverages the solutions to the forward problem. In each iteration, the algorithm computes the soft-Bellman equilibrium given the current reward parameters and then updates those parameters with a projected-gradient method based on the implicit function theorem [9].

The proposed inverse game algorithm outperforms a baseline algorithm that ignores the coupling between players. Experiments in a predator-prey OpenAI Gym environment [10] show that the reward parameters inferred by the proposed algorithm reduce the Kullback-Leibler divergence between the equilibrium policies and observed policies by at least two orders of magnitude than the baseline algorithm.

II. RELATED WORK

In single-agent settings, literature in inverse reinforcement learning studies the problem of inferring reward parameters from human experts' trajectories. The principle of maximum entropy is a popular approach in this direction [11]. Subsequent studies further extend this principle to accommodate stochastic transitions using causal entropy [12]. For example, recent work extends the maximum causal entropy framework in inverse reinforcement learning to an infinite time horizon setting and proposes the concept of stationary soft-Bellman policy [13]. This policy concept inspires the formulation of the soft-Bellman equilibrium to account for the players' bounded rationality, a feature lacking in the Nash equilibrium concept.

In multi-agent settings, most existing works that try to address the limitation of Nash equilibrium assume specific game structures, including matrix games [2], fully cooperative games [3–5], two-player zero-sum games [6], and two-player general-sum games [7]. This paper generalizes the

S. Chen, Y. Yu, D. Fridovich-Keil, and U. Topcu are with the Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, TX, 78712, USA (emails: shenghui.chen@utexas.edu, yueyu@utexas.edu, dfk@utexas.edu, utopcu@utexas.edu).

existing works to *multi-player, general-sum* Markov games.

First formulated in normal-form and extensive-form games, the quantal response equilibrium is a solution concept to model the bounded rationality of human players [14, 15]. Inspired by this solution concept, recent work proposes the entropic cost equilibrium to extend the quantal response equilibrium to games with *continuous* states and actions [8].

The current work, on the other hand, proposes the soft-Bellman equilibrium to support stochastic transitions in affine Markov games with *discrete* state and action spaces. Although both the entropic cost equilibrium and the soft-Bellman equilibrium extend the quantal response equilibrium, the soft-Bellman equilibrium is different in choosing the state-action frequency matrix, instead of the policy, as the variable to optimize for each player. This subtle difference changes the expected reward from a nonconvex function to a convex one, laying the groundwork for establishing conditions that ensure the existence and uniqueness of solutions.

III. MODELS

We present our main theoretical models: a special class of Markov games, along with a novel equilibrium concept that accounts for bounded rationality.

A. Affine Markov Games

We consider a Markov game [1] where each player solves an MDP with independent dynamics and an affine reward function that couples the players' actions. We let $p \in \mathbb{N}$ denote the number of players. Player $i \in [p]$ solves an MDP specified by a tuple which includes a set of states, a set of actions, a transition kernel, an initial state distribution, a reward matrix, and a discount factor. We let $n^i \in \mathbb{N}$ and $m^i \in \mathbb{N}$ denote the number of states and actions for player i , respectively. We let $S_t^i \in [n^i]$ and $A_t^i \in [m^i]$ denote the state and action of player i at time $t \in \mathbb{N}$. Each action triggers a stochastic transition between the current state to the next state. We let $T^i \in \mathbb{R}^{n^i \times m^i \times n^i}$ denote the *transition kernel* of player i such that

$$T_{saj}^i := \mathbb{P}(S_{t+1}^i = j | S_t^i = s, A_t^i = a) \quad (1)$$

for all $t \in \mathbb{N}$, $s, j \in [n^i]$ and $a \in [m^i]$. We let $q^i \in \mathbb{R}^{n^i}$ denote the *initial state distribution* of player i such that

$$q_s^i := \mathbb{P}(S_0^i = s) \quad (2)$$

for all $s \in [n^i]$. We let $R^i \in \mathbb{R}^{n^i \times m^i}$ denote the reward matrix, where R_{sa}^i denotes the reward of player i for choosing choosing action a in state s . Finally, we let $\gamma \in [0, 1)$ denote a reward discount factor. For each player $i \in [p]$, a stationary policy maps each state to a probability distribution over actions. We denote such a policy as a matrix $\Pi^i \in \mathbb{R}^{n^i \times m^i}$ where

$$\Pi_{sa}^i := \mathbb{P}(A_t^i = a | S_t^i = s) \quad (3)$$

for all $t \in \mathbb{N}$, $s \in [n^i]$, $a \in [m^i]$. An optimal stationary policy in an MDP minimizes the following expected total discounted state-action reward

$$\sum_{t=0}^{\infty} \sum_{s=1}^n \sum_{a=1}^m \gamma^t \mathbb{P}(S_t^i = s, A_t^i = a) R_{sa}^i. \quad (4)$$

We let $Y^i \in \mathbb{R}^{n^i \times m^i}$ denote the state-action frequency matrix of player $i \in [p]$ such that

$$Y_{sa}^i := \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(S_t^i = s, A_t^i = a). \quad (5)$$

for all $s \in [n^i]$ and $a \in [m^i]$.

We now introduce the definition of a p -player affine Markov game.

Definition 1. A p -player affine Markov game is a collection of MDPs $\{\mathcal{M}^i = \{[n^i], [m^i], q^i, T^i, R^i, \gamma\}\}_{i=1}^p$ such that there exists $b^i \in \mathbb{R}^{m^i n^i}$ and $C^{ij} \in \mathbb{R}^{m^i n^i \times m^j n^j}$ for each $i, j \in [p]$ such that

$$\text{vec}(R^i) = b^i + \sum_{j=1}^p C^{ij} \text{vec}(Y^j) \quad (6)$$

for all $i \in [p]$, where $Y^i \in \mathbb{R}^{m^i \times n^i}$ satisfies (5).

The affine reward structure in (6) couples different players' decisions together: the reward for a player is not a fixed number, but depends on the other players' state-action frequencies. Similar coupling appears in matrix games where each player has a finite number of candidate options [16, 17]. This function has two parameters: b pertains to the individual player, and C considers the coupling between the players.

B. Soft-Bellman Equilibrium

We now introduce the notion of *soft-Bellman equilibrium*. It extends the notion of quantal response equilibrium in games with deterministic dynamics to Markov games with stochastic dynamics [14, 15]. Unlike Nash equilibrium, it states that all players choose a soft-Bellman policy—rather than the optimal policy that satisfies the Bellman equations—given other players' actions.

Definition 2. Let $\{\mathcal{M}^i = \{[n^i], [m^i], q^i, T^i, R^i, \gamma\}\}_{i=1}^p$ be an affine Markov game. Let $\Pi^i \in \mathbb{R}^{n^i \times m^i}$ be a stationary policy matrix of player $i \in [p]$. If there exists $v^i \in \mathbb{R}^{n^i}$ and $Q^i \in \mathbb{R}^{n^i \times m^i}$ such that

$$\Pi_{sa}^i = \frac{\exp(Q_{sa}^i)}{\sum_{j=1}^{m^i} \exp(Q_{sj}^i)}, \quad (7a)$$

$$Q_{sa}^i = R_{sa}^i + \gamma \sum_{j=1}^{n^i} T_{saj}^i v_j^i, \quad (7b)$$

$$v_s^i = \log \left(\sum_{a=1}^{m^i} \exp(Q_{sa}^i) \right), \quad (7c)$$

for all $s \in [n^i]$ and $a \in [m^i]$, then $\{\Pi^i\}_{i=1}^p$ is a *soft-Bellman equilibrium* for $\{\mathcal{M}^i\}_{i=1}^p$.

Previous studies have proposed similar notions of equilibrium, e.g., Markov quantal response equilibrium in [18]. However, unlike Equation (7c) in Definition 2, their formulations are inconsistent with the characterization of soft-Bellman policies. There is a close connection between the

soft-Bellman equilibrium and the following optimization over state-action frequency matrix:

$$\begin{aligned} & \underset{Y \in \mathbb{R}^{n^i \times m^i}}{\text{maximize}} && \ell^i(Y) + h(Y) \\ & \text{subject to} && \sum_{a=1}^{m^i} Y_{sa} = q_s^i + \gamma \sum_{j=1}^{n^i} \sum_{a=1}^{m^i} T_{jas}^i Y_{ja}, \quad s \in [n^i], \end{aligned} \quad (8)$$

where

$$\begin{aligned} \ell^i(Y) &:= \text{vec}(Y)^\top b^i + \frac{1}{2} \text{vec}(Y)^\top C^{ii} \text{vec}(Y) \\ &+ \sum_{j=1, j \neq i}^p \text{vec}(Y^j)^\top C^{ij} \text{vec}(Y), \end{aligned} \quad (9a)$$

$$h(Y) := \sum_{s=1}^n \sum_{a=1}^m Y_{sa} \left(\log \left(\sum_{j=1}^m Y_{sj} \right) - \log(Y_{sa}) \right). \quad (9b)$$

The following theorem shows that, if each player chooses its policy by solving optimization (8), then the resulting policies form a soft-Bellman equilibrium.

Theorem 1. Let $\{\mathcal{M}^i = \{[n^i], [m^i], q^i, T^i, R^i, \gamma\}\}_{i=1}^p$ be an affine Markov game. Suppose that $C^{ii} \preceq 0$ and $Y^i \in \mathbb{R}_{>0}^{n^i \times m^i}$ is an optimal solution of optimization (8) for all $i \in [p]$. Let $\Pi^i \in \mathbb{R}^{n^i \times m^i}$ be such that

$$\Pi_{sa}^i = \frac{Y_{sa}^i}{\sum_{j=1}^{m^i} Y_{sj}^i} \quad (10)$$

for all $i \in [p]$, $s \in [n^i]$, and $a \in [m^i]$. Then $\{\Pi^i\}_{i=1}^p$ is a soft-Bellman equilibrium for $\{\mathcal{M}^i\}_{i=1}^p$.

Proof. First, $h(Y)$ is a concave function of matrix Y [13, 19], and $\ell^i(Y)$ is also a concave function of Y since $C^{ii} \preceq 0$. Next, by applying the chain rule to (9a) and (9b) we can show the following:

$$\partial_{Y_{sa}} \ell^i(Y) = R_{sa}^i, \quad \partial_{Y_{sa}} h(Y) = \log \left(\sum_{j=1}^m Y_{sj} \right) - \log(Y_{sa}),$$

where $R^i \in \mathbb{R}^{n^i \times m^i}$ satisfies (6). Since $Y \in \mathbb{R}_{>0}^{n^i \times m^i}$ is an optimal solution for optimization (8), there exists $v^i \in \mathbb{R}^{n^i}$ such that the following Karush-Kuhn-Tucker conditions hold:

$$\sum_{a=1}^{m^i} Y_{sa}^i = q_s^i + \gamma \sum_{j=1}^{n^i} \sum_{a=1}^{m^i} T_{jas}^i Y_{ja}^i, \quad (11a)$$

$$\log(Y_{sa}^i) - \log \left(\sum_{j=1}^{m^i} Y_{sj}^i \right) = Q_{sa}^i - v_s^i, \quad (11b)$$

for all $s \in [n^i]$ and $a \in [m^i]$, where Q_{sa}^i is given by (7b). Let Π_{sa}^i be given by (10), then (11b) implies that

$$\Pi_{sa}^i = \exp(Q_{sa}^i - v_s^i), \quad (12a)$$

$$1 = \sum_{j=1}^{m^i} \Pi_{sj}^i = \sum_{j=1}^{m^i} \exp(Q_{sj}^i - v_s^i), \quad (12b)$$

for all $s \in [n^i]$ and $a \in [m^i]$. By combining (12a) with (12b) one can obtain the condition in (7a). Finally, multiplying both sides of (12b) by $\exp(v_s^i)$ gives (7c), which completes the proof. \square

IV. FORWARD SOLUTION VIA NONLINEAR LEAST-SQUARES

We now establish the existence and uniqueness of a soft-Bellman equilibrium, followed by a discussion on how to compute it by solving a nonlinear least-squares problem. To this end, we introduce the following notation:

$$\begin{aligned} l &:= \sum_{i=1}^p m^i n^i, \quad b := [(b^1)^\top \quad (b^2)^\top \quad \dots \quad (b^p)^\top]^\top, \\ r &:= \sum_{i=1}^p n^i, \quad q := [(q^1)^\top \quad (q^2)^\top \quad \dots \quad (q^p)^\top]^\top. \end{aligned} \quad (13)$$

Let matrices $D^i, E^i \in \mathbb{R}^{n^i \times m^i n^i}$ be such that

$$D^i = I_{n^i} \otimes (\mathbf{1}_{m^i}^\top), \quad E_{kj}^i = T_{\text{quo}(j,m)+1, \text{rem}(j,m), k}^i, \quad (14)$$

for all $k \in [n^i]$ and $j \in [m^i n^i]$. Furthermore, let

$$H := \text{blkdiag}(D^1 - \gamma E^1, D^2 - \gamma E^2, \dots, D^p - \gamma E^p),$$

$$K := \text{blkdiag}((D^1)^\top D^1, (D^2)^\top D^2, \dots, (D^p)^\top D^p),$$

$$C := \begin{bmatrix} C^{11} & C^{12} & \dots & C^{1p} \\ C^{21} & C^{22} & \dots & C^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C^{p1} & C^{p2} & \dots & C^{pp} \end{bmatrix}. \quad (15)$$

With these notations, we are ready to establish the following results on the existence and uniqueness of the soft-Bellman equilibrium.

Theorem 2. Let $\{\mathcal{M}^i = \{[n^i], [m^i], q^i, T^i, R^i, \gamma\}\}_{i=1}^p$ be an affine Markov game where $C^{ii} \preceq 0$ for all $i \in [p]$. Then $Y^i \in \mathbb{R}_{>0}^{n^i \times m^i}$ is an optimal solution of optimization (8) for all $i \in [p]$ if and only if there exists $v \in \mathbb{R}^r$ such that

$$\begin{aligned} \log(y) &= \log(Ky) + b + Cy - H^\top v, \\ Hy &= q, \end{aligned} \quad (16)$$

where

$$y = [\text{vec}(Y^1)^\top \quad \text{vec}(Y^2)^\top \quad \dots \quad \text{vec}(Y^p)^\top]^\top. \quad (17)$$

Furthermore, there exists $y \in \mathbb{R}^l$ such that (16) holds for some $v \in \mathbb{R}^r$. If $C + C^\top \preceq 0$, then such a y is unique.

Proof. First of all, the conditions (16) are the union of the KKT conditions for optimization (8) for all $i \in [p]$. Due to the assumption that $C^{ii} \preceq 0$, $C + C^\top \preceq 0$, and the strict concavity of logarithm function, one can verify that $Y^i \in \mathbb{R}^{n^i \times m^i}$ is an optimal solution of optimization (8) for all $i \in [p]$ if and only if $\{Y^i\}_{i=1}^p$ is a Nash equilibrium of a p -player diagonally strictly concave game, which exists and is unique [16]. \square

As a result of Theorem 2, one can compute a soft-Bellman equilibrium by solving the following nonlinear least-squares problem:

$$\begin{aligned} & \underset{y, v}{\text{minimize}} && \|\log(Ky) + b + Cy - H^\top v - \log(y)\|^2 \\ & && + \|Hy - q\|^2 \end{aligned} \quad (18)$$

Notice that the optimal value of the above optimization is zero, since there exists at least one solution for the nonlinear equations in (16).

V. INVERSE LEARNING VIA IMPLICIT DIFFERENTIATION

Given the parameters of an affine Markov game, one can compute a soft-Bellman equilibrium of this game by solving the nonlinear least-squares problem in (18). The question remains, however, of how to infer these parameters such that they best explain observed decisions, a problem also known as the *inverse game*. Next, we answer this question by developing a projected-gradient method for parameter calibration.

The inverse game problem is a parameter optimization problem defined as follows. We start with a set of empirically observed equilibrium state-action frequencies,

$$\hat{Y}^1, \hat{Y}^2, \dots, \hat{Y}^p, \quad (19)$$

where $\hat{Y}_{sa}^i \in \mathbb{R}^{m^i \times n^i}$ denotes the empirical probability for player i to choose action a in state s . Let

$$\hat{y} := [\text{vec}(\hat{Y}^1)^\top \quad \text{vec}(\hat{Y}^2)^\top \quad \dots \quad \text{vec}(\hat{Y}^p)^\top]^\top. \quad (20)$$

To find the best parameters that explain the observed state-action frequency matrices in (20), one can solve the following optimization problem

$$\begin{aligned} & \underset{y, v, b, C}{\text{minimize}} && \|y - \hat{y}\|^2 \\ & \text{subject to} && \log(y) = \log(Ky) + b + Cy - H^\top v, \quad (21) \\ & && Hy = q, \quad b \in \mathbb{B}, \quad C \in \mathbb{D}, \end{aligned}$$

where $\mathbb{B} \subset \mathbb{R}^l$ and $\mathbb{D} \subset \mathbb{R}^{l \times l}$ are closed convex constraint sets for vector b and matrix C , respectively.

Solving problem (21) is numerically challenging because this optimization contains both nonlinear equation constraints and possible positive semi-definite cone constraints in set \mathbb{D} . As a remedy, we propose an approximate projected-gradient method that combines nonlinear equation solving with efficient projections. To this end, we let

$$J = \begin{bmatrix} K \text{diag}(Ky)^{-1} - \text{diag}(y)^{-1} + C & -H^\top \\ H & 0_{r \times r} \end{bmatrix} \quad (22)$$

By using the chain rule and the implicit function theorem [9] one can show that, if (16) holds and matrix J is nonsingular, then

$$\partial_b \|y - \hat{y}\|^2 = -2 [(y - \hat{y})^\top \quad 0_r] J^{-1} \begin{bmatrix} I_l \\ 0_{r \times l} \end{bmatrix}, \quad (23a)$$

$$\partial_C \|y - \hat{y}\|^2 = -2 y_j [(y - \hat{y})^\top \quad 0_r] J^{-1} \begin{bmatrix} I_l \\ 0_{r \times l} \end{bmatrix}, \quad (23b)$$

for all $j \in [l]$, where $C_j \in \mathbb{R}^l$ is the j -th column of matrix C . Hence one can compute the approximate gradient for vector b and matrix C that locally decreases the value of the objective function in (21) as follows:

$$\tilde{\nabla}_b \|y - \hat{y}\|^2 := -2 [I_l \quad 0_{l \times r}] (J^\dagger)^\top \begin{bmatrix} y - \hat{y} \\ 0_r \end{bmatrix}, \quad (24a)$$

$$\tilde{\nabla}_C \|y - \hat{y}\|^2 := -2 [I_l \quad 0_{l \times r}] (J^\dagger)^\top \begin{bmatrix} y - \hat{y} \\ 0_r \end{bmatrix} y^\top. \quad (24b)$$

Notice that we approximate J^{-1} with the Moore-Penrose pseudoinverse J^\dagger in (24). Such an approximation is exact if J is nonsingular, and still well-defined even if J is singular or J^{-1} is numerically unstable to compute.

Based on the formulas in (24), we propose an approximate projected-gradient method, summarized in Algorithm 1, to solve optimization (21), where we let

$$\text{Proj}_{\mathbb{B}}(b) := \underset{z \in \mathbb{B}}{\text{argmin}} \|z - b\|, \quad (25a)$$

$$\text{Proj}_{\mathbb{D}}(C) := \underset{X \in \mathbb{D}}{\text{argmin}} \|X - C\|_F, \quad (25b)$$

for all $b \in \mathbb{R}^l$ and $C \in \mathbb{R}^{l \times l}$. Each iteration of this method first solves the nonlinear least-squares problem in (18), then performs a projected-gradient step on b and C .

Algorithm 1 Approximated projected-gradient method.

Input: Step size $\alpha \in \mathbb{R}_{>0}$, number of iterations $k_{\max} \in \mathbb{N}$, random initial parameters $b_{\text{init}} \in \mathbb{R}^l, C_{\text{init}} \in \mathbb{R}^{l \times l}$, tolerance $\epsilon \in \mathbb{R}$.

- 1: Initialize $k = 1, b = b_{\text{init}}, C = C_{\text{init}}$.
- 2: **while** $k < k_{\max}$ **do**
- 3: Solve optimization (18) for y .
- 4: **if** change in $\|y - \hat{y}\|^2 < \epsilon$ **then**
- 5: terminate.
- 6: **end if**
- 7: $b \leftarrow \text{Proj}_{\mathbb{B}}(b - \alpha \tilde{\nabla}_b \|y - \hat{y}\|^2)$ ▷ cf. (24a)
- 8: $C \leftarrow \text{Proj}_{\mathbb{D}}(C - \alpha \tilde{\nabla}_C \|y - \hat{y}\|^2)$ ▷ cf. (24b)
- 9: $k \leftarrow k + 1$
- 10: **end while**

Output: Vector b and matrix C .

VI. EXPERIMENTS

We evaluate the performance of the proposed algorithm against a baseline algorithm that neglects the fact that players' reward functions depend upon each other's actions in a predator-prey OpenAI Gym environment [10]. We solve the forward problem by specifying the nonlinear least-squares problem (18) in Julia [20] using the JuMP [21] interface and the COIN-OR IPOPT [22] optimizer. The source code is publicly available at https://github.com/vivianchen98/Inverse_MDPGame.

A. Baseline

The baseline algorithm is a decoupled version of Algorithm 1, that is, it solves optimization (8) with the coupling parameter $C^{ij} = 0$ for all players $i, j \in [p]$. Dropping this parameter frees the baseline algorithm to solve an optimization for each player independently, similar to many existing multi-agent inverse reinforcement learning algorithms.

B. Algorithm Parameters

For the projected-gradient method in Algorithm 1, we use a backtracking line search technique to fine-tune the step size in line 7 and 8 based upon the Armijo (sufficient decrease) condition [23]. Each iteration starts with an initial step size

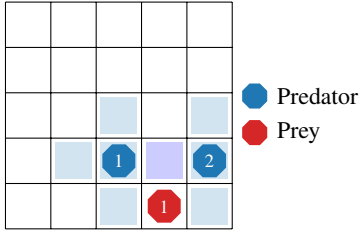


Fig. 1: Two predators (blue) and one prey (red) moving in a 5x5 GridWorld. The light blue cells represent the catching region of the predators, and the light purple cell represents the overlapping of both predators’ catching regions. This episode terminates when the prey is inside a light purple cell.

$\alpha = 1$, and the algorithm reduces the step size by half until it meets the sufficient decrease condition. Both algorithms terminate when the change in $\|y - \hat{y}\|^2$ is below a given tolerance $\epsilon = 0.005$. The maximum number of iterations k_{\max} is 100, and the discount factor γ is 0.99. We sample the values of the vector b_{init} and the matrix C_{init} from a random number generator given a seed. We run both algorithms from seed 1 to 10.

C. Predator-Prey Environment

We consider a predator-prey environment from a collection of multi-agent environments based on OpenAI Gym [10]. As shown in Fig. 1, two predators attempt to capture one randomly moving prey in a 5×5 GridWorld. Each predator has observations of all players and the coordinates of the prey relative to itself and selects one of five actions: left, right, up, down, or stop. The prey is caught when it is within the catching region (light blue cells in Fig. 1) of at least one predator. An episode terminates when the prey is caught by more than one predator (inside a light purple cell in Fig. 1), resulting in a positive reward. For every new episode, the environment initializes the prey into random locations and the prey never moves voluntarily into the predators’ neighborhood. In this environment, only the two predators are controllable, but we collect the trajectories of all three players, including the prey, to solve the inverse game problem.

D. Observed Dataset Collection

We collect all players’ trajectories as the observed interactions. Each trajectory is a sequence of states and actions until termination for the current episode. We train a policy using a multi-agent reinforcement learning algorithm [24] and sample trajectories from this policy. The players in this policy exhibit uncertainties in their decision-making process that are difficult to articulate explicitly, much like humans. As a result, the data from these models can serve as a proxy for human datasets.

We process the collected trajectories from all three players by first pruning those shorter than the 50th percentile of trajectory lengths and then capping the remaining trajectories to the same length. After processing, we attain 100 useful trajectories of length 6. We compute the collection of state-action frequencies for all three players \hat{y} and approximate

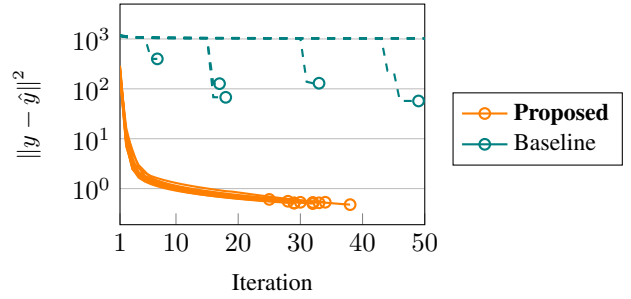


Fig. 2: Algorithms for the inverse game problem with termination marked in circles (the lower the better).

the initial state distributions and the transition probabilities for all players using the observed data.



Fig. 3: Heatmaps showing Kullback–Leibler divergence $D_{KL}(\Pi_s^i \parallel \hat{\Pi}_s^i)$ between the equilibrium policy Π_s^i and the observed policy $\hat{\Pi}_s^i$ at each state s in the GridWorld for all three players. All values rounded to two decimal places, the smaller (lighter color) the better.

E. Numerical Results

We demonstrate Algorithm 1 and the baseline algorithm on the predator-prey environment introduced in Section VI-C. Fig. 2 shows $\|y - \hat{y}\|^2$, the squared norm of the difference between the computed state-action frequency y and the observed state-action frequency matrices \hat{y} , with respect to the number of iterations. Results show the proposed algorithm ends in 31.0 ± 3.6 iterations, while the baseline algorithm takes 50.0 ± 31.3 iterations to terminate. As shown in Fig. 2, the final iterate of the proposed algorithm has $\|y - \hat{y}\|^2$ below 1, while the baseline algorithm on average terminates with a value above 590.7. This comparison highlights the

importance of accounting for the coupling between the players.

Given a state-action frequency matrices y^i for player i , we compute the corresponding policies Π_{sa}^i by (10), and denote the equilibrium policy at each state s as a probability distribution

$$\Pi_s^i = [\Pi_{s,\text{left}}^i \quad \Pi_{s,\text{right}}^i \quad \Pi_{s,\text{up}}^i \quad \Pi_{s,\text{down}}^i \quad \Pi_{s,\text{stop}}^i].$$

We report the Kullback–Leibler divergence $D_{\text{KL}}(\Pi_s^i \parallel \hat{\Pi}_s^i)$ between the equilibrium policy Π_s^i , computed using the proposed and the baseline algorithms, and the observed policy $\hat{\Pi}_s^i$ at each state s for all three players. Fig. 3 shows that Algorithm 1 arrives at an equilibrium policy closer to the observed policy than the baseline algorithm does.

VII. CONCLUSION & FUTURE WORK

We proposed soft-Bellman equilibrium as a novel solution concept in affine Markov games, a class of Markov games where an affine reward function couples the players’ actions, to capture interactions of boundedly rational players in stochastic, dynamic environments. We provided conditions for the existence and uniqueness of the soft-Bellman equilibrium. We solved the forward problem of computing such an equilibrium for a given affine Markov game and proposed an algorithm to tackle the inverse game problem of inferring players’ reward parameters from observed interactions.

Future work should validate the effectiveness of the proposed algorithms using human datasets instead of synthetic datasets. For example, the INTERACTION dataset contains human driving trajectories in interactive traffic scenes [25], and can serve as a more representative dataset for the inverse game problem.

ACKNOWLEDGMENT

The authors would like to thank Negar Mehr and Xiao Xiang for their constructive feedback.

REFERENCES

- [1] Lloyd S Shapley. “Stochastic games”. In: *Proceedings of the National Academy of Sciences* (1953).
- [2] Kevin Waugh, Brian D Ziebart, and J Andrew Bagnell. “Computational rationalization: The inverse equilibrium problem”. In: *International Conference on Machine Learning (ICML)* (2013).
- [3] Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. “Making friends on the fly: Cooperating with new teammates”. In: *Artificial Intelligence* (2017).
- [4] Hoang M Le, Yisong Yue, Peter Carr, and Patrick Lucey. “Coordinated multi-agent imitation learning”. In: *International Conference on Machine Learning (ICML)*. 2017.
- [5] Adrian Sošić, Wasiur R KhudaBukhsh, Abdelhak M Zoubir, and Heinz Koeppl. “Inverse reinforcement learning in swarm systems”. In: *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2016.
- [6] Xiaomin Lin, Peter A Beling, and Randy Cogill. “Multi-agent inverse reinforcement learning for two-person zero-sum games”. In: *IEEE Trans. on Games* (2017).
- [7] Xiaomin Lin, Stephen C Adams, and Peter A Beling. “Multi-agent inverse reinforcement learning for certain general-sum stochastic games”. In: *Journal of Artificial Intelligence Research* (2019).

- [8] Negar Mehr, Mingyu Wang, Maulik Bhatt, and Mac Schwager. “Maximum-entropy multi-agent dynamic games: Forward and inverse solutions”. In: *IEEE Trans. on Robotics* (2023).
- [9] Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings: A view from variational analysis*. Springer, 2014.
- [10] Anurag Koul. *ma-gym: Collection of multi-agent environments based on OpenAI gym*. <https://github.com/koulanurag/ma-gym>. 2019.
- [11] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. “Maximum entropy inverse reinforcement learning”. In: *Association for the Advancement of Artificial Intelligence (AAAI)*. 2008.
- [12] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. “The principle of maximum causal entropy for estimating interacting processes”. In: *IEEE Trans. on Information Theory* (2013).
- [13] Zhengyuan Zhou, Michael Bloem, and Nicholas Bambos. “Infinite time horizon maximum causal entropy inverse reinforcement learning”. In: *IEEE Trans. Autom. Control* (2017).
- [14] Richard D McKelvey and Thomas R Palfrey. “Quantal response equilibria for normal form games”. In: *Games and Economic Behavior* (1995).
- [15] Richard D McKelvey and Thomas R Palfrey. “Quantal response equilibria for extensive form games”. In: *Experimental Economics* (1998).
- [16] J Ben Rosen. “Existence and uniqueness of equilibrium points for concave n-person games”. In: *Econometrica: J. Econ. Soc.* (1965).
- [17] Yue Yu, Jonathan Salfity, David Fridovich-Keil, and Ufuk Topcu. “Inverse matrix games with unique quantal response equilibrium”. In: *IEEE Control Syst. Lett.* (2022).
- [18] Steffen Eibelshäuser and David Poensgen. “Markov quantal response equilibrium and a homotopy method for computing and selecting markov perfect equilibria of dynamic stochastic games”. In: *Available at SSRN 3314404* (2019).
- [19] Michael Bloem and Nicholas Bambos. “Infinite time horizon maximum causal entropy inverse reinforcement learning”. In: *Proc. IEEE Conference on Decision Control (CDC)*. 2014.
- [20] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. “Julia: A fresh approach to numerical computing”. In: *SIAM Review* (2017).
- [21] Iain Dunning, Joey Huchette, and Miles Lubin. “JuMP: A modeling language for mathematical optimization”. In: *SIAM Review* (2017).
- [22] Andreas Wächter and Lorenz T Biegler. “On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming”. In: *Mathematical Programming* (2006).
- [23] Larry Armijo. “Minimization of functions having Lipschitz continuous first partial derivatives”. In: *Pacific Journal of Mathematics* (1966).
- [24] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. “Value-decomposition networks for cooperative multi-agent learning”. In: *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2018.
- [25] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Königshof, Christoph Stiller, Arnaud de La Fortelle, et al. “Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps”. In: *arXiv preprint arXiv:1910.03088* (2019).