Stochastic Games with Sensing Costs

Mohamadreza Ahmadi, Suda Bharadwaj, Takashi Tanaka, and Ufuk Topcu

***Abstract*— In real-world games involving autonomous agents making decisions under uncertainty [1], the agents are often subject to sensing and communication limitations. In these cases, it is desirable to win the game, while also minimizing an agent's sensing budget. In particular, in two-player uncertain adversarial environments, where one player enters the opponent's territory, we seek a wining strategy with minimum sensing. In this paper, we consider finite two-player stochastic games, wherein in addition to the conventional cost over states and actions of each player, we include the sensing budget in terms of transfer entropy. We find a set of pure and mixed strategies for such a game via dynamic programming. The application of dynamic programming leads to a set of coupled nonlinear equations that we solve using the modified Arimoto-Blahut algorithm. The efficacy of the proposed method is illustrated by a stochastic unmanned aerial vehicle (UAV) pursuit-evasion game example using the tool AMASE.**

I. INTRODUCTION

Stochastic games, introduced by Shapely [2], are dynamic games over a sequence of stages with probabilistic transitions. At the beginning of each stage the game is in some state. The players select actions and each player receives a payoff that depends on the current state and the chosen actions. The game then moves to a new random state, whose distribution depends on the previous state and the actions chosen by the players. Stochastic games can be used to model and analyze discrete systems operating in an unknown (adversarial) environment. Applications of such games run the gamut of computer networks [3] to economics [4].

In practice, however, the players have a limited sensing or information budget, since information acquisition, processing, and transmission are costly operations [1]. Hence, within a stochastic game setting, the players should design strategies while taking into account sensing and information resources, such as sensors, satellites, and etc.

A Markov decision process (MDPs) [5] is a stochastic game with just one player. Sensing and information optimal policies have been studied in the framework of MDPs. Entropy maximization have been used in reinforcement learning to maximize exploration in the presence of multimodal objectives through randomized policies [6], [7], [8], and was applied to carry out robotic tasks [9]. In another vein, recently in [10], [11], the authors considered a notion from information theory called the *transfer entropy* [12] to characterizes the information flow from the states of the MDP to the policy of the agent (see also [13] and [14] where transfer entropy was used to analyze communication channels with feedback and memoryless networks, respectively). It was also demonstrated that introducing the additional transfer entropy cost to the MDP leads to a randomized policy.

In this paper, we build upon the results in [10], [11] and consider two-player, finite stochastic games with sensing costs in terms of transfer entropy. We demonstrate that such a game has an optimal mixed strategy for the good player and an optimal pure strategy for the adversary. We apply dynamic programming and obtain a set of nonlinear equations, involving a finite number of variables, that the optimal strategies satisfy. We propose a modified version of the Arimoto-Blahut algorithm [15] for solving these nonlinear equations. We show the efficiency of the proposed methodology by applying it to a stochastic UAV pursuit-evasion game example based on the tool AMASE.

The rest of the paper is organized as follows. We describe the stochastic game we study in this paper in Section II. In Section III, we calculate a set of coupled nonlinear equations that the optimal strategies satisfy and propose an iterative algorithm for solving them. In Section IV, we apply the proposed method to a case study of UAVs tracking a ground vehicle. Finally, in Section V, we conclude the paper and provide directions for future research.

Notation: The notations employed in this paper are relatively straightforward. $\mathbb{R}_{\geq 0}$ denotes the set $[0, \infty)$. For a sequence x, we write x^t to denote $(x_1, x_2, \ldots, x_t)$. Upper case symbols such as X are used to represent random variables, while lower case symbols such as x are used to represent a specific realization. We use the natural logarithm $\log(\cdot) = \log_e(\cdot)$ throughout the paper.

II. PROBLEM FORMULATION

Let $\mathcal{T} = \{1, 2, \ldots, T\}$ be a stage (time) index set. We consider a finite, two-player stochastic game given by a tuple $(\{\mathcal{X}_t\}_{t\in\mathcal{T}}, \{\mathcal{U}_t\}_{t\in\mathcal{T}}, \{\mathcal{W}_t\}_{t\in\mathcal{T}}, \{p\}_{t\in\mathcal{T}}, \{c_t\}_{t\in\mathcal{T}})$, where $\mathcal{X}_t$ denotes the state space at stage t, $\mathcal{U}_t$ is the (good) player action space at stage t, $\mathcal{W}_t$ is the adversary action space at stage t, $p_{t+1}(x_{t+1} \mid x_t, u_t, w_t)$ are transition probabilities to a new state given current state and actions at stage t, and $c_t : \mathcal{X}_t \times \mathcal{U}_t \times \mathcal{W}_t \to \mathbb{R}$ the payoff function at stage t. We assume that the sets $\mathcal{X}_t, \mathcal{U}_t$, and $\mathcal{W}_t$ are all finite. We consider mixed control strategies that can be represented by conditional probability distributions of the form $q_t(u_t \mid x_t, u_{t-1})$. Similarly, we consider mixed adversary strategies represented by $q_t(w_t \mid x_t, w_{t-1})$. The joint distribution $\mu_{t+1}(x_{t+1}, u_t, w_t)$ of the state, control and adversary trajectories is uniquely determined by the initial state distribution $p_1(x_1)$, the state transition probability $p_{t+1}(x_{t+1} \mid x_t, u_t, w_t)$, player strategy

The authors are with the Department of Aerospace Engineering and Engineering Mechanics, and the Institute for Computational Engineering and Sciences (ICES), University of Texas, Austin, 201 E 24th St, Austin, TX 78712. e-mail: ({mrahmadi, sbharadwaj,ttanaka,utopcu}@utexas.edu).

distribution $q_t(u_t \mid x_t, u_{t-1})$, and adversary strategy distribution $q_t(w_t \mid x_t, w_{t-1})$ by the recursive formula

$$\mu_{t+1}(x_{t+1}, u_t, w_t) = p_{t+1}(x_{t+1} \mid x_t, u_t, w_t) \times q_t(u_t \mid x_t, u_{t-1}) q_t(w_t \mid x_t, w_{t-1}) \mu_t(x_t, u_{t-1}, w_{t-1}), \tag{1}$$

which represents the game dynamics. In a standard, finite, two-player stochastic game, we seek to find strategies that (antagonistically) operate on the total payoff function

$$J(X^{T+1}, U^T, W^T) := \sum_{t \in \mathcal{T}} \mathbb{E}\ c_t(X_t, U_t, W_t) + \mathbb{E}\ c_{T+1}(X_{T+1}). \tag{2}$$

Then, the game has a value if

$$\min_{\{q_t^u\}_{t\in\mathcal{T}}} \max_{\{q_t^w\}_{t\in\mathcal{T}}} J(X^{T+1}, U^T, W^T) = \max_{\{q_t^w\}_{t\in\mathcal{T}}} \min_{\{q_t^u\}_{t\in\mathcal{T}}} J(X^{T+1}, U^T, W^T). \tag{3}$$

where $q_t^u = q_t(u_t \mid x_t, u_{t-1})$ and $q_t^w = q_t(w_t \mid x_t, w_{t-1})$ for notational convenience. It was shown in [16], [17] that any finite stochastic game has a value. Furthermore, if there is a finite number of players and the action sets and the set of states are finite, then a stochastic game with a finite number of stages always has a Nash equilibrium.

In this study, we additionally consider sensing costs in terms of transfer entropy [18] given by

$$I(X^T \to U^T) = \sum_{t \in \mathcal{T}} I(X^t; U_t \mid U^{t-1}),$$

where the conditional mutual information term $I(X^t; U_t \mid U^{t-1})$ is described as

$$I(X^t; U_t \mid U^{t-1}) = \sum_{\mathcal{X}^{t+1}} \sum_{\mathcal{U}^t} \mu_{t+1}(x_{t+1}, u_t) \log \frac{\mu_{t+1}(u_t \mid x_t, u_{t-1})}{\mu_{t+1}(u_t \mid u_{t-1})}.$$

Our goal is to find strategies $\{q_t^u\}_{t\in\mathcal{T}}$ and $\{q_t^w\}_{t\in\mathcal{T}}$ that solve the following class of stochastic games

$$\min_{\{q_t^u\}_{t\in\mathcal{T}}} \max_{\{q_t^w\}_{t\in\mathcal{T}}} J(X^{T+1}, U^T, W^T) + \frac{1}{\gamma_1} I(X^T \to U^T) - \frac{1}{\gamma_2} I(X^T \to W^T), \tag{4}$$

where $\gamma_1, \gamma_2 > 0$. In particular, in this paper, we are interested in the case $\gamma_2 \to \infty$, since we assume the adversary has full access over its states. That is, the following stochastic game

$$\min_{\{q_t^u\}_{t\in\mathcal{T}}} \max_{\{q_t^w\}_{t\in\mathcal{T}}} J(X^{T+1}, U^T, W^T) + \frac{1}{\gamma_1} I(X^T \to U^T). \tag{5}$$

The above stochastic game formulation can be compared to the one in [19], where the authors considered two-player stochastic games with bounded rationality represented by Kullback-Leibler (KL) constraints between each agent's strategy and a reference one. Unlike stochastic games with bounded rationality, there is no need to specify a predefined reference strategy in our framework and we are rather interested in penalizing sensing.

III. Computing Optimal Strategies

In this section, we calculate the optimal strategies to (5) and the corresponding optimal payoff function at each stage. These solutions are used to find a set of nonlinear equations for the dynamic game based on dynamic programming.

In order to apply dynamic programming, we use the following result.

Proposition 1: Let X, U, and W be random variables assuming values in $\mathcal{X}$, $\mathcal{U}$, and $\mathcal{W}$, respectively. Let $c : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathbb{R}$ be an arbitrary function and a probability distribution $p(x)$ on $\mathcal{X}$ be given. Consider the following optimization problem

$$\min_{q(u|x)} \max_{q(w|x)} \mathbb{E}c(X, U, W) + \frac{1}{\gamma_1} I(X; U) - \frac{1}{\gamma_2} I(X; W), \tag{6}$$

where $I(X;U) = \sum_{\mathcal{X},\mathcal{U}} p(x) q(u \mid x) \log \frac{q(u|x)}{\sum_{\mathcal{X}} p(x) q(u|x)}$ and $I(X;W) = \sum_{\mathcal{X},\mathcal{U}} p(x) q(w \mid x) \log \frac{q(w|x)}{\sum_{\mathcal{X}} p(x) q(w|x)}$. Then, there exist an optimal solution to optimization problem (6), and the optimal solutions satisfy the following equalities $p(x)$-almost everywhere

$$q^*(w \mid x) = \frac{\mu^*(w) \exp\left(\gamma_2 \sum_{\mathcal{U}} q^*(u \mid x) c(x, u, w)\right)}{\sum_{\mathcal{W}} \mu^*(w) \exp\left(\gamma_2 \sum_{\mathcal{U}} q^*(u \mid x) c(x, u, w)\right)} \tag{7}$$

$$q^*(u \mid x) = \frac{\nu^*(u) \exp\left(-\gamma_1 \sum_{\mathcal{W}} q^*(w \mid x) c(x, u, w)\right)}{\sum_{\mathcal{U}} \nu^*(u) \exp\left(-\gamma_1 \sum_{\mathcal{W}} q^*(w \mid x) c(x, u, w)\right)}, \tag{8}$$

$$\mu^*(w) = \sum_{\mathcal{X}} p(x) q^*(w \mid x), \tag{9}$$

$$\nu^*(u) = \sum_{\mathcal{X}} p(x) q^*(u \mid x). \tag{10}$$

Furthermore, as $\gamma_2 \to \infty$, the optimal solution satisfies

$$q^*(w \mid x) = \begin{cases} 1, & w = w^* = \arg\max_{w \in \mathcal{W}} c(x, \cdot, w), \\ 0, & w \in \mathcal{W} \setminus \{w^*\}, \end{cases} \tag{11}$$

$$q^*(u \mid x) = \frac{\nu^*(u) \exp\left(-\gamma_1 c(x, u, w^*)\right)}{\sum_{\mathcal{U}} \nu^*(u) \exp\left(-\gamma_1 c(x, u, w^*)\right)}, \tag{12}$$

and the optimal value of (6) is given by

$$\frac{-1}{\gamma_1} \mathbb{E}^{p(x)} \log \left(\sum_{\mathcal{U}} \nu^*(U) \exp\left(-\gamma_1 c(X, U, W^*)\right) \right). \tag{13}$$

Proof: Please refer to Appendix. ∎

Equation (11) in Proposition 1 implies that, at $\gamma_2 \to \infty$, the adversary adopts a pure strategy as each stage. This observation is consistent with the previous results on finite stochastic games, that, without the an entropy cost, the finite stochastic game problem has pure strategies [2].

It is straightforward to generalize the sufficient conditions of Proposition 1 to the following optimization

$$\min_{q(u|x,z)} \max_{q(w|x)} \mathbb{E}c(X, U, W, Z) + \frac{1}{\gamma_1} I(X; U \mid Z), \tag{14}$$

where Z is a random variable taking values in $\mathcal{Z}$ and $c : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \times \times \mathcal{Z} \to \mathbb{R}$. Using Proposition 1, we can find the following necessary conditions

$$q^*(u|x,z) := \frac{\nu^*(u \mid z) \exp\left(-\gamma_1 \; c(x,u,w^*,z)\right)}{\sum_{\mathcal{U}} \nu^*(u \mid z) \exp\left(-\gamma_1 \; c(x,u,w^*,z)\right)}, \tag{15}$$

wherein, $w^* = \arg\max_{w \in \mathcal{W}} c(x,\cdot,w,\cdot)$, and

$$\nu^*(u \mid z) := \sum_{\mathcal{X}} p(x,z) q^*(u|x,z). \tag{16}$$

Defining the following *partition* function

$$\phi^*(x,z) := \sum_{\mathcal{U}} \nu^*(u \mid z) \exp\left(-\gamma_1 \; c(x,u,w^*,z)\right), \tag{17}$$

then (13), in this case, changes to $\frac{-1}{\gamma_1}\mathbb{E}^{p(x,z)} \log\left(\phi^*(X,Z)\right)$.

A. Necessary Optimality Conditions via Dynamic Programming

Proposition 1 brings forward expressions for the optimal strategies and the optimal cost function at a single stage of the dynamic stochastic game (5). In what follows, we demonstrate how we can use dynamic programming to find a set of nonlinear equations, which solves the stochastic game problem over the game horizon $\mathcal{T}$. We carry out this by backward iteration. Let

$$\mathscr{V}_t(\mu_t(x^t,u^t,w^t)) = \min \max \sum_{l=t}^{T} \left\{ \mathbb{E}^{\mu_t} \; c_l(X_l,U_l,W_l) + \frac{1}{\gamma_1} I(X^l; U_l \mid U^{l-1}) \right\}, \tag{18}$$

For $t = 1,\ldots,T$, the cost-to-go function satisfies the following Bellman equation

$$\mathscr{V}_t\left(\mu_t\left(x^t,u^{t-1},w^{t-1}\right)\right) = \min_{q_t^u} \max_{q_t^w} \left\{ \mathbb{E}^{\mu_t, q_t^u, q_t^w} \, c_t(X_t,U_t,W_t) + \frac{1}{\gamma_1} I(X^t; U_t \mid U^{t-1}) + \mathscr{V}_{t+1}\left(\mu_{t+1}\left(x^{t+1},u^t,w^t\right)\right) \right\}, \tag{19}$$

with terminal condition at stage $t = T+1$ given by

$$\mathscr{V}_{T+1}\left(\mu_T\left(x^{T+1},u^T,w^T\right)\right) = \mathbb{E}^{\mu_{T+1}} c_{T+1}\left(X_{T+1}\right). \tag{20}$$

The next proposition indicates that the the cost-to-go function has a special structure, which will be used in the sequel.

Proposition 2: For each $t \in \mathcal{T} \cup \{T+1\}$, there exists a function $\phi_t(\cdot)$ such that

$$\mathscr{V}_t\left(\mu_t\left(x^t,u^{t-1},w^{t-1}\right)\right) = \frac{-1}{\gamma_1}\mathbb{E}^{\mu_t} \log\left(\phi_t(X_t,U^{t-1})\right). \tag{21}$$

Proof: We prove by induction. The terminal condition (20) implies that $\phi_{T+1}(x_{T+1}) = \exp(-\gamma_1 c_{T+1}(x_{T+1}))$. Thus, (21) holds at stage $T+1$. At this point, assume that there exists a function $\phi_{t+1}(\cdot)$ such that

$$\mathscr{V}_{t+1}\left(\mu_{t+1}\left(x^{t+1},u^t,w^t\right)\right) = \frac{-1}{\gamma_1}\mathbb{E}^{\mu_{t+1}} \log\left(\phi_t(X_{t+1},U^t)\right).$$

Since $\mathbb{E}^{\mu_{t+1}}(\cdot) = \mathbb{E}^{\mu_t, q_t^u, q_t^w, p_{t+1}}(\cdot)$, the righthand side of the above equation can be re-written as

$$\frac{-1}{\gamma_1}\mathbb{E}^{\mu_t,q_t^u,q_t^w} \sum_{\mathcal{X}_{t+1}} p_{t+1}(x_{t+1} \mid X_t,U_t,W_t) \times \log\left(\phi_{t+1}(x_{t+1},U^t)\right).$$

Introducing the function

$$\rho_t(x_t,u^t,w^t) := c_t(x_t,u_t,w_t) - \frac{1}{\gamma_1} \sum_{\mathcal{X}_{t+1}} p_{t+1}(x_{t+1} \mid x_t,u_t,w_t) \log\left(\phi_{t+1}(x_{t+1},u^t)\right), \tag{22}$$

then the Bellman equation (19) can be written as

$$\min_{q_t^u} \max_{q_t^w} \left\{ \mathbb{E}^{\mu_t,q_t^u,q_t^w} \rho_t(X_t,U^t,W^t) + \frac{1}{\gamma_1} I(X^t; U_t \mid U^{t-1}) \right\}, \tag{23}$$

which is the same optimization problem as (14). Therefore, necessary conditions for optimality are as follows

$$q_t^o(w_t|x_t) = \begin{cases} 1, & w_t = w_t^* = \arg\max_{w_t \in \mathcal{W}_t} c(x_t,\cdot,w_t), \\ 0, & w_t \in \mathcal{W}_t \setminus \{w_t^*\}, \end{cases} \tag{24}$$

$$q_t^o(u_t|x_t,u^{t-1}) = \frac{\nu^*(u_t \mid u^{t-1}) \exp\left(-\gamma_1 \; \rho(x_t,u^t,w_*^t)\right)}{\sum_{\mathcal{U}_t} \nu^*(u_t \mid u^{t-1}) \exp\left(-\gamma_1 \; \rho(x_t,u^t,w_*^t)\right)}, \tag{25}$$

$$\nu_t^o(u_t \mid u^{t-1}) := \sum_{\mathcal{X}_t} \mu_t(x_t|u^{t-1}) q_t^o(u_t|x_t,u^{t-1}), \tag{26}$$

$\mu_t(x^t \mid u^{t-1})$-almost everywhere. Note that (24) is derived by noting that $c_t(x_t,u_t,w_t)$ is concave in w_t and therefore $\rho_t(x_t,u^t,w^t)$ is also concave in w^t. Then, with the partition function defined as

$$\phi_t(x_t,u^{t-1}) = \sum_{\mathcal{U}_t} \nu_t^o(u_t \mid u^{t-1}) \exp\left(-\gamma_1 \; \rho(x_t,u^t,w_*^t)\right), \tag{27}$$

we have the optimal value

$$-\frac{1}{\gamma_1}\mathbb{E}^{\mu_t}\left\{\log\left(\phi_t(X^t,U^{t-1})\right)\right\}.$$

■

From the expressions for (25) and (26), we can deduce that we can find the optimal strategy by just considering

$$\mu_{t+1}(x_{t+1},u^t,w^t) = \sum_{\mathcal{X}_t} p_{t+1}(x_{t+1} \mid x_t,u_t,w_t) \times q_t(u_t \mid x_t,u_{t-1}) q_t(w_t \mid x_t) \mu_t(x_t,u^{t-1},w^{t-1}).$$

All in all, we can infer that the optimal solution of the stochastic game (5) satisfies the set of nonlinear coupled equations given in (28).

$$\mu^*_{t+1}(x_{t+1}, u^t, w^t) = \sum_{\mathcal{X}_t} p_{t+1}(x_{t+1} \mid x_t, u_t, w_t) q_t(u_t \mid x_t, u_{t-1}) q_t(w_t \mid x_t) \mu_t(x_t, u^{t-1}, w^{t-1}), \tag{28a}$$

$$\nu^*_t(u_t \mid u^{t-1}) = \sum_{\mathcal{X}_t} \mu^*_t(x^t|u^{t-1}) q^*_t(u_t|x^t, u^{t-1}), \tag{28b}$$

$$\rho^*_t(x_t, u^t, w^t) = c_t(x_t, u_t, w_t) - \frac{1}{\gamma_1} \sum_{\mathcal{X}_{t+1}} p_{t+1}(x_{t+1} \mid x_t, u_t, w_t) \log\left(\phi^*_{t+1}(x_{t+1}, u^t)\right), \tag{28c}$$

$$q^*_t(w_t|x^t) = \begin{cases} 1, & w_t = w^*_t = \arg\max_{w_t \in \mathcal{W}_t} c_t(x_t, \cdot, w_t), \\ 0, & w_t \in \mathcal{W}_t \setminus \{w^*_t\}. \end{cases} \tag{28d}$$

$$\phi^*_t(x_t, u^{t-1}) = \sum_{\mathcal{U}_t} \nu^*_t(u_t \mid u^{t-1}) \exp\left(-\gamma_1 \sum_{\mathcal{W}_t} q^*_t(w_t \mid x_t)\ \rho^*_t(x_t, u^t, w^t)\right), \tag{28e}$$

$$q^*_t(u_t|x^t, u^{t-1}) = \frac{\nu^*_t(u_t \mid u^{t-1}) \exp\left(-\gamma_1 \sum_{\mathcal{W}_t} q^*_t(w_t|x^t)\ \rho^*_t(x_t, u^t, w^t)\right)}{\sum_{\mathcal{U}_t} \nu^*_t(u_t \mid u^{t-1}) \exp\left(-\gamma_1 \sum_{\mathcal{W}_t} q^*_t(w_t|x^t)\ \rho^*_t(x_t, u^t, w^t)\right)}. \tag{28f}$$

B. Solution Using the Modified Arimoto-Blahut Algorithm

In this section, following the footsteps of [20], [11], we propose an algorithm for solving the coupled nonlinear equations (28) with unknowns μ^*, ν^*, ρ^*, ϕ^*, q^{u*} and q^{w*}. We group the equations in (28) into two sets of equations, i.e., equations (28a)-(28b) in one group and the rest in another group. Given ρ^*, ϕ^*, q^{u*} and q^{w*}, equations (28a)-(28b) can be seen as the forward Kolmogorov equation, which can be solved forward in time to obtain μ^* and ν^*. On the other hand, when μ^* and ν^* are fixed, equations (28c)-(28f) represent the backward Bellman equation, which can be solved backward in time to compute ρ^*, ϕ^*, q^{u*} and q^{w*}.

We propose the following boot-strapping technique: first, the forward computation of equations is carried out based on the current best guess of ρ^*, ϕ^*, q^{u*} and q^{w*}, and then the backward computation of equations (28c)-(28f) is performed based on the updated guess of μ^* and ν^*. We then repeat the forward-backward iteration until convergence. We summarize this method in Algorithm 1. The convergence and complexity properties of Algorithm 1 are discussed in [20].

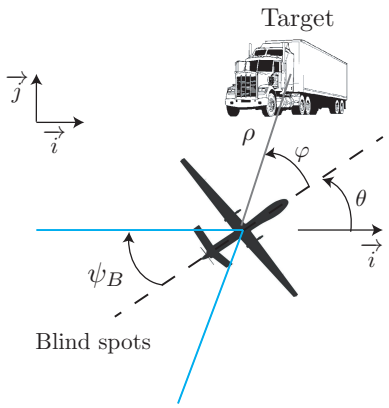


Fig. 1: Schematic diagram of a UAV with a mounted camera.

IV. Case Study: UAV Pursuit-Evasion Scenario

We present a case study of a UAV vision-based target tracking task of a ground vehicle which can be modeled as a stochastic pursuit-evasion game. The goal of the UAV is to keep the adversarial ground vehicle in the field of view of its camera. We use a modified version of the target tracking problem presented in [21] and [22]. In [21], the vision-based target tracking problem was set up as a two-payer stochastic game and solved using dynamic programming. Here, we include the additional sensing cost of transfer entropy to the problem formulation.

A. Game Dynamics

We assume fixed altitude, fixed velocity UAV dynamics also used in [22] described by the Dubins vehicle model.

$$dx = v\cos(\theta)dt \tag{29}$$

$$dy = v\sin(\theta)dt \tag{30}$$

$$d\theta = udt \tag{31}$$

where x, y are the coordinates of UAV position, θ is the heading of the UAV, v is the constant velocity of the UAV, and $u \leq \overline{u}$ is the turn rate. The target is assumed to be a ground vehicle that can also turn and has constant velocity v_g. The state of the target is x_g, y_g, and θ_g with the same dynamics as the UAV with maximum turn rate $\overline{u}_g$.

The action space of the UAV is given by $\mathcal{U}_t = \{-\overline{u}, 0, \overline{u}\}$, and the action space of the ground vehicle is $\mathcal{W}_t = \{-\overline{u}_g, 0, \overline{u}_g\}$.

We define the joint state of the game $\eta_t = (\rho, \psi) \in \mathcal{X}_t$ to be the *relative* distance ρ and angle between the ground vehicle and UAV ψ. As shown in [22], the relative position is given by

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \cos\theta_g & \sin\theta_g \\ -\sin\theta_g & \cos\theta_g \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix}, \tag{32}$$

and the relative angle or *bearing* is given by $\psi = \tan^{-1}(\frac{x_g - x}{y_g - y})$. We then define the relative distance between

Algorithm 1 Forward-Backward Arimoto-Blahut Algorithm

$q_t^{(0)}(w_t \mid x_t)$ and $q_t^{(0)}(u_t \mid x_t, u_{t-1})$ for $t \in \mathcal{T}$, ▷ Initialize

$\phi_{T+1}^{(k)}(x_{T+1}, u_{T+1}) = \exp(-\gamma_1 c_{T+1}(x_{T+1}))$ for $k = 1, 2, \ldots, K$;

for $k = 1, 2, ..., K$ (until convergence) **do**

  for $t = 1, 2, ..., T$ **do** ▷ Forward Path

    $\mu_{t+1}^{(k)}(x_{t+1}, u^t, w^t) = \sum_{\mathcal{X}_t} p_{t+1}(x_{t+1} \mid x_t, u_t, w_t) q_t^{(k-1)}(u_t \mid x_t, u_{t-1}) q_t^{(k-1)}(w_t \mid x_t) \mu_t^{(k)}(x_t, u^{t-1}, w^{t-1})$;

    $\nu_t^{(k)}(u_t \mid u^{t-1}) = \sum_{\mathcal{X}_t} \mu_t^{(k)}(x^t | u^{t-1}) q_t^{(k-1)}(u_t | x^t, u^{t-1})$;

  for $t = T, T-1, ..., 1$ **do** ▷ Backward Path

    $$q_t^{(k)}(w_t|x^t) = \begin{cases} 1, & w_t = w_t^* = \arg\max_{w_t \in \mathcal{W}_t} \; c_t(x_t, \cdot, w_t), \\ 0, & w_t \in \mathcal{W}_t \setminus \{w_t^*\} \end{cases};$$

    $\phi_t^{(k)}(x_t, u^{t-1}) = \sum_{\mathcal{U}_t} \nu_t^{(k)}(u_t \mid u^{t-1}) \exp\left(-\gamma_1 \sum_{\mathcal{W}_t} q_t^{(k)}(w_t \mid x_t) \; \rho_t^{(k)}(x_t, u^t, w^t)\right)$;

    $\rho_t^{(k)}(x_t, u^t, w^t) = c_t(x_t, u_t, w_t) - \frac{1}{\gamma_1} \sum_{\mathcal{X}_{t+1}} p_{t+1}(x_{t+1} \mid x_t, u_t, w_t) \log\left(\phi_{t+1}^{(k)}(x_{t+1}, u^t)\right)$;

    $$q_t^{(k)}(u_t|x^t, u^{t-1}) = \frac{\nu_t^{(k)}(u_t|u^{t-1}) \exp\left(-\gamma_1 \sum_{\mathcal{W}_t} q_t^{(k)}(w_t|x^t) \; \rho_t^{(k)}(x_t, u^t, w^t)\right)}{\sum_{\mathcal{U}_t} \nu_t^{(k)}(u_t|u^{t-1}) \exp\left(-\gamma_1 \sum_{\mathcal{W}_t} q_t^{(k)}(w_t|x^t) \; \rho_t^{(k)}(x_t, u^t, w^t)\right)};$$

return $q_t^{(K)}(w_t \mid x_t)$ and $q_t^{(K)}(u_t \mid x_t, u_{t-1})$ for $t \in \mathcal{T}$.

the target and UAV as $\rho = \sqrt{\eta_1^2 + \eta_2^2}$. We note that the overall dynamics of the state (η_1, η_2, η_3) is captured in the dynamics of (x, y, θ), and (x_g, y_g, θ_g).

B. Cost objective

We model the UAV as having a camera on a gimbal mounted on the underside of its body. If the camera is pointed at the back of the UAV, it will be blocked by the landing gear and hence there exists a blind spot in that region. This is illustrated in Figure 1. The requirements for the vision-based target tracking are two-fold. We want the UAV to follow the ground vehicle at a prescribed distance ρ_c and we want to ensure the ground vehicle is not in the blind spot region of the UAV. We capture this using the cost function $c_t = c_1(\rho) + c_2(\psi)$ where

$$c_1(\rho) = \beta_1(\rho - \rho_c)^2 \tag{33}$$

$$c_2(\psi) = \begin{cases} 0 & \psi \notin \psi_b \\ \beta_2 & \psi \in \psi_b \end{cases} \tag{34}$$

where ψ_b is the set angles that constitute the blind spot of the UAV and $\beta_1, \beta_2 \in \mathbb{R}$ are constants. Hence, c_t incentivises the UAV to keep the target as close to distance ρ_c as possible and not allow the target to enter the blind spot.

C. Results

We use the UAV simulation environment *OpenAMASE* developed at Air Force Research Laboratory[1]. The problem formulation deals with a discrete and finite state space. We thus discretize the state space as follows. $\rho = \{0, \Delta_\rho + 2\Delta_\rho + \ldots, 10\rho_c\}$ and $\psi = \{0, \Delta_\psi, \Delta_\psi, \ldots, 2\pi\}$ where Δ_ρ, Δ_ψ are the discretization parameters. The parameters we use for the test are summarized in Table I. We note that it is typically assumed that $v > v_g$ as otherwise the tracking task is not possible as the target will always be able to escape. We also assume there is no wind or other sources of noise. The UAV is penalized in knowledge of the state of the game, i.e., strategies that rely more heavily on knowledge of the relative distance and heading are penalized. However, the target always has full knowledge of the system. We compare this case to the situation where this is no information cost and the UAV also has full state information.

[1]Available online at https://github.com/afrl-rq/OpenAMASE.

$\mathbf{v}$	$\mathbf{v_g}$	$\overline{\mathbf{u}}$	$\overline{\mathbf{u}}_\mathbf{g}$	$\rho_\mathbf{c}$	$\mathbf{\Delta}_\rho$	$\mathbf{\Delta}_\psi$
20	10	$0.34\ rad/s$	$0.5\ rad/s$	$300\ m$	$10\ m$	$\frac{\pi}{8}\ rads$

TABLE I: Experiment parameter values.

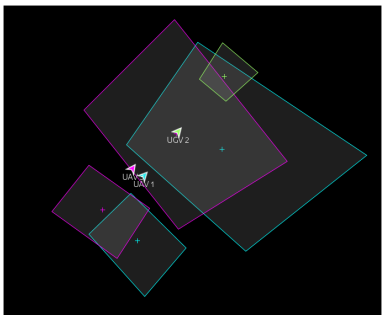
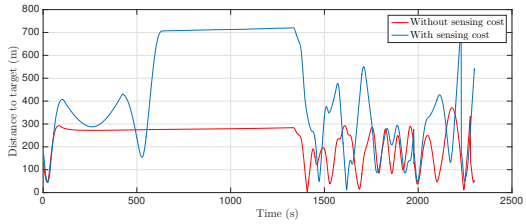
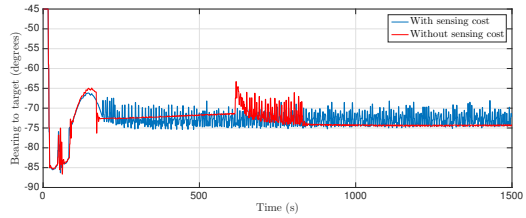


Fig. 2: Snapshot of simulation on AMASE. The purple and blue UAVs are tracking the green target ground vehicle. The purple UAV has no sensing cost. The corresponding colored polygons in front of the respective UAVs are the sensor footprint of the camera and the polygons behind the UAV represent the camera blindspots.

In the example, we set $\beta_1 << \beta_2$. As can be seen in Figure 3a, in the presence of sensing cost, the UAV typically maintains a larger distance from the target. This is because this the target entering the blind spot of the gimbal mounted camera is penalized more heavily than maintaining the ρ_c distance. As the target has a faster turn rate and the UAV



(a) Distance from the UAVs to the target.



(b) Bearing from the UAVs to the target over time.

Fig. 3: Graphs of distance and bearing from the UAVs to the target in both the without-sensing-cost (in red) and with-sensing-cost (in blue) cases.

has a minimum turn radius of $\frac{v}{\bar{u}}$ m, not getting too close gives the UAV more time and space to react to the target potentially manouevering towards the blind spot. Note that the results without any sensing cost are similar to those seen in [21] and [22].

V. CONCLUSIONS AND FUTURE WORK

We considered two-player stochastic games with additional sensing costs in terms of transfer entropy. We derived a set of nonlinear equations that the optimal strategies satisfy and presented a method for computing them using the modified Arimoto-Blahut algorithm. We applied the proposed methodology to a UAV pursuit-evasion stochastic game.

Prospective research will focus on following the footsteps of our previous work [11] to additionally allow for high-level mission specifications in terms of co-safe linear temporal logic formulae. Moreover, extending the current work to partially observable stochastic games and stochastic games with bounded rationality are interesting open problems.

REFERENCES

[1] A. J. Newman, C. L. Richardson, S. M. Kain, P. G. Stankiewicz, P. R. Guseman, B. A. Schreurs, and J. A. Dunne, "Reconnaissance blind multi-chess: an experimentation platform for ISR sensor fusion and resource management," pp. 9842 – 9842 – 20, 2016.

[2] L. S. Shapley, "Stochastic games," *Proceedings of the National Academy of Sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[3] E. Altman and A. Hordijk, "Zero-sum Markov games and worst-case optimal control of queueing systems," *Queueing Systems*, vol. 21, no. 3, pp. 415–447, Sep 1995.

[4] I. Karatzas, M. Shubik, and W. D. Sudderth, "A strategic market game with secured lending," *Journal of Mathematical Economics*, vol. 28, no. 2, pp. 207 – 247, 1997.

[5] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1994.

[6] E. Todorov, "Linearly-solvable Markov decision problems," in *Advances in neural information processing systems*, 2007, pp. 1369–1376.

[7] K. Rawlik, M. Toussaint, and S. Vijayakumar, "On stochastic optimal control and reinforcement learning by approximate inference," in *Robotics: science and systems*, vol. 13, no. 2, 2012, pp. 3052–3056.

[8] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," *arXiv preprint arXiv:1702.08165*, 2017.

[9] C. Daniel, G. Neumann, and J. Peters, "Hierarchical relative entropy policy search," in *Artificial Intelligence and Statistics*, 2012, pp. 273–281.

[10] T. Tanaka, H. Sandberg, and M. Skoglund, "Transfer-Entropy-Regularized Markov Decision Processes," *ArXiv e-prints*, Aug. 2017.

[11] S. Bharadwaj, M. Ahmadi, T. Tanaka, and U. Topcu, "Transfer entropy MDPs with temporal logic constraints," in *The 57th IEEE Conference on Decision and Control*, 2018, submitted.

[12] J. L. Massey, "Causality, feedback and directed information," in *In Proc. 1990 Int. Symp. on Info. Th. & its Appls.*, Hawaii, USA, 1990.

[13] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 644–662, Feb 2009.

[14] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theor.*, vol. 49, no. 1, pp. 4–21, Jan. 2003.

[15] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, Jul 1972.

[16] T. Bewley and E. Kohlberg, "The asymptotic theory of stochastic games," *Mathematics of Operations Research*, vol. 1, no. 3, pp. 197–208, 1976. [Online]. Available: http://www.jstor.org/stable/3689563

[17] J. F. Mertens and A. Neyman, "Stochastic games," *International Journal of Game Theory*, vol. 10, no. 2, pp. 53–66, Jun 1981.

[18] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, pp. 461–464, Jul 2000.

[19] J. Grau-Moya, F. Leibfried, and H. Bou-Ammar, "Balancing two-player stochastic games with soft Q-learning," *arXiv preprint arXiv:1802.03216*, 2018.

[20] T. Tanaka, H. Sandberg, and M. Skoglund, "Finite state Markov decision processes with transfer entropy costs," *ArXiv*, vol. abs/1708.09096, 2017.

[21] S. A. Quintero and J. P. Hespanha, "Vision-based target tracking with a small UAV: Optimization-based control strategies," *Control Engineering Practice*, vol. 32, pp. 28 – 42, 2014.

[22] D. Milutinović, D. W. Casbeer, D. Kingston, and S. Rasmussen, "A stochastic approach to small UAV feedback control for target tracking and blind spot avoidance," in *2017 IEEE Conference on Control Technology and Applications (CCTA)*, Aug 2017, pp. 1031–1037.

[23] D. Du and P. Pardalos, *Minimax and Applications*, ser. Nonconvex Optimization and Its Applications. Springer US, 1995.

APPENDIX

PROOF OF PROPOSITION 1

Since we consider a finite number of states and decisions, we rewrite (6) as

$$L = \sum_{ijk} p^i Q^{ij} Q^{ik} c_{ijk} + \frac{1}{\gamma_1} \sum_{ij} p^i Q^{ij} \log \frac{Q^{ij}}{\sum_{i'} p^{i'} Q^{i'j}} - \frac{1}{\gamma_2} \sum_{ik} p^i Q^{ik} \log \frac{Q^{ik}}{\sum_{i'} p^{i'} Q^{i'k}}, \quad (35)$$

where I, J, and K are the index sets of states, control actions, and adversary actions, respectively, and $Q^{ij} = q(u \mid x)$, $Q^{ik} = q(w \mid x)$, $c_{ijk} = c(x, u, w)$, and $p^i = p(x)$. Then, the optimization problem we ought to solve can be described as

$$\min_{\{Q^{ij}\}_{i\in I, j\in J}} \max_{\{Q^{ik}\}_{i\in I, k\in K}} L. \quad (36)$$

The above optimization problem is also subject to the following constraints on the controller and the adversary probability

distributions

$$\sum_{j\in J} Q^{ij} = 1, \quad (37)$$

$$\sum_{k\in K} Q^{ik} = 1. \quad (38)$$

We split the proof into two parts. In the first part, we study the problem when $\gamma_1, \gamma_2 \neq \infty$, and then, in the second part, we show how the results change as $\gamma_2 \to \infty$.

A. Case 1: $\gamma_1, \gamma_2 \neq \infty$.

Objective function L is a continuous function, convex in Q^{ij} and concave in Q^{ik}, and Q^{ij} and Q^{ik} take values in a compact set (any finite set is a compact). By von Neumann's minimax theorem [23], the following strong minimax property holds

$$\min_{\{Q^{ij}\}_{i\in I, j\in J}} \max_{\{Q^{ik}\}_{i\in I, k\in K}} L = \max_{\{Q^{ik}\}_{i\in I, k\in K}} \min_{\{Q^{ij}\}_{i\in I, j\in J}} L.$$

In order to solve optimization problem (36) subject to equality constraints (37) and (38), we write down the Lagrangian as $F = L + \lambda^i p^i \left(\sum_j Q^{ij} - 1\right) + \alpha^i p^i \left(\sum_k Q^{ik} - 1\right)$, where λ_i and α_i, $i \in I$ are the Lagrange multipliers. We begin by maximizing F with respect to Q^{ik}. Taking the derivative of F with respect to Q^{ik} yields

$$\frac{\partial F}{\partial Q^{ik}} = p^i \sum_j Q^{ij} c^{ijk} - \frac{1}{\gamma_2}\Bigg(p^i \log \frac{Q^{ik}}{\sum_{i'} p^{i'} Q^{i'k}} - p^i Q^{ik} \frac{1}{Q^{ik}} + \underbrace{p^i Q^{ik} \frac{p^i}{\sum_{i'} p^{i'} Q^{i'k}} + \sum_{i\neq i'} p^{i'} Q^{i'k} \frac{p^i}{\sum_{i'} p^{i'} Q^{i'k}}}_{p^i} \Bigg) + \alpha^i p^i.$$

Simplifying the above terms gives

$$\frac{\partial F}{\partial Q^{ik}} = p^i \left(\sum_j Q^{ij} c^{ijk} - \frac{1}{\gamma_2} \log \frac{Q^{ik}}{\mu^k} + \alpha^i \right)$$

where $\mu^k = \sum_{i'} p^{i'} Q^{i'k}$. If we equate $\frac{\partial F}{\partial Q^{ik}} = 0$, we obtain

$$p^i \left(\sum_j Q^{ij} c^{ijk} - \frac{1}{\gamma_2} \log \frac{Q^{ik}}{\mu^k} + \alpha^i \right) = 0.$$

If $p^i = 0$, that state is of no concern in maximizing or minimizing the objective function. Therefore, we are concerned with the cases when $p^i > 0$. We have then

$$\sum_{j'} Q^{ij'} c^{ij'k} - \frac{1}{\gamma_2} \log \frac{Q^{ik}}{\mu^k} + \alpha^i = 0,$$

and, by re-arranging the terms, we obtain an expression for the optimal adversary distribution

$$Q^{ik} = \mu^k \exp\left(\gamma_2 \alpha^i + \gamma_2 \sum_{j'} Q^{ij'} c^{ij'k} \right).$$

Summing both sides of the above expression over $k \in K$ and applying constraint (38), we get

$$Q^{ik} = \frac{\mu^k \exp\left(\gamma_2 \sum_{j'} Q^{ij'} c^{ij'k}\right)}{\sum_{k'} \mu^{k'} \exp\left(\gamma_2 \sum_{j'} Q^{ij'} c^{ij'k'}\right)}, \quad (39)$$

which is the same as equation (7). Plugging this solutions back into the term $\sum_{ik} p^i Q^{ik} \log \frac{Q^{ik}}{\sum_{i'} p^{i'} Q^{i'k}}$ in L as in (35) gives

$$\begin{aligned}
&\sum_{ik} p^i Q^{ik} \log \frac{Q^{ik}}{\sum_{i'} p^{i'} Q^{i'k}} \\
&= \sum_{ik} p^i Q^{ik} \log \frac{\exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)}{\sum_k \mu^k \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)} \\
&= \sum_{ik} \Bigg(\gamma_2 p^i Q^{ik} \sum_j Q^{ij} c^{ijk} \\
&\quad - p^i Q^{ik} \log\Big(\sum_{k'} \mu^{k'} \exp\big(\gamma_2 \sum_j Q^{ij} c^{ijk'}\big)\Big)\Bigg) \\
&= \gamma_2 \sum_{ijk} p^i Q^{ik} Q^{ij} c^{ijk} \\
&\quad - \sum_i p^i \log\left(\sum_{k'} \mu^{k'} \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk'} \right)\right).
\end{aligned}$$

Substituting the above term back in L leads to

$$\begin{aligned}
L = \min_{Q^{ij}} \quad & \frac{1}{\gamma_1} \sum_{ij} p^i Q^{ij} \log\left(\frac{Q^{ij}}{\sum_{i'} p^{i'} Q^{i'j}} \right) \\
& + \frac{1}{\gamma_2} \sum_i p^i \log\left(\sum_k \mu^k \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk} \right)\right) \quad (40)
\end{aligned}$$

In order to find Q^{ij}_*, we compute the derivative of F with respect to Q^{ij} with L as in (40) as follows

$$\begin{aligned}
\frac{\partial F}{\partial Q^{ij}} = \frac{1}{\gamma_1}\Bigg(& p^i \log(Q^{ij}) + p^i Q^{ij} \frac{1}{Q^{ij}} - p^i \log(\nu^j) \\
& \underbrace{p^i Q^{ij} \frac{p^i}{\sum_{i'} p^{i'} Q^{i'j}} + \sum_{i\neq i'} p^{i'} Q^{i'j} \frac{p^i}{\sum_{i'} p^{i'} Q^{i'j}}}_{p^i} \Bigg) \\
& + p^i \frac{\sum_k \mu^k c^{ijk} \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)}{\sum_k \mu^k \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)} + \lambda^i p^i \\
& \underbrace{+ p^i \frac{\partial Q^{ik}_*}{Q^{ij}} \log\left(\sum_k \mu^k \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk} \right)\right)}_{=0}
\end{aligned}$$

where, $\nu^j = \sum_{i'} p^{i'} Q^{i'j}$. Then, we have

$$p^i \left(\frac{1}{\gamma_1} \log \left(\frac{Q^{ij}}{\nu^j} \right) + \frac{\sum_k \mu^k c^{ijk} \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)}{\sum_k \mu^k \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)} + \lambda^i \right) = 0.$$

Since we are interested in the cases when $p^i > 0$, we obtain

$$\frac{1}{\gamma_1} \log \left(\frac{Q^{ij}}{\nu^j} \right) + \frac{\sum_k \mu^k c^{ijk} \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)}{\sum_k \mu^k \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)} + \lambda^i = 0.$$

Solving for Q^{ij} gives

$$Q^{ij} = \nu^j \exp\left(-\gamma_1 \lambda^i - \gamma_1 \frac{\sum_k \mu^k c^{ijk} \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)}{\sum_k \mu^k \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)} \right).$$

Summing both sides of the above expression over j and applying the constraint (37), we find the optimal strategy distribution

$$Q_*^{ij} = \frac{\nu^j \exp\left(-\gamma_1 \frac{\sum_k \mu^k c^{ijk} \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)}{\sum_k \mu^k \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)} \right)}{\sum_j \nu^j \exp\left(-\gamma_1 \frac{\sum_k \mu^k c^{ijk} \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)}{\sum_k \mu^k \exp\left(\gamma_2 \sum_j Q^{ij} c^{ijk}\right)} \right)}.$$

From (39), we have

$$Q_*^{ij} = \frac{\nu^j \exp\left(-\gamma_1 \sum_{k'} Q_*^{ik'} c^{ijk'}\right)}{\sum_{j'} \nu^{j'} \exp\left(-\gamma_1 \sum_{k'} Q_*^{ik'} c^{ij'k'}\right)}, \tag{41}$$

which is the same as equation (8). Substituting the Q_*^{ij} into (40) yields

$$\begin{aligned} L^* = & \sum_{ijk} p^i Q_*^{ij} Q_*^{ik} c^{ijk} \\ & - \frac{1}{\gamma_1} \sum_i p^i \log \left(\sum_j \nu_*^j \exp\left(-\gamma_1 \sum_k Q_*^{ik} c^{ijk} \right) \right) \\ & + \frac{1}{\gamma_2} \sum_i p^i \log \left(\sum_k \mu_*^k \exp\left(\gamma_2 \sum_j Q_*^{ij} c^{ijk} \right) \right). \end{aligned} \tag{42}$$

B. Case 2: $\gamma_2 \to \infty$.

We compute the limit of the calculated quantities in the previous section as $\gamma_2 \to \infty$.

- Calculating Q_*^{ik}:

$$\begin{aligned} \lim_{\gamma_2 \to \infty} Q_*^{ik} &= \lim_{\gamma_2 \to \infty} \frac{\mu^k \exp\left(\gamma_2 \sum_{j'} Q^{ij'} c^{ij'k}\right)}{\sum_{k'} \mu^{k'} \exp\left(\gamma_2 \sum_{j'} Q^{ij'} c^{ij'k'}\right)} \\ &= \frac{\mu^k}{\sum_{k'} \mu^{k'}} \lim_{\gamma_2 \to \infty} \exp\left(\gamma_2 \sum_{j'} Q_*^{ij'} (c^{ijk} - c^{ij'k^*}) \right), \end{aligned}$$

where $k^* = \arg\max_k c^{i,\cdot,k}$ or equivalently $w^* = \arg\max_{w \in \mathcal{W}} c(x, \cdot, w)$. Since c is concave in w such k always exists. We can find the optimal adversary strategy by solving the following equation

$$\begin{aligned} Q_*^{ik} &= \underbrace{\frac{\sum_{i'} p^{i'} Q_*^{i'k}}{\sum_k \sum_{i'} p^{i'} Q_*^{i'k}}}_{=1} \\ &\quad \times \lim_{\gamma_2 \to \infty} \exp\left(\gamma_2 \sum_{j'} Q_*^{ij'} (c^{ijk} - c^{ij'k^*}) \right) \\ &= \begin{cases} \mu^{k^*} = 1, & k = k^* = \arg\max_{k \in K} c^{i,\cdot,k}, \\ 0, & k \in K \setminus \{k^*\}, \end{cases} \end{aligned} \tag{43}$$

which is indeed a pure strategy as expected.

- Computing the optimal value function L^*: To this end, we compute the limit for the last term on the right hand side of (42) as $\gamma_2 \to \infty$. That is,

$$\lim_{\gamma_2 \to \infty} \frac{1}{\gamma_2} \sum_i p^i \log \left(\sum_k \mu_*^k \exp\left(\gamma_2 \sum_j Q_*^{ij} c^{ijk} \right) \right) = \frac{\infty}{\infty}. \tag{44}$$

Thus, we apply the L'Hospital's rule

$$\begin{aligned} \lim_{\gamma_2 \to \infty} \sum_i p^i & \frac{\sum_k \mu^k \sum_j Q^{ij} c^{ijk} \exp\left(\gamma_2 \sum_j Q_*^{ij} c^{ijk}\right)}{\sum_k \mu_*^k \exp\left(\gamma_2 \sum_j Q_*^{ij} c^{ijk}\right)} \\ &= \sum_i \sum_j p^i Q^{ij} \sum_k \frac{\mu^k}{\sum_{k'} \mu^{k'}} \\ &\quad \times \lim_{\gamma_2 \to \infty} \exp\left(\gamma_2 \sum_{j'} Q_*^{ij'} (c^{ijk} - c_{ij'k^*}) \right) c^{ijk}. \end{aligned}$$

From (33), we infer that the right-hand side of the above expression equals

$$\sum_{ij} p^i Q^{ij} Q^{ik^*} c^{ijk^*}$$

If we plug in the above expression into (42), we find the optimal value function

$$L^* = -\frac{1}{\gamma_1} \sum_i p^i \log \left(\sum_j \nu_*^j \exp\left(-\gamma_1 Q_*^{ik^*} c^{ijk^*} \right) \right),$$

which is the same as equation (13).

- Computing the optimal protagonist strategy Q_*^{ij}: We substitute Q_*^{ik} as computed above in (33) obtaining

$$Q_*^{ij} = \frac{\nu^j \exp\left(-\gamma_1 Q_*^{ik^*} c^{ijk^*}\right)}{\sum_{j'} \nu^{j'} \exp\left(-\gamma_1 Q_*^{ik^*} c^{ij'k^*}\right)},$$

which is identical to equation (12).