Structure in the value function of zero-sum games of incomplete information

Auke J. Wiggers
Universiteit van Amsterdam
Amsterdam, The Netherlands
wiggers.auke@gmail.com

Frans A. Oliehoek
Universiteit van Amsterdam
University of Liverpool
fao@liverpool.ac.uk

Diederik M. Roijers
Universiteit van Amsterdam
Amsterdam, The Netherlands
d.m.roijers@uva.nl

ABSTRACT

In this paper, we introduce plan-time sufficient statistics, representing probability distributions over joint sets of private information, for zero-sum games of incomplete information. We define a *family of zero-sum Bayesian Games (zs-BGs)*, of which the members share all elements but the plan-time statistic. Using the fact that the statistic can be decomposed into a marginal and a conditional term, we prove that the value function of the family of zs-BGs exhibits concavity in marginal-space of the maximizing agent and convexity in marginal-space of the minimizing agent. We extend this result to sequential settings with a dynamic state, i.e., *zero-sum Partially Observable Stochastic Games (zs-POSGs)*, in which the statistic is a probability distribution over joint action-observation histories. First, we show that the final stage of a zs-POSG corresponds to a family of zs-BGs. Then, we show by induction that the convexity and concavity properties can be extended to every time-step of the zs-POSG.

1. INTRODUCTION

Modeling decision making for strictly competitive settings with incomplete information is a field with many promising applications for AI. Examples include games such as poker [26], and more importantly, security settings [7]. When the environment can be influenced via actions and the decision making is sequential, the problem can be modeled as a *zero-sum Partially Observable Stochastic Game (zs-POSG)*.

Reasoning about zs-POSGs poses a challenge for strategic agents. Not only do they need to take their own uncertainty about the state of the environment into account, but also uncertainty regarding the opposing agent. Furthermore, because one agent is trying to minimize the reward that the other is maximizing, behaving strategically typically requires stochastic strategies. A factor that further complicates the reasoning is that agents can not only influence the *future state of the environment* through their own actions, but also what they will observe, as well as what the other agent will observe.

In this paper, we prove the existence of useful properties of zs-POSGs, that may be exploited to make reasoning about these models more tractable. We take inspiration from recent work for collaborative settings which has shown that it is possible to summarize the past joint policy using so called plan-time sufficient statistics [16], which can be interpreted as the belief of a special type of Partially Observable Markov Decision Process (POMDP) to which the collaborative Decentralized POMDP can be reduced [2, 9, 14]. This is helpful, since it allows the problem to be solved using solution methods for POMDPs, leading to increases in scalability [2]. We extend these results to the zs-POSG setting by making the following four contributions:

1. The introduction of the concept of a *family of zero-sum Bayesian Games* and proof that its value function exhibits certain concave/convex properties.
2. A zs-POSG value function formulation based on the past joint policy.
3. A zs-POSG value function formulation based on the new plan-time sufficient statistic, and subsequent reduction of the POSG model to a centralized model.
4. A proof that the result for the family of zs-BGs extends to the finite horizon zs-POSG case. More specifically, we show that final stage of the zs-POSG is equivalent to a family of zs-BGs and then prove that the value function exhibits similar concave/convex properties on every stage.

As far as the authors are aware, this is the first work that gives insight in the shape of the value function of the zs-POSG specified over the space of plan-time sufficient statistics. We believe that this result about the shape of the value function will open up the route for effective solution methods like *dynamic programming*, and will explain this in more detail.

2. BACKGROUND

In this work, we focus on zero-sum games of incomplete information where the number of states, actions, observations and the horizon are finite. We examine games with a static hidden state, and games with a dynamic hidden state that may evolve over time. We first provide background on zero-sum games and formally define the zs-BG and zs-POSG frameworks. We assume *perfect recall*, i.e., agents recall past actions and observations, and assume that all elements of the game are *common knowledge* among the agents.

2.1 Zero-sum Normal Form Games

We first shortly discuss zero-sum games before defining the Normal Form Game framework. In the two-player, zero-sum setting, rewards for both agents sum to zero. The *value* of the game is the value attained when both agents play rationally, i.e. they follow their respective *maxmin-strategies*.

Definition 1. A **maxmin-strategy** for agent i is the strategy that gives the best payoff for agent i, given that the opposing agent aims to minimize it.

Appears in: *The 10th Annual Workshop on Multiagent Sequential Decision-Making Under Uncertainty (MSDM-2015)*, held in conjunction with *AAMAS*, May 2015, Istanbul, Turkey.

By convention, let agent 1 be the maximizing agent and agent 2 the minimizing agent. Throughout this work the value refers to the value for agent 1, as the value for agent 2 is its additive inverse.

To find the maxmin-strategies for a zero-sum game of complete information, we model it as a Strategic Game, also known as Normal Form Game (NFG), which is a framework for multi-agent decision making in one-shot games. The zero-sum NFG framework is defined as a tuple $\langle I, \Xi, R \rangle$:

- $I = \{1, 2\}$ is the set of 2 agents.
- $\Xi_1 \times \Xi_2$ is the set of pure strategies $\xi = \{\xi_1, \xi_2\}$.
- $R(\xi)$ is the reward function for agent 1.

We distinguish between deterministic or pure strategies ξ_i, and mixed strategies μ_i that specify a probability distribution over pure strategies. Given a mixed joint strategy $\mu = \langle \mu_1, \mu_2 \rangle$, the value for agent 1 is:

$$V_{\text{NFG}}(\mu) = \sum_{\xi \in \Xi} \mu(\xi) R(\xi).$$

Using this function, maxmin-strategies and the corresponding value can be found. A commonly used solution concept is the *Nash Equilibrium* (NE), which is a joint strategy from which no agent has an incentive to unilaterally deviate. In the zs-NFG, a mixed NE is guaranteed to exist [12], and happens to coincide with the mixed maxmin-strategies μ_1^* and μ_2^*. That is, for every NE $\langle \mu_1^*, \mu_2^* \rangle$, the following holds:

$$V_{\text{NFG}}(\mu_1^* \mu_2^*) = \max_{\mu_1} \min_{\mu_2} V_{\text{NFG}}(\mu_1 \mu_2) = \min_{\mu_2} \max_{\mu_1} V_{\text{NFG}}(\mu_1 \mu_2). \tag{2.1}$$

For proof of the validity of (2.1) the reader is referred to [20]. A mixed NE can be found using Linear Programming [1].

2.2 Zero-sum Bayesian Games

The zero-sum Bayesian Game (zs-BG) is a model for multi-agent decision making under uncertainty in a one-shot zero-sum game. Two agents simultaneously select an action based on an individual observation (their *type*), and aim to maximize expected individual reward, knowing that the opposing agent aims to minimize it. It is a tuple $\langle I, \Theta, \mathcal{A}, R, \sigma \rangle$:

- $I = \{1, 2\}$ is the set of 2 agents.
- $\Theta_1 \times \Theta_2$ is the set of joint types $\theta = \{\theta_1, \theta_2\}$.
- $\mathcal{A}_1 \times \mathcal{A}_2$ is the set of joint actions $a = \{a_1, a_2\}$.
- $R(\theta, a)$ is the reward function for agent 1.
- $\sigma \in \Delta(\Theta)$ is the probability distribution over types.

A pure strategy in a Bayesian Game is a mapping from types to actions, to which we refer as a *pure decision rule* δ_i. A *stochastic decision rule* maps from types to probability distributions over the set of actions, denoted as $\delta_i(a_i|\theta_i)$. Given a joint decision rule, which is a tuple containing decision rules for both agents $\delta = \langle \delta_1, \delta_2 \rangle$, the value for agent 1 is:

$$Q_{\text{BG}}(\delta) \triangleq \sum_{\theta} \sigma(\theta) \sum_{a} \delta(a|\theta) R(\theta, a). \tag{2.2}$$

The goal in zs-BGs is to find a rational joint decision rule and the corresponding value. This value is defined as follows:

$$V_{\text{BG}} \triangleq \max_{\delta_1} \min_{\delta_2} Q_{\text{BG}}(\delta_1 \delta_2). \tag{2.3}$$

To solve a zs-BG, it can be converted to an NFG by treating every pure decision rule as a pure strategy. The mixed NE in this NFG corresponds to a stochastic joint decision rule $\delta^* = \langle \delta_1^*, \delta_2^* \rangle$. The value function is then given by (2.1). The NE in the BG is referred to as the *Bayesian Nash Equilibrium* (BNE), i.e. a joint decision rule from which no agent has an incentive to unilaterally deviate [15]. A more efficient solution method for zs-BGs involves converting the game to sequence form and solving it accordingly [8].

2.3 Zero-sum POSGs

A zero-sum Partially Observable Stochastic Game (zs-POSG) is a model for multi-agent decision making under uncertainty in zero-sum sequential games where the state evolves over time. It is a tuple $\langle h, I, \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, b^0 \rangle$:

- h is the horizon.
- $I = \{1, 2\}$ is the set of 2 agents.
- $\mathcal{S}$ is the set of states s.
- $\mathcal{A}$ is the set of joint actions $a = \{a_1, a_2\}$.
- $\mathcal{O}$ is the set of joint observations $o = \{o_1, o_2\}$.
- T is the transition function that specifies $\Pr(s^{t+1}|s^t, a^t)$.
- O is the observation function: $\Pr(o^{t+1}|s^{t+1}, a^t)$.
- $R(s^t, a^t)$ is the reward function for agent 1.
- $b^0 \in \Delta(\mathcal{S})$ is the initial probability distribution over joint states.

In the zs-POSG, we aim to find a rational joint policy[1] and corresponding value. Let a *pure policy* for agent i be a mapping from individual action-observation histories (AOHs) $\vec{\theta}_i^t = \langle \theta_i^0 \dots \theta_i^t \rangle$ to actions. Let a *stochastic policy* for agent i be a mapping from individual AOHs to a probability distribution over actions, denoted as $\pi_i(a_i^t|\vec{\theta}_i^t)$. An individual policy defines action selection of one agent on every stage of the game, and is essentially a sequence of individual decision rules $\pi_i = \langle \delta_i^0 \dots \delta_i^{h-1} \rangle$. We define the *past joint policy* as a tuple of decision rules $\varphi^t = \langle \delta^0, \dots, \delta^{t-1} \rangle$, and define the tuple containing decision rules from stage t to h as the *partial policy* $\pi^t = \langle \delta^t, \dots, \delta^{h-1} \rangle$.

Similar to the BG, a zs-POSG can be converted to normal form by treating every pure joint policy as a pure strategy. As such, a mixed NE exists, which corresponds to a stochastic joint policy. However, solving this NFG quickly becomes intractable, as the number of pure joint policies is exponential in the number of agents, actions *and* observations. In the zero-sum case, it is more efficient to convert the zs-POSG to sequence form and solve it accordingly.

3. FAMILIES OF BAYESIAN GAMES

We introduce the concept of a family of Bayesian Games. In Section 3.1, we prove that in the zero-sum case its value function exhibits certain concave/convex properties. In Section 5, we will use this result to give a similar proof for the zs-POSG.

Definition 2. A **family of Bayesian Games**, defined as a tuple $\mathcal{F} = \langle I, \Theta, \mathcal{A}, R \rangle$, is the set of Bayesian Games for which all elements but the probability distribution over joint types σ are identical.

Let $\mathcal{F}$ be a family of *zero-sum* Bayesian Games. We define its value function in terms of the type distribution:

$$V_{\mathcal{F}}^*(\sigma) \triangleq \sum_{\theta} \sigma(\theta) \sum_{a} \delta^*(a|\theta) R(\theta, a). \tag{3.1}$$

were δ^* is the rational joint decision rule. Note that this gives the value for the BG in $\mathcal{F}$ that has type distribution σ.

[1]While the terms 'strategy' and 'policy' are interchangeable, the first is often used in the field of game theory whereas the latter is the standard in AI. We will use 'strategy' for NFGs, and 'policy' for POSGs.

We generalize (2.2) and (2.3) as follows:

$$Q_{\mathcal{F}}(\sigma,\delta) \triangleq \sum_{\theta}\sigma(\theta)\sum_{a}\delta(a|\theta)R(\theta,a), \tag{3.2}$$

$$V^*_{\mathcal{F}}(\sigma) \triangleq \max_{\delta_1}\min_{\delta_2} Q_{\mathcal{F}}(\sigma,\delta_1\delta_2). \tag{3.3}$$

We define *best-response value functions* that give the best-response value to a decision rule of the opposing agent:

$$V^{\mathrm{BR1}}_{\mathcal{F}}(\sigma,\delta_2) \triangleq \max_{\delta_1} Q_{\mathcal{F}}(\sigma,\delta_1\delta_2), \tag{3.4}$$

$$V^{\mathrm{BR2}}_{\mathcal{F}}(\sigma,\delta_1) \triangleq \min_{\delta_2} Q_{\mathcal{F}}(\sigma,\delta_1\delta_2). \tag{3.5}$$

It follows from equations 3.3, 3.4 and 3.5 that

$$V^*_{\mathcal{F}}(\sigma) = \min_{\delta_2} V^{\mathrm{BR1}}_{\mathcal{F}}(\sigma,\delta_2) = \max_{\delta_1} V^{\mathrm{BR2}}_{\mathcal{F}}(\sigma,\delta_1). \tag{3.6}$$

3.1 Concavity/convexity of the value function

We will prove that $V^*_{\mathcal{F}}$ exhibits a certain concave/convex shape. We decompose σ into a *marginal* term,

$$\sigma_{m,1}(\theta_1) \triangleq \sum_{\theta_2}\sigma(\theta_1\theta_2), \tag{3.7}$$

and a *conditional* term,

$$\sigma_{c,1}(\theta_2|\theta_1) \triangleq \frac{\sigma(\theta_1\theta_2)}{\sum_{\theta_2}\sigma(\theta_1\theta_2)} = \frac{\sigma(\theta_1\theta_2)}{\sigma_{m,1}(\theta_1)}. \tag{3.8}$$

The terms $\sigma_{m,2}$ and $\sigma_{c,2}$ are defined similarly. We refer to the simplex $\Delta(\Theta_i)$ containing marginals $\sigma_{m,i}$ as the *marginal-space* of agent i. We define $V^*_{\mathcal{F}}(\sigma_{m,1}|\sigma_{c,1}) = V^*_{\mathcal{F}}(\sigma)$ and use this notation to indicate that we consider the value function in $\Delta(\Theta_1)$ for a single conditional $\sigma_{c,1}$.

We will show that the best-response value functions defined in (3.4) and (3.5) are linear in their respective marginal-spaces. Using this result, we prove that $V^*_{\mathcal{F}}$ exhibits concavity in $\Delta(\Theta_1)$ for every $\sigma_{c,1}$, and convexity in $\Delta(\Theta_2)$ for every $\sigma_{c,2}$. For this purpose, let us define a vector that contains the reward for agent 1 for each individual type θ_1, given $\sigma_{c,1}$ and given that agent 2 follows decision rule δ_2:

$$r_{[\sigma_{c,1},\delta_2]}(\theta_1) \triangleq \max_{a_1}\sum_{\theta_2}\sigma_{c,1}(\theta_2|\theta_1)\sum_{a_2}\delta_2(a_2|\theta_2)R(\theta,a). \tag{3.9}$$

The vector $r_{[\sigma_{c,2},\delta_1]}$ is defined similarly.

LEMMA 1. *(1) $V^{BR1}_{\mathcal{F}}$ is linear in $\Delta(\Theta_1)$ for all $\sigma_{c,1}$ and δ_2, and (2) $V^{BR2}_{\mathcal{F}}$ is linear in $\Delta(\Theta_2)$ for all $\sigma_{c,2}$ and δ_1. More specifically, we can write the best-response value functions as the inner product of a marginal $\sigma_{m,i}$ and a vector:*

1. $V^{BR1}_{\mathcal{F}}(\sigma_{m,1},\delta_2|\sigma_{c,1}) = \sigma_{m,1}\cdot r_{[\sigma_{c,1},\delta_2]}$, (3.10)
2. $V^{BR2}_{\mathcal{F}}(\sigma_{m,2},\delta_1|\sigma_{c,2}) = \sigma_{m,2}\cdot r_{[\sigma_{c,2},\delta_1]}$. (3.11)

PROOF. The full proof is listed in appendix A. □

THEOREM 1. *$V^*_{\mathcal{F}}$ is (1) concave in $\Delta(\Theta_1)$ for a given conditional distribution $\sigma_{c,1}$, and (2) convex in $\Delta(\Theta_2)$ for a given conditional distribution $\sigma_{c,2}$. More specifically, $V^*_{\mathcal{F}}$ is respectively a minimization over linear functions in $\Delta(\Theta_1)$ and a maximization over linear functions in $\Delta(\Theta_2)$:*

1. $V^*_{\mathcal{F}}(\sigma_{m,1},|\sigma_{c,1}) = \min_{\delta_2}\left[\sigma_{m,1}\cdot r_{[\sigma_{c,1},\delta_2]}\right]$,
2. $V^*_{\mathcal{F}}(\sigma_{m,2}|\sigma_{c,2}) = \max_{\delta_1}\left[\sigma_{m,2}\cdot r_{[\sigma_{c,2},\delta_1]}\right]$.

PROOF. Filling in the result of Lemma 1 gives:

$$V^*_{\mathcal{F}}(\sigma_{m,1}|\sigma_{c,1}) \overset{\{3.5\}}{=} \min_{\delta_2} V^{\mathrm{BR1}}_{\mathcal{F}}(\sigma_{m,1},\delta_2|\sigma_{c,1}) \overset{\{3.10\}}{=} \min_{\delta_2}\left[\sigma_{m,1}\cdot r_{[\sigma_{c,1},\delta_2]}\right].$$

The proof for item 2 is analogous to that of item 1. □

4. VALUE FUNCTION OF THE ZS-POSG

In this section, we give two formulations for the value function of the zs-POSG, one in terms of the past joint policy (in Section 4.1) and one in terms of a *plan-time sufficient statistic* (in Section 4.2). These formulations are relatively straightforward extensions from previous work on POSGs with identical payoffs, called Dec-POMDPs [9, 16]. They allow us, in Section 5, to prove certain properties of the structure of the value function. In Section 4.3 we show that, using plan-time sufficient statistics, the POSG model can be reduced to a centralized stochastic game with hidden state and without observations, which we refer to as the Non-Observable Stochastic Game (NOSG).

4.1 Value in Terms of Past Joint Policies

In order to facilitate the formulation in the next subsection, we first express the value of the POSG at a stage t in terms of a *past joint policy* φ^t (as defined in Section 2.3) attained when all agents follow the joint decision rule δ^t and assuming that in future stages agents will act rationally, i.e., they follow a rational joint future policy $\pi^{t+1*}_i = \langle\delta^{t+1*}\dots\delta^{h-1*}\rangle$.

Conceptually, this enables us to treat the problem of finding a rational joint policy as a series of smaller problems, namely identification of a rational joint decision rule δ^{t*} at every stage. However, as we will show, a circular dependency exists: selection of δ^t is dependent on the future rational decision rule δ^{t+1*}, which in turn is dependent on φ^t and thus on δ^t.

We define the Q-value function at the final stage $t = h-1$, and give an inductive definition of the Q-value function at preceding stages. We then define the value function at every stage. Let the reward function in terms of a joint AOH and joint action be defined as $R(\vec{\theta}^t,a^t) \triangleq \sum_{s^t}\Pr(s^t|\vec{\theta}^t,b^0)R(s^t,a^t)$. The immediate reward for a joint AOH and a joint decision rule is then:

$$R(\vec{\theta}^t,\delta^t) \triangleq \sum_{a^t}\delta^t(a^t|\vec{\theta}^t)R(\vec{\theta}^t,a^t). \tag{4.1}$$

For the final stage $t = h-1$, the Q-value function reduces to this immediate reward, as there is no future value:

$$Q^*_{h-1}(\varphi^{h-1},\vec{\theta}^{h-1},\delta^{h-1}) \triangleq R(\vec{\theta}^{h-1},\delta^{h-1}). \tag{4.2}$$

Given an AOH and decision rule at stage t, it is possible to find a probability distribution over AOHs at the next stage, as an AOH at $t+1$ is the AOH at t concatenated with action a^t and observation o^{t+1}:

$$\begin{aligned}\Pr(\vec{\theta}^{t+1}|\vec{\theta}^t,\delta^t) &= \Pr(\langle\vec{\theta}^t,a^t,o^{t+1}\rangle|\vec{\theta}^t,\delta^t)\\ &= \Pr(o^{t+1}|\vec{\theta}^t,a^t)\delta^t(a^t|\vec{\theta}^t).\end{aligned} \tag{4.3}$$

For all stages except the final stage $t = h-1$, the value at future stages is propagated to the current stage using (4.3):

$$\begin{aligned}Q^*_t(\varphi^t,\vec{\theta}^t,\delta^t) \triangleq R(\vec{\theta}^t,\delta^t) + \\ \sum_{a^t}\sum_{o^{t+1}}\Pr(\vec{\theta}^{t+1}|\vec{\theta}^t,\delta^t)Q^*_{t+1}(\varphi^{t+1},\vec{\theta}^{t+1},\delta^{t+1*}),\end{aligned} \tag{4.4}$$

$$Q^*_t(\varphi^t,\delta^t) \triangleq \sum_{\vec{\theta}^t}\Pr(\vec{\theta}^t|b^0,\varphi^t)Q^*_t(\varphi^t,\vec{\theta}^t,\delta^t). \tag{4.5}$$

We use (4.5) to find rational decision rules for both agents. Consistent with (4.4), we show how to find δ^{t+1*}_i:

$$\delta^{t+1*}_1 = \operatorname*{argmax}_{\delta^{t+1}_1}\min_{\delta^{t+1}_2} Q^*_{t+1}(\varphi^{t+1},\delta^{t+1}_1\delta^{t+1}_2), \tag{4.6}$$

$$\delta^{t+1*}_2 = \operatorname*{argmin}_{\delta^{t+1}_2}\max_{\delta^{t+1}_1} Q^*_{t+1}(\varphi^{t+1},\delta^{t+1}_1\delta^{t+1}_2). \tag{4.7}$$

Using (4.2), (4.6) and (4.7), a rational joint decision rule δ^{h-1*} can be found by performing a maxminimization over immediate reward. Evaluation of $Q^*_{h-1}(\varphi^t, \delta^{h-1*})$ gives us the value at stage $t = h-1$, and (4.4) propagates the value to the preceding stages. As such, rationality for all stages follows by induction. We can now define the value function in terms of the past joint policy as:

$$V^*_t(\varphi^t) = \max_{\delta^t_1} \min_{\delta^t_2} Q^*_t(\varphi^t, \delta^t_1 \delta^t_2). \quad (4.8)$$

By (4.2), (4.6) and (4.7), δ^{t*} is dependent on δ^{t+1*}, and thus on the rational future joint policy. However, δ^{t+1*} can only be found if past joint policy φ^{t+1}, which includes δ^t, is known. This circular dependency on both the future and past joint policy makes multi-agent decision making in POSGs a difficult problem. Furthermore, in the Dec-POMDP case it is possible to find an exact solution using dynamic programming because we then search in the *finite* space of pure joint policies. In the zs-POSG case, we search in the infinitely large stochastic policy-space, rendering this impossible.

4.2 Plan-Time Sufficient Statistics

Even though there are infinitely many past joint policies, we do not expect that their effects on the game at a particular stage are completely arbitrary. In fact, in this section we propose to replace the dependence of the value function on past joint policies by a plan-time sufficient statistic that summarizes many past joint policies. As we will show, this new statistic allows us to break the circular dependency discussed in the previous subsection: with decision rule selection at stage t dependent on the new plan-time statistic rather than the past joint policy, agents can determine the rational partial policy π^{t*} if they know the statistic, regardless of choices made on stages 0 to t. Furthermore, as we will show in Section 5, the value function exhibits a certain concave/convex shape in statistic-space that may be exploitable.

Definition 3. The **plan-time sufficient statistic** for a general past joint policy φ^t, assuming b^0 is known, is a distribution over joint AOHs: $\sigma^t(\vec{\theta}^t) \triangleq \Pr(\vec{\theta}^t | b^0, \varphi^t)$.

In the collaborative Dec-POMDP case, these plan-time sufficient statistics fully capture the influence of the past joint policy in the zs-POSG case. We will prove that this also holds for the zs-POSG case, by showing that use of these statistics allows for redefinition of the equations from Section 4.1. We aim to express the value for a given decision rule δ^t in terms of a plan-time sufficient statistic, given that the agents act rationally at later stages. We first define the update rule for plan-time sufficient statistics:

$$\sigma^{t+1}(\vec{\theta}^{t+1}) \triangleq \Pr(o^{t+1} | \vec{\theta}^t, a^t) \delta^t(a^t | \vec{\theta}^t) \sigma^t(\vec{\theta}^t). \quad (4.9)$$

At the final stage $t = h-1$, the Q-value function reduces to the immediate reward, as there is no future value:

$$Q^*_{h-1}(\sigma^{h-1}, \vec{\theta}^{h-1}, \delta^{h-1}) \triangleq R(\vec{\theta}^{h-1}, \delta^{h-1}). \quad (4.10)$$

We then define the Q-value for all other stages as:

$$Q^*_t(\sigma^t, \vec{\theta}^t, \delta^t) \triangleq R(\vec{\theta}^t, \delta^t) + \sum_{a^t} \sum_{o^{t+1}} \Pr(\vec{\theta}^{t+1} | \vec{\theta}^t, \delta^t) Q^*_{t+1}(\sigma^{t+1}, \vec{\theta}^{t+1}, \delta^{t+1*}), \quad (4.11)$$

$$Q^*_t(\sigma^t, \delta^t) \triangleq \sum_{\vec{\theta}^t} \sigma^t(\vec{\theta}^t) Q^*_t(\sigma^t, \vec{\theta}^t, \delta^t). \quad (4.12)$$

We use (4.12) to find rational decision rules for both agents:

$$\delta^{t+1*}_1 = \operatorname*{argmax}_{\delta^{t+1}_1} \min_{\delta^{t+1}_2} Q^*_{t+1}(\sigma^{t+1}, \delta^{t+1}_1 \delta^{t+1}_2), \quad (4.13)$$

$$\delta^{t+1*}_2 = \operatorname*{argmin}_{\delta^{t+1}_2} \max_{\delta^{t+1}_1} Q^*_{t+1}(\sigma^{t+1}, \delta^{t+1}_1 \delta^{t+1}_2). \quad (4.14)$$

We formally prove the equivalence of (4.5) and (4.12).

LEMMA 2. *σ^t is a sufficient statistic for the value of the zs-POSG, i.e. $Q^*_t(\sigma^t, \vec{\theta}^t, \delta^t) = Q^*_t(\varphi^t, \vec{\theta}^t, \delta^t), \forall t \in 0 \dots h-1, \forall \vec{\theta}^t \in \Delta(\vec{\Theta}^t), \forall \delta^t$.*

PROOF. The proof is listed in appendix A. □

We define the value function for a two-player, zs-POSG, similar to (4.8), but in terms of σ^t, as follows:

$$V^*_t(\sigma^t) \triangleq \max_{\delta^t_1} \min_{\delta^t_2} Q^*_t(\sigma^t, \delta^t_1 \delta^t_2). \quad (4.15)$$

Although we have now identified the value at a single stage of the game, we can not implement a backward inductive approach directly, as decisions on stages before t affect σ^t. However, given σ^t, we can compute π^{t*} without knowing φ^t.

4.3 Reduction to NOSG

A recent development in the field of Dec-POMDPs, is that the (decentralized) Dec-POMDP model can be reduced to a (centralized) special case of POMDP (a non-observable MDP, or NOMDP) [9, 14, 2, 17], which allows POMDP solution methods to be employed in the context of Dec-POMDPs. The proposed plan-time statistics for Dec-POMDPs [16] precisely correspond to the belief in the centralized model.

Since we have shown that it is possible to generalize the sufficient plan-time statistics to zero-sum POSGs, it is reasonable to expect that there also is a reduction to a centralized model possible. Here we present this reduction to a centralized model, to which we refer as a *Non-Observable Stochastic Game* (NOSG). We do not provide the full background of the reduction for the Dec-POMDP case, but refer to [17]. The difference between that reduction and the one we present next, is that the zs-POSG is reduced to a centralized model where the joint AOH acts as the state (in order to support stochastic policies, see also the discussion in [16]), and that a maxmin-operator is used to compute the value of the game instead of a max-operator.

Definition 4. A **plan-time Non-Observable Stochastic Game** for a zs-POSG is a tuple $\langle \dot{S}, \dot{A}, \dot{\mathcal{O}}, \dot{T}, \dot{O}, \dot{R}, \dot{b}^0 \rangle$:

- A set of augmented states $\dot{S}$. Each state $\dot{s}^t$ corresponds to a joint AOH $\vec{\theta}^t$.
- A set of joint actions $\dot{A}$. Each action $\dot{a}^t$ corresponds to a joint decision rule δ^t.
- A set of joint observations $\dot{\mathcal{O}} = \{\text{NULL}\}$ that only contains the NULL observation.
- A transition function as specified in (4.3):
$$\dot{T}(\dot{s}^{t+1} | \dot{s}^t, \dot{a}^t) = \Pr(\vec{\theta}^{t+1} | \vec{\theta}^t, \delta^t).$$
- An observation function $\dot{O}$ that specifies that observation NULL is received with probability 1.
- A reward function as specified in (4.1):
$$\mathcal{R}(\dot{s}^t, \dot{a}^t) = R(\vec{\theta}^t, \delta^t).$$
Note that this gives the reward for agent 1 in the POSG, and that the centralized agent aims to *maxminimize* it.
- The initial belief over states $\dot{b}^0 \in \Delta(\dot{S})$.

In the NOSG model, a centralized agent conditions its choice on belief over the augmented states $\dot{b} \in \Delta(\dot{S})$, which corresponds to the belief over joint AOHs captured in the statistic $\sigma^t \in \Delta(\Theta^t)$. As such, a value function formulation for the NOSG can be given in accordance with (4.15). Note that while the NULL observation is shared, the state and action contain entries for both agents in the original zs-POSG.

A zs-POSG can also be converted to a POMDP by fixing the policies of one agent [11], which leads to a model where the information state $b(s, \theta_j)$ is a distribution over states and AOHs of the other agent. In contrast, our NOSG formulation maintains a belief over joint AOHs, and does not require fixing the policy of any agent. That is, where the approach of Nair leads to a single-agent model that can be used to compute a best-response (which of course can be employed to compute an equilibrium, see, e.g., [18]), our conversion leads to a multi-agent model that can potentially be used to compute a Nash equilibrium directly.

Observation 1. Ghosh et al. [4] treat a special type of zs-POSG in which both agents receive the same (i.e., shared) observations. They show that, for such a problem, a reduction to a completely observable model is possible and that in the infinite horizon case a value and rational joint policy must exist. As the NOSG is a specific case of such a zs-POSG (with one shared NULL observation), these results directly extend to the plan-time NOSG for our general zs-POSG. As such, our reduction shows that the properties established by Ghosh et al. for a limited subset of zs-POSGs, in fact extend to all zs-POSGs.

5. CONCAVITY/CONVEXITY

Similar to the decomposition of the distribution over types in BGs (in equations 3.7 and 3.8), the plan-time sufficient statistic can be decomposed into a marginal term $\sigma^t_{m,i}$ and conditional term $\sigma^t_{c,i}$. We will prove formally that the value function of the two-player zs-POSG exhibits two properties: It is, at every stage t, concave in marginal-space for agent 1, $\Delta(\vec{\Theta}^t_1)$, and convex in marginal-space for agent 2, $\Delta(\vec{\Theta}^t_2)$. We first define best-response value functions V^{BR1}_t and V^{BR2}_t. Using these definitions, we prove the concave/convex properties of V^*_t.

Figure 1 provides intuition on how the best-response value functions relate to the concave/convex value function. For sample statistics $\sigma^t_{(1)} - \sigma^t_{(4)}$, the best-response value functions to some partial policies π^t_2 are shown in blue. If these are the minimal best-response value functions, then they can be used to construct the concave value function (in red). As we will show, selecting a 'slice' in statistic-space corresponding to a single conditional $\sigma^t_{c,1}$ ($\sigma^t_{c,2}$) guarantees the concave (convex) shape of the value function.

We will give a recursive definition of the value in terms of a statistic and a joint partial policy, V_t, that makes the distinction between the immediate reward and the value propagated from future stages explicit. Let us first define the *immediate reward* function Q^R_t that gives the reward attained at the current stage given a statistic and joint decision rule:

$$Q^R_t(\sigma^t, \delta^t) \triangleq \sum_{\vec{\theta}^t} \sigma^t(\vec{\theta}^t) R(\vec{\theta}^t, \delta^t) = \sigma^t \cdot R_{\delta^t}. \quad (5.1)$$

Here, R_{δ^t} is a vector containing reward for every joint AOH, attained when agents follow the given joint decision rule δ^t.

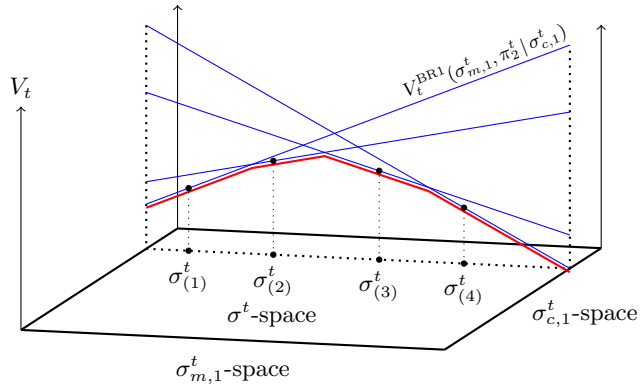

Figure 1: An abstract visualisation of the decomposition of statistic-space into marginal-space and conditional-space.

We formally define V_t as follows:

$$V_t(\sigma^t, \pi^t) \triangleq \begin{cases} Q^R_t(\sigma^t, \delta^t) & \text{if } t = h-1 \\ Q^R_t(\sigma^t, \delta^t) + V_{t+1}(\text{U}_{\text{ss}}(\sigma^t, \delta^t), \pi^{t+1}) & \text{otherwise.} \end{cases} \quad (5.2)$$

Here, U_{ss} is the statistic update function derived from (4.9). Note that through the statistic update, the future value is dependent on the decision rule δ^t. Using (5.2), we define the value function of the zs-POSG as a maxminimization or minmaximization over π^t_i, similar to (2.1):

$$V^*_t(\sigma^t) \triangleq \max_{\pi^t_1} \min_{\pi^t_2} V_t(\sigma^t, \pi^t_1 \pi^t_2) = \min_{\pi^t_2} \max_{\pi^t_1} V_t(\sigma^t, \pi^t_1 \pi^t_2). \quad (5.3)$$

Best-response value functions in terms of σ^t and π^t_i are defined as V^{BR1}_t and V^{BR2}_t, similar to (3.4) and (3.5), for agent 1 and agent 2 respectively. Similar to (3.6), we have:

$$V^*_t(\sigma^t) = \min_{\pi^t_2} V^{\text{BR1}}_t(\sigma^t, \pi^t_2) = \max_{\pi^t_1} V^{\text{BR2}}_t(\sigma^t, \pi^t_1). \quad (5.4)$$

We will use these definitions to prove the concave/convex properties of the value function. We have already shown, in Section 3.1, that the value function of a family of zero-sum Bayesian Games exhibits concavity and convexity in terms of the marginals parts of the type distribution σ for respectively agent 1 and 2. We formally prove that this result directly extends to the final stage of the zs-POSG $t = h-1$.

LEMMA 3. *For a family of BGs $\mathcal{F}$, if*

1. *joint actions are equal to joint actions of the POSG,*
2. *types θ correspond to AOHs $\vec{\theta}^{h-1}$,*
3. *the initial distribution over types σ is equal to σ^{h-1},*

then the value function at the final stage of the POSG, V^*_{h-1}, and the value function of the family of Bayesian Games, $V^*_{\mathcal{F}}$, are equivalent.

PROOF. The proof is listed in Appendix A. □

By the results of Theorem 1 (i.e. that the value function $V^*_{\mathcal{F}}$ exhibits concavity and convexity in marginal-spaces of agent 1 and 2 respectively) and Lemma 3, V^*_{h-1} is concave in $\Delta(\vec{\Theta}^{h-1}_1)$ for all $\sigma^{h-1}_{c,1}$, and convex in $\Delta(\vec{\Theta}^{h-1}_2)$ for all $\sigma^{h-1}_{c,2}$.

Even though the final stage is equal to a family of Bayesian Games, our approach is substantially different from approaches that represent a POSG as a *series* of BGs [3] and derivative works [19]. In fact, all other stages (0 to $h-2$) cannot be represented as a family of BGs as defined in Sec-

tion 3, as the relation between the decision rule δ^t and the value is non-linear. Nevertheless, we show that the analysis of the value function extends to these stages as well.

We first show that the best-response value functions V_t^{BR1} and V_t^{BR2} are linear in their respective marginal-spaces. Using this result, we prove that V_t^* exhibits concavity in $\Delta(\vec{\Theta}_1^t)$ at every stage for every $\sigma_{c,1}^t$, and convexity in $\Delta(\vec{\Theta}_2^t)$ for every $\sigma_{c,2}^t$. For this purpose, let us define a vector that contains the value (immediate reward *and* future value) for agent 1 for each individual AOH $\vec{\theta}_1^t$, given that agent 2 follows the partial policy π_2^t:

$$\nu_{[\sigma_{c,1}^t,\pi_2^t]}(\vec{\theta}_1^t) \triangleq \max_{a_1^t}\left[\sum_{\vec{\theta}_2^t}\sigma_{c,1}^t(\vec{\theta}_2^t|\vec{\theta}_1^t)\sum_{a_2^t}\delta_2^t(a_2^t|\vec{\theta}_2^t)\left(R(\vec{\theta}^t,a^t)+\sum_{o^{t+1}}\Pr(o^{t+1}|\vec{\theta}^t,a^t)\nu_{[\sigma_{c,1}^{t+1},\pi_2^{t+1}]}(\vec{\theta}_1^{t+1})\right)\right] \quad (5.5)$$

Note that this is a recursive definition, and that the next AOH is $\vec{\theta}_1^{t+1} = \langle\vec{\theta}_1^t, a_1^t, o_1^{t+1}\rangle$. We have established in Lemma 3 that the value function at the final stage $t = h-1$ is equivalent to the value function of a family of zs-BGs. Thus, the zs-POSG value vector from (5.5) reduces to the value vector for a family of zs-BGs from (3.9) when we make the substitutions of Lemma 3:

$$\nu_{[\sigma_{c,1}^{h-1},\pi_2^{h-1}]}(\vec{\theta}_1^{h-1}) = r_{[\sigma_{c,1}^{h-1},\pi_2^{h-1}]}(\vec{\theta}_1^{h-1}). \quad (5.6)$$

Intuitively, this also makes sense, as at the final stage the future value is zero and the partial policy π_2^{h-1} only contains a single decision rule δ_2^{h-1}. The vector $\nu_{[\sigma_{c,2}^t,\pi_1^t]}$ is defined similarly.

LEMMA 4. *(1) V_t^{BR1} is linear in $\Delta(\vec{\Theta}_1^t)$ for a given $\sigma_{c,1}^t$ and π_2^t, and (2) V_t^{BR2} is linear in $\Delta(\vec{\Theta}_2^t)$ for a given $\sigma_{c,2}^t$ and π_1^t, for all stages $t = 0,\ldots,h-1$. More specifically, we can write these functions as the inner products of a marginal $\sigma_{m,i}^t$ and a vector:*

1. $V_t^{BR1}(\sigma_{m,1}^t,\pi_2^t|\sigma_{c,1}^t) = \sigma_{m,1}^t\cdot\nu_{[\sigma_{c,1}^t,\pi_2^t]}$, (5.7)
2. $V_t^{BR2}(\sigma_{m,2}^t,\pi_1^t|\sigma_{c,2}^t) = \sigma_{m,2}^t\cdot\nu_{[\sigma_{c,2}^t,\pi_1^t]}$. (5.8)

PROOF. We prove this by induction. By the result of Lemma 3, we know the value function at stage $t = h-1$ to be equivalent to that of a family of BGs. As such, the result of Lemma 1 is a base case for the proof. The full proof is listed in Appendix A. □

THEOREM 2. *V_t^* is (1) concave in $\Delta(\vec{\Theta}_1^t)$ for a given $\sigma_{c,1}^t$, and (2) convex in $\Delta(\vec{\Theta}_2^t)$ for a given $\sigma_{c,2}^t$. More specifically, V_t^* is respectively a minimization over linear functions in $\Delta(\vec{\Theta}_1^t)$ and a maximization over linear functions in $\Delta(\vec{\Theta}_2^t)$:*

1. $V_t^*(\sigma_{m,1}^t,|\sigma_{c,1}^t) = \min_{\pi_2^t}\left[\sigma_{m,1}^t\cdot\nu_{[\sigma_{c,1}^t,\pi_2^t]}\right]$,
2. $V_t^*(\sigma_{m,2}^t|\sigma_{c,2}^t) = \max_{\pi_1^t}\left[\sigma_{m,2}^t\cdot\nu_{[\sigma_{c,2}^t,\pi_1^t]}\right]$.

PROOF. Filling in the result of Lemma 4 gives:

$$V_t^*(\sigma_{m,1}^t,|\sigma_{c,1}^t) \overset{\{5.4\}}{=} \min_{\pi_2^t} V_t^{\text{BR1}}(\sigma_{m,2}^t,\pi_2^t|\sigma_{c,1}^t) \overset{\{5.7\}}{=} \min_{\pi_2^t}\left[\sigma_{m,1}^t\cdot\nu_{[\sigma_{c,1}^t,\pi_2^t]}\right].$$

The proof for item 2 is analogous to that of item 1. □

The importance of this theorem is that it may enable the development of new solution methods for these classes of games. To draw the parallel, many POMDP solution methods successfully exploit the fact that a POMDP value function is piecewise-linear and convex in belief-space [21, 28] (which is similar to the statistic-space in the zs-POSG), and recently such results have been extended to the decentralized cooperative case [9, 2].

In future work, we aim to exploit the found structure. Our results indicate that within 'slices' of statistic-space that correspond to conditional plan-time statistics, it is possible to approximate the value function using piecewise-linear and concave/convex functions (as can be seen in Figure 1). Therefore, using the value vector definition in (5.5), it may be possible to adapt solution methods for POMDPs [28] or Dec-POMDPs [9], and apply it at (sampled) conditionals at the final stage. Corresponding marginals can be sampled, or selected in an intelligent way [21, 25]. For every marginal, a value vector can be computed, and, using (5.5), an exhaustive backup can be performed. This gives us a set of value vectors at stage 0, which describe a piecewise linear and convex (for $i = 1$) or concave (for $i = 2$) value function in $\sigma_{m,i}^0$-space. As the initial marginal is known, we can then compute the approximate value of the zs-POSG.

While it is conceivable that a similar reduction to a centralized model as the one in Section 4.3 is possible for general-sum POSGs, we point out that our results do not extend to the general-sum case directly. In general, there can be many equilibria, and as such one will need to reason about sets of possible outcomes rather than just one value.

6. RELATED WORK

In this section we describe related methods for (zero-sum) POSGs. We do not treat the work by Ghosh et al. [4], which was already treated in Section 4.3.

A recent paper that is similar in spirit to ours is by Nayyar et al. [13] who introduce a so-called Common Information Based Conditional Belief — a probability distribution over AOHs and the state conditioned on common information — and use it to design a dynamic-programming approach for zs-POSGs. This method converts stages of the zs-POSG to Bayesian Games for which the type distribution corresponds to the statistic at that stage. However, since their proposed statistic is a distribution over joint AOHs *and states*, the statistic we propose in this paper is more compact. Furthermore, Nayyar et al. do not provide any results regarding the structure of the value function, which is the main contribution of our paper.

Hansen et al. [6] present a dynamic-programming approach for finite-horizon (general sum) POSGs that works by iteratively constructing sets of one-step-longer (pure) policies for all agents. At every iteration, the sets of individual policies are pruned by removing dominated policies. This pruning is based on a different statistic called *multi-agent belief*: a distribution over states and policies of other agents. Such a multi-agent belief is sufficient from the perspective of an individual agent to determine its best response (or whether some of its policies are dominated). However, it is not a sufficient statistic for the past joint policy from a designer perspective (as is the proposed plan-time sufficient statistic in this paper).

There are a number of papers from the game theory literature that do present structural (concave/convex) results

on the value function for zero-sum games. The models for which these results have been proven are substantially less general than the zs-POSG model that we treat in this paper.

One line of work targets 'zero-sum sequential game with incomplete information' [10, 22, 23, 24]. These can best be understood as a particular class of two-player extensive games that lie in between Bayesian games and POSGs: at the start of the game, nature determines the state (essentially a joint type) from which each agent makes a private observation (i.e., individual type), and subsequently the agents take actions in turns thereby observing the actions of the other player. For various flavors of such games, it has been shown that a value function exists and has a concave/convex structure: incomplete information on one side [24], and cases with incomplete information on both sides where 'observations are independent' (i.e., where the distribution over joint types is a product of individual type distributions) [22] or dependent (general joint type distributions) [10, 23]. These results, however, crucially depend on the alternating actions and the static state and therefore do not extend to zs-POSGs.

Another class of models for which structural concave/convex results are known are 'repeated zero-sum games with incomplete information on one side' [27]. In these games action selection is simultaneous and agents observe the *past* actions of the opposing agent. However, found results do not directly extend to the zs-POSG setting, where there is incomplete information on both sides and agents do not necessarily observe the actions of the opposing agent.

A game-theoretic model that is closely related to the POSG model is the *Interactive POMDP* or *I-POMDP* [5]. In I-POMDPs, a (subjective) belief is constructed from the perspective of a single agent as a probability distribution over states and the *types*, ζ, of all other agents. As the agents are rational, each individual AOH induces one type in the I-POMDP. Therefore, the belief of agent i in the I-POMDP, which is a distribution $b(s, \zeta_j)$, can be seen to correspond to a conditional $\sigma^t_{c,i}(\vec{\theta}^t_j|\vec{\theta}^t_i)$ in the zs-POSG.

7. CONCLUSIONS

This paper gives a structural result on the shape of the value function of two-player zero-sum games of incomplete information, for games of static state and dynamic state, typically modeled as a Bayesian Game (BG) and Partially Observable Stochastic Game (POSG) respectively. We introduced the concept of a family of zero-sum BGs, $\mathcal{F}$, the members of which share all elements but the plan-time statistic: a probability distribution over joint types. Using the fact that this probability distribution can be decomposed into a marginal and a conditional term, we gave proof that in the zero-sum case its value function $V^*_{\mathcal{F}}$ exhibits concavity (convexity) in the marginal-space of the maximizing (minimizing) agent. We gave two formulations for the value function of the zero-sum POSG: One in terms of the past joint policy, and one in terms of the recently introduced *sufficient plan-time statistics* (originally used in the collaborative setting [16]). We proved the equivalence of both formulations, and gave formal proof that at the final stage of the game, the latter formulation is equivalent to $V^*_{\mathcal{F}}$, meaning that it exhibits a similar structure. Using this result, we gave inductive proof that the zero-sum POSG value function exhibits similar concave/convex properties at all stages. Finally, we discussed possible uses of the found properties.

Acknowledgments.
This research is supported by the NWO Innovational Research Incentives Scheme Veni (#639.021.336) and NWO DTC-NCAP (#612.001.109) project.

APPENDIX

A. PROOF OF LEMMA'S

PROOF OF LEMMA 1. We expand (3.2) in order to bring the marginal term to the front of the equation:

$$Q_{\mathcal{F}}(\sigma,\delta) \overset{\{3.2\}}{=} \sum_{\theta}\sigma(\theta)\sum_{a}\delta(a|\theta)R(\theta,a) \tag{A.1}$$

$$= \sum_{\theta_1}\sigma_{m,1}(\theta_1)\sum_{\theta_2}\sigma_{c,1}(\theta_2|\theta_1)\sum_{a_1}\delta_1(a_1|\theta_1) \sum_{a_2}\delta_2(a_2|\theta_2)R(\theta_1\theta_2,a_1a_2). \tag{A.2}$$

A maximization over stochastic decision rules conditioned on the AOH θ_1 is equal to choosing a maximizing action for each of these AOHs. Thus, we can rewrite the best-response value function from (3.4) as follows:

$$V^{\mathrm{BR1}}_{\mathcal{F}}(\sigma,\delta_2) = \max_{\delta_1} Q^R_t(\sigma,\delta_1\delta_2)$$

$$\overset{\{A.2\}}{=} \max_{\delta_1}\left[\sum_{\theta_1}\sigma_{m,1}(\theta_1)\sum_{\theta_2}\sigma_{c,1}(\theta_2|\theta_1)\sum_{a_1}\delta_1(a_1|\theta_1) \sum_{a_2}\delta_2(a_2|\theta_2)R(\theta_1\theta_2,a_1a_2)\right]$$

$$= \sum_{\theta_1}\sigma_{m,1}(\theta_1)\max_{a_1}\left[\sum_{\theta_2}\sigma_{c,1}(\theta_2|\theta_1)\sum_{a_2}\delta_2(a_2|\theta_2)R(\theta,a)\right]$$

$$\overset{\{3.9\}}{=} \sum_{\theta_1}\sigma_{m,1}(\theta_1)r_{[\sigma_{c,1},\delta_2]}(\theta_1) = \sigma_{m,1}\cdot r_{[\sigma_{c,1},\delta_2]}.$$

As it is possible to write $V^{\mathrm{BR1}}_{\mathcal{F}}$ as an inner product of the marginal distribution $\sigma_{m,1}$ and a vector, $V^{\mathrm{BR1}}_{\mathcal{F}}$ is linear in $\Delta(\Theta_1)$ for all $\sigma_{c,1}$ and δ_2. Analogously, $V^{\mathrm{BR2}}_{\mathcal{F}}$ is linear in $\Delta(\Theta_2)$ for all $\sigma_{c,2}$ and δ_1. □

PROOF OF LEMMA 2. The proof is largely identical to the proof of correctness of sufficient statistics in the collaborative setting [16]. For the final stage $t = h-1$, we have the following:

$$Q^*_t(\varphi^t,\vec{\theta}^t,\delta^t) = R(\vec{\theta}^t,\delta^t) = Q^*_t(\sigma^t,\vec{\theta}^t,\delta^t)$$

. As induction hypothesis we assume that at stage $t+1$, σ^{t+1} is a sufficient statistic, i.e.:

$$Q^*_{t+1}(\varphi^{t+1},\vec{\theta}^{t+1},\delta_{t+1}) = Q^*_{t+1}(\sigma^{t+1},\vec{\theta}^{t+1},\delta_{t+1}). \tag{A.3}$$

We aim to show that at stage t, σ^t is a sufficient statistic:

$$Q^*_t(\varphi^t,\vec{\theta}^t,\delta^t) = Q^*_t(\sigma^t,\vec{\theta}^t,\delta^t). \tag{A.4}$$

We substitute the induction hypothesis into (4.4):

$$Q^*_t(\varphi^t,\vec{\theta}^t,\delta^t)$$

$$\overset{\{4.4\}}{=} R(\vec{\theta}^t,\delta^t) + \sum_{a^t}\sum_{o^{t+1}}\Pr(\vec{\theta}^{t+1}|\vec{\theta}^t,\delta^t)Q^*_{t+1}(\varphi^{t+1},\vec{\theta}^{t+1},\delta^{t+1*})$$

$$\overset{\{A.3\}}{=} R(\vec{\theta}^t,\delta^t) + \sum_{a^t}\sum_{o^{t+1}}\Pr(\vec{\theta}^{t+1}|\vec{\theta}^t,\delta^t)Q^*_{t+1}(\sigma^{t+1},\vec{\theta}^{t+1},\delta^{t+1*})$$

$$\overset{\{4.11\}}{=} Q^*_t(\sigma^t,\vec{\theta}^t,\delta^t).$$

Furthermore, decision rules $\delta_{1,pp}^{t+1*}$ (based on the past joint policy) and $\delta_{1,ss}^{t+1*}$ (based on the sufficient statistic) are equal:

$$\begin{aligned}
\delta_{1,pp}^{t+1*} &\overset{\{4.6\}}{=} \\
&\operatorname*{argmax}_{\delta_1^{t+1}} \min_{\delta_2^{t+1}} \sum_{\vec{\theta}^{t+1}} \Pr(\vec{\theta}^{t+1}|b^0, \varphi^{t+1}) Q_{t+1}^*(\varphi^{t+1}, \vec{\theta}^{t+1}, \delta^{t+1}) \\
&\overset{\{\text{Def.3}\}}{=} \operatorname*{argmax}_{\delta_1^{t+1}} \min_{\delta_2^{t+1}} \sum_{\vec{\theta}^{t+1}} \sigma^{t+1}(\vec{\theta}^{t+1}) Q_{t+1}^*(\sigma^{t+1}, \vec{\theta}^{t+1}, \delta^{t+1}) \\
&\overset{\{4.13\}}{=} \delta_{1,ss}^{t+1*}.
\end{aligned}$$

Analogous reasoning holds for $\delta_{2,ss}^{t+1*}$. Thus, by induction, σ^t is a sufficient statistic for φ^t, $\forall t \in 0 \dots h-1$. □

Proof of Lemma 3. Given is that:

1. joint actions are equal to joint actions of the POSG,
2. types θ correspond to AOHs $\vec{\theta}^{h-1}$,
3. the initial distribution over types σ is equal to σ^{h-1},

At stage $t = h-1$, the partial policy π^{h-1} contains only the joint decision rule δ^{h-1}. As such, we have:

$$\begin{aligned}
&V_t(\sigma^{h-1}, \pi^{h-1}) = V_{h-1}(\sigma^{h-1}, \delta^{h-1}) \\
&\overset{\{5.2\}}{=} Q_t^R(\sigma^{h-1}, \delta^{h-1}) \overset{\{5.1\}}{=} \sum_{\vec{\theta}^{h-1}} \sigma^{h-1}(\vec{\theta}^{h-1}) R(\vec{\theta}^{h-1}, \delta^{h-1}) \\
&\overset{\{4.1\}}{=} \sum_{\vec{\theta}^{h-1}} \sigma^{h-1}(\vec{\theta}^{h-1}) \sum_{a^{h-1}} \delta(a^{h-1}|\vec{\theta}^{h-1}) R(\vec{\theta}^{h-1}, a^{h-1}).
\end{aligned}$$

By premises 1-3, this is equal to $Q_{\mathcal{F}}$:

$$Q_{\mathcal{F}}(\sigma, \delta) \overset{\{3.2\}}{=} \sum_{\theta} \sigma(\theta) \sum_{a} \delta(a|\theta) R(\theta, a).$$

As such, we have $Q_{\mathcal{F}}(\sigma, \delta) = V_{h-1}(\sigma^{h-1}, \delta^{h-1})$. From (3.3) and (5.3), it follows trivially that $V_{\mathcal{F}}^*(\sigma) = V_{h-1}^*(\sigma^{h-1})$. □

Proof of Lemma 4. By the results of Lemma 1 and Lemma 3, we know the best-reponse value function V_{h-1}^{BR1} to be linear in $\Delta(\vec{\Theta}_1^{h-1})$ For all other stages, we assume the following induction hypothesis:

$$V_{t+1}^{\text{BR1}}(\sigma_{m,1}^{t+1}, \pi_2^{t+1}|\sigma_{c,1}^{t+1}) = \sigma_{m,1}^{t+1} \cdot \nu_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}. \quad \text{(A.5)}$$

For the inductive step we aim to prove that at the current stage t the following holds:

$$V_t^{\text{BR1}}(\sigma_{m,1}^t, \pi_2^t|\sigma_{c,1}^t) = \sigma_{m,1}^t \cdot \nu_{[\sigma_{c,1}^t, \pi_2^t]}. \quad \text{(A.6)}$$

We expand the definition of V_t^{BR1}. For notational convenience, we write σ^t instead of $\sigma_{m,i}^t|\sigma_{c,i}^t$, but keep in mind that we only consider the statistics corresponding to the conditional $\sigma_{c,i}^t$.

$$\begin{aligned}
&V_t^{\text{BR1}}(\sigma^t, \pi_2^t) \overset{\{5.4\}}{=} \max_{\pi_1^t} V_t(\sigma^t, \pi_1^t \pi_2^t) \\
&\overset{\{5.2\}}{=} \max_{\pi_1^t} \left[Q_t^R(\sigma^t, \delta_1^t \delta_2^t) + V_{t+1}(\mathrm{U}_{\text{ss}}(\sigma^t, \delta^t), \pi_1^{t+1} \pi_2^{t+1}) \right] \\
&= \max_{\delta_1^t} \max_{\pi_1^{t+1}} \left[Q_t^R(\sigma^{t+1}, \delta_1^t \delta_2^t) + V_{t+1}(\sigma^{t+1}, \pi_1^{t+1} \pi_2^{t+1}) \right] \\
&= \max_{\delta_1^t} \left[Q_t^R(\sigma^t, \delta_1^t \delta_2^t) + \max_{\pi_1^{t+1}} \left[V_{t+1}(\sigma^{t+1}, \pi_1^{t+1} \pi_2^{t+1}) \right] \right] \\
&\overset{\{5.4\}}{=} \max_{\delta_1^t} \left[Q_t^R(\sigma^t, \delta_1^t \delta_2^t) + V_{t+1}^{\text{BR1}}(\sigma^{t+1}, \pi_1^{t+1} \pi_2^{t+1}) \right]. \quad \text{(A.7)}
\end{aligned}$$

We make the decomposition of σ^t into the marginal and conditional terms explicit again. Immediate reward Q_t^R can be expanded similar to (A.2):

$$\begin{aligned}
Q_t^R(\sigma_{m,1}^t, \delta_1^t \delta_2^t|\sigma_{c,1}^t) = \sigma_{m,1}^t(\vec{\theta}_1^t) \sigma_{c,1}^t(\vec{\theta}_2^t|\vec{\theta}_1^t) \sum_{a_1^t} \delta_1^t(a_1^t|\vec{\theta}_1^t) \\
\sum_{a_2^t} \delta_2^t(a_2^t|\vec{\theta}_2^t) R(\vec{\theta}_1^t \vec{\theta}_2^t, a_1^t a_2^t). \quad \text{(A.8)}
\end{aligned}$$

We expand V_{t+1}^{BR1} using the induction hypothesis in order to bring the marginal distribution $\sigma_{m,1}^t$ to the front:

$$\begin{aligned}
&V_{t+1}^{\text{BR1}}(\sigma_{m,1}^{t+1}, \pi_2^{t+1}|\sigma_{c,1}^{t+1}) \overset{\{A.5\}}{=} \sigma_{m,1}^{t+1} \cdot \nu_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]} \\
&= \sum_{\vec{\theta}_1^{t+1}} \sigma_{m,1}^{t+1}(\vec{\theta}_1^{t+1}) \nu_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}(\vec{\theta}_1^{t+1}) \\
&\overset{\{4.9\}}{=} \sum_{\vec{\theta}_1^t} \sigma_{m,1}^t(\vec{\theta}_1^t) \sum_{\vec{\theta}_2^t} \sigma_{c,1}^t(\vec{\theta}_2^t|\vec{\theta}_1^t) \sum_{a_1^t} \delta_1^t(a_1^t|\vec{\theta}_1^t) \sum_{a_2^t} \delta_2^t(a_2^t|\vec{\theta}_2^t) \\
&\quad \sum_{o_1^{t+1}} \sum_{o_2^{t+1}} \Pr(o_1^{t+1} o_2^{t+1}|\vec{\theta}_1^t \vec{\theta}_2^t, a_1^t a_2^t) \nu_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}(\vec{\theta}_1^{t+1}).
\end{aligned} \quad \text{(A.9)}$$

Filling the expanded equations into (A.7) and factorizing gives:

$$\begin{aligned}
&V_t^{\text{BR1}}(\sigma_{m,1}^t, \pi_2^t|\sigma_{c,1}^t) \overset{\{5.4\}}{=} \\
&\max_{\delta_1^t} \left[Q_t^R(\sigma_{m,}^t, \delta_1^t \delta_2^t|\sigma_{c,1}^t) + V_{t+1}^{\text{BR1}}(\sigma_{m,1}^{t+1}, \pi_2^{t+1}|\sigma_{c,1}^{t+1}) \right] \overset{\{A.8, A.9\}}{=} \\
&\max_{\delta_1^t} \Bigg[\sum_{\vec{\theta}_1^t} \sigma_{m,1}^t(\vec{\theta}_1^t) \sum_{\vec{\theta}_2^t} \sigma_{c,1}^t(\vec{\theta}_2^t|\vec{\theta}_1^t) \sum_{a_1^t} \delta_1^t(a_1^t|\vec{\theta}_1^t) \sum_{a_2^t} \delta_2^t(a_2^t|\vec{\theta}_2^t) \\
&\left(R(\vec{\theta}^t, a^t) + \sum_{o^{t+1}} \Pr(o^{t+1}|\vec{\theta}^t, a^t) \nu_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}(\vec{\theta}_1^{t+1}) \right) \Bigg] \quad \text{(A.10)}
\end{aligned}$$

Note that the vector is indexed by the conditional $\sigma_{c,1}^{t+1}$. While this conditional is dependent on δ_2^t, it is not dependent on δ_1^t, allowing us to remove the maximization over decision rules δ_1^t from the equation. As a maximization over decision rules conditioned on individual AOH $\vec{\theta}_1^t$ is equal to choosing the maximizing action for each of these AOHs, we can rewrite (A.10) as follows:

$$\begin{aligned}
&V_t^{\text{BR1}}(\sigma^t, \pi_2^t) \\
&= \sum_{\vec{\theta}_1^t} \sigma_{m,1}^t(\vec{\theta}_1^t) \max_{a_1^t} \Bigg[\sum_{\vec{\theta}_2^t} \sigma_{c,1}^t(\vec{\theta}_2^t|\vec{\theta}_1^t) \sum_{a_2^t} \delta_2^t(a_2^t|\vec{\theta}_2^t) \\
&\quad \left(R(\vec{\theta}^t, a^t) + \sum_{o^{t+1}} \Pr(o^{t+1}|\vec{\theta}^t, a^t) \nu_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}(\vec{\theta}_1^{t+1}) \right) \Bigg] \\
&\overset{\{5.5\}}{=} \sum_{\vec{\theta}_1^t} \sigma_{m,1}^t(\vec{\theta}_1^t) \nu_{[\sigma_{c,1}^t, \pi_2^t]}(\vec{\theta}_1^t) = \sigma_{m,1}^t \cdot \nu_{[\sigma_{c,1}^t, \pi_2^t]}. \quad \text{(A.11)}
\end{aligned}$$

This corresponds to (A.6). Therefore, by induction, best-response value function V_t^{BR1} is linear in $\Delta(\vec{\Theta}_1^t)$ for a given $\sigma_{c,1}^t$ and π_2^t, for all stages $t = 0 \dots h-1$. Analogously, V_t^{BR2} is a linear function in $\Delta(\vec{\Theta}_2^t)$ for a given $\sigma_{c,2}^t$ and π_1^t, for all stages $t = 0 \dots h-1$. □

REFERENCES

[1] G. B. Dantzig. *Games and Linear Programs*. Princeton University Press, 1998.

[2] J. S. Dibangoye, C. Amato, O. Buffet, and F. Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. In *In Proceedings of the International Joint Conference on Artificial Intelligence*, 2013.

[3] R. Emery-Montemerlo, G. Gordon, J. Schneider, and S. Thrun. Approximate solutions for partially observable stochastic games with common payoffs. In *In Proceedings of International Joint Conference on Autonomous Agents and Multi Agent Systems*, 2004.

[4] M. Ghosh, D. McDonald, and S. Sinha. Zero-sum stochastic games with partial information. *Journal of Optimization Theory and Applications*, 121(1):99–118, 2004.

[5] P. J. Gmytrasiewicz and P. Doshi. Interactive pomdps: Properties and preliminary results. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pages 1374–1375. IEEE Computer Society, 2004.

[6] E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic Programming for Partially Observable Stochastic Games. In *In Proceedings of the National Conference on Artificial Intelligence*, volume 4, pages 709–715, 2004.

[7] M. Jain, D. Korzhyk, O. Vaněk, V. Conitzer, M. Pěchouček, and M. Tambe. A double oracle algorithm for zero-sum security games on graphs. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 327–334, 2011.

[8] D. Koller, N. Megiddo, and B. Von Stengel. Fast algorithms for finding randomized strategies in game trees. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 750–759. ACM, 1994.

[9] L. C. MacDermed and C. Isbell. Point based value iteration with optimal belief compression for Dec-POMDPs. In *Advances in Neural Information Processing Systems 26*, pages 100–108, 2013.

[10] J.-F. Mertens and S. Zamir. The value of two-person zero-sum repeated games with lack of information on both sides. *International Journal of Game Theory*, 1(1):39–64, 1971.

[11] R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella. Taming Decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *In Proceedings of the International Joint Conference on Artificial Intelligence*, pages 705–711, 2003.

[12] J. F. Nash. Equilibrium points in N-person games. *In Proceedings of the National Academy of Sciences of the United States of America*, 36:48–49, 1950.

[13] A. Nayyar, A. Gupta, C. Langbort, and T. Basar. Common information based markov perfect equilibria for stochastic games with asymmetric information: Finite games. *Automatic Control, IEEE Transactions on*, 59(3):555–570, 2014.

[14] A. Nayyar, A. Mahajan, and D. Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58:1644–1658, July 2013.

[15] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. Bayesian nash implementation. In *Algorithmic Game Theory*, volume 1. Cambridge University Press Cambridge, 2007.

[16] F. A. Oliehoek. Sufficient plan-time statistics for decentralized POMDPs. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 302–308, 2013.

[17] F. A. Oliehoek and C. Amato. Dec-POMDPs as non-observable MDPs. IAS technical report IAS-UVA-14-01, Intelligent Systems Lab, University of Amsterdam, Amsterdam, The Netherlands, Oct. 2014.

[18] F. A. Oliehoek, E. D. de Jong, and N. Vlassis. The parallel Nash memory for asymmetric games. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 337–344. ACM, 2006.

[19] F. A. Oliehoek, M. T. Spaan, and N. A. Vlassis. Optimal and approximate q-value functions for Decentralized POMDPs. *Journal of AI Research*, 32:289–353, 2008.

[20] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*, chapter Nash Equilibrium, pages 21 – 27. The MIT Press, July 1994.

[21] J. Pineau, G. J. Gordon, and S. Thrun. Anytime point-based approximations for large POMDPs. *Journal of AI Research*, 27:335–380, 2006.

[22] J.-P. Ponssard. Zero-sum games with "almost" perfect information. *Management Science*, 21(7):794–805, 1975.

[23] J.-P. Ponssard and S. Sorin. Some results on zero-sum games with incomplete information: The dependent case. *International Journal of Game Theory*, 9(4):233–245, 1980.

[24] J.-P. Ponssard and S. Zamir. Zero-sum sequential games with incomplete information. *International Journal of Game Theory*, 2(1):99–107, 1973.

[25] D. M. Roijers, S. Whiteson, and F. Oliehoek. Computing convex coverage sets for faster multi-objective coordination. *Journal of Artificial Intelligence Research*, 52:399–443, 2015.

[26] J. Rubin and I. Watson. Computer poker: A review. *Artificial Intelligence*, 175(5):958–987, 2011.

[27] S. Sorin. Stochastic Games with incomplete information. In A. Neyman and S. Sorin, editors, *Stochastic Games and applications*, volume 570. Springer, 2003.

[28] M. T. Spaan and N. Vlassis. Perseus: Randomized point-based value iteration for POMDPs. *Journal of AI Research*, 24:195–220, 2005.