Structure in the Value Function of Two-Player Zero-Sum Games of Incomplete Information

Wiggers, A.J.; Oliehoek, F.A.; Roijers, D.M.

Publication date
2016

Document Version
Submitted manuscript

Structure in the Value Function of Two-Player Zero-Sum Games of Incomplete Information

Auke J. Wiggers[1] and **Frans A. Oliehoek**[2] and **Diederik M. Roijers**[3]

Abstract. Zero-sum stochastic games provide a rich model for competitive decision making. However, under general forms of state uncertainty as considered in the Partially Observable Stochastic Game (POSG), such decision making problems are still not very well understood. This paper makes a contribution to the theory of zero-sum POSGs by characterizing structure in their value function. In particular, we introduce a new formulation of the value function for zs-POSGs as a function of the 'plan-time sufficient statistics' (roughly speaking the information distribution in the POSG), which has the potential to enable generalization over such information distributions. We further delineate this generalization capability by proving a structural result on the shape of value function: it exhibits concavity and convexity with respect to appropriately chosen marginals of the statistic space. This result is a key pre-cursor for developing solution methods that may be able to exploit such structure. Finally, we show how these results allow us to reduce a zs-POSG to a 'centralized' model with shared observations, thereby transferring results for the latter, narrower class, to games with individual (private) observations.

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 [24] and security settings [8]. In strictly competitive sequential games in which the environment can be influenced through actions, the problem of behaving rationally can be modeled as a *zero-sum Partially Observable Stochastic Game (zs-POSG)*.

Reasoning about zs-POSGs poses a challenge for strategic agents: they need to reason about their own uncertainty regarding the state of the environment as well as uncertainty regarding the opposing agent. As this opponent is trying to minimize the reward that they are maximizing, behaving strategically typically requires stochastic strategies. A factor that further complicates the reasoning is that agents not only influence their immediate rewards, but also the *future state of the environment* and *both agents' future observations*.

In this paper, we prove the existence of structural 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 [15], 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 [3, 10, 14]. This enabled tackling these problems using solution methods for POMDPs, leading to increases in scalability [3].

This paper provides a theoretical basis for enabling similar advancements for zs-POSGs. In particular, we extend results for Dec-POMDPs to the zs-POSG setting by presenting three contributions:

1. Two novel formulations of the value functions in POSGs, one based on past joint policies, and one based on distributions of information in the game called *plan-time sufficient statistics*.
2. A proof that the latter formulation allows for a generalization over the statistics: on every stage, the value function exhibits concavity and convexity in different *subspaces* of statistic space.
3. A reduction of the zs-POSG to a *Non-Observable Stochastic Game*, which in turn allows us to shows that certain properties previously proven for narrower classes of games generalize to the more general zs-POSG considered here.

This is the first work that gives insight in how the value function of a zs-POSG generalizes over the space of sufficient statistics. We argue that this result may open up the route for new solution methods.

2 Background

In this section we provide the necessary background to explain our contributions. We defer a treatment of related work to Section 6, where we can more concisely point out the differences to our work.

This paper focuses on zero-sum games of incomplete information where the number of states, actions, observations and the horizon are finite. We examine games where the hidden state is static, and games with dynamic state (i.e., it changes over time). We assume *perfect recall*, i.e., agents recall their own past actions and observations, and assume that all elements of the game are *common knowledge* among the agents [19, Chapter 5].

2.1 Zero-Sum One-Shot Games

We start by describing one-shot (static) games.

Definition 1. A **normal form game** (NFG) is a tuple $\mathcal{N} = \langle I, \mathcal{A}, \mathcal{R} \rangle$:

- $I = \{1, 2\}$ is the set of 2 agents,
- $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ is the set of joint actions $a = \langle a_1, a_2 \rangle$, also called (joint) strategies,
- $\mathcal{R} = \{R_1, R_2\}$ is the set of reward (or 'payoff') functions for the agents: $R_i : \mathcal{A} \to \mathbb{R}$ is the reward function for agent i,

[1] Scyfer B.V., email: auke@scyfer.nl
[2] University of Liverpool, Universiteit van Amsterdam, email: Frans.Oliehoek@liverpool.ac.uk
[3] University of Oxford, email: diederik.roijers@cs.ox.ac.uk

In the case of a *zero-sum* NFG (zs-NFG), we have that $R_1(a) = -R_2(a), \forall a$. In a zs-NFG, we will define the following quantities (and the associated strategies):

Definition 2. The **maxmin value** (for agent 1) of a 2 player zero-sum game is defined as $V_{maxmin}(\mathcal{N}) = \max_{a_1} \min_{a_2} R_1(a_1, a_2)$.

Definition 3. The **minmax value** (for agent 1) of a 2 player zero-sum game is defined as $V_{minmax}(\mathcal{N}) = \min_{a_2} \max_{a_1} R_1(a_1, a_2)$.

The min-max theorem states that $\forall \mathcal{N} \quad V_{maxmin}(\mathcal{N}) \leq V_{minmax}(\mathcal{N})$ [2]. In case of equality, we say that $V(\mathcal{N}) = V_{maxmin}(\mathcal{N}) = V_{minmax}(\mathcal{N})$ is the **value** of the game.

A *Nash Equilibrium* (NE) is a joint strategy from which no agent has an incentive to unilaterally deviate. In zs-NFGs, if an NE exists, then it coincides with the value of the game. That is, the value of a game is the value attained when both agents follow the strategy specified by the NE. Moreover any NEs, also called saddle points in this context, correspond to maxmin-strategies for the players [19, Proposition 22.2]. We will refer to such strategies as *rational* strategies.

The value is not guaranteed to exist in all zs-NFGs. If the action sets are convex and compact, if $\forall a_2,$ the mapping $a_1 \rightarrow R_1(a_1, a_2)$ is a concave function, and if $\forall a_1, \ a_2 \rightarrow R_1(a_1, a_2)$ is a convex function, then the value exists and the zs-NFG has a Nash equilibrium (i.e., a saddle point) [1, p. 134], [19, Proposition 20.3]. In games where actions are discrete, an NE (and corresponding value) may not exist. However, when mixed strategies are allowed, we can convert to a continuous action NFG: $\langle I, M_1 \times M_2, \mathcal{R}' \rangle$, where $\mu_i \in M_i$ specifies a probability distribution over actions, and where $R'_i(\mu_1, \mu_2) = \sum_{a_1} \mu_1(a_1) \sum_{a_2} \mu_2(a_2) R_i(a_1, a_2)$. Furthermore, it can now be shown that the utility function $R'(\mu_1, \mu_2)$ in terms of mixed strategies is concave in μ_1 for each μ_2, and convex in μ_2 for each μ_1, such that the aforementioned assumptions needed for the existence of a saddle-point hold (i.e., an NE exists [2, p.239]).

2.2 Zero-Sum Bayesian Games

Here we consider *zero-sum Bayesian Games* (zs-BGs), in which agents simultaneously select an action based on an individual observation (often referred to as their *type*).

Definition 4. A **zs-BG** is defined as a tuple $\mathcal{B} = \langle I, \Theta, \mathcal{A}, R, \sigma \rangle$:

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

In this paper we treat finite zs-BGs, where the sets of actions and types are finite. A pure strategy, to which we refer as a *pure decision rule*, is a mapping from types to actions. A *stochastic decision rule* $\delta_i \in \Delta_i^S$ is a mapping from types to probability distributions over the set of actions, denoted as $\delta_i(a_i|\theta_i)$, Δ_i^S is the space of such mappings. Given a joint decision rule $\delta = \langle \delta_1, \delta_2 \rangle$, the value is:

$$Q_{\text{BG}}(\mathcal{B}, \delta) \triangleq \sum_{\theta} \sigma(\theta) \sum_{a} \delta(a|\theta) R(\theta, a). \tag{1}$$

where $\delta(a|\theta) \triangleq \delta_1(a_1|\theta_1)\delta_2(a_2|\theta_2)$. The case of pure decision rules is covered by treating them as degenerate stochastic policies.

There are two ways to reduce a zs-BG to a zs-NFG. First, we can simply reinterpret $Q_{\text{BG}}(\mathcal{B}, \delta)$ as a payoff function $R(\delta_1, \delta_2)$, such that the BG corresponds to an NFG with continuous action sets Δ_i^S for each player i. The conditions for the existence of the value for such a game can be shown to hold, so if we suppose that $\langle \delta_1^*, \delta_2^* \rangle$ is an NE, then the value of the game can be defined as the maxmin (=minmax) value:

$$V_{\text{BG}}(\mathcal{B}) \triangleq Q_{\text{BG}}(\mathcal{B}, \langle \delta_1^*, \delta_2^* \rangle) =$$
$$\max_{\delta_1 \in \Delta_1^S} \min_{\delta_2 \in \Delta_2^S} Q_{\text{BG}}(\mathcal{B}, \langle \delta_1, \delta_2 \rangle) = \min_{\delta_2 \in \Delta_2^S} \max_{\delta_1 \in \Delta_1^S} Q_{\text{BG}}(\mathcal{B}, \langle \delta_1, \delta_2 \rangle). \tag{2}$$

Alternatively, one can reinterpret the BG as an NFG with finite action sets corresponding to the pure decision rules. This then leads to a similar $R'(\mu_1, \mu_2)$ formulation where the mixed strategies μ_i now are distributions over pure decision rules. Again this formulation will satisfy the required assumptions on concavity/convexity, such that this reinterpretation leads to the same logical conclusion that the value of the finite zs-BG exists (these dual perspectives are possible due to the one-one correspondence between stochastic decision rules and mixed strategies).

These reductions imply that solution methods for zs-NFGs (e.g., via linear programming [25]) can be used to solve zs-BGs. However, such an approach does not scale well — a more efficient solution method is to convert the zs-BG to *sequence form* [9].

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 changes over time, and the agents simultaneously choose actions at every stage.

Definition 5. A **finite zs-POSG** is defined as a tuple $\mathcal{P} = \langle h, I, \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, b^0 \rangle$:

- h is the horizon,
- $I = \{1, 2\}$ is the set of 2 agents,
- $\mathcal{S}$ is the finite set of states s,
- $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ is the finite set of joint actions $a = \langle a_1, a_2 \rangle$,
- $\mathcal{O} = \mathcal{O}_1 \times \mathcal{O}_2$ is the finite set of joint observations $o = \langle o_1, o_2 \rangle$,
- T is the transition function $\Pr(s^{t+1}|s^t, a^t)$,
- O is the observation function $\Pr(o^{t+1}|s^{t+1}, a^t)$,
- $R : S \times \mathcal{A} \times S \rightarrow \mathbb{R}$ is the reward function for agent 1,
- $b^0 \in \Delta(\mathcal{S})$ is the probability distribution over states.

In the zs-POSG, we aim to find maxmin-strategies and corresponding value. Let a *pure policy* for agent i be a mapping from individual action-observation histories (AOHs) $\vec{\theta}_i^t = \langle a_i^0, o_i^1, \ldots, a_i^{t-1}, o_i^t \rangle$ 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 $\pi_i = \langle \delta_i^0 \ldots \delta_i^{h-1} \rangle$. We define the *past individual policy* as a tuple of decision rules $\varphi_i^t = \langle \delta_i^0, \ldots, \delta_i^{t-1} \rangle$, and define the tuple containing decision rules from stage t to h as the *partial individual policy* $\pi_i^t = \langle \delta_i^t, \ldots, \delta_i^{h-1} \rangle$.

As in zs-BGs, it is theoretically possible to convert a zs-POSG to a zs-NFG and solve using standard methods, but this is infeasible in practice. An alternative is to converting the zs-POSG to an extensive form game (EFG) and solve it in sequence form [9]. While this is more efficient than the NFG route, it is still intractable: the resulting EFG is huge since its size depends on the number of full histories (trajectories of joint actions, joint observations, and states) [18].

With the goal of opening paths to completely new approaches of tackling zs-POSGs, this paper focuses on the description of the **value function** of zs-POSGs, which (in contrast to the *value* of a game as defined above) is a function mapping from some notion of 'state' of a game to the expected value (for agent 1). Similar to how the 'value' is defined by rational strategies, we will use the term 'value function' for a function that captures the future expected rewards under a rational joint policy. However, since one can also reason about the value of non-rational policies, we will refer to the 'rational value function' if there is a need to clarify.

3 Structure in One-Shot Value

In order to provide a value function definition for the sequential setting (in Section 4), we will rely on an intermediate result for one-shot games developed in this section. In particular, we introduce the concept of a Family of zero-sum Bayesian Games, for which we will introduce the joint type distribution as a suitable notion of 'state' and prove that its value function exhibits certain concave/convex properties with respect to this notion.

3.1 Families of Bayesian Games

Here we consider the notion of a *family* of (zs)-BGs. Intuitively, different stages t of a POSG are similar to a BG: each agent has a privately observed history, which corresponds to its type. However, the probabilities of these histories might depend on how the game was played in earlier stages. As such, we will need to reason about families of BGs.

Definition 6. A **Family of Bayesian Games**, $\mathcal{F} = \langle I, \Theta, \mathcal{A}, R \rangle$, is the set of Bayesian Games of the form $\langle I, \Theta, \mathcal{A}, R, \sigma \rangle$ for which $I, \Theta, \mathcal{A}$ and R are identical.

Let $\mathcal{F}$ be a Family of *zero-sum* Bayesian Games. By providing a joint type distribution, $\mathcal{F}(\sigma)$ indicates a particular zs-BG. We generalize (1) and (2) as follows:

$$Q_{\mathcal{F}}(\sigma, \delta) \triangleq Q_{\text{BG}}(\mathcal{F}(\sigma), \delta), \tag{3}$$

$$V^*_{\mathcal{F}}(\sigma) \triangleq V_{\text{BG}}(\mathcal{F}(\sigma)). \tag{4}$$

As each $\mathcal{F}(\sigma)$ is a zs-BG, the rational value function $V^*_{\mathcal{F}}(\sigma)$ can be written as:

$$V^*_{\mathcal{F}}(\sigma) = \max_{\delta_1 \in \Delta_1^S} \min_{\delta_2 \in \Delta_2^S} Q_{\mathcal{F}}(\sigma, \langle \delta_1, \delta_2 \rangle). \tag{5}$$

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

$$V^{\text{BR1}}_{\mathcal{F}}(\sigma, \delta_2) \triangleq \max_{\delta_1 \in \Delta_1^S} Q_{\mathcal{F}}(\sigma, \langle \delta_1, \delta_2 \rangle), \tag{6}$$

$$V^{\text{BR2}}_{\mathcal{F}}(\sigma, \delta_1) \triangleq \min_{\delta_2 \in \Delta_2^S} Q_{\mathcal{F}}(\sigma, \langle \delta_1, \delta_2 \rangle). \tag{7}$$

We remind the reader that, for each σ, $Q_{\mathcal{F}}(\sigma, \langle \delta_1, \delta_2 \rangle)$ exhibits concavity and convexity in the space of *decision rules*, as discussed in Section 2.2.

3.2 Concavity/Convexity of the Value Function

We saw in the previous section that the concave/convex shape of the utility function expressed in the space of decision rules follows directly from known results. Here we provide a novel result: $V^*_{\mathcal{F}}$ exhibits a similar concave/convex shape *in terms of type distributions*. This is important, as it provides insight on how the *distribution of information* (in addition to the distribution of actions) affects the value of the game, which is key to enabling generalization of the value in different parts of sequential games. To provide this formulation, we decompose σ into a *marginal* term $\sigma_{m,i}$ and a *conditional* term $\sigma_{c,i}$, here shown for $i = 1$:

$$\sigma_{m,1}(\theta_1) \triangleq \sum_{\theta_2 \in \Theta_2} \sigma(\theta_1 \theta_2), \qquad \sigma_{c,1}(\theta_2 | \theta_1) \triangleq \frac{\sigma(\theta_1 \theta_2)}{\sigma_{m,1}(\theta_1)}. \tag{8}$$

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 write $\sigma^t = \sigma^t_{m,i}\sigma^t_{c,i}$ for short.

We will first show that the best-response value functions defined in (6) and (7) 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:

$$\vec{r}_{[\sigma_{c,1}, \delta_2]}(\theta_1) \triangleq \max_{a_1 \in \mathcal{A}_1} \left[\sum_{\theta_2} \sigma_{c,1}(\theta_2 | \theta_1) \sum_{a_2} \delta_2(a_2 | \theta_2) R(\theta, a) \right]. \tag{9}$$

The vector $\vec{r}_{[\sigma_{c,2}, \delta_1]}$ is defined analogously.

Now we can state an important lemma that shows that the best-response value functions are linear functions in the marginal space.

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:*

1. $V^{BR1}_{\mathcal{F}}(\sigma_{m,1}\sigma_{c,1}, \delta_2) = \vec{\sigma}_{m,1} \cdot \vec{r}_{[\sigma_{c,1}, \delta_2]}$, (10)
2. $V^{BR2}_{\mathcal{F}}(\sigma_{m,2}\sigma_{c,2}, \delta_1) = \vec{\sigma}_{m,2} \cdot \vec{r}_{[\sigma_{c,2}, \delta_1]}$. (11)

Proof. The proof is listed in Appendix A. □

This lemma can now be used to show that the rational value function $V^*_{\mathcal{F}}$ possess concave and convex structure with respect to (appropriately chosen marginals of) the space of joint type distributions.

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. $V^*_{\mathcal{F}}(\sigma_{m,1}\sigma_{c,1}) = \min_{\delta_2 \in \Delta_2^S} \left[\vec{\sigma}_{m,1} \cdot \vec{r}_{[\sigma_{c,1}, \delta_2]} \right]$,
2. $V^*_{\mathcal{F}}(\sigma_{m,2}\sigma_{c,2}) = \max_{\delta_1 \in \Delta_1^S} \left[\vec{\sigma}_{m,2} \cdot \vec{r}_{[\sigma_{c,2}, \delta_1]} \right]$.

Proof. Filling in the result of Lemma 1 gives:

$$V^*_{\mathcal{F}}(\sigma_{m,1}\sigma_{c,1}) \overset{\{5,6\}}{=} \min_{\delta_2 \in \Delta_2^S} V^{\text{BR1}}_{\mathcal{F}}(\sigma_{m,1}\sigma_{c,1}, \delta_2) \overset{\{10\}}{=} \min_{\delta_2 \in \Delta_2^S} \left[\vec{\sigma}_{m,1} \cdot \vec{r}_{[\sigma_{c,1}, \delta_2]} \right].$$

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

The importance of this theorem is that it gives direct ways to approximately generalize the value function using piecewise-linear and convex (concave) functions.

4 Structure in zs-POSG Value

Given the result for one-shot games established in the previous section, we are now in the position to present our main contributions: novel formulations for the value function, and generalization of the structural result of Theorem 1 to the sequential setting. First, we will introduce a description of the rational value function based on past joint policies.

4.1 Past Joint Policies

To make rational decisions at stage t in the zs-POSG, it is sufficient to know the *past decisions*, which are captured in the past joint policy φ^t. To show this, the zs-POSG value function in terms of the past joint policy φ^t can be defined in terms of φ^t by extending the formulation by Oliehoek [15].

We define the value function of the zs-POSG at a stage t in terms of a *past joint policy* φ^t. The definition that follows gives the value attained when all agents follow the joint decision rule δ^t, assuming that in future stages agents will act rationally. That is, the agents follow a rational joint future policy $\pi_i^{t+1*} = \langle \delta^{t+1*} \dots \delta^{h-1*} \rangle$.

We first 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 immediate reward for a joint AOH and a joint decision rule be defined as:

$$R(\vec{\theta}^t, \delta^t) \triangleq \sum_{a^t} \delta^t(a^t|\vec{\theta}^t) \sum_{s^t} \Pr(s^t|\vec{\theta}^t, b^0) R(s^t, a^t).$$

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

$$Q^*_{h-1}(\varphi^{h-1}, \vec{\theta}^{h-1}, \delta^{h-1}) \triangleq R(\vec{\theta}^{h-1}, \delta^{h-1}). \tag{12}$$

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

$$\Pr(\vec{\theta}^{t+1}|\vec{\theta}^t, \delta^t) = \Pr(\langle \vec{\theta}^t, a^t, o^{t+1} \rangle | \vec{\theta}^t, \delta^t) = \Pr(o^{t+1}|\vec{\theta}^t, a^t) \delta^t(a^t|\vec{\theta}^t).$$

For all stages except the final stage $t = h-1$, the value at future stages is propagated to the current stage as follows:

$$Q^*_t(\varphi^t, \vec{\theta}^t, \delta^t) \triangleq R(\vec{\theta}^t, \delta^t) + \sum_{a^t \in \mathcal{A}} \sum_{o^{t+1}} \Pr(\vec{\theta}^{t+1}|\vec{\theta}^t, \delta^t) \; Q^*_{t+1}(\varphi^{t+1}, \vec{\theta}^{t+1}, \delta^{t+1*}). \tag{13}$$

$$Q^*_t(\varphi^t, \delta^t) \triangleq \sum_{\vec{\theta}^t \in \vec{\Theta}^t} \Pr(\vec{\theta}^t|b^0, \varphi^t) Q^*_t(\varphi^t, \vec{\theta}^t, \delta^t). \tag{14}$$

We use (14) to find rational decision rules for both agents. Consistent with (13), we show how to find $\delta^{t+1*} = \langle \delta_1^{t+1*}, \delta_2^{t+1*} \rangle$:

$$\delta_1^{t+1*} = \operatorname*{argmax}_{\delta_1^{t+1} \in \Delta_1^S} \min_{\delta_2^{t+1} \in \Delta_2^S} Q^*_{t+1}(\varphi^{t+1}, \langle \delta_1^{t+1}, \delta_2^{t+1} \rangle), \tag{15}$$

$$\delta_2^{t+1*} = \operatorname*{argmin}_{\delta_2^{t+1} \in \Delta_2^S} \max_{\delta_1^{t+1} \in \Delta_1^S} Q^*_{t+1}(\varphi^{t+1}, \langle \delta_1^{t+1}, \delta_2^{t+1} \rangle). \tag{16}$$

Using (12), (15) and (16), 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 (13) propagates the value to the preceding stages. As such, rationality for all stages follows by induction. The value function can now be defined in terms of the past joint policy as:

$$V^*_t(\varphi^t) = \max_{\delta_1^t \in \Delta_1^S} \min_{\delta_2^t \in \Delta_2^S} Q^*_t(\varphi^t, \langle \delta_1^t, \delta_2^t \rangle). \tag{17}$$

By (12), (15) and (16), 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. As such, (17) can not be used to form a backward inductive approach directly.

4.2 Plan-Time Sufficient Statistics

A disadvantage of the value function definition from the previous section is that the value at a stage t depends on all the *joint* decisions from stage 0 to t. We propose to define the value function in terms of a plan-time sufficient statistic that summarizes many past joint policies. Furthermore, we give formal proof that this value function exhibits a concave/convex shape in statistic-space that may be exploitable.

Definition 7. 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, a plan-time sufficient statistic fully captures the influence of the past joint policy. We prove that this also holds for the zs-POSG case, thereby validating the previous definition and the name 'sufficient statistic'. 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. First, we define the update rule for plan-time sufficient statistics:

$$\sigma^{t+1}(\vec{\theta}^{t+1}) \triangleq \Pr(o^{t+1}|\vec{\theta}^t, a^t) \delta^t(a^t|\vec{\theta}^t) \sigma^t(\vec{\theta}^t). \tag{18}$$

The Q-value function at stage $h-1$ reduces to the immediate reward:

$$Q^*_{h-1}(\sigma^{h-1}, \vec{\theta}^{h-1}, \delta^{h-1}) \triangleq R(\vec{\theta}^{h-1}, \delta^{h-1}). \tag{19}$$

We then define the Q-value for all other stages similar to (13), (14):

$$Q^*_t(\sigma^t, \vec{\theta}^t, \delta^t) \triangleq R(\vec{\theta}^t, \delta^t) + \sum_{a^t} \sum_{o^{t+1}} \Pr(\vec{\theta}^{t+1}|\vec{\theta}^t, \delta^t) \; Q^*_{t+1}(\sigma^{t+1}, \vec{\theta}^{t+1}, \delta^{t+1*}). \tag{20}$$

$$Q^*_t(\sigma^t, \delta^t) \triangleq \sum_{\vec{\theta}^t} \sigma^t(\vec{\theta}^t) Q^*_t(\sigma^t, \vec{\theta}^t, \delta^t). \tag{21}$$

Rational decision rules can be found using (21):

$$\delta_1^{t+1*} = \operatorname*{argmax}_{\delta_1^{t+1} \in \Delta_1^S} \min_{\delta_2^{t+1} \in \Delta_2^S} Q^*_{t+1}(\sigma^{t+1}, \langle \delta_1^{t+1}, \delta_2^{t+1} \rangle), \tag{22}$$

$$\delta_2^{t+1*} = \operatorname*{argmin}_{\delta_2^{t+1} \in \Delta_2^S} \max_{\delta_1^{t+1} \in \Delta_1^S} Q^*_{t+1}(\sigma^{t+1}, \langle \delta_1^{t+1}, \delta_2^{t+1} \rangle). \tag{23}$$

We formally prove that the statistic σ^t provides sufficient information for rational decision-making in the zs-POSG.

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

Proof. The proof is listed in Appendix A. □

This allows us to define the value function of the zs-POSG in terms of the sufficient statistic as follows:

$$V_t^*(\sigma^t) \triangleq \max_{\delta_1^t \in \Delta_1^S} \min_{\delta_2^t \in \Delta_2^S} Q_t^*(\sigma^t, \langle \delta_1^t, \delta_2^t \rangle). \tag{24}$$

Although we have now identified the value at a single stage of the game, implementing a backwards inductive approach directly is still not possible, since the space of statistics is continuous and we do not know how to represent $V_t^*(\sigma^t)$. This paper takes a first step at resolving this problem by investigating the structure of $V_t^*(\sigma^t)$.

4.3 Equivalence Final Stage Zero-Sum POSG and Family of Zero-Sum Bayesian Games

We have already proven that the value function of a Family of zero-sum Bayesian Games exhibits concavity and convexity in terms of the marginal parts of the type distribution σ for respectively agent 1 and 2. We show that the final stage of the zs-POSG $\mathcal{P}$, $t = h-1$, can be defined as a Family of zs-BGs as follows:

- $I = \{1, 2\}$ is the set of agents,
- $\Theta = \vec{\Theta}^{h-1}$ is the set of joint types corresponding to AOHs in zs-POSG $\mathcal{P}$ at stage $h-1$,
- $\mathcal{A}$ is the set of joint actions in zs-POSG $\mathcal{P}$,
- $R(\vec{\theta}^{h-1}, a)$ follows directly from the immediate reward function of $\mathcal{P}$ (as described in Section 4.1).

By the result of Theorem 1 (i.e., that $V_{\mathcal{F}}^*$ exhibits concavity and convexity in marginal-spaces of agent 1 and 2 respectively) the value function at the final stage of the zs-POSG, V_{h-1}^*, is concave in $\Delta(\vec{\Theta}_1^{h-1})$ for all $\sigma_{c,1}^{h-1}$, and convex in $\Delta(\vec{\Theta}_2^{h-1})$ for all $\sigma_{c,2}^{h-1}$.

Note that, 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 [4] and derivative works [17]. In fact, all other stages (0 to $h-2$) cannot be represented as a Family of BGs, as the rational value function for stages $t = 0, ..., h-2$ cannot be expressed as a function $R(\vec{\theta}^t, a)$.

4.4 Concavity/Convexity of the Value Function

We continue to show that the value function for any stage exhibits the same type of structure. In particular, the plan-time sufficient statistic can be decomposed in marginals and conditionals, and the value function is concave in marginal-space for agent 1, $\Delta(\vec{\Theta}_1^t)$, and convex in marginal-space for agent 2, $\Delta(\vec{\Theta}_2^t)$. Figure 1 provides intuition on how the best-response value functions relate to the concave/convex value function: a 'slice' in statistic-space corresponds to a single conditional $\sigma_{c,1}^t$ ($\sigma_{c,2}^t$) and exhibits concave (convex) shape of the value function made up by linear segments that each correspond to a partial policy of the other agent. As we will show, each segment corresponds exactly to a best-response value function.

Best-response value functions in terms of σ^t and π_i^t are defined as V_t^{BR1} and V_t^{BR2}, similar to (6) and (7). Let Π_i^t be the space of all stochastic partial policies π_i^t. We then have:

$$V_t^*(\sigma^t) = \min_{\pi_2^t \in \Pi_2^t} V_t^{\text{BR1}}(\sigma^t, \pi_2^t) = \max_{\pi_1^t \in \Pi_1^t} V_t^{\text{BR2}}(\sigma^t, \pi_1^t). \tag{25}$$

We first show that the best-response value functions V_t^{BR1} and V_t^{BR2} are linear in their respective marginal-spaces. Let us define a vector that contains the value (immediate reward *and* future value)

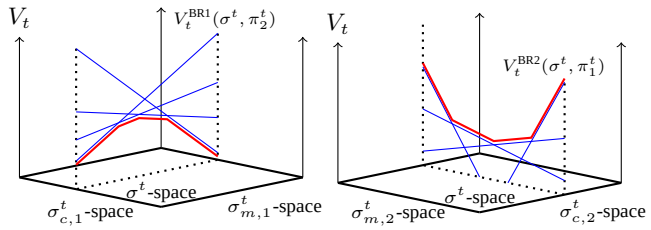

Figure 1: An abstract visualization of the decomposition of statistic-space into marginal-space and conditional-space.

for agent 1 for each individual AOH $\vec{\theta}_1^t$, given that agent 2 follows the partial policy π_2^t:

$$\vec{\nu}_{[\sigma_{c,1}^t, \pi_2^t]}(\vec{\theta}_1^t) \triangleq \max_{a_1^t \in \mathcal{A}_1} \Bigg[\sum_{\vec{\theta}_2^t \in \vec{\Theta}_2^t} \sigma_{c,1}^t(\vec{\theta}_2^t | \vec{\theta}_1^t) \sum_{a_2^t \in \mathcal{A}_2} \delta_2^t(a_2^t | \vec{\theta}_2^t) \Big(R(\vec{\theta}^t, a^t) + \sum_{o^{t+1} \in \mathcal{O}_1} \Pr(o^{t+1} | \vec{\theta}^t, a^t) \vec{\nu}_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}(\vec{\theta}_1^{t+1}) \Big) \Bigg] \tag{26}$$

Note that this is a recursive definition, and that $\vec{\theta}_1^{t+1} = \langle \vec{\theta}_1^t, a_1^t, o_1^{t+1} \rangle$. The vector $\vec{\nu}_{[\sigma_{c,2}^t, \pi_1^t]}$ is defined analogously.

Lemma 3. *(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 $t = 0, \ldots, h-1$:*

1. $V_t^{BR1}(\sigma_{m,1}^t \sigma_{c,1}^t, \pi_2^t) = \vec{\sigma}_{m,1}^t \cdot \vec{\nu}_{[\sigma_{c,1}^t, \pi_2^t]}$, (27)
2. $V_t^{BR2}(\sigma_{m,2}^t \sigma_{c,2}^t, \pi_1^t) = \vec{\sigma}_{m,2}^t \cdot \vec{\nu}_{[\sigma_{c,2}^t, \pi_1^t]}$. (28)

Proof. We prove this by induction. We know the zs-POSG value function at stage $t = h-1$ to be equivalent to that of a Family of zs-BGs, which is concave/convex (Lemma 1). This 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. $V_t^*(\sigma_{m,1}^t \sigma_{c,1}^t) = \min_{\pi_2^t \in \Pi_2^t} \left[\vec{\sigma}_{m,1}^t \cdot \vec{\nu}_{[\sigma_{c,1}^t, \pi_2^t]} \right]$,
2. $V_t^*(\sigma_{m,2}^t \sigma_{c,2}^t) = \max_{\pi_1^t \in \Pi_1^t} \left[\vec{\sigma}_{m,2}^t \cdot \vec{\nu}_{[\sigma_{c,2}^t, \pi_1^t]} \right]$.

Proof. Filling in the result of Lemma 3 gives:

$$V_t^*(\sigma_{m,1}^t \sigma_{c,1}^t) \overset{\{25\}}{=} \min_{\pi_2^t \in \Pi_2^t} V_t^{\text{BR1}}(\sigma_{m,1}^t \sigma_{c,1}^t, \pi_2^t) \overset{\{27\}}{=} \min_{\pi_2^t \in \Pi_2^t} \left[\vec{\sigma}_{m,1}^t \cdot \vec{\nu}_{[\sigma_{c,1}^t, \pi_2^t]} \right].$$

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

The importance of this theorem is that it suggests ways to (approximately) represent $V_t^*(\sigma^t)$. Thus, it may enable the development of new solution methods for zs-POSGs. To draw the parallel, many POMDP solution methods exploit the fact that a POMDP value function is piecewise-linear and convex (PWLC) in belief-space [20, 27] (which is similar to the statistic-space we consider), and recently

such results have been extended to the decentralized cooperative (i.e., Dec-POMDP) case [3, 10].

Note that our result is similar to, but different from the saddle-point function that is typically associated with min-max equilibria. In particular, the value function in is defined in the space of plan-time sufficient statistics, while the well-known saddle point function (see Section 2.1) is a function of complete strategies. This means that the latter is only defined for the reduction to a normal form game, which per definition destroys any of the specific structure of the game under concern. In contrast, our formulation preserves such structure and thus allows us to make statements about how *value generalizes as a function of the information distribution*. That is, the formulation allows us to give approximation bounds that generalize within each conditional statistic.

5 Reduction to NOSG

The application of methods that exploit the PWLC structure of the value function of Dec-POMDPs was enabled by a reduction from Dec-POMDP to a *non-observable MDP (NOMDP)*, which is a special type of (centralized) POMDP [3, 10, 14, 16]. This allows POMDP solution methods to be employed in the context of Dec-POMDPs. The proposed plan-time statistics for Dec-POMDPs [15] precisely correspond to the belief in the centralized model.

Since we have shown that it is possible to generalize the plan-time statistics to the zs-POSG case, it is reasonable to expect that zs-POSGs can be reduced similarly. Here we present a reduction from zs-POSG to a special type of stochastic game where information is centralized, 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 [16]. The difference between the reduction from Dec-POMDP to NOMDP and the one we present next, is that the zs-POSG is reduced to a stochastic game where the joint AOH acts as the state.

Definition 8. A **plan-time Non-Observable Stochastic Game** for a zs-POSG is a tuple $\langle \dot{S}, \dot{\mathcal{A}}_1, \dot{\mathcal{A}}_1, \dot{\mathcal{O}}, \dot{T}, \dot{O}, \dot{R}, \dot{b}^0 \rangle$:

- $I = \{1, 2\}$ is the set of agents,
- $\dot{S}$ is the set of augmented states $\dot{s}^t$, each corresponding to a joint AOH $\vec{\theta}^t$,
- $\dot{\mathcal{A}} = \dot{\mathcal{A}}_1 \times \dot{\mathcal{A}}_2$ is the continuous action-space, containing stochastic decision rules $\delta^t = \langle \delta_1^t, \delta_2^t \rangle$,
- $\dot{\mathcal{O}} = \{\text{NULL}\}$ is set of joint observations that only contains the NULL observation.
- $\dot{T}$ is the transition function that specifies $\dot{T}(\dot{s}^{t+1} \mid \dot{s}^t, \langle \dot{a}_1^t, \dot{a}_2^t \rangle) = \Pr(\vec{\theta}^{t+1} \mid \vec{\theta}^t, \langle \delta_1^t, \delta_2^t \rangle)$.
- $\dot{O}$ is the observation function that specifies that observation NULL is received with probability 1.
- $\dot{\mathcal{R}} : \dot{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function $R(\vec{\theta}^t, \langle \delta_1^t, \delta_2^t \rangle)$,
- $\dot{b}^0 \in \Delta(\dot{S})$ is the initial belief over states.

In the NOSG model, agents condition their choices on the joint belief over 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 (24).

In order to avoid potential confusion, let us point out that a zs-POSG can also be converted to a best response POMDP by fixing the policies of one agent [12], 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 only joint AOHs. More importantly, however, our approach does not require fixing the policy of any agent, which would necessitate recomputing of all values when the fixed policy changes (if the past policy changes, the distributions over paths change, and thus future values are affected). Where the approach of Nair et al. [12] leads to a single-agent model that can be used to compute a best-response, our conversion leads to a multi-agent model that can be used to compute a Nash equilibrium directly.

A key contribution of our NOSG formulation is that it directly indicates that properties of 'zero-sum stochastic games with shared observations' [5] also hold for zs-POSGs.

Definition 9. A **zero-sum Shared Observation Stochastic Game (zs-SOSG)** is a zs-POSG (cf. Def. 5) in which each agent receives the joint observation and can observe the other's actions.

Ghosh et al. [5] show that, under some technical assumptions, a zs-SOSG can be converted into a completely observable model (similar to the conversion of a POMDP into a belief MDP), and that, in the infinite-horizon case, both the value and a rational joint policy exists.

Our claim is that these results in fact transfer to the more general class of zs-POSGs via our NOSG construction. We start by noting:

Lemma 4. *The plan-time NOSG of a finite-horizon zs-POSG is a zs-SOSG.*

Proof. In the finite-horizon case, the set of states in our plan-time NOSG is discrete. The action space, while continuous, is of finite dimensionality, and is a closed and bounded set. The set of shared observations only consists of a trivial NULL observation and by assuming rationality for both players, we can assume that they observe each others actions (corresponding to decision rules of the POSG). □

In the infinite-horizon case, some assumptions on the class of decision rules δ_i^t are needed to be able to formulate the plan-time NOSG model and assuming infinite policy trees as the policies would violate certain technical requirements on which Ghosh's results depend (e.g., the action-spaces are required to be metric and compact spaces). However, a straightforward extension of our reduction for the case where the agents use finite-state controllers (analogous to such formulations for Dec-POMDPs [10]), would satisfy these requirements, and as such we can infer the existence of a value for such games:

Corollary 1. *For infinite-horizon zs-POSGs where agents are restricted to use finite-state controllers for their policies, the value of the game exists.*

In that way, our reduction shows that some of the properties established by Ghosh et al. for a limited subset of zero-sum stochastic games, in fact extend to a much broader class of problems.

6 Related Work

There is rich body of literature on zero-sum games, and we provide pointers to the most relevant works here. The concave and convex structure we have found for the zs-POSG value function is similar to the saddle point structure associated with min-max equilibria [1]. Note, however, that we have defined the value function in terms of the distribution over information, rather than the substantially different space of *joint strategies*, solving of which requires flattening the game to normal form. That is, our results tell us something about the value of acting using the current information, thus they may give insight into games of general partial observability.

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. More importantly, 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. [7] present a dynamic-programming approach for finite-horizon (general sum) POSGs that iteratively constructs 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). A more generalized investigation of individual statistics in decentralized settings is given by Wu & Lall [28]. However, these notions are not a sufficient statistic for the past joint policy from a *designer perspective* (as is the proposed plan-time sufficient statistic in this paper). In fact, they are complementary and we hypothesize that they can be fruitfully combined in future work.

There are works from game theory literature that present structural results on the value function of so-called 'repeated zero-sum games with incomplete information' [11, 21, 22, 23]. These can best be understood as a class of two-player extensive form games that lie in between Bayesian games and POSGs: at the start of the game, nature determines the state (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 opponent. The models for which these results have been proven are therefore substantially less general than the zs-POSG model we consider.

For various flavors of such games, it has been shown that the value function has a concave/convex structure: cases with incomplete information on one side [23, 26], 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) [21] or dependent (general joint type distributions) [11, 22]. These results, however, crucially depend on the alternating actions and the static state and therefore do not extend to zs-POSGs.

A game-theoretic model that is closely related to the POSG model is the *Interactive POMDP* or *I-POMDP* [6]. In I-POMDPs, a (subjective) belief $b_i(s, \zeta_j)$ is constructed from the perspective of a single agent as a probability distribution over states and the *types*, ζ_j, of the other agent. A *level-k* I-POMDP agent i reasons about level-$(k-1)$ types ζ_j. Since each ζ_j fully determines the future policy of the other agent j, an I-POMDP can be interpreted as a best-response POMDP similar to the one introduced by Nair et al. [12] (discussed in Section 5), with the difference that the state also represents which policy the other agent uses. The differences mentioned in Section 5 also apply here; where an I-POMDP can be used to compute a best-response, our formulation is aimed at computing a Nash equilibrium.

7 Conclusions and Future Work

This paper presents 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 formally defined the value function for both types of games in terms of an information distribution called the sufficient plan-time statistic: a probability distribution over joint sets of private information (originally used in the collaborative setting [15]). Using the fact that this probability distribution can be decomposed into a marginal and a conditional term, we presented that in the zero-sum case value functions of both types of games exhibit concavity (convexity) in the space of marginal statistics of the maximizing (minimizing) agent, for every conditional statistic. In the multi-stage game, this structure of the value function is preserved on every stage. Thus, our formulation enables us to make statements about how value generalizes as a function of the information distribution. Lastly, we showed how the results allow us to reduce our zs-POSG to a model with shared observations, thereby transferring properties of this narrower class of games to the zs-POSG.

We hope that this result leads to solution methods that exploit the structure of the value function at every stage, as recently such developments have been made in the field of cooperative multi-agent problems [10]. In particular, we believe that heuristic methods that identify useful (conditional) statistics to explore, or point-based methods that iteratively select statistics to evaluate [10, 27] may be adapted for the zs-POSG case.

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

A Appendix

Proof of Lemma 1. We will prove item 1:

$$V^{\text{BR1}}_{\mathcal{F}}(\sigma_{m,1}\sigma_{c,1}, \delta_2) = \vec{\sigma}_{m,1} \cdot \vec{r}_{[\sigma_{c,1},\delta_2]}$$

The Q-value definition is expanded in order to bring the marginal term $\sigma_{m,1}$ to the front of the equation:

$$\begin{aligned} Q_{\mathcal{F}}(\sigma, \delta) &\overset{\{1,3\}}{=} \sum_{\theta \in \Theta} \sigma(\theta) \sum_{a \in \mathcal{A}} \delta(a|\theta) R(\theta, a) \\ &= \sum_{\theta_1 \in \Theta_1} \sigma_{m,1}(\theta_1) \sum_{\theta_2 \in \Theta_2} \sigma_{c,1}(\theta_2|\theta_1) \sum_{a_1 \in \mathcal{A}_1} \delta_1(a_1|\theta_1) \\ &\quad \sum_{a_2 \in \mathcal{A}_2} \delta_2(a_2|\theta_2) R(\langle \theta_1, \theta_2 \rangle, \langle a_1, a_2 \rangle). \end{aligned} \tag{A.29}$$

A maximization over stochastic decision rules conditioned on θ_1 is equivalent to choosing a maximizing action for each θ_1. Thus, we can rewrite the best-response value function as follows:

$$\begin{aligned} V^{\text{BR1}}_{\mathcal{F}}(\sigma, \delta_2) &= \max_{\delta_1 \in \Delta_1^S} Q^R_t(\sigma, \langle \delta_1, \delta_2 \rangle) \\ &\overset{\{A.29\}}{=} \max_{\delta_1 \in \Delta_1^S} \Bigg[\sum_{\theta_1 \in \Theta_1} \sigma_{m,1}(\theta_1) \sum_{\theta_2 \in \Theta_2} \sigma_{c,1}(\theta_2|\theta_1) \sum_{a_1 \in \mathcal{A}_1} \delta_1(a_1|\theta_1) \\ &\qquad \sum_{a_2 \in \mathcal{A}_2} \delta_2(a_2|\theta_2) R(\langle \theta_1, \theta_2 \rangle, \langle a_1, a_2 \rangle) \Bigg] \\ &= \sum_{\theta_1 \in \Theta_1} \sigma_{m,1}(\theta_1) \max_{a_1} \Bigg[\sum_{\theta_2 \in \Theta_2} \sigma_{c,1}(\theta_2|\theta_1) \sum_{a_2 \in \mathcal{A}_2} \delta_2(a_2|\theta_2) R(\theta, a) \Bigg] \\ &\overset{\{9\}}{=} \sum_{\theta_1 \in \Theta_1} \sigma_{m,1}(\theta_1) \vec{r}_{[\sigma_{c,1},\delta_2]}(\theta_1) \overset{\{\text{vec. not.}\}}{=} \vec{\sigma}_{m,1} \cdot \vec{r}_{[\sigma_{c,1},\delta_2]}. \end{aligned}$$

As it is possible to write $V^{\text{BR1}}_{\mathcal{F}}$ as an inner product of two vectors, $V^{\text{BR1}}_{\mathcal{F}}$ is linear in $\Delta(\Theta_1)$ for all $\sigma_{c,1}$ and δ_2. Analogously, $V^{\text{BR2}}_{\mathcal{F}}$ 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 [15]. 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.:

$$Q_{t+1}^*(\varphi^{t+1}, \vec{\theta}^{t+1}, \delta_{\text{pp}}^{t+1}) = Q_{t+1}^*(\sigma^{t+1}, \vec{\theta}^{t+1}, \delta_{\text{ss}}^{t+1}). \quad \text{(A.30)}$$

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

$$Q_t^*(\varphi^t, \vec{\theta}^t, \delta_{\text{pp}}^t) = Q_t^*(\sigma^t, \vec{\theta}^t, \delta_{\text{ss}}^t). \quad \text{(A.31)}$$

We substitute the induction hypothesis into (21):

$$\begin{aligned} Q_t^*(\varphi^t, \vec{\theta}^t, \delta_{\text{pp}}^t) &\overset{\{21\}}{=} R(\vec{\theta}^t, \delta_{\text{pp}}^t) + \sum_{a^t \in \mathcal{A}} \sum_{o^{t+1}} \Pr(\vec{\theta}^{t+1} | \vec{\theta}^t, \delta_{\text{pp}}^t) \\ &\quad Q_{t+1}^*(\varphi^{t+1}, \vec{\theta}^{t+1}, \delta_{\text{pp}}^{t+1*}) \\ &\overset{\{A.30\}}{=} R(\vec{\theta}^t, \delta_{\text{ss}}^t) + \sum_{a^t \in \mathcal{A}} \sum_{o^{t+1}} \Pr(\vec{\theta}^{t+1} | \vec{\theta}^t, \delta_{\text{ss}}^t) \\ &\quad Q_{t+1}^*(\sigma^{t+1}, \vec{\theta}^{t+1}, \delta_{\text{ss}}^{t+1*}) = Q_t^*(\sigma^t, \vec{\theta}^t, \delta_{\text{ss}}^t). \end{aligned}$$

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:

$$\begin{aligned} \delta_{1,pp}^{t+1*} &\overset{\{(22)\}}{=} \operatorname*{argmax}_{\delta_1^{t+1} \in \Delta_1^S} \min_{\delta_2^{t+1} \in \Delta_2^S} \Bigg[\sum_{\vec{\theta}^{t+1} \in \vec{\Theta}^{t+1}} \Pr(\vec{\theta}^{t+1} | b^0, \varphi^{t+1}) \\ &\quad Q_{t+1}^*(\varphi^{t+1}, \vec{\theta}^{t+1}, \langle \delta_1^{t+1}, \delta_2^{t+1} \rangle) \Bigg] \\ &\overset{\{(7)\}}{=} \operatorname*{argmax}_{\delta_1^{t+1} \in \Delta_1^S} \min_{\delta_2^{t+1} \in \Delta_2^S} \Bigg[\sum_{\vec{\theta}^{t+1} \in \vec{\Theta}^{t+1}} \sigma^{t+1}(\vec{\theta}^{t+1}) \\ &\quad Q_{t+1}^*(\sigma^{t+1}, \vec{\theta}^{t+1}, \langle \delta_1^{t+1}, \delta_2^{t+1} \rangle) \Bigg] \overset{\{22\}}{=} \delta_{1,ss}^{t+1*}. \end{aligned}$$

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

Proof of Lemma 3. By the results of Lemma 1 and the results from Section 4.3, we know the best-response 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:

$$V_{t+1}^{\text{BR1}}(\sigma_{m,1}^{t+1} \sigma_{c,1}^{t+1}, \pi_2^{t+1}) = \sigma_{m,1}^{t+1} \cdot \vec{\nu}_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}. \quad \text{(A.32)}$$

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

$$V_t^{\text{BR1}}(\sigma_{m,1}^t \sigma_{c,1}^t, \pi_2^t) = \vec{\sigma}_{m,1}^t \cdot \vec{\nu}_{[\sigma_{c,1}^t, \pi_2^t]}. \quad \text{(A.33)}$$

Let Q_t^R be a function similar to (3) that defines reward for a given statistic and joint decision rule. 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 consider statistics corresponding to conditional $\sigma_{c,i}^t$.

$$\begin{aligned} V_t^{\text{BR1}}(\sigma^t, \pi_2^t) &\overset{\{25\}}{=} \max_{\pi_1^t \in \Pi_1^t} V_t(\sigma^t, \langle \pi_1^t, \pi_2^t \rangle) \\ &= \max_{\pi_1^t \in \Pi_1^t} \left[Q_t^R(\sigma^t, \langle \delta_1^t, \delta_2^t \rangle) + V_{t+1}(\text{U}_{\text{ss}}(\sigma^t, \delta^t), \langle \pi_1^{t+1}, \pi_2^{t+1} \rangle) \right] \\ &= \max_{\delta_1^t \in \Delta_1^S} \max_{\pi_1^{t+1}} \left[Q_t^R(\sigma^{t+1}, \langle \delta_1^t, \delta_2^t \rangle) + V_{t+1}(\sigma^{t+1}, \langle \pi_1^{t+1}, \pi_2^{t+1} \rangle) \right] \\ &= \max_{\delta_1^t \in \Delta_1^S} \left[Q_t^R(\sigma^t, \langle \delta_1^t, \delta_2^t \rangle) + \max_{\pi_1^{t+1}} \left[V_{t+1}(\sigma^{t+1}, \langle \pi_1^{t+1}, \pi_2^{t+1} \rangle) \right] \right] \\ &\overset{\{25\}}{=} \max_{\delta_1^t \in \Delta_1^S} \left[Q_t^R(\sigma^t, \langle \delta_1^t, \delta_2^t \rangle) + V_{t+1}^{\text{BR1}}(\sigma^{t+1}, \langle \pi_1^{t+1}, \pi_2^{t+1} \rangle) \right]. \quad \text{(A.34)} \end{aligned}$$

Here, U_{ss} is the statistic update rule, defined in accordance with (18). 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.29):

$$\begin{aligned} Q_t^R(\sigma_{m,1}^t \sigma_{c,1}^t, \langle \delta_1^t, \delta_2^t \rangle) &= \sigma_{m,1}^t(\vec{\theta}_1^t) \sigma_{c,1}^t(\vec{\theta}_2^t | \vec{\theta}_1^t) \sum_{a_1^t \in \mathcal{A}_1} \delta_1^t(a_1^t | \vec{\theta}_1^t) \\ &\quad \sum_{a_2^t \in \mathcal{A}_2} \delta_2^t(a_2^t | \vec{\theta}_2^t) R(\langle \vec{\theta}_1^t, \vec{\theta}_2^t \rangle, \langle a_1^t, a_2^t \rangle). \quad \text{(A.35)} \end{aligned}$$

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

$$\begin{aligned} V_{t+1}^{\text{BR1}}(\sigma_{m,1}^{t+1} \sigma_{c,1}^{t+1}, \pi_2^{t+1}) &\overset{\{A.32\}}{=} \sigma_{m,1}^{t+1} \cdot \vec{\nu}_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]} \\ &= \sum_{\vec{\theta}_1^{t+1} \in \vec{\Theta}_1^{t+1}} \sigma_{m,1}^{t+1}(\vec{\theta}_1^{t+1}) \vec{\nu}_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}(\vec{\theta}_1^{t+1}) \\ &\overset{\{18\}}{=} \sum_{\vec{\theta}_1^t \in \vec{\Theta}_1^t} \sigma_{m,1}^t(\vec{\theta}_1^t) \sum_{\vec{\theta}_1^t \in \vec{\Theta}_2^t} \sigma_{c,1}^t(\vec{\theta}_2^t | \vec{\theta}_1^t) \sum_{a_1^t \in \mathcal{A}_1} \delta_1^t(a_1^t | \vec{\theta}_1^t) \\ &\quad \sum_{a_2^t \in \mathcal{A}_2} \delta_2^t(a_2^t | \vec{\theta}_2^t) \sum_{o^{t+1} \in \mathcal{O}_1} \sum_{o^{t+1} \in \mathcal{O}_2} \\ &\quad \Pr(o_1^{t+1} o_2^{t+1} | \langle \vec{\theta}_1^t, \vec{\theta}_2^t \rangle, \langle a_1^t, a_2^t \rangle) \vec{\nu}_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}(\vec{\theta}_1^{t+1}). \quad \text{(A.36)} \end{aligned}$$

Filling the expanded equations into (A.34) and factorizing gives:

$$\begin{aligned} V_t^{\text{BR1}}(\sigma_{m,1}^t \sigma_{c,1}^t, \pi_2^t) &\overset{\{25\}}{=} \max_{\delta_1^t \in \Delta_1^S} \Big[Q_t^R(\sigma_{m,}^t, \langle \delta_1^t, \delta_2^t \rangle | \sigma_{c,1}^t) + \\ &\quad V_{t+1}^{\text{BR1}}(\sigma_{m,1}^{t+1}, \pi_2^{t+1} | \sigma_{c,1}^{t+1}) \Big] \overset{\{A.35, A.36\}}{=} \\ &\quad \max_{\delta_1^t \in \Delta_1^S} \Bigg[\sum_{\vec{\theta}^t \in \vec{\Theta}_1^t} \sigma_{m,1}^t(\vec{\theta}_1^t) \sum_{\vec{\theta}^t \in \vec{\Theta}_2^t} \sigma_{c,1}^t(\vec{\theta}_2^t | \vec{\theta}_1^t) \sum_{a_1^t \in \mathcal{A}_1} \\ &\quad \delta_1^t(a_1^t | \vec{\theta}_1^t) \sum_{a_1^t \in \mathcal{A}_2} \delta_2^t(a_2^t | \vec{\theta}_2^t) \Bigg(R(\langle \vec{\theta}_1^t, \vec{\theta}_2^t \rangle, \langle a_1^t, a_2^t \rangle) + \\ &\quad \sum_{o^{t+1} \in \mathcal{O}} \Pr(o^{t+1} | \vec{\theta}^t, a^t) \vec{\nu}_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}(\vec{\theta}_1^{t+1}) \Bigg) \Bigg]. \quad \text{(A.37)} \end{aligned}$$

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.37) as follows:

$$\begin{aligned} &V_t^{\text{BR1}}(\sigma^t, \pi_2^t) \\ &= \sum_{\vec{\theta}_1^t \in \vec{\Theta}_1^t} \sigma_{m,1}^t(\vec{\theta}_1^t) \max_{a_1^t} \Bigg[\sum_{\vec{\theta}^t \in \vec{\Theta}_2^t} \sigma_{c,1}^t(\vec{\theta}_2^t | \vec{\theta}_1^t) \sum_{a_1^t \in \mathcal{A}_2} \delta_2^t(a_2^t | \vec{\theta}_2^t) \\ &\quad \left(R(\vec{\theta}^t, a^t) + \sum_{o^{Ft+1} \in \mathcal{O}} \Pr(o^{t+1} | \vec{\theta}^t, a^t) \vec{\nu}_{[\sigma_{c,1}^{t+1}, \pi_2^{t+1}]}(\vec{\theta}_1^{t+1}) \right) \Bigg] \\ &\overset{\{26\}}{=} \sum_{\vec{\theta}_1^t \in \vec{\Theta}_1^t} \sigma_{m,1}^t(\vec{\theta}_1^t) \vec{\nu}_{[\sigma_{c,1}^t, \pi_2^t]}(\vec{\theta}_1^t) \overset{\{\text{vec. not.}\}}{=} \vec{\sigma}_{m,1}^t \cdot \vec{\nu}_{[\sigma_{c,1}^t, \pi_2^t]}. \quad \text{(A.38)} \end{aligned}$$

This corresponds to (A.33). 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] Jean-Pierre Aubin, *Optima and equilibria: an introduction to nonlinear analysis*, volume 140, Springer Science & Business Media, 1998.

[2] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[3] Jilles S. Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet, 'Optimally solving Dec-POMDPs as continuous-state MDPs', in *Proceedings of the International Joint Conference on Artificial Intelligence*, (2013).

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

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

[6] Piotr J Gmytrasiewicz and Prashant Doshi, 'Interactive POMDPs: Properties and preliminary results', in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pp. 1374–1375. IEEE Computer Society, (2004).

[7] Eric A Hansen, Daniel S. Bernstein, and Shlomo Zilberstein, 'Dynamic Programming for Partially Observable Stochastic Games', in *Proceedings of the National Conference on Artificial Intelligence*, volume 4, pp. 709–715, (2004).

[8] Manish Jain, Dmytro Korzhyk, Ondřej Vaněk, Vincent Conitzer, Michal Pěchouček, and Milind Tambe, 'A double oracle algorithm for zero-sum security games on graphs', in *Proceedings of the International Joint Conference on Autonomous Agents and Multi Agent Systems*, pp. 327–334, (2011).

[9] Daphne Koller, Nimrod Megiddo, and Bernhard Von Stengel, 'Fast algorithms for finding randomized strategies in game trees', in *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pp. 750–759. ACM, (1994).

[10] Liam C. MacDermed and Charles Isbell, 'Point based value iteration with optimal belief compression for Dec-POMDPs', in *Advances in Neural Information Processing Systems 26*, pp. 100–108, (2013).

[11] Jean-Francois Mertens and Shmuel 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).

[12] Ranjit Nair, Milind Tambe, Makoto Yokoo, David Pynadath, and Stacy Marsella, 'Taming Decentralized POMDPs: Towards efficient policy computation for multiagent settings', in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 705–711, (2003).

[13] Ashutosh Nayyar, Abhishek Gupta, Cedric Langbort, and Tamer 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] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis, 'Decentralized stochastic control with partial history sharing: A common information approach', *IEEE Transactions on Automatic Control*, **58**, 1644–1658, (July 2013).

[15] Frans A. Oliehoek, 'Sufficient Plan-Time Statistics for Decentralized POMDPs', in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 302–308, (2013).

[16] Frans A. Oliehoek and Christopher Amato, 'Dec-POMDPs as non-observable MDPs', IAS technical report IAS-UVA-14-01, Intelligent Systems Lab, University of Amsterdam, Amsterdam, The Netherlands, (October 2014).

[17] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos A Vlassis, 'Optimal and approximate q-value functions for Decentralized POMDPs.', *Journal of AI Research*, **32**, 289–353, (2008).

[18] Frans A. Oliehoek and Nikos Vlassis, 'Dec-POMDPs and extensive form games: equivalence of models and algorithms', *Ias technical report IAS-UVA-06-02, University of Amsterdam, Intelligent Systems Lab, Amsterdam, The Netherlands*, (2006).

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

[20] Joelle Pineau, Geoffrey J Gordon, and Sebastian Thrun, 'Anytime point-based approximations for large POMDPs.', *Journal of AI Research*, **27**, 335–380, (2006).

[21] Jean-Pierre Ponssard, 'Zero-sum games with "almost" perfect information', *Management Science*, **21**(7), 794–805, (1975).

[22] Jean-Pierre Ponssard and Sylvain Sorin, 'Some results on zero-sum games with incomplete information: The dependent case', *International Journal of Game Theory*, **9**(4), 233–245, (1980).

[23] Jean-Pierre Ponssard and Shmuel Zamir, 'Zero-sum sequential games with incomplete information', *International Journal of Game Theory*, **2**(1), 99–107, (1973).

[24] Jonathan Rubin and Ian Watson, 'Computer poker: A review', *Artificial Intelligence*, **175**(5), 958–987, (2011).

[25] Yoav Shoham and Kevin Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*, Cambridge University Press, 2008.

[26] Sylvain Sorin, 'Stochastic Games with incomplete information', in *Stochastic Games and applications*, eds., Abraham Neyman and Sylvain Sorin, volume 570, Springer, (2003).

[27] Matthijs T.J. Spaan and Nikos Vlassis, 'Perseus: Randomized point-based value iteration for POMDPs.', *Journal of AI Research*, **24**, 195–220, (2005).

[28] Jeffrey Wu and Sanjay Lall, 'A theory of sufficient statistics for teams', in *Proc. of the 53rd Annual Conference on Decision and Control*, pp. 2628–2635. IEEE, (2014).