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. In this paper, we introduce a new formulation for the value function of a zero-sum Partially Observable Stochastic Game (zs-POSG) in terms of a 'plan-time sufficient statistic', a distribution over joint sets of information. We prove that this value function exhibits concavity and convexity with respect to appropriately chosen subspaces of the statistic space. We anticipate that this result is a key pre-cursor for developing solution methods that exploit such structure. Finally, we show that the formulation allow us to reduce a finite zs-POSG to a 'centralized' model with shared observations, thereby transferring results for the latter (narrower) class of games to games with individual observations.

1 Introduction

The 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 based on individual observations. In this work, we prove the existence of structural properties of the zs-POSG value function, which 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 [6], which can be interpreted as the belief of a special type of Partially Observable Markov Decision Process (POMDP) to which the collaborative Decentralized POMDP (Dec-POMDP) can be reduced [1, 4, 5]. This enabled tackling these problems using solution methods for POMDPs, leading to increases in scalability [1].

We extend these results for Dec-POMDPs to the zs-POSG setting by presenting three contributions. First, a definition of the value function of a zs-POSG in terms of distributions over information called *plan-time sufficient statistics*. Second, a proof that the formulation allows for a generalization over the statistics: on every stage, the value function exhibits concavity and convexity in different *subspaces* of statistic-space. Third, a reduction of the zs-POSG to a *Non-Observable Stochastic Game*, which in turn allows us to show 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 plan-time sufficient statistics. We argue that this result may open up the route for new solution methods.

[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

2 Model definition

Definition 1. 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 (finite) 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 $\Pr(o^{t+1}|s^{t+1}, a^t)$,
- $R : S \times \mathcal{A} \times S \to \mathbb{R}$ is the reward function for agent 1,
- $b^0 \in \Delta(\mathcal{S})$ is the initial probability distribution over states.

In the zs-POSG, we aim to find *rational strategies* (i.e., maxmin-strategies) and the *value* of the game (i.e., the expected sum of rewards when agents follow the said strategies). 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* (one-stage policies) $\pi_i = \langle \delta_i^0 \ldots \delta_i^{h-1} \rangle$. 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* $\pi_i^t = \langle \delta_i^t, \ldots, \delta_i^{h-1} \rangle$. Rational policies are denoted π^*.

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. We will use the term 'value function' for a function that captures the future expected rewards under a rational joint policy.

3 Structure in zs-POSG Value Function

It is theoretically possible to convert a zs-POSG to *normal form* and solve it using standard methods, but this is infeasible in practice. An alternative is to convert the zs-POSG to an extensive form game (EFG) and solve it in sequence form [3]. While this is more efficient than the NFG route, it is often 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) [8]. Instead, in this section, we give a value function formulation in terms of a so-called plan-time sufficient statistic (originally used in the collaborative Dec-POMDP setting [6]). We show that this value function exhibits a potentially exploitable structure at every stage of the game.

Definition 2. The **plan-time sufficient statistic** for a general past joint policy φ^t, assuming b^0 is known, is a distribution over joint AOHs: $\sigma^t(\vec{\theta}^t) \triangleq \Pr(\vec{\theta}^t|b^0, \varphi^t)$.

The zs-POSG Q-value function at the final stage $h-1$ reduces to the immediate reward function: $Q^*_{h-1}(\sigma^{h-1}, \vec{\theta}^{h-1}, \delta^{h-1}) \triangleq R(\vec{\theta}^{h-1}, \delta^{h-1})$. We define the Q-value for all other stages as:

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

$$Q^*_t(\sigma^t, \delta^t) \triangleq \sum_{\vec{\theta}^t} \sigma^t(\vec{\theta}^t) Q^*_t(\sigma^t, \vec{\theta}^t, \delta^t). \quad (2)$$

Here, σ^{t+1} is found using the statistic update rule: $\sigma^{t+1}(\vec{\theta}^{t+1}) \triangleq \Pr(o^{t+1}|\vec{\theta}^t, a^t)\delta^t(a^t|\vec{\theta}^t)\sigma^t(\vec{\theta}^t)$.

Lemma 1. *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. See [9]. □

We can now define the value function of the zs-POSG in terms of σ^t:

$$V^*_t(\sigma^t) \triangleq \max_{\delta^t_1 \in \Delta^S_1} \min_{\delta^t_2 \in \Delta^S_2} Q^*_t(\sigma^t, \langle \delta^t_1, \delta^t_2 \rangle). \quad (3)$$

Although we have 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 $V^*_t(\sigma^t)$.

We decompose the plan-time sufficient statistic in marginal and conditional terms for agent 1 as $\sigma^t(\langle \vec{\theta}^t_1, \vec{\theta}^t_2 \rangle) = \sigma^t_{m,1}(\vec{\theta}^t_1)\sigma^t_{c,1}(\vec{\theta}^t_2|\vec{\theta}^t_1)$ (similar for agent 2), and show that the value function is concave in marginal-space for agent 1, $\Delta(\vec{\Theta}^t_1)$, and convex in marginal-space for agent 2, $\Delta(\vec{\Theta}^t_2)$.

Theorem 1. *V^*_t is concave in $\Delta(\vec{\Theta}^t_1)$ for a given $\sigma^t_{c,1}$, and convex in $\Delta(\vec{\Theta}^t_2)$ for a given $\sigma^t_{c,2}$.*

Proof. See [9]. □

Figure 1 provides intuition on how the concepts are related. Each 'slice' in statistic-space corresponds to a single conditional $\sigma^t_{c,1}$ ($\sigma^t_{c,2}$) and a marginal-space. The value function exhibits a concave (convex) shape on this slice, and is comprised of linear segments that each correspond to a partial policy π^t_i of the opposing agent.

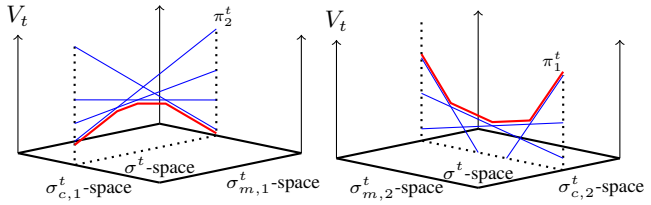

Figure 1: An abstract visualization of the decomposition of statistic-space into marginal-space and conditional-space.

The importance of this theorem is that it suggests ways to (approximately) represent $V^*_t(\sigma^t)$. As our formulation preserves the structure of the game, it allows us to make statements about how *value generalizes as a function of the information distribution*. Thus, it may enable the development of new solution methods for zs-POSGs.

4 Reduction to NOSG

Similar to how the use of sufficient statistics allows a Dec-POMDP to be reduced to a special type of (centralized) POMDP [1, 4, 5, 7], it turns out that, through the use of sufficient statistics, any finite zs-POSG can be reduced to a game to which we refer as a *Non-Observable Stochastic Game* (NOSG): a stochastic game with a single, shared NULL observation, where the joint AOH acts as the state. We give the full NOSG definition in [9].

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 (3). This indicates that properties of 'zero-sum stochastic games with shared observations' [2] also hold for finite zs-POSGs.

5 Conclusions

We present a structural result on the shape of the value function of two-player zero sum Partially Observable Stochastic Games (zs-POSG). We define the zs-POSG value function in terms of an information distribution called the sufficient plan-time statistic, and prove that this value function exhibits concavity (convexity) in the space of marginal statistics of the maximizing (minimizing) agent. 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 finite zs-POSG to a stochastic game with shared observations, thereby transferring properties of this narrower class of games to the finite 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.

REFERENCES

[1] 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).

[2] 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).

[3] 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).

[4] 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).

[5] 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).

[6] 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).

[7] 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).

[8] 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).

[9] A. J. Wiggers, F. A. Oliehoek, and D. M. Roijers, 'Structure in the Value Function of Two-Player Zero-Sum Games of Incomplete Information', *ArXiv e-prints*, (June 2016). http://arxiv.org/abs/1606.06888.