Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Sufficient Plan-Time Statistics for Decentralized POMDPs

Frans A. Oliehoek
Maastricht University
The Netherlands
frans.oliehoek@maastrichtuniversity.nl

Abstract

Optimal decentralized decision making in a team of cooperative agents as formalized by decentralized POMDPs is a notoriously hard problem. A major obstacle is that the agents do not have access to a sufficient statistic during execution, which means that they need to base their actions on their histories of observations. A consequence is that even during off-line planning the choice of decision rules for different stages is tightly interwoven: decisions of earlier stages affect how to act optimally at later stages, and the optimal value function for a stage is known to have a dependence on the decisions made up to that point. This paper makes a contribution to the theory of decentralized POMDPs by showing how this dependence on the 'past joint policy' can be replaced by a sufficient statistic. These results are extended to the case of k-step delayed communication. The paper investigates the practical implications, as well as the effectiveness of a new pruning technique for MAA* methods, in a number of benchmark problems and discusses future avenues of research opened by these contributions.

1 Introduction

Multiagent planning under uncertainty has attracted considerable attention in the last decade. In many applications, a team of agents will be facing a great number of uncertainties—e.g., due to unpredictable outcomes of actions, limited and noisy sensors, and failing or absence of communication—that have to be dealt with in a principled manner. Decentralized partially-observable Markov decision processes (Dec-POMDPs) [Seuken and Zilberstein, 2008; Oliehoek, 2012] have been put forward as a framework for such problems. However, optimal decentralized decision making for a team of cooperative agents as formalized in the framework is a notoriously hard problem [Bernstein *et al.*, 2002].

A major obstacle is that the agents only have access to their individual observations during execution. It is currently not known, however, if such individual observation histories can be summarized using more compact (sufficient) statistics, without fixing the policies of the other agents. This means that even if we perform the planning in advance (i.e., in an off-line planning phase) we need to assume that agents base their actions on their entire histories of observations. A consequence is that even during off-line planning the choice of decision rules for different stages is tightly interwoven: as in (single-agent) MDPs [Puterman, 1994], the best future decisions affect optimal decisions for earlier stages, but in Dec-POMDPs *decisions of earlier stages affect how to act optimally at later stages*. That is, the optimal value function for a stage t is known to have a dependence on the policies followed at stages $0, \dots, t-1$.

This paper makes a contribution to the theory of decentralized POMDPs by showing how this dependence on the 'past joint policy' can be replaced by a probability distribution over joint action-observation histories, or a joint distribution over joint observation histories and states. Thereby, it introduces sufficient *plan-time* statistics for the past joint policy. We show how the optimal value functions can be formulated in terms of these statistics and prove their correctness, hence establishing the sufficiency of the statistics. Moreover, we also show how these results can be extended to settings with k-step delayed communication. In an empirical evaluation, we investigate the potential for practical implications in a number of benchmark problems, showing that in certain problems the use of sufficient statistics can allow for a much more compact representation of the optimal value function. Additionally, a new *sufficient statistic-based pruning* technique for heuristic search methods is shown to have the potential to improve planning efficiency, although it does not directly address the bottleneck of current state-of-the-art methods. Finally, we discuss avenues for future research opened by the identification of the proposed plan-time statistics.

This paper is organized as follows. First, in Section 2, we provide the necessary background on Dec-POMDPs and their value functions. The main contribution, the identification of sufficient statistics is presented in Section 3. Section 4 extends these results delayed-communication settings. The empirical evaluation is described in Section 5. Finally, Section 6 discusses opportunities for future work, and Section 7 concludes.

2 Background

Here we provide a concise review of the necessary background on Decentralized POMDPs and their value functions. For a more extensive introduction see [Oliehoek, 2012].

2.1 Dec-POMDPs

A Dec-POMDP is a model for multiagent planning under uncertainty, in which, at every time step or *stage*, each agent selects an action based on its individual observations (we assume no communication unless mentioned explicitly).

Definition 1 (Dec-POMDP). A *decentralized partially observable Markov decision process (Dec-POMDP)* is a tuple $\langle \mathcal{D}, \mathcal{S}, \mathcal{A}, T, R, \mathcal{O}, O, h, b_0 \rangle$, where

- $\mathcal{D} = \{1, \ldots, n\}$ is the set of n agents,
- $\mathcal{S}$ is the finite set of states s,
- $\mathcal{A}$ is the set of joint actions $a = \langle a_1, \ldots a_n \rangle$,
- T is the transition function that specifies $\Pr(s_{t+1}|s_t, a_t)$,
- $R(s,a)$ is the immediate reward function,
- $\mathcal{O}$ is the set of joint observations $o = \langle o_1, \ldots, o_n \rangle$,
- O the observation function: $\Pr(o_{t+1}|a_t, s_{t+1})$,
- h is the horizon of the problem,
- $b_0 \in \triangle(\mathcal{S})$, is the initial state distribution at time $t = 0$.

The goal in Dec-POMDPs is to find an optimal joint policy π^* that maximizes the expected sum (over stages) of rewards. A key difficulty that sets Dec-POMDPs apart from frameworks as multiagent MDPs [Boutilier, 1996], is that this joint policy is *decentralized:* it is a tuple $\langle \pi_i, \ldots, \pi_n \rangle$ such that the individual policy π_i of every agent i maps individual observations histories (OH) $\vec{o}_{i,t} = (o_{i,1}, \ldots, o_{i,t})$ to actions $\pi_i(\vec{o}_{i,t}) = a_{i,t}$. The *joint OH* is denoted $\vec{o}_t = \langle \vec{o}_{1,t}, \ldots, \vec{o}_{n,t} \rangle$. We also consider stochastic policies, which map from action-observation histories (AOH) $\vec{\theta}_{i,t} = (a_{i,0}, o_{i,1}, \ldots, a_{i,t-1}, o_{i,t})$ to probability distributions over actions: $\pi_i(a_{i,t}|\vec{\theta}_{i,t})$. Joint AOHs are denoted $\vec{\theta}_t$.

A policy is a sequence $\pi_i = (\delta_{i,0}, \ldots, \delta_{i,h-1})$ of decision rules that map length-t observation histories to actions $\delta_{i,t}(\vec{o}_{i,t}) = a_{i,t}$. We also consider stochastic decision rules $\delta_{i,t}(a_{i,t}|\vec{\theta}_{i,t})$. A joint decision rule δ_t specifies a decision rule for each agent. We define a joint policy that is partially specified $\varphi_t = (\delta_0, \ldots, \delta_{t-1})$ as the *past joint policy* at stage t.

2.2 Optimal Value Functions

As for MDPs [Puterman, 1994; Bertsekas, 2005] and POMDPs [Kaelbling *et al.*, 1998; Spaan, 2012], for Dec-POMDPs, it is possible to identify optimal value functions. To define them, we will need the following preliminary definitions:

$$R(\vec{\theta}_t, \delta_t) = \sum_{s_t} \Pr(s_t|b_0, \vec{\theta}_t) \sum_{a_t} R(s_t, a_t) \delta_t(a_t|\vec{\theta}_t), \quad (2.1)$$

$$\Pr(\vec{\theta}_{t+1}|\vec{\theta}_t, \delta_t) = \sum_{s_t} \Pr(s_t|b_0, \vec{\theta}_t) \sum_{s_{t+1}} \Pr(o_{t+1}|a_t, s_{t+1}) \Pr(s_{t+1}|s_t, a_t) \delta_t(a_t|\vec{\theta}_t). \quad (2.2)$$

Theorem 1 ([Oliehoek *et al.*, 2008b]). *The optimal value function for a Dec-POMDP is defined as*

$$Q_t(b_0, \varphi_t, \vec{\theta}_t, \delta_t) = R(\vec{\theta}_t, \delta_t) + \sum_{a_t} \sum_{o_{t+1}} \Pr(\vec{\theta}_{t+1}|\vec{\theta}_t, \delta_t) Q_{t+1}(b_0, \varphi_{t+1}, \vec{\theta}_{t+1}, \delta^*_{t+1}) \quad (2.3)$$

(for the last stage the second term is omitted) with $\varphi_{t+1} = (\varphi_t, \delta_t)$ the past joint policy formed by concatenating φ_t and δ_t. This equation in turn defines the optimal decision rule via

$$Q_t(b_0, \varphi_t, \delta_t) \triangleq \sum_{\vec{\theta}_t} \Pr(\vec{\theta}_t|b_0, \varphi_t) Q_t(b_0, \varphi_t, \vec{\theta}_t, \delta_t), \quad (2.4)$$

$$\delta^*_t = \arg\max_{\delta_t} Q_t(b_0, \varphi_t, \delta_t). \quad (2.5)$$

A number of remarks are in order:

- Note that (2.5) defines δ^*_{t+1} in (2.3).
- In contrast to other descriptions, this set of equations, referred to as the 'sequentially rational' optimal value function, determines the optimal value also for joint AOHs that will never be realized under an optimal joint policy [Oliehoek *et al.*, 2008b].
- Here, we follow the notation of [Oliehoek, 2010; 2012], which makes explicit the dependence on b_0. The definitions of (2.1), (2.2) used here are modified to allow for stochastic policies.
- While the above formulations do not resemble traditional Q-function for MDPs, the use of the letter 'Q' can understood by interpreting δ_t as an action in a meta-level MDP for the planning process [Oliehoek, 2010]. This meta-MDP has 'states' (b_0, φ_t) with values $V_t(b_0, \varphi_t)$ corresponding to the maximum of (2.5).

Even though the above description is (relatively) concise, using these equations to compute the optimal joint policy is cumbersome, as it requires evaluating (2.4), (2.5) for all past joint policies φ_{h-1} at the last stage. As such, even if the maximization in (2.5) could be performed efficiently, this algorithm would at best gain one horizon on brute force search. Therefore, in practice, researchers have resorted to heuristic search over this space of joint policies [Szer *et al.*, 2005; Oliehoek *et al.*, 2013], dynamic programming [Hansen *et al.*, 2004; Boularias and Chaib-draa, 2008; Amato *et al.*, 2009] or approximate methods [Nair *et al.*, 2003; Emery-Montemerlo *et al.*, 2004; Oliehoek *et al.*, 2008a; Seuken and Zilberstein, 2008; Kumar and Zilberstein, 2010; Wu *et al.*, 2010].

3 Sufficient Plan-Time Statistics

In this section we present our main contribution: the identification of sufficient statistics of the past joint policy for Dec-POMDPs. That is, we show how the equations in Theorem 1 can be reformulated such that they no longer depend on the past joint policy φ_t, but rather on a sufficient statistic σ_t that summarizes it. Since many φ_t may correspond to the same statistic, this can lead to substantially more compact representations of the optimal value function.

3.1 Statistics for General Policies

Although it is well-known that a Dec-POMDP has at least one deterministic optimal joint policy, there is no reason to exclude the more general case of stochastic policies from the description of optimal value functions. Moreover, this assumption will lead to simplest description of a sufficient statistic σ_t as follows.

Definition 2 (Sufficient statistic for general policies). The sufficient statistic for a general φ_t, assuming b_0 is known, is a the distribution over joint AOHs: $\sigma_t(\vec{\theta}_t) \triangleq \Pr(\vec{\theta}_t|b_0,\varphi_t)$.

Given this definition, we will now posit the equations that are the equivalent of Theorem 1. The proof of their correctness follows. The optimal value can be expressed as

$$Q_t(b_0,\sigma_t,\vec{\theta}_t,\delta_t) = R(\vec{\theta}_t,\delta_t)+ \sum_{a_t}\sum_{o_{t+1}} \Pr(\vec{\theta}_{t+1}|\vec{\theta}_t,\delta_t)Q_{t+1}(b_0,\sigma_{t+1},\vec{\theta}_{t+1},\delta^{sg*}_{t+1}) \quad (3.1)$$

with the updated statistic—note $\vec{\theta}_{t+1} = (\vec{\theta}_t,a_t,o_{t+1})$—

$$\sigma_{t+1}(\vec{\theta}_{t+1}) = \Pr(o_{t+1}|\vec{\theta}_t,a_t)\delta_t(a_t|\vec{\theta}_t)\sigma_t(\vec{\theta}_t). \quad (3.2)$$

Optimal decision rules can be derived from

$$Q_t(b_0,\sigma_t,\delta_t) \triangleq \sum_{\vec{\theta}_t} \sigma_t(\vec{\theta}_t)Q_t(b_0,\sigma_t,\vec{\theta}_t,\delta_t), \quad (3.3)$$

$$\delta^{sg*}_t = \arg\max_{\delta_t} Q_t(b_0,\sigma_t,\delta_t). \quad (3.4)$$

We formally proof the correctness of the above equations, starting with the sufficiency of σ_t for predicting the optimal value $Q_t(b_0,\varphi_t,\vec{\theta}_t,\delta_t)$.

Theorem 2. *For all φ_t, the distribution over AOHs $\sigma_t(\vec{\theta}_t)$ is sufficient to predict the optimal value:*

$$\forall_{b_0,\vec{\theta}_t,\delta_t} \quad Q_t(b_0,\varphi_t,\vec{\theta}_t,\delta_t) = Q_t(b_0,\sigma_t,\vec{\theta}_t,\delta_t).$$

Proof. The proof is listed in the appendix. □

The following conclusions follow immediately.

Corollary 1. *The non-history-based 'meta MDP' Q-functions given by (2.4) and (3.3) are identical: $Q_t(b_0,\varphi_t,\delta_t) = Q_t(b_0,\sigma_t,\delta_t)$.*

Proof. This follows directly from Theorem 2 and the definitions of the 'meta MDP' Q-functions in (2.4) and (3.3). □

Corollary 2. *The system of equations given by (3.1) and (3.4) express the optimal value function.*

Proof. This follows directly from their equality to equations (2.3) and (2.5). □

3.2 Deterministic Policies

The above definition of the statistic σ_t leads to the most straightforward formulation. However, in the context of deterministic policies the statistic is not directly useful; when restricting to deterministic policies, per definition, each φ_t induces a different $\sigma_t(\vec{\theta}_t)$. In this sub-section, we fix this problem by introducing a second statistic that additionally takes away the dependence on the initial belief b_0.

Definition 3 (Sufficient statistic for deterministic policies). The sufficient statistic for a tuple (b_0,φ_t), with φ_t deterministic, is a the distribution over joint OHs and states: $\sigma_t(s_t,\vec{o}_t) \triangleq \Pr(s_t,\vec{o}_t|b_0,\varphi_t)$.

In the following, we will also write $\sigma_t(s_t|\vec{o}_t)$ and $\sigma_t(\vec{o}_t)$ for the conditional and marginal computed from σ_t. Again, we will need preliminary definitions for the rewards and observation probabilities:

$$R(\sigma_t,\vec{o}_t,\delta_t) = \sum_{s_t} R(s_t,\delta_t(\vec{o}_t))\sigma_t(s_t|\vec{o}_t), \quad (3.5)$$

$$\Pr(o_{t+1}|\sigma_t,\vec{o}_t,\delta_t)=\sum_{s_t}\sum_{s_{t+1}} \Pr(o_{t+1},s_{t+1}|s_t,\delta_t(\vec{o}_t))\sigma_t(s_t|\vec{o}_t). \quad (3.6)$$

The next statistic (a function of σ_t and δ_t) is given by

$$\sigma_{t+1}(s_{t+1},\vec{o}_{t+1}) = \sum_{s_t} \Pr(s_{t+1},o_{t+1}|s_t,\delta_t(\vec{o}_t))\sigma_t(s_t,\vec{o}_t). \quad (3.7)$$

We are now in a position to give optimal value functions based on this new sufficient statistic.

Theorem 3. *Using the sufficient statistic for deterministic past joint policies, the optimal value function of a finite-horizon Dec-POMDP can be written as*

$$Q_t(\sigma_t,\vec{o}_t,\delta_t) = R(\sigma_t,\vec{o}_t,\delta_t)+ \sum_{o_{t+1}} \Pr(o_{t+1}|\sigma_t,\vec{o}_t,\delta_t)Q_{t+1}(\sigma_{t+1},\vec{o}_{t+1},\delta^{sd*}_{t+1}), \quad (3.8)$$

where optimal decision rules are defined via

$$Q_t(\sigma_t,\delta_t) \triangleq \sum_{\vec{o}_t} \sigma_t(\vec{o}_t)Q_t(\sigma_t,\vec{o}_t,\delta_t), \quad (3.9)$$

$$\delta^{sd*}_t = \arg\max_{\delta_t} Q_t(\sigma_t,\delta_t). \quad (3.10)$$

Proof. The proof is similar to that of Theorem 2. A sketch of the proof is in the appendix. □

3.3 Restricted-length Policies

In the equations for the optimal value function, the role of the observation history is purely in terms of providing accurate distributions over states $\sigma_t(s_t|\vec{o}_t)$ and providing the basis for action selection. In cases where it is possible to restrict the class of considered policies to policies that map from the last k observations, it is possible to maintain more compact statistics $\sigma_t(s_t,\vec{o}^k_t)$ over length-k observation histories.[1]

For Dec-POMDPs, such a restriction in general is suboptimal: the most accurate distribution $\sigma_t(s_t|\vec{o}_t)$ over states is given by the complete history, and as such, policies should in general condition on the entire history. Nevertheless, there may be situations where we can prove that conditioning on full history is not necessary. An example is the sub-class of transition independent Dec-MDPs (TI-Dec-MDPs) [Becker *et al.*, 2003]. For such problems, it can be shown that an optimal joint policy exists in decentralized mappings from

[1]The technicalities of such a statistic are similar to the situation of k-step delayed communication (treated in the next section) with the difference that o_{t-k+1} is not observed, but must be averaged over: $\sigma_{t+1}(s_{t+1},\vec{o}^k_{t+1}) = \sum_{s_t}\sum_{o_{t-k+1}} \Pr(s_{t+1},o_{t+1}|s_t,\delta^k_t(\vec{o}^k_t))\sigma_t(s_t,\vec{o}^k_t)$.

the last ($k = 1$) observation to actions. As such, it is possible to maintain a more compact statistic $\sigma_t(s_t,o_t)$. Furthermore, since in TI-Dec-MDPs the joint observation identifies the state and vice versa, this statistic simply reduces to a distribution over states $\sigma_t(s_t)$, and therefore corresponds exactly to the so-called *state-occupancy* that was recently identified as a sufficient statistic for planning for TI-Dec-MDPs and that has led to significant improvements in their solutions [Dibangoye *et al.*, 2012].

4 Delayed Communication

In this section we consider Dec-POMDPs with k-step delayed communication. That is, we assume that, at every stage t, all the agents broadcast their individual observations, but that this information only arrives at stage $t+k$. The descriptions of optimal value functions introduced in Section 3 can be generalized to delayed communication. Essentially, this integrates the insight of the previous section in existing descriptions of optimal value functions for delayed communication [Ooi and Wornell, 1996; Oliehoek *et al.*, 2008b]. In this section we will concentrate on the deterministic past joint policy formulation, but extension to stochastic policies follows trivially.

Delayed Communication Value Functions. We will follow the description of value functions given in [Oliehoek *et al.*, 2008b]. The main idea behind the descriptions of value functions for Dec-POMDPs with k-step delayed communication is that every stage t is similar to a horizon-k Dec-POMDP: since the communicated individual observations of stage $t-k$ will have arrived, each agent knows the joint AOH $\vec{\theta}_{t-k}$ and can compute b_{t-k}, the distribution over states at that stage:

$$b_{t-k}(s_{t-k}) = \Pr(s_{t-k}|b_0,\vec{\theta}_{t-k}).$$

This distribution serves the same role as the initial belief b_0 in a Dec-POMDP without communication. In addition, each agent will know the sequence $\vec{o}^{\,k}_{i,t} = (o_{i,t-k+1},\ldots,o_{i,t})$ of its last k private observations. Therefore, to act at stage t, the agents have to use a joint decentralized decision rule $\delta^k_t = \langle \delta^k_{1,t},\ldots,\delta^k_{n,t}\rangle$ that maps length-k observation histories to joint actions $\delta^k_t(\vec{o}^{\,k}_t) = a_t$. However, the optimal δ^k_t depends on φ^k_t, the past joint policy *since* stage $t-k$. This φ^k_t fulfills the same role as φ_t in the normal Dec-POMDP formulation and also has similar shape: it simply is a tuple of horizon-k policy trees, one for each agent.

Note that we still assume that planning takes place in advance, so each agent will be able to determine what φ^k_t is (given $\vec{\theta}_{t-k}$). This means that we can form the length-$(k+1)$ policy $\varphi^{k+1}_t = (\varphi^k_t,\delta^k_t)$ in exactly the same way as for normal Dec-POMDPs. The difference, however, is in the way it will be used, rather than directly plugging φ^{k+1}_t in the value function for stage $t+1$ (cf. equation 2.3), it will be used to track the length-k past joint policy at the next stage. In particular, at the next stage, each agent will receive o_{t-k+1} (via communication). Therefore they will know which part of φ^{k+1}_t has been executed during the last k stages $t-k+1,\ldots,t$ and they discard the part not needed further. We will write discarding the part of φ^{k+1}_t that is not consistent with o_{t-k+1} as

$$\varphi^k_{t+1} = \varphi^{k+1}_t \Downarrow_{o_{t-k+1}}. \tag{4.1}$$

The optimal value function for a finite-horizon Dec-POMDP with k-step delayed, cost and noise free communication [Oliehoek *et al.*, 2008b] is given by:

$$Q_t(b_{t-k},\varphi^k_t,\vec{\theta}^k_t,\delta^k_t) = R(b_t,\delta^k_t(\vec{\theta}^k_t)) + \sum_{o_{t+1}} \Pr(o_{t+1}|b_t,\delta^k_t(\vec{\theta}^k_t))Q_{t+1}(b_{t-k+1},\varphi^k_{t+1},\vec{\theta}^k_{t+1},\delta^{k*}_{t+1}) \tag{4.2}$$

where b_t is the joint belief that results from b_{t-k} and $\vec{\theta}^k_t$, and where the definitions of $R(\ldots)$ and $\Pr(o_{t+1}|\ldots)$ follow from trivial adaptations of equations (2.1) and (2.2). The next-stage length-k past joint policy is φ^k_{t+1} is given by (4.1). To better interpret (4.2), it is informative to compare this equation to the equation for Dec-POMDPs without communication (2.3). Analogous to that setting, also in the case of k-step delayed communication, we can define the optimal decision rules via:

$$Q_t(b_{t-k},\varphi^k_t,\delta^k_t) \triangleq \sum_{\vec{\theta}^k_t} \Pr(\vec{\theta}^k_t|b_{t-k},\varphi^k_t)Q_t(b_{t-k},\varphi^k_t,\vec{\theta}^k_t,\delta^k_t), \tag{4.3}$$

$$\delta^{k*}_t = \max_{\delta^k_t} Q_t(b_{t-k},\varphi^k_t,\delta^k_t). \tag{4.4}$$

Sufficient Statistics. While the number of past joint policies considered is only doubly exponential in k and not the full horizon, this number is very large for longer delays. As such, also in this case, having descriptions of value functions based on sufficient plan-time statistics can be valuable.

Definition 4 (Sufficient statistic for k-step delayed communication). A sufficient statistic for a tuple $\langle b_{t-k},\varphi^k_t\rangle$, with φ^k_t deterministic, is the distribution over joint OHs and states: $\sigma_t(s_t,\vec{o}^{\,k}_t) \triangleq \Pr(s_t,\vec{o}^{\,k}_t|b_{t-k},\varphi^k_t)$.

This allows us to define $R(\sigma_t,\vec{o}^{\,k}_t,\delta^k_t)$ and $\Pr(o_{t+1}|\sigma_t,\vec{o}^{\,k}_t,\delta^k_t)$, analogous to (3.5), (3.6). The next statistic is a function of σ_t, δ_t and the communicated joint observation o_{t-k+1}. Let $\vec{o}^{\,k}_t = (o_{t-k+1},\vec{o}^{\,k-1}_t)$ and $\vec{o}^{\,k}_{t+1} = (\vec{o}^{\,k-1}_t,o_{t+1})$, then the updated statistic is given by

$$\sigma_{t+1}(s_{t+1},\vec{o}^{\,k}_{t+1}) = \frac{\sum_{s_t} \Pr(s_{t+1},o_{t+1}|s_t,\delta^k_t(\vec{o}^{\,k}_t))\sigma_t(s_t,\vec{o}^{\,k}_t)}{P(o_{t-k+1}|\sigma_t)},$$

with $P(o_{t-k+1}|\sigma_t)$ a normalization constant.

Theorem 4. *The optimal value function of a Dec-POMDP with k-step delayed communication can be written as*

$$Q_t(\sigma_t,\vec{o}^{\,k}_t,\delta^k_t) = R(\sigma_t,\vec{o}^{\,k}_t,\delta^k_t) + \sum_{o_{t+1}} \Pr(o_{t+1}|\sigma_t,\vec{o}^{\,k}_t,\delta^k_t)Q_{t+1}(\sigma_{t+1},\vec{o}^{\,k}_{t+1},\delta^{k*}_{t+1}), \tag{4.5}$$

where the next-stage statistic σ_{t+1} is as defined above, and where optimal decision rules are defined via

$$Q_t(\sigma_t,\delta^k_t) \triangleq \sum_{\vec{o}^{\,k}_t} \sigma_t(\vec{o}^{\,k}_t)Q_t(\sigma_t,\vec{o}^{\,k}_t,\delta^k_t), \tag{4.6}$$

	$t=1$		$t=2$		$t=3$	
	φ_1	σ_1	φ_2	σ_2	φ_3	σ_3
tiger	9	2	729	20	$4.78e6$	4520
broadcast	4	4	64	56	$1.63e4$	$1.16e4$
recycling	9	9	729	441	$4.78e6$	X
FF	9	9	729	729	$4.78e6$	X
gridsmall	25	16	$1.56e4$	4096	$6.10e9$	X
hotel1	9	1	$5.90e4$	4	$1.7e19$	–

Table 1: Number of σ_t vs. number of φ_t .

$$\delta_t^{k*} = \arg\max_{\delta_t^k} Q_t(\sigma_t, \delta_t^k). \tag{4.7}$$

Proof. The proof—omitted due lack of space—shows that $Q_t(\sigma_t, \vec{o}_t^k, \delta_t^k) = Q_t(b_{t-k}, \varphi_t^k, \delta_t^k)$ following the the same steps as the proof of Theorem 2. □

5 Experiments

Here we report on an empirical evaluation directed at investigating the potential practical impact of the proposed sufficient statistics. A potential important consequence of using sufficient statistics is that it allows representing the optimal value function more compactly. However, the extent to which this is the case depends on how many φ_t map to the same distribution $\sigma_t(s_t, \vec{o}_t)$. Therefore, to investigate this potential in practice, we have examined the number of unique distributions $\sigma_t(s_t, \vec{o}_t)$ in a number of standard benchmark problems.[2]

The results are shown in Table 1. It shows the number of past joint policies φ_t for different stages t, as well as the number of unique distributions $\sigma_t(s_t, \vec{o}_t)$ that those histories induce, given the initial belief b_0. Entries marked '–' ran out of time (>1h), and marked 'X' ran out of memory (2GB). This clearly indicates a limitation of using sufficient statistics: caching the distributions themselves can take a considerable amount of memory. However, the results also show that there can be considerable reductions in size, although this is very much problem dependent. For instance, firefighting (FF) does not allow for any reduction: every φ_t induces a unique statistic σ_t. In contrast, tiger and hotel1 allow for a reductions of respectively three and four orders of magnitude. Note that these result clearly illustrate that the reduction due to sufficient statistics is quite different from the clustering technique used in [Oliehoek *et al.*, 2009]: it is not necessarily the case that problems that exhibit high clustering (such as broadcast channel and recycling) also lead to the highest reductions by using sufficient statistics.[3]

Since certain problems (tiger and hotel1) give large reductions, we examined whether it is possible to use these statistics to increase performance of policy search for these problems. In particular, we augmented the state-of-the-art GMAA*-ICE solver [Oliehoek *et al.*, 2013] with *sufficient statistic-based pruning (SSBP):* a procedure that checks if the statistic σ_t induced by the current φ_t was already encountered, allowing for pruning in the search. Effectively this transforms the GMAA* search tree into a DAG: when reaching a node that was visited before, it is only further expanded if the value along the new path is higher than before (i.e., the cost of reaching it is lower). We compare the number of nodes that are created when pruning based on σ_t versus when not.

		nodes created at depth t					
	SSBP	1	2	3	4	5	6
tiger							
QMDP, h5	yes	1	10	615	28475	4	
	no	9	69	2319	41130	4	
QBG,h6	yes	1	2	8	18	162	1
	no	9	2	8	18	166	1
hotel1							
QMDP, h4	yes	1	4	6	3		
	no	9	252	11178	10935		
QMDP, h5	yes	1	4	12	15	7	
	no	not solvable (out of 2GB mem.)					
QBG, h5	no	9	4	3	3	1	

Table 2: Number of created child nodes in GMAA-ICE, when using sufficient statistic-based pruning (SSBP).

The results are shown in Table 2. It clearly shows that when using QMDP, many nodes can be pruned. This translates to improvements in run time, e.g., 1.5s vs 39.9s for QMDP horizon 4. When using tighter heuristics as QBG, however, we see that these are already perform very well at guiding the search over past joint policies, such that the effect of using sufficient statistics is limited. For longer horizons, however, these heuristics are often difficult to compute and lose their tightness, meaning that there might still be a practical role for sufficient statistics in heuristic search algorithms. However, the current bottleneck that these methods experience—the complexity of expansion of the nodes for later stages—will need to be tackled with a different approach.

6 Future Work

This work lies the foundation for a number of future research directions. Potentially a great advantage of using sufficient statistics over past joint policies is that the difference between two statistics can be measured. An important direction of research is therefore to see if a bound on the difference in statistics can imply a bound on difference in value. This would directly provide a starting point for developing approximate versions of Dec-POMDP algorithms, which can give guarantees on the error. In fact, [Dibangoye *et al.*, 2013] simultaneously to this work identified a similar statistic of the form $\sigma_t(s, \vec{\theta}_t)$, and showed that the value function is piecewise linear and convex over this space, a property they show that can be exploited very effectively by adapting POMDP solution methods. Future work, should determine whether exploiting the same property of the more compact $\sigma_t(s, \vec{o}_t)$ statistic presented here can lead to further improvements.

Another promising direction of research is enabled by the insight that restricted-length policies allow for more compact statistics. This means that not only the maximization (3.10) becomes more tractable, but also that there is a larger

[2] Available from http://www.masplan.org/.

[3] Clustering tests if $P(s, \vec{o}_{-i} | \vec{o}_i, \varphi) = P(s, \vec{o}_{-i} | \vec{o}_i', \varphi)$ and merges OHs, thereby collapsing many extensions of φ. SSBP tests $P(s, \vec{o} | \varphi) = P(s, \vec{o} | \varphi')$ and avoids going down a branch φ completely, provided that it went down an equivalent branch φ' earlier.

chance that past joint policies will result in the same statistic. As such, artificially restricting the complexity of the policies gains traction in two complementary ways. Future research should investigate if this can be leveraged by searching in spaces of policies of incrementally increasing complexity.

7 Conclusions

This paper introduced sufficient plan-time statistics for Dec-POMDPs that allow for more compact description of the optimal value function. We formally proved that these descriptions are correct, i.e., the proposed statistics are indeed sufficient to predict the future value, and extended these descriptions to the case of k-step delayed communication in Dec-POMDPs. An empirical evaluation investigated the numerical impact on the description of optimal value functions for a number of benchmark problems, showing a potentially large, but problem-dependent reduction in size. Moreover, it was demonstrated that using sufficient statistic-based pruning can potentially speed up heuristic search for Dec-POMDPs, but that they do not address the current bottleneck for such methods. Finally, we discussed a number of promising directions of future work that are enabled by this work.

Acknowledgments

I want to thank Christopher Amato for valuable comments. Supported in part by NWO CATCH project #640.005.003.

A Appendix

Proof of Theorem 2

The proof is by induction over the stages of the problem.

Base Case. For the last stage $t = h - 1$, we have that

$$Q_t(b_0,\varphi_t,\vec{\theta}_t,\delta_t) = R(\vec{\theta}_t,\delta_t) = Q_t(b_0,\sigma_t,\vec{\theta}_t,\delta_t).$$

Induction Hypothesis. Per induction hypothesis, we assume that, for stage $t+1$, σ_{t+1} is a sufficient statistic. I.e., 'past-joint-policy form' Q-values are equal to 'statistic form':

$$Q_{t+1}(b_0,\varphi_{t+1},\vec{\theta}_{t+1},\delta_{t+1}) = Q_{t+1}(b_0,\sigma_{t+1},\vec{\theta}_{t+1},\delta_{t+1}), \tag{A.1}$$

where $\forall_{\vec{\theta}_{t+1}} \quad \sigma_{t+1}(\vec{\theta}_{t+1}) \triangleq \Pr(\vec{\theta}_{t+1}|b_0,\varphi_{t+1})$.

Induction Step. We need to show that for stage t, σ_t is a sufficient statistic. That is, if $\Pr(\vec{\theta}_t|b_0,\varphi_t) = \sigma_t(\vec{\theta}_t)$ then the Q-values are equal:

$$\forall_{\vec{\theta}_t} \left[\Pr(\vec{\theta}_t|b_0,\varphi_t) = \sigma_t(\vec{\theta}_t)\right] \quad \Longrightarrow \quad Q_t(b_0,\varphi_t,\vec{\theta}_t,\delta_t) = Q_t(b_0,\sigma_t,\vec{\theta}_t,\delta_t), \tag{A.2}$$

Proof: We assume that $\forall_{\vec{\theta}_t} \left[\Pr(\vec{\theta}_t|b_0,\varphi_t) = \sigma_t(\vec{\theta}_t)\right]$ (Assumpt.1). Now we need to show the identity of the Q-values of the r.h.s. of (A.2). Using their respective definitions (2.3) and (3.1), we need to show

$$Q_{t+1}(b_0,\varphi_{t+1},\vec{\theta}_{t+1},\delta^{pp*}_{t+1}) = Q_{t+1}(b_0,\sigma_{t+1},\vec{\theta}_{t+1},\delta^{sg*}_{t+1}). \tag{A.3}$$

This is the case if

1. the optimal next decision rule under 'past-joint-policy form', δ^{pp*}_{t+1}, and the optimal one under 'sufficient-statistic form', δ^{sg*}_{t+1}, are equal.
2. $\forall_{\vec{\theta}_{t+1}} \quad \Pr(\vec{\theta}_{t+1}|b_0,\varphi_{t+1}) = \sigma_{t+1}(\vec{\theta}_{t+1})$,

since then the IH applies. We first proof item 2):

$$\sigma_{t+1}(\vec{\theta}_{t+1}) \stackrel{\{(3.2)\}}{=} \Pr(o_{t+1}|\vec{\theta}_t,a_t)\delta_t(a_t|\vec{\theta}_t)\sigma_t(\vec{\theta}_t) \stackrel{\{\text{Assumpt.1}\}}{=} \Pr(o_{t+1}|\vec{\theta}_t,a_t)\delta_t(a_t|\vec{\theta}_t)\Pr(\vec{\theta}_t|b_0,\varphi_t) = \Pr(\vec{\theta}_{t+1}|b_0,\varphi_{t+1}).$$

Using this result, we prove item 1) via the I.H.: $\delta^{pp*}_{t+1} \triangleq$

$$\arg\max_{\delta_{t+1}} \sum_{\vec{\theta}_{t+1}} \Pr(\vec{\theta}_{t+1}|b_0,\varphi_{t+1})Q_{t+1}(b_0,\varphi_{t+1},\vec{\theta}_{t+1},\delta_{t+1})$$

$$= \arg\max_{\delta_{t+1}} \sum_{\vec{\theta}_{t+1}} \sigma_{t+1}(\vec{\theta}_{t+1})Q_{t+1}(b_0,\sigma_{t+1},\vec{\theta}_{t+1},\delta_{t+1}) \triangleq \delta^{sg*}_{t+1}$$

Therefore (A.3) holds true, proving the induction step. □

Proof Sketch of Theorem 3

The proof strategy is—similarly to that of Theorem 2—to show that, for all deterministic φ_t, for all b_0,δ_t, $Q_t(\sigma_t,\vec{o}_t,\delta_t) = Q_t(b_0,\varphi_t,\vec{\theta}_t,\delta_t)$, with $\vec{\theta}_t$ the joint AOH resulting from $\vec{o}_t$ and φ_t. In this case, to show the equality of (2.3) and (3.8) it is necessary to additionally show that the immediate reward terms (2.1), (3.5), and observation probability terms (2.2), (3.6) are equal. This requires showing that, for all $\vec{o}_t$, for the $\vec{\theta}_t$ resulting from $\vec{o}_t$ and φ_t, $\Pr(s_t|b_0,\vec{\theta}_t) = \sigma_t(s_t|\vec{o}_t)$. Moreover, to prove the induction step we will need to show that $\Pr(\vec{\theta}_t|b_0,\varphi_t) = \sigma_t(\vec{o}_t)$. We prove these additional requirements here, starting with the latter. For a deterministic φ_t we can write $\Pr(\vec{\theta}_t|b_0,\varphi_t) = \Pr(\vec{o}_t|b_0,\varphi_t)C(\vec{\theta}_t,\varphi_t)$, where $C(\vec{\theta}_t,\varphi_t)$ is a term that is 1 iff $\vec{\theta}_t$ is consistent with φ_t. Clearly, since $\sigma_t(\vec{o}_t) \triangleq \Pr(\vec{o}_t|b_0,\varphi_t)$ and since (for $\vec{\theta}_t$ resulting from $\vec{o}_t$) $C(\vec{\theta}_t,\varphi_t) = 1$ we can conclude $\Pr(\vec{\theta}_t|b_0,\varphi_t) = \sigma_t(\vec{o}_t)$. Using this result, we can write the joint distribution $\Pr(s_t,\vec{\theta}_t|b_0,\varphi_t)$

$$\Pr(s_t|b_0,\vec{\theta}_t)\Pr(\vec{o}_t|b_0,\varphi_t)C(\vec{\theta}_t,\varphi_t) = \sigma_t(s_t,\vec{o}_t)C(\vec{\theta}_t,\varphi_t),$$

from which we can deduce that $\Pr(s_t|b_0,\vec{\theta}_t)$

$$= \frac{\sigma_t(s_t,\vec{o}_t)C(\vec{\theta}_t,\varphi_t)}{\Pr(\vec{o}_t|b_0,\varphi_t)C(\vec{\theta}_t,\varphi_t)} = \frac{\sigma_t(s_t,\vec{o}_t)}{\sigma_t(\vec{o}_t)} = \sigma_t(s_t|\vec{o}_t),$$

for all $\vec{\theta}_t$ consistent with φ_t. Given these results, the remainder of the proof follows the proof of Theorem 2. □

References

[Amato *et al.*, 2009] Christopher Amato, Jilles S. Dibangoye, and Shlomo Zilberstein. Incremental policy generation for finite-horizon DEC-POMDPs. In *Proc. of the Int. Conference on Automated Planning and Scheduling*, pages 2–9, 2009.

[Becker *et al.*, 2003] Raphen Becker, Shlomo Zilberstein, Victor Lesser, and Claudia V. Goldman. Transition-independent decentralized Markov decision processes. In *Proc. of the Int. Conference on Autonomous Agents and Multi Agent Systems*, pages 41–48, 2003.

[Bernstein *et al.*, 2002] Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[Bertsekas, 2005] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, 3rd edition, 2005.

[Boularias and Chaib-draa, 2008] Abdeslam Boularias and Brahim Chaib-draa. Exact dynamic programming for decentralized POMDPs with lossless policy compression. In *Proc. of the Int. Conference on Automated Planning and Scheduling*, 2008.

[Boutilier, 1996] Craig Boutilier. Planning, learning and coordination in multiagent decision processes. In *Proc. of the 6th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 195–210, 1996.

[Dibangoye *et al.*, 2012] Jilles S. Dibangoye, Christopher Amato, and Arnaud Doniec. Scaling up decentralized MDPs through heuristic search. In *Proc. of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 217–226, 2012.

[Dibangoye *et al.*, 2013] Jilles S. Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. In *Proc. of the Int. Joint Conference on Artificial Intelligence*, 2013. (To appear).

[Emery-Montemerlo *et al.*, 2004] Rosemary Emery-Montemerlo, Geoff Gordon, Jeff Schneider, and Sebastian Thrun. Approximate solutions for partially observable stochastic games with common payoffs. In *Proc. of the Int. Conference on Autonomous Agents and Multi Agent Systems*, pages 136–143, 2004.

[Hansen *et al.*, 2004] Eric A. Hansen, Daniel S. Bernstein, and Shlomo Zilberstein. Dynamic programming for partially observable stochastic games. In *Proc. of the National Conference on Artificial Intelligence*, pages 709–715, 2004.

[Kaelbling *et al.*, 1998] Leslie P. Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

[Kumar and Zilberstein, 2010] Akshat Kumar and Shlomo Zilberstein. Point-based backup for decentralized POMDPs: Complexity and new algorithms. In *Proc. of the Int. Conference on Autonomous Agents and Multi Agent Systems*, pages 1315–1322, 2010.

[Nair *et al.*, 2003] Ranjit Nair, Milind Tambe, Makoto Yokoo, David V. Pynadath, and Stacy Marsella. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *Proc. of the Int. Joint Conference on Artificial Intelligence*, pages 705–711, 2003.

[Oliehoek *et al.*, 2008a] Frans A. Oliehoek, Julian F.P. Kooi, and Nikos Vlassis. The cross-entropy method for policy search in decentralized POMDPs. *Informatica*, 32:341–357, 2008.

[Oliehoek *et al.*, 2008b] Frans A. Oliehoek, Matthijs T. J. Spaan, and Nikos Vlassis. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, 32:289–353, 2008.

[Oliehoek *et al.*, 2009] Frans A. Oliehoek, Shimon Whiteson, and Matthijs T. J. Spaan. Lossless clustering of histories in decentralized POMDPs. In *Proc. of the Eighth Int. Joint Conference on Autonomous Agents and Multiagent Systems*, pages 577–584, 2009.

[Oliehoek *et al.*, 2013] Frans A. Oliehoek, Matthijs T. J. Spaan, Christopher Amato, and Shimon Whiteson. Incremental clustering and expansion for faster optimal planning in decentralized POMDPs. *Journal of Artificial Intelligence Research*, 46:449–509, 2013.

[Oliehoek, 2010] Frans A. Oliehoek. *Value-Based Planning for Teams of Agents in Stochastic Partially Observable Environments*. PhD thesis, Informatics Institute, University of Amsterdam, 2010.

[Oliehoek, 2012] Frans A. Oliehoek. Decentralized POMDPs. In Marco Wiering and Martijn van Otterlo, editors, *Reinforcement Learning: State of the Art*, pages 471–503. Springer Berlin Heidelberg, 2012.

[Ooi and Wornell, 1996] James M. Ooi and Gregory W. Wornell. Decentralized control of a multiple access broadcast channel: Performance bounds. In *Proc. of the 35th Conference on Decision and Control*, pages 293–298, 1996.

[Puterman, 1994] Martin L. Puterman. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[Seuken and Zilberstein, 2008] Sven Seuken and Shlomo Zilberstein. Formal models and algorithms for decentralized decision making under uncertainty. *Journal of Autonomous Agents and Multi-Agent Systems*, 17(2):190–250, 2008.

[Spaan, 2012] Matthijs T. J. Spaan. Partially observable Markov decision processes. In Marco Wiering and Martijn van Otterlo, editors, *Reinforcement Learning: State of the Art*, pages 387–414. Springer Berlin Heidelberg, 2012.

[Szer *et al.*, 2005] Daniel Szer, François Charpillet, and Shlomo Zilberstein. MAA*: A heuristic search algorithm for solving decentralized POMDPs. In *Proc. of Uncertainty in Artificial Intelligence*, pages 576–583, 2005.

[Wu *et al.*, 2010] Feng Wu, Shlomo Zilberstein, and Xiaoping Chen. Point-based policy generation for decentralized POMDPs. In *Proc. of the Int. Conference on Autonomous Agents and Multi Agent Systems*, pages 1307–1314, 2010.