Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Structural Results for Cooperative Decentralized Control Models

Jilles S. Dibangoye
Inria — CITI
Lyon, France
jilles.dibangoye@inria.fr

Olivier Buffet
Inria
Nancy, France
olivier.buffet@inria.fr

Olivier Simonin
Inria — CITI
Lyon, France
olivier.simonin@insa-lyon.fr

Abstract

The intractability in cooperative, decentralized control models is mainly due to prohibitive memory requirements in both optimal policies and value functions. The complexity analysis has emerged as the standard method to estimating the memory needed for solving a given computational problem, but complexity results may be somewhat limited. This paper introduces a general methodology—structural analysis—for the design of optimality-preserving concise policies and value functions, which will eventually lead to the development of efficient theory and algorithms. For the first time, we show that memory requirements for policies and value functions may be asymmetric, resulting in cooperative, decentralized control models with exponential reductions in memory requirements.

1 Introduction

Decentralized partially observable Markov decision processes (Dec-POMDPs) have emerged as the standard framework for sequential decentralized decision-making [Radner, 1962; Yoshikawa and Kobayashi, 1978; Bernstein *et al.*, 2002]. This general model involves multiple agents with different observations that cooperate to achieve a common objective, but cannot communicate with one another. Unfortunately, its worst-case complexity has limited its applicability. The finite-horizon case is in NEXP [Bernstein *et al.*, 2002]. The infinite-horizon case is undecidable, and ϵ-approximations remain intractable [Rabinovich *et al.*, 2003]. These negative complexity results are mainly due to the exponential growth in the size of both optimal policy and value function spaces with time. This results in limited scalability and applicability [Hansen *et al.*, 2004; Szer *et al.*, 2005; Oliehoek *et al.*, 2008; Bernstein *et al.*, 2009].

To allow further scalability and applicability, much attention has been devoted to cooperative decentralized control models with restrictive assumptions [Goldman and Zilberstein, 2004; Becker *et al.*, 2004; Nair *et al.*, 2005; Melo and Veloso, 2011]. These assumptions concern restrictions on both the dynamics and the rewards. In particular, [Goldman and Zilberstein, 2004] demonstrated that the complexity goes down from NEXP to NP when agents influence one another only through rewards, and are otherwise fully independent. Unfortunately, not all restrictions result in scalability gains, let alone complexity drops. So far, the asymptotic complexity analysis suggests that memory reductions due to restrictions on rewards are unlikely, if not impossible [Bernstein *et al.*, 2002; Goldman and Zilberstein, 2004; Allen and Zilberstein, 2009]. This negative complexity has limited the attractiveness of models involving reward restrictions.

This paper introduces a general methodology referred to as *structural analysis*, which helps designing optimality-preserving concise policies and value functions for the cooperative decentralized control problem at hand. For the first time, we show that memory requirements for policies and value functions may be asymmetric, resulting in significant memory reduction on certain models. Another novel and important result is the proof that under mild conditions on rewards, the optimal value function consists of linear functions from hidden states to reals. To allow a wide applicability, the structural analysis builds upon a recent theory to solving Dec-POMDPs by recasting them as a continuous-state deterministic Markov decision process with a piecewise-linear and convex optimal value function [Dibangoye *et al.*, 2013a]. Overall, it seems like the asymptotic complexity analysis provides a useful hierarchy of problems, while the structural analysis is geared to guide the automatic characterization of optimality-preserving concise policies and value functions, which will eventually lead to the development of scalable, applicable and widely adaptable theory and algorithms.

The remainder of this paper is organized as follows. Section 2 focuses on Dec-POMDPs and the recent approach to solving them. Next, we introduce the criteria we rely on to characterize optimality-preserving concise policies and value functions. Next, we discuss certain subclasses of Dec-POMDPs with mild conditions, demonstrating the usefulness of the structural analysis.

2 Cooperative Decentralized Control Models

We consider a *decentralized partially observable Markov decision process* (Dec-POMDP) involving n cooperative agents

$$M \equiv (S, (A_i), (Z_i), (P^{as}), (R^a), \zeta^0).$$

Here, S is a finite set of hidden states; A_i is a finite set of private actions of agent i; and Z_i is a finite set of private observations of agent i. Given any hidden state and joint action, s and a, the probability of each possible next hidden state and joint observation, s' and o, is $P^{as}(s', o)$. Similarly, given any current state and joint action, s and a, the reward is $R^a(s)$. This model is parametrized by ζ^0 the initial state distribution and T the planning horizon.

The goal of optimal decentralized decision-making is to find a joint policy $\pi = (\pi_i)$ maximizing the expected sum of rewards starting in initial distribution ζ^0. A joint policy $\pi = (\pi_i)$ is an n-tuple of agent policies π_i, one for each agent. Alternatively, a joint policy $\pi = (\pi^t)$ is a sequence of T joint decision rules, one for each time step. Similarly, an agent policy $\pi_i = (\pi_i^t)_{t\in\{0,1,\ldots,T-1\}}$ is a sequence of T agent decision rules. A step-t agent decision rule π_i^t is a mapping from length-t private (observation) history $\theta_i = (a_i^0, z_i^1, \ldots, a_i^{t-1}, z_i^t)$ to private actions $\pi_i^t(\theta_i) \in A_i$.

Example 1. *To illustrate the main result of this paper, consider a simple cooperative, decentralized control problem in which three agents must navigate to their terminal locations (denoted s_G) as soon as possible. Agents need to coordinate since they affect each other through both rewards and dynamics. In this world, possible actions for each agent include moving north, south, west, east and staying in the same place. Each time, agents take an action, they receive a reward and a noisy observation about the current state of the world. The penalty is* 0 *if the current state is s_G and* $+1$ *otherwise.*

For simplicity, in the remainder of the paper, we use majuscule bold symbols to denote random variables, and minuscule symbols to denote associated realizations — *e.g.,* for any action-observation histories, we use $\boldsymbol{\Theta}$ and θ to denote random variable and its realization, respectively.

2.1 Occupancy Markov Decision Process

The reformulation relies on a common assumption in decentralized decision-making that planning can be achieved in a centralized manner while still preserving decentralized execution [Szer *et al.*, 2005; Oliehoek *et al.*, 2008; 2013]. Building on this insight, [Dibangoye *et al.*, 2013a] introduced the concept of *occupancy state*, denoted ζ, to represent a distribution over hidden states and joint histories.

Occupancy states $\zeta^t(\boldsymbol{S}^t, \boldsymbol{\Theta}^t) \overset{\text{def}}{=} \mathbb{P}^{\zeta^0, \pi^{0:t-1}}(\boldsymbol{S}^t, \boldsymbol{\Theta}^t)$ are **sufficient statistics** of data $(\zeta^0, \pi^{0:t-1})$ for the characterization of optimal joint policy $\pi^{0:t-1}$ starting in initial occupancy ζ^0. Before proceeding any further, let us provide a formal definition of a sufficient statistics by [Fisher, 1922].

Definition 1. *A statistic $\chi(\boldsymbol{X})$ is* ***sufficient*** *for parameter $\boldsymbol{Y}$ precisely if the conditional probability of $\boldsymbol{Y}$, given the statistic $\chi(\boldsymbol{X})$, does not depend on data $\boldsymbol{X}$ —* i.e.,

$$P(\boldsymbol{Y}|\chi(\boldsymbol{X}), \boldsymbol{X}) = P(\boldsymbol{Y}|\chi(\boldsymbol{X})), \tag{1}$$

where $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\chi(\boldsymbol{X})$ are random variables.

Intuitively, a sufficient statistic captures all important information contained in data about a given parameter to be estimated, no further information can be obtained from the data. Notice, however, that sufficient statistics pertain to *data reduction*, not merely parameter estimation. As an example, in MDPs, the current joint observation is sufficient for the estimate of the current state. Once the current joint observation is known, no further information about the current state can be obtained from the history of an MDP — *i.e.,* $P(\boldsymbol{S}^t|\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^t) = P(\boldsymbol{S}^t|\boldsymbol{Z}^t)$ [Puterman, 1994].

Unfortunately, similarly to the joint policy space, in the worst case the occupancy-state space (denoted $\triangle$) grows exponentially as time goes on. In particular, the next-step occupancy states $\zeta^{t+1} = P^{\pi^t}\zeta^t$ is updated at each time-step t to incorporate the latest joint decision rule π^t:

$$\zeta^{t+1}(s', (\theta, a, z)) = \sum_{s\in S} \zeta^t(s, \theta) \cdot P^{as}(s', z), \tag{2}$$

Not surprisingly, in the worst case the space of joint decision rules (denoted Π) grows exponentially with time as well. From the Bayesian update-rule in (2), occupancy states describe a process that is Markov.

The immediate reward $R^{\pi^t}\zeta^t$ obtained by taking joint decision rule π^t at any occupancy state ζ^t is written:

$$R^{\pi^t}\zeta^t \overset{\text{def}}{=} \sum_{s,\theta} \zeta^t(s, \theta) \cdot R^{\pi^t(\theta)}(s). \tag{3}$$

Equations (2) and (3) directly lead to the definition of the *occupancy-state Markov decision process* (OMDP):

$$\hat{M} \equiv (\triangle, \Pi, (R^{\pi^t}), (P^{\pi^t}), \zeta^0)$$

of the following meaning: $\triangle$ is a continuous-state space; Π is a multidimensional action space; R^{π^t} is the immediate reward vector for joint decision rule π^t; and P^{π^t} the deterministic transition-matrix for joint decision rule π^t. The goal then is to find a joint policy π whose value function is the solution of the [Bellman, 1957] optimality equation: $\forall \zeta^t \in \triangle$,

$$V_M^t(\zeta^t) = \max_{\pi^t\in\Pi} R^{\pi^t}\zeta^t + V^{t+1}(P^{\pi^t}\zeta^t), \tag{4}$$

with boundary condition $V_M^t(\cdot) = 0$. Interestingly, [Dibangoye *et al.*, 2013a] demonstrate the optimal value function $(V_M^t)_{t\in\{0,1,\ldots,T-1\}}$ solution of (5) is a piecewise-linear and convex function of occupancy states. In other words, there exists a finite set of high-dimensional vectors (called γ-*vectors*), $(\Gamma^t = \{\gamma^t\})_{t\in\{0,1,\ldots,T-1\}}$, such that the optimal value at any occupancy state $\zeta^t \in \triangle$ is:

$$V_M^t(\zeta^t) = \max_{\gamma^t\in\Gamma^t} \gamma^t\zeta^t. \tag{5}$$

This property is particularly important as it allows the value function to generalize over the entire occupancy space. The step-t set of γ-vectors Γ^t can be built from the next-step set Γ^{t+1} using the [Bellman, 1957] backup operator, denoted $\mathbb{H}$. We shall use notation $\mathbb{H}(\zeta^t, \Gamma^{t+1}) = \gamma_*^t$ to denote the exact point-based backup:

$$\gamma_*^t = \operatorname*{arg\,max}_{\gamma^t = R^{\pi^t} + (P^{\pi^t})^\top\gamma^{t+1} \,:\, \pi^t\in\Pi, \gamma^{t+1}\in\Gamma^{t+1}} \gamma^t\zeta^t. \tag{6}$$

A repeated application of the point-based backup over reachable occupancy states ensures optimality of algorithms.

Algorithm 1: The OHSVI algorithm.

```
function OHSVI ($(L^t)_t, (U^t)_t$)
    while GAP($\zeta^0$) > 0 do EXPLORE ($\zeta^0$).

function GAP ($\zeta^t$)
    return $U^t(\zeta^t) - L^t(\zeta^t)$.

function EXPLORE ($\zeta^t$)
    if GAP($\zeta^t$) > 0 then
        action-selection given $U^{t+1}$ and $\zeta^t$ yields $\pi^t_*$.
        update upper-bound $U^t$ at occupancy state $\zeta^t$.
        EXPLORE ($P^{\pi^t_*}\zeta^t$) .
        update lower-bound $L^t$ at occupancy state $\zeta^t$.
```

2.2 Optimally Solving Dec-POMDPs as OMDPs

[Dibangoye *et al.*, 2013a] demonstrate an optimal joint policy for $\hat{M}$ is also optimal for M. Hence, we focus on optimally solving $\hat{M}$. Since $\hat{M}$ is essentially a continuous-state MDP with a piecewise-linear and convex optimal value function, methods for solving POMDPs can also solve $\hat{M}$.

In particular, [Dibangoye *et al.*, 2013a] extend the heuristic search value iteration (HSVI) algorithm for POMDPs [Smith and Simmons, 2004] to solving for OMDPs. The resulting OHSVI algorithm 1 is a trial-based algorithm, which proceeds by generating trajectories of occupancy states, starting at occupancy state ζ^0. It maintains both upper and lower bounds over the optimal value function, we denote (L^t) and (U^t), respectively. It guides exploration towards occupancy states that are more relevant to the upper bounds by greedily selecting joint decision rules with respect to the upper bounds, which reduces the difference between bounds at those points. The algorithm terminates when the gap at the initial occupancy state ζ^0 is zero. In such a case, the algorithm has converged.

The OHSVI algorithm has demonstrated impressive results on medium-sized problems from the literature of Dec-POMDPs [Dibangoye *et al.*, 2013a], but its scalability remains a serious challenge. There are two reasons for the limited scalability of methods to solving $\hat{M}$ due to the dimensionality of joint decision rules π^t, occupancy state ζ^t and γ-vectors γ^{t+1}. Computing ζ^t, $R^{\pi^t}\zeta^t$ and $P^{\pi^t}\zeta^t$ requires time complexities *polynomial* in the dimensionality of ζ^t, which itself grows exponentially with time. Even more importantly, the *exponential* complexity requirement of $\mathbb{H}(\zeta^t, \Gamma^{t+1})$ is the major source of the intractability of $\hat{M}$. This highlights the impetus for *automatic and general purpose methods* for designing optimality-preserving concise policies, occupancy states and γ-vectors. To date, determining structural results (*e.g.,* sufficient statistics) seems more like an art than a science [Goldman and Zilberstein, 2004; Dibangoye *et al.*, 2012; 2013a; Oliehoek, 2013]. We will distinguish between four categories of sufficient statistics ranging along two dimensions: *objective* and *perspective* dimensions.

3 Separation of Policy and Value Statistics

Along the objective dimension, this section investigates the separation of sufficient statistics for the estimation of expected cumulated rewards and those for the design of optimal joint policies — namely, **policy-** and **value-sufficient statistics**, respectively. Along the perspective dimension, we will differentiate between data provided from an agent to that of a centralized planner point of views — namely, **agent-** and **planner-centric** sufficient statistics.

3.1 Separation Principle

The goal of planning in cooperative, decentralized control models is to find an optimal joint policy — *i.e.,* the one that maximizes the long-term expected cumulated rewards. Since the design of optimality-preserving concise policies and the estimation of expected cumulated rewards are interconnected objectives, a common assumption is that both share the same sufficient statistic. But this common assumption is not always true. Notice that any agent policy depends on agent action-observation histories (in the worst case):

$$\pi_i(\boldsymbol{\Theta}_i) = \mathbb{P}_M^{\zeta^0,\pi_j}(\boldsymbol{A}_i|\boldsymbol{\Theta}_i),$$

where $\boldsymbol{A}_i$ and $\boldsymbol{\Theta}_i$ denote random variables associated to agent actions and histories. Instead, the immediate expected reward depends on agent action-observation histories and actions,

$$\mathbb{R}_M^{\zeta^0,\pi_j}(\boldsymbol{\Theta}_i, \boldsymbol{A}_i) = \mathbb{E}^{\zeta^0,\pi_j}\left[\boldsymbol{R}|\boldsymbol{\Theta}_i, \boldsymbol{A}_i\right],$$

where $\boldsymbol{R} = R^{\pi_j(\boldsymbol{\Theta}_j),\boldsymbol{A}_i}(\boldsymbol{S})$ denotes random variable associated to immediate reward agents received upon agent i taking action $\boldsymbol{A}_i$ in action-observation history $\boldsymbol{\Theta}_i$. Hence, while being interdependent, policy- and value-sufficient statistics are possibly different, as stated in the following.

Lemma 1 (Separation Principle)**.** *In cooperative, decentralized control models, the expected value $V^\pi_M(\zeta^0)$ of any arbitrary joint policy π starting at initial state ζ^0 depends on two possibly different sufficient statistics:*

$$V^\pi_M(\zeta^0) = \sum_{\theta_i, a_i} \mathbb{P}_M^{\zeta^0,\pi_j}(\chi(\theta_i)) \cdot \mathbb{P}_M^{\zeta^0,\pi_j}(a_i|\chi(\theta_i)) \cdot \mathbb{R}_M^{\zeta^0,\pi_j}(\chi'(\chi(\theta_i), a_i)),$$

where $\chi(\boldsymbol{\Theta}_i)$ and $\chi'(\chi(\boldsymbol{\Theta}_i), \boldsymbol{A}_i)$ denote policy- and value-sufficient statistics, respectively.

Lemma 1 provides a useful characterization of agent-centric policy- and value-sufficient statistics. It further suggests that value-sufficient statistics have a dimensionality lower than or equal to policy-sufficient statistics. This is a major discovery as, so far, we assumed both were identical. The immediate implication of this property is the ability to improve (1) the scalability of operations $P^{\pi^t}\zeta^t$, $R^{\pi^t}\zeta^t$ and even more importantly $\mathbb{H}(\zeta^t, \Gamma^{t+1})$; and (2) the estimation accuracy by enhancing the generalization between policy-sufficient statistics that share the same corresponding (lower-dimensional) value-sufficient statistics.

Example 2. *Clearly, example 1 is a three-agent goal-directed Dec-POMDP [Amato and Zilberstein, 2009; Goldman and Zilberstein, 2004]. In practice, [Amato and Zilberstein, 2009] used the complete history of actions and observations*

θ_i for each agent $i \in N$ as policy- and value-sufficient statistics in such a setting. However, the separation principle suggests a value-sufficient statistic can be made significantly different than a policy-sufficient statistic. Based on the reward function, one can indeed show that if the probability of being in s_G is p_{s_G} then the immediate expected reward will be $(1 - p_{s_G}) = 0 \times p_{s_G} + (+1) \times (1 - p_{s_G})$. Hence, the conditional probability of being in s_G, denoted $\mathbb{P}^{\zeta^0,\pi_j}(s_G|\theta_i, a_i)$, is a sufficient statistic for the estimation of immediate expected reward $\mathbb{R}^{\zeta^0,\pi_j}(\theta_i, a_i)$ — i.e., $\mathbb{R}^{\zeta^0,\pi_j}(\theta_i, a_i) = 1 - \mathbb{P}^{\zeta^0,\pi_j}(s_G|\theta_i, a_i)$. *Once the conditional probability $\mathbb{P}^{\zeta^0,\pi_j}(s_G|\theta_i, a_i)$ of being in s_G is known, no further information can be obtained from agent history and action, θ_i and a_i, respectively.*

4 Policy-Sufficient Statistics

The previous section demonstrated that policy- and value-sufficient statistics are possibly different. This section introduces criteria that provide a convenient characterization of concise, if not minimal, policy-sufficient statistics required to solving M (respectively $\hat{M}$).

4.1 Markov Property

Here, we define the property, that is a necessary condition for the characterization of policy-sufficient statistics.

Criterion 1. *For any arbitrary cooperative, decentralized control model M, agent i and teammates j, a statistic $\chi(\mathbf{\Theta}_i)$ satisfies the Markov property if, and only if, the future statistics depend solely on the current one:*

$$\mathbb{P}_M^{\zeta^0,\pi_j}(\chi(\mathbf{\Theta}'_i)|\mathbf{\Theta}_i, \mathbf{A}_i, \mathbf{Z}_i) = \mathbb{P}_M^{\zeta^0,\pi_j}(\chi(\mathbf{\Theta}'_i)|\chi(\mathbf{\Theta}_i), \mathbf{A}_i, \mathbf{Z}_i),$$

where $\mathbf{\Theta}'_i$ denotes the next-step random variable of $\mathbf{\Theta}_i$.

Informally, a statistic satisfies the policy criterion if, and only if, it contains all information about its offsprings, no further information can be obtained from the complete history. It is worth noticing that the Markov property is not sufficient for the design of an optimal agent policy. But all policy-sufficient statistics satisfies the policy criterion. So, one can use this criterion to check whether or not a statistic is a good candidate policy-sufficient statistic, as discussed below.

Example 3. *Consider our running example 1. We showed in example 2 that a concise value-sufficient statistic is the probability of being in the terminal state. Using policy criterion, one can show that this statistic cannot serve as a policy-sufficient statistic. This is mainly because the current probability of being in the terminal state discarded important information for an accurate estimation of the next-step probability of being in the terminal state. As a consequence this statistic does not satisfy the policy criterion, and this cannot be a policy-sufficient statistic. Examples of statistics that are Markovian include: the complete action-observation history $\chi(\theta_i) = \theta_i$; or the current observation $\chi(\theta_i) = z_i$, etc. Markov property trivially holds when using complete action-observation histories. Similarly, it holds when using the current observation, since the next-step observation does not depend on the complete history.*

4.2 Value-Preserving Property

Here, we define the value-preserving property, that is a necessary condition for the characterization of agent-centric policy- and value-sufficient statistics.

Criterion 2. *For any arbitrary cooperative, decentralized control model M, agent i and teammates j, a statistic $\chi(\mathbf{\Theta}_i)$ satisfies the value-preserving property if, and only if, there exists agent-centric value-sufficient statistic $\chi'(\chi(\mathbf{\Theta}_i), \mathbf{A}_i)$:*

$$\mathbb{R}_M^{\zeta^0,\pi_j}(\mathbf{\Theta}_i, \mathbf{A}_i) = \mathbb{R}_M^{\zeta^0,\pi_j}(\chi'(\chi(\mathbf{\Theta}_i), \mathbf{A}_i)),$$

Intuitively, this criterion requires statistic $\chi(\mathbf{\Theta}_i)$ along with $\mathbf{A}_i$ to be an agent-centric value-sufficient statistic themselves. Similarly to the Markov property, statistics that are value-preserving need not be policy-sufficient. Yet, all policy-sufficient statistics need to be value-preserving in order the preserve ability of being policy-sufficient, as illustrated next.

Example 4. *Back to our running example 1, we know from example 3 that the current observation z_i is Markovian and the probability of being in the terminal state $\mathbb{P}^{\zeta^0,\pi_j}(s_G|\theta_i, a_i)$ describes the minimal value-sufficient statistic. A natural question then is whether the current observation is value-preserving. In other words, does the current observation describes a sufficient statistic for the probability of being in the terminal state starting given the complete action-observation history? The answer is negative since the history contains additional information than are crucial in estimating this probability. Hence, the current observation is Markovian but it is not value-preserving, which precludes z_i from being a policy-sufficient statistic.*

4.3 Characterizing Policy-Sufficient Statistics

We are now ready to state and provide a convenient characterization of agent- and planner-centric policy-sufficient statistics based on Markov and value-preserving properties.

Theorem 1. *For any arbitrary cooperative, decentralized control model M, agent i and teammates j, a statistic $\chi(\mathbf{\Theta}_i)$ is an agent-centric policy-sufficient statistic if it satisfies both Markov and value-preserving properties.*

An immediate implication of this theorem is a useful characterization of a minimal sufficient statistic for policies. A minimal sufficient statistic for policies is the lowest-dimensional statistic that satisfies the transition criterion. This characterization will eventually lead to the design of efficient automatic methods for identifying minimal sufficient statistics for policies of any decentralized control application at hand, as demonstrated in Section 6. Theorem 1 focuses on policy-sufficient statistics from each agent's perspective. Another important implication pertains to the design of the minimal policy-sufficient statistic from the centralized planner's perspective. This is still an open problem, but the following provides a convenient way to build good planner-centric policy-sufficient statistics.

Corollary 1. *For any arbitrary cooperative, decentralized control model M, joint policy π and initial state ζ^0, a statistic $\chi(\zeta^0, \pi)$ is a planner-centric policy-sufficient statistic if, and only if, it describes a probability distribution* *over hidden states and agent-centric policy-sufficient statistics* $(\chi_1(\mathbf{\Theta}_1), \cdots, \chi_N(\mathbf{\Theta}_N))$*:*

$$\chi(\zeta^0, \pi) = \mathbb{P}^{\zeta^0,\pi}(\chi_1(\mathbf{\Theta}_1), \cdots, \chi_N(\mathbf{\Theta}_N)),$$

where $\chi_i(\mathbf{\Theta}_i)$ *denotes a policy-sufficient statistic of agent i.*

The total information available to the centralized planner is joint policy π and initial state ζ^0. Hence, a planner-centric policy-sufficient statistic summarizes data π and ζ^0. The primary goal of Theorem 1 and Corollary 1 is to provide a convenient way to check whether a statistic is actually a policy-sufficient statistic, as illustrated below.

Example 5. *Consider statistics* $\chi_i(\mathbf{\Theta}_i) = \mathbb{P}_M^{\zeta^0,\pi_j}(\mathbf{S}, \mathbf{\Theta}_j | \mathbf{\Theta}_i)$ *describing the conditional probability of hidden states and other agent histories given history of agent i. One can show that statistic* $\chi_i(\mathbf{\Theta}_i)$ *is Markovian, since* $\mathbb{P}_M^{\zeta^0,\pi_j}(\mathbf{S}', \mathbf{\Theta}'_j | \mathbf{\Theta}_i, \mathbf{A}_i, \mathbf{Z}_i) = \mathbb{P}_M^{\zeta^0,\pi_j}(\mathbf{S}', \mathbf{\Theta}'_j | \chi_i(\mathbf{\Theta}_i), \mathbf{A}_i, \mathbf{Z}_i)$*. Moreover, statistic* $\mathbb{P}_M^{\zeta^0,\pi_j}(\mathbf{S}, \mathbf{\Theta}_j | \mathbf{\Theta}_i)$ *is value-preserving because* $\mathbb{P}_M^{\zeta^0,\pi_j}(s_G | \mathbf{\Theta}_i, \mathbf{A}_i)$ *is obtained by marginalizing out over non-terminal states and other agent histories. Hence,* $\chi_i(\mathbf{\Theta}_i)$ *is a policy-sufficient statistic. Similarly, one can demonstrate that an occupancy state* $\zeta(\mathbf{S}, \mathbf{\Theta}) = \mathbb{P}^{\zeta^0,\pi}(\mathbf{S}, \mathbf{\Theta})$ *is a policy-sufficient statistic from the centralized planner's perspective if we let* $\chi_i(\mathbf{\Theta}_i) = \mathbf{\Theta}_i$ *and apply Corollary 1.*

5 Value-Sufficient Statistics

This section introduces criteria that provide a convenient characterization of concise, if not minimal, agent- and planner-centric value-sufficient statistics required to solving M (respectively $\hat{M}$). We distinguish between models in which agents have independent rewards to those in which agents influence each other through rewards. The following lemma builds upon the definition of a sufficient statistic to prove that if agents have independent rewards, then the central planner requires only an agent-centric value-sufficient statistic.

Lemma 2. *For any arbitrary cooperative, decentralized control model M, agent i and teammates j, a statistic* $\chi'(\chi(\mathbf{\Theta}_i), \mathbf{A}_i)$ *is agent-centric value-sufficient if, and only if, there exists an agent-centric policy-sufficient statistic* $\chi(\mathbf{\Theta}_i)$ *such that:* $\mathbb{R}_M^{\zeta^0,\pi_j}(\mathbf{\Theta}_i, \mathbf{A}_i) = \mathbb{R}_M^{\zeta^0,\pi_j}(\chi'(\chi(\mathbf{\Theta}_i), \mathbf{A}_i))$.

Intuitively, this lemma states that under independent rewards, the centralized planner needs to maintain one value function for each agent, hence only agent-centric value-sufficient statistics are required. Section 6.3 provides structural results for an important subclass of models with independent rewards. Next, we consider a much broader settings in which agents have a common reward function.

Lemma 3. *For any arbitrary cooperative, decentralized control model M, joint policy* π *and initial state* ζ^0*, a statistic* $\chi(\zeta^0, \pi)$ *is planner-centric value-sufficient if, and only if, it is sufficient for any agent-centric value-sufficient statistic* $\chi'_i(\chi_i(\mathbf{\Theta}_i), \mathbf{A}_i)$*.*

Intuitively, this lemma states that a statistic that is sufficient for agent-centric value-sufficient statistic for all agents is a planner-centric value-sufficient statistic.

6 New Structural Results

This section presents a general automatic method, referred to as a *structural analysis*, which eases the design of concise, if not minimal, sufficient statistics for policies and γ-vectors. We apply the methodology on a selection of decentralized control models, on which asymptotic complexity analysis suggests significant memory gains were unlikely, if not impossible. Surprisingly, our structural analysis demonstrates that significant memory gains can be achieved.

6.1 Structural Analysis

Algorithm 2 describes the structural analysis method. It starts with the selection of a subset of possible statistics $\tilde{\Omega}$ from the unknown set of all possible statistics Ω. Each statistic $\tilde{O}$ from the subset is an *observable random variable* of the decentralized control application at hand. The methodology proceeds by testing statistics in the order of increasing dimensionality. The test consists in checking whether the given sufficient criterion C, depending on the type of sufficient statistic we target, is satisfied. The methodology terminates when there are no more statistics to be tested, or one statistic passed the test; in the latter case, a concise, if not minimal, sufficient statistic has been found.

Algorithm 2: The Structural Analysis.

```
function SA($\Omega$,$C$)
    (1) Select subset $\tilde{\Omega} \subset \Omega$ of statistics $\tilde{O}$.
    (2) Extract the lower-dimensional one $\tilde{O}^*$.
    (3) If $\tilde{O}^*$ fails on criterion $C$, go back to (2).
    (4) Otherwise return sufficient statistic $\tilde{O}^*$.
```

The selection of a representative subset of statistics is not a trivial task as many subsets may end up with no sufficient statistics. As an example, one can select sufficient statistics in the space of histories as did [Dibangoye *et al.*, 2013a] with success. Though preliminary, this method enhances automatic characterization of concise sufficient statistics, policies and value functions (via γ-vectors). Of course, the method could also help verifying previously exhibited statistics were indeed sufficient [Goldman and Zilberstein, 2004; Becker *et al.*, 2004; Nair *et al.*, 2005; Oliehoek, 2013; Dibangoye *et al.*, 2012; 2013b; 2014].

In demonstrating the practical usefulness of the methodology, the following exhibits sufficient statistics for general decentralized control models under mild restrictions on the system rewards.

6.2 Action-Independent Rewards

We start with decentralized control models with action-independent rewards. The system rewards (R^a) are *action-independent* iff, there exists a function R, for any arbitrary joint action a, such that $R^a = R$. This mild condition appears in a number of important applications ranging from robotics (*e.g.,* collective information gathering tasks) to network security (*e.g.,* intrusion detection problems) [Blin *et al.*, 2006]. In such a setting, the structural analysis reveals that the *empty history* is the sufficient statistic of any private history for optimality-preserving and concise γ-vectors, which leads to the following lemma.

Lemma 4. *Under action-independent rewards, the distribution over hidden states $\mathbb{P}^{\zeta_0,\pi}(\boldsymbol{S})$ is a planner-centric value-sufficient statistic.*

From the above lemma, it directly follows that γ-vectors are mappings from hidden states to reals, as stated in the following theorem.

Theorem 2. *Under action-independent rewards, optimality-preserving concise γ-vectors are $|S|$-dimensional real vectors: $\tilde{\gamma}^{\tilde{\pi}^{t:T-1}}(\boldsymbol{S}) = \boldsymbol{E}_{\tilde{\pi}^{t:T-1}}[\sum_{\tau=t}^{T-1} R(\boldsymbol{S})]$.*

In general, γ-vectors $\gamma^{\pi^{t:T-1}}$ have dimensionality exponential with time, $\gamma^{\pi^{t:T-1}}(\boldsymbol{S}, \boldsymbol{\Theta}) = \boldsymbol{E}_{\pi^{t:T-1}}[\sum_{\tau=t}^{T-1} R^{\pi^\tau(\boldsymbol{\Theta})}(\boldsymbol{S})]$.

Theorem 2 reveals that the resources needed to accurately maintain value functions through γ-vectors is exponentially smaller than that of the general setting. Even more importantly, under action-independent rewards, the dimensionality of γ-vectors is bounded by the number of hidden states. Notice that this result goes beyond the scope of Dec-POMDPs. It can apply on all subclasses of Dec-POMDPs with action-independent rewards — *e.g.,* POMDPs [Kaelbling *et al.*, 1998] with action-independent rewards. Besides the enormous memory savings, this property can significantly enhance generalization and evaluation of the value function, as previously discussed.

6.3 Additive Rewards

Another important condition pertains to the additivity of agent rewards. The system rewards R are *fully additive* iff there exists n functions $R_1, R_2, \ldots, R_n$, such that: $\forall s \in S, a = (a_i) \in A$, $R^a(s) = R_1^{a_1}(s) + R_2^{a_2}(s) + \ldots + R_n^{a_n}(s)$. A notable example of this family is the foraging problem, in which agents collaborate to search for disseminated resources. The reward is often defined as the sum of resources agents transport back to home.

In such a setting, the structural analysis reveals that a value-sufficient statistic $\chi_i(\boldsymbol{\Theta}_i)$ of any private history $\boldsymbol{\Theta}_i$ for optimality-preserving and concise γ-vectors is both a *state-distribution* $\mathbb{P}^{\zeta_0,\pi_j}(\boldsymbol{S}|\boldsymbol{\Theta}_i, \boldsymbol{A}_i)$ and independent of histories of other agents. This leads to the following lemma.

Lemma 5. *Under additive rewards, state-distribution $\mathbb{P}^{\zeta_0,\pi_j}(\boldsymbol{S}|\boldsymbol{\Theta}_i, \boldsymbol{A}_i)$ is an agent-centric value-sufficient statistic.*

Another important result pertaining to the characterization of optimality-preserving concise value functions follows from the above lemma.

Theorem 3. *Under additive rewards, optimality-preserving concise value functions consist of tuples of private value functions $(\Gamma_i^t)_{i\in\{1,2,\ldots,n\},t\in\{0,1,\ldots,T\}}$, one for each time step and agent. Even more importantly, each γ-vector $\tilde{\gamma}_{\boldsymbol{\Theta}_i}^{\tilde{\pi}^{t:T-1}} \in \Gamma_i^t$ is an $|S|$-dimensional real vector: $\tilde{\gamma}_{\boldsymbol{\Theta}_i}^{\tilde{\pi}^{t:T-1}}(\boldsymbol{S}) = \boldsymbol{E}_{\tilde{\pi}^{t:T-1}}\left[\sum_{\tau=t}^{T-1} R_i^{\pi_i^\tau(\boldsymbol{\Theta}_i)}(\boldsymbol{S})\right]$.*

In general, γ-vectors $\gamma_{\boldsymbol{\Theta}_i}^{\pi^{t:T-1}}$ have dimensionality exponential with time, $\gamma_{\boldsymbol{\Theta}_i}^{\pi^{t:T-1}}(\boldsymbol{S}, \boldsymbol{\Theta}_j) = \boldsymbol{E}_{\pi^{t:T-1}}[\sum_{\tau=t}^{T-1} R^{\pi^\tau(\boldsymbol{\Theta})}(\boldsymbol{S})]$. This theorem shows that, in models with additive rewards, each agent holds its private value function although they jointly optimize the sum of these functions. Even more importantly, each optimality-preserving concise private value function is represented by a set of $|S|$-dimensional real vectors. Nonetheless, agents cannot choose their private actions on their own as value-sufficient statistics also depend on the choices of the other agents.

7 Discussion and Conclusion

This paper introduces a general methodology, referred to as structural analysis, as a means of characterizing a concise if not minimal, model for the applications at hand — demonstrating adaptability. Under mild conditions, the structural analysis of optimality-preserving concise policies and value functions often results in exponential reductions in memory and time requirements — demonstrating scalability gains. We developed the structural analysis as an analysis tool on top of occupancy-state Markov decision processes, a general model to represent a large variety of Markov models — demonstrating the wide applicability.

The structural analysis aims at characterizing concise sufficient statistics for decentralized control models. The idea of describing minimal sufficient statistics is not new. It can be traced back to Fisher-Neyman factorization [Fisher, 1922], since then numerous authors applied this method to derive sufficient statistics for optimal decision-making: [Smallwood and Sondik, 1973] demonstrate the belief state (distribution over hidden states) is a sufficient statistic of the history of agent for characterizing an optimal policy in general POMDPs; similarly, [Dibangoye *et al.*, 2012] show that the distribution over states is a sufficient statistic of the initial belief and past joint policy for the design of concise and optimal policies in transitional and observational Dec-MDPs; and [Dibangoye *et al.*, 2013a; Oliehoek, 2013] prove that the distribution over hidden states and joint histories is a sufficient statistic for initial belief and past policies; similar results hold for ND-POMDPs [Nair *et al.*, 2005; Dibangoye *et al.*, 2014]. To date, however, determining sufficient statistics seems more like an art than a science, explaining the importance of the structural analysis. Overall, it looks like the asymptotic complexity analysis provides a useful hierarchy of problems, while the structural analysis is geared to guide the automatic characterization of optimality-preserving concise policies and value functions, which will eventually lead to the development of scalable, applicable and widely adaptable theory and algorithms.

References

[Allen and Zilberstein, 2009] Martin Allen and Shlomo Zilberstein. Complexity of decentralized control: Special cases. In *Proceedings of the Twenty-Third Neural Information Processing Systems Conference*, pages 19–27, Vancouver, British Columbia, Canada, 2009.

[Amato and Zilberstein, 2009] Christopher Amato and Shlomo Zilberstein. Achieving goals in decentralized pomdps. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '09, pages 593–600, 2009.

[Becker *et al.*, 2004] Raphen Becker, Shlomo Zilberstein, Victor R. Lesser, and Claudia V. Goldman. Solving transition independent decentralized Markov decision processes. *Journal of Artificial Intelligence Research*, 22:423–455, 2004.

[Bellman, 1957] Richard E. Bellman. *Dynamic Programming*. Dover Publications, Incorporated, 1957.

[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), 2002.

[Bernstein *et al.*, 2009] Daniel S. Bernstein, Christopher Amato, Eric A. Hansen, and Shlomo Zilberstein. Policy iteration for decentralized control of Markov decision processes. *Journal of Artificial Intelligence Research*, 34:89–132, 2009.

[Blin *et al.*, 2006] Lelia Blin, Pierre Fraigniaud, Nicolas Nisse, and Sandrine Vial. Distributed chasing of network intruders. In *Proceedings of the 13th Colloquium on Structural Information and Communication Complexity (SIROCCO)*, 2006.

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

[Dibangoye *et al.*, 2013a] Jilles S. Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. In *IJCAI*, 2013.

[Dibangoye *et al.*, 2013b] Jilles S. Dibangoye, Christopher Amato, Arnaud Doniec, and François Charpillet. Producing efficient error-bounded solutions for transition independent decentralized MDPs. In *Proceedings of the Twelfth International Conference on Autonomous Agents and Multiagent Systems*, 2013.

[Dibangoye *et al.*, 2014] Jilles S. Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Exploiting separability in multi-agent planning with continuous-state MDPs. In *Proceedings of the Thirteenth International Conference on Autonomous Agents and Multiagent Systems*, pages 1281–1288, 2014.

[Fisher, 1922] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222:309–68, 1922.

[Goldman and Zilberstein, 2004] Claudia V. Goldman and Shlomo Zilberstein. Decentralized control of cooperative systems: Categorization and complexity analysis. *J. Artif. Int. Res.*, 22(1):143–174, November 2004.

[Hansen *et al.*, 2004] Eric A. Hansen, Daniel S. Bernstein, and Shlomo Zilberstein. Dynamic programming for partially observable stochastic games. In *Proceedings of the Nineteenth 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.

[Melo and Veloso, 2011] Francisco S. Melo and Manuela M. Veloso. Decentralized MDPs with sparse interactions. *Artificial Intelligence*, 175(11):1757–1789, 2011.

[Nair *et al.*, 2005] Ranjit Nair, Pradeep Varakantham, Milind Tambe, and Makoto Yokoo. Networked distributed POMDPs: A synthesis of distributed constraint optimization and POMDPs. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*, pages 133–139, 2005.

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

[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 Dec-POMDPs. *Journal of Artificial Intelligence Research*, 46:449–509, 2013.

[Oliehoek, 2013] Frans A. Oliehoek. Sufficient plan-time statistics for decentralized POMDPs. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2013.

[Puterman, 1994] Matrin L. Puterman. *Markov Decision Processes, Discrete Stochastic Dynamic Programming*. Wiley-Interscience, Hoboken, New Jersey, 1994.

[Rabinovich *et al.*, 2003] Zinovi Rabinovich, Claudia V. Goldman, and Jeffrey S. Rosenschein. The complexity of multiagent systems: the price of silence. In *Proceedings of the Second International Conference on Autonomous Agents and Multiagent Systems*, pages 1102–1103, 2003.

[Radner, 1962] R. Radner. Team decision problems. *Ann. Math. Statist.*, 33(3):857–881, 09 1962.

[Smallwood and Sondik, 1973] Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable Markov decision processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

[Smith and Simmons, 2004] Trey Smith and Reid Simmons. Heuristic search value iteration for POMDPs. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pages 520–527, Arlington, Virginia, United States, 2004.

[Szer *et al.*, 2005] Daniel Szer, François Charpillet, and Shlomo Zilberstein. MAA*: A heuristic search algorithm for solving decentralized POMDPs. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 568–576, 2005.

[Yoshikawa and Kobayashi, 1978] Tsuneo Yoshikawa and Hiroaki Kobayashi. Separation of estimation and control for decentralized stochastic control systems. *Automatica*, 14(6):623–628, 1978.