COLLABORATION VS. OBSERVATION: AN EMPIRICAL STUDY IN MULTIAGENT PLANNING UNDER RESOURCE CONSTRAINT

Emmanuel Benazera
LAAS-CNRS, Université de Toulouse, 7 av. du Colonel Roche, 31077 Cedex 4, Toulouse, France
ebenazer@laas.fr

Keywords: Multiagent planning, Decentralized Markov Decision Processes, Resource constraints, Empirical study.

Abstract: A team of robots and an exploratory mission are modeled as a multiagent planning problem in a decentralized decision theoretical framework. In this application domain, agents are constrained by resources such as their remaining battery power. In this context, there is an intrinsic relation between collaboration, computation and the need for the agents to observe their resource level. This paper reports on an empirical study of this relationship.

1 Introduction

Among formal models for the control of collaborative multiagent systems, decision-theoretic planning has focused on Markov Decision Problems (MDPs) (Boutilier et al., 1999). There exist several multiagent extensions to the MDP framework. The decentralized MDP framework (DEC-MDP) represents agents whose knowledge is partial and that act relatively to their local models of the world. An even more general framework is the decentralized partially observable MDPs (DEC-POMDPs) where individual agents do not fully observe their portions of the world (Bernstein et al., 2002). The multiagent control problem where agents have both stochastic actions and internal continuous state-spaces can be represented as decentralized hybrid MDP (DEC-HMDP). By hybrid it is meant that it involves both continuous and discrete variables. DEC-HMDPs are related to DEC-POMDPs with the difference that the former decide in the observation space whereas the later decide in the belief space.

Many real-world planning applications that involve teams of agents can be modeled as DEC-HMDPs. Our application domain is that of teams of exploratory robots. The interest in building and controlling teams of agents is motivated by an expected increase in both the overall capabilities and the robustness of the system. In our application domain, the continuous state-space of an agent represents its available level of resource, such as battery power and remaining time for the mission. A consequence of the resource constrained nature of the agents is that each of them is rarely able to achieve all the tasks in a mission. It follows that agents have to pay close attention to their resource levels before taking decisions such as achieving one task or the other. In this context, what the designers of multiagent robotic missions and systems may not foreknow is that there is an intrinsic relation between collaboration, computation and the need for the agents to observe their local world. In other words, the amount by which the agents constrain each others, that is the level of collaboration allowed, affects the need for observation of the agent local worlds, as well as the difficulty of computing an optimal controller for the team. This has potential consequences on the design of both missions and robots themselves.

Modern algorithms allow solving DEC-HMDPs with a small number of agents (Becker et al., 2004; Petrik and Zilberstein, 2007). This paper's focus is not on the computational techniques for DEC-HMDPs but rather on the form of their solutions and the light they shed on the relation between collaboration, computation and the need for agents to observe their local world. Through simulations and tests, empirical evidences are given of the structured relationship between collaboration, computation and

the need for observation. The first half of the paper lay the required background for understanding the resource constrained DEC-HMDP framework. The second half reports on a series of experiments and empirically establishes a few useful facts that connect collaboration, computation and the need for observation.

2 Background

We give with a brief overview of decentralized hybrid Markov decision processes, and its resource constrained variant. The reader interested in more detailed description is referred to (Becker et al., 2004).

2.1 Decentralized hybrid Markov decision process and resource constraints

A team of m agents is modeled as a DEC-HMDP. It is defined by a tuple (N,X,A,ϕ,R,N_0). N is a set of m discrete variables $N_i, i = 1,\cdots,m$, that refer to each agent i discrete component, and n_i denotes a discrete state in N_i. $X = \bigotimes_{i=1}^{m} X_i$ is the continuous state space, and x_i denotes a continuous state in state-space X_i. $A = A_1 \times \cdots \times A_m$ is a finite set of joint actions. $\phi = \phi_1 \times \cdots \times \phi_m$ are joint transition functions. ϕ is decomposed into the discrete marginals $P(n' \mid n,x,a)$ and the continuous conditionals $P(x' \mid n,x,a,n')$. For all (n,x,a,x') it holds $\sum_{n' \in N} P(n' \mid n,x,a) = 1$ and $\int_{x \in X} P(x' \mid n,x,a,n')dx = 1$.

$R_n(x)$ denotes the reward obtained in joint state $(n;x)$ where $n \in N, x \in X$. N_0 is the initial discrete state, with initial distributions $P_0(x_i)$ for each agent $i = 1,\cdots,m$ and $P_0(x) = \bigotimes_{i=1}^{m} P_0(x_i)$.

In our application domain, continuous variables model non-replenishable resources. This translates into the assumption that the value of the continuous variables is non increasing. Each resource is internal to an agent and is thus independent of other agent resources. It is thus assumed that an agent action has no effect on other agent resource states. In this work we rely on the stronger assumption that the DEC-HMDP is *transition independent*, that is an agent action has no effects on other agent discrete state as well (Becker et al., 2004). This assumption greatly simplifies the computation and adds useful properties to the framework.

An m-agents transition independent DEC-HMDP (TI-DEC-HMDP) is a DEC-HMDP such that for $a \in A$, $n \in N$, $x \in X$, ϕ is such that

$$\forall i = 1,\cdots,m, \begin{cases} P(n'_i \mid n,x,a) & = P(n'_i \mid n_i,x_i,a) \\ P(x'_i \mid n,x,a,n') & = P(x'_i \mid n_i,x_i,a,n'_i). \end{cases}$$

In the rest of the paper, we consider an m-agents resource constrained TI-DEC-HMDP (RC-TI-DEC-HMDP).

2.2 Goals, policy and reward

Agents operate in a decentralized manner, and choose their actions according to their local view of the world. They do not communicate but are cooperative, i.e. there is a single value function for all agents. We assume a set of identified global goals $\{g_1,\cdots,g_k\}$, each of which is known and corresponds to an achiement by one or more agents. Each $g_j \in N$ is such that $g_j = \{g_{ij}\}_{i=\alpha_j^1,\cdots,\alpha_j^{q_j}}$ where $g_{ij} \in N_i$ and $\alpha_j^q \in \{1,\cdots,m\}$. For simplifying notations, we note $i \in g_j$ the fact that $g_{ij} \in g_j$, i.e. agent i is involved in goal state g_j. The reward function for a RC-TI-DEC-HMDP is decomposed into two components for each goal j: a set of individual local reward functions for each agent, the $R_{g_{ij}}(x_i)$; a joint reward $c_j(x)$ the team receives and that depends on the actions of more than one agent.

The joint reward articulates the interaction among agents. In general agents seek to maximized the local and joint reward functions. In this case negative c_j such as in our case study (see 2.3) can be seen as a penalty put on some agent interactions. Positive c_j naturally favor certain other interactions.

Given a RC-TI-DEC-HMDP, we define a policy $\pi = \{\pi_1,\cdots,\pi_m\} : (N,X) \to A$ to be a mapping from the spate space to the action space. A global value function $GV : (N,X) \to \Re$ gives the expected total reward of the system starting from an initial state and acting according to the policy π until termination. Termination occurs whenever all goals are achieved or all agents have run out of resources. Similarly, the local value function $V^i : (N_i,X_i) \to \Re$ gives the expected total reward for agent i, and the joint value function $JV : (N,X) \to \Re$ gives the joint expected total reward of the system. The joint value function is given by

$$JV(x \mid \pi_1,\cdots,\pi_m) = \sum_{j=1}^{k} c_j(x) \prod_{i \in g_j} P_{g_{ij}}(x_i \mid \pi_i) \quad (1)$$

where the c_j are the joint rewards, and $P_{g_{ij}}(x_i \mid \pi_i)$ is the probability agent i has to achieve goal j according to the policy π_i. Often the joint rewards are in fact

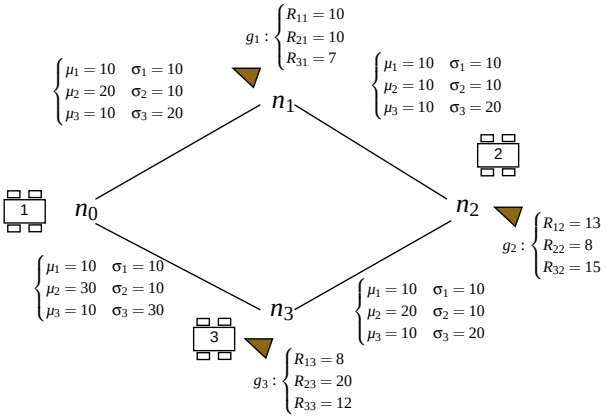


Figure 1: Case-study: a 3 agents / 3 goals problem.

penalties. The global value function is given by

$$GV(x \mid \pi_1, \cdots, \pi_m) = \sum_{i=1}^{m} V_0^i(x_i) + JV(x \mid \pi_1, \cdots, \pi_m) \tag{2}$$

where $V_0^i(x_i)$ is the local value function for the initial state of agent i. The optimal joint policy is noted $\pi^* = \{\pi_1^*, \cdots, \pi_m^*\}$, given by

$$\pi^* = \underset{\pi_1, \cdots, \pi_m}{\text{argmax}} E_X[GV(x \mid \pi_1, \cdots, \pi_m)] \tag{3}$$

where E_X denotes the expectation over state space X. The optimal value function is $GV^*(x \mid \pi_1^*, \cdots, \pi_m^*) = \max_{\pi_1, \cdots, \pi_m} GV(x \mid \pi_1, \cdots, \pi_m)$. Note that action Abort_i ends the policy of agent i.

2.3 Case-study

Consider the problem on figure 1. Three rovers share three rocks over four locations. The (μ_i, σ_i) are the standard mean and variance of the Gaussian distribution that models resource consumption $P(x_i' \mid x_i, n, n', a)$ of rover i on each path (n, n'). $R_{g_{ij}}$ is the reward function for achieving goal j and agent i. Each rover starts from a different initial location.
We study a joint value of the form

$$JV(x \mid \pi_1, \cdots, \pi_m) = -\beta \max_{i \in \{0, \cdots, m\}} \sum_{j=1}^{k} R_{g_{ij}}(x) \prod_{i \in g_j} P_{g_j}(x_i \mid \pi_i) \tag{4}$$

where $\beta \in [0, 1]$. In other words, the joint reward signal $c_j(x)$ for goal j is a negative factor, or penalty, of value the maximum reward possibly obtained by an agent for that goal. The rational behind this model is that it allows to parametrize the collaboration among agents. Thus for $\beta = 1$ no collaboration is beneficial, and each agent has a high incentive of avoiding goals already achieved by other agents. When $\beta < 1$, there is incentive for all agents to consider all goals, with the amount collaboration inversely proportional to β. For this reason, in the following, we refer as *the collaboration factor* of a team of agents as a function that is *inversely proportional to* β.

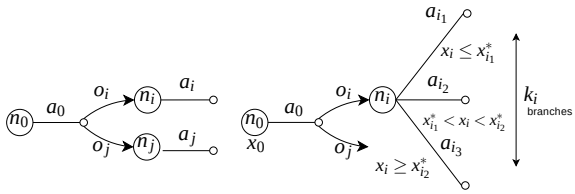


Figure 2: MDP policy (left) and HMDP policy (right).

2.4 Oversubscription, conditional policies, branches and observations

2.4.1 Oversubscription in goals

In a RC-TI-DEC-HMDP, the resource constrained nature of each agent gives rise to *over-subscribed* planning problems. These are problems in which not all the goals are feasible by each agent under its internal resource constraints and the initial distribution over its resources. Their solutions are characterized by the existence of different achievable goal sets for different resource levels. In our application domain, it is assumed that each goal can be achieved only once (no additional utility is achieved by repeating the task).

2.4.2 Conditional policies

As defined earlier, the policy solution to a RC-TI-DEC-HMDP is a set of individual policies, one per agent. In fact, an agent policy is solution to an underlying HMDP (Becker et al., 2004). There is no need to define this HMDP here. What we are interested in is the form taken by an HMDP policy, in general.

Most traditional planners assume a discrete state-space and a small number of action outcomes. When the model is formalized as an MDP, the planner can decide on discrete action outcomes. The policy *branches* on discrete action outcomes. A policy thus reads *from state n_0 and action a_0, when action outcome is o_i, do action a_i; else when action outcome is o_j, do action a_j; ....* A MDP solution policy is pictured on the left-hand side of figure 2.

When the model includes continuous resources and is modeled as an HMDP, a consequence of over-

Table 1: Empirical measures and related information on the underlying planning problem.

Empirical measure	Information
OCS	Dependency of one agent on the other agents' policies
Local plans studied	Dependency of agent on the other agents' policies. Complexity of the local decision problem.
Joint policies studied	Computational difficulty of finding of the joint controller. Dependencies among all agents.
Branches in the optimal joint policy	Observations of their local resource states by the agents.
Size of the optimal joint policy	Repartition of goals (rocks) among agents.
Discretization (number of pieces)	Complexity of the decision problem (local or global).

subscription is that a HMDP policy is conditional upon resources. Thus the planner must branch not only on the discrete action outcomes, but on the availability of continuous resources. In this case, a policy reads *from discrete state* n_0, *continuous resource* x_0 *and action* a_0, *when action outcome is* o_i, *then if continuous resource is* $x_i \leq x^*_{i_1}$, *do action* a_{i_1}, *else if continuous resource is* $x^*_{i_1} < x_i < x^*_{i_2}$, *do action* a_{i_2}, *etc...; else when action outcome is* o_j.... The right-hand side of figure 2 pictures a portion of an HMDP policy.

2.4.3 Observation, collaboration and computation

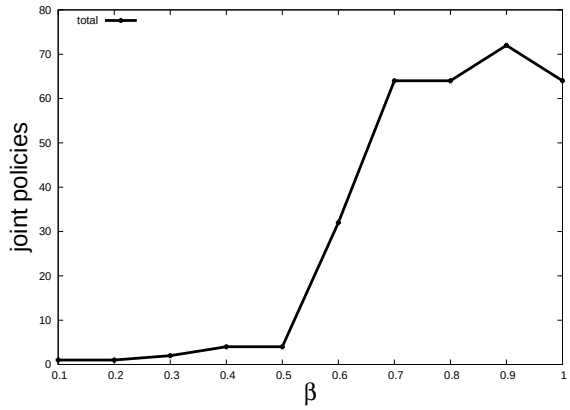
Now, the important point is that each branch of an HMDP policy calls for an observation of the agent resource state. Each observation is to be carried out at execution time. Because of the oversubscribed nature of the planning problem, each agent has to make a certain number of observations before deciding which goals to achieve. In the multiagent framework, the collaboration among agents and its possible penalties affects the repartition of goals, and thus the need for observation of its resource state by each agent. As a consequence, this also affects the computational weight of finding an optimal policy for a team of agents. The rest of this paper reports on the results of a series of simulations and tests that yield empirical evidences of the relation between collaboration, observation and computation.

3 Computation and Complexity

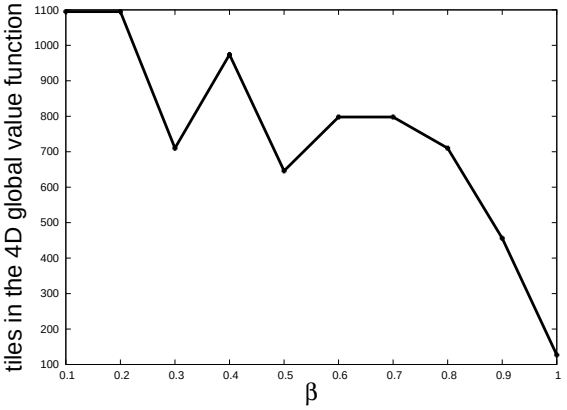
3.1 Solving RC-TI-DEC-HMDPs

Here we give a little background on the solving of an m-agents RC-TI-DEC-HMDPs. The Cover Set Algorithm (CSA) is an efficient algorithm that finds optimal policies (Becker et al., 2004; Petrik and Zilberstein, 2007). It is a two steps algorithm. The first step consists in finding a set of policies for each of $(m-1)$ agents, called the optimal cover set (OCS). Each agent's OCS is such that for any of the other agent's policies it contains at least a policy that is optimal. In other words, the OCS of an agent is guaranteed to contain the optimal policy for this agent that belongs to the optimal policy for the team. In computing the OCS for an agent, the CSA has to study a number of competing local policies for this agent. This number yields an information on the dependency of the agent w.r.t. the other agent policies. The second step iterates all combinations of policies in the $(m-1)$ OCS, computes an optimal policy for the m-th agent, and returns the combination of m policies that yields the maximal expected global value (2). Table 1 sums up the empirical measures and their information on the underlying planning problem.

Computationally, the challenging aspect of solving an HMDP is the handling of continuous variables, and particularly the computation of the so-called Bellman optimality equation. At least two approaches, (Feng et al., 2004) and (Li and Littman, 2005) exploit the structure in the continuous value functions of HMDPs. Typically these functions appear as a collection of humps and plateaus, where the later correspond to a region in the continuous state space where similar goals are pursued by the policy. The steepness of the slope between plateaus reflects the uncertainty in achieving the underlying goals. The algorithms used for producing the results analyzed in this paper exploit a problem structure where the continuous state can be partitioned into a finite set of regions. Taking advantage of the structure relies on grouping those states that belong to the same plateau, while dynamically scaling the discretization for the regions of the state space where it is most useful such as in between plateaus. It follows that the dynamic discretization of the continuous state-space reflects the complexity of the decision problem: the less discretized pieces, the easiest the decision, see Table 1.



(a) Number of studied joint policies.



(b) Dynamic discretization of the optimal global value function.

Figure 3: Case-study: joint policy computation.

3.2 Empirical evidences

This section reports on planning for our case-study. It helps understanding the relation between collaboration and computation. Figure 3 reports on the computation of the optimal joint policy. Figure 3(a) shows the number of joint policies studied for selecting the optimal joint policy. This number jumps with the reduction of the collaboration factor among agents that is implicitely carried by the joint reward structure. One hypothesis is that the problem becomes globally more computational when the amount of collaboration among agents is reduced. In fact, this hypothesis is confirmed by the results on figure 3(b). The number of discretized regions in the optimal four-dimensional global value function reflects the discretization of the optimal value functions of individuals. The finer the discretization, the more complex and thus the more difficult the decision problems at the level of individual agents. The very clear drop in the number of regions with the reduction of the collaboration factor among agents corroborates our hypothesis: low collaboration puts the stress on the global controller and relieves the individuals. On the opposite, when β moves toward 0 and collaboration is high, each agent has an incentive to visit all rocks.

Fact 1 *The computational difficulty of finding the global controller for a team of resource constrained agents is a decreasing function of their collaboration factor.*

Table 2 characterizes the optimal joint policy for our case-study. These numbers confirm the trend observed in other figures: agents involved in less collaborative problems (i.e. $\beta \approx 1$) are more dependent on the strategies of others since they are forced to avoid the goals possibly achieved by others.

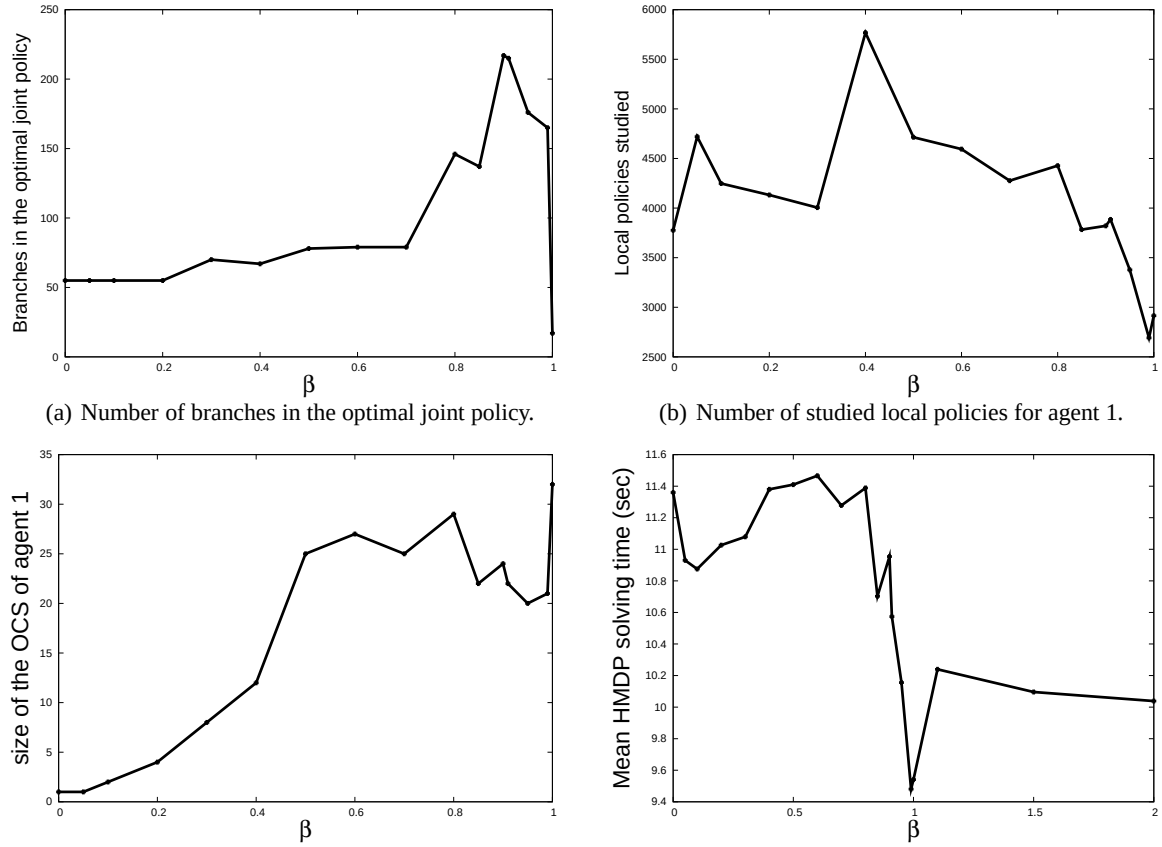
Table 2: Case-study: Optimal joint policy.

β	local policies studied	joint policy size	branches	unused agents
1	113256	17	5	1
0.9	89232	20	5	0
<0.9	68640	>20	5	0

Interestingly, for $\beta = 1$, agent 1 is useless, that is its policy is empty. In fact, agent 1 is dominated, i.e. the two other agents do what it does, and do it better. This is an indication that in larger problems, with more agents, heavy computations might lead to a empty optimal policies for certain agents. The next section studies the number of branches and the relation between collaboration and observation.

4 Collaboration and Observation

In this section we empirically study the relation between collaboration and the need for observation at agent level. To this end, we have considered a five rocks, two agents problem in the Mars rovers domain. We have varied the collaboration factor β for this problem. Figure 4(a) reports on the number of branches in the optimal joint policy for $\beta \in [0, 1]$. The number of branches is the number of times the policy asks for an observation of the level of continuous resources before acting. We see a four times increase of the number of branches, which reflects a growing need for the individual agents to observe their internal



(a) Number of branches in the optimal joint policy.

(b) Number of studied local policies for agent 1.

(c) Number of policies in the OCS of agent 1.

(d) Mean time needed to solve agent 1 local policies.

Figure 4: Collaboration vs. Observation: report on a 2 agents, 5 rocks problem from the Mars rovers domain.

resource state and cast away uncertainty[1].

Fact 2 *The number of observations required by the optimal policy of a team of resource constrained agents is a decreasing function of their collaboration factor.*

Now, consider figure 4(b) that shows the number of local policies that are studied decreases when β increases. This number is a function of two variables: i/the discretization of the continuous space; ii/the structural dependency on other agent strategies. Figure 4(c) shows that the size of the coverage set of agent 1. It shows that the number of policies in the OCS of agent 1 augments, and does not significantly decreases when β increases. This means that agent 1 grows more dependent on agent 2 when β increases. This is because when collaboration becomes more penalized, agent 1 has to be increasingly aware of agent 2's strategy before it takes action, thus mitigating its potentially negative effect on the global reward.

It follows from i/ and ii/ that this is the discretization of the continuous space that becomes less dense when β increases. Equivalently, this indicates that the underlying local decision problems are less complex. Less formally, this means that the world becomes increasingly sharper for the individual agents, with value functions that exhibit more plateaus and less slopes. Metaphorically, the world, as seen by each individual agent, turns into a *black & white* decisional space, where rocks must be clearly partitioned among agents, and collaboration tends to be avoided. In other words, there is less room for uncertainty, and risk is agressively eliminated, as early as possible. This correlates naturally with the higher number of observations required by the optimal joint policy. Each observation disambiguates the reachability of each rock and sharpens the view of an otherwise very stochastic world.

Fact 3 *The complexity of the agent local decision problems is an increasing function of their collabo-*

[1]The ratio (branches/number of actions in the optimal joint policy) remains rather constant, thus reflecting the structure of the problem: in most cases decision is taken before navigation to a rock.

ration factor.

In parallel, since local decision is sharper, it increasingly needs to be articulated with that of other agents. A consequence is that with increasing β, the decisional stress is increasingly shifted to the global controller. We had already noticed this behavior in section 3. Here we choose to observe the side-effect that is a relief of the computational weight that is put on individual agents. Figure 4(d) shows the mean time needed to solve an augmented HMDP for different values of β. The sudden decrease indicates the decisional shift from the local controllers to the global controller.

Fact 4 *A decrease in the collaboration factor for a team of resource constrained agents implies a shift of the computational weight from the local controllers to the global team controller.*

To summarize, agents, each with eclectic abilities, acting in a specialized world in which collaboration is not well valued, are forced to agressively decide upon their objectives more often, while the final computational burden is shifted to the global controller that governs them.

Conclusion

We have reported on an empirical study of the connections between collaboration, computation and the need for observation in optimal policies for resource constrained multiagent problems. These problems well model number of real-world situations for modern teams of robots. This includes our application domain, that of team of exploratory rovers.

We have defined the collaboration as the positive value given to interactions among agents in a team. Interestingly, we could show that the need for observation is a decreasing function of the collaboration among agents. We can sum up our empirical finding by considering a world where the division of labor is extreme, and collaboration not much valuable. In this world, resource constrained individuals with eclectic abilities (i.e. that are equally able with every task), are stressed to take sharp decisions, more often, and based on recurrent observations of their own resources.

REFERENCES

Becker, R., Zilberstein, S., Lesser, V., and Goldman, C. (2004). Solving transition independent decentralized markov decision processes. *Journal of Artificial Intelligence Research*, 22.

Bernstein, D., Givan, R., Immerman, N., and Zilberstein, S. (2002). The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, 27(4).

Boutilier, C., Dean, T., and Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11.

Feng, Z., Dearden, R., Meuleau, N., and Washington, R. (2004). Dynamic programming for structured continuous Markov decision problems. In *Proceedings of the Twentieth International Conference on Uncertainty In Artificial Intelligence*, pages 154–161.

Li, L. and Littman, M. (2005). Lazy approximation for solving continuous finite-horizon mdps. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*.

Petrik, M. and Zilberstein, S. (2007). Anytime coordination using separable bilinear programs. In *Proceedings of the Twenty Second National Conference on Artificial Intelligence*.