Multiagent Planning under Uncertainty with Stochastic Communication Delays

Matthijs T.J. Spaan
Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal

Frans A. Oliehoek
Informatics Institute
University of Amsterdam
The Netherlands

Nikos Vlassis
Department of Production Engineering and Management
Technical University of Crete
Chania, Greece

Abstract

We consider the problem of cooperative multiagent planning under uncertainty, formalized as a decentralized partially observable Markov decision process (Dec-POMDP). Unfortunately, in these models optimal planning is provably intractable. By communicating their local observations before they take actions, agents synchronize their knowledge of the environment, and the planning problem reduces to a centralized POMDP. As such, relying on communication significantly reduces the complexity of planning. In the real world however, such communication might fail temporarily. We present a step towards more realistic communication models for Dec-POMDPs by proposing a model that: (1) allows that communication might be delayed by one or more time steps, and (2) explicitly considers future probabilities of successful communication. For our model, we discuss how to efficiently compute an (approximate) value function and corresponding policies, and we demonstrate our theoretical results with encouraging experiments.

Introduction

In this paper we consider planning for multiagent systems (MASs), formalized in a decision-theoretic framework to tackle various forms of uncertainty a multiagent team can encounter. As in the single-agent case, two main sources of uncertainty are each agent's imperfect sensors and the uncertain effects of its actions. Moreover, planning in MASs is significantly harder than for a single agent, since when considering the plan for an agent, one also has to consider the effects of the actions of other agents. Especially when agents have to base their decisions on local observations (sensor readings), each agent has a different view of the environment, making it hard to predict the actions of other agents. Optimal planning in such partially observable and decentralized scenarios is provably intractable, which limits the scalability of optimal solutions to a very small number of agents and a planning horizon of a few time steps.

Communication capabilities can mitigate these issues of partial observability, as they allow agents to share information such as sensor readings. In this way, communication of the local observations makes each agent better informed regarding the state of the environment, as well as providing a way for the agents to coordinate their actions. In particular, assuming instantaneous and cost-free communication effectively reduces the problem to a centralized problem (Pynadath and Tambe 2002), modeled as a partially observable Markov decision process (POMDP).

This approach requires synchronization within every time step: after reading their sensors, each agent broadcasts its local observation to the team, and waits for incoming messages. Instantaneous communication does not exist, so this synchronization step requires some time. Moreover, communication can fail temporarily, in which case the agent still has to select an action. However, many current approaches for planning for decentralized POMDPs (Dec-POMDPs) with communication assume that communication is instantaneous and without failure (Roth, Simmons, and Veloso 2005; Becker, Lesser, and Zilberstein 2005; Roth, Simmons, and Veloso 2007), and do not provide a mechanism to deal with less-than-perfect communication. Other approaches in literature examined MASs in which communication arrives with a delay of one time step (Schoute 1978; Grizzle, Hsu, and Marcus 1982), however, guaranteed communication is still assumed (but with a fixed delay). Moreover, when communication is not delayed these methods are not able to exploit this. We address these shortcomings by explicitly reasoning about the probability of successful communication (with variable delays) in the future. Our work provides a significant step towards more realistic communication models for planning in Dec-POMDPs with unreliable communication.

In previous work we have shown how Bayesian games can be used to plan for a MAS in which communication arrives with a delay of one time step (Oliehoek, Spaan, and Vlassis 2007). To select an action at each stage in this 1-step delayed (1TD) setting we proposed to use the Q_{BG}-value function and we demonstrated how this Q_{BG}-value function can be computed efficiently. We extend upon this work by considering the setting in which there is stochastically delayed communication (SDC). That is, when communication will usually succeed within a stage, but might fail with some probability. In particular we formalize (1) at what points in time synchronization (i.e., communication) is expected to be completed, (2) the probability with which this occurs, and (3) what happens if synchronization does not finish within the allotted wall-clock time frame. For this SDC setting

we propose a planning method that is exact when the delay of communication is at most one stage. We show that the resulting value function can be compactly represented and how it can be computed efficiently by transferring POMDP solution techniques. We apply our theoretical results by demonstrating them in encouraging experiments, showing the potential benefit of the proposed model.

The rest of the paper is organized as follows. First we introduce the general Dec-POMDP model without communication. Then we describe three communication models, namely instantaneous communication (0TD), communication with a one time step delay (1TD), and with a variable stochastic delay (SDC). Next, we show how we can plan in these models using value iteration. Finally, we perform several experiments, we provide conclusions and we discuss future work.

Dec-POMDPs without communication

First we will briefly review the Dec-POMDP model; for a detailed introduction we refer to (Oliehoek, Spaan, and Vlassis 2008). In this standard Dec-POMDP setting no communication is possible, and optimal planning is provably intractable (NEXP-complete (Bernstein et al. 2002)).

Definition 1 A *decentralized partially observable Markov decision process (Dec-POMDP)* with m agents is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \mathcal{O}, O, b^0 \rangle$, where:

- $\mathcal{S}$ is a finite set of states.
- $\mathcal{A} = \times_i \mathcal{A}_i$ is the set of *joint actions*, where $\mathcal{A}_i$ is the set of actions available to agent i. Every time step, one joint action $\mathbf{a} = \langle a_1, \ldots, a_m \rangle$ is taken. Agents do not observe each other's actions.
- $T : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the transition function, a mapping from states and joint actions to probability distributions over states, specifying $P(s'|s, \mathbf{a})$.
- R is the immediate reward function, which maps states and joint actions to real numbers: $R(s, \mathbf{a})$.
- $\mathcal{O} = \times_i \mathcal{O}_i$ is the set of joint observations, where $\mathcal{O}_i$ is a finite set of observations available to agent i. Every time step one joint observation $\mathbf{o} = \langle o_1, \ldots, o_m \rangle$ is received, from which each agent i observes its own component o_i.
- O is the observation function, which specifies the probability of joint observations given taken joint actions and successor states: $P(\mathbf{o}|\mathbf{a}, s')$.
- $b^0 \in \mathcal{P}(\mathcal{S})$ is the initial state distribution at $t = 0$.

When there is only one agent in a Dec-POMDP, the model reduces to a POMDP (Kaelbling, Littman, and Cassandra 1998). The planning problem is to compute a plan, or *policy*, for each agent that is optimal for a particular number of time steps h, also referred to as the *horizon* of the problem. We denote the interval of wall-clock time that passes between two decision points by Δ_t, and assume it to be constant without loss of generality. A common optimality criterion is the expected cumulative (discounted) future reward $E(\sum_{t=0}^{h-1} \gamma^t R(t))$, where $R(t)$ denotes the reward at time step t, and $0 < \gamma \leq 1$ is a discount factor.

A tuple of policies $\pi = \langle \pi_1, \ldots, \pi_m \rangle$ is referred to as a *joint policy*. In general, each individual deterministic (*pure*) policy π_i is a mapping from histories of observations to actions: $\pi_i((o_i^1, \ldots, o_i^t)) = a_i$. Here, $(o_i^1, \ldots, o_i^t)$ is the sequence of observations received by agent i up to time step t, which we refer to as the *observation history* $\vec{o}_i^t$. We also use a different notion of history, namely the *action-observation history* $\vec{\theta}_i^t$ which consists of all observations received and actions taken up to time step t: $\vec{\theta}_i^t = (a_i^0, o_i^1, a_i^1, \ldots, a_i^{t-1}, o_i^t)$. We also consider corresponding *joint* histories, respectively denoted as $\vec{\mathbf{o}}^t$ and $\vec{\theta}^t$.

Figure 1(a) demonstrates the no-communication setting as modeled by a Dec-POMDP. It illustrates that the agents select actions based on their individual observations only.

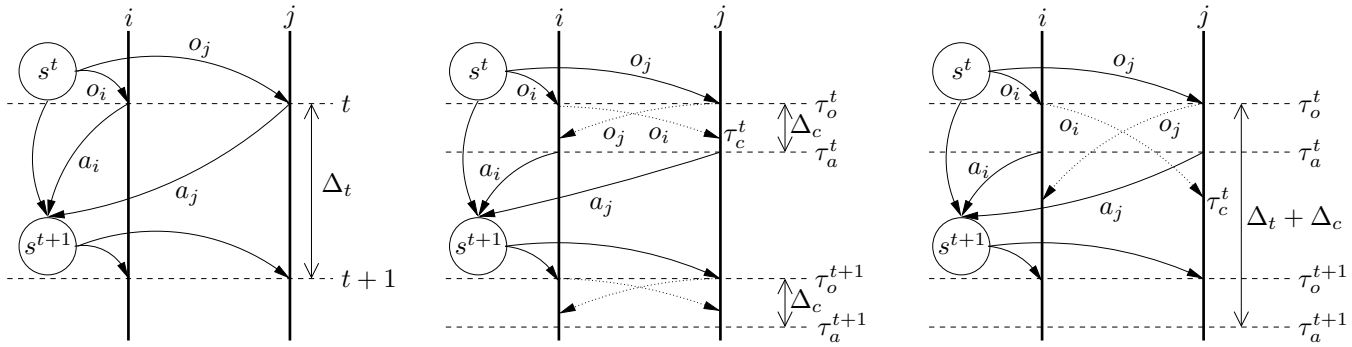
Instantaneous communication

A natural approach to tackle the problem of decentralized observations is to allow the agents to communicate their observations. In the case of cost-free, instantaneous and noiseless communication, sharing local observations at each time step is optimal (Pynadath and Tambe 2002). Of course, true instantaneous communication does not exist, but when communication is guaranteed to be very fast, the assumption can be applied as demonstrated in Figure 1(b). Once the previous actions are completed and the state transitions to s^t, the agents get their new observations. We assume this happens at (real-world, wall-clock) time τ_o^t.[1] At that point each agent broadcasts its individual observation, resulting in synchronization at τ_c. The specifics of how synchronization can be achieved are beyond the scope of this paper, and we assume the agents have synchronized clocks. For an in-depth treatment of related common-knowledge issues, we refer to (Halpern and Moses 1990). The agents will have to act at time τ_a^t, which means that synchronization must be guaranteed to be completed in $\Delta_c^t = \tau_a^t - \tau_o^t$ time units. We assume without loss of generality that the communication periods Δ_c are of equal length for all stages and we drop the t index.

In this 0TD case the planning problem reduces to a centralized POMDP, as one can assume there is a centralized agent, say a 'puppeteer', that receives joint observations and takes joint actions in response. During execution all agents will communicate their observations, look up the optimal *joint* action for the resulting *joint* observation history, and execute their own action component. For such a POMDP the (joint action-observation) history of the process can be summarized by a probability distribution over states called a *joint belief* b. We will write $b^{\vec{\theta}^t}$ for the joint belief as it would be computed by the puppeteer after action-observation history $\vec{\theta}^t$. The joint belief $b^{\vec{\theta}^{t+1}}$ resulting from $b^{\vec{\theta}^t}$ by joint action $\mathbf{a}$ and joint observation $\mathbf{o}$ can be calculated by Bayes' rule:

$$b^{\vec{\theta}^{t+1}}(s') = \frac{P(\mathbf{o}|\mathbf{a}, s')}{P(\mathbf{o}|b^{\vec{\theta}^t}, \mathbf{a})} \sum_s P(s'|s, \mathbf{a}) b^{\vec{\theta}^t}(s). \quad (1)$$

[1] Throughout the paper, t indices refer to the discrete time steps of the decision process, while τ denotes a point in wall-clock time.



(a) No communication. (b) Instantaneous communication (0TD). (c) Delayed communication (1TD).

Figure 1: Illustration of different communication models used in this paper. For two consecutive time steps, and two agents, i and j, we show on what information each agent bases its action choice. In (a) we show the general Dec-POMDP setting which lacks communication. In the 0TD setting (b) agents can only decide on their actions (at τ_a^t) after receiving the local observation of the other agents, shortly after synchronization at τ_c^t. Finally, (c) shows the 1TD setting, in which agents act upon receiving their local observation at τ_o^t (Δ_c can be short), without waiting until τ_c^t. However, at τ_a^{t+1} they have received all local observations from time t, i.e., $\tau_a^{t+1} > \tau_c^t$.

The optimal Q-value function for POMDPs is based on such beliefs and satisfies the following Bellman equation:

$$Q^*_{\text{P}}(b^{\vec{\theta}^t}, \mathbf{a}) = R(b^{\vec{\theta}^t}, \mathbf{a}) + \sum_{\mathbf{o}\in\mathcal{O}} P(\mathbf{o}|b^{\vec{\theta}^t}, \mathbf{a}) \max_{\mathbf{a}^{t+1}} Q^*_{\text{P}}(b^{\vec{\theta}^{t+1}}, \mathbf{a}^{t+1}), \quad (2)$$

where $R(b^{\vec{\theta}^t}, \mathbf{a}) = \sum_s R(s, \mathbf{a}) b^{\vec{\theta}^t}(s)$ is the expected immediate reward.

It is well known that the value function (2) is a piecewise-linear and convex (PWLC) function over the joint belief space (Kaelbling, Littman, and Cassandra 1998). This means we can use sets of vectors $\mathcal{V}^t_{\mathbf{a}}$ to represent the value of a joint belief b and $\mathbf{a}$:

$$Q^t(b, \mathbf{a}) = \max_{v^t_{\mathbf{a}} \in \mathcal{V}^t_{\mathbf{a}}} b \cdot v^t_{\mathbf{a}}, \quad (3)$$

where $(\cdot)$ denotes inner product. The fact that the POMDP value function is PWLC allows for a compact representation for optimal algorithms, as well as many opportunities for fast approximate ones. Note that for the Dec-POMDP case, i.e., without communication, no such convenient policy representation exists.

Communication delayed by one time step

We now consider communication that is delayed by one time step as illustrated in Figure 1(c). In this 1TD setting synchronization will not be completed before the agents select an action, i.e., $\tau_a^t < \tau_c^t$. Rather, synchronization must be completed before the decision at the next stage $\tau_c^t < \tau_a^{t+1}$, i.e., synchronization is achieved within $\Delta_t + \Delta_c$ time units. Note that, since the agents do not wait for communication within a stage, they can act (almost) immediately when receiving their observation and Δ_c can be set to 0 (or be short).

When the agents communicate their individual observations and last taken action, then at every time step t each agent knows the previous joint observation history $\vec{\mathbf{o}}^{t-1}$ and action $\mathbf{a}^{t-1}$. However, each agent i has only observed its individual last observation o_i^t and is uncertain regarding the last joint observation $\mathbf{o}^t$. For every $(\vec{\theta}^{t-1}, \mathbf{a}^{t-1})$, this situation can be modeled using a *Bayesian game* (BG) (Oliehoek, Spaan, and Vlassis 2007), a strategic game with imperfect information and identical payoffs. In particular, the private information of each agent (which defines its *type*), is its last local observation o_i^t. As such, the policies for the BG map a single individual observation to an individual action $\beta_i : \mathcal{O}_i \to \mathcal{A}_i$. A joint BG-policy is a tuple $\beta = \langle \beta_1, \ldots, \beta_m \rangle$. The probabilities of the joint observations in this BG are known: $P(\mathbf{o}) \equiv P(\mathbf{o}^t | b^{\vec{\theta}^{t-1}}, \mathbf{a}^{t-1})$.

When a payoff function $Q(\mathbf{o}, \mathbf{a})$ is available, the solution of a BG with identical payoffs is given by the optimal joint BG policy β^*:

$$\beta^* = \arg\max_{\beta} \sum_{\mathbf{o}\in\mathcal{O}} P(\mathbf{o}) Q(\mathbf{o}, \beta(\mathbf{o})), \quad (4)$$

where $\beta(\mathbf{o}) = \langle \beta_1(o_1), \ldots, \beta_m(o_m) \rangle$ is the joint action specified by β for joint observation (type) $\mathbf{o}$. β^* is a Pareto-optimal Bayesian Nash equilibrium (Oliehoek, Spaan, and Vlassis 2008).

We have shown before that when considering the solutions of all such BGs (for all stages t and all $\vec{\theta}^{t-1}, \mathbf{a}^{t-1}$), the optimal payoff function Q for the 1TD setting is recursively defined and corresponds to the Q_{BG} value function (Oliehoek, Spaan, and Vlassis 2007):

$$Q^*_{\text{BG}}(\vec{\theta}^t, \mathbf{a}) = R(\vec{\theta}^t, \mathbf{a}) + \max_{\beta} \sum_{\mathbf{o}} P(\mathbf{o}|\vec{\theta}^t, \mathbf{a}) Q^*_{\text{BG}}(\vec{\theta}^{t+1}, \beta(\mathbf{o})), \quad (5)$$

where $R(\vec{\theta}^t, \mathbf{a}) = R(b^{\vec{\theta}^t}, \mathbf{a})$, as defined at (2).

We now prove that we can write the Q_{BG}-value function as a function over the joint belief space, and we prove that it is

PWLC over this space[2]. As such, we unify the mathematical frameworks for both 0TD and 1TD communication.

Lemma 1 *The Q_{BG}-value function* (5) *is a function over the joint belief space, i.e., we can re-write it as:*

$$Q_{BG}(b^{\vec{\theta}^t}, \mathbf{a}) = R(b^{\vec{\theta}^t}, \mathbf{a}) + \max_{\beta} \sum_{\mathbf{o}^{t+1}} P(\mathbf{o}^{t+1} | b^{\vec{\theta}^t}, \mathbf{a}) Q_{BG}(b^{\vec{\theta}^{t+1}}, \beta(\mathbf{o}^{t+1})), \quad (6)$$

where $b^{\vec{\theta}^t}$ denotes the joint belief induced by $\vec{\theta}^t$.

Sketch of proof Converting a joint action-observation history $\vec{\theta}^t$ to the corresponding joint belief $b^{\vec{\theta}^t}$ is a matter of applying Bayes' rule (1). What needs to be shown is that, if two different joint action-observation histories $\vec{\theta}^{t,a}, \vec{\theta}^{t,b}$ correspond to the same joint belief, they also have the same Q_{BG}-values: $\forall_{\mathbf{a}}\ Q_{BG}(\vec{\theta}^{t,a}, \mathbf{a}) = Q_{BG}(\vec{\theta}^{t,b}, \mathbf{a})$. The proof is inductive, with the base case given by the last time step $t = h - 1$. In this case (5) reduces to $R(b^{\vec{\theta}^{h-1}}, \mathbf{a})$. Clearly $\forall_{\mathbf{a}}\ Q_{BG}(\vec{\theta}^{t,a}, \mathbf{a}) = Q_{BG}(\vec{\theta}^{t,b}, \mathbf{a})$ holds in this case. Proof that $\forall_{\mathbf{a}}\ Q_{BG}(\vec{\theta}^{t,a}, \mathbf{a}) = Q_{BG}(\vec{\theta}^{t,b}, \mathbf{a})$ given that the property holds for $t + 1$ is given in (Oliehoek, Vlassis, and Spaan 2007). □

Theorem 1 *The Q_{BG}-value function for a finite-horizon Dec-POMDP with 1 time step delayed, free and noiseless communication, as defined in* (6) *is piecewise-linear and convex over the joint belief space.*

Sketch of proof The proof is by induction. The base case is the last time step $t = h - 1$, and again, for this case (6) tells us that $Q_{BG}(b^{\vec{\theta}^t}, \mathbf{a}) = \sum_s R(s, \mathbf{a}) b^{\vec{\theta}^t}(s)$. Clearly Q_{BG} is trivially PWLC for the last time step. The induction step proves that when Q_{BG} is PWLC for $t+1$, it is also PWLC for t. The full proof is listed in (Oliehoek, Vlassis, and Spaan 2007). □

These results are in accordance with the fact that a decentralized system under one step delayed communication (also "one step delay sharing patterns") is separable (Varaiya and Walrand 1978). Hsu and Marcus (1982) presented a, rather involved, application of dynamic programming and mention that the resulting value function (which is different from the Q_{BG}-value function), is PWLC. Our results here should be taken as a reformulation of this dynamic program which can be interpreted in the context of BGs, and show a clear analogy between the setting with 0TD communication and the 1TD communication setting.

The implications of Theorem 1 are that, unlike the general Dec-POMDP case, the value function for a Dec-POMDP with 1TD communication can be compactly represented. Also, it is possible to transfer POMDP solution methods that exploit the PWLC property to the computation of the Q_{BG} value function. Moreover, the identified analogy between the 0TD and the 1TD setting, allows us to blend them in the same methodological framework, as shown next.

[2]These results were already mentioned in (Oliehoek, Spaan, and Vlassis 2007), however, formal proofs were not presented.

Stochastically delayed communication

The 0TD and 1TD models described in the previous sections assume guarantees on communication delay. However, in the real world, communication may temporarily fail, and such guarantees are hard to come by. For instance, consider a team of robots connected via a wireless network. Such wireless links can be unreliable, requiring retransmissions of packets and resulting in variable delays. This makes guaranteed synchronization within a particular time window hard to achieve.

We propose an approach for MASs with stochastically delayed communication (SDC): systems where communication will be available most of the time, i.e., where synchronization succeeds within a stage with a particular probability. We start by formalizing the probability that communication succeeds, and we assume that successful communication results in synchronization. For the 0TD model, the agents need to synchronize their observations within Δ_c time units and for the 1TD model within $\Delta_t + \Delta_c$, as described before (see Figure 1). Suppose we have a cumulative distribution function (cdf) $f_c(\Delta)$ which provides $P(\tau_c \leq \tau_o + \Delta)$: the probability that communication succeeds within Δ time units after the begin of the communication phase (τ_o). This allows us to write

$$p^{0\text{TD}} = f_c(\Delta_c) \quad (7)$$

for the probability that communication is instantaneous;

$$p^{1\text{TD}} = f_c(\Delta_t + \Delta_c) - f_c(\Delta_c) \quad (8)$$

for the probability that communication is 1-step delayed;

$$p^{2\text{TD}} = f_c(2\Delta_t + \Delta_c) - f_c(\Delta_t + \Delta_c) \quad (9)$$

for the probability of two steps delay, and so on. The optimal value function for such a setting would consider a weighted sum of these different settings. Using simplified notation, we have that the value of SDC can be expressed as

$$Q^*_{\text{SD}} = R + p^{0\text{TD}} F_{0\text{TD}} + p^{1\text{TD}} F_{1\text{TD}} + p^{2\text{TD}} F_{2\text{TD}} + \ldots, \quad (10)$$

where $F_{i\text{TD}}$ is the expected future reward given that communication will be delayed i stages.

To evaluate (10) exactly is impractical, and when f_c only reaches 1 in the limit even impossible. Rather we propose to approximate it. In particular we will assume (during the offline planning phase) that the communication is at most 1TD. That is, we define the probability of delayed communication as

$$p^D = p^{1\text{TD}} + p^{2\text{TD}} + \cdots = 1 - p^{0\text{TD}}, \quad (11)$$

and our approximate value function as

$$\widetilde{Q}^*_{\text{SD}}(\vec{\theta}^t, \mathbf{a}) = R(\vec{\theta}^t, \mathbf{a}) + p^{0\text{TD}} F_{0\text{TD}}(\vec{\theta}^t, \mathbf{a}) + p^D F_{1\text{TD}}(\vec{\theta}^t, \mathbf{a}), \quad (12)$$

where

$$F_{0\text{TD}}(\vec{\theta}^t, \mathbf{a}) = \sum_{\mathbf{o} \in \mathcal{O}} P(\mathbf{o} | \vec{\theta}^t, \mathbf{a}) \max_{\mathbf{a}^{t+1}} \widetilde{Q}^*_{\text{SD}}(\vec{\theta}^{t+1}, \mathbf{a}^{t+1}), \quad (13)$$

$$F_{1\mathrm{TD}}(\vec{\theta}^t, \mathbf{a}) = \max_{\beta} \sum_{\mathbf{o}\in\mathcal{O}} P(\mathbf{o}|\vec{\theta}^t, \mathbf{a}) \widetilde{Q}^*_{\mathrm{SD}}(\vec{\theta}^{t+1}, \beta(\mathbf{o})), \tag{14}$$

correspond to the expected future reward for the case of 0TD resp. 1TD communication at the next stage. Formally, our assumption is that the probability of 1TD communication is $1 - f_c(\Delta_c)$. Note that such an approach is exact when $f_c(\Delta_t + \Delta_c) = 1$ and may be an accurate approximation when it is close to 1. When computing the $\widetilde{Q}_{\mathrm{SD}}$-value function we determine a joint action $\mathbf{a}$ as well as a joint BG policy β for each belief. If in a stage t synchronization occurs within Δ_c, the agents can compute b^t and use $\mathbf{a}$. If not, they choose their actions according to the β of b^{t-1}.

The PWLC property of the 0TD and 1TD value functions allows us to assert the PWLC-property for $\widetilde{Q}_{\mathrm{SD}}$.

Corollary 1 *The value function for the stochastically delayed communication setting* (12) *is a function over the joint belief space, i.e., it can be written as*

$$\widetilde{Q}^*_{SD}(b^{\vec{\theta}^t}, \mathbf{a}) =$$
$$R(b^{\vec{\theta}^t}, \mathbf{a}) + p^{0\mathrm{TD}} F_{0\mathrm{TD}}(b^{\vec{\theta}^t}, \mathbf{a}) + p^D F_{1\mathrm{TD}}(b^{\vec{\theta}^t}, \mathbf{a}). \tag{15}$$

Moreover, for a finite horizon, it is PWLC over this space of joint beliefs.

Proof Proof that $\widetilde{Q}_{\mathrm{SD}}$ is a function over the belief space is analogue to the proof of Lemma 1. PWLC is proven as follows. Using simplified notation, we have

$$\widetilde{Q}^t_{\mathrm{SD}} = R + p^{0\mathrm{TD}} F^t_{0\mathrm{TD}} + p^D F^t_{1\mathrm{TD}}.$$

Using our knowledge of POMDPs, we know that $F^t_{0\mathrm{TD}}$ given by (13) is PWLC if $\widetilde{Q}^{t+1}_{\mathrm{SD}}$ is PWLC. The PWLC property of Q_{BG} also indicates that $F^t_{1\mathrm{TD}}$ given by (14) is PWLC when $\widetilde{Q}^{t+1}_{\mathrm{SD}}$ is PWLC. A weighted sum of two PWLC functions and adding a third (R) yields a PWLC function. Therefore $\widetilde{Q}^t_{\mathrm{SD}}$ is PWLC when $\widetilde{Q}^{t+1}_{\mathrm{SD}}$ is. Once again the base case is given by the last time step $t = h - 1$, for which $\widetilde{Q}_{\mathrm{SD}}(b^{\vec{\theta}^{h-1}}, \mathbf{a}) = R(b^{\vec{\theta}^{h-1}}, \mathbf{a})$ is PWLC. □

We note that it is possible to make the probability of communication state-dependent by using a cdf $f_c(\Delta; s)$. This flexibility allows us to model scenarios in which communication links are strong in certain locations, for instance when robots are close, and weaker in others. Let us write $p^{0\mathrm{TD}}$ for the probability that there is 0TD communication in the next stage. Then

$$p^{0\mathrm{TD}}(s, \mathbf{a}) \equiv \sum_{s'} P(s'|s, \mathbf{a}) f_c(\Delta_c; s'), \tag{16}$$

$$p^{0\mathrm{TD}}(\vec{\theta}^t, \mathbf{a}) = \sum_{s} b^{\vec{\theta}^t}(s) p^{0\mathrm{TD}}(s, \mathbf{a}). \tag{17}$$

The probability of one (or more) steps delayed communication in the next stage is given by $p^D(\vec{\theta}^t, \mathbf{a}) = 1 - p^{0\mathrm{TD}}(\vec{\theta}^t, \mathbf{a})$, which can be directly substituted in (12). Also, including a dependence of $p^{0\mathrm{TD}}$ on a particular stage t is trivial (in this finite-horizon setting).

Delays of more than one time step

The $\widetilde{Q}_{\mathrm{SD}}$-value function is exact when the communication delays are at most one time step. However, we would also like to be able to act successfully in models with longer delays, i.e., in which it may occur that the communication of $t - k, k > 1$ is not always received at stage t. If this happens, the agents should still select actions in some meaningful way. Given the complexity of computing (10) for delays > 1 (it grows doubly exponential in k (Ooi and Wornell 1996)), we propose an approximate on-line algorithm. In order to take into account the probability of 0TD communication in the future, it uses the $\widetilde{Q}_{\mathrm{SD}}$-value function, which has been computed off-line. The proposed open-loop method ensures agents take coordinated decisions in situations with delays longer than 1 one time step, however, other types of approximations are possible as well.

The main idea is that even when communication is failing, the agents know at what $t-k$ they have synchronized for the last time, i.e., all know $\vec{\theta}^{t-k}$. Basing decisions exclusively on information that is common knowledge ensures that the agents will continue to act in a coordinated manner. We propose to use an algorithm similar to Dec-COMM (Roth, Simmons, and Veloso 2005), in which each agent models the distribution $P(\vec{\theta}^{t-k+l}|\vec{\theta}^{t-k}, \mathbf{a}^{t-k+l}), 1 < l \leq k$ over *possible* $\vec{\theta}$ at subsequent time steps. To ensure that the joint actions taken at intermediate time steps t' (until communication is restored) will be known, the agents base their decisions only on information that is common knowledge such as the probability that a $\vec{\theta}^{t'}$ has been realized, denoted by $p(\vec{\theta}^{t'})$. Each agent computes

$$\arg\max_{\mathbf{a}} \sum_{\vec{\theta}^{t'}} p(\vec{\theta}^{t'}) \widetilde{Q}^*_{\mathrm{SD}}(\vec{\theta}^{t'}, \mathbf{a}), \tag{18}$$

and executes its component action a_i. When communication is restored, the agents fully synchronize their knowledge by sending all local observations since $t - k$.

Finite-horizon value iteration

So far we have discussed the existence of value functions corresponding to different communication models and we have shown some of their properties. Next we detail how a value-iteration algorithm for finite-horizon problem settings can be defined for all these PWLC value functions. Analogous to the POMDP case (Kaelbling, Littman, and Cassandra 1998), we will define how to compute $\mathrm{Q}^t_{\mathrm{BG}}$ from $\mathrm{Q}^{t+1}_{\mathrm{BG}}$, and $\widetilde{Q}^t_{\mathrm{SD}}$ from $\widetilde{Q}^{t+1}_{\mathrm{SD}}$, by computing new sets of vectors $\mathcal{V}^t_{\mathbf{a}}$ from those representing the next stage Q-function $\mathcal{V}^{t+1}_{\mathbf{a}'}$. This operation is called the backup operator H and can be performed in roughly two ways. The first way is to exhaustively generate all $|\mathcal{A}| \times |\mathcal{V}^{t+1}|^{|\mathcal{O}|}$ possible vectors and use those as $\mathcal{V}^t$. The second one is to compute a set of joint beliefs for stage t, generating a vector for each of them, resulting in $\mathcal{V}^t$. We will focus on the latter method as it is used by recent approximate POMDP solvers (e.g., (Pineau, Gordon, and Thrun 2003; Spaan and Vlassis 2005)).

The basis of the new vectors is formed by 'back-projected' vectors $\mathbf{g_{ao}}$ from the next time step. For a particular $\mathbf{a}, \mathbf{o}$ and $v \in \mathcal{V}^{t+1}_{\mathbf{a}'}$ they are defined as

$$\mathbf{g}^{v}_{\mathbf{ao}}(s^t) = \sum_{s^{t+1} \in \mathcal{S}} P(\mathbf{o}|\mathbf{a}, s^{t+1}) P(s^{t+1}|s^t, \mathbf{a}) v(s^{t+1}). \tag{19}$$

We denote the set of $\mathbf{g_{ao}}$ for a particular $\mathbf{a}, \mathbf{o}$ (but different next-stage vectors $v^{t+1}_{\mathbf{a}'}$) by $\mathcal{G}_{\mathbf{ao}}$. In the POMDP case, we can define the (finite-horizon and thus not discounted) backup $H^{\mathrm{P}}_{\mathbf{a}b}$ for a particular joint action $\mathbf{a}$ and for the joint belief b of stage t as

$$H^{\mathrm{P}}_{\mathbf{a}b} Q^{t+1}_{\mathrm{P}} = \mathbf{r_a} + \mathbf{f}^{\mathrm{P}}_{\mathbf{a}b}, \tag{20}$$

with $\mathbf{r_a}$ is an $|S|$-dimensional vector, $\mathbf{r_a}(s) = R(s, \mathbf{a})$ and $\mathbf{f}^{\mathrm{P}}_{\mathbf{a}b}$ a vector that expresses the expected future reward

$$\mathbf{f}^{\mathrm{P}}_{\mathbf{a}b} = \sum_{\mathbf{o}} \arg\max_{\mathbf{g_{ao}} \in \mathcal{G}_{\mathbf{ao}}} b \cdot \mathbf{g_{ao}}, \tag{21}$$

where $\mathcal{G}_{\mathbf{ao}}$ is the set of gamma vectors constructed from the sets $\mathcal{V}^{t+1}_{\mathbf{a}'}, \forall_{\mathbf{a}'}$ that represent the next-stage value function Q^{t+1}_{P} (Kaelbling, Littman, and Cassandra 1998).

The $\mathrm{Q_{BG}}$ backup operator uses the same back-projected vectors, but instead of maximizing over all, it only maximizes over those whose next time-step joint action $\mathbf{a}'$ is consistent with a particular joint BG-policy (Oliehoek, Spaan, and Vlassis 2008). This set is defined as

$$\mathcal{G}_{\mathbf{ao}\beta} \equiv \left\{ \mathbf{g}^{v^{t+1}_{\mathbf{a}'}}_{\mathbf{ao}} \mid v^{t+1}_{\mathbf{a}'} \in \mathcal{V}^{t+1}_{\mathbf{a}'} \wedge \beta(\mathbf{o}) = \mathbf{a}' \right\}. \tag{22}$$

The $\mathrm{Q_{BG}}$-backup $H^{\mathrm{B}}_{\mathbf{a}b}$ is completed by maximizing over the BG-policies:

$$H^{\mathrm{B}}_{\mathbf{a}b} Q^{t+1}_{\mathrm{BG}} = \mathbf{r_a} + \mathbf{f}^{\mathrm{B}}_{\mathbf{a}b}, \tag{23}$$

with

$$\mathbf{f}^{\mathrm{B}}_{\mathbf{a}b} = \max_{\beta} \sum_{\mathbf{o}} \arg\max_{\mathbf{g_{ao}} \in \mathcal{G}_{\mathbf{ao}\beta}} b \cdot \mathbf{g_{ao}}. \tag{24}$$

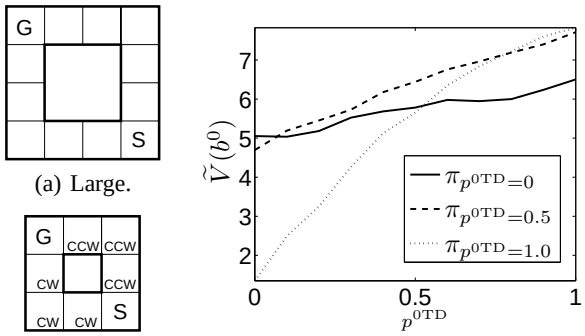
At this point, we can introduce the backup operator $H^{\mathrm{SD}}_{\mathbf{a}b}$ for the $\widetilde{Q}_{\mathrm{SD}}$-value function. It can be seen as a weighted sum of the POMDP and $\mathrm{Q_{BG}}$ backup operators and is defined as

$$H^{\mathrm{SD}}_{\mathbf{a}b} \widetilde{Q}^{t+1}_{\mathrm{SD}} = \mathbf{r_a} + \mathbf{p}^{0\mathrm{TD}}_{\mathbf{a}} \mathbf{f}^{\mathrm{P}}_{\mathbf{a}b} + \mathbf{p}^{D}_{\mathbf{a}} \mathbf{f}^{\mathrm{B}}_{\mathbf{a}b}, \tag{25}$$

where $\mathbf{p}^{0\mathrm{TD}}_{\mathbf{a}}$ is a vector defined as $\mathbf{p}^{0\mathrm{TD}}_{\mathbf{a}}(s) \equiv p^{0\mathrm{TD}}(s, \mathbf{a})$ and similar for $\mathbf{p}^{D}_{\mathbf{a}}$. Note that in this equation the sets $\mathcal{G}_{\mathbf{ao}}, \mathcal{G}_{\mathbf{ao}\beta}$ used by respectively $\mathbf{f}^{\mathrm{P}}_{\mathbf{a}b}, \mathbf{f}^{\mathrm{B}}_{\mathbf{a}b}$ are computed from $\widetilde{Q}^{t+1}_{\mathrm{SD}}$ and therefore different from those in the pure 1TD, 2TD settings. Also, when $\forall s, \mathbf{a}\ p^{0\mathrm{TD}}(s, \mathbf{a}) = 1$, $H^{\mathrm{SD}}_{\mathbf{a}b}$ reduces to $H^{\mathrm{P}}_{\mathbf{a}b}$, while if $\forall s, \mathbf{a}\ p^{0\mathrm{TD}}(s, \mathbf{a}) = 0$, $H^{\mathrm{SD}}_{\mathbf{a}b}$ reduces to $H^{\mathrm{B}}_{\mathbf{a}b}$. The off-line computational effort of $H^{\mathrm{SD}}_{\mathbf{a}b}$ is similar to that of $\mathrm{Q_{BG}}$, as computing the $\mathcal{G}_{\mathbf{ao}\beta}$ sets (22) is the main computational burden.

Experiments

In previous work we have provided experimental results on comparing the POMDP and $\mathrm{Q_{BG}}$ value functions, in the 1TD context (Oliehoek, Spaan, and Vlassis 2007), as well as when used as heuristics for non-communicative Dec-POMDP solving (Oliehoek, Spaan, and Vlassis 2008). Here, we demonstrate how the $\widetilde{Q}_{\mathrm{SD}}$-value function can be applied in SDC scenarios. As the scenarios we target are too large to be solved optimally, we applied an approximate point-based technique, based on (Pineau, Gordon, and Thrun 2003; Spaan and Vlassis 2005). The main idea here is to maintain a PWLC approximate value function, computed on a sampled set of beliefs. All reported sampled control quality values were averaged over $1,000$ independent runs.



(a) Large. (b) Small. (c) Comparing different $p^{0\mathrm{TD}}$.

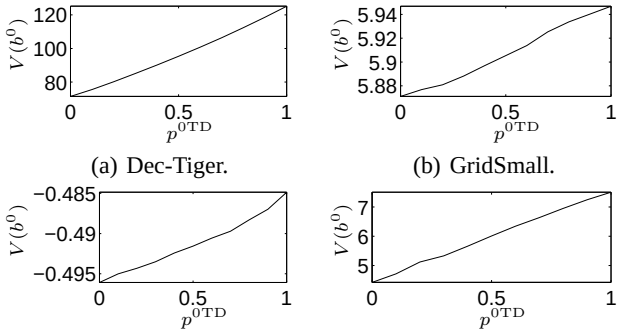
Figure 2: The "Meet in corner" test domain: (a) shows the map of the large variation, and (b) the small version. In (c) we compare different $p^{0\mathrm{TD}}$ in the large version. We show the empirically determined value ($\widetilde{V}(b^0)$, y-axis) of three policies computed for a particular $p^{0\mathrm{TD}}$, evaluated using a varying range of actual $p^{0\mathrm{TD}}$ values (x-axis).

Problem domains

We use a number of two-agent domains, of which Dec-Tiger and GridSmall are standard test problems (details provided by Oliehoek, Spaan, and Vlassis (2008), for instance). One-Door is a noisy version of the OneDoor environment introduced by Oliehoek, Spaan, and Vlassis (2007). "Meet in corner" is a problem in which two robots have to reach a particular corner of their maze, denoted by G in Figures 2(a) and 2(b), after starting in S. They can move clockwise (CW) or counter-clockwise (CCW) with the intended effect 80% of the time, or declare goal when both have reached G, in which case they receive a reward of 10 (and are transported to an absorbing state). Declaring goal at another location or not coordinated is penalized with reward -1 for each agent. When at the goal, agents observe the goal with $p = 1$, in all other states they receive the same non-goal observation.

Delays of up to one time step

In this section we consider settings in which synchronization will be achieved with either 0TD or 1TD, i.e., the setting in which our approximation (12) is exact. First we perform tests when $p^{0\mathrm{TD}}(s, \mathbf{a})$ (and hence $p^{D}(s, \mathbf{a})$) is uniform, i.e., the probability that synchronization does not occur within Δ_c time units is equal $\forall s, \mathbf{a}$. Figure 3 shows the $\widetilde{Q}_{\mathrm{SD}}$-value for the initial belief b^0 for a range of values of $p^{0\mathrm{TD}}$, and $h = 20$. We see that the value increases monotonically with



(a) Dec-Tiger. (b) GridSmall.

(c) OneDoor. (d) Meet in corner (large).

Figure 3: The $\widetilde{Q}_{\mathrm{SD}}$-value $V(b^0)$ (y-axis) for the initial belief b^0, computed for $h = 20$ and varying $p^{0\mathrm{TD}}$ (x-axis), ranging from 0 to 1 in increments of 0.1.

an increasing $p^{0\mathrm{TD}}$. This is expected: when communication is more likely to complete within Δ_c time units, the agents can expect to take more informed decisions. Also this figure clearly illustrates that the effect of communication delays is very much problem dependent. The relative impact in the Dec-Tiger and "Meet in corner" problem is much larger.

Figure 2(c) shows the performance of a joint policy computed for a particular value of $p^{0\mathrm{TD}}$ in the large "Meet in corner" environment in which the actual value of $p^{0\mathrm{TD}}$ is different. We computed $\widetilde{Q}_{\mathrm{SD}}$-value functions for $p^{0\mathrm{TD}} = 0$, 0.5 and 1, and tested them by simulating trajectories in the same environment, but where the value of $p^{0\mathrm{TD}}$ ranged from 0 to 1. It demonstrates how the control quality of a policy that assumes perfect communication ($\pi_{p^{0\mathrm{TD}}=1.0}$) can deteriorate severely when in fact synchronization is never achieved within Δ_c time units (at $p^{0\mathrm{TD}} = 0$ on the x-axis). This highlights the risk of assuming perfect communication, largely ignored in relevant literature, as discussed in the introduction. On the other hand, the policy $\pi_{p^{0\mathrm{TD}}=0.5}$ (which assumes $p^{0\mathrm{TD}} = 0.5$ at computation time) performs well at run time for all values of $p^{0\mathrm{TD}}$ tested.

We also empirically verified that the $\widetilde{Q}_{\mathrm{SD}}$-value function considers potential future communication capabilities. In a small "Meet in corner" variation ($h = 10$), we penalized move actions in the locations labeled CW in Figure 2(b) with reward -0.1, and the CCW ones with -0.15. Hence, the CW route is cheaper, and policies computed for uniform $p^{0\mathrm{TD}} = 1$ and 0 take that route, resulting in values of 5.73 resp. 2.61 (actual sampled control performance). However, when we use a non-uniform communication model, setting $p^{0\mathrm{TD}}(s, \mathbf{a}) = 0, \forall \mathbf{a}, \forall s \in \mathrm{CW}$ and to 1 everywhere else, the $\widetilde{Q}_{\mathrm{SD}}$ policy takes the CCW route obtaining a expected reward of 5.22. The proposed model hence successfully trades off the heavier move penalty with more accurate information resulting from 0TD communication.

Longer delays

Next, we tested our approach for settings in which delays of more than one time step occur. In particular, in Table 1

	a	b	c	d	e	f
$p^{0\mathrm{TD}}$	0.8	0.8	0.6	0.6	0.6	0.6
$p^{1\mathrm{TD}}$	0.2	0.1	0.4	0.3	0.2	0.2
$p^{2\mathrm{TD}}$		0.1		0.1	0.2	0.1
$p^{3\mathrm{TD}}$						0.1

Table 1: Six communication models, "a" through "f", defined by the probability communication will succeed within a certain number of time steps (uniform for all states and joint actions). Empty entries indicate the probability is zero.

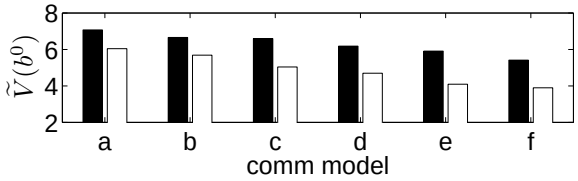


Figure 4: Sampled control performance $\widetilde{V}(b^0)$ for the large "Meet in corner" domain, for the communication models detailed in Table 1. Black bars show the proposed method, and white bars the baseline algorithm.

we defined a set of communication models, ordered roughly by decreasing quality. For instance, we would expect model "f" to have the worst performance, as there is a higher probability of longer delays. Note that communication models with the same $p^{0\mathrm{TD}}$ use the same $\widetilde{Q}_{\mathrm{SD}}$-value function, as it approximates $p^{1\mathrm{TD}}, p^{2\mathrm{TD}}, p^{3\mathrm{TD}}, \ldots$ as $1 - p^{0\mathrm{TD}}$.

We compared algorithm (18), which keeps track of all possible joint beliefs and uses the $\widetilde{Q}_{\mathrm{SD}}$-value function, with the performance of a baseline algorithm. This baseline is identical to (18), except that it uses the POMDP (0TD) value function, and already needs to consider $P(\vec{\theta}^{t-k+l} | \vec{\theta}^{t-k}, \mathbf{a}^{t-k+l})$ when $l = 1$, while the proposed algorithm uses the $\widetilde{Q}_{\mathrm{SD}}$-value function for a delay of 1 time step (and when there is no delay). Figure 4 shows that algorithm (18) consistently outperforms the baseline algorithm, as it takes into account the probability $p^{0\mathrm{TD}}$ of instantaneous communication in the future. As expected, the control quality goes down as communication becomes worse, i.e., as longer delays are more likely.

Conclusions

In this paper we presented an overview of different communication assumptions that are often made in decision-theoretic multiagent planning. In particular we discussed the assumption of instantaneous communication (0TD) (Roth, Simmons, and Veloso 2005; Becker, Lesser, and Zilberstein 2005; Roth, Simmons, and Veloso 2007), as well as one-step delayed communication (1TD) (Schoute 1978; Grizzle, Hsu, and Marcus 1982; Oliehoek, Spaan, and Vlassis 2007). Such models assume that communication is guaranteed to complete within 0 or 1 time steps. However, in the real world such guarantees may be hard to enforce. We in-

troduced a model for stochastically delayed communication (SDC) which more realistically models unreliable communication for such Dec-POMDP settings. The model can handle variable delays, and the delays can be dependent on the state (e.g., agents that are physically close might have stronger communication links). Because computing optimal solutions is impractical, we proposed an approximation for this SDC setting and demonstrated how this approximation can be computed efficiently. Finally, we performed experiments that indicate that a joint policy constructed with an overly positive assumption on communication may result in a severe drop in value. We also empirically demonstrated that, in settings where 0TD communication is beneficial, the joint policy computed by our methods specifies actions that are likely to lead to 0TD communication in the future. Finally, we demonstrated also that delays of more than one time step can be tackled successfully, outperforming a POMDP-based baseline.

There are quite a number of directions for future research of which we mention a few here. In settings where communication delays are typically longer than one stage, the proposed approximation can be crude due to its open-loop nature. For such settings alternative methods should be developed that do take into account $F_{\mathrm{kTD}}, k \geq 2$. However, since k-steps delay problem with $k \geq 2$ are not separable (Varaiya and Walrand 1978), the proposed value-iteration method will not transfer directly to such settings. Another direction is to extend this work to the infinite-horizon SDC setting. As the infinite-horizon $\mathrm{Q_{BG}}$-value function can be approximated with arbitrary accuracy using a PWLC function (Oliehoek, Vlassis, and Spaan 2007), we should be able to naturally extend our results for the SDC setting to the infinite horizon.

Our framework depends on the ability of the agents to synchronize, i.e., to establish common knowledge regarding the individual observation histories. In some settings where not only delayed, but also noisy communication is considered, this may be non-trivial. It may be possible to augment our model to explicitly incorporate the probabilities of transmitting error-free messages. Finally, in our work we have assumed that a model of communication is available. However, in some cases it may be hard to obtain an accurate estimation of communication probabilities a priori. It would be interesting to consider methods that allow the communication model to be learned on-line.

Acknowledgments

We are thankful to Francisco Melo for his useful comments. This work was supported by Fundação para a Ciência e a Tecnologia (ISR/IST pluriannual funding) through the POS_Conhecimento Program that includes FEDER funds, and through grant PTDC/EEA-ACR/73266/2006. The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

References

Becker, R.; Lesser, V.; and Zilberstein, S. 2005. Analyzing myopic approaches for multi-agent communication. In *Proc. of Intelligent Agent Technology*.

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research* 27(4):819–840.

Grizzle, J. W.; Hsu, K.; and Marcus, S. I. 1982. A decentralized control strategy for multiaccess broadcast networks. *Large Scale Systems* 3:75–88.

Halpern, J. Y., and Moses, Y. 1990. Knowledge and common knowledge in a distributed environment. *Journal of the ACM* 37(3).

Hsu, K., and Marcus, S. 1982. Decentralized control of finite state Markov processes. *IEEE Transactions on Automatic Control* 27(2):426–431.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101:99–134.

Oliehoek, F. A.; Spaan, M. T. J.; and Vlassis, N. 2007. Dec-POMDPs with delayed communication. In *Multi-agent Sequential Decision Making in Uncertain Domains*. Workshop at AAMAS07.

Oliehoek, F. A.; Spaan, M. T. J.; and Vlassis, N. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32:289–353.

Oliehoek, F. A.; Vlassis, N.; and Spaan, M. T. J. 2007. Properties of the QBG-value function. IAS technical report IAS-UVA-07-03, University of Amsterdam.

Ooi, J. M., and Wornell, G. W. 1996. Decentralized control of a multiple access broadcast channel: Performance bounds. In *Proc. 35th Conf. on Decision and Control*.

Pineau, J.; Gordon, G.; and Thrun, S. 2003. Point-based value iteration: An anytime algorithm for POMDPs. In *Proc. Int. Joint Conf. on Artificial Intelligence*.

Pynadath, D. V., and Tambe, M. 2002. The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of Artificial Intelligence Research* 16:389–423.

Roth, M.; Simmons, R.; and Veloso, M. 2005. Reasoning about joint beliefs for execution-time communication decisions. In *Proc. of Int. Joint Conf. on Autonomous Agents and Multi Agent Systems*.

Roth, M.; Simmons, R.; and Veloso, M. 2007. Exploiting factored representations for decentralized execution in multi-agent teams. In *Proc. of Int. Joint Conf. on Autonomous Agents and Multi Agent Systems*.

Schoute, F. C. 1978. Decentralized control in packet switched satellite communication. *IEEE Transactions on Automatic Control* 23(2):362–371.

Spaan, M. T. J., and Vlassis, N. 2005. Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research* 24:195–220.

Varaiya, P., and Walrand, J. 1978. On delayed sharing patterns. *IEEE Transactions on Automatic Control* 23(3):443–445.