Balancing coordination and synchronization cost in Cooperative Situated Multi-agent Systems with Imperfect Communication

Andréa I. Tavares and **Mário F. M. Campos**[1]

Abstract. We propose a new Markov team decision model to the decentralized control of cooperative multi-agent systems with imperfect communication. Informational classes capture system's communication semantics and uncertainties about transmitted information and stochastic transmission models, including delayed and lost messages, summarize characteristics of communication devices and protocols. This model provides a quantitative solution to the problem of balancing coordination and synchronization cost in cooperative domains, but its exact solution is computationally infeasible. We propose a generic heuristic approach, based on a off-line centralized team plan. Decentralized decision-making relies on Bayesian dynamic system estimators and decision-theoretic policy generators. These generators use system estimators to express agent's uncertainty about system state and also to quantify expected effects of communication on local and external knowledge. Probabilities of external team behavior, a byproduct of policy generators, are used into system estimators to infer state transition. Experimental results concerning two previously proposed multi-agent tasks are presented, including limited communication range and reliability.

1 INTRODUCTION

In this paper we address the problem that arises when an agent in team must autonomously define its actions to implement a pre-established coordinated plan (pre-plan). This agent acquires information about the state of the environment through sensors that are noisy and with limited capabilities and is able to act upon it with its error prone actuators. Incoherence among agents local perceptions is the main cause of coordination problems in decentralized decision processes. The classical solution is to resort to interagent communication in order to synchronize distinct perceptual information, which brings forth key issues of what, when and to whom communicate. Synchronization incurs in a cost overhead and due to range and load limitations it may be unachievable or delayed. Rational decision making should establish a tradeoff between conformity to pre-plan and synchronization cost.

One possible approach is to derive synchronization policies from a centralized pre-plan by prescribing communication whenever local knowledge is not sufficient to determine the right action [7, 11]. As reliable and costless communication is assumed, this can be seen as a qualitative approach. The quantification of synchronization's actual gain is harder since all costs and uncertainties involved should be considered. In this case, decentralized Markov decision processes (MDPs) [3, 2], and more specifically those with communicating agents [12, 10, 6], are appropriate to quantitatively settle this tradeoff, but they all consider reliable communication. Under total observability assumption, the communication is used in [3] to arrive in a good joint action when ties exist. In our approach, this is solved before execution using lexicographic order. The use of costed communication to synchronize perceptions in partial observable domains where first investigated in [12]. This model assumes perfect communication, that every agent has access to all transmitted messages and that external information is acquired only by communication. Deeper description of the communication process in this model is provided in [6]. Authors define a communication language and the types of messages that agents may transmitted, but reliable communication assumption still remains. There is also no means to model uncertainty about transmitted information and the relation between it and agent's previous knowledge is not explicit. The COM-MTDP [10] also consider interagent communication, relaxing the assumptions about observability. Despite its wide expression power, unreliable communication must be treated by solution's designer since, in the model, the only represented information are the messages that an agent *sends*. Delayed information was treated in the context of decentralized control [1], but not from interagent communication viewpoint.

Since in situated systems failures inhere communication process and information, this article proposes an extension of COM-MTDP that explicitly considers unreliable communication. It also extends previous models by defining a communication language that is associated with system state and allows expressing uncertainty about transmitted information. The communication process resembles the control process. Messages' repositories are the communication state, including a system repository that stores *delayed* messages. Agents change communication state by performing communicative actions (*c-actions*). These c-actions generate messages whose payload depends upon their *informational classes*, the system's communication language. Message transmission features are captured by *delay models* which establish the probability of reception cycles. This model is presented at Section 2.

Previously discussed models [12, 10, 6] have all being solved by heuristic approaches based on expected communication gain, but they relies on external description of communication effect and are not directly applicable with unreliable communication. Generic solutions to decentralized MDPs based on Game Theory equilibrium concepts have been previously proposed [9, 8]. The decision-theoretic approach to Game Theory provides an alternative approach where agents may reason about the process to achieve equilibrium. The Recursive Modelling Method (RMM) [4] defines a framework to per-

[1] Federal University of Minas Gerais, Belo Horizonte, Brasil email: iabrudi@dcc.ufmg.br

form punctual decision making by representing uncertainties about external behavior through alternative models with associated probabilities. It provides a rational acting policy in general MAS and is also suitable for cooperative systems. Communication policies, considering losses in messages but not delays, are also defined in the model [5], based on *information value* which is actually the trade-off between coordination gain and synchronization costs. Section 3 presents a feasible and generic heuristic solution to decentralized decision making, based on RMMs. It relies on state-based utilities matrices (the *pre-plan*), defined before execution and shared among the team. An agent models the *joint* decision making process and defines its control action as the one with maximum expected utility[2]. Bayesian inference is used to summarize agent's knowledge and also teammates' external knowledge. The effect of a c-action onto the decision processes, with delays and losses, is quantified by predicting future knowledge throw Bayesian estimators and evaluating the new utility. The expected gain in coordination is the difference of these utilities and, if it is greater than communication cost, a synchronization takes place.

We have applied our approach to two different multi-agent tasks [12, 10] and experimental results asses its adequacy to generically solve the problem. In Section 4, we analyze these results under the assumption of reliable and unreliable communication (limited range and failures in message transmission). Our concluding remarks are presented in Section 5.

2 THE EMTDP MODEL

The Explicit communicative Markov Team Decision Problem (EMTDP) is a W-step Markovian system for a team $\mathcal{R}$ of n agents, where W is the cycle-threshold to a message's loss. At cycle t, system is in state $s^t \in \mathcal{S}$. Agents have access only to incomplete information about state, basing their acting on local histories. An EMTDP cycle is split into four synchronous stages: *Observation*, where acquired local partial observations of system's state are incorporated into local history; *Communicative Action*, when agents send messages to specific teammates according to their communication policies and cause a transition into communication state; *Communicative Observation*, where received messages are incorporated onto agents' local histories; and *Action*, where control actions are defined according to acting policies. System transits to a new state defined by the joint control action and agents receive a instantaneous *shared* reward.

An EMTDP $\langle \mathcal{R}, W, \mathcal{S}, \mathbf{A}, T_{\mathsf{p}}, \mathbf{O}, o_{\mathsf{p}}, \Sigma_S, \mathbf{\Phi}, e_{\mathsf{p}}, \mathbf{\Sigma}, m_{\mathsf{p}}, r \rangle$ has similar control and communication components:

	Control	Communication
State	$\mathcal{S}$	Σ_S,$\mathbf{\Sigma}$ (repositories)
Action	$\mathbf{A}$	$\mathbf{\Phi}$ (c-action)
Transition	T_{p}	e_{p} (transmission)
Observation	$\mathbf{O}$	$\mathbf{\Sigma}$ (messages)
Likelihood	o_{p}	m_{p} (language)

and control components maintain their original definition [10]. *Joint* action[3] $\boldsymbol{a}^{t-1}$, defined in *Action* stage based on local histories h_x^{t-1}, causes system to evolve to new state s^t according with transition function $T_{\mathsf{p}}(s^{t-1}, \boldsymbol{a}^{t-1}, s^t)$. At cycle t, each agent $\mathrm{R}_x \in \mathcal{R}$ receives local sensorial evidence $o_x^t \in \mathcal{O}_x$ about this state. It is incorporated to local history $h^t_{x\bullet\mathbf{\Sigma}}$ using the joint observation likelihood $o_{\mathsf{p}}(s^t, \boldsymbol{a}^{t-1}, \boldsymbol{o}^t) = \prod_x o_{\mathsf{p}}^x(s^t, a_x^{t-1}, o_x^t)$. Communicative stages add information to local histories in order to improve decision making at *Action* stage. Communication data, which comprises performed c-actions and received messages, are incorporated to the history resulting in h_x^t, that is used to define control action.

The first communication component of EMTDP is the *communication state* $\langle \boldsymbol{\sigma}^t, \sigma_S^t \rangle$, defined by message *repositories*. *System repository* $\sigma_S^t \in \Sigma_S$ (abstractly) stores *delayed* messages, that is, those sent at previous cycles but not received yet. *Agent repository* $\sigma_x^t \in \Sigma_x$ stores messages that are *addressed* to Agent R_x received at cycle t. This state is changed by the performance of c-actions, which generate messages that are store at repositories. Messages' payloads are interpreted by agents using their communication language.

2.1 C-actions and Communication Language

Like in [12], there are three types of c-actions: *inform(I)*, where a local content is transmitted; *request(R)*, where an external content is requested and an answer message is received; and *synchronize(S)*, that is equivalent to perform both previous types, but results in synchronization after one cycle. Messages' semantic (or communication language [6]) is modelled by *informational classes* set $\mathcal{I}$ and messages' transmission, including delays or losses of messages, is modelled by *communication networks* set $\mathcal{D}$ that represents the communication media (physical devices and protocols) employed.

A c-action comprises the choice of a type, a receiver, a informational class and a communication network. The set of c-actions $\mathcal{E}_x$ of Agent R_x is defined as $\mathcal{E}_x = \{I, R, S\} \times \mathcal{R}_{-x} \times \mathcal{D} \times \mathcal{I}$. Each agent may perform up to n_s c-action in a cycle, then $\Phi_x = \{\mathcal{E}_x \bigcup \emptyset\}^{n_s}$. A c-action $e = (o, \mathrm{R}_y, d, i) \in \mathcal{E}_x$ performed by R_x at cycle t generates a message m with a header composed by emitter, recipient, type, informational class, communication network and a *timestamp* t. When c-action's type is R or S, if m actually arrives at the recipient agent, it will be answered in the next cycle. Message's payload depends on its informational class (R messages have no payload).

Communication language is composed by informational classes. Each informational class i, with a payload domain $\mathcal{F}_i$, represents one type of information that may be exchanged among agents, for instance agent's or obstacles' positions, the accomplishment of a subgoal. A probability model defines the relationship among transmitted information and system state, analogously to domain observation likelihood,providing a mechanism to automatically incorporate external data onto agent's knowledge. The *payload likelihood* $f_{\mathsf{p}}^i(s^t, \mathrm{R}_x, c)$ of $c \in \mathcal{F}_i$ is the probability that Agent R_x generates this payload when system state is s^t. It allows that an agent expresses its uncertainty about transmitted information that arises, for instance, when it is acquired using noisy sensors. When informational classes comprises features of system state, *informative* and *world information* types of messages [6] are modelled. Auxiliary procedures, like inference of state given a sub-goal accomplishment or a reward, are necessary to model other types of messages.

The *communicative observation likelihood* $m_{\mathsf{p}}(s^{W:t}, \boldsymbol{\sigma}^t)$[4] defines the probability that payloads of messages in $\boldsymbol{\sigma}^t$ are observed, given the last W states $s^{W:t}$. This is necessary since message's payload is influenced by system's state at the generation cycle and messages may be delayed up to $W-1$ cycles. It is reasonable to assume that the payloads of two messages are conditionally independent given the system state and the emitter agent. The communication observation

[2] Lexicographic order is employed to break ties.

[3] Bold symbols are used for the cartesian product of individual agent's sets, i.e., $\boldsymbol{a} \in \mathbf{A} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_n$ and $a_x \in \mathcal{A}_x$.

[4] Notation $s^{k:t}$ is used for the values assumed by variable s in the cycles $t-k+1$ to t, that is $s^{t-k+1}, \ldots, s^t$.

likelihood can be defined directly from payload likelihoods as

$$m_{\mathsf{p}}(s^{W:t}, \boldsymbol{\sigma}^t) = \prod_{m \in \boldsymbol{\sigma}^t} f_{\mathsf{p}}^{i_m}(s^{t_m}, \mathrm{R}_m, c_m),$$

where m is transmitted by R_m at cycle t_m, i_m is its informational class and e c_m is its payload.

2.2 Messages Transmission and Communicative Transition

As message transmission only depends the employed communication media, we define the *delay model* of each communication network d. Let m be a message sent by Agent R_x to Agent R_y at cycle t using communication network d. The probability that it is received after exactly w cycles is $w_{\mathsf{p}}^d(s^t, \mathrm{R}_x, \mathrm{R}_y, l_m, w)$. Message's delay is conditioned at its length l_m (that depends on its informational class), since the greater a message, the more it is error prone; at system state s^t, specially if a wireless device is used, since topology changes with state and there may be signal interference; and by the communicating agents, as physical distance may be determinant for transmission time. The *loss probability* of m is $w_{\mathsf{p}}^d(s^t, \mathrm{R}_x, \mathrm{R}_y, l_m, W)$ and the probability that it is instantaneously received is $w_{\mathsf{p}}^d(s^t, \mathrm{R}_x, \mathrm{R}_y, l_m, 0)$. We assume that: (a) the delay of a message is defined at its generation cycle and does not change during its transmission; e (b) the delays of two messages are conditionally independent given their generation state, their lengths and their participants. This assumption is true if either transmission capacity is greater enough to avoid collisions or the instantaneous communication load is expressed in system state.

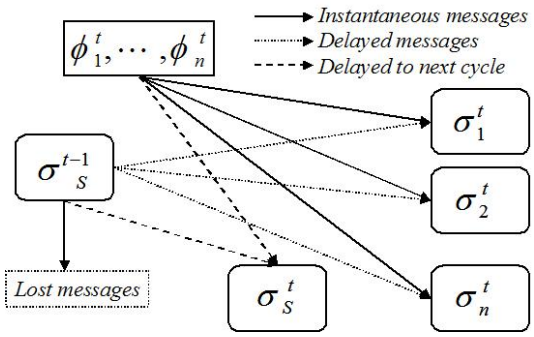

Figure 1. Diagram for communicative transition in EMTDP.

The *communicative transition function* defines the evolution of communication state $\langle \boldsymbol{\sigma}^t, \sigma_S^t \rangle$. Figure 1 shows possible messages' flows among message repositories. At cycle t, a message m generated by R_x's c-action in ϕ_x^t is either instantaneously received (even $E1$ of $m \in \sigma_y^t$ and $m \notin \sigma_S^{t-1}$) or will be delayed or lost (event $E2$, when $m \in \sigma_S^t$ and $m \notin \sigma_S^{t-1}$). Messages in σ_S^{t-1} may be received with delay at the present cycle (event $E3$ of $m \in \sigma_y^t$ and $m \in \sigma_S^{t-1}$), may be delayed by one more cycle (event $E4$ of $m \in \sigma_S^t$ and $m \in \sigma_S^{t-1}$) or is lost (event $E5$). Given the conditional independe assumptions, the communicative transition function $e_{\mathsf{p}}(\sigma_S^{t-1}, s^{W:t}, \boldsymbol{\phi}^t, \boldsymbol{\sigma}^t, \sigma_S^t)$ is defined as the product of individual messages' delay probabilities for each of the five events described above. Event $E1$ has probability $w_{\mathsf{p}}(s^t, \mathrm{R}_x, \mathrm{R}_y, l_m, 0)$ and event $E2$ is complementary it. The the lest three events, the message has been generated at cycle $t - w, w \geq 1$ and has already been delayed for $w - 1$ cycles. If $w = W$, event $E5$ has probability 1 and events $E3$ and $E4$ have both probability 0. Otherwise, event $E5$ has probability 0, and event $E3$ has probability $\frac{w_{\mathsf{p}}(s^{t-w}, \mathrm{R}_x, \mathrm{R}_y, l_m, w)}{\sum_{u=w}^{W} w_{\mathsf{p}}(s^{t-w}, \mathrm{R}_x, \mathrm{R}_y, l_m, u)}$. Event $E4$ is complementary to $E3$ and we define the transition function $e_{\mathsf{p}}(\sigma_S^{t-1}, s^{W:t}, \boldsymbol{\phi}^t, \boldsymbol{\sigma}^t, \sigma_S^t)$ as the product of individual messages' events.

2.3 Instantaneous Reward and Exact Solution

The last component is the *shared instantaneous reward* function $r(s^t, \boldsymbol{a}^t, \boldsymbol{\phi}^t) = v(s^t) - c^{\mathbf{A}}(\boldsymbol{a}^t, s^t) - c^{\mathbf{\Phi}}(\boldsymbol{\phi}^t, s^t)$. The instantaneous gain $v(s)$ is a measure of state quality with respect to task accomplishment. The joint action cost $c^{\mathbf{A}}(\boldsymbol{a}, s)$ allows the designer to express variations in action's cost given the system's state and also the cooperative nature of actions. The joint c-action cost $c^{\mathbf{\Phi}}(\boldsymbol{\phi}^t, s^t)$ captures communication cost dependence on environment and internal features (for instance, this cost may be higher if the agent is in a low power situation). The cooperative nature of communication may be expressed by assigning greater costs to redundant c-actions (two synchronize between the same agents) and to states where the recipient is performing subtasks that are critical.

A team policy $\boldsymbol{\pi}$ is composed by individual *communicative* and *acting* policies π_{Cx}^t and π_{Ax}^t for each cycle t. The communicative policy is based on pre-communication history $h_{x\bullet\mathbf{\Sigma}}^t$ and acting policy, on the pos-communication history h_x^t, where information about communicative stages are already available:

$$\pi_{Cx}^t : (\mathcal{O}_x \times \Phi_x \times \Sigma_x \times \mathcal{A}_x)^{t-1} \times \mathcal{O}_x \to \Phi_x$$
$$\pi_{Ax}^t : (\mathcal{O}_x \times \Phi_x \times \Sigma_x \times \mathcal{A}_x)^{t-1} \times \mathcal{O}_x \times \Phi_x \times \Sigma_x \to \mathcal{A}_x$$

The value $V_{\boldsymbol{\pi}}(s)$ of a team policy with initial state s is a function of the expected shared instantaneous rewards, over all possible system's evolutions[5] e, given an optimization criteria. Under finite horizon T, for instance, it is evaluated as:

$$V_{\boldsymbol{\pi}}(s) = \sum_{e} \sum_{t=1}^{T} T_{\mathsf{p}}(s^{t-1}, \boldsymbol{a}_{\boldsymbol{\pi}}^t, s^t) o_{\mathsf{p}}(s^t, \boldsymbol{a}_{\boldsymbol{\pi}}^t, \boldsymbol{o}^t)$$
$$e_{\mathsf{p}}(\sigma_S^{t-1}, s^{W:t}, \boldsymbol{\phi}_{\boldsymbol{\pi}}^t, \boldsymbol{\sigma}^t, \sigma_S^t) m_{\mathsf{p}}(s^{W:t}, \boldsymbol{\sigma}^t) r(s^t, \boldsymbol{a}_{\boldsymbol{\pi}}^t, \boldsymbol{\phi}_{\boldsymbol{\pi}}^t),$$

where $\boldsymbol{a}_{\boldsymbol{\pi}}^t = (\pi_{A1}^t(h_1^t), \ldots, \pi_{An}^t(h_n^t))$ and $\boldsymbol{\phi}_{\boldsymbol{\pi}}^t = (\pi_{C1}^t(h_{1\bullet\mathbf{\Sigma}}^t), \ldots, \pi_{Cn}^t(h_{n\bullet\mathbf{\Sigma}}^t))$. The optimal solution of EMTDP is the policy $\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} V_{\boldsymbol{\pi}}(s)$ that maximizes the value. As a COM-MTDP may be reduced to an EMTDP, defining the optimal policy is computationally prohibitive [2, 10].

3 BaQuaRA HEURISTIC

Our generic heuristic approach to solve EMTDP is based on myopically evaluating communication gain. It is called BaQuaRA [6] (***Ba**yesian **Qua**ntitative **R**ational **A**cting*) heuristic, for its based on Bayesian inference, quantitative a priori planning and rational decision-making based on modeling joint process by RMM. A RMM M_x for Agent R_x is composed by a n-dimensional game matrix U_x,

[5] A evolution e is the sequence $\langle s, \boldsymbol{o}^1, \boldsymbol{\phi}_{\boldsymbol{\pi}}^1, \boldsymbol{\sigma}^1, \sigma_S^1, \boldsymbol{a}_{\boldsymbol{\pi}}^1, \ldots \boldsymbol{a}_{\boldsymbol{\pi}}^T, s^{T+1} \rangle$.
[6] From tupi, a Brazilian indigene language, *mbae´kwara*, ´things knower´.

the joint action utilities, and alternative models for team external behavior with associated probabilities. Each such model is either a recursive definition of an external agent behavior, by its own game matrix and the models it assigns to the other agents (including R_x), or a probability distribution over its actions. M_x solution results in *Intentional Probability Distributions* (IPDs) $i_x(\boldsymbol{a}_{-x})$ over external actions $\boldsymbol{a}_{-x}$. They summary expected team behavior from R_x's viewpoint when this team has a rational reasoning. The rational policy for R_x is to choose the action that maximizes its utility, given the IPDs for the rest of the team, that is:

$$\pi_{Ax} = \arg u^*(M_x) = \arg \max_{a \in \mathcal{A}_x} \sum_{\boldsymbol{a}_{-x}} U_x(\langle a, \boldsymbol{a}_{-x} \rangle) i_x(\boldsymbol{a}_{-x}),$$

where $u^*(M_x)$ is the value that R_x expects to gain performing its best action. In BaQuaRA approach, agents' game matrix is derived from a pre-plan, which provides $U^t(s)$ utilities (global measurements) of joint actions in each cycle and state of an EMTDP. In this work, pre-plan is obtained from the Q-values of a centralized solution of a fully observable EMTDP, which is equivalent to an MDP. Under full observability, the decentralized policy $\pi^t_{Ax}(s) = \arg_x \max_{\boldsymbol{a} \in \mathbf{A}} U^t(s)(\boldsymbol{a})$ implements the pre-plan. Under partial observability, as the agent does not known actual system state, it maintains probability distributions over $\mathcal{S}$, named *belief state*. Let $b^t_x(s)$ be R_x's belief state at cycle t. Its RMM M^t_x for cycle t decision making has as game matrix the expected value of pre-plan over this belief state: $U_x = \sum_{s \in \mathcal{S}} b^t_x(s) \times U^t(s)$.

Pre-plan also serves to model external behavior models and we proposed three RMM versions: Optimistic, Coordinated State and External Observation. In the first two, there is an alternative team model for each possible state s with probability $b^t_x(s)$. R_x assumes that its teammates actually know system state and expected it to also know system state, at the Optimistic version, that is, they implement $U^t(s)$. In *Coordinated State*, R_x assumes that its teammates expect that system's state is the one defined by implementing pre-plan without error in the execution of actions[7]. These versions are simple, but do not express the actual decision making. In the External Observation version, the agent maintains estimates about external belief states, using EMTDP model and Bayesian inference to assign probabilities to teammates' local histories. It assumes that the team implements an Optimistic RMM over these belief states.

Communication decision making is based on c-actions' *information value* [5], which predicts the effect of a c-action onto agent's RMM. Since communication is unreliable, it is necessary to consider every possible message's reception cycle and its loss ($t+w, 0 \leq w \leq W$). Message's payload is also stochastic and agent has to consider every $f \in \mathcal{F}_i$. The pos-communication value $u^*_{e\bullet}(M^t_x)$ of a c-action e is $\sum_{(w,f)} p_{e\bullet}(w,f) u^*(M^t_x(w,f))$. The probability $p_{e\bullet}(w,f)$ depends on the payload likelihood and the delay model. The value of RMM $M^t_x(w,f)$ is defined after e's message(s) is(are) incorporated into emitter and/or on recipient. The knowledge change is evaluated by prospective Bayesian inference of agents' belief state (for Coordinate State, just the of the information being transmitted). The difference between pre- and pos-communication RMM's values is the *expected coordination gain*, while $\overline{c}(e)$ is the expected communication cost. The information value $i^t(e)$ establishes the tradeoff between coordination and synchronization cost and BaQuaRA agent's communicative policy is the c-action(s) with maximum *positive* information value:

$$\pi^t_{Cx} = \arg \max_{e \in \mathcal{E}_x} i^t(e) = \arg \max_{e \in \mathcal{E}_x} (u^*_{e\bullet}(M^t_x) - u^*(M^t_x)) - \overline{c}(e).$$

Belief states are obtained by Bayesian inference using EMTDP probabilistic. Pre-communication belief state $b^{t+1}_{x\bullet\boldsymbol{\Sigma}}$ summarizes $h^{t+1}_{x\bullet\boldsymbol{\Sigma}}$, incorporating o^{t+1}_x e a^t_x to a priori belief state b^t_x. The main difficulty is that external actions are not directly accessible and state transition depends on *joint* action. We solve that by using IDPs from RMMs and defining the *external action* likelihood $\gamma^t_x(\boldsymbol{a}_{-x}) = (1-\alpha_x) \times N + \alpha_x \times {i_x}^t(\boldsymbol{a}_{-x})$. It is the probability that R_x assigns to its teammates' actions at cycle t, using Bayesian average with a non-informative distribution N to avoid inconsistencies. $b^{t+1}_{x\bullet\boldsymbol{\Sigma}}(s)$ is proportional to:

$$o^x_{\mathsf{p}}(s, a^t_x, o^{t+1}_x) \sum_{s_a \in \mathcal{S}} \left[b^t_x(s_a) \sum_{\boldsymbol{a}^t_{-x}} \left[T_{\mathsf{p}}(s_a, \boldsymbol{a}^t, s) \gamma^t_x(\boldsymbol{a}^t_{-x}) \right] \right]. \quad (1)$$

R_x may maintain an estimative of R_y's belief state $b^t_{x,y\bullet\boldsymbol{\Sigma}}$. For this, external observation probabilities are evaluated as $\mathsf{Pr}(o^t_y) = \sum_s \sum_{a_y} o^y_{\mathsf{p}}(s, a_y, o^t_y) \gamma^{t-1}_x(a_y) b^t_{x\bullet\boldsymbol{\Sigma}}(s)$. Using these probabilities, Equation 1 is applied for every possible observation o^t_y, averaging by their probabilities and using a priori values $b^{t-1}_{x,y}$ e $\gamma^{t-1}_{x,y}(\boldsymbol{a}^{t-1}_{-y})$. IPDs $\gamma^{t-1}_{x,y}$ are obtained from the local RMM model of R_y.

After communicative stages, R_x evaluates its pos-communication belief state b^t_x using its messages σ^t_x and its a priori knowledge $b^t_{x\bullet\boldsymbol{\Sigma}}$. Due to conditional independence of messages, they may be sequentially incorporated into belief state and it is sufficient to define a estimator for a single message e. An instantaneous message carries information about system state s^t and a w-delayed message, about state s^{t-w}. In both cases, there are two sources of information: the transmission and the payload probabilities. The probability that a message m with payload f_m is sent from R_y to R_x and received after exactly w cycles when system state is s is $w_{\mathsf{p}}(s, \mathrm{R}_y, \mathrm{R}_x, l_m, w) f_{\mathsf{p}}(s, \mathrm{R}_y, f_m)$. Let $\mathsf{S_g}$ and $\mathsf{S_r}$ the random variables for the generation and reception states of a message. The belief state $b^t_x(s)$ after the reception of m is:

$$\propto \mathsf{Pr}(\mathsf{S_r} = s) \sum_{s_g \in \mathcal{S}} \left[\mathsf{Pr}(m | \mathsf{S_g} = s_g) \mathsf{Pr}(\mathsf{S_r} = s | \mathsf{S_g} = s_g) \mathsf{Pr}(\mathsf{S_g} = s_g) \right]$$

$$\propto b^t_{x\bullet\boldsymbol{\Sigma}}(s) \sum_{s_g \in \mathcal{S}} \Big[w_{\mathsf{p}}(s_g, \mathrm{R}_y, \mathrm{R}_x, l_m, w) f_{\mathsf{p}}(s_g, \mathrm{R}_y, c_m)$$

$$\mathsf{Pr}(s_g \text{ evolves to } s \text{ in } w \text{ cycles}) \times b^{t-w}_x(s_g) \Big].$$

where the evolution probability is evaluated using agent's actions and its external action likelihoods for the last $w-1$ cycles. The same inference procedure can be prospectively used to infer the effect of a c-action onto agent's future belief state. Every pair (w,f) rises a specific belief state and an associated RMM. When $w > 0$, future system's transitions are obtained by evaluating RMMs for future cycles and thus changing belief states according to external actions likelihood and not taking future observations into account.

4 EXPERIMENTAL RESULTS

The TER system [10] is a two helicopter-agent flight across enemy territory to reach a destination position as soon as possible. Helicopter T has no firepower but Helicopter E can destroy the enemy

[7] Suitable for transition independent domains with known initial state [12, 6].

radar unit located in an unknown position (from 1 to 8) along their 10-unit length path. To espace detection by the radar, T travels at lower speed. When E, that always flies at the same speed, arrives at the radar, it destroys it with certainty. After this, T starts to fly faster if it observes the destruction, event whose likelihood is a function of the *observability* parameter λ (within the range $[0, 1]$) and T's from the radar. E may communicate the destruction to T at fixed cost c within the range $[0, 1]$. A COM-MTDP model [10] has been purposed for TER to optimally balance out coordination improvement and communication cost. EMTDP model is an extension of it consisting of one informational class for deterministic radar's destruction, a unique communication network and E's Inform c-action. Delay model introduces communication unreliability in the sense that a message may be lost with probability p_c. BaQuaRA's solution is composed by an External Observation version to E and an Optimistic version to T. Pre-plan is obtained from a centralized MDP for TER.

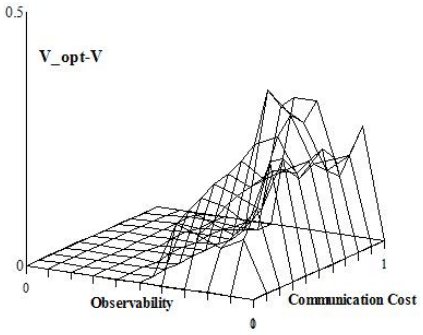

Figure 2. Suboptimality of BaQuaRA under reliable communication.

As noted in [10], optimal solution to TER implements a kind of plan recognition: sometimes it is better to E to delay its communication in order to infer if T has observed radar's destruction. Though plan recognition is not pre-defined in BaQuaRA solution, it is a consequence from RMM modelling. Still, our heuristic gain is sometimes suboptimal, as it is shown by its difference from optimal's gain in Figure 2. This is mainly due to little inconsistencies between the estimated T's belief state maintained by E and its actual value, when T does not receive an observation. Pre-plan indicates, from E's viewpoint, that T will fly faster, but it actually happens that it flies slowly and the team looses gain. Another source of suboptimality follows from the agent's incapability to reason about the lack of communication. If it remains just one possible position for the radar, RMM prescribes faster motion to T, even at the expense of being destroyed. Knowing that E would have informed it about the radar destruction, T could better define its action.

A typical performance degradation for low observability[8], due to unreliable communication, is shown in Figure 3, for $\lambda = 0.3$. For high observability settings, results indicates that unreliability has little effect. The most interesting point is that performance does degrade with greater loss probability, but communication cost has little effect as we fix p_c. This is a consequence of automatically adjusting the number of sent messages, as shown in right graphic of Figure 3. For low communication costs, the expected number of messages increases with loss probability. E retransmits of a message if it is lost

[8] In the graphic, we are compare gains with reliable optimal ones, so the suboptimality may be slower.

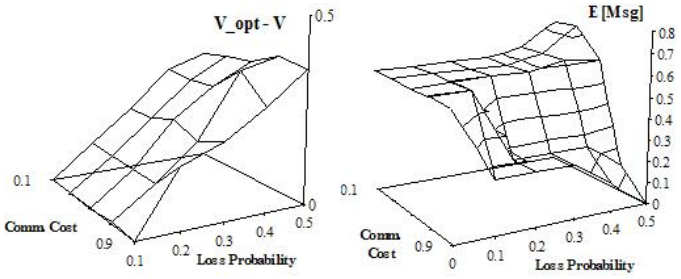

Figure 3. Performance under unreliable communication ($\mu = 0.3$).

(fact that it can implicitly detect) and the information value of resending it is positive. As communication cost increases, the number of expected messages *decreases*: message loss probability is a part of communication decision and the higher is the loss probability, the lower is information value until a threshold where no communication is performed.

5 CONCLUSIONS

The EMTDP model subsumes previous purposes as it provides the capability to analyze the optimality of team decisions under *imperfect* communication. As communication in situated systems is inherently unreliable, EMTDP enlarges the class of formally tractable systems. Messages' transmission and communication language are part of the model, allowing reasoning about information gain and its incorporation in local knowledge avoiding the use of auxiliary external procedures. We have proposed a novel *generic* heuristic approach that provides rational decentralized policies dealing with *lost* and *delayed* messages. Such a generic solution is important since unreliable communication is just starting to be systematically treated and insights can be gained from the prescribed solutions. In particular, we have shown that BaQuaRA's performance is suitable, despite its generality, to define almost optimal policies for a specific system. Ongoing work with different systems [12, 6] shows that this quality remains. Our future research includes the investigation of message delay impact on team decision and the extension of BaQuaRA to fulfills the reasoning about the lack of communication.

References

[1] M. Aicardi, D. Franco, and R. Minciardi, 'Decentralized optimal control of Markov chains with a common past information set', *IEEE Transactions on Automatic Control*, **32**(11), 1028–1031, (1987).

[2] D. S. Bernstein, S. Zilberstein, and N. Immerman, 'The complexity of decentralized control of Markov decision processes', in *UAI-2000*, pp. 32–37, (2000).

[3] C. Boutilier, 'Sequential optimality and coordination in multiagent systems', in *IJCAI-99*, pp. 478–485, (1999).

[4] P. J. Gmytrasiewicz and E. H. Durfee, 'Rational coordination in multi-agent systems', *Autonomous Agents and Multi-Agent Systems Journal*, **3**(4), 319–350, (2000).

[5] P. J. Gmytrasiewicz and E. H. Durfee, 'Rational communication in multi-agent systems', *Autonomous Agents and Multi-Agent Systems Journal*, **4**(3), 233–272, (2001).

[6] C. V. Goldman and S. Zilberstein, 'Optimizing information exchange in cooperative multi-agent systems', in *AAMAS-03*, (2003).

[7] H. Jung, M. Tambe, and S. Kulkarni, 'Argumentation as distributed constraint satisfaction: Applications and results', in *AGENTS01*, (2001).

[8] R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella, 'Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings', in *IJCAI-2003*, (2003).
[9] L. Peshkin, K. Kim, N. Meuleau, and L. P. Kaelbling, 'Learning to cooperate via policy search', in *UAI-2000*, (2000).
[10] D. V. Pynadath and M. Tambe, 'The communicative multiagent team decision problem: Analyzing teamwork theories and models', *Journal of Artificial Intelligence Research*, **16**, (2002).
[11] P. Xuan and V. Lesser, 'Multi-agent policies: From centralized ones to decentralized ones', in *AAMAS02*, (2002).
[12] P. Xuan, V. Lesser, and S. Zilberstein, 'Communication decisions in multi-agent cooperation: Model and experiments', in *AGENTS01*, (2001).