A POMDP approach to cooperative localization in sparse environments

Francisco S. Melo[*] Isabel Ribeiro[*]

Abstract

In this paper we discuss how communication can be used advantageously for cooperative navigation in sparse environments. Specifically, we analyze the tradeoff between the cost of communication cost and the efficient completion of the navigation task. We make use of a partially observable Markov decision process (POMDP) to model the navigation task, since this model allows to explicitly consider the tradeoff between information-gathering actions and actions that move the robot towards the goal. By explicitly including communication in the POMDP model as an information-gathering action with an associated cost, we are able to optimally settle this tradeoff between the gain in information arising from the use of communication and the corresponding cost. We illustrate our results in a small test application.

1 Introduction

Consider a group of robots moving in a sparse environment described by a topological map with M nodes. We refer to the nodes in the map as *states*. Each robot must navigate from an initial state to a goal state, known only to that single robot. We admit that several states in the environment have *distinctive landmarks* that the robots can generally perceive through its sensors. However, until a robot is able to reach one such state and observe the corresponding landmark, it must generally navigate through several other states receiving no sensorial feedback from the environment (*e.g.*, using only dead-reckoning).

In this paper we focus only on the problems of global localization and navigation/planning, disregarding other problems such as motion control or obstacle avoidance. We model the navigation task as a sequence of decisions: at each decision instant, each robot must choose from a set of action primitives that control its movements. The robots are allowed to communicate with each other and *share* sensorial information. This received sensorial data can then be used by the robot to improve its localization in the environment.

However, if communication can be used to improve the sensorial capabilities of each robot, it is also true that the communication process generally takes time and consumes resources.

[*] The authors are with the Institute for Systems and Robotics, Instituto Superior Técnico, Lisboa, Portugal. E-mail: [fmelo,mir]@isr.ist.utl.pt. This work was partially supported by POS_C that includes FEDER funds. The first author acknowledges the PhD grant SFRH/BD/3074/2000.

Furthermore, it may happen occasionally that no useful information is received. Therefore, it is important to realize in which situations is communication advantageous and in which situations it should be avoided. For example, it may happen that the cost of communication is much higher than the benefit obtained from it. To settle this problem, we make use of a *partially observable Markov decision process* (POMDP) to model each robot in the environment. POMDPs have successfully been used for topological navigation [1–3] and are particularly amenable to the use of Markov localization methods [4]. Furthermore, POMDPs explicitly consider the tradeoff between choosing actions to *disambiguate the state of the robot* (information-gathering) and actions to *move the robot towards the goal.* By explicitly including *communication* in the POMDP model as an information-gathering action with an associated cost, we are able to optimally settle this tradeoff between the gain in information arising from the use of communication and the corresponding cost. This appealing feature of the POMDP framework has led some researchers to address *active sensing* using POMDP models [5].

The paper is organized as follows. In Section 2 we introduce the general POMDP framework. In Section 3 we apply this framework to model the particular problems addressed in the paper. We introduce a simple illustrative example that is used throughout the paper to illustrate the main ideas. Finally, Section 4 concludes the paper with a summary of the main conclusions and points out several possible directions for future work.

2 Partially observable Markov decision processes

A POMDP is a tuple $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathsf{P}, \mathsf{O}, r, \gamma)$, where $\mathcal{X}$ is the finite set of possible states of the system, $\mathcal{A}$ is a finite set of control primitives and $\mathcal{Z}$ corresponds to a finite set of possible observations. At each time instant t, a decision-maker chooses an action A_t depending on the past history of events, causing the system to move from its current state X_t to state X_{t+1}. We denote by $\mathsf{P}_a(i, j)$ the probability of moving from state i to state j under action a. As soon as the transition occurs, the decision-maker receives an observation Z_{t+1} that depends on the new state of the system. We denote by $\mathsf{O}_a(j, z)$ the probability of $Z_{t+1} = z$ when $X_{t+1} = j$ and $A_t = a$. Also, as soon as the transition occurs, the decision-maker is granted a numerical reward $r(i, a, j)$, verifying $|r(i, a, j)| \leq R_{\max}$. The purpose of the decision-maker is to choose the control sequence $\{A_t\}$ so as to maximize the functional

$$V(b_0, \{A_t\}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t, X_{t+1}) \mid X_0 \sim b_0\right], \tag{1}$$

where $\gamma < 1$ is a positive discount factor, b_0 is the *initial belief* for the process and $X_0 \sim b_0$ denotes the fact that X_0 is distributed according to b_0. The initial belief b_0 is a probability vector describing the initial distribution of the state of the system.

In this paper, we are interested in using a POMDP model to describe a robot moving in a sparse environment described topologically. Using a POMDP model, the state-space $\mathcal{X}$ corresponds to the set of sites in the environment. The state of the system at time t, X_t, corresponds to the position of the robot in the environment at time t. The control primitives, or *actions*, correspond to the high-level navigation commands that allow the robot to move between sites in the environment and the observations correspond to the sensorial information. With a POMDP model, Markov localization [4] can be implemented in a straightforward way: at each time instant t the robot maintains a *belief-vector* b_t, each component $b_t(i)$ describing the probability of being in a particular state $i \in \mathcal{X}$. This belief-vector is updated componentwise as

$$b_{t+1}(j) = \mathsf{B}_a(b, z)_j = \frac{\sum_{i \in \mathcal{X}} b_t(i) \mathsf{P}_a(i, j) \mathsf{O}_a(j, z)}{\sum_{i,k \in \mathcal{X}} b_t(i) \mathsf{P}_a(i, k) \mathsf{O}_a(k, z)},$$

where $A_t = a$ and $Z_{t+1} = z$ and $\mathsf{B}_a(b, z)_j$ represents the jth component of the vector $\mathsf{B}_a(b, z)$.

The *optimal value function* for a POMDP is defined as $V^*(b) = \max_{\{A_t\}} V(\{A_t\}, b)$ and verifies the following recursive relation

$$V^*(b) = \max_{a \in \mathcal{A}} \sum_{i,j \in \mathcal{X}} b(i) \mathsf{P}_a(i, j) \left[r(i, a, j) + \gamma \sum_{z \in \mathcal{Z}} \mathsf{O}_a(j, z) V^*\big(\mathsf{B}_a(b, z)\big) \right].$$

The value $V^*(b)$ represents the total expected discounted reward received along an optimal trajectory starting from the initial state distribution b. The optimal decision rule can be defined by means of the mapping

$$\pi^*(b) = \arg\max_{a \in \mathcal{A}} \sum_{i,j \in \mathcal{X}} b(i) \mathsf{P}_a(i, j) \left[r(i, a, j) + \gamma \sum_{z \in \mathcal{Z}} \mathsf{O}_a(j, z) V^*\big(\mathsf{B}_a(b, z)\big) \right].$$

is called the *optimal policy* for the POMDP $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathsf{P}, \mathsf{O}, r, \gamma)$.

There are numerous methods in the literature to compute the optimal policy for a POMDP (see the survey works [6, 7]). In this paper we adopt the incremental prunning (IP) algorithm [8]. Further details can be found in an extended version of this paper.

3 The POMDP model

We are interested in addressing the situation in which a group of robots moves in a sparse environment, each robot trying to reach its goal location. We use a POMDP model to analyze how the robots can benefit from efficiently using communication, by explicitly including *communication* in the POMDP model as an information-gathering action with an associated cost. The optimal POMDP policy optimally settle this tradeoff between the gain in information arising from the use of communication and the corresponding cost.

Two important observations are in order. First of all, communication between two robots is modeled as a *directed exchange of sensorial information* and we assume communication to be *peer-to-peer* and not broadcasted.

Secondly, even though we consider the existence of multiple communicating robots in the environment, *we do not address the interaction between these robots.* In particular, the actions of one robot do not affect the behavior of any other robot nor its ability to reach the goal. Therefore, each robot can be modeled independently of the other robots and there is no need to consider multi-agent decision models, such as Dec-POMDPs or stochastic games.

From the discussion above, it should be clear that we can focus our analysis on the behavior of a *single robot*, considering the other robots as part of the environment. We resort to a simplified model that encompasses all the fundamental features of the class of navigation problems considered in the paper. This model is represented in Figure 1.

In this simplified model, a robot departs from the “Start” state by choosing any of the two available actions a or b. It then moves to either state 1 or state 2 with equal probability. The

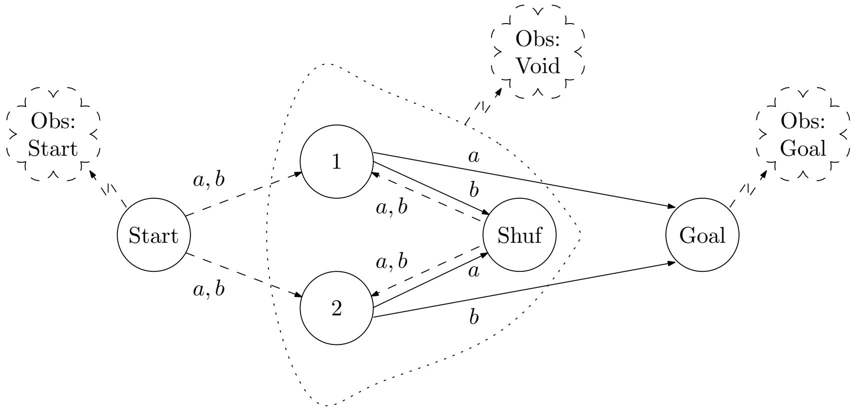

Figure 1: A general sparse environment.

Table 1: Comparison between the performance of the cooperative and the non-cooperative robots.

Test	**Total disc. reward**
Non-cooperative	58.118 ± 22.185
Cooperative	81.498 ± 9.867

robot will then move to the "Goal" state by choosing a in state 1 or b in state 2 and to the state marked as "Shuf", otherwise. Upon reaching the Goal state, the robot receives a reward of +20 and upon reaching the Shuf state, it receives a reward of -5. At the Shuf state, independently of the robot's action, its position is randomly reset to either state 1 or 2, with equal probability. In this model, the robot has 3 available observations: "Start", in the Start state, "Void", in the states inside the dotted line, and "Goal", in the goal state.

In terms of navigation in sparse environments, the set of three undistinguishable states describes those situations in which the robot gets lost due to long periods of dead-reckoning. Upon reaching the Goal state, the robot is back to a location with distinguishing features and can use this information to localize once again. In this simplified model, the robot merely ignores the existence of other robots and just chooses its actions so as to maximize its total reward (reaching the Goal state as quiclly as possible).

We ran IP to compute the optimal policy for the problem above and tested the performance of the obtained policy by running 2000 independent Monte Carlo trials, each consisting of a 10-time step trajectory. We then computed the average total discounted reward. The results are reported in Table 1.

Notice that, after the first action, the robot will get lost between states 1 and 2 and can only "bet" in one of the two possible actions a and b, hoping that it will lead to the desired outcome. However, whatever action the robot chooses, it lead to a successful outcome only 50% of the time. A particularly "lucky" run can bring quite a large reward, while a particularly "unlucky" run can bring an alarmingly large penalty. This justifies the large standard deviation observed in the results portrayed in Table 1. An important aspect of the behavior just described is that it matches the one expected from a robot navigating in a sparse environment: when it gets lost, it keeps moving in a direction where a recognizable state is expectable.

We now describe how the model in Figure 1 is modified to include the existence of another robot in the environment. In particular, and unlike the situation analyzed before, the robot

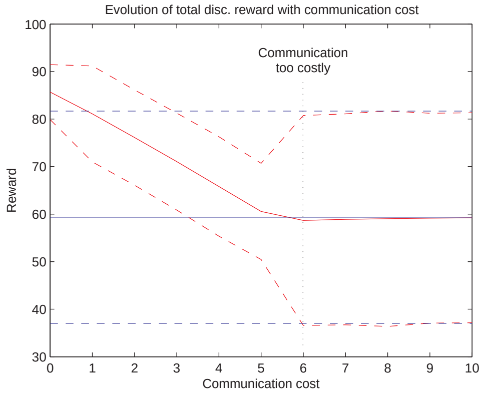

Figure 2: Tradeoff between the cost of communication, and the value of information. Blue line – Non-cooperative robot; Red line – Cooperative robot.

has now two further actions available, dubbed as “Comm 1” and “Comm 2”. None of the latter actions affects the position of the robot in the environment. Instead, each action “Comm i”, $i = 1, 2$, sends a message “from” state i. This message is sent to a second robot (Robot B) who, upon receiving it, will send the first robot (Robot A) its sensorial information. Notice that such communication actions only suceed in the corresponding states. This means that, if the robot sends a message “Comm i” when at state $j \neq i$, there is a high probability of receiving no reply. Finally, the robot receives a reward of -1 for every communication action.

Two important observations are in order. First of all, notice that there is a cost involved in communication, even if less significant than getting into the “Shuff” state. Secondly, the communication may not succeed. This can happen either because Robot A chose the “wrong” communication action, or simply because Robot B can give Robot A no information on its position (*e.g.*, Robot B cannot “see” Robot A).

We ran IP and computed the optimal policy for this new scenario to compare the performance of the robot with the one observed from the non-cooperative robot. As before, we tested the performance of the optimal policy by running 2000 independent Monte Carlo trials, each consisting of a 10-time step trajectory, and computed the average total discounted reward. The results are reported in Table 1.

Notice, first of all, the tremendous difference in performance between the two robots. The robot relying on communication exhibited an average increase in performance of about 30%, even receiving the negative rewards arising from communication. Furthermore, the optimal policy in the presence of communication leads to a much more reliable performance, since the observed standard deviation is much smaller.

Finally, to assess the explicit tradeoff between the cost of communication and the “value of information”, we have conducted similar tests, varying the communication cost, r_{comm}, between 0 and -10. The corresponding results are summarized in Figure 2, where the solid lines correspond to the mean total discounted reward over 2000 independent Monte-Carlo runs and the dotted line the corresponding standard deviation.

Notice that, as the cost of communication increases, the performance of the cooperative (communicating) robot approaches that of the non-cooperative robot. The two performances reach a similar level when the cost of communication is similar to that of getting lost (*i.e.*, $r_{\text{comm}} = -5$). An important aspect to emphasize is that, when $r_{\text{comm}} = -5$, the performance of the optimal policy in the cooperative case is much more *reliable* than that of the non-cooperative case. This is easily seen from the standard deviation observed. Therefore, even at a high cost, the robot does rely on communication to navigate and this actually translates in an actual improvement in terms of performance. Finally, for $r_{\text{comm}} \geq 6$, communication becomes too costly, as indicated in Figure 2. This means that the optimal policy in both the cooperative and non-cooperative case are similar, as seen from the performance observed in Figure 2.

4 Conclusions and future work

In this paper we addressed the problem of cooperative localization and navigation in sparse environments. We showed that, even if it is possible for a robot moving in such an environment to reach its goal without ever considering the existence of other robots in the environment, communication can greatly improve its overall performance. We made use of a POMDP model to explicitly consider the tradeoff between choosing actions to disambiguate the state of the robot (information-gathering) and actions to move the robot towards the goal. By explicitly including communication in the POMDP model as an information-gathering action with an associated cost, we were able to optimally settle this tradeoff between the gain in information arising from the use of communication and the corresponding cost.

Our results suggest several interesting avenues for future research. First of all, decision-theoretic models as the one described in this paper can be further explored to analyze situations in which communication is used to combine the sensing information from different sources, so as to optimize the total information obtained from the sensorial data. Secondly, multi-agent variations of the model used here can also be used to address *active sensor networks*, by casting such networks as communicating, cooperative multi-agent systems.

References

[1] N. Roy and S. Thrun, "Coastal navigation with mobile robot," in *NIPS 13*, 1999, pp. 1043–1049.

[2] R. Simmons and S. Koenig, "Probabilistic robot navigation in partially observable environments," in *IJCAI'95*, 1995, pp. 1080–1087.

[3] F. Melo and I. Ribeiro, "Transition entropy in partially observable Markov decision processes," in *IAS-9*, 2005, pp. 282–289.

[4] D. Fox, "Markov localization: A probabilistic framework for mobile robot localization and navigation," Ph.D. dissertation, University of Bonn, 1998.

[5] S. Whitehead and D. Ballard, "Learning to perceive and act by trial and error," *Machine Learning*, vol. 7, no. 1, pp. 45–83, 1991.

[6] D. Aberdeen, "A (revised) survey of approximate methods for solving partially observable Markov decision processes," National ICT Australia, Tech. Rep., 2003.

[7] A. Cassandra, "Optimal policies for partially observable Markov decision processes," Dep. Computer Sciences, Brown University, Tech. Rep. CS-94-14, 1994.

[8] A. Cassandra, M. Littman, and N. Zhang, "Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes," in *UAI'97*, 1997, pp. 54–61.