Coordination vs. information in multi-agent decision processes

Maike Kaufman and Stephen Roberts
Department of Engineering Science
University of Oxford
Oxford, OX1 3PJ, UK
{maike, sjrob}@robots.ox.ac.uk

ABSTRACT
Agent coordination and communication are important issues in designing decentralised agent systems, which are often modelled as flavours of Markov Decision Processes (MDPs). Because communication incurs an overhead, various scenarios for sparse agent communication have been developed. In these treatments, coordination is usually considered more important than making use of local information. We argue that this is not always the best thing to do and provide alternative approximate algorithms based on local inference. We show that such algorithms can outperform the guaranteed coordinated approach in a benchmark scenario.

Categories and Subject Descriptors
I.2.11 [Computing Methodologies]: Distributed Artificial Intelligence—Multiagent systems

General Terms
Algorithms

Keywords
Multiagent communication, Multiagent coordination, local decision-making

1. INTRODUCTION
Various flavours of fully and partially observable Markov Decision Processes (MDPs) have gained increasing popularity for modelling and designing cooperative decentralised multi-agent systems [11, 18, 23]. In such systems there is a trade-off to be made between the extent of decentralisation and the tractability and overall performance of the optimal solution. Communication plays a key role in this as it increases the amount of information available to agents but creates an overhead and potentially incurs a cost. At the two ends of the spectrum lie fully centralised multi-agent (PO)MDPs and completely decentralised (PO)MDPs. Decentralised (PO)MDPs have been proven to be NEXP-complete [5, 18] and a considerable amount of work has gone into finding approximate solutions, e.g. [1, 2, 4, 7, 8, 10, 12, 15, 16, 14, 21, 22, 25].

Most realistic scenarios arguably lie somewhere in between full and no communication and some work has focused on scenarios with more flexible amounts of communication. Here, the system usually alternates between full communication among all agents and episodes of zero communication, either to facilitate policy computation for decentralised scenarios [9, 13], to reduce communication overhead by avoiding redundant communications [19, 20] and/or to determine a (near-)optimal communication policy in scenarios where communication comes at a cost [3, 26]. Some of these approaches require additional assumptions, such as transition independence. The focus in most of this work lies on deciding when to communicate, either by pre-computing communication policies or by developing algorithms for on-line reasoning about communication. Such treatment is valuable for scenarios in which inter-agent communication is costly but reliably available. In many real-world systems on the other hand, information exchange between agents will not be possible at all times. This might, for example, be due to faulty communication channels, security concerns or limited transmission ranges. As a consequence agents will not be able to plan ahead about when to communicate but will have to adapt their decision-making algorithms according to whichever opportunities are available. We would therefore like to view the problem of sparse communication as one of good decision-making under differing beliefs about the world. Agents might have access to local observations, which provide some information about the global state of the system. However, these local observations will in general lead to different beliefs about the world and making local decision-choices based on them could potentially lead to uncoordinated collective behaviour. Hence, there is again a trade-off to be made: should agents make the most of their local information, or should overall coordination be valued more highly?

Existing work seems to suggest that coordination should in general be favoured over more informed local beliefs, see for example [8, 19, 20, 24], although the use of local observations has shown some improvement of performance to an existing algorithm for solving DEC-POMDPs [7]. We would like to argue more fundamentally here that focusing on guaranteed coordination will often lead to lower performance and that a strong case can be made for using what local information is available in the decision-making process. For simplicity we will concentrate on jointly observable systems with uniform transition probabilities and free communication, in which agents must sometimes make decisions without being able to communicate observations. Such simple 1-step scenarios could be solved using a dec-POMDP or dec-MDP, but in more complicated settings (e.g. when varying subsets of
agents communicate or when communication between agents is faulty) application of a dec-(PO)MDP is not straightforward and possibly not even intractable. Restricting the argument to a subset of very simple scenarios arguably limits its direct applicability to more complex settings, especially those with non-uniform transition functions. However, it allows us to study the effects of different uses of information in the decision-making algorithms in a more isolated way. Taking other influencing factors such as the approximation of infinite-horizon policy computation into account at this stage, would come at the cost of a less rigorous treatment of the problem. The argument for decision-making based on local information is in principle extendable to more general systems and we believe that understanding the factors which influence the trade-off between coordination and local information gain for simple cases ultimately enable the treatment of more complicated scenarios. In that sense this paper is intended as a first proof of concept.

In the following we will describe exact decision-making based on local beliefs and discuss three simple approximations by which it can be made tractable. Application of the resulting local decision-making algorithms to a benchmark scenario show that they can outperform an approach based on guaranteed coordination for a variety of reward matrices.

2. MULTI-AGENT DECISION PROCESS

Cooperative multi-agent systems are often modelled by a Multi-agent MDP (MMDP) [6], Multi-agent POMDP, [18], decentralized MDP (dec-MDP) [5] or decentralized POMDP (dec-POMDP) [5]. A summary of all these approaches can be found in [18].

Let \( \{N, S, A, O, p_T, p_O, \Theta, R, B\} \) be a tuple where:

- \( N \) is a set of \( n \) agents indexed by \( i \)
- \( S = \{S^1, S^2, \ldots \} \) is a set of global states
- \( A_i = \{a^1_i, a^2_i, \ldots \} \) is a set of local actions available to agent \( i \)
- \( A = \{A^1, A^2, \ldots \} \) is a set of joint actions with \( A = A_1 \times A_2 \times \ldots \times A_n \)
- \( O_i = \{\omega^1_i, \omega^2_i, \ldots \} \) is a set of local observations available to agent \( i \)
- \( O = \{O^1, O^2, \ldots \} \) is a set of joint observations with \( O = O_1 \times O_2 \times \ldots O_n \)
- \( p_T : S \times S \times A \rightarrow [0, 1] \) is the joint transition function where \( p_T(S^0|A^k, S^p) \) is the probability of arriving in a state \( S^q \) when taking action \( A^k \) in state \( S^p \)
- \( p_O : S \times O \rightarrow [0, 1] \) is a mapping from states to joint observations, where \( p_O(O^k|S^i) \) is the probability of observing \( O^k \) in state \( S^i \)
- \( \Theta : O \rightarrow S \) is a mapping from joint observations to global states
- \( R : S \times A \times S \rightarrow \mathbb{R} \) is a reward function, where \( R(S^p, A^k, S^q) \) is the reward obtained for taking action \( A^k \) in a state \( S^p \) and transitioning to \( S^q \)
- \( B = (b_1, \ldots, b_n) \) is the vector of local belief states

A local policy \( \pi_i \) is commonly defined as a mapping from local observation histories to individual actions, \( \pi_i : \omega_i \rightarrow A_i \). For the purpose of this work, let a local policy more generally be a mapping from local belief states to local actions, \( \pi_i : b_i \rightarrow A_i \), and let a joint policy \( \pi \) be a mapping from global (belief-)states to joint actions, \( \pi : S \rightarrow A \) and \( \pi : B \rightarrow A \) respectively. Depending on the information exchange between agents, this model can be assigned to one of the following limit cases:

- **Multi-Agent MDP** If agents have guaranteed and free communication among each other and \( \Theta \) is a surjective mapping, the system is collectively observable. The problem simplifies to finding a joint policy \( \pi \) from global states to joint actions.
- **Multi-Agent POMDP** If agents have guaranteed and free communication but \( \Theta \) is not a surjective mapping, the system is collectively partially observable. Here the optimal policy is defined as a mapping from belief states to actions.
- **DEC-MDP** If agents do not exchange their observations and \( \Theta \) is a surjective mapping, the process is jointly observable but locally only partially observable. The aim is to find the optimal joint policy consisting of local policies \( \pi = (\pi_1, \ldots, \pi_n) \).
- **DEC-POMDP** If agents do not exchange their observations and \( \Theta \) is not a surjective mapping, the process is both jointly and locally partially observable. As with the DEC-MDP the problem lies in finding the optimal joint policy comprising local policies.

In all cases the measure for optimality is the discounted sum of expected future rewards. For systems with uniform transition probabilities in which successive states are equally likely and independent of the actions taken, finding the optimal policy simplifies to maximising the immediate reward:

\[
V_{\pi}(S) = R(S, \pi(S))
\]  

3. EXACT DECISION-MAKING

Assume that agents are operating in a system in which they rely on regular communication, e.g. a MMDP, and that at a certain point in time they are unable to fully synchronise their respective observations. This need not mean that no communication at all takes place, only that not all agents can communicate with all others. In such a situation their usual means of decision-making (the centralised policy) will not be of use, as they do not hold sufficient information about the global state. As a result they must resort to an alternative way of choosing a (local) action. Here, two general possibilities exist: agents can make local decisions in a way that conserves overall coordination or by using some or all of the information which is only locally available to them.

3.1 Guaranteed coordinated

Agents will be guaranteed to act coordinately if they ignore their local observations and use the commonly known prior distribution over states to calculate the optimal joint policy by maximising the expected reward:

\[
V_{\pi} = \sum_S p(S)R(S, \pi(S))
\]
However, this guaranteed coordination comes at the cost of discarding potentially valuable information, thus making a decision which is overall less informed.

3.2 Local

Consider instead calculating a local solution $\pi_i$ to $V_{v_i}(b_i)$, the local expected reward given agent $i$’s belief over the global state:

$$V_{v_i}(b_i) = \sum_S \sum_{B_{-i}} p(B_{-i}|b_i)p(S|B) \sum_{\pi_{-i}} R(S, \pi(B))$$

(3)

where $B = (b_1, \ldots, b_n)$ is the vector comprising local beliefs and $\pi(B) = (\pi_1(b_1), \ldots, \pi_n(b_n))$ is the joint policy vector and we have implicitly assumed that without prior knowledge all other agents’ policies are equally likely. With this the total reward under policy $\pi_i$ as expected by agent $i$ is given by

$$V_{v_i} = \sum_S \sum_{b_i} p(S)p(b_i|S)V_{v_i}(b_i)$$

(4)

Calculating the value function in equation 3 requires marginalising over all possible belief states and policies of other agents and will in general be intractable. However, if it were possible to solve this equation exactly, the resulting local policies should never perform worse than an approach which guarantees overall coordination by discarding local observations. This is because the coordinated policies are a subset of all policies considered here and should emerge as the optimal policies in cases where coordination is absolutely crucial. As a result the overall reward $V_{v_i}$ expected by any agent $i$ will always be greater or equal to the expected reward under a guaranteed coordinated approach as given by equation 2. The degree to which this still holds and hence to which a guaranteed coordinated approach is to be favoured over local decision-making therefore depends on the quality of any approximate solution to equation 3 and the extent to which overall coordination is rewarded.

4. APPROXIMATE DECISION-MAKING

The optimal local one-step policy of an individual agent is simply the best response to the possible local actions the others could be choosing at that point in time. The full marginalisation over others’ local beliefs and possible policies therefore amounts to a marginalisation over all others’ actions. Calculating this requires knowing the probability distribution over the current state and remaining agents’ action choices, given agent $i$’s local belief $b_i$, $p(S, A_{-i}=b_i)$.

Together with equation 3 the value of a local action given the current local belief over global state then becomes

$$V_i(a_i, b_i) = \sum_S \sum_{A_{-i}} p(S, A_{-i}=b_i)R(S, a_i, A_{-i})$$

(5)

This re-formulation in terms of $p(S, A_{-i}=b_i)$ significantly reduces the computational complexity compared to iterating over all local beliefs and policies. However, its exact form will in general not be known without performing the costly iteration over others actions and policies. To solve equation 5 we therefore need to find a suitable approximation to $p(S, A_{-i}=b_i)$.

Agent $i$’s joint belief over the current state and other agents’ choice of actions can be expanded as

$$p(S, A_{-i}=b_i) = p(A_{-i}|S)p(S|b_i) = p(A_{-i}|S)b_i(S)$$

(6)

Finding the local belief state $b_i(S)$ is a matter of straightforward Bayesian inference based on the knowledge of the system’s dynamics. One convenient way of solving this calculation is by casting the scenario as a graphical model and using standard solution algorithms to obtain the marginal distribution $b_i(S)$. For the systems considered in this work, where observations only depend on the current state we can use the sum-product algorithm [17], which makes the calculation of local beliefs particularly easy.

Obtaining an expression for the local belief over all other agents’ actions is less simple: Assuming $p(A_{-i}|S)$ were known agent $i$ could calculate its local expectation of future rewards according to equation 6 and choose the local action which maximises this value. All remaining agents will be executing the same calculation simultaneously. This means that agent $i$’s distribution over the remaining agents’ actions is influenced by the simultaneous decision-making of the other agents, which in turn depends on agent $i$’s action choice. Finding a solution to these interdependent distributions is not straightforward. In particular, an iterative solution based on reasoning over others’ choices will lead to an infinite regress of one agent trying to choose its best local policy based on what it believes another agent’s policy to be even though that action is being decided on at the same time. Below we describe three heuristic approaches by which the belief over others actions could be approximated in a quick, simple way.

4.1 Optimistic approximation

From the point of agent $i$ an optimistic approximation to $p(A_{-i}|S)$ is to assume that all other agents choose the local action given by the joint centralised policy for a global state, that is

$$p(A_{-i}|S) = \begin{cases} 1 & \text{if } A_{-i} = \pi(S)_{-i} \\ 0 & \text{otherwise} \end{cases}$$

(7)

This is similar to the approximation used in [7].

4.2 Uniform approximation

Alternatively, agents could assume no prior knowledge about the actions others might choose at any point in time by putting a uniform distribution over all possible local actions:

$$p(a_j = a_j^k|S) = \frac{1}{|A_j|}$$

(8)

and

$$p(A_{-i}|S) = \prod_{j \neq i} p(a_j^k|S) | a_j^k \in A$$

(9)

4.3 Pessimistic approximation

Finally, a pessimistic agent could assume that the local decision-making will lead to sub-optimal behaviour and that the other agents can be expected to choose the worst possible action in a given state.

$$p(A_{-i}|S) = \begin{cases} 1 & \text{if } A_{-i} = (\arg \min_A V_{\text{centralised}}(S))_{-i} \\ 0 & \text{otherwise} \end{cases}$$

(10)

Each of these approximations can be used to implement local decision-making by calculating the expected value of a local action according to equation 5. Ideally we would like to compare the overall expected reward (see equation 4) under each of the approximate local algorithms and compare
Rewards

<table>
<thead>
<tr>
<th>Actions</th>
<th>Rewards</th>
</tr>
</thead>
<tbody>
<tr>
<td>both choose tiger</td>
<td>-50</td>
</tr>
<tr>
<td>both choose reward</td>
<td>100</td>
</tr>
<tr>
<td>both choose nil</td>
<td>0</td>
</tr>
<tr>
<td>both wait</td>
<td>-2</td>
</tr>
<tr>
<td>one tiger, one nil</td>
<td>-100</td>
</tr>
<tr>
<td>one tiger, one reward</td>
<td>-50</td>
</tr>
<tr>
<td>one tiger, one wait</td>
<td>101</td>
</tr>
<tr>
<td>one nil, one wait</td>
<td>-1</td>
</tr>
<tr>
<td>one nil, one reward</td>
<td>50</td>
</tr>
<tr>
<td>one reward, one wait</td>
<td>49</td>
</tr>
</tbody>
</table>

(a) Reward setting 1: some reward for uncoordinated actions

<table>
<thead>
<tr>
<th>Actions</th>
<th>Rewards</th>
</tr>
</thead>
<tbody>
<tr>
<td>both choose tiger</td>
<td>-20</td>
</tr>
<tr>
<td>both choose reward</td>
<td>100</td>
</tr>
<tr>
<td>both choose nil</td>
<td>0</td>
</tr>
<tr>
<td>both wait</td>
<td>-2</td>
</tr>
<tr>
<td>one tiger, one nil</td>
<td>-100</td>
</tr>
<tr>
<td>one tiger, one reward</td>
<td>-100</td>
</tr>
<tr>
<td>one tiger, one wait</td>
<td>101</td>
</tr>
<tr>
<td>one nil, one wait</td>
<td>-1</td>
</tr>
<tr>
<td>one nil, one reward</td>
<td>20</td>
</tr>
<tr>
<td>one reward, one wait</td>
<td>19</td>
</tr>
</tbody>
</table>

(b) Reward setting 2: small reward for uncoordinated actions

<table>
<thead>
<tr>
<th>Actions</th>
<th>Rewards</th>
</tr>
</thead>
<tbody>
<tr>
<td>both choose tiger</td>
<td>-20</td>
</tr>
<tr>
<td>both choose reward</td>
<td>100</td>
</tr>
<tr>
<td>both choose nil</td>
<td>0</td>
</tr>
<tr>
<td>both wait</td>
<td>-2</td>
</tr>
<tr>
<td>one tiger, one nil</td>
<td>-100</td>
</tr>
<tr>
<td>one tiger, one reward</td>
<td>-100</td>
</tr>
<tr>
<td>one tiger, one wait</td>
<td>101</td>
</tr>
<tr>
<td>one nil, one wait</td>
<td>-1</td>
</tr>
<tr>
<td>one nil, one reward</td>
<td>0</td>
</tr>
<tr>
<td>one reward, one wait</td>
<td>-1</td>
</tr>
</tbody>
</table>

(c) Reward setting 3: no reward for uncoordinated actions

Table 1: Reward matrices for the Tiger Scenario with varying degrees by which uncoordinated actions are rewarded. Joint actions for which the rewards were varied are shaded.

![Graph](image1.png)

Figure 1: Average obtained reward (red diamonds) compared to expected reward (green squares) for different approximate decision-making algorithms. Data points were obtained by averaging over 500 time-steps. The uniform algorithm consistently under-estimates the expected reward, while the pessimistic algorithm both under- and over-estimates, depending on the setting of the reward matrix. The optimistic algorithm tends to over-estimate the reward but has the smallest deviation and in particular approximates it well for the setting which is most favourable to uncoordinated actions.

To adapt the scenario to be an example of a dec-MDP with uniform transition probabilities as discussed above, we have modified this scenario in the following way: Two agents are faced with three doors, behind which sit a tiger, a reward or nothing. At each time step both agents can choose to open one of the doors or to do nothing and wait. These actions are carried out deterministically and after both agents have chosen their actions, an identical reward is received (according to the commonly known reward matrix) and the configuration behind the doors is randomly re-set to a new state. Prior to choosing their actions the agents are both informed about the contents behind one of the doors, but never both about the same door. If agents can exchange their observations prior to making their decisions, the problem becomes fully observable and the optimal choice of action is straightforward. If, on the other hand, they cannot exchange their observations, they will both hold differing, incomplete information about the global state which will lead to differing beliefs over where the tiger and the reward are located.

5. EXAMPLE SIMULATION

We have applied the different decision-making algorithms to a modified version of the Tiger Problem, which was first introduced by Kaelbling et al. [11] in the context of single-agent POMDPs and has since been used in modified forms as a benchmark problem for dec-POMDP solution techniques [2, 12, 13, 16, 19, 20, 22, 25]. For a comprehensive description of the initial multi-agent formulation of the problem see [12].
above for different reward matrices and have compared the performance of the various approximate algorithms to a guaranteed coordinated approach in which agents discard their local observations and use their common joint belief over the global state of the system to determine the best joint action. Each scenario run consisted of 500 time-steps. In all cases the highest reward (lowest penalty) was given to coordinated joint actions. The degree by which agents received partial awards for uncoordinated actions varied for the different settings. For a detailed listing of the reward matrices used see table 1.

Figure 4 shows the expected and average obtained rewards for the different reward settings and approximate algorithms described above. As expected the average reward gained during the simulation differs from the expected reward as predicted by an individual agent. While this difference is quite substantial in some cases, it is consistently smallest for the optimistic algorithm.

Figure 2 shows the performance of the approximate algorithms compared to the performance of the guaranteed coordinated approach. The pessimistic approach consistently performs worse than any of the other algorithms, while the optimistic and the uniform approach achieve similar performance. Interestingly, the difference between the expected and actual rewards under the different approximate algorithms (figure 4) does not provide a clear indicator for the performance of an algorithm. Compared to the guaranteed coordinated algorithm the performance of the optimistic/uniform algorithms depends on the setting of the reward matrix. They clearly outperform it for setting 1, while achieving less average reward for setting 3. In the intermediate region all three algorithms obtain similar rewards. It is important to remember here that even for setting 1 the highest reward is awarded to coordinated actions and that setting 3 is the absolute limit case in which no reward is gained by acting uncoordinatedly. We would argue that the latter is a somewhat artificial scenario and that many interesting applications are likely to have less extreme reward matrices. The results in figure 2 suggest that for such intermediate ranges even a simple approximate algorithm for decision-making based on local inference might outperform an approach which guarantees agent coordination.

6. CONCLUSIONS

We have argued that coordination should not automatically be favoured over making use of local information in multi-agent decision processes with sparse communication and have described three simple approximate approaches that allow local decision-making based on individual beliefs. We have compared the performance of these approximate local algorithms to that of a guaranteed coordinated approach on a modified version of the Tiger Problem. Some of the approximate algorithms showed comparable or better performance than the coordinated algorithm for some settings of the reward matrix. Our results can thus be understood as first evidence that strictly favouring agent coordination over the use of local information can lead to lower collective performance than using an algorithm for seemingly uncoordinated local decision making. More work is needed to fully understand the influence of the reward matrix, system dynamics and belief approximations on the performance of the respective decision-making algorithms. Future work will also include the extension of the treatment to truly sequential decision processes where the transition function is no longer uniform and independent of the actions taken.

7. REFERENCES


