Supplementary Material for NIPS 2011 Paper #1083: "Efficient Offline Communication Policies for Factored Multiagent POMDPs"

João V. Messias
Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal
jmessias@isr.ist.utl.pt

Matthijs T.J. Spaan
Delft University of Technology
Delft ,The Netherlands
m.t.j.spaan@tudelft.nl

Pedro U. Lima
Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal
pal@isr.ist.utl.pt

This document contains additional technical details related to the work described in [1]. For brevity, any constructs which here are left undefined are assumed to have the same meaning as in the main document, and are referred, where necessary, to its appropriate sections.

1 Detecting Conflicting Actions Through Convex Optimization

As described in section 3.2.2 of [1], actions a and a' are said to be *conflicting* in local space if there are two points b, b' such that $M_{\mathcal{L}}^{\mathcal{X}} b = M_{\mathcal{L}}^{\mathcal{X}} b'$, and where $\phi(\arg\max_{\alpha\in\Gamma} \alpha \cdot b)_i \neq \phi(\arg\max_{\alpha\in\Gamma} \alpha \cdot b')_i$ for agent i. This means that b and b' are associated with different individual actions for that agent, but are undistinguishable given only their marginalization $b_{\mathcal{L}}$. We shall consider the action with the highest value bound at $b_{\mathcal{L}}$ (hereafter defined as a) as the expected best action to take. For this action there already is at least one joint belief point for which it is maximal - the point which generated the maximum value bound. The problem is then to find some b where $a' \neq a$ is the maximal action. Given $\mathbf{v} = Ab$ and $\mathbf{v}' = A'b$, the vectors describing all possible values associated with a and a', the problem can then be described as the constrained optimization:

$$\begin{array}{lll} \text{minimize} & \max_i \mathbf{v}_i - \max_j \mathbf{v}'_j & \\ \text{subject to} & \mathbf{v} = Ab & b \succeq \mathbf{0}_n \\ & \mathbf{v}' = A'b & \mathbf{1}_n^T b = 1 \\ & M_{\mathcal{L}}^{\mathcal{X}} b = b_{\mathcal{L}} & \end{array} \tag{1}$$

If the solution to this problem is negative, then we know that a' is maximal at some point b, which means that neither action can be taken whithout further information. Unfortunately, the target function in this optimization is non-convex. Taking the epigraph of the first term of the target function, the problem becomes:

$$\begin{array}{lll} \text{minimize} & s - \max_j \mathbf{v}'_j & \\ \text{subject to} & Ab \preceq \mathbf{1}_k s & b \succeq \mathbf{0}_n \\ & \mathbf{v}' = A'b & \mathbf{1}_n^T b = 1 \\ & M_{\mathcal{L}}^{\mathcal{X}} b = b_{\mathcal{L}} & \end{array} \tag{2}$$

If the vectors in $|\Gamma^{a'}|$ (rows of A') are then taken individually, the problem trivially becomes the LP:

$$\begin{array}{llll} \forall i = 1, \ldots, |\Gamma^{a'}| & \text{maximize} & \Gamma_i^{a'} b - s & \\ & \text{subject to} & Ab \preceq \mathbf{1}_k s & b \succeq \mathbf{0}_n \\ & & M_{\mathcal{L}}^{\mathcal{X}} b = b_{\mathcal{L}} & \mathbf{1}_n^T b = 1 \end{array} \tag{3}$$

An alternative is to introduce the slack variable ξ in the constraints of (2):

$$\begin{array}{ll} Ab \preceq \mathbf{1}_k s & b \succeq \mathbf{0}_n \\ A'b = \mathbf{1}_{k'} s + \xi & \mathbf{1}_n^T b = 1 \\ M_{\mathcal{L}}^{\mathcal{X}} b = b_{\mathcal{L}} & \end{array} \tag{4}$$

If the maximum element of ξ is positive at some b, then we can safely conclude that $\max_i \mathbf{v}_i \leq \max_j \mathbf{v}'_j$ and therefore the actions are undecidable. The problem of maximizing the maximum element of ξ, however, is only solvable by splitting ξ into its positive and negative components, ξ^+ and ξ^-, and requiring that $(\xi^+)^T \cdot \xi^- = 0$. The latter constraint is itself non-convex, and at best it increases the complexity of the optimization procedure beyond that of the exhaustive LP (3). In order to contain this problem as an LP, we must then relax these constraints, and describe the problem as:

$$\begin{array}{lll} \text{maximize} & \mathbf{1}_{k'}^T \xi & \\ \text{subject to} & Ab \preceq \mathbf{1}_k s & b \succeq \mathbf{0}_n \\ & A'b = \mathbf{1}_{k'} s + \xi & \mathbf{1}_n^T b = 1 \\ & M_{\mathcal{L}}^{\mathcal{X}} b = b_{\mathcal{L}} & \end{array} \tag{5}$$

The target function in this optimization is not the same as in the original problem (1), since it instead seeks to find the point b with the highest average difference between the maximum element of $\mathbf{v}$ and the values of A' (highest mean value of ξ). While the optimal solution to this problem is typically achieved at a point where ξ has positive components, this is not necessarily so, and therefore we must consider this as an approximate solution to the original problem. Since the vectors in A and A' are arbitrary as long as the full value function is convex, it is also difficult to establish a bound on the quality of this approximation. In practice, for the examples studied in the results of [1], we found that using (5) instead of (3) does not noticeably affect the quality of the resulting communication map, and allows us to scale better to larger domains.

The extension of (5) to the problem of finding a set of factors $\mathcal{G}$ with no conflicting actions is straightforward: we need only to simultaneously consider two symmetric problems, that of finding a point b where a is maximal, and that of finding b' where a' is maximal, while requiring that these points are undistinguishable when projected to $\mathcal{G}$. The full optimization is then:

$$\begin{array}{llll} \text{maximize} & \mathbf{1}_{k'}^{\mathrm{T}} \xi' + \mathbf{1}_k^{\mathrm{T}} \xi & & \\ \text{subject to} & Ab \preceq \mathbf{1}_k s & A'b = \mathbf{1}_{k'} s + \xi & M_{\mathcal{L}}^{\mathcal{X}} b = b_{\mathcal{L}} \\ & A'b' \preceq \mathbf{1}_{k'} s' & Ab' = \mathbf{1}_k s' + \xi' & M_{\mathcal{L}}^{\mathcal{X}} b' = b_{\mathcal{L}} \\ & b \succeq \mathbf{0}_n & b' \succeq \mathbf{0}_n & M_{\mathcal{G}}^{\mathcal{X}} b = M_{\mathcal{G}}^{\mathcal{X}} b' \end{array} \tag{6}$$

2 Building the Communication Map

Below is a more detailed, pseudo-code description of the algorithm suggested in section 3.2.3 of [1]. The inputs to this algorithm are the set of local factors, $\mathcal{L}$, the set of non-local factors $\mathcal{F}$, the value function V, and the number of desired samples N. The output is a set of pairs $\langle b_{\mathcal{L}}, \mathcal{G} \rangle$ of local belief points and associated communication decisions.

Algorithm 1 CreateCommunicationMap($\mathcal{L}, \mathcal{F}, V, N$)

$\{Single_LP(b_{\mathcal{L}}, a, a')$ refers to (5)$\}$
$\{Full_LP(factors, b_{\mathcal{L}}, a')$ refers to (6)$\}$
$Samples \leftarrow$ sample N reachable local belief points $b_{\mathcal{L}}$;
$bounds \leftarrow$ obtain local value bounds of V; $Map \leftarrow \emptyset$;
for all $b_{\mathcal{L}} \in Samples$ **do**
  $\alpha' \leftarrow \arg\max_{\alpha} \overline{V_{\alpha}}(b_{\mathcal{L}})$;
  if $\underline{V_{\alpha'}}(b_{\mathcal{L}}) \geq \overline{V_{\alpha}}(b_{\mathcal{L}}) \quad \forall \alpha \neq \alpha'$ or $Single_LP(b_{\mathcal{L}}, a, a') > 0 \quad \forall \alpha \neq \alpha'$ **then**
    $Map \leftarrow Map \cup \langle b_{\mathcal{L}}, \emptyset \rangle$;
  else
    $\mathcal{G} \leftarrow \emptyset$; $\mathcal{H} \leftarrow \mathcal{F}/\mathcal{L}$;
    while $\mathcal{H}$ is not empty **do**
      $temp \leftarrow$ remove factor from $\mathcal{H}$; $factors \leftarrow \mathcal{H} \cup \mathcal{G}$;
      if $Full_LP(factors, b_{\mathcal{L}}, a')$ returns both negative solutions **then**
        $\mathcal{G} \leftarrow temp$;
      end if
    end while
    $Map \leftarrow Map \cup \langle b_{\mathcal{L}}, \mathcal{G} \rangle$;
  end if
end for
return Map

3 The OneDoor Scenario

We here provide further description of the OneDoor environment used in the results of [1]. In this problem, originally introduced in [2], two agents operate in a grid-like world, represented in fig. 1, and may each be in one of 7 possible positions. One of the agents is know to be in positions 1, 2 or 3 (with uniform probability) and has the goal of reaching position 5. The other starts in positions 5, 6 or 7 and must reach position 3. Each agent can move in any of the four directions, with an associated probability of ending up in an unintended neighbor state, and can observe positions 2, 4 and 6 with no noise. The remaining positions all produce the same observation, albeit also deterministically. Therefore $|\mathcal{O}_i| = 4$. The robots may share the same position, and they receive a penalty for being both in position 4 at the same time. They receive a positive reward for reaching their goal, and no reward otherwise. The agents are uncoupled expect through the reward function (i.e. we here assume a transition-observation independent version of the problem). Even so, this means that an acceptable policy in this problem must be such that one of the agents waits for the other to clear the "door" in position 4 until it attempts to move there.

If a sufficiently large horizon is considered, as shown in [1], this problem allows for a significant reduction in communication, using our method. This is because a near-optimal joint policy defines wich agent should take priority, and since that agent always moves first, it rarely needs to communicate (only when the other agent has a sufficient probability of moving to position 4 due to the noise in its actions). The other agent, in turn, must communicate until its partner clears the door, and afterwards, its local actions can also be taken independently and so it ceases communication. For horizons smaller than 10, however, the agents may not have enough decisions left to gather any positive reward, and in these cases they both communicate in order to avoid any possible collisions.

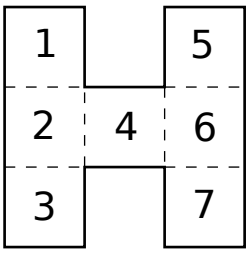

Figure 1: Representation of the OneDoor scenario.

References

[1] J.V. Messias, M.T.J. Spaan, and P. U. Lima. Efficient offline communication policies for factored multiagent POMDPs. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, 2011.

[2] Frans A. Oliehoek, Matthijs T. J. Spaan, and Nikos Vlassis. Dec-POMDPs with delayed communication. In *Multi-agent Sequential Decision Making in Uncertain Domains*, 2007. Workshop at AAMAS07.