

Theorem 4 *The decision problem of determining whether there exist policies, π_Σ and π_A , for a given COM-MTDP with free communication and collective observability, that yield a total reward at least K over some finite horizon T is P-complete.*

Proof: To prove that the problem is P-hard, we reduce the single-agent MDP to a COM-MTDP. In particular, if we are given a POMDP, $\langle S, A, P, R \rangle$, we can construct a COM-MTDP for a single-agent team (i.e., $\alpha = \{1\}$):

$$S' = S$$

$$A'_1 = A$$

$$\Sigma'_1 = \emptyset$$

$$P'(s_i, \langle a_1 \rangle, s_f) = P(s_i, a_1, s_f)$$

$$\Omega'_1 = S$$

$$\mathcal{O}'(s, \langle a_1 \rangle, \langle s \rangle) = 1$$

$$B'_1 = S$$

$$R'_A(s, \langle a_1 \rangle) = R(s, a_1)$$

$$R'_\Sigma(s, \sigma) = 0$$

This COM-MTDP meets our assumptions of free communication and collective observability (in fact, it is individually observable). Just as in the proof of Theorem 1, we can show that there exists a policy with expected utility greater than K for this COM-MTDP if and only if there exists one for the MDP. The decision problem for the MDP is known to be P-hard, so the COM-MTDP problem under free communication and collective observability must be P-hard.

To show that the problem is in P, we take a COM-MTDP under free communication and collective observability and reduce it to a single-agent MDP. In particular, if we are given a COM-MTDP, $\langle S, \mathbf{A}, \Sigma, P, \Omega, \mathcal{O}, \mathbf{B}, R \rangle$, we can construct a single-agent MDP as follows:

$$S' = S$$

$$A' = \mathbf{A}$$

$$P'(s_i, \mathbf{a}, s_f) = P(s_i, \mathbf{a}, s_f)$$

$$R'(s, \mathbf{a}) = R_A(s, \mathbf{a})$$

From Theorem 2, we need to consider only the complete-communication policy for the COM-MTDP and this policy has a zero reward. Therefore, the decision problem for the COM-MTDP is simply to find a domain-level policy that produces an expected reward exceeding K . Given full communication and collective observability, the state estimator functions for the COM-MTDP (using the result in the proof of Theorem 2) reduce to:

$$SE_{i\Sigma^\bullet}(\langle \Omega^0, \dots, \Omega^{t-1}, \Omega_i^t \rangle, \Sigma^t) = \langle \Omega^0, \dots, \Omega^{t-1}, \Omega^t \rangle = S^t \quad (1)$$

A policy for our MDP specifies an action for each and every state of the world: $\pi' : S' \rightarrow A'$. The state of the world is exactly the belief state of our COM-MTDP under full communication. Therefore, we can translate a MDP-policy, π' , into an equivalent domain-level policy for the COM-MTDP:

$$\pi_A(s) \equiv \pi'(s) \quad (2)$$

A team following π_A will perform the exact same domain-level actions as a single agent following π' . Thus, there exists a policy with expected utility greater than K for the COM-MTDP if and only if there exists one for the MDP. The decision problem for a MDP is known to be in P, so the COM-MTDP problem (under free communication and collective observability) must be in P as well. \square