

Theorem 3 *The decision problem of determining whether there exist policies, π_Σ and π_A , for a given COM-MTDP with free communication, that yield a total reward at least K over some finite horizon T is PSPACE-complete.*

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

$$S' = S$$

$$A'_1 = A$$

$$\Sigma'_1 = \emptyset$$

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

$$\Omega'_1 = \Omega$$

$$O'(s, \langle a_1 \rangle, \langle \omega_1 \rangle) = O(s, a_1, \omega_1)$$

$$B'_1 = (\Omega)^*$$

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

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

This COM-MTDP satisfies our assumption of free communication. The POMDP assumes perfect recall, so we use the following state estimator functions:

$$SE_1^0() = \langle \rangle \tag{1}$$

$$SE_{1 \bullet \Sigma}(\langle \Omega_1^0, \dots, \Omega_1^{t-1} \rangle, \Omega_1^t) = \langle \Omega_1^0, \dots, \Omega_1^{t-1}, \Omega_1^t \rangle \tag{2}$$

$$SE_{1 \Sigma \bullet}(\beta, \text{null}) = \beta \tag{3}$$

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 POMDP. The decision problem for the POMDP is known to be PSPACE-hard, so the COM-MTDP problem under free communication must be PSPACE-hard.

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

$$S' = S$$

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

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

$$\Omega' = \Omega$$

$$O'(s, \mathbf{a}, \omega) = \mathbf{O}(s, \mathbf{a}, \omega)$$

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

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, the state estimator functions for the COM-MTDP (as shown in the proof of Theorem 2) reduce to:

$$SE_{i \Sigma \bullet}(\langle \Omega^0, \dots, \Omega^{t-1}, \Omega^t \rangle, \Sigma^t) = \langle \Omega^0, \dots, \Omega^{t-1}, \Omega^t \rangle \tag{4}$$

A policy for our POMDP specifies an action for each and every history of observations: $\pi' : (\Omega')^+ \rightarrow A'$. The history of observations for the single-agent POMDP corresponds to the belief states of our COM-MTDP under full communication. Therefore, we can translate a POMDP-policy, π' , into an equivalent domain-level policy for the COM-MTDP:

$$\pi_A(\langle \omega_0, \omega_1, \dots, \omega_t \rangle) \equiv \pi'(\langle \omega_0, \omega_1, \dots, \omega_t \rangle) \quad (5)$$

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 POMDP. The decision problem for a POMDP is known to be in PSPACE, so the COM-MTDP problem (under free communication) must be in PSPACE as well. \square