

Theorem 1 *The decision problem of whether there exist policies, π_Σ and π_A , for a given COM-MTDP, under general communication, that yield a total reward at least K over some finite horizon T is NEXP-complete if $|\alpha| \geq 2$ (i.e., more than one agent).*

Proof: To prove that the COM-MTDP problem is NEXP-hard, we reduce a DEC-POMDP [1] to a COM-MTDP with no communication by copying all of the other model features from the given DEC-POMDP. In particular, if we are given a DEC-POMDP, $\langle S, \{A^i\}_{i=1}^m, P, \{\Omega^i\}_{i=1}^m, O, R \rangle$, we can construct a COM-MTDP as follows:

$$S' = S$$

$$A'_i = A^i$$

$$\Sigma' = \emptyset$$

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

$$\Omega'_i = \Omega^i$$

$$O'(s, \langle a_1, \dots, a_m \rangle, \langle \omega_1, \dots, \omega_m \rangle) = O(\omega_1, \dots, \omega_m | a_1, \dots, a_m, s)$$

$$B'_i = (\Omega^i)^*$$

$$R'(s, \sigma, \langle a_1, \dots, a_m \rangle) = R(s, a_1, \dots, a_m)$$

The DEC-POMDP assumes perfect recall, so we use the following state estimator functions:

$$SE_i^0() = \langle \rangle \quad (1)$$

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

$$SE_{i \Sigma \bullet}(\beta, \text{null}) = \beta \quad (3)$$

Since there is no communication for this COM-MTDP, we have a fixed silent policy, π_Σ . We can translate any domain-level policy, π_A , into a DEC-POMDP joint policy, δ , as follows:

$$\delta^i(o_1^i, \dots, o_i^i) \equiv \pi_{iA}(\langle o_1^i, \dots, o_i^i \rangle) \quad (4)$$

The expected utility of following this joint policy, δ , within the DEC-POMDP is identical to that of following π_Σ and π_A within the constructed COM-MTDP. Thus, there exists a policy with expected utility greater than K for the COM-MTDP if and only if there exists one for the DEC-POMDP. The decision problem for a DEC-POMDP is known to be NEXP-complete, so the COM-MTDP problem must be NEXP-hard.

To show that the COM-MTDP is in NEXP, our proof proceeds similarly to that of the DEC-POMDP. In other words, we guess the joint policy, π , and write it down in exponential time (we assume that $T \leq |S|$). We can take the COM-MTDP plus the policy and generate (in exponential time) a corresponding MDP where the state space is the space of all possible combined belief states of the agents. We can then use dynamic programming to determine (in exponential time) whether π generates an expected reward of at least K . \square

References

- [1] Daniel S. Bernstein, Shlomo Zilberstein, and Neil Immerman. The complexity of decentralized control of Markov decision processes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 32–37, 2000.