tempted to scale to larger problems and horizons by not generating the full set of policies that may be optimal. These approaches, known as memory-bounded algorithms, were introduced by Seuken and Zilberstein [24] and then successively refined [10, 15]. Memory-bounded algorithms sample forward a bounded number of belief states, and back up (i.e., generate next step policies for) one decentralized history-dependent policy for each belief state. To avoid the explicit enumeration of all possible policies, Kumar and Zilberstein [15] perform the backup by solving a corresponding constraint optimization problem (COP) [9], that represents the decentralized backup. Although, memory-bounded techniques are suboptimal, the decentralized backup can be applied in exact settings as we demonstrate in our algorithm.

More specifically, the decentralized backup can build a horizon-τ decentralized policy that is maximal with respect to a belief state and horizon-(τ + 1) policies available for each agent. The associated COP is given by: a set of variables, one for each local observation of each agent; a set of domains, where the domain for the variables corresponding to an agent is the set of horizon-(τ + 1) policies available for that agent; a set of soft constraints, one for each joint observation. The soft constraint maps assignments to real values. Intuitively, these values represent the expected reward accrued when agents together perceive a given joint observation and follow a given horizon-(τ + 1) decentralized policy. Since horizon-τ decentralized policies consist of horizon-(τ + 1) policies, it is easy to see that maximizing the sum of the soft constraints yields a maximal horizon-τ decentralized policy.

Closer to our model is the ND-POMDP framework [19]. It aims at modeling multiagent teamwork where agents have strong locality of interaction, often through binary interactions. That is, the reward model in such domains is decomposed among sets of agents. There has been a substantial body of work that extend general Dec-POMDP techniques (discussed above) to exploit the locality of interaction [19, 14, 16]. Nair et al. [19] introduced the only optimal algorithm for this model, namely the General Optimal Algorithm (GOA). When the domain does not contain binary interactions, there is no reason to expect GOA to outperform general Dec-POMDP algorithms, as all methods use similar strategies in selecting policy candidates. However, when the domain contains primarily binary interactions (or more generally when each agent’s rewards are not dependent on many other agents), GOA is likely to outperform general Dec-POMDP algorithms.

It is worth noting that ND-POMDPs and transition and observation independent Dec-MDPs make the same assumptions about transition and observation independence, but make different assumptions about the reward model and partial observability. More specifically, ND-POMDPs assume the reward can be decomposed into the sum of local reward models for sets of agents, while the reward model for transition and observation independent Dec-MDPs is more general, allowing global rewards for all agents (i.e., considering all agents to be in one set). Dec-MDPs assume that the state is jointly fully observable (i.e., that the state is fully determined by the combination of local observations of all agents), while ND-POMDPs do not make this limiting assumption. Both models therefore make different assumptions to address complexity and the choice of model depends on which assumptions best match the domain being solved.

4 Theoretical Properties

In this section, we demonstrate the main theoretical results of this paper.

4.1 Optimal Policies

A decentralized MDP solver aims to calculate an optimal decentralized policy π∗ that maximizes the expected cumulative reward:

\[ \pi^* = \arg \max_{\pi} E[\sum_{t=0}^{T-1} r(s_t, a_t) | \pi, \eta_0]. \]  

The following theorem proves that decentralized Markov policies yield the optimal performance in decentralized MDPs with independent transitions and observations. Goldman et al. [12] established the optimality of Markov policy for an agent under the assumption that the other agents choose Markov policies. Here, we state the optimality of Markov policies for an agent no matter what its teammates’ policies are. We also construct the proof in a manner that more directly relates policies to values (rather than information sets). This may be more clear to some readers.

**Theorem 1 (Optimality of decentralized Markov policies)**

**In Dec-MDPs with independent transitions and observations, optimal policies for each agent depend only on the local state and not on agent histories.**

**Proof** Without loss of generality, we construct a proof by induction for two agents, 1 and 2, from agent 1’s perspective. We first show that in the last step of the problem, agent 1’s policy does not depend on its local history.

Agent 1’s local policy on the last step is:

\[ \sigma_{T-1}^1(h_{T-1}^1) = \arg \max_{\pi_1} \sum_{h_{T-1}^1} P(h_{T-1}^1|h_{T-1}^1) \cdot R(s, a^1, \sigma_{T-1}^2(h_{T-1}^2)), \]

which chooses a local action to maximize value based on the possible local histories of agent 2 and resulting states of the system \( s = \langle z_{T-1}^1, z_{T-1}^2 \rangle \).

Based on transition and observation independence and the use of decentralized policies, it can be shown that

\[ P(h_{T-1}^1|h_{T-1}^1) = P(h_{T-1}^2). \]

Due to space limitations, we