solve problems that are multiple orders of magnitude larger and those that include up to 10 agents.

2 Motivating examples

To illustrate the characteristics of decentralized partially observable Markov decision processes (Dec-POMDPs) that we are interested in, consider a simple two-agent “meeting-in-a-grid under uncertainty” domain in Figure 1. In this scenario, two agents want to meet as soon as possible on a two-dimensional grid. In this world, each agent’s possible actions include moving north, south, west, east and staying in the same place. The actions of a given agent do not affect the other agents. After taking an action, each agent can sense some information, which in this case corresponds to its own location. Here, each agent’s own partial information is insufficient to determine the global state of the world. This is mainly because agents are not permitted to explicitly communicate their local locations with each other. However, if this (instantaneous and noise-free) communication were allowed the agents’ partial information together would reveal the true state of the world, (i.e., the agents’ joint location). It is the presence of this joint full observability property that differentiates Dec-MDPs from Dec-POMDPs.

More generally, in partially observable models including Dec-POMDPs, the agents’ partial information together can map to multiple different states of the world. As a consequence, decisions in such models depend on the entire past histories of actions and observations that the agents ever experienced. In the meeting-in-a-grid under uncertainty problem, since both transitions and observations are not affected by the other agents, each agent’s decision depends only on its last piece of partial information, (i.e., the agent’s own location) [5]. These characteristics appear in many real-world applications including:

Mars exploration rovers. The meeting-in-a-grid domain was motivated by a real problem of controlling the operation of multiple space exploration rovers, such as the ones used by NASA to explore the surface of Mars [27].

Distributed sensor net surveillance. The sensor net domain [17], where a team of stationary or moveable UAVs, satellites, or other sensors must coordinate to track targets while sensors have independent transitions and observations, is particularly suited to our model.

Distributed smart-grid domains. This application aims at finding the optimal schedules and amounts of generated power for a collection of generating units, given demands, and operational constraints over a time horizon.

3 Background and Related Work

In this section, we review the decentralized MDP model, the assumptions of transition and observation independence, the associated notation, and related work.

Definition 1 (The decentralized MDP) A n-agent decentralized MDP $(S,A,p,r)$ consists of:

- A finite set $S = Z^1 \times Z^2 \times \cdots \times Z^n$ of states $s = (z^1, z^2, \cdots, z^n)$, where $Z^i$ denotes the set of local observations $z^i$ of agent $i = 1, 2, \ldots, n$.

- A finite set $A = A^1 \times A^2 \times \cdots A^n$ of joint actions $a = (a^1, a^2, \cdots, a^n)$, where $A^i$ is the set of local actions $a^i$ of agent $i = 1, 2, \ldots, n$.

- A transition function $p(s,a,s')$, which denotes the probability of transitioning from state $s = (z^1, z^2, \cdots, z^n)$ to state $s' = (z'^1, z'^2, \cdots, z'^n)$ when taking joint action $a = (a^1, a^2, \cdots, a^n)$

- A reward function $r: S \times A \rightarrow \mathbb{R}$, where $r(s,a)$ denotes the reward received when executing joint action $a$ in state $s$

As noted above, decentralized MDPs are distinguished by the state being jointly fully observable. This property ensures that the global state would be known if all agents shared their observations at a given step (i.e., there is no external uncertainty in the problem) and follows trivially from the definition of states as observations for each agent. The Dec-MDP is parameterized by the initial state distribution $\eta_0$. When the agents operate over a bounded number of steps (typically referred to as the problem horizon $T$), the model is referred to as a finite-horizon decentralized MDP. Solving a decentralized MDP for a given planning horizon $T$ and start state distribution $\eta_0$ can be seen as finding $n$ individual policies that maximize the expected cumulative reward over the steps of the problem.