Planning and Learning For Decentralized MDPs With Event Driven Rewards

Tarun Gupta¹, Akshat Kumar²†, Praveen Paruchuri¹†
¹Machine Learning Lab, Kohli Center on Intelligent Systems, IIIT Hyderabad
²School of Information Systems, Singapore Management University
tarun.gupta@research.iiit.ac.in, akshatkumar@smu.edu.sg, praveen.p@iiit.ac.in

Abstract

Decentralized (PO)MDPs provide a rigorous framework for sequential multiagent decision making under uncertainty. However, their high computational complexity limits the practical impact. To address scalability and real-world impact, we focus on settings where a large number of agents primarily interact through complex joint-rewards that depend on their entire histories of states and actions. Such history-based rewards encapsulate the notion of events or tasks such that the team reward is given only when the joint-task is completed. Algorithmically, we contribute — 1) A nonlinear programming (NLP) formulation for such event-based planning model; 2) A probabilistic inference based approach that scales much better than NLP solvers for a large number of agents; 3) A policy gradient based multiagent reinforcement learning approach that scales well even for exponential state-spaces. Our inference and RL-based advances enable us to solve a large real-world multiagent coverage problem modeling schedule coordination of agents in a real urban subway network where other approaches fail to scale.

1 Introduction

Decentralized MDPs and POMDPs (Dec-(PO)MDPs) provide a rigorous framework for collaborative multiagent sequential decision making under uncertainty and partial observability (Bernstein et al. 2002). They model settings where agents act based on different partial observations about the environment and each other to maximize a global objective. Applications of Dec-POMDPs include coordinating planetary rovers (Becker et al. 2004), multi-robot coordination (Amato et al. 2015) and throughput optimization in a wireless network (Pajarinen, Hottinen, and Peltonen 2014). Solving Dec-POMDPs is computationally challenging with NEXP-Hard complexity (Bernstein et al. 2002).

For scalability and practical application, several variants of Dec-POMDPs have been developed including state transition and observation independence (Becker et al. 2004; Nair et al. 2005; Kumar, Zilberstein, and Toussaint 2011; Dibangoye et al. 2013), weak coupling among agents (Witwicki and Durfee 2010) and collective interactions (Sonu, Chen, and Doshi 2015; Robbel, Oliehoek, and Kochenderfer 2016; Nguyen, Kumar, and Lau 2017a; 2017b). Our focus is on agent interactions with complex event-based rewards which depend on entire state-action histories of multiple agents. We use the transition independent Dec-MDP model (TIDec-MDP) (Becker et al. 2004) where agents have their independent local MDPs. Agents are coupled through event-based global rewards dependent on their execution histories. Intuitively, events capture if agents accomplished some high-level task. Becker et al. show such event-driven rewards are highly expressive being able to model temporal relations among agents (e.g., agent A’s activity facilitates or hinders agent B’s activity rewards). Joint-rewards are given when at least \( x \) or at most \( x \) events occur. Such expressiveness is useful to model several practical problems such as multiagent coverage (Yehoshua and Agmon 2016). Several approaches have been developed to solve TIDec-MDPs and related models. (Dibangoye et al. 2013) use occupancy measures over the joint state-space of agents to compute agent policies. Their focus is on standard reward setting, and do not address event-based rewards. Bilinear programming based approach of (Petrik and Zilberstein 2011) is limited to two agents. (Scharpff et al. 2016) solve transition independent multiagent MDPs where an agent’s policy can depend on the global state while the global state is unobservable in TIDec-MDPs. Thus, our work addresses this gap by developing scalable approaches for large multiagent settings (>2 agents) and event-based rewards.

Our contributions are: 1) we present a nonlinear programming (NLP) formulation for TIDec-MDPs with event-based rewards. The NLP formulation is more scalable than the approach for 2-agent TIDec-MDPs in (Becker et al. 2004). Becker et al.’s approach models some types of joint-rewards, e.g., at least \( x \) or at most \( x \) events, by adding additional bits to states leading to exponential state-space increase. Our NLP formulation does not require any modifications to the state-space and is highly compact; 2) we present an inference-based approach that translates the NLP to that of inference in a graphical model by extending the planning-as-inference strategy (Toussaint, Harmeling, and Storkey 2006; Kumar, Zilberstein, and Toussaint 2015). A key difference is that we directly solve the underlying nonconvex, nonlinear program without requiring expensive forward-backward message-passing. Unlike the NLP, our inference approach always involves solving one small convex program per agent

¹Equal advising.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Factored action space

Factored state space defined as

Given the joint state

We define an

real-world problems.

Coverage problems have received a lot of attention in multiagent and robotics literature with several applications such as vacuum cleaning robots, search-and-rescue, intrusion detection, mine clearing among others (Yehoshua and Agmon 2016; Galceran and Carreras 2013). We focus on multiagent coverage by patrol units across a mass rapid transit (MRT) system. In contrast to previous multiagent coverage settings (Yehoshua and Agmon 2016) which require full communication during execution time, our approach assumes partial observability where agents observe only their local state and cannot observe the status of other agents. This is also closer to reality in underground MRT systems where full communication is impossible. Each agent is responsible for some private locations. Inspecting a private location at least once every k time steps gives a local reward to the agent. Thus, the agent is incentivized to spread its inspections to different locations across time. There are shared locations also which denote interchange stations where multiple agents can inspect. Thus, agents should coordinate their inspections to shared locations given that any one agent-inspection is enough to claim the reward. We test our multiagent RL approach on the real MRT map of Singapore. The state-space is exponential for this setting — given n locations, state space is \(O(2^n)\) per agent. We show that our multiagent RL scales well for this problem whereas EM and NLP fails, and also provides much better solution quality than independently optimizing agent policies, confirming the effectiveness of incorporating joint-events for computing gradients. Thus, our work significantly advances the scalability of multiagent planning for real-world problems.

2 Model Definition

We define an n-agent transition independent Dec-MDP using the tuple \((S, A, P, R)\) (Becker et al. 2004):

- Factored state space defined as \(S = \times_{i=1}^n S^i\), where \(S^i\) is the state space for agent \(i\).
- Factored action space \(A = \times_{i=1}^n A^i\), where \(A^i\) is the action space for agent \(i\).
- Given the joint state \(s = (s^i)_{i=1}^n\) and joint-action \(a = (a^i)_{i=1}^n\), the transition to next state \(\bar{s}\) has probability \(P(\bar{s}|s, a) = \times_{i=1}^n P^i(s^i|s^i, a^i)\), where \(P^i\) is agent \(i\)'s local state transition function. This factorization of the transition function results in the transition independence property of the model.

- Local observability: Each agent fully observes its own local state \(s^i\) at each time step \(t\). Agent \(i\) does not observe the local state of any other agent during execution time.
- Local rewards: Each agent has its own local reward function \(r(s^i, a^i)\), and the global reward is additively defined as \(r(s, a) = \sum_{i=1}^n r_i(s^i, a^i)\).

The above model defines a set of \(n\)-independent MDPs as agents’ transition, observation, and the reward functions are all independent. We next describe how joint-rewards are defined that depend on actions of multiple agents. Such event-based joint-rewards are the key to defining a rich class of non-linear interaction among agents that can model several practical scenarios. Solving Dec-MDPs with such rich joint-reward structure is the main focus of our work.

Events: Joint-rewards are defined based on the high-level notion of events. We first review their definitions with detailed treatment in (Becker et al. 2004).

Definition 1. A primitive event for an agent \(i\), \(e = (\hat{s}^i, a^i, \hat{s}'^i)\), is a tuple including agent’s local state, an action, and an outcome state. An event \(E = \{e_1, \ldots, e_h\}\) is a set of primitive events.

Let \(\Phi^i\) denote a valid local state-action execution sequence \([s^i_1, a^i_1, s^i_2, a^i_2, \ldots]\) for an agent \(i\) (subscripts denote time).

Definition 2. A primitive event \(e = (\hat{s}^i, a^i, \hat{s}'^i)\) occurs in history \(\Phi^i\), denoted as \(\Phi^i \models e\), iff the tuple \((\hat{s}^i, a^i, \hat{s}'^i)\) appears as a sub-sequence of \(\Phi^i\). An event \(E\) occurs in the history \(\Phi^i\) (or \(\Phi^i \models E\)) iff \(\exists e \in E : \Phi^i \models e\).

Intuitively, events model the accomplishment of some high-level task by the agent. An event \(E\) is composed of multiple primitive events to account for the uncertainty in the domain and multiple ways of accomplishing the same task. E.g., consider the multiagent MRT patrolling domain. We define the high-level task for the agent is to visit a particular station once between time steps \(t\) and \(t + k\). Primitive events would correspond to visiting the station at time \(t\), \((t + 1)\) through \((t + k)\). As long as any one of these primitive events occurs in a history, it implies the accomplishment of the task.

Definition 3. A primitive event \(e\) is proper if it can occur at most once in each possible history of a given MDP:

\[\forall \Phi^i = \Phi^i_1 \land \Phi^i_2 : (\neg(\Phi^i_1 \models e) \land \neg(\Phi^i_2 \models e))\]

Definition 4. An event \(E\) is proper if it consists of mutually exclusive proper primitive events w.r.t. a given MDP.

Mutually exclusive condition implies that any two primitive events \(e, e' \in E\) cannot occur together within any possible history \(\Phi^i\) or we shall never observe \(\Phi^i \models e \land \Phi^i \models e'\) for any \(\Phi^i\). Becker et al. show how non-proper events can be cast as proper events using techniques such as making time as part of the state or including additional bits in the state to memorize the occurrence of some primitive events. For algorithmic development, we focus on proper events.

Joint-Reward: A joint-reward is described as a constraint among agents that specifies how the interaction among
agents affects the global value of the system. A constraint \( k \) exists among a subset of agents \( G_k \) (\(|G_k| \geq 2\)). It is defined as a tuple \( \langle E^j_k, \forall j \in G_k \rangle, c_k \). Semantically, the constraint \( k \) specifies that if each involved agent in \( G_k \) satisfies its part of the constraint, then the global reward \( c_k \) is given. Formally, let \( \Phi^j \) through \( \Phi^n \) denote histories for all the agents. Constraint \( \langle \langle E^j_k, \forall j \in G_k \rangle, c_k \rangle \) specifies that the reward \( c_k \) is added to the global value iff \( \Phi^j \models E^j_k \) for each agent \( j \in G_k \). Let \( \rho \) be the set of all constraints; same logic is followed for each constraint \( k \in \rho \).

**Expressiveness:** As per the above constraint semantics, global reward \( c_k \) is given if all events in a constraint occur. Several other types of constraint structures are possible. E.g., the global reward can be given at the occurrence of at most \( x \) events, exactly \( x \) events or at least \( x \) events. Becker et al. show how such variants can be reformulated using all event occurring semantics. However, in their approach, translating different constraint types to the standard all event syntax requires adding additional bits to states, which increases the state-space exponentially. In contrast, our work addresses different constraint types directly without the need to expand the state-space. Thus, our modeling and algorithmic techniques are significantly more effective and scalable.

**Policy and joint-value function:** For TIDec-MDPs, the optimal local policy depends on agent's local observed state (Goldman and Zilberstein 2004). We represent agent \( i \)'s stochastic policy as mapping from local state to a distribution over actions or \( \pi^i(a^i|s^i) \). We have fixed-horizon histories, say \( H \). Given local policies \( \pi^j \), the probability \( P(e; \pi^j) \) of a proper, primitive event \( e = (s^i, a^i, s^i') \) occurring during any execution of \( \pi^i \) is:

\[
P(e; \pi^i) = \sum_{t=1}^{H} P(s^t = s^i; \pi^i) \pi^i(a^i|s^i) P^j(s'^t|s^t, a^i)
\]

As all primitive events in a proper event \( E \) are mutually exclusive, we have \( P(E; \pi^i) = \sum_{e \in E} P(e; \pi^i) \). Given the starting state \( s^1 \) for an agent \( i \), \( \rho \) as the set of constraints, the global value function is defined as:

\[
GV(s_1; \pi) = \sum_{i=1}^{n} V^i(s^i_1; \pi^i) + JV(\rho; \pi)
\]

\[
JV(\rho; \pi) = \sum_{k \in \rho} c_k \prod_{j \in G_k} P(E^j_k; \pi^j)
\]

where \( V^i \) is the value function of agent \( i \)'s local MDP. Our goal is to compute the joint-policy \( \pi \) that optimizes the global value function (7).

### 3 Optimization Based Formulation

An optimal algorithm is presented in (Becker et al. 2004) to solve TIDec-MDPs with event-based rewards. Given the NP-Hardness of the problem, the optimal approach is not scalable to a large number of several agents. We therefore first present a nonlinear math programming (NLP) formulation for TIDec-MDPs. The NLP formulation has been successfully used for approximately solving POMDPs (Amato, Bernstein, and Zilberstein 2007b; Charlin, Poupart, and Shioda 2007) and Dec-POMDPs (Amato, Bernstein, and Zilberstein 2007a; 2010) using off-the-shelf solvers such as SNOPT (Gill, Murray, and Saunders 2005).

Table 1 shows the NLP formulation for TIDec-MDPs. This program is similar to the standard MDP dual LP formulation using occupancy variables for state-action pairs (Puterman 1994). Variables \( x^j_k(s, a) \) denote the probability of being in state \( s^j \) and taking action \( a^j \) at time \( t \) for agent \( i \). The first part of the objective computes the total local reward of all agents, and the second summation computes the total joint-value based on (8). Constraint (3) is the standard flow conservation constraint for each state of each agent; (4) connects the initial belief \( b^j_1 \) at time step \( 1 \) with \( x \); variables (5) computes the probability of events using (6). The stochastic policy \( \pi(a|s) \) can be computed as \( x(s, a) / \sum x(s, a) \). The multilinear term modeling event-based rewards in the objective make this program nonlinear and nonconvex.

We can modify the objective (2) to handle other global reward types such as at least one event. Consider the setting \( \langle \langle E^j_k, \forall j \in G_k \rangle, c_k \rangle \) where the reward \( c_k \) is given if at least one event out of \( |G_k| \) happens. The probability of at least one event happening is one minus the probability that none of the events happen. It is modeled as:

\[
\sum_{k \in \rho} c_k \left( 1 - \prod_{j \in G_k} (1 - x(E^j_k)) \right)
\]

We can similarly handle other joint-reward types such as exactly \( x \) events or at most \( x \) events by deriving the contribution of such joint-rewards to the objective analogous to (9). In addition, we can also have multiple types of joint-rewards for our model by simply adding their contribution to the objective (2) of the math program. Notice that our approach of directly modifying nonlinear objective terms does not add additional bits to states, thus do not result in exponential state-space increase as in (Becker et al. 2004).

Program in Table 1 can in principle be solved using standard NLP solvers. However, we observed empirically that

<table>
<thead>
<tr>
<th>Table 1: Nonlinear program (NLP) for event-based TIDec-MDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables: ( x^j_k(s, a) ) ( \forall t, \forall s^j, a^j \forall i );</td>
</tr>
<tr>
<td>Variables: ( x(E^j_k) ) ( \forall j \in G_k, \forall k \in \rho ) ( (1) );</td>
</tr>
<tr>
<td>[ \max_\rho \sum_{i=1}^{n} \sum_{t=1}^{H} \sum_{s^j, a^j} x^j_k(s, a) r^j(s, a) + \sum_{k \in \rho} c_k \prod_{j \in G_k} x(E^j_k) ] ( (2) );</td>
</tr>
<tr>
<td>[ x^j_k(s, a) - \sum_{s'^j, a'^j} x^j_k(s', a') P^j(s'^j</td>
</tr>
<tr>
<td>[ \sum_{s^j, a^j} x^j_k(s, a) = b^j_1(s^j) ; \forall s^j, \forall i ] ( (4) );</td>
</tr>
<tr>
<td>[ x(E^j_k) = \sum_{s^j, a^j} x^j_k(s, a) ] ( \langle \langle E^j_k, \forall j \in G_k \rangle, c_k \rangle ) ( (5) ).</td>
</tr>
</tbody>
</table>
with the increasing number of agents, NLP solvers were unable to scale, slow and often ran out of memory. To address this, we next present a probabilistic inference based formulation of the program in Table 1 that results in a much more scalable approach for large multiagent systems.

4 Inference for TIDec-MDPs

Planning-as-inference is a recent paradigm that translates the planning problem to that of probabilistic inference. Policy optimization in MDPs and POMDPs can be recast to that of maximum likelihood estimation (MLE) problem in a graphical model (Toussaint, Harmeling, and Storkey 2006), and have also been applied to multiagent planning (Pajariinen and Peltonen 2011; Wu, Zilberstein, and Jennings 2013; Kumar, Zilberstein, and Toussaint 2015). Recently, the problem of marginal MAP, which calculates the mode of the marginal posterior distribution of a subset of variables with the remaining variables marginalized is shown to be equivalent to decision making further establishing close connections between planning and inference (Liu and IJssel 2012; 2013). Expectation-maximization (EM) (Dempster, Laird, and Rubin 1977) is a commonly used technique to solve such planning-as-inference formulations.

We also develop here an inference based approach to solve TI-Dec-MDPs. We present a graphical model such that MLE in this model is equivalent to solving the program in Table 1. We use the EM algorithm for MLE in this graphical model. Our work differs from previous applications of EM to planning — 1) Previous works model the sequential aspect of decision making by using dynamic Bayes nets (DBNs) of varying length for reformulation as an inference problem. E.g., Figure 1(a) shows a T-length DBN for an MDP. Notice that we enforce constraints (3),(4) over parameters SAT and r as shown in Figure 1(b). The variable r is a binary random variable. The variable SAT has domain \{⟨s, a⟩|∀s ∈ S, ∀a ∈ A, ∀t = 1: H\}.

We set the parameters of this BN as follows:

\[
P(\hat{r} = 1|s, a, t) = \hat{r}_{sa} = \frac{r(s, a) - r_{\min}}{r_{\max} - r_{\min}} \quad (10)
\]

\[
P(s, a, t) = x_t(s, a)/H \quad (11)
\]

Intuitively, the conditional probabilities \(P(\hat{r} = 1|\cdot)\) model the scaled reward of the original MDP, and the probabilities \(P(s, a, t)\) model the occupancy measures or \(x_t\) variables of the dual LP. The following result shows the equivalence of MLE and the dual LP for MDP.

**Theorem 1.** Let the parameters of the BN be set as per (10),(11). Maximizing the likelihood \(P(\hat{r} = 1; x)\) of observing \(\hat{r} = 1\) in the BN subject to constraints (3),(4) over parameters \(x\) optimally solves the dual LP for the MDP.

**Proof.** The dual LP objective is given as \(\text{dualObj} = \sum_{s,a,t} x_t(s, a)r(s, a)\). We have the probability in the BN of Figure 1(b) as:

\[
P(\hat{r} = 1; x) = \sum_{s,a,t} P(s, a, t)P(\hat{r} = 1|s, a, t)
\]

\[
= \frac{1}{H} \sum_{s,a,t} x_t(s, a)\hat{r}_{sa} = \frac{1}{H} \sum_{s,a,t} x_t(s, a)r(s, a) - \frac{r_{\min}}{r_{\max} - r_{\min}}
\]

The last expression implies \(P(\hat{r} = 1; x) \propto \text{dualObj}\). Thus, maximizing likelihood would also solve the dual LP for MDP. Notice that we enforce constraints (3),(4) over parameters \(x\) while maximizing the likelihood making sure that our solution is always feasible.

Theorem 1 provides the standard setting for the EM algorithm application. The observed data is \(\hat{r} = 1\), variable SAT is hidden, and parameters to optimize are \(x\). EM is an iterative approach that starts with random initial parameters \(x\) and maximizes the following expected log-likelihood to obtain a better estimate \(x^*\).

\[
Q(x; x^*) \propto \sum_{s,a,t} P(\hat{r} = 1, s, a, t; x) \log P(\hat{r} = 1, s, a, t; x^*)
\]

\[
\propto \sum_{s,a,t} \hat{r}_{sa} x_t(s, a) \log x_t^*(s, a)
\]
where we have omitted terms independent of \( \mathbf{x}^* \). The convex optimization problem EM solves in each iteration is:

\[
\begin{align*}
\max_{\theta} & \quad \sum_{x,a,t} r_{sa} x_t(s,a) \log x_t^*(s,a) \\
\text{s.t.} & \quad \sum_a x_{t+1}^*(s',a) - \sum_s x_t^*(s,a) P(s'|s,a) = 0 \quad \forall s' \in S, \ t = 1:H \\
& \quad \sum_a x_t^*(s,a) = b_t(s) \quad \forall s \in S
\end{align*}
\]

The above program directly operates with occupancy measures \( x \), and does not require any forward-backward message-passing like other applications of EM to planning (Toussaint, Harmeling, and Storkey 2006). Even though EM algorithm is not guaranteed to converge to a global optima, we show that it is globally optimal for MDPs:

**Proposition 1.** EM algorithm converges to the global optimum of the log-likelihood for the MDP inference model in Figure 1(b).

Proof is in the appendix. We next show how to solve the NLP in Table 1 for TIDec-MDPs analogous to the inference model and reasoning developed in this section.

4.2 Inference Model for TIDec-MDP

We now present a mixture of BNs such that MLE in the mixture is equivalent to solving the NLP in Table 1. In EM, optimizing the expected log-likelihood (or the M-step) becomes decoupled resulting in a separate optimization problem for each agent regardless of the number of joint rewards or the number of agents in a joint-reward. This is a significant scalability boost as NLP solvers directly optimize the monolithic program in Table 1 which quickly becomes unscaleable due to a large number of variables/nonlinear terms, whereas EM solves an independent convex program for each agent.

Figure 2 shows the mixture of BNs for TIDec-MDPs. We create one BN for each agent \( i \) (left Figure 2) and one BN for each global reward \( k \in \rho \). The left BN simulates the total local reward from agent \( i \)'s MDP, similar to the BN for an MDP. The structure and interpretation of this BN is same as in previous section. The mixture variable \( M \) can take any integer value in range \([1,n+|\rho|]\) to index each BN, and has a fixed uniform distribution \(=1/(n+|\rho|)\).

To simulate event-based rewards, the right BN (in Figure 2) is created for each global reward \( k \in \rho \). The variable \( \hat{r} \) is binary as before. In addition, we create one binary variable \( y^j_k \) for each agent \( j \in G_k \) involved in the global reward \( \langle E^j_k \forall j \in G_k \rangle \). Intuitively, \( y^j_k = 1 \) implies agent \( j \)'s event \( E^j_k \) occur; vice versa for \( y^j_k = 0 \). W.l.o.g. we assume all rewards \( c^j_k \) are positive (otherwise we can subtract \( r_{\text{min}} \) from them), and \( \theta \) is a positive constant. Let \( y_k \) denote the random vector \((y^j_k \forall j \in G_k)\). Probabilities for this BN are as:

\[
P(y^j_k = 1) = x(E^j_k) \quad \text{and} \quad P(y^j_k = 0) = 1-x(E^j_k) \quad \forall j \in G_k
\]

where positive constant \( r_{\text{max}} \) is chosen such that \((\theta + c^j_k)/r_{\text{max}} \) is less than one (i.e. valid probabilities). Intuitively, (13) connect probabilities of variables \( y^j_k \) with parameters \( x(E^j_k) \) used in the NLP of Table 1, (14) model the condition that global reward \( c^j_k \) is given only when all events \( E^j_k \) happen, otherwise a default reward is awarded.

**Theorem 2.** Let the parameters of the BNs in the mixture model (Figure 2) be set as per (10),(11),(13),(14). Maximizing the likelihood \( P(\hat{r} = 1; x) \) of observing \( \hat{r} = 1 \) in the BN subject to constraints (3),(4),(5) over parameters \( x \) optimally solves the NLP for the TIDec-MDP.

Proof is provided in the appendix.

**M-step** Given theorem 2, we can again use the EM algorithm similar to section 4.1 to iteratively optimize the expected log-likelihood. We directly show the M-step optimization problem (analogous to problem (12)) for TIDec-MDPs. It involves solving one independent convex optimization problem for each agent \( i \).

\[
\begin{align*}
\max_{\mathbf{x}^*, \mathbf{s}^i, \mathbf{a}^i, \mathbf{t}} & \quad \sum_{k \in \rho(i)} \hat{r}^i_{sa} \log x^*_t(s,a) \\
& \quad \sum_{k \in \rho(i)} \log x^*_t(E^i_k) - \theta \sum_{k \in \rho(i)} H_k(x^*_t, x^*_t) \\
& \quad \sum_{a^i} x^*_{t+1}(s',a^i) - \sum_{s',a^i} x^*_t(s,a^i) P(s'|s,a) = 0 \forall t, s, a^i \\
& \quad \sum_{a^i} x^*_t(s,a^i) = b^i_t(s) \quad \forall s, a^i \\
& \quad x^*(E^i_k) = \sum_{i \in K} \sum_{t = 1}^H x^*(s_c, a_c) P^i(s^i_c|s_c, a_c) \forall k \in \rho(i)
\end{align*}
\]

where \( \hat{r}^i_{sa} \) is the normalized local reward for the local MDP of agent \( i \); \( \hat{c}_k \) is the normalized global reward \( (c^j_k/r_{\text{max}}) \); \( \hat{\theta} = \theta/r_{\text{max}} \); \( \rho(i) \) denotes the set of joint-rewards in which agent \( i \) participates. For a global reward \( k \), \( H_k \) denote the cross entropy between previous iteration’s parameters \( x^* \) and current parameter \( x^* \): \(-[x^*(E^i_k) \log x^*(E^i_k) + x^*(E^i_k) \log x^*(E^i_k)]\), where \( x^*(E^i_k) = 1 - x^*(E^i_k) \).

The above M-step optimization highlights the scalability of the EM algorithm which solves one separate optimization problem for each agent \( i \). Note that even though the M-step solves each agent’s sub-problem separately, these sub-problems are still correlated as each sub-problem involves information from other agents and events from previous iteration’s solution. Since previous iteration’s solution is fixed,
different sub-problems can still be solved separately. Each sub-problem is convex and much smaller than the large NLP in Table 1. Thus, the EM algorithm significantly boosts the scalability for a large number of agents.

Other joint-reward types: The EM approach is also extendable to other joint-reward types. In Figure 2, we construct one separate Bayesian network (BN) for each joint-reward \( k \in \rho \). Based on the type of the joint-reward \( k \) (at most \( x \) events or at least \( x \) events), we can set the parameters of its corresponding BN appropriately by modifying (14).

5 RL for Event-Based Rewards

The previous section presented a scalable EM algorithm for TIDec-MDPs. However, the scalability still suffers when the state-space of each agent \( i \) is exponential, which is often the case for several patrolling and coverage problems (Yehoshua, Agmon, and Kaminka 2015). To address such settings, we develop a reinforcement learning (RL) approach that uses function approximators such as deep neural nets (NN) to represent agent policies and optimizes them using the policy gradient approach (Williams 1992; Sutton et al. 1999). Policy gradient is a natural counterpart to the EM algorithm—Schulman et al. (2015) show that EM is the surrogage loss function to the EM algorithm—Schulman et al. (2015) show that EM optimizes a surrogate loss function by minimizing an upper bound on this function, whereas policy gradient uses gradient descent to minimize the same loss function. Empirically, we observed that for small and medium-sized problems, EM is faster (RL is slower but comparable in solution quality). The main advantage of RL lies for large problems where EM cannot scale well due to its tabular policies.

Given the start state \( s_1 \) at time step 1 for all the agents; each agent \( i \)'s policy parameterized using \( \theta^i \) (which represent NN parameters), our goal is to compute the gradient of global-value function (7):

\[
\nabla_{\theta^i} GV(s_1; \pi) = \nabla_{\theta^i} V^i(s_1^i; \pi^i) + \nabla_{\theta^i} \sum_{k \in \rho(i)} \sum_{j \in G_k} P(E_k^i) \tag{19}
\]

\[
= \nabla_{\theta^i} V^i(s_1^i; \pi^i) + \sum_{k \in \rho(i)} \nabla_{\theta^i} P(E_k^i) \prod_{j \in G_{k \setminus \{i\}}} P(E_j^i) \tag{20}
\]

where we have used the fact that agent \( i \)'s policy does not affect agent \( j \)'s local value function \( V^j \) and event probabilities \( P(E_j^i) \) to simplify the expression; \( \rho(i) \) denotes the set of joint-rewards in which agent \( i \) participates. The gradient of local MDP value function \( \nabla_{\theta^i} V^i(s_1^i; \pi^i) \) can be computed using the technique in (Sutton et al. 1999). We next focus on how the gradient backpropagates from event-based rewards. Given that we have \( P(E; \pi^i) = \sum_{e \in E} P(e; \pi^i) \), gradient is:

\[
\nabla_{\theta^i} P(E; \pi^i) = \sum_{e \in E} \nabla_{\theta^i} P(e; \pi^i) \tag{21}
\]

We next compute the gradient of a proper, primitive event or \( \nabla_{\theta^i} P(e; \pi^i) \). Let \( e = (s^i, a^i, \hat{s}^i) \) be a primitive event for agent \( i \). The probability of this event is given as:

\[
P(e) = \prod_{t=1}^{H} P^e(s^i_t | s^i_0) P^e(a^i_t | s^i_t, \hat{s}^i_t) = \sum_{t=1}^{H} \prod_{s^i_t, a^i_t} P^e(s^i_{t+1}, a^i_t) I(\langle s^i, a^i, s^i_{t+1} \rangle = \langle \hat{s}^i, \hat{a}^i, \hat{s}^i_{t+1} \rangle)
\]

where \( I \) is the indicator function returning one when input logical condition is true, zero otherwise. The gradient is:

\[
\nabla_{\theta^i} P(e) = \sum_{t=1}^{H} \prod_{s^i_t, a^i_t} P^e(s^i_{t+1}, a^i_t) \left[ \sum_{e' | e} \nabla_{\theta^i} \log p_i(e' | s^i_t) \right] \times I(\langle s^i_t, a^i_t, s^i_{t+1} \rangle = \langle \hat{s}^i, \hat{a}^i, \hat{s}^i_{t+1} \rangle) \tag{22}
\]

where we used the log derivative trick (Schulman et al. 2015) that allows sampling-based evaluation of \( \nabla_{\theta^i} P(e) \). Consider a complete state-action trajectory \( \Phi^i = (s_1^i, a_1^i) \). For any \( \Phi^i \), the proper, primitive event \( e \) can only occur at most once (see definition 3). Based on this fact, we derive the stochastic gradient estimate as (complete derivation in paper’s extended version):

\[
\nabla_{\theta^i} P(e) = \frac{1}{|\xi|} \sum_{e \in \xi, \Phi^i(e) = e} \left( \sum_{t=1}^{t(e, \Phi^i)} \nabla_{\theta^i} \log p_i(e' | s^i_t) \right) \tag{23}
\]

\( \xi \) is the set of complete state-action samples from \( P^\pi(s_1^i, a_1^i, \Phi^i) \); sum over those samples \( \Phi^i \) such that event \( e \) occurs in \( \Phi^i \) at some point in time (denoted using \( t(e, \Phi^i) \)). We can use the above gradient within (21) to compute the gradient of an event, and use it in (20) to compute the required gradient. Given sample set \( \xi \) for each agent, we can also empirically compute the probability estimates of primitive events \( e \), and use it to compute empirical estimate of events \( P(E_k^i) \), which can be used in (20). Thus, we have shown how the policy gradient \( \nabla_{\theta^i} GV \) for TIDec-MDPs can be computed using sampling.

Other joint-reward types: We can also derive policy gradient for other types of events mentioned in section 2. E.g., we show the gradient of \( JV(\rho; \pi) \) (in eq (8)) when constraints \( k \in \rho \) are of the type at least one event:

\[
\nabla_{\theta^i} JV = \sum_{k \in \rho(i)} \sum_{j \in G_{k \setminus \{i\}}} \nabla_{\theta^i} P(E_k^i) \prod_{j \in G_{k \setminus \{i\}}} (1 - P(E_j^i)) \tag{24}
\]

One can compute \( \nabla_{\theta^i} P(E_k^i) \) as earlier.

Multiagent credit assignment: A key problem while learning from global rewards in multiagent setting is that the gradient computed for an agent \( i \) does not explicitly reason about the contribution of that agent to the global team reward. As a result, the gradient becomes noisy given that other agents are also exploring, leading to poor quality solutions (Foerster et al. 2017; Bagnell and Ng 2005). Fortunately, creating a separation among local MDPs of agents and joint-event-based-rewards automatically addresses this problem of noisy gradient in TIDec-MDPs. E.g., consider the gradient (24) for at least one event w.r.t. agent \( i \)'s parameters. Intuitively, if another agent \( j \) has a high probability of finishing event \( E_k^j \) (or \( P(E_k^j) \approx 1 \)), then the term \( (1 - P(E_k^j)) \) in \( \nabla_{\theta^i} JV \) would discourage agent \( i \) to increase its probability \( P(E_k^j) \). Similarly, if all other agents \( j \) have a very low probability of performing \( E_k^j \) then \( (1 - P(E_k^j)) \) would be high encouraging agent \( i \) to perform its event \( E_k^i \) for higher joint-value. Thus, event-based rewards help make policy gradients precise and accurate in TIDec-MDPs.
6 Experiments

We tested on two domains—Mars rover and the Multiagent coverage problem. Domain details and other experimental settings are provided in the extended version of the paper.

Mars rover domain: This domain is adapted from (Becker et al. 2004). Agents have private and shared locations from which they can collect data. Joint-rewards are given when all the involved agents perform their corresponding events on the shared location.

Runtime: Figure 3a shows the runtime results for NLP and EM for the four categories, ⟨Easy, Medium, Hard, All⟩, of problems on the x-axis. The results are averaged over ⟨30, 40, 35, 105⟩ instances corresponding to the four categories (‘All’ includes all the instances). The categories ⟨Easy, Medium, Hard⟩ include ⟨10, 30, 40⟩ agents, ⟨4, 5, 7⟩ locations per agent and ⟨10, 18, 25⟩ agents sharing a common site respectively.

Figure 3a shows that EM has a much lower runtime on an average. Only those problems that got solved by NLP within the cutoff time of 30 minutes were included in the figure (EM being an anytime algorithm, always returns a solution). For the hard instances, EM was much faster than the NLP solver showing that EM’s strategy of solving independent program per agent results in significant speedups over NLP. Figure 3b highlights the anytime nature of EM by showing the normalized solution quality of EM for each iteration averaged over 35 Hard instances. The solution quality increases monotonically with iterations till convergence and reaches to ⟨98.7%, 99.6%⟩ of final converged quality within the first ⟨100, 200⟩ iterations. Similar trends hold for Easy and Medium problems.

Solution Quality: Figures 3c and 3d present solution quality results for NLP and EM. In particular, the x-axis of both the figures show the cutoff time in minutes, i.e., maximum allowable time for the algorithms to return a solution. The y-axis of Figure 3c shows the ratio (#{ of instances solved by NLP}/#{ of instances solved by EM}) (in %) for the time limit indicated on the x-axis. Figure 3c shows that the ratio increases monotonically as the time limit increases for all categories. For Hard instances, NLP could provide a solution for only 40% of problems, and did not terminate for the rest of the problems. In contrast, EM always converged within 30 min for all the instances, confirming its better scalability.

The y-axis of Figure 3d shows the ratio (in %) of total average rewards obtained by NLP w.r.t. the EM within the cutoff time on the x-axis. This figure captures the reward for all the experiments performed. That is, if NLP does not return a solution within the cutoff time, a valid random policy was assigned for evaluation purposes. Trends remain similar as in Figure 3c. For Easy category, the ratio becomes 100% within 15 minutes while it remains below 71% for Hard cases even after 30 minutes (and does not reach 100% for any other category). To summarize, the anytime property of EM provides a significant runtime as well as solution quality advantage (with performance gap widening significantly as we move from Easy to the Hard category of problems).

Multiagent coverage problem: For testing the scalability of our multiagent RL (MARL), we experimented with the multiagent coverage problem introduced in section 1 (detailed domain settings are in the extended version of the paper). The problem involves inspecting different locations within a mass rapid transit (MRT) network. Reward is given when a location is inspected at least once within a fixed timeframe, such as every 1 hour or half hour. These problems are challenging for EM/NLP as the state-space is exponential in the number of locations in the MRT system. We first tested MARL on relatively tractable instances that were solvable by EM. These instances involved up to 2 lines in the MRT map, with a maximum of 3 private locations per line, and a maximum of 3 shared locations. Figure 5a shows that MARL achieves similar solution quality as EM for all the instances (within 97% of EM’s quality on average). However, the average runtime of MARL (5 hours) was significantly higher than EM’s runtime (10 minutes). Thus, for relatively smaller instances both EM and MARL provide similar quality, but EM is preferable because of its lower runtime.

For hard settings, we tested on the Singapore MRT map with 5 lines, each having 20 private locations per line, and 20 shared locations. Each line has a single agent able to move among locations on the line. Shared locations correspond to interchange stations where multiple lines meet. Thus, shared locations can be inspected by multiple agents. The time horizon was 1024 minutes (17 hours). The joint reward to inspect any shared location was much higher than private locations given that shared locations are heavily crowded and thus
more important. To claim the reward, agents must success-
fully inspect a location once every \( k \) time steps (also called \textit{reset} time), and \( k \) (in hours) varied in the range \{0.5, 1, 2, 4\}. The inspect action consumes 15 minutes, and moving to the
next location on the line takes 3 minutes. Inspecting a location
multiple times within the reset time window does not fetch an additional reward. Thus, agents are required to in-
crease their coverage to gain additional rewards. Similarly, for shared locations, inspection by any single agent is suf-
cient to get the reward. Thus, agents are also incentivized to coordinate with each other to avoid multiple agents inspect-
ing the same shared location within the reset time window.
We model such global rewards using the \textit{at least one event} semantics.

For these problems, EM and NLP were unable to scale
due to the large state-space of the problem. However, MARL
scaled well for these problems due to its neural network
based parametric policies and policy gradient based opti-
mization (settings for neural network and gradient ascent
are provided in the extended version of the paper). For fig-
ures 4 and 5b, y-axis shows average reward value at each
iteration which is computed as a moving average of last
100 iteration values (i.e., value at that iteration and previ-
ous 99 values). Figure 5b shows the average reward quality
achieved by MARL for different settings of the reset time \( k \).
A smaller value of \( k \) implies agents can claim rewards fre-
cently. Therefore, we can observe that the reward is higher
for lower \( k \) values. For all the settings, despite the large
problem size with a long horizon, convergence was achieved
in about 5000 iterations.

Figure 4: Quality comparisons between coordinated RL (MARL) and independent policy optimization (I-RL) for varying reset time \( k \) (in hours). Gain is % quality improvement by MARL over I-RL upon convergence.

Coordinated vs. Uncoordinated RL: We tested the MARL
approach against a baseline method that independently opti-
mized the policy of each agent using policy gradient based RL (‘uncoordinated’ RL). We also observed that when
shared locations are fewer, agents need to coordinate tightly
to gain the reward from inspecting these shared locations.

To test this hypothesis, we reduced the number of shared loca-
tions down to 10, from 20 in the real map. Figure 4 shows
quality improvements by MARL over independent policy
optimization (I-RL). The figure shows four plots for values
of \( k \) (in hours) varied in the range \{0.5, 1, 2, 4\}. The un-
certainty in the figure represents the standard deviation in a
moving interval of 100 iterations. From the plots we can see
that the benefit of MARL is higher for lower reset time \( k \)
as it provides more opportunities to inspect shared locations
and claim higher reward from coordinated actions. These re-
sults clearly illustrate that our MARL approach provides sig-
nificant benefits when agents learn in a coordinated fashion
versus learning independently.

7 Conclusion

We addressed multiagent decision making in settings where
joint-rewards may depend upon entire state-action histories
of agents. Such history-dependent rewards can capture the
notion of events and tasks in multiagent planning. We de-
veloped a scalable approach for this setting by translating the
problem to that of inference in a graphical model. The re-
sulting EM algorithm was shown to be more scalable than
the standard nonlinear program. We also developed a multi-
agent RL approach that optimized agent policies represented
as neural networks using stochastic policy gradients. Our in-
ference and RL-based advances enabled us to solve large
synthetic problems and a realistic multiagent coverage prob-
lem for schedule coordination of agents in an MRT network
where other approaches failed to scale.

Acknowledgements

This research was supported by the Singapore Ministry of
Education (MOE) Academic Research Fund (AcRF) Tier 1
grant. We also thank Kohli Center on Intelligent Systems, In-
ternational Institute of Information Technology, Hyderabad
for the generous support.
References
Appendix

1 Proposition 1

EM algorithm converges to the global optimum of the log-likelihood for the MDP inference model in figure 1(b).

Proof. From the established convergence results by Wu, C. F. J. 1983. On the convergence properties of the EM algorithm. Annals Of Statistics, it is known that the EM algorithm would converge to a stationary point of the log-likelihood function if the expected log-likelihood $Q(x, x^\star)$ is continuous in both the parameters $x$ and $x^\star$. This is the case for our $Q(\cdot)$ function.

The likelihood function for our case, $P(\hat{r} = 1; x)$, is given in Theorem 1. The likelihood function is linear in parameters $x$. Thus, the log of this function (or log-likelihood) is concave in parameters $x$ as the log of an affine function is concave. It is a standard result in convex optimization that a stationary point of a concave function subject to linear constraints is also a global optima. Hence, EM would converge to a global optima of the log-likelihood, which is equivalent to the optimal policy.

2 Theorem 2

Let the parameters of the BNs in the mixture model (figure 2) be set as per (10),(11),(13) and (14). Maximizing the likelihood $P(\hat{r} = 1; x)$ of observing $\hat{r} = 1$ in the BN subject to constraints (3),(4),(5) over parameters $x$ optimally solves the NLP for the TIDec-MDP.

Proof. The prob $P(\hat{r} = 1; x)$ in the mixture model (fig 2) given as:

$$= \frac{1}{n + |\rho|} \left[ \sum_{i=1}^{n} P(\hat{r} = 1|M = i) + \sum_{k=n+1}^{n+|\rho|} P(\hat{r} = 1|M = k) \right]$$

(25)

We know from section 4.1 that $P(\hat{r} = 1)$ in the BN for each agent $i$ is directly proportional to the value function of agent $i$’s MDP (see theorem 1). Therefore, the summation $\sum_{i=1}^{n} P(\hat{r} = 1|M = i)$...
$1|M = i)$ simulates the first summation in the NLP. We next focus on $P(\hat{r} = 1|M = k)$ in the BN for the global reward $k$.

$$P(\hat{r} = 1|M = k) = \sum_{y_k} P(y_k) P(\hat{r} = 1|y_k)$$

$$= P(y_k = 1)P(\hat{r} = 1|y_k = 1) + \sum_{y_k \neq 1} P(y_k \neq 1)P(\hat{r} = 1|y_k \neq 1)$$

where $y_k = 1$ implies that each variable in the random vector $y_k$ is one, and $y_k \neq 1$ implies there is at least one $y^i$ that is zero.

$$P(\hat{r} = 1|M = k) = P(y_k = 1) \theta + c_k \frac{\theta}{r_{\text{max}}} + \sum_{y_k \neq 1} P(y_k \neq 1) \theta$$

$$\propto P(y_k = 1)(\theta + c_k) + \theta(1 - P(y_k = 1)) \quad (26)$$

$$= P(y_k = 1)c_k + \theta \prod_{j \in G_k} x(E^j_t) + \theta \quad (27)$$

where we have used the fact that $\sum_{y_k \neq 1} P(y_k \neq 1) = 1 - P(y_k = 1)$ in (26). From (27), we can clearly see that the $P(\hat{r} = 1|M = k)$ in the BN for the global reward $k$ is directly proportional to the actual contribution $c_k \prod_{j \in G_k} x(E^j_t)$. Therefore, the likelihood in (25) is directly proportional to the objective of NLP in table 1 proving the theorem.

## 3 M-step Objective

In this section, we will derive the M-step objective for TIDec-MDPs. From equation 12 in the paper, we know the M-step objective for agent $i$’s local MDP using BN in figure 2 is as follows:

$$\max_{x^{i*}} \sum_{s^i, a^i, t} \sum_{i^j} r_{i^j | a^i, t} x_i^*(s^i, a^i) \log x_i^*(s, a) \quad (28)$$

Now let us consider the BNs for joint rewards in figure 2. Our objective becomes:

$$Q(x, x^*) = \sum_{k \in \rho} \sum_{y_k} P(\hat{r} = 1, y_k|x) \log P(\hat{r} = 1, y_k|x^*)$$

Consider a fixed constraint $\rho_k$. We simplify the above summation for this constraint as below:

$$Q(x, x^*) = P(\hat{r} = 1, y_k = 1|x) \log P(\hat{r} = 1, y_k = 1|x^*) +$$

$$\sum_{y_k \neq 1} P(\hat{r} = 1, y_k \neq 1|x) \log P(\hat{r} = 1, y_k \neq 1|x^*) \quad (29)$$
where $y_k = 1$ implies that each variable in the random vector $y_k$ is one, and $y_k \neq 1$ implies there is at least one $y_k^i$ that is zero. The first term can be simplified as:

$$P(\hat{r} = 1, y_k = 1 | x) \log P(\hat{r} = 1, y_k = 1 | x) = (\hat{\theta} + \hat{c}_k) P(y_k = 1 | x) \sum_{i \in G_k} \log P(y_k^i = 1 | x^*)$$  \hspace{1cm} (30)$$

We know that $\log P(\hat{r} = 1, y_k \neq 1 | x^*) = \log P(\hat{r} = 1 | y_k \neq 1; x^*) + \log P(y_k \neq 1; x^*)$. We focus on simplifying on the second log term with $\log P(y_k \neq 1; x^*)$ as below:

$$\sum_{y_k \neq 1} P(\hat{r} = 1, y_k \neq 1; x) \log P(y_k \neq 1; x^*) = \sum_{y_k \neq 1} P(\hat{r} = 1, y_k \neq 1; x) \left[ \sum_{i \in G_k} \log P(y_k^i; x^*) \right]$$  \hspace{1cm} (31)$$

$$= \sum_{i \in G_k} \left[ \sum_{y_k \neq 1} P(\hat{r} = 1, y_k \neq 1; x) \log P(y_k^i; x^*) \right]$$  \hspace{1cm} (32)$$

The summation $\sum_{y_k \neq 1}$ can be described intuitively as the summation over all the possible settings of $y_k^i \forall i$ such that at least one of the $y_k^i$ is zero. We use this fact to break the above summation as:

$$= \sum_{i \in G_k} \left[ \sum_{y_k^i = 0} P(\hat{r} = 1, y_k^i = 0, y_k^{-i}; x) \log P(y_k^i = 0; x^*) + \sum_{y_k^i = 1} P(\hat{r} = 1, y_k^i = 1, y_k^{-i}; x) \log P(y_k^i = 1; x^*) \right]$$  \hspace{1cm} (33)$$

Considering left term in Equation 33:

$$= \sum_{i \in G_k} \left[ \sum_{y_k^i = 0} \hat{\theta} P(y_k^i = 0; x) P(y_k^{-i}; x) \log P(y_k^i = 0; x^*) + \sum_{y_k^i = 1} \hat{\theta} P(y_k^i = 0; x) P(y_k^{-i}; x) \log P(y_k^i = 1; x^*) \right]$$  \hspace{1cm} (34)$$

$$= \sum_{y_k^{-i}} \hat{\theta} P(y_k^{-i}; x) \log P(y_k^i = 0; x^*) = \hat{\theta} P(y_k^i = 0; x) \log P(y_k^i = 0; x^*)$$  \hspace{1cm} (35)$$
Considering right term in Equation 33:

\[
= \sum_{y_{k_i} \neq 1} P(\hat{r} = 1, y_{k_i} = 1, y_{k_i}^{-i}; x) \log P(y_{k_i} = 1; x^*)
\]

\[
= \sum_{y_{k_i} \neq 1} P(\hat{r} = 1|y_{k_i} = 1, y_{k_i}^{-i}, x)P(y_{k_i} = 1, y_{k_i}^{-i}; x) \log P(y_{k_i} = 1; x^*)
\]

\[
= \sum_{y_{k_i} \neq 1} \hat{\theta}P(y_{k_i} = 1; x)P(y_{k_i}^{-i}; x) \log P(y_{k_i} = 1; x^*)
\]

\[
= \hat{\theta}P(y_{k_i} = 1; x) \log P(y_{k_i} = 1; x^*) \sum_{y_{k_i} \neq 1} P(y_{k_i}^{-i}; x)
\]

\[
= \hat{\theta}P(y_{k_i} = 1; x) \log P(y_{k_i} = 1; x^*) \left( 1 - P(y_{k_i}^{-i} = 1; x) \right)
\]

Therefore, the simplified form of (33) is:

\[
\sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 0; x) \log P(y_{k_i} = 0; x^*) + \sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 1; x) \left( 1 - P(y_{k_i}^{-i} = 1; x) \right) \log P(y_{k_i} = 1; x^*)
\]

\[
= \sum_{k \in \rho} \left[ \hat{c}_k P(y_{k} = 1; x) \left( \sum_{i \in G_k} \log P(y_{k_i} = 1; x^*) \right) + \hat{\theta}P(y_{k} = 1; x) \log P(y_{k} = 1; x^*) \right] + \sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 0; x) \log P(y_{k_i} = 0; x^*) + \sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 1; x) \log P(y_{k_i} = 1; x^*)
\]

\[
= \sum_{k \in \rho} \left[ \hat{c}_k P(y_{k} = 1; x) \left( \sum_{i \in G_k} \log P(y_{k_i} = 1; x^*) \right) + \hat{\theta}P(y_{k} = 1; x) \log P(y_{k} = 1; x^*) \right] - \hat{\theta} \sum_{i \in G_k} H_k(x^i, x^{i*})
\]

\[Q(x, x^*) = \sum_{k \in \rho} \left[ (\hat{\theta} + \hat{c}_k)P(y_{k} = 1; x) \left( \sum_{i \in G_k} \log P(y_{k_i} = 1; x^*) \right) + \sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 0; x) \log P(y_{k_i} = 0; x^*) + \sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 1; x) \log P(y_{k_i} = 1; x^*) \right] \]

\[+ \hat{\theta} \sum_{i \in G_k} H_k(x^i, x^{i*}) \]

\[\sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 0; x) \log P(y_{k_i} = 0; x^*) + \sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 1; x) \log P(y_{k_i} = 1; x^*) \]

\[= \sum_{k \in \rho} \left[ \hat{c}_k P(y_{k} = 1; x) \left( \sum_{i \in G_k} \log P(y_{k_i} = 1; x^*) \right) + \hat{\theta}P(y_{k} = 1; x) \log P(y_{k} = 1; x^*) \right] - \hat{\theta} \sum_{i \in G_k} H_k(x^i, x^{i*}) \]

\[\sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 0; x) \log P(y_{k_i} = 0; x^*) + \sum_{i \in G_k} \hat{\theta}P(y_{k_i} = 1; x) \log P(y_{k_i} = 1; x^*) \]

\[= \sum_{k \in \rho} \left[ \hat{c}_k P(y_{k} = 1; x) \left( \sum_{i \in G_k} \log P(y_{k_i} = 1; x^*) \right) + \hat{\theta}P(y_{k} = 1; x) \log P(y_{k} = 1; x^*) \right] - \hat{\theta} \sum_{i \in G_k} H_k(x^i, x^{i*}) \]
Let $\rho(i)$ denote the joint rewards in which agent $i$ participates in. Without loss of generality, we also assume that for each joint reward constraint, agent $i$ has only one event. Equation (42) can be decomposed for agent $i$ as follows:

$$\sum_{k \in \rho(i)} \hat{c}_k \left( \prod_{j \in G_k} \sum_{i \in \rho(i)} x(E^i_k) \right) \log x^*(E^i_k) - \hat{\theta} \sum_{k \in \rho(i)} H_k(x^*, x^{i^*})$$

(43)

Combining (28) and (43), we get Mstep for agent $i$ as shown in the paper:

$$\max_{x^{i^*}} \sum_{s^i, a^i, t} r^i_{s^i,a^i} x^i_{s^i,a^i}(s^i, a^i) + \sum_{k \in \rho(i)} \hat{c}_k \left( \prod_{j \in G_k} x(E^j_k) \right) \log x^*(E^j_k) - \hat{\theta} \sum_{k \in \rho(i)} H_k(x^*, x^{i^*})$$

(44)

4 RL Gradient

We first consider the case for standard TI-Dec-MDPs, and later extend the analysis to The global-value function for a TI-Dec-MDP is given as:

$$GV(\pi) = \sum_{i=1}^{n} V_{\pi_i}(s^i_0) + JV(\rho; \pi)$$

(45)

where $JV$ is the joint value function for all the events. It is defined as:

$$JV(\rho|\pi) = \sum_{k=1}^{|ho|} \sum_{j=1}^{|G_k|} c_k \prod_{j=1}^{|G_k|} P(E^{g_{kj}}_k)$$

(46)

Let the policy of an agent $i$ be parameterized by $\theta^i$. Let $\rho(i)$ denotes all the constraints in which agent $i$ is involved. The gradient of the global value function is given as:

$$\nabla_{\theta^i} GV(\pi) = \nabla_{\theta^i} V_{\pi_i}(s^i_0) + \sum_{k \in \rho(i)} \prod_{j=1}^{|G_k|} P(E^{g_{kj}}_k)$$

(47)

$$= \nabla_{\theta^i} V_{\pi_i}(s^i_0) + \sum_{k \in \rho(i)} c_k \nabla_{\theta^i} P(E^{g_{kj}}_k) \prod_{j \neq i} P(E^{g_{kj}}_k)$$

(48)

Therefore, to compute the global gradient, we need to compute the gradient of a proper event $E$. Let us fix the agent to be $i$, and focus on a event $E$ that can occur during the agent’s
execution history. The probability (and gradient) of the event is given as:

\[ P(E; \pi^i) = \sum_{e \in E} P(e; \pi^i) \]  

\[ \nabla_{\theta^i} P(E; \pi^i) = \sum_{e \in E} \nabla_{\theta^i} P(e; \pi^i) \]

Therefore, now our problem becomes computing the gradient of primitive event \( e \). This computation is shown below.

### 4.1 Primitive event gradient

Let \( e = \{ \hat{s}_i, a_i, \hat{s}' \} \) be a primitive event for agent \( i \). It can occur only once in any state-action history. The probability of this event is given as:

\[ P^\pi(e) = \sum_{t=1}^{H} P^\pi(\hat{s}_i|t) \pi(a_i|\hat{s}_i) T(\hat{s}'|\hat{s}, a_i) \]  

\[ = \sum_{t=1}^{H} \sum_{s_{1:t+1}, a_{1:t}} P^\pi(s_{1:t+1}, a_{1:t}) \mathbb{I}(\langle s_t, a_t, s_{t+1} \rangle = \langle \hat{s}, \hat{a}, \hat{s}' \rangle) \]  

\[ \nabla_{\theta^i} P^\pi(e) = \sum_{t=1}^{H} \sum_{s_{1:t+1}, a_{1:t}} P^\pi(s_{1:t+1}, a_{1:t}) \left[ \sum_{t'=1}^{t} \nabla_{\theta} \log \pi(a_{t'}|s_{t'}) \right] \mathbb{I}(\langle s_t, a_t, s_{t+1} \rangle = \langle \hat{s}, \hat{a}, \hat{s}' \rangle) \]  

\[ = \sum_{s_{1:H+1}, a_{1:H}} P^\pi(s_{1:H+1}, a_{1:H}) \sum_{t=1}^{H} \left( \left[ \sum_{t'=1}^{t} \nabla_{\theta} \log \pi(a_{t'}|s_{t'}) \right] \mathbb{I}(\langle s_t, a_t, s_{t+1} \rangle = \langle \hat{s}, \hat{a}, \hat{s}' \rangle) \right) \]  

\[ P^\pi(e) = \sum_{s_{1:H+1}, a_{1:H}} P^\pi(s_{1:H+1}, a_{1:H}) \sum_{t=1}^{H} \mathbb{I}(\langle s_t, a_t, s_{t+1} \rangle = \langle \hat{s}, \hat{a}, \hat{s}' \rangle) \]  

Notice that a proper primitive event happens at most once for any execution history \( \Phi = (s_{1:H+1}, a_{1:H}) \). If event \( e \) does not happen in \( \Phi \), then the inner summation would be zero. Therefore, we sum only over those execution histories in which event \( e \) happens:

\[ \nabla_{\theta^i} P^\pi(e) = \sum_{\Phi: \Phi\models e} P^\pi(\Phi) \sum_{t=1}^{H} \left( \left[ \sum_{t'=1}^{t} \nabla_{\theta} \log \pi(a_{t'}|s_{t'}) \right] \mathbb{I}(\langle s_t, a_t, s_{t+1} \rangle = \langle \hat{s}, \hat{a}, \hat{s}' \rangle) \right) \]  

\[ 6 \]
Notice also that a proper primitive event happens at most once. Let the time at which event $e$ happens in $\Phi$ be $t(e, \Phi)$. We simplify using this fact:

$$\nabla_\theta P_\pi(e) = \sum_{\Phi: \Phi \models e} P_\pi(\Phi) \left[ \sum_{t'=1}^{t(e, \Phi)} \nabla_\theta \log \pi(a_{t'}|s_{t'}) \right]$$  \hspace{1cm} (58)

A practical way to compute the above gradient is to sample first the entire trajectories $s_1^\xi, a_1^\xi, s_2^\xi$. For each sample $\xi$, compute the only time $t$ at which the event $e$ takes place. The gradient estimate is:

$$\nabla_\theta P_\pi(e) = \frac{1}{M} \sum_{k=1}^{M} \sum_{t'=1}^{t(\xi^k)} \nabla_\theta \log \pi(a_{t'}|s_{t'})$$  \hspace{1cm} (59)

5 Experimentation Details

We ran our experiments on a 24 core machine with 48 Gb RAM. Non linear program and Mstep for EM was modelled using AMPL $^1$ and solved using Bonmin $^2$ solver.

5.1 Domain 1 - Mars Rover

There are multiple rovers (say three), each with their own region of space to explore. Let the total time to mission is $T$. Each region has some private locations $P$ as well as some shared locations $SP$ with boundary regions of other agent. Therefore the total locations available for agent are $L = P \cup SP$. Each location $x$ has a value for the data collected there, $r(x)$, and a time to collect the data, $t(x)$. For going from location $x$ to location $y$, time incurred by agent is $\hat{t}(x, y)$. Suppose each region has 3 (can be any variable quantity) locations to search. The local state set $S_i$ for agent $i$ is defined as the current site $l$ within its own region where the agent is, time left for the mission $t$, and data to be collected from the sites $d_1, d_2, d_3$, where $d_i$ is either 0 or 1. Formally, a state for agent $i$ is defined as a tuple:

$$S = \{ (l, t, d_i) \mid l \in L, t \in \{0..T\}, d_i \in \{0, 1\} \forall i \in \{1..|L|\} \}$$


$^3$Domain derived from Becker et al. 2004
The actions available to agents are \textit{goto} location \(x\), \(\forall x \in L\) and \textit{collect} at current location. Both actions are stochastic with probabilities \(\alpha\) and \(\beta\) respectively and whether action is successful or not, time remaining for the mission \(t\) always reduces (reduces by \(t(x)\) if \textit{collect} action is taken and \(\hat{t}(x, y)\) if \textit{goto} action is taken from location \(x\) to location \(y\).) If \textit{collect} action is successful at site \(x\), then agent will get a reward \(r(x)\) once and \(d_x\) will toggle to 1. Agent will not get any reward for successful collection at already collected sites.

For each shared location on the map, there is a joint reward structure containing events for agents where they have to collect data at shared site \(x\), where site \(x\) is the shared location \((x \in SP)\) within region of that agent. Event \(E^i_x\) would correspond to agent \(i\) collecting data at site \(x\). The primitive events would be of the form:

\[
\{(s, a, s') | s = (l = x, t, d_l = 0), a = \text{collect}, s' = (l = x, t - t(x), d_l = 1)\}
\]

Agents would have to optimize expected reward deciding which site to visit (since it takes up time to travel a distance) and which site to collect (takes up time to collect and also based on different rewards \(r(x)\) for different sites).

\section*{5.1.1 Experimentation}

We tested out EM and NLP algorithms on three different settings i.e. Easy, Medium and Hard. Easy settings include upto 10 agents and 4 locations per agent and maximum 10 agents sharing a shared site. [This implies a product of 10 variables in NLP, which Becker et. al was not able solve. Medium settings include upto 30 agents and 5 locations per agent and maximum 18 agents sharing a shared site. Hard settings include upto 40 agents and 7 locations per agent and maximum 25 agents sharing a shared site.

\section*{5.2 Domain 2 - Multiagent Coverage Problem}

Metro transit systems in many places across the world are divided into lines e.g., the Singapore metro transit system has 5 lines (red, green, yellow, blue and purple). Each line \(l_i\) covers a set of stations \(X_i\). Lines can have overlapping stations i.e. shared stations (stations not shared are private to \(l_i\)). Therefore \(X_i\) comprises of set \(P\) of some locations private to \(l_i\) and set \(SP\) of shared locations with other lines. Each line has a single agent able to move among locations on the line. Shared locations correspond to interchange stations where multiple lines meet.

- The local state set \(S_i\) for agent \(i\) is same as in domain 1, with set of locations \(L = X_i\). Each agent in a patrol unit can choose to either \textit{goto next} or \textit{goto previous} site from current site or take an \textit{inspect} action on current site. The agents are given a patrol time of \(k\) time units, in which they have to inspect maximum number of locations. Both
the inspect and go to action are probabilistic and consumes agent’s time. There can be multiple such patrols possible. The total horizon for the problem is $T$.

- When agent takes an inspect action on site $s$, it gets an MDP reward $r(s)$ associated with site $s$ on successful inspection. This inspect action marks the site as inspected until the next patrol begins, where all sites are again marked un inspected. If agent tries to re-inspect an already inspected site in the same patrol, it will get no reward. However other agents can still choose to inspect the site.

- From Singapore MRT map, each shared location on the map lies at an intersection of maximum 3 lines. For every patrol and for each such shared location on the map, there is a joint reward structure, containing events for agents with the following task: *At least one agent out of all agents who share the location should successfully inspect the shared location for joint reward being added to the system*. This marks an essential difference with respect to the definition of joint reward structure from (Becker et al. 2004), where all agents need to collect at shared location for getting the joint reward. In this domain, even if one agent successfully inspects a shared location in current patrol, an extra constraint reward for that location $c(s)$ is added to the system. Let event $E^i_x$ correspond to agent $i$ inspecting at location $x$. The primitive events would be of the form:

$$\{(s, a, s') | s = \langle l = x, t, d_l = 0 \rangle , a = inspect, s' = \langle l = x, t - t(x), d_l = 1 \rangle \},$$

Figure 1: Singapore MRT map
where $t(x)$ is time taken by inspection action.

5.2.1 Experimentation

For verifying our RL implementation, we compared learned RL policy against EM policy for 20 experiments with easy settings. Easy settings involved upto 2 lines in an MRT map, with a maximum 3 private locations per line, and a maximum of 3 shared locations. For hard settings, we tested out our problem on Singapore MRT map with 5 lines, having 20 private locations per line and 20 shared locations shared by a minimum of 2 lines and a maximum of 3 lines. It is not possible to generate state spaces for these settings in a reasonable time, since state space is exponential in number of locations. Time allotted to the agents was 17 hours, with each inspection at a location taking 16 minutes and movement between locations taking 4 minutes. The joint reward was much bigger than the MDP reward for every shared location. The repetition of patrols was varied among different experiments as every $k$ hours, where $k \in \{0.5, 1, 2, 4\}$. For instance, $k = 0.5$ implies that for all shared locations, at least one agent who shares it should successfully inspect it every 0.5 hours. In other words, there will be a new patrol every $k$ hours.

5.2.2 Network Architecture

We use two neural networks per agent, one for learning the value function and other for learning the policy for the agent. Learned value function also acts as a baseline \(^4\) for policy gradient. We will describe the architecture in detail now.

- **Input Layers:** The input to both the neural networks is a state represented as an array of binary bits. The input representation consists of one-hot representation of current location, one-hot representation of remaining time in the mission and the inspection bits for all locations $L$ of the agent. For example, given a domain with 4 time units as total horizon, a single agent with 3 locations, the state representation for agent with current location 2, remaining time in the mission as 2 units will be written as:

$$S_{rep} = \{0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0\}$$

The first three bits indicate the current location as 2, current time remaining in the mission as 2 time units and last three bits indicate the inspection status for the current patrol i.e. location 1 is already inspected by the agent in current patrol. Please note, that once next patrol begins, all inspection bits will be reverted to 0 again.

\(^4\)Sutton et al. 1999
• Output Layers: The output from value function NN is $V(s)$ for the input state using a linear activation. Output from policy NN is $\pi(a|s)$ i.e. probabilities for all possible actions from a given state. In our case, we have 3 output units for each possible action (i.e. inspect, goto next, goto previous). Policy NN uses a softmax activation for output layer.

• Hidden Layers: Out network has four fully connected hidden layers with 128 units each using ReLU $[\phi(x) = \max(0, x)]$ activation.

5.2.3 Training Details and Hyperparameters

Training was done in batches with batch size of $N = 50$. We used a learning rate of 0.0001 with Adam\textsuperscript{5} optimizer. We used Glorot Uniform initialization for NN parameters. The algorithm for training using policy gradient is explained in Algorithm 1. The training cycle progresses as follows:

• Agents use their own simulators (environments) to collect $N$ episodes starting from a start state and ending at terminal state. While executing an episode, policy NN provides an action based on the learned policy for current state. The procedure COLLECTSAMPLES collects $N$ samples for each for an agent $i$, each containing an episode of horizon $T$.

• Agents train their value function neural network by computing the target based on observations, rewards and new observations empirically and using it to minimize the mean squared error loss.

• Agents train their policy network using the gradient of the global objective function as defined in (20) and (24) for at least one event in a constraint:

$$
\nabla_{\theta}GV(s_1; \pi) = \nabla_{\theta}V(s_1^i; \pi^i) + \sum_{k \in \rho(i)} c_k \nabla_{\theta}P(E_k) \prod_{j \in G_k \setminus \{i\}} (1 - P(E_j^i))
$$

Algorithm 1 explains the training algorithm used for training policy and VF neural networks as follows:

– Line 5 and 6 initializes neural networks for both value function (VF) and policy neural networks for all the agents. The two functions InitializeVFNetwork() and InitializePolicyNetwork() defines the NN layers as described above and initialize the parameters. This gives us parameters for both VF and policy NN as $\hat{\theta}_{agent}$ and

\( \theta_{agent} \) respectively. It also gives us functions \( VF_{agent}(obs) \) and \( \pi_{agent}(obs) \), as value function and policy function based on initial weights.

- Lines 10 - 12 stores \( N \) episodes for each agent in \( SAMPLES_{agent} \), containing observations \( (obs_t^j) \), new observations \( (newobs_t^j) \), rewards \( (R_t^j) \), and action information \( (a_t^j) \) for each episode \( j \in \{1..N\} \) and \( t \in \{0..T - 1\} \). In our case \( T \) is the total mission time (horizon).

- Lines 15 - 17 calculates a target value function based on observations \( obs_t^j \) in the samples as reward received \( R_t^j + VF_{agent}(newobs_t^j) \) for each episode \( j \in \{1..N\} \) and \( t \in \{0..T - 1\} \). Then parameters of VF neural network \( \hat{\theta}_{agent} \) are updated based on minimum squared loss between targeted values and current values.

- The gradient for local MDP can be taken from (Sutton et al. 1999) for agent \( i \) as follows:

\[
\nabla_{\theta_{i}} V^{i}(s_{i}^{1}; \pi_{i}) = \frac{1}{NT} \sum_{j=1}^{N} \sum_{t=0}^{T-1} \nabla_{\theta_{i}} \log \pi_{i}(a_{t}^{j}|s_{t}^{j})(R_{t}^{j} - VF_{i}(s_{t}^{j}))
\]

\[
= \frac{1}{NT} \nabla_{\theta_{i}} \sum_{j=1}^{N} \sum_{t=0}^{T-1} \log \pi_{i}(a_{t}^{j}|s_{t}^{j})(R_{t}^{j} - VF_{i}(s_{t}^{j}))
\]

(61)

\( s_{i}^{1} \) is the start state for agent \( i \). Now we simplify joint value gradient as follows:

\[
\nabla_{\theta_{i}} JV = \sum_{k \in \rho(i)} c_k \nabla_{\theta_{i}} P(E_{k}^{i}) \prod_{j \in G_{k} \setminus \{i\}} (1 - P(E_{k}^{j})))
\]

\[
\sum_{k \in \rho(i)} c_k \left( \sum_{e \in E_{k}^{i}} \nabla_{\theta_{i}} P(e; \pi_{i}) \right) \prod_{j \in G_{k} \setminus \{i\}} (1 - P(E_{k}^{j})))
\]

\[
\sum_{k \in \rho(i)} c_k \left( \sum_{e \in E_{k}^{i}} \frac{1}{|\xi|} \sum_{\Phi_{i} \in \xi: \Phi_{i} = e} \left( \sum_{t'=1}^{t(e,\Phi_i)} \nabla_{\theta_{i}} \log \pi_{i}(a_{t'}|s_{t'}) \right) \right) \prod_{j \in G_{k} \setminus \{i\}} (1 - P(E_{k}^{j})))
\]

(62)

- Line 20 computes (46) and Lines 21 - 26 computes (47). Finally line 27 updates the policy NN parameters \( \theta_{agent} \).
Algorithm 1 TIDec-MDPs Policy Gradient Algorithm

Input:
- $n \uparrow$ Number of agents
- $N \uparrow$ Batch Size
- $\alpha \uparrow$ Learning Rate
- $\text{iter} \uparrow$ Number of iterations

1: procedure COLLECTSAMPLES($\text{agent}$) \hspace{1cm} $\triangleright$ Collects $N$ episodes for $\text{agent}$
2: procedure TRAIN
3: $l \leftarrow 0$ \hspace{1cm} $\triangleright$ Iteration Counter
4: for $\text{agent} \in \{1, \ldots, n\}$ do \hspace{1cm} $\triangleright$ Initialize the Neural Networks
5: \hspace{1cm} $\pi_{\text{agent}}, \theta_{\text{agent}} \leftarrow \text{InitializePolicyNetwork}()$
6: \hspace{1cm} $\hat{\theta}_{\text{agent}} \leftarrow \text{InitializeVFNetwork}()$
7: while $l \neq \text{iter}$ do
8: \hspace{1cm} $l \leftarrow l + 1$
9: \hspace{1cm} $\text{SAMPLES} \leftarrow \text{None}$ \hspace{1cm} $\triangleright$ Holds samples for all agents in $l^{th}$ iteration
10: for $\text{agent} \in \{1, \ldots, n\}$ do \hspace{1cm} $\triangleright$ Collect samples from all agents (in parallel)
11: \hspace{1cm} $\text{SAMPLES}_{\text{agent}} \leftarrow \text{COLLECTSAMPLES}(\text{agent})$
12: \hspace{1cm} $\text{SAMPLES}_{\text{agent}} \leftarrow \text{Normalize(SAMPLES}_{\text{agent}})$ \hspace{1cm} $\triangleright$ Normalize rewards
13: for $\text{agent} \in \{1, \ldots, n\}$ do \hspace{1cm} $\triangleright$ Train VF and Policy NN (in parallel)
14: \hspace{1cm} $\Phi \leftarrow \text{SAMPLES}_{\text{agent}}$
15: \hspace{1cm} for episode $\in \Phi$ do
16: \hspace{2cm} $\text{targetVF}(\text{obs}^{\text{episode}}_{t}) \leftarrow R_{t}^{\text{episode}} + VF_{\text{agent}}(\text{newobs}^{\text{episode}}_{t})$
17: \hspace{2cm} objective$_{\text{vf}} \leftarrow \sum_{\text{episode} \in \Phi} \sum_{t=0}^{T-1} (\text{targetVF}(\text{obs}^{\text{episode}}_{t}) - VF_{\text{agent}}(\text{obs}^{\text{episode}}_{t}))^2$
18: \hspace{2cm} $\hat{\theta}_{\text{agent}} \leftarrow \hat{\theta}_{\text{agent}} - \alpha \nabla_{\hat{\theta}_{\text{agent}}}(\text{objective}_{\text{vf}})$
19: \hspace{2cm} $\nabla_{\theta_{\text{agent}}}VF_{\text{agent}}(s_{1}^{\text{agent}}) \leftarrow \frac{1}{|\Phi|} \nabla_{\theta_{\text{agent}}} \sum_{j=1}^{[\Phi]} \sum_{t=0}^{T-1} \log \pi_{\text{agent}}(a_{j}^{t}|s_{j}^{t})(R_{j}^{t} - VF_{\text{agent}}(s_{j}^{t}))$
20: \hspace{2cm} $\nabla_{\theta_{\text{agent}}}JV \leftarrow 0$
21: \hspace{2cm} for $k \in \rho(\text{agent})$ do \hspace{1cm} $\triangleright$ Using SAMPLES of all agents.
22: \hspace{2cm} product $\leftarrow \prod_{j \in G_{k} \setminus \{\text{agent}\}} (1 - P(E_{k}^{j}))$
23: \hspace{2cm} $\nabla_{\theta_{\text{agent}}}P(E_{k}^{\text{agent}}) \leftarrow \frac{1}{|\xi|} \nabla_{\theta_{\text{agent}}} \sum_{e \in E} \sum_{\Phi \in \xi: \Phi = e} \left( \sum_{t' = 1}^{t(e, \Phi)} \log \pi_{\text{agent}}(a_{t'}|s_{t'}) \right)$
24: \hspace{2cm} $\nabla_{\theta_{\text{agent}}}JV \leftarrow \nabla_{\theta_{\text{agent}}}JV + (c_{k} \nabla_{\theta_{\text{agent}}}P(E_{k}^{\text{agent}}) \times \text{product})$
25: \hspace{2cm} $\nabla_{\theta_{\text{agent}}}\text{objpolicy} \leftarrow \nabla_{\theta_{\text{agent}}}VF_{\text{agent}}(s_{1}^{\text{agent}}) + \nabla_{\theta_{\text{agent}}}JV$
26: \hspace{2cm} $\theta_{\text{agent}} \leftarrow \theta_{\text{agent}} - \alpha \nabla_{\theta_{\text{agent}}}(\text{objpolicy})$