Combining Propositional Logic Based Decision Diagrams with Decision Making in Urban Systems

Jiajing Ling,* Kushagra Chandak,* Akshat Kumar
School of Information Systems
Singapore Management University
{jjling.2018, kushagrac, akshatkumar}@smu.edu.sg

Abstract

Solving multiagent problems can be an uphill task due to uncertainty in the environment, partial observability, and scalability of the problem at hand. Especially in an urban setting, there are more challenges since we also need to maintain safety for all users while minimizing congestion of the agents as well as their travel times. To this end, we tackle the problem of multiagent pathfinding under uncertainty and partial observability where the agents are tasked to move from their starting points to ending points while also satisfying some constraints, e.g., low congestion, and model it as a multiagent reinforcement learning problem. We compile the domain constraints using propositional logic and integrate them with the RL algorithms to enable fast simulation for RL.

1 Introduction

The emergence and continued rise of autonomous and semi-autonomous vehicles in the urban landscape has made its way to a number of areas for transportation and mobility like self-driving cars and delivery trucks, railways, unmanned aerial vehicles, delivery drones fleet etc. Several key challenges remain to manage such agents like maintaining safety (no collisions among vehicles), avoiding congestion and minimizing travel time to better serve the users and reduce pollution. To model such scenarios, we leverage cooperative sequential multiagent decision making, where agents acting in a partially observable and uncertain environment are required to take coordinated decisions towards a long term goal (Durfee and Zilberstein 2013). Decentralized partially observable MDPs (Dec-POMDPs) provide a rich framework for multiagent planning (Bernstein et al. 2002; Oliehoek and Amato 2016), and are applicable in domains such as vehicle fleet optimization (Nguyen, Kumar, and Lau 2017), cooperative robotics (Amato et al. 2019), and multiplayer video games (Rashid et al. 2018). However, scalability remains a key challenge with even a 2-agent Dec-POMDP NEXP-Hard to solve optimally (Bernstein et al. 2002). To address the challenge of scalability, several frameworks have been introduced that model restricted class of interactions among agents such as transition independence (Becker et al. 2004; Nair et al. 2005), event driven and population-based interactions (Becker, Zilberstein, and Lesser 2004; Varakantham et al. 2012). Recently, several multiagent reinforcement learning (MARL) approaches are developed that push the scalability envelop (Lowe et al. 2017; Foerster et al. 2018; Rashid et al. 2018) by using simulation-driven optimization of agent policies.

Key limitations of several MARL approaches include sample inefficiency, and difficulty in learning when rewards are sparse, which is often the case in problems with combinatorial flavor. We address such a combinatorial problem of multiagent path finding (MAPF) under uncertainty and partial observability. Even the deterministic MAPF setting where multiple agents need to find collision-free paths from their respective sources to destinations in a shared environment is NP-Hard (Yu and LaValle 2013).

The MAPF problem is a general formulation that is capable of addressing several applications in the domain of urban mobility like autonomous vehicle fleet optimization (Ling, Gupta, and Kumar 2020; Sartoretti et al. 2019), taxiway path planning for aircrafts (Li et al. 2019), and train rescheduling (Nygren and Mohanty 2020). Figure 1 shows the airspace of a city divided into multiple geofenced airblocks. Such structured airspace can be used by drones to safely travel to their destinations (Ling, Gupta, and Kumar 2020). Since such spaces can have a lot of constraints, they can be modelled using our framework to manage the traf-
fic. Deep RL approaches have been applied to MAPF under uncertainty and partial observability (Sartoretti et al. 2019; Ling, Gupta, and Kumar 2020). A key challenge faced by RL algorithms is that it takes several simulations to find even a single route to destination as model-free RL does not explicitly exploit the underlying graph connectivity. Furthermore, agents can move in cycles, specially during initial training episodes, which makes the standard RL approaches highly sample inefficient. Recent approaches combine underlying graph structure with deep neural nets for combinatorial problems such as minimum vertex cover and traveling salesman problem (Dai et al. 2017; Bello et al. 2019). However, the knowledge compilation framework that we present provides much more explicit domain knowledge to RL approaches for MAPF.

To address the challenges of delayed rewards, and difficulty of finding feasible routes to destinations, we compile the graph over which agents move in MAPF using propositional logic based probabilistic sentential decision diagrams (psdd) (Kisa et al. 2014). A psdd represents probability distributions defined over the models of a given propositional theory. We use psdd to represent distribution over all simple paths (without loops) for a given source-destination pair. A key benefit is that any random sample from a psdd is guaranteed to be a valid simple path from the given source to destination. Furthermore, psdd are also equipped with associated inference methods (Shen, Choi, and Darwiche 2016a) (such as computing conditional probabilities) that significantly aid RL methods (e.g., given the current partial path, we are the possible next edges that are guaranteed to lead to the destination via a simple path). Using psdd significantly helps in pruning the search space, and generate high quality training samples for the underlying learning algorithm. However, integrating psdd with different RL methods is challenging, as the standard psdd inference methods are too slow to be used in the simulation-driven RL setting where one needs to query psdd at each time step. Therefore, we also develop highly efficient psdd inference methods that specifically aid RL by enabling fast sampling of training episodes, and are more than an order of magnitude faster than generic psdd inference. Given that number of paths between a source-destination can be exponential, we also use hierarchical decomposition of the graph to enable a tractable psdd representation (Choi, Shen, and Darwiche 2017a).

To summarize, our main contributions are as follows. First, we compile static domain information such as underlying graph connectivity using psdd for the MAPF problem under uncertainty and partial observability. Second, we develop techniques to integrate such decision diagrams within diverse deep RL algorithms based on policy gradient and Q-learning. Third, we develop fast algorithms to query compiled decision diagrams to enable fast simulation for MARL. We integrate our psdd-based framework with previous MARL approaches (Sartoretti et al. 2019; Ling, Gupta, and Kumar 2020), and show that the resulting algorithms significantly outperform the original algorithms both in terms of sample complexity and solution quality on a number of instances. We also highlight that psdd is a general framework for incorporating constraints in decision making, and discuss extensions of the standard MAPF that can be addressed using psdd.

2 The Dec-POMDP Model and MAPF

A Dec-POMDP is defined using the tuple \((S, A, T, O, Z, r, n, \gamma)\). There are \(n\) agents in environment (indexed using \(i = 1 : n\)). The environment can be in one of the states \(s \in S\). At each time step, agent \(i\) chooses an action \(a^i \in A\), resulting in the joint action \(a \in A^i \equiv A^n\). As a result of the joint action, the environment transitions to a new state \(s'\) with probability \(T(s, a, s')\). The joint-reward to the agent team is given as \(r(s, a)\). The reward discount factor is \(\gamma < 1\).

We assume a partially observable setting in which agent \(i\)’s observation \(z^i \in Z\) is generated using the observation function \(O(a, s', z^i) = P(z^i|a, s')\) where the last joint action taken was \(a\) and the resulting state was \(s'\) (for simplicity, we have assumed the observation function is the same for all agents). As a result, different agents can receive different observations from the environment.

An agent’s policy is a mapping from its action-observation history \(t^i \in (Z \times A)^*\) to actions or \(\pi^i(a^i|t^i; \theta^i)\), where \(\theta^i\) parameterizes the policy. Let the discounted future return be denoted by \(R_t = \sum_{k=0}^{\infty} \gamma^k r_{k+t}\). The joint-value function induced by the joint-policy of all the agents is denoted as \(V^\pi(s_t) = \mathbb{E}_{a_{t+1}\ldots a_{t+n} \sim \pi}[R_t|s_t, a_t]\), and joint action-value function as \(Q^\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}\ldots s_{t+n} \sim \pi}[R_t|s_t, a_t]\). The goal is to find the best joint-policy \(\pi\) to maximize the value for the starting belief \(b_0\) \(V^\pi(\pi) = \sum_s b_0(s)V^\pi(s)\).

Learning from simulation: In the RL setting, we do not have access to transition and observation functions \(T, O\). Instead, multiagent RL approaches (MARL) learn via interacting with the environment simulator. The simulator, given the joint-action input \(a_t\) at time \(t\), provides the next environment state \(s_{t+1}\), generates observation \(z_{t+1}^i\) for each agent, and provides the reward signal \(r_t\). Similar to several previous MARL approaches, we assume a centralized learning and decentralized policy execution (Foerster et al. 2018; Lowe et al. 2017). During centralized training, we assume access to extra information (such as environment state, actions of different agents) that help in learning value functions \(V^\pi, Q^\pi\). However, during policy execution, agents rely on their local action-observation history. An agent’s policy \(\pi^i\) is typically implemented using recurrent neural nets to condition on action-observation history (Hausknecht and Stone 2015). However, our developed results are not affected by a particular implementation of agent policies.

MARL for MAPF: MAPF can be mapped to a Dec-POMDP instance in multiple ways to address different variants (Ma, Kumar, and Koenig 2017; Sartoretti et al. 2019; Ling, Gupta, and Kumar 2020). We therefore present the MAPF problem under uncertainty and partial observability using minimal assumptions to ensure the generality of our knowledge compilation framework. There is a graph \(G = (V, E)\) where the set \(V\) denotes the locations where agents can move, and edges connect different locations. An agent \(i\) has a start vertex \(s_i\) and final goal vertex \(g_i\). At any time step, an agent can be located at a vertex \(v \in V\), or
in-transit on an edge \((u, v)\) (i.e., moving from vertex \(u\) to \(v\)).

An agent’s action set is denoted by \(A = A_{\text{mov}} \cup A_{\text{st}}\). Intuitively, \(A_{\text{mov}}\) denotes actions that intend to change the location of agent from the current vertex to a neighboring directly connected vertex in the graph (e.g., move up, right, down, left in a grid graph). The set \(A_{\text{st}}\) denotes other actions that do not intend to change the location of the agent (e.g., \(\text{noop}\) that intends to make agent stay at the current vertex).

Note that we do not make any assumptions regarding the actual transition after taking the action (i.e., move/stay actions may succeed or fail as per the specific MAPF instance).

Depending on the states of all the agents, an agent \(i\) receives observation \(z_i\). We assume that an agent is able to fully observe its current location (i.e., the vertex it is currently located at). Other information can also be part of the observation (e.g., location of agents in the local neighborhood of the agent), but we make no assumptions about such information. We make no specific assumptions about the joint-reward \(r\), other than assuming that an agent prefers to reach its destination as fast as possible if the agent’s movement does not conflict with other agents’ movements. Typical examples of reward \(r\) include penalty for every time step an agent is not at its goal vertex, positive reward at the goal vertex, a high penalty for creating congestion at vertices or edges of the graph (Ling, Gupta, and Kumar 2020), or for blocking other agents from moving to their destination (Sartoretti et al. 2019).

### 3 Incorporating Compiled Knowledge in RL

A key challenge for RL algorithms for MAPF is that often finding feasible paths to destinations require a large number of samples. For example, figure 2(a) shows the case when an agent loops back to one of its earlier vertex. Figure 2(b) shows another scenario where an agent moves towards a deadend. Such scenarios increase the training episode length in RL. Our key intuition is to develop techniques that ensure that RL approaches only sample paths that are (i) simple, (ii) always originate at the source vertex \(s_i\) and end at the goal vertex \(g_i\) for any agent \(i\).

Let \(p_i^t\) denote the path taken by an agent \(i\) until time \(t\) (or the sequence of vertices visited by an agent starting from source \(s_i\)). We also assume that it does not contain any cycle. This cycle can be extracted from agent’s history \(\tau_i^t\). Let \(a \in A_{\text{mov}}\) be a movement action towards vertex \(a_v\). We assume the existence of a function \(\text{feasibleActions}(p_i^t; s_i, d_i)\) that takes as input an agent’s current path \(p_i\) and returns the set \(\text{nextActions} = \{a \in A_{\text{mov}} \mid \text{s.t. } [p_i, a_v] \rightarrow d_i\}\). The condition \([p_i, a_v] \rightarrow d_i\) implies there exists at least one simple path from source \(s_i\) to destination \(d_i\), that includes the path segment \([p_i, a_v]\). Thus, starting with \(p_0 = [s_i]\), the RL approach would only sample simple paths that are guaranteed to reach an agent’s destination, thereby significantly pruning the search space, and resulting in trajectories that have good potential to generate high rewards. The information required for implementing \(\text{feasibleActions}\) can be compiled offline even before training and execution of policy starts (explained in next section, using decision diagrams), and does not include any communication overhead during policy execution. Using this abstraction, we next present simple and easy-to-implement modifications to a variety of deep multiagent RL algorithms.

**Policy gradient based MARL:** We first provide a brief background of policy gradient approaches for single agent case (Sutton et al. 2000). An agent’s policy \(\pi^\theta\) is parameterized using \(\theta\). The policy is optimized using gradient ascent on the total expected reward \(V(\theta) = E_{\pi^\theta}[R_0]\). The gradient is given as:

\[
\nabla_\theta V(\theta) = E_{s_0, a_0, \ldots} \left[ \sum_{t=0}^{\infty} R_t \nabla_\theta \log \pi^\theta(a_t|s_t) \right]
\]

(1)

Above gradient expression is also extendible to the multiagent case in an analogous manner (Peshkin et al. 2000; Forerster et al. 2018). In multiagent setting, we can compute gradient of the joint-value function \(V\) w.r.t. an agent’s policy \(\pi_i\)’s policy parameters \(\theta^i\) or \(\nabla_i V\). The expectation is w.r.t. the joint state-action trajectories \(E_{s_0, a_0, \ldots}\), and \(R_t\) denotes future return for the agent team. The input to policy are some features of the agent’s observation history or \(\phi(\tau^i)\). The function \(\phi\) can be either hard-coded (e.g., only last two observations), or can be learned using recurrent neural networks.

For using compiled knowledge using the function \(\text{feasibleActions}\), the only change we require is in the structure of an agent’s policy \(\pi\) (we omit superscript \(i\) for brevity). The main challenge is addressing the variable sized output of the policy in a differentiable fashion. Assuming a deep neural net based policy \(\pi\), given the discrete action space \(A\), the last layer of the policy has \(|A|\) outputs using the softmax layer (to normalize action probabilities \(\pi(a|\cdot)\)). However, when using \(\text{feasibleActions}\), the probability of actions not in \(\text{feasibleActions}\) needs to be zero. However, the set \(\text{feasibleActions}\) changes as the observation history \(\tau\) of the agent is updated. Therefore, a fixed sized output layer appears to create difficulties. However, we propose an easy fix. We use \(\tilde{\pi}\) to denote the standard way policy \(\pi\) is constructed with last layer having fixed \(|A|\) outputs. However, we do not require the last layer to be a softmax layer. Instead, we re-define the policy \(\pi\) as:

\[
\pi(a|\tau) = \begin{cases} 
0 & \text{if } a \notin \text{feasibleActions}(p(\tau); s, d) \\
\exp \left( \phi(a|\phi(\tau)) \right) & \sum_{a' \in \text{feasibleActions}(p(\tau); s, d)} \exp \left( \phi(a'|\phi(\tau)) \right)
\end{cases}
\]

(2)

where \(p(\tau)\) denotes the path taken by the agent so far, and \(s, d\) are its source and destination. Sampling from \(\pi\) guarantees that invalid actions are not sampled. Furthermore, \(\pi\)
is differentiable even when feasibleActions gives different length outputs at different time steps. The above operation can be easily implemented in autodiff libraries such as Tensorflow without requiring a major change in the policy structure \( \pi \).

**Q-learning based MARL:** Deep Q-learning for the single agent case (Volodymyr et al. 2015) has been extended to the multiagent setting also (Rashid et al. 2018). In the QMIX approach (Rashid et al. 2018), the joint action-value function \( Q_{\text{tot}}(\tau; a; \psi) \) is factorized as (non-linear) combination of action-value functions \( Q_i(\tau^i; a^i; \theta^i) \) of each agent \( i \). A key operation when training different parameters \( \theta^i \) and \( \psi \) involves maximizing \( \max_{a \in \mathcal{A}} Q_{\text{tot}}(\tau; a; \psi) \) (for details we refer to Rashid et al.). This operation is intractable in general, however, under certain conditions, it can be approximated by maximizing individual Q functions \( \max_{a \in \mathcal{A}} Q_i(\tau^i; a^i; \theta^i) \) in QMIX. We require two simple changes to incorporate our knowledge compilation scheme in QMIX. First, instead of maximizing over all the actions, we maximize only over feasible actions of an agent as \( \max_{a \in \text{feasibleActions}} \max_{a \in \mathcal{A}} Q_i(\tau^i; a^i; \theta^i) \). Second, in Q-learning, typically a replay buffer is also used which stores samples from the environment as \((\tau, a, \tau', r)\). In our case, we also store additionally the set of feasible actions for the next observation history \( \tau'^i \) for each agent \( i \) as feasibleActions\((p(\tau'^i); s^i, d^i)\) along with the tuple \((\tau, a, \tau', r)\). The reason is when this tuple is replayed, we have to maximize \( Q'(\tau'^i, a) \) over \( a \in \text{feasibleActions}(p(\tau'^i); s^i, d^i) \), and storing the set feasibleActions\((p(\tau'^i); s^i, d^i)\) would reduce computation.

We have integrated our knowledge compilation framework with two policy gradient approaches proposed in (Sartoretti et al. 2019; Ling, Gupta, and Kumar 2020) (one using feedforward neural net, another using recurrent neural network based policy), and a QMIX-variant (Fu et al. 2019) for MAPF, demonstrating the generalization power of the framework for a range of MARL solution methods.

### 4 Compiling and Querying Decision Diagrams for MAPF

We now present our decision diagram based approach to implement the feasibleActions function. Let upper case letters \((X)\) denote variables and lowercase letters \((x)\) denote their instantiations. Bold upper case letter \((\mathbf{X})\) denotes a set of variables and their lower case counterparts \((\mathbf{x})\) denote the instantiations.

**Paths as a Boolean formula:** A path \( p \) from a given source \( s \) to the destination \( d \) in the underlying undirected graph \( G = (V, E) \) can be represented as a Boolean formula as follows. Consider Boolean random variables \( X_{i,j} \) for each edge \((i, j) \in E\). If an edge \((i, j)\) occurs in \( p \), then \( X_{i,j} \) is set to true, otherwise it’s set to false. Hence, conjunction of these literals denotes path \( p \), and the Boolean formula representing all paths is obtained by simply disjoining formulas for all such paths (Choi, Tuvabi, and Darwiche 2016). An example path in a graph is given in fig 3(a).

**Sentential decision diagrams:** Since the number of paths between two nodes can be exponential, we need a compact representation of the Boolean formula representing paths. To this end, we use sentential decision diagram or sdd (Darwiche 2011). It is a Boolean function \( f(X, Y) \) on some non-overlapping variable sets \( X, Y \) and is written as a decomposition in terms of functions on \( X \) and \( Y \). In particular, \( f = (p_1(X) \land s_1(Y)) \lor \ldots \lor (p_n(X) \land s_n(Y)) \), with each element \((p_i, s_i)\) of the decomposition composed of a prime \( p_i \) and a sub \( s_i \). An sdd represented as a decision diagram describes members of a combinatorial space (e.g., paths in a graph) using propositional logic in a tractable manner. It has two kinds of nodes:

- **Terminal node**, which can be a literal \((X\) or \(\neg X\)), always true \((\top)\) or always false \((\bot)\), and

- **Decision node**, which is represented as \((p_1 \land s_1) \lor \ldots \lor (p_n \land s_n)\) where all \((p_i, s_i)\) pairs are recursively sdds and the primes are always consistent, mutually exclusive and exhaustive.

Figure 3(b) represents an sdd for the graph in fig 3(a) encoding all paths from \( n_1 \) to \( n_5 \). The encircled node is a decision node with two elements \((D, E)\) and \((\neg D, \bot)\). The primes are \( D \) and \( \neg D \) and the subs are \( E \) and \( \bot \). The Boolean formula representing this sdd node is \((D \land E) \lor (\neg D \land \bot)\) which is equivalent to \( D \lor E \). The Boolean formula encoded by the whole sdd is given by the root node of the sdd.

An sdd is characterized by a full binary tree, called a vtree, which induces a total order on the variables from a left-right traversal of the vtree. E.g., for the vtree in figure 3(c), the variable order is \((A, B, C, D, E)\). Given a fixed vtree, the sdd is unique. An sdd node \( n \) is normalized (or associated with) for a vtree node \( v \) as follows:

- If \( n \) is a terminal node, then \( v \) is a leaf vtree node which contains the variable of \( n \) (if any).
- If \( n \) is a decision node, then \( n \)’s primes (subs) are normalized for the (right) child of \( v \).
- If \( n \) is the root node, then \( v \) is the root vtree node.

Intuitively, a decision node \( n \) being normalized for vtree node \( v \) implies that the Boolean formula encoded by \( n \) contains only those variables contained in the sub-tree rooted at \( v \). We will use this normalization property for our analysis later. The Boolean formula encoding the domain knowledge can be compiled into a decision diagram using the sdd compiler (Oztok and Darwiche 2015). The resulting sdd may not be exponential in size even though it is representing an exponential number of objects.

**Probabilistic Sentential decision diagrams:** In our case, for computing feasibleActions, we also need to associate a probability distribution with the sdd that encodes all the paths from a given source to destination. The key benefit is that we can exploit associated inference methods such as computing conditional probabilities, which will help in computing feasibleActions.

If we parameterize each of the decision nodes of the sdd, such that the local parameters form a distribution, the resulting probabilistic structure is called a psdd or a probabilistic sdd (Kisa et al. 2014). It can be used to represent discrete
probability distributions $Pr(X)$ where several instantiations $x$ have zero probability $Pr(x) = 0$ because of the constraints imposed on the space. More concretely, a psdd normalized for an sdd is defined as follows:
- For each decision node $(p_1, s_1), \ldots, (p_n, s_n)$, there’s a positive parameter $\theta_i$ such that $\sum_{i=1}^{n} \theta_i = 1$ and $\theta_i = 0$ iff $s_i = \bot$.
- For each terminal node $T$, there’s a parameter $0 < \theta < 1$.

psdds are tractable models of probability distributions as several probabilistic queries can be performed in poly-time such as computing marginal probabilities, or conditional probabilities.

**NZ (Non-Zero) Inference for feasibleActions:** Given an sdd encoding all simple paths from a source $s$ to a destination $d$, we uniformly parameterize this sdd as noted earlier. That is, for a decision node $(p_1, s_1), \ldots, (p_n, s_n)$, each $\theta_i$ is the same (except when $s_i = \bot$, then $\theta_i = 0$). And we also enforce that non-zero $\theta_i$s normalize to 1. This strategy makes sure that the probability of each simple path from $s$ to $d$ is non-zero. Assume that the current sampled path by the agent is $p$ (in the context of psdd, we assume that $p$ is a set of edges in graph $G$ traversed from source $s$ by the agent). Let $v_p$ denote the current vertex of the agent (and assume $v_p$ is not the destination). Let $\text{Nb}(v_p)$ denote all direct neighbors of $v_p$. The feasibleActions set is given as:

$$\text{feasibleActions}(p) = \{v' \in \text{Nb}(v_p) \land (v_p, v') \notin p \land Pr((v_p, v') | p) > 0\} \quad (3)$$

That is, if the conditional probability $Pr((v_p, v') | p) = 0$, then $v'$ can be pruned from the action set as it implies there is no simple path to destination $d$ that takes the edge $(v_p, v')$ after taking the path $p$. This strategy seems straightforward to implement as psdd is equipped with inference methods to compute conditional probabilities. However, in RL, this inference needs to be done at each time step for each training episode. We observed empirically that this method was extremely slow, and it was impractical to scale it for multiple agents. We therefore next develop our customized inference technique that is much faster than this generic inference.

**Sub-context connectivity analysis for NZ Inference:** We note that all the discussion below is for a psdd that encodes all simple paths from a source $s$ to destination $d$, and the psdd is normalized for some right linear vtree. Proofs for different results are provided in the supplementary material in the full paper available on Arxiv.

**Lemma 1.** In a psdd normalized for a right linear vtree, each prime is a literal $X$ or $\neg X$ or $\top$.

The above result is a direct consequence of the manner in which the underlying sdd is constructed using a right linear vtree.

We sample a path from such a psdd by traversing it in a top-down fashion and selecting one branch at a time for each of the decision nodes according to the probability for that branch and then selecting the prime and recursively going down the sub (Kisa et al. 2014). As all the prime nodes are terminal as per lemma 1, if the prime node is a positive literal $X$, then we select the edge $e$ corresponding to $X$ for our path (say $e_X$). If prime node is $\neg X$, then we do not select edge $e_X$. We show in the supplement that the prime nodes encountered during such sampling procedure for a psdd that encodes simple paths cannot be $\top$.

As an example, consider the graph in fig 3(a) and its corresponding psdd in fig 3(d). We start at the root of the psdd and select the left branch with probability 1. We then select the prime $A$ in our sample and recursively go down its sub as shown by the red arrows. The final sampled path is $A \land \neg B \land C \land \neg D \land E$.

**Definition 1.** (S-Path) Let $n$ be a psdd node normalized either for $\hat{v}_l$ and $\hat{v}_r$, the two deepest vtree nodes. Let $(p_1, s_1), \ldots, (p_k, s_k)$ be the elements appearing on some path from the psdd root to node $n$ (i.e., $n = p_k$ or $n = s_k$). Then $p_1 \land \ldots \land p_k \land n$ is called an s-path for node $n$, and is feasible iff $s_k \neq \bot$.

In figure 3(c), $\hat{v}_l = D$, and $\hat{v}_r = E$. There can be multiple s-paths for a node $n$. Let $\text{spset}$ denote the set of all feasible s-paths for all psdd nodes $n$ normalized either for $\hat{v}_l$ or $\hat{v}_r$.
Lemma 2. There is a one-to-one mapping between s-paths in the set \( \text{spset} \) and the set of all simple paths in \( G \) from source \( s \) to destination \( d \).

The above lemma states that if we find a feasible s-path in the \( \text{psdd} \), then it would correspond to a valid simple path from source \( s \) to destination \( d \) in the graph \( G \) which will also have nonzero probability as per our \( \text{psdd} \). Reading off the path in \( G \) given a feasible s-path is straightforward. A feasible s-path is also a conjunction of literals (using lemma 1, and if \( n \) is a sub, it will also be a literal as \( n \) is normalized for deepest node in vtree). For each positive literal \( X \) in s-path, we include its corresponding edge \( e_X \) in the path in \( G \). The set of resulting edges would form a simple path in \( G \).

This result also provides a strategy for our fast \( \text{NZ} \) inference. Given a path \( p \) in graph \( G \), our goal is to find whether \( Pr((v_p, v')|p) > 0 \). If we can prove that there exists an s-path \( sp \in \text{spset} \) such that its corresponding path in graph \( G \) (using lemma 2) contains all the edges in \( p \) and \((v_p, v')\), then \( Pr((v_p, v')|p) \) must be nonzero. We need few additional results below to turn this insight into an efficient algorithmic procedure.

Definition 2. (Sub-context (Kisa et al. 2014)) Let \((p_1, s_1), \ldots, (p_k, s_k)\) be the elements appearing on some path from the sdd root to node \( n \) (i.e., \( n = p_k \) or \( n = s_k \)). Then \( p_1 \wedge \ldots \wedge p_k \) is called a sub-context \( sc \) for node \( n \), and is feasible iff \( s_i \neq \bot \).

Notice that a \( \text{psdd} \) node \( n \) can have multiple (feasible) sub-context as a \( \text{psdd} \) is a directed acyclic graph (DAG). Essentially, each sub-context corresponds to one possible way of reaching node \( n \) from the \( \text{psdd} \) root. For a right linear vtree, a feasible sub-context is a conjunction of literals as all primes are literals (lemma 1).

Given two \( \text{psdd} \) nodes \( n \) and \( n' \), we say that \( n' \) is deeper than \( n \) if the vtree node \( v' \) for which \( n' \) is normalized is deeper than vtree node \( v \) for which \( n \) is normalized.

Definition 3. (Sub-context set) Let \( X \) be a positive literal, and let \( p_1, \ldots, p_k \) be \( \text{psdd} \) prime nodes such that each \( p_i = X \). Let \( ssc_1, \ldots, ssc_k \) be sets such that each \( ssc_j \) contains all the feasible sub-contexts of \( p_j \). Then the sub-context set of \( X \) denoted by \( \text{sset}(X) \) is defined as \( \text{sset}(X) = \cup_{j=1}^k \text{ssc}_j \).

We now show the procedure to perform sub-context connectivity analysis for \( \text{NZ} \) inference. Assume that the current sampled path in graph \( G \) is \( p = \{e_1, \ldots, e_k\} \) (each \( e_i \) is the edge traversed by the agent so far). Let the current vertex of the agent be \( v_p \). Let \( e = (v_p, v') \) be one possible edge in \( G \) that the agent can traverse next. Let \( X_{e_1}, \ldots, X_{e_k}, X_e \) be the respective Boolean variables for the different edges. We wish to determine whether \( P(X_e|X_{e_1}, \ldots, X_{e_k}) \) is greater than zero. We follow the following steps to determine this.

1. Find the variable \( \hat{X} \in \{X_{e_1}, \ldots, X_{e_k}, X_e\} \) that is deepest in the vtree order.
2. Check if there exists a sub-context \( sc \in \text{sset}(\hat{X}) \) such that \( sc \) contains all the positive literals \( \{X_{e_1}, \ldots, X_{e_k}, X_e\} \).
3. Since \( sc^* \) is the sub-context of the variable deepest in the vtree order among \( \{X_{e_1}, \ldots, X_{e_k}, X_e\} \), it can be extended to a feasible s-path \( sp \in \text{spset} \) such that \( sp \) contains \( sc^* \) (or \( sc^* \wedge sp = sp \)). (Proved formally in supplementary).

Hierarchical clustering for large graphs: For increasing the scalability of the \( \text{psdd} \) framework and \( \text{NZ} \) inference for large graphs, we take motivation from (Choi, Shen, and Darwiche 2017b; Shen et al. 2019). These previous results show that by suitably partitioning the graph \( G \) among clusters, we can keep the size of the \( \text{psdd} \) tractable even for very large graphs. Such partitioning does result in the loss of expressiveness as the \( \text{psdd} \) for the partitioned graph may omit some simple paths, but empirically, we found that this partitioning scheme still improved efficiency of the underlying RL algorithms significantly. This partitioning method is described in the supplementary material in more detail.

5 Extensions and Modeling Other Logical Constraints

The framework that we presented can be used to compile a number of different kinds of constraints. For example, the agent has to first go to a pickup location and then to a delivery location (Liu et al. 2019), or TSP-like constraints where the agent has to visit some locations before reaching the destination while avoiding collisions. An example is explained below in more detail.

Landmark Constraints: This framework can be extended to settings where an agent is required to visit some landmarks before reaching the destination. We can construct the Boolean formula representing such a constraint by taking incident edge variables for each of the landmarks and allowing at least one of them to be true. We can then multiply (Shen, Choi, and Darwiche 2016b) the \( \text{PSDD} \) representing such a formula with the \( \text{PSDD} \) representing simple path constraint. For example, if \( n_i \) is a node representing a landmark and \( A, B, C, D \) are Boolean variables representing the edges incident on \( n_i \), then we can represent the constraint for \( n_i \) as \( \beta_i = A \lor B \lor C \lor D \). For \( k \) such landmarks, we can similarly represent the constraints \( \beta_1, \ldots, \beta_k \). Then the Boolean formula for all the landmarks would be \( \beta = \bigwedge_{i=1}^k \beta_i \) and can be compiled as a \( \text{PSDD} \). Now, if \( \alpha \) is a \( \text{PSDD} \) representing simple paths between a source and a destination, then we can multiply \( \alpha \) and \( \beta \) to get the final \( \text{PSDD} \) representing simple paths where an agent is required to visit some land-
marks before the destination. This strategy can be scaled up by hierarchical partitioning of the graph (Choi, Shen, and Darwiche 2017b; Shen et al. 2019) and can be used to represent complex constraints by multiplying them. This process is also modular since the constraints are separately modeled from the underlying graph connectivity.

Furthermore, this framework can also be used in cases where the underlying graph connectivity is dynamic; e.g., in scenarios where edges are dynamically getting blocked over time or the graph is revealed with time like the Canadian Traveller Problem (Liao and Huang 2014). Any observation about blocked edges at a time can become the evidence, and by conditioning on this evidence, the agent can rule out routes via such blocked edges. The generalizability and flexibility of this framework make it a promising approach in combining domain knowledge with models for RL, pathfinding, and other areas.

## 6 Empirical Evaluation

We present results to show how the integration of our framework with previous multiagent deep-RL approaches based on policy gradient and Q-learning (Sartoretti et al. 2019; Ling, Gupta, and Kumar 2020) performs better in MAPF problems in terms of both sample efficiency and solution quality on a number different maps with different number of agents.

### Simulation Speed

We show comparisons between our method and psdd inference method for calculating marginal probabilities. Our approach is more than an order of magnitude faster.

| Approach          | No Clustering | Clustering |
|-------------------|---------------|------------|
| source            | 3x3 | 4x4 | 5x5 | 10x10 |
| SCANZ             | 1.84 | 3.86 | 19.82 | 407.95 |
| psdd inference    | 26.55 | 158.41 | 979.71 | 402665.98 |

Table 1: Simulation speed comparison (in seconds)

### Open Grid Maps

We next evaluate the integration of our knowledge-based framework with policy gradient and Q-learning based approaches. We combine our framework with DCRL (Ling, Gupta, and Kumar 2020) and MAPQN (Fu et al. 2019) on several open grid maps with varying number of agents. DCRL is a policy gradient based algorithm, and MAPQN is a Q-learning based algorithm. We follow the same MAPF model as (Ling, Gupta, and Kumar 2020) where each node has its own capacity (maximum number of agents that can be accommodated), and agents can take multiple time steps to move between two contiguous nodes. The total objective is to minimize sum of costs (SOC) of all agents combined with penalties for congestion. More details on the experiments, the neural network structure and the hyperparameters are noted in the supplementary material.

The environment setting is varying from 4x4, 2 agents up to 10x10, 30 agents. We generated 10 instances for each setting. In each setting, we follow (Ling, Gupta, and Kumar 2020) to randomly select the sources and destinations and to specify the capacity of each node. We also specify the min and max time ($t_{min}$, $t_{max}$) to move between two nodes. We run for each instance three times, and we terminate the runs either after 500 iterations or 10 hours. Each episode has a maximum length of 500 steps. For each instance, we choose the run with the best performance. We compute the total objective averaged over all agents and the cumulative number of samples averaged over all agents during training. Finally, we plot the average total objective vs the average cumulative sample count over all instances.

Figure 4 shows the results comparing DCRL with Knowledge Compilation (DCRL+KCO) and DCRL on 4x4, 8x8, and 10x10 grids (plots for 4x4, 2 agents, 8x8, 6 agents, and 10x10, 10 agents are deferred to the supplementary). Although all agents are able to reach their respective destinations (no stranded agents) in both DCRL+KCO and DCRL, agents are trained to reach destinations cooperatively with significantly fewer samples in DCRL+KCO. It means that agents are exploring the environment more efficiently in DCRL+KCO than in DCRL especially during the initial few training episodes. This is also reflected in the plot as the average total objective in DCRL+KCO is significantly higher during initial training phase compared to DCRL.

Figure 5 shows the comparison of sample efficiency between MAPQN+KCO and MAPQN on 4x4 and 8x8 grids. We did not evaluate MAPQN+KCO on 10x10 grid since MAPQN itself is not able to train a large number of agents on large grid maps (more details in (Ling, Gupta, and Kumar 2020)). We observe that MAPQN+KCO converges faster and to a better quality than MAPQN especially on 8x8 grid. This is because several agents did not reach their destination within the episode cutoff in MAPQN, in contrast, all agents reach their destination in MAPQN+KCO.

### Obstacles

We evaluate KCO with DCRL and MAPQN on a 10x10 obstacle map with varying number of agents (from 2 agents up to 10 agents). The obstacles are randomly generated with density 0.35. We generate 10 instances for this setting. For each instance, sources and destinations are randomly generated from the non-blocked nodes from the top and bottom rows (each source and destination pair is guaranteed to be reachable). Other parameters are specified in the same way as the above experiments. This set of experiments is quite challenging especially when there are several agents since they can go into dead ends easily while cooperating with each other to reduce the congestion level. Figure 6 clearly shows that DCRL and MAPQN can converge much faster with the integration of KCO and confirms that our approach is more sample efficient. Specifically for MAPQN, several agents did not reach their destinations (8.8 agents on average, for N10 case), whereas in MAPQN+KCO, all agents reached destination, which explains much better solution quality by MAPQN+KCO.

We also evaluate KCO integrated with the PRIMAL framework (Sartoretti et al. 2019) which is based on asynchronous advantage actor-critic or A3C (Mnih et al. 2016) combined with imitation learning. We test it on a 10x10 map with obstacles, keeping the density high (0.35). We generated 10 instances and tested with 2 and 4 agents. As noted in (Sartoretti et al. 2019), high obstacle density is particularly problematic for PRIMAL. Our results in Figure 7 show...
that PRIMAL+KCO clearly outperforms PRIMAL in terms of sample efficiency. With 2 agents, the average SOC in PRIMAL is fluctuating around 250 during the initial episodes (maximum episode length is 256). However, the average SOC in PRIMAL+KCO is quite low during the initial training phase as expected (lower is better). With 4 agents, although the average SOC is quite high in both PRIMAL and PRIMAL+KCO during the initial episodes, the average SOC by PRIMAL+KCO is still lower than that by PRIMAL. The reason for the initial high average SOC is that the agents are trying to avoid collisions by taking a lot of no-op actions because of the high density. Overall, our framework is flexible enough to be integrated with different MARL approaches and consistently improve the performance.

7 Conclusion

We addressed the problem of cooperative multiagent pathfinding under uncertainty. Our work compiled static domain information such as underlying graph connectivity using propositional logic based decision making diagrams. We developed techniques to integrate such diagrams with deep RL algorithms such as Q-learning and policy gradient. Furthermore, to make simulation faster for RL, we developed an algorithm by analyzing the sub-context connectivity. We showed that the simulation speed of our algorithm is faster than the generic psdld method. We demonstrated the effectiveness of our approach both in terms of sample efficiency and solution quality on a number of instances.

Acknowledgments

This research is supported by the Ministry of Education, Singapore under its Academic Research Fund Tier 2 (MOE2018-T2-1-179).

References

Amato, C.; Konidaris, G.; Kaelbling, L. P.; and How, J. P. 2019. Modeling and planning with macro-actions in decentralized POMDPs. Journal of Artificial Intelligence Research 64: 817–859.

Becker, R.; Zilberstein, S.; and Lesser, V. 2004. Decentralized Markov decision processes with event-driven interactions. In International Joint Conference on Autonomous Agents and Multiagent Systems, 302–309. ISBN 1581138644.

Becker, R.; Zilberstein, S.; Lesser, V.; and Goldman, C. V. 2004. Solving transition independent decentralized Markov decision processes. Journal of Artificial Intelligence Research 22: 423–455. ISSN 10769757.

Bello, I.; Pham, H.; Le, Q. V.; Norouzi, M.; and Bengio, S. 2019. Neural combinatorial optimization with reinforcement learning. In International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings.

Bernstein, D. S.; Givan, R.; Immelman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of Markov decision processes. Mathematics of Operations Research 27(4): 819–840.

Choi, A.; Shen, Y.; and Darwiche, A. 2017a. Tractability in structured probability spaces. In Advances in Neural Information Processing Systems, 3478–3486.
Choi, A.; Shen, Y.; and Darwiche, A. 2017b. Tractability in structured probability spaces. In Advances in Neural Information Processing Systems, 3477–3485.

Choi, A.; Tavabi, N.; and Darwiche, A. 2016. Structured Features in Naive Bayes Classification. In AAAI.

Dai, H.; Khalil, E. B.; Zhang, Y.; Dilkina, B.; and Song, L. 2017. Learning combinatorial optimization algorithms over graphs. In Advances in Neural Information Processing Systems, 6349–6359.

Darwiche, A. 2011. SDD: A new canonical representation of propositional knowledge bases. In Twenty-Second International Joint Conference on Artificial Intelligence.

Durfee, E.; and Zilberstein, S. 2013. Multiagent Planning, control, and execution. In Weiss, G., ed., Multiagent Systems, chapter 11, 485–546. Cambridge, MA, USA: MIT Press. ISBN 978-0-262-01889-0.

Foerster, J. N.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual multi-agent policy gradients. In AAAI Conference on Artificial Intelligence, 2974–2982.

Fu, H.; Tang, H.; Hao, J.; Lei, Z.; Chen, Y.; and Fan, C. 2019. Deep multi-agent reinforcement learning with discrete-continuous hybrid action spaces. In International Joint Conference on Artificial Intelligence, 2329–2335.

Hausknecht, M.; and Stone, P. 2015. Deep recurrent Q-learning for partially observable MDPs. In AAAI Fall Symposium - Technical Report, 29–37.

Hio, L. 2016. Traffic System for Drones. URL https://www.straitstimes.com/singapore/traffic-system-for-drones.

Kisa, D.; Van Den Broeck, G.; Choi, A.; and Darwiche, A. 2014. Probabilistic sentential decision diagrams. In Principles of Knowledge Representation and Reasoning, 558–567.

Li, J.; Zhang, H.; Gong, M.; Liang, Z.; Liu, W.; Tong, Z.; Yi, L.; Morris, R.; Pasareanu, C.; and Koenig, S. 2019. Scheduling and Airport Taxiway Path Planning under Uncertainty .

Liao, C.-S.; and Huang, Y. 2014. The Covering Canadian Traveller Problem. Theoretical Computer Science 530: 80 – 88. ISSN 0304-3975. doi:https://doi.org/10.1016/j.tcs.2014.02.026. URL http://www.sciencedirect.com/science/article/pii/S0304397514001327.

Ling, J.; Gupta, T.; and Kumar, A. 2020. Reinforcement Learning for Zone Based Multiagent Pathfinding under Uncertainty. In International Conference on Automated Planning and Scheduling, 551–559.

Liu, M.; Ma, H.; Li, J.; and Koenig, S. 2019. Task and Path Planning for Multi-Agent Pickup and Delivery. In AAMAS, 1152–1160.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In Advances in Neural Information Processing Systems, 6380–6391.

Ma, H.; Kumar, T. K.; and Koenig, S. 2017. Multi-agent path finding with delay probabilities. In AAAI Conference on Artificial Intelligence, 3605–3612.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In International conference on machine learning, 1928–1937.

Nair, R.; Varakantham, P.; Tambe, M.; and Yokoo, M. 2005. Networked distributed POMDPs: A synthesis of distributed constraint optimization and POMDPs. In AAAI Conference on Artificial Intelligence, volume 1, 133–139.

Nguyen, D. T.; Kumar, A.; and Lau, H. C. 2017. Collective multiagent sequential decision making under uncertainty. In AAAI Conference on Artificial Intelligence, 3036–3043.

Nygren, E.; and Mohanty, S. 2020. Flatland Challenge: Multi Agent Reinforcement Learning on Trains. URL https://www.aicrowd.com/challenges/flatland-challenge.

Oliehoek, F. A.; and Amato, C. 2016. A Concise Introduction to Decentralized POMDPs.

Oztok, U.; and Darwiche, A. 2015. A top-down compiler for sentential decision diagrams. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

Peshkin, L.; Kim, K.-E.; Meuleau, N.; and Kaelbling, L. P. 2000. Learning to Cooperate via Policy Search. In Conference in Uncertainty in Artificial Intelligence, 489–496.

Rashid, T.; Samvelyan, M.; de Witt, C. S.; Farquhar, G.; Foerster, J. N.; and Whiteson, S. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning, 4292–4301.

Sartoretti, G.; Kerr, J.; Shi, Y.; Wagner, G.; Satish Kumar, T. K.; Koenig, S.; and Choset, H. 2019. PRIMAL: Pathfinding via Reinforcement and Imitation Multi-Agent Learning. IEEE Robotics and Automation Letters 4(3): 2378–2385.

Shen, Y.; Choi, A.; and Darwiche, A. 2016a. Tractable operations for arithmetic circuits of probabilistic models. In Advances in Neural Information Processing Systems, 3943–3951.

Shen, Y.; Choi, A.; and Darwiche, A. 2016b. Tractable operations for arithmetic circuits of probabilistic models. In Advances in Neural Information Processing Systems, 3936–3944.

Shen, Y.; Goyanka, A.; Darwiche, A.; and Choi, A. 2019. Structured bayesian networks: From inference to learning with routes. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 7957–7965.

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In Advances in Neural Information Processing Systems, 1057–1063.

Varakantham, P.; Cheng, S. F.; Gordon, G.; and Ahmed, A. 2012. Decision support for agent populations in uncertain and congested environments. In AAAI Conference on Artificial Intelligence, 1471–1477.

Volodymyr, M.; Koray, K.; David, S.; Rusu Andrei A; Joel, V.; Bellemare Marc G; Alex, G.; Martin, R.; Fidjeland Andreas K; and Georg, O. 2015. Human-level control through deep reinforcement learning. Nature 518(7540): 529.
Yu, J.; and LaValle, S. M. 2013. Structure and intractability of optimal multi-robot path planning on graphs. In AAAI Conference on Artificial Intelligence, 1443–1449.