6.845 Final Project: Quantum POMDPs

Jenny Barry

December 12, 2012

Abstract

We present quantum observable Markov decision processes (QOMDPs), the quantum analog of the partially observable Markov decision process (POMDP). In a QOMDP, an agent's state is represented as a quantum superposition and the agent can choose a superoperator to apply. This is similar to the POMDP belief state, which is a probability distribution over world states and evolves via a stochastic matrix. We show that the existence of a policy of at least a certain value has the same complexity for QOMDPs and POMDPs in the polynomial and infinite horizon cases. However, we also prove that the existence of a policy that can reach a goal state is decidable for POMDPs and undecidable for QOMDPs.

1 Introduction

Planning under uncertainty is a key problem in robotics. Because sensors and actuators are much less capable than human muscles or eyes, robots have a very limited understanding of their world. One way of modeling uncertainty is to use a partially observable Markov decision process (POMDP). POMDPs model an agent acting in a world of discrete states. The agent is never told its state, but can make actions and receive observations about the world. Both the action and the observation model are known to the agent but are non-deterministic. The agent is rewarded for its actual, hidden state at each time step, but, although it knows the reward model, it is not told its reward.

As we will discuss further in Section 2, it is possible to maximize future expected reward in a POMDP by only maintaining a probability distribution, or belief state, over the agent's current state. By carefully updating this state after every action and observation, we can ensure that it reflects the underlying probability that the agent is in each state. We can make decisions using only the agent's belief about its state without ever needing to reason more directly about its exact state.

The evolution of a POMDP appears at a high level to be related to the evolution of a quantum system; in both, we can change the system but only observe the state indirectly. However, classical probability distributions evolve stochastically while quantum superpositions evolve unitarily. Uncertainty can only be introduced into a quantum system via an a priori lack of knowledge or a measurement. Therefore, we must define more precisely what we mean by a "quantum observable" Markov decision process (QOMDP) before we can explore their properties.

In Section 3, we give a definition of a quantum observable Markov decision process. We then show that, while solving a QOMDP has the same complexity as solving a POMDP, there are other related decision problems that are decidable in the classical POMDP case but undecidable for QOMDPs.

2 Partially Observable Markov Decision Processes (POMDPs)

For the reader's convenience, we begin with an overview of Markov decision processes and partially observable Markov decision processes. A reader familiar with these concepts may wish to skip to Section 3.

2.1 Fully Observable Case

We begin by defining fully observable Markov decision processes, usually just called Markov decision processes (MDPs). This will facilitate us in our discussion of POMDPs as POMDPs can be reduced to continuous state MDPs. The bulk of this discussion was taken from Russell and Norvig, Chapter 17 [7].

A Markov Decision Process (MDP) is a model of an agent acting in an uncertain, but observable world. An MDP is a tuple $\langle S, A, T, R, \gamma \rangle$ consisting of a set of states S, a set of actions A, a state transition function $T(s_i, a, s_j) : S \times A \times S \to [0, 1]$ giving the probability that taking action a in state s_i results in state s_j, a reward function $R(s_i, a) : S \times A \to \mathbb{R}$ giving the reward of taking action a in state s_i, and a discount factor $\gamma \in [0, 1)$ that discounts the importance of reward gained later in time. Note that having the reward depend on current state and action is WLOG; it is always possible to define another MDP in which the reward depends only on state or on the current state and next state with only a polynomial increase in input size. At each time step, the agent is in exactly one, known state, chooses to take a single action, and transitions to a new state according to T. The objective is to act in such a way as to maximize future expected reward. We formalize this objective in the Policy Existence Problem.

A *policy* $\pi(s_i, t) : S \times \mathbb{Z}^+ \to A$ is a function mapping states at time t to actions. The *value* of a policy at state s_i over horizon h is the future expected reward of acting according to π for h time steps:

$$V_\pi(s_i, h) = R(s_i, \pi(s_i, h)) + \gamma \sum_{s_j \in S} T(s_i, \pi(s_i, h), s_j) V_\pi(s_j, h - 1).$$

The *solution* to an MDP of horizon h is the policy that maximizes future expected reward over horizon h. The associated decision problem is the policy existence problem:

Definition 1 (Policy Existence Problem): The *policy existence problem* is to decide, given a decision process, whether there is a policy of horizon h that achieves value at least V for the given starting state.

For MDPs, we will evaluate the infinite horizon case. In this case, we will drop the time argument from the policy since it cannot matter (the optimal policy at time infinity is the same as the optimal policy at time infinity minus one). The optimal policy over an infinite horizon is the one inducing the value function

$$V^*(s_i) = \max_{a \in A} \left[R(s_i, a) + \gamma \sum_{s_j \in S} T(s_i, a, s_j) V^*(s_j) \right]. \tag{1}$$

This is the Bellman equation and there is a unique solution for V^*. V^* is non-infinite if $\gamma < 1$.

When the input size is polynomial in $|S|$, finding an ϵ-optimal policy for an MDP is in P. One algorithm for finding the optimal value function is to simply to initialize $V = 0$ and then iterate Bellman's equation. This is called *value iteration*:

ValueIteration$(S, A, T, R, \gamma, \epsilon)$

$V_0(s_i) \leftarrow 0$, $\pi(s_i) \leftarrow$ NULL, $\delta \leftarrow \infty$, $t \leftarrow 1$
while $\delta > \frac{\epsilon(1-\gamma)}{2\gamma}$
    for $s_i \in S$
        $\pi(s_i) \leftarrow \arg\max_{a \in A} \left[R(s_i, a) + \gamma \sum_{s_j \in S} T(s_i, a, s_j) V_{t-1}(s_j) \right]$
        $V_t(s_i) \leftarrow R(s_i, \pi(s_i)) + \gamma \sum_{s_j \in S} T(s_i, \pi(s_i), s_j) V_{t-1}(s_j)$
    $\delta \leftarrow \max_{s_i \in S} |V_t(s_i) - V_{t-1}(s_i)|$, $t \leftarrow t + 1$
return π

The running time of value iteration is $O\left(|S|^2 |A| \log(\epsilon R_{max}/(1-\gamma))\right)$, where $R_{max} = \max_{s_i \in S, a \in A} R(s_i, a)$. Therefore, the policy existence problem for MDPs is in P.

Goal MDP: A derivative of the MDP of interest to us is the *goal MDP*. A goal MDP is a tuple $M = \langle S, A, T, g \rangle$ where S, A, and T are as before and $g \in S$ is an absorbing goal state so $T(g, a, g) = 1$ for all $a \in A$. The objective in a goal MDP is to find the policy that reaches the goal with the highest probability. The associated decision problem is the Goal-State Reachability Problem:

Definition 2 (Goal-State Reachability Problem for Decision Processes): The *goal-state reachability problem* is to decide, given a goal decision process, whether there exists a policy that can reach the goal state from the starting state with probability at least p.

When solving goal decision processes, we never need to consider time-dependent policies because nothing changes with the passing of time. Therefore, when analyzing the goal-state reachability problem, we will only consider *stationary policies* that depend only upon the current state.

2.2 Partially Observable Case

A partially observable Markov decision process (POMDP) generalizes an MDP to the case where the world is also not observable. We follow the work of Kaelbling et al. [3] in explaining POMDPs.

In a partially observable world, the agent does not know its own state but receives information about it in the form of observations. Formally, a POMDP is a tuple $\left\langle S, A, \Omega, T, R, O, \vec{b}_0, \gamma \right\rangle$ where S is a set of states, A is a set of actions, Ω is a set of observations, $T(s_i, a, s_j) : S \times A \times S \to [0, 1]$ is the probability of transitioning to state s_j given that action a was taken in state s_i, $R(s_i, a) : S \times A \to \mathbb{R}$ is the reward for taking action a in state s_i, $O(s_j, a, o) : S \times A \times \Omega \to [0, 1]$ is the probability of making observation o given that action a was taken and ended in state s_j, $\vec{b}_0$ is a probability distribution over possible initial states, and $\gamma \in [0, 1)$ is the discount factor. In a POMDP the agent's state is "hidden" meaning that the agent does not know its state, but the dynamics of the world behave according to agent's actual state. At each time step, the agent chooses an action, transitions to a new state according to its hidden starting state and T, and receives an observation according to its hidden ending state and O. Again having the reward and observations depend upon the state and action in the way they do is WLOG; it is always possible to define another POMDP in which the observation depends on the current state or the reward depends on the ending state, etc. As with MDPs, the goal is to maximize future expected reward.

The easiest way to understand POMDPs is to consider the *belief MDP*. A *belief state* $\vec{b}$ is a probability distribution over possible states. For $s_i \in S$, $\vec{b}_i$ is the probability that the agent is in state s_i. Since $\vec{b}$ is

a probability distribution, $0 \leq \vec{b}_i \leq 1$ and $\sum_{i=1}^{|S|} \vec{b}_i = 1$. If the agent is in belief state $\vec{b}$, takes action a, and receives observation o the new belief state is

$$\begin{aligned} \vec{b'}_j &= \Pr(s_j|o,a,\vec{b}) \\ &= \frac{\Pr(o|s_j,a,b)\Pr(s_j|a,b)}{\Pr(o|a,b)} \\ &= \frac{O(s_j,a,o)\sum_{s_i \in S} T(s_i,a,s_j)\vec{b}_i}{\Pr(o|a,b)}. \end{aligned} \tag{2}$$

This is the belief update equation. $\Pr(o|a,b) = \sum_j O(s_j,a,o)\sum_{s_i \in S} T(s_i,a,s_j)\vec{b}_i$ is independent of s' and usually just computed afterwards as a normalizing factor that causes $\vec{b'}$ to sum to one. We define the matrix

$$(\tau^{ao})_{ij} = O(s_j,a,o)T(s_i,a,s_j). \tag{3}$$

The belief update for seeing observation o after taking action a is

$$\vec{b'} = \frac{\tau^{ao}\vec{b}}{\left|\tau^{ao}\vec{b}\right|_1} \tag{4}$$

where $\left|\vec{b}\right|_1 = \sum_i \vec{b}_i$ is the L1-norm. The probability of transitioning from belief state $\vec{b}$ to belief state $\vec{b'}$ when taking action a is

$$\tau(\vec{b},a,\vec{b'}) = \sum_{o \in \Omega} \Pr(b'|a,b,o)\Pr(o|a,b) \tag{5}$$

where

$$\Pr(\vec{b'}|a,\vec{b},o) = \begin{cases} 1 & \text{if } \vec{b'} = \frac{\tau^{ao}\vec{b}}{\left|\tau^{ao}\vec{b}\right|_1} \\ 0 & \text{else.} \end{cases}$$

The expected reward of taking action a in belief state $\vec{b}$ is

$$r(b,a) = \sum_{s_i \in S} \vec{b}_i R(s_i,a). \tag{6}$$

Now the agent always knows its belief state so the belief space is fully observable. This means we can define the *belief MDP* $\langle B, A, \tau, r, \gamma \rangle$ where B is the set of all possible belief states. The optimal solution to the MDP is also the optimal solution to the POMDP. The only problem is that the state space is continuous and all known algorithms for solving MDPs optimally in polynomial time are polynomial in the size of the state space.

However, it is still possible to solve a POMDP as an MDP (although not polynomially) because its optimal value function over beliefs has a specific structure. We will show that the value function for a POMDP is piecewise-linear and convex.

The easiest way to think about policies for a POMDP is to consider a *policy tree* like the one shown in Figure 1. A policy tree specifies the action to take with t steps remaining given the entire history of observations and actions (the belief state is a sufficient statistic for this history, which is why the belief dynamics are Markovian). Let p be a t-step policy tree, let $p(s_i)$ be the action to take if the agent starts in state s_i, and let $p(o)$ be the $t-1$ sub-tree associated with having seen observation o after taking $p(s_i)$. The expected value of following p from state s_i is

$$\begin{aligned} V_p(s_i) &= R(s_i,p(s_i)) + \gamma \sum_{s_j \in S} \Pr(s_j|s_i,p(s_i)) \sum_{o \in \Omega} \Pr(o|s_j,p(s_i))V_{p(o)}(s_j) \\ &= R(s_i,p(s_i)) + \gamma \sum_{s_j \in S} T(s_i,p(s_i),s_j) \sum_{o \in \Omega} O(s_j,p(s_i),o)V_{p(o)}(s_j). \end{aligned} \tag{7}$$

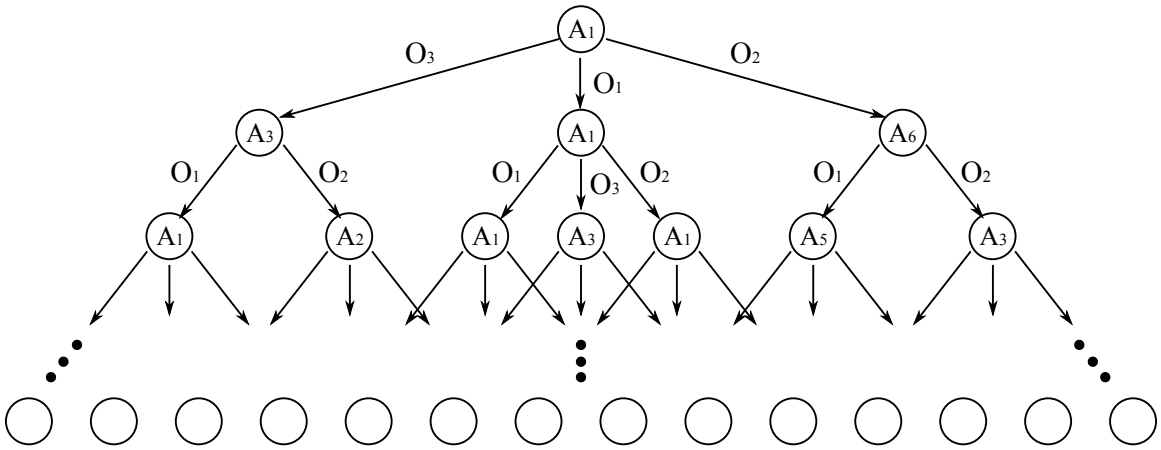

Figure 1: A policy tree is a policy for a POMDP. This shows the action to take given the entire history of actions previously taken and observations previously made.

The value of a belief state is $V_p(\vec{b}) = \sum_{s_i \in S} \vec{b}_i V_p(s_i)$. We let the *alpha vector* of a policy be $\vec{\alpha_p} = \left[V_p(s_1), ..., V_p(s_{|S|})\right]$ so $V_p(\vec{b}) = \vec{b} \cdot \vec{\alpha_p}$. The optimal t-step value function is the maximum over the possible t-step policy trees $\mathcal{P}$,

$$V_t(\vec{b}) = \max_{p \in \mathcal{P}} \vec{b} \cdot \vec{\alpha_p}. \tag{8}$$

Equation 8 tells us the form of the value function. Each policy tree p gives a value function $V_p(\vec{b}) = \vec{b} \cdot \vec{\alpha_p}$ that is linear in $\vec{b}$. The optimal value function V_t is a collection of these individual value functions, choosing the one that optimizes each $\vec{b}$. Therefore, V_t is piecewise-linear. Moreover, V_t is the top surface of the V_p's so V_t is also convex. This makes intuitive sense as the agent could certainly pick the best actions for a state if the agent knew its current state (the "edges" of the belief space where the probability of a particular state is 1). The closer the agent is to knowing its current state, the better it can adapt its choices to its actual state. In the "middle" range of belief values, the belief is close to uniform so the agent has to make a very uninformed decision. Therefore, the value function is lowest in the middle and highest at the edges of the space. This has a connection to value of information (the value of moving to a more uncertain belief).

Now the problem of finding the optimal value function is reduced to assigning policy trees to sets of beliefs. Given a set of policy trees, it is always possible to define a subset of those trees that give rise to the same optimal value function. (In theory, this set could be improper - it is possible that every policy tree could be optimal for some belief - but luckily this appears not to be the case in most problems.) Let $\mathcal{R}(\vec{\alpha}_p)$ be the region of space in which a particular alpha vector dominates,

$$\mathcal{R}(\vec{\alpha}_p) = \left\{ \vec{b} \,\middle|\, \forall p' \in \mathcal{P},\ \vec{b} \cdot \vec{\alpha}_p \geq \vec{b} \cdot \vec{\alpha_{p'}} \right\}.$$

A linear program can be used to find if $\mathcal{R}(\alpha_p)$ is empty or not. We call a policy tree p *useful* if $\mathcal{R}(\vec{\alpha}_p)$ is non-empty.

The simplest method for solving a POMDP is to start with the useful set of policies with one step to go and use dynamic programming to find the set of useful policies with t steps to go. This is similar to the value iteration method for solving MDPs. Let $\mathcal{P}_t$ be the set of useful policy trees with t steps to go. We calculate $\mathcal{P}_t$ from $\mathcal{P}_{t-1}$ by considering the set $\mathcal{P}_t^+$ of all policy trees with t steps to go such that all $t-1$ sub-trees are in $\mathcal{P}_{t-1}$. Clearly $\mathcal{P}_t \subseteq \mathcal{P}_t^+$ and we can find $\mathcal{P}_t$ by identifying the $p \in \mathcal{P}_t^+$ such that $\mathcal{R}(\vec{\alpha}_p)$

is not empty. The problem is that $|\mathcal{P}_t^+| = |A||\mathcal{P}_{t-1}|^{|\Omega|}$ so $\mathcal{P}_t^+$ cannot be enumerated in polynomial time, but the value function can be evaluated in polynomial space if we are only interested in a polynomial horizon. Therefore, the policy existence problem for POMDPs is in PSPACE. Note that we cannot represent the actual policy in PSPACE, but we can evaluate the value function for a given belief state. It was shown in 1987 that the policy existence problem for POMDPs is in fact PSPACE-Complete [5]. The policy existence problems for POMDPs in the infinite horizon case, however, is undecidable [4].

Goal POMDP: A *goal POMDP* is a tuple $P = \left\langle S, A, \Omega, T, O, \vec{b}_0, g \right\rangle$ where S, A, Ω, T, and O are defined as before but instead of a reward function, we assume that $g \in S$ is a goal state. This state g is absorbing so we are promised that for all $a \in A$, that $T(g, a, g) = 1$. Moreover, the agent receives an observation $o_{|\Omega|} \in \Omega$ telling it that it has reached the goal so for all $a \in A$, $O(g, a, o_{|\Omega|}) = 1$. This observation is only received in the goal state so for all $s_i \neq g$, and all $a \in A$, $O(s_i, a, o_{|\Omega|}) = 0$. The solution to a goal POMDP is a policy that reaches the goal state with the highest possible probability starting from $\vec{b}_0$.

Because the goal is absorbing and known, the observable belief space corresponding to a goal POMDP is a goal MDP $M(P) = \left\langle B, A, \tau, \vec{b}_0, \vec{b}_g \right\rangle$. Here $\vec{b}_g$ is the state in which the agent knows it is in g with probability 1. We show that this is absorbing. Firstly the probability of observing o after taking action a is

$$\Pr(o|a, \vec{b}_g) = \sum_{s_j \in S} O(s_j, a, o) \sum_{s_i \in S} T(s_i, a, s_j) \left(\vec{b}_g\right)_i = \sum_{s_j \in S} O(s_j, a, o) T(g, a, s_j) = O(g, a, o) = \delta_{o o_{|\Omega|}}.$$

Therefore, if we are in state $\vec{b}_g$, regardless of the action taken, we see observation $o_{|\Omega|}$. Assume we take action a and see observation $o_{|\Omega|}$. The next belief state is

$$\vec{b'}_j = \Pr(s_j|o_{|\Omega|}, a, \vec{b_g}) = \frac{O(s_j, a, o_{|\Omega|}) \sum_{s_i \in S} T(s_j, a, s_i) \vec{b}_i}{\Pr(o_{|\Omega|}|a, b)} = \frac{O(s_j, a, o_{|\Omega|})}{\Pr(o_{|\Omega|}|a, b)} T(s_j, a, g) = \delta_{s_j g}.$$

Therefore, regardless of the action taken, the next belief state is $\vec{b}_g$ so we have a goal MDP.

3 QOMDPs: Quantum Observable Markov Decision Processes

In this section, we formulate a quantum observable Markov decision process (QOMDP). A QOMDP differs from an MDP in that its states are always continuous and are allowed to be entangled. A QOMDP can simulate both an MDP or the belief space of a POMDP.

We will first give the necessary background in quantum superoperators and then give a formal definition of a QOMDP.

3.1 Notation

We briefly review the notation we will use in this paper. We let

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{else} \end{cases}$$

be the Kronecker delta function. $\mathbb{I}_n$ is the $n \times n$ identity matrix

$$(\mathbb{I}_n)_{ij} = \delta_{ij}.$$

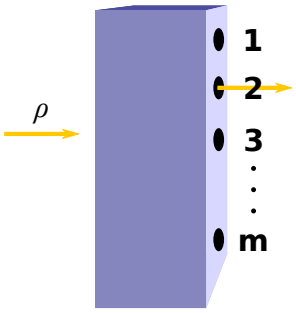

Figure 2: A quantum superoperator with m possible outcomes. We view a superoperator as a box with m ports. When the state ρ is input, it chooses an output port non-deterministically and we receive the corresponding observation.

We use standard bra-ket notation, in which $|\psi\rangle$ is a column vector while $\langle\psi|$ is a its conjugate transpose row vector. The notation $\langle\psi|\phi\rangle$ denotes the inner product between $|\psi\rangle$ and $|\phi\rangle$ while $|\psi\rangle\langle\phi|$ is the outer. The state $|i\rangle$ is the ith basis state

$$|i\rangle_j = \delta_{ij}.$$

We will use basis states to pick out elements of other vectors so the ith element of $|\psi\rangle$ is $\langle i|\psi\rangle$ with the ijth element of M is $\langle i|M|j\rangle$.

3.2 Quantum Operators and Kraus Matrices

In an MDP or POMDP we have actions with probabilistic outcomes. We can create probabilistic outcomes in the quantum mechanics framework using measurements to probabilistically collapse the state. We have chosen to formulate actions as quantum operators, which are operators that probabilistically produce quantum states. Intuitively, we view superoperators as a box with several possible output ports. When a state is input, the output port is chosen non-deterministically and the associated observation is returned. This is shown in Figure 2.

A quantum superoperator $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ acting on states of dimension d is defined by $\mathcal{K}$ $d \times d$ Kraus matrices[1] [1]. Given a density matrix ρ, there are $\mathcal{K}$ possible next states for ρ. Specifically the next state is

$$\rho'_i \rightarrow \frac{K_i \rho K_i^\dagger}{\mathrm{Tr}\left(K_i \rho K_i^\dagger\right)} \tag{9}$$

with probability

$$\Pr(\rho'_i|\rho) = \mathrm{Tr}\left(K_i \rho K_i^\dagger\right). \tag{10}$$

The superoperator returns observation i if the ith Kraus matrix was applied.

Since there must be probability 1 that ρ transitions somewhere, we have

$$1 = \sum_{i=1}^{\mathcal{K}} \mathrm{Tr}(K_i \rho K_i^\dagger) = \sum_{i=1}^{\mathcal{K}} \mathrm{Tr}(K_i^\dagger K_i \rho) = \mathrm{Tr}\left(\sum_{i=1}^{\mathcal{K}} K_i^\dagger K_i \rho\right) \tag{11}$$

[1] Actually, the quantum operator acts on a product state of which the first dimension is d. In order to create quantum states of dimension d probabilistically, the superoperator entangles the possible next states with a measurement register and then measures that register. Thus the operator actually acts on the higher-dimensional product space, but for the purposes of this discussion, we can treat it as an operator that probabilistically maps states of dimension d to states of dimension d.

for any density matrix ρ. Let

$$S = \sum_{i=1}^{\mathcal{K}} K_i^\dagger K_i,$$

and let $\rho = |i\rangle\langle i|$ so

$$\rho_{jk} = \begin{cases} 1 & j = k = i \\ 0 & \text{else} \end{cases}$$

is the matrix with a one at position ρ_{ii} and zeros everywhere else. Then Equation 11 gives us that $S_{ii} = 1$ so S has ones along the diagonal. Now let ρ be any density matrix. Then

$$1 = \text{Tr}(S\rho) = \sum_{i,j=1}^{\mathcal{K}} S_{ij}\rho_{ji} = \sum_{i=1}^{\mathcal{K}} \rho_{ii} + \sum_{i \neq j} S_{ij}\rho_{ji} = 1 + \sum_{i \neq j} S_{ij}\rho_{ju}$$

so $S_{ij} = 0$ if $i \neq j$. Therefore S is the identity matrix. A set of matrices $\{K_1, ..., K_\mathcal{K}\}$ of dimension d is a set of Kraus matrices if and only if

$$\sum_{i=1}^{\mathcal{K}} K_i^\dagger K_i = \mathbb{I}_d.$$

Note that this sum is $K_i^\dagger K_i$ while the state evolution is $K_i \rho K_i^\dagger$.

3.3 QOMDP Formulation

We can now define the quantum observable Markov decision process (QOMDP). A QOMDP is a tuple $\langle S, \Omega, \mathcal{A}, R, \gamma, \rho_0 \rangle$ where

- S is a Hilbert space. We allow pure and mixed quantum states so we will represent states in S as density matrices.
- $\Omega = \{\Omega_1, ..., \Omega_{|\Omega|}\}$ is a set of possible observations.
- $\mathcal{A} = \left\{A^1, ..., A^{|\mathcal{A}|}\right\}$ is a set of superoperators. Each superoperator $A^i = \left\{A_1^i, ..., A_{|\Omega|}^i\right\}$ has $|\Omega|$ Kraus matrices, some of which may be the all-zeros matrix. The return of o_i indicates the application of the ith Kraus matrix.
- $R : S \times \mathcal{A} \to \mathbb{R}$ is a reward function.
- $\gamma \in [0, 1)$ is a discount factor.
- $\rho_0 \in S$ is the starting state.

At each time step, the agent chooses a superoperator and receives an observation. As with MDPs and POMDPs, the objective is for the agent to act in such a way as to maximize future expected reward.

QOMDPs are fully observable in the sense that we always know the current quantum superposition or mixed state (this is very similar to "knowing" the probability distribution over the possible world states in the belief space MDP). Since we are given the initial state, we can deduce the state of the system after n steps given the sequence $\{(a_1, o_1), ..., (a_n, o_n)\}$ of actions taken and observations made. The idea of a partially observable QOMDP in which we are not given the initial state is out of the scope of this paper, but a possible direction for future research.

As with MDPs, a policy for a QOMDP is a function $\pi : S \times \mathbb{Z} \to \mathcal{A}$ mapping states at time t to actions. The value of the policy over horizon h starting from state ρ_0 is

$$V^{\pi}(\rho_0) = \sum_{t=0}^{h} E\left[\gamma^t R(\rho_t, \pi(\rho_t)) \middle| \pi\right].$$

The solution to a QOMDP is the policy that maximizes future expected reward

$$\pi^* = \arg\max_{\pi} V^{\pi}.$$

For a general reward function, the Bellman equation (Equation 1) clearly does not hold. Whether there is a reward function type that guarantees the Bellman equation will hold for QOMDPs is unclear.

Goal QOMDPs A *goal QOMDP* is a tuple $\langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ where S, Ω, $\mathcal{A}$, and ρ_0 are as defined above. The goal state ρ_g must be absorbing so that for all $A^i \in \mathcal{A}$ and all $A^i_j \in A^i$ if $\mathrm{Tr}\left(A^i_j \rho_g {A^i_j}^\dagger\right) > 0$ then

$$\frac{A^i_j \rho_g {A^i_j}^\dagger}{\mathrm{Tr}\left(A^i_j \rho_g {A^i_j}^\dagger\right)} = \rho_g.$$

As with goal MDPs and POMDPs, the objective for a goal QOMDP is to maximize the probability of achieving the goal state.

3.4 QOMDP Policy Existence Complexity

As we can always simulate classical evolution with a quantum system[2], the definition of QOMDPs contains POMDPs. Therefore we immediately find that the policy existence problem for the infinite horizon case is undecidable. We also have that the polynomial horizon case is at least PSPACE-Complete. We can, in fact, prove that the polynomial horizon case is in PSPACE.

Theorem 1: The policy existence problem (Definition 1) for QOMDPs with a polynomial horizon is in PSPACE.

Proof: Given a QOMDP $\langle S, \Omega, \mathcal{A}, R, \gamma, \rho_0 \rangle$ and horizon h, consider the set of possible policies. The state is observable and Markovian, so we need only consider policies dependent on the current state and time-to-go. From the starting state we can reach $O\left((|\mathcal{A}||\Omega|)^h\right)$ possible states giving us $O\left(h\,|\mathcal{A}|\,(|\mathcal{A}||\Omega|)^h\right)$ possible policies. This number is only exponential so we can represent it exactly in PSPACE. Therefore, we can assign every policy and every state a number allowing us to determine the value of policy i at state j and time step k. The value of the policy can be at most $h \max_{s_i \in S} \max_{a \in A} R(s_i, a)$ so the value will also be representable in PSPACE. Thus we can evaluate every policy and find the best one in PSPACE. □

4 A Complexity Separation in Goal-State Reachability

However, although the policy existence problem has the same complexity for QOMDPs and POMDPs, we can show that the goal-state reachability problem with probability 1 (Definition 2) is decidable for

[2]I certainly believe this is true, but never formulated the reduction from POMDPs to QOMDPs. If this is actually incorrect, the lower bounds for QOMDPs do not follow from POMDPs, but are almost certainly true anyway.

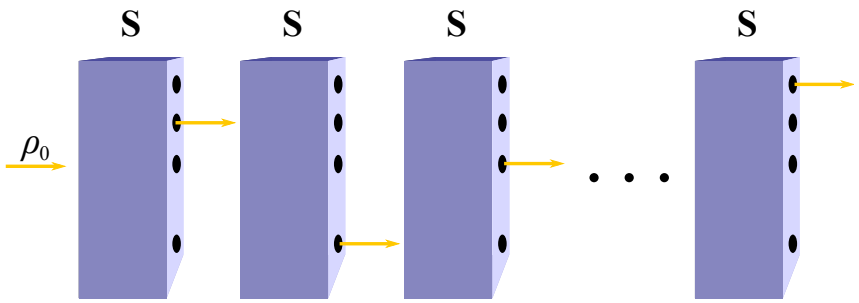

Figure 3: The quantum measurement occurrence problem. The starting state ρ_0 is fed into the superoperator **S**. The output is then fed iteratively back into **S**. The question is whether there is some finite sequence of observations that can never be observed.

goal POMDPs but undecidable for goal QOMDPs. Throughout this section when we discuss goal-state reachability problem it is assumed we mean the goal-state reachability problem with probability 1.

4.1 Undecidability of Goal-State Reachability for QOMDPs

We will show that the goal-state reachability problem is undecidable for QOMDPs by showing that we can reduce the quantum measurement occurrence problem proposed by Eisert et al. [2] to it.

Definition 3 (Quantum Measurement Occurrence Problem): The *quantum measurement occurrence problem* (QMOP) is to decide, given a quantum superoperator described by $\mathcal{K}$ Kraus operators $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ whether there is some finite sequence $\{i_1, ..., i_n\}$ such that $K_{i_1}^\dagger ... K_{i_n}^\dagger K_{i_n} ... A_{i_1} = 0$.

The setting for this problem is shown in Figure 3. We assume that the system starts in state ρ_0. This state is fed into $\mathcal{S}$. We then take the output of $\mathcal{S}$ acting on ρ_0 and feed that again into $\mathcal{S}$ and iterate. QMOP is equivalent to asking whether there is some finite sequence of observations $\{i_1, ..., i_n\}$ that can never be observed even if ρ_0 is full rank. We will reduce from the version of the problem given in Definition 3, but will use the language of measurement occurrence in providing intuition.

Theorem 2 (Undecidability of QMOP): The quantum measurement occurrence problem is undecidable.

Proof: This can be shown using a reduction from the matrix mortality problem. For the full proof see Eisert et al[2]. □

We first describe a method for creating a goal QOMDP from an instance of QMOP. The main ideas behind the choices we make here are shown in Figure 4.

Definition 4 (QMOP Goal QOMDP): Given an instance of QMOP $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ with Kraus matrices of dimension d, we create a goal QOMDP $Q(\mathcal{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ as follows:

- S is $d+1$-dimensional Hilbert space.
- $\Omega = \{o_1, o_2, ..., o_{d+2}\}$ is a set of $d+2$ possible observations. Observations o_1 through o_{d+1} correspond to At-Goal while o_{d+2} is Not-At-Goal.

- $\mathcal{A} = \left\{A^1, ..., A^{\mathcal{K}}\right\}$ is a set of $\mathcal{K}$ superoperators each with $d+2$ Kraus matrices $A^i = \{A^i_1, ..., A^i_{d+2}\}$. We set

$$A^i_{d+2} = K_i \oplus 0 = \begin{bmatrix} & & 0 \\ & K_i & \vdots \\ & & 0 \\ 0 & ... & 0 \end{bmatrix},$$

the ith Kraus matrix from the QMOP instance with the $d+1$st column and row all zeros. Now let

$$Z^i = \mathbb{I}_{d+1} - {A^i_{d+2}}^\dagger A^i_{d+2} = \left(\sum_{j \neq i} K_j^\dagger K_j\right) \oplus 1 = \begin{bmatrix} & & & 0 \\ & \sum_{i \neq j} K_j & & 0 \\ & & & \vdots \\ 0 & 0 & ... & 1 \end{bmatrix}.$$

Now $\left(K_j^\dagger K_j\right)^\dagger = K_j^\dagger K_j$ so Z^i is Hermitian. Moreover, $K_j^\dagger K_j$ is positive semi-definite for all j so Z^i is positive semi-definite. Let an orthonormal eigendecomposition of Z^i be

$$Z^i = \sum_{j=1}^{d+1} z^i_j |z^i_j\rangle\langle z^i_j|.$$

Since Z^i is a positive semi-definite Hermitian matrix, z^i_j is non-negative and real so $\sqrt{z^i_j}$ is also real. We let A^i_j for $j < d+2$ be the $d+1 \times d+1$ matrix in which the first d rows are all 0s and the bottom row is $\sqrt{z^i_j}\langle z^i_j|$:

$$\begin{aligned} A^i_{j>0pq} &= \sqrt{z^i_j}\langle z^i_j|q\rangle\delta_{p(d+1)}, \\ A^i_{j>0} &= \begin{bmatrix} 0 & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & 0 \\ \sqrt{z^i_j}\langle z^i_j| \end{bmatrix}. \end{aligned}$$

(Note that if $z^i_j = 0$ then A^i_j is the all-zero matrix but it is cleaner to allow each action to have the same number of Kraus matrices.)

- ρ_0 is the maximally mixed state $\rho_{0ij} = \frac{1}{d+1}\delta_{ij}$.
- ρ_g is the state $|d+1\rangle\langle d+1|$.

The intuition behind the definition of $Q(\mathcal{S})$ is shown in Figure 4. Although each action actually has $d+2$ choices, we will show that $d+1$ of those choices (every one except A^i_{d+2}) always transitions to the goal state. Therefore each action A^i really only has two choices:

1. Transition to goal state.
2. Evolve according to K_i.

Now consider choosing some sequence of actions $A^{i_1}, ..., A^{i_n}$. The probability that we transition to the goal state is the same as the probability that we do not evolve according to first K_{i_1} then K_{i_2} etc. Therefore, we transition to the goal state with probability 1 if and only if it is impossible to transition

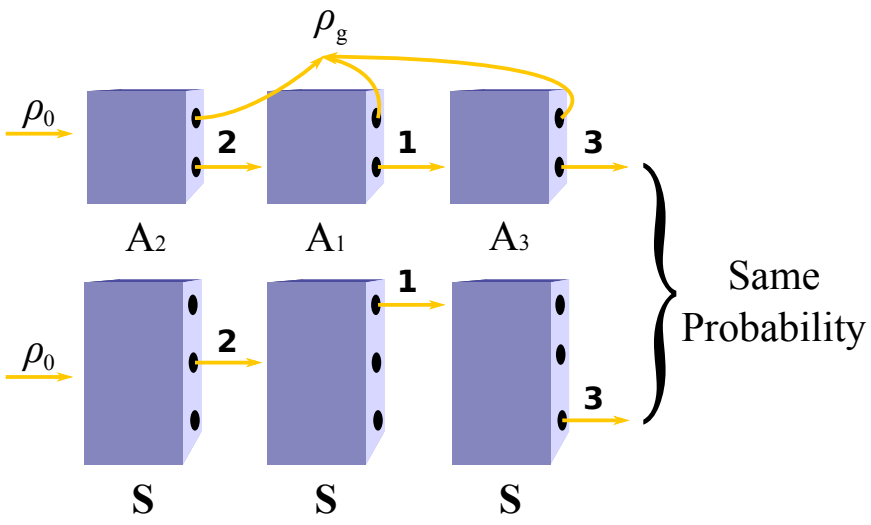

Figure 4: A goal QOMDP for a QMOP instance with superoperator **S**. We create 3 actions to correspond to the 3 outputs of the superoperator. Each action A_i has two options for ρ: either it transitions according to K_i from **S** or it transitions to the goal state. Intuitively, we can think of A_i as either outputting the observation "transitioned to goal" or observation i from **S**. Then it is clear that if the action sequence $\{A_2, A_1, A_3\}$ is taken, for instance, the probability that we do *not* see the observation sequence 2, 1, 3 is the probability that the system transitions to the goal state somewhere in this sequence. Therefore, the probability that an action sequence reaches the goal state is the probability that the corresponding observation sequence is not observed.

according to first K_{i_1} then K_{i_2} etc. Thus in the original problem, it must have been impossible to see the observation sequence $\{i_1, ..., i_n\}$. In other words, we can reach a goal state with probability 1 if and only if there is some sequence of observations in the QMOP instance that can never be observed. So we can use goal-state reachability in QOMDPs to solve QMOP giving us that goal-state reachability for QOMDPs must be undecidable.

We now prove formally the sketch we just gave. Before we can do anything else, we must show that $Q(\mathcal{S})$ is in fact a goal QOMDP. We start by showing that ρ_g is absorbing in two lemmas. In the first, we prove that $A^i_{j<d+2}$ transitions all density matrices to the goal state. In the second we show that ρ_g has zero probability of evolving according to A^i_{d+2}.

Lemma 3: Let $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ with Kraus matrices of dimension d be an instance of QMOP and let $Q(\mathcal{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ be the corresponding goal QOMDP. For any density matrix $\rho \in S$, if A^i_j is the jth Kraus matrix of the ith action of $Q(\mathcal{S})$ and $j < d + 2$ then

$$\frac{A^i_j \rho {A^i_j}^\dagger}{\mathrm{Tr}\left(A^i_j \rho {A^i_j}^\dagger\right)} = |d+1\rangle\langle d+1|.$$

Proof: Consider

$$\begin{aligned}\left(A^i_j \rho {A^i_j}^\dagger\right)_{pq} &= \sum_{h,l} {A^i_j}_{ph} \rho_{hl} {A^i_j}^\dagger_{lq} \\ &= \sum_{h,l} {A^i_j}_{ph} \rho_{hl} {A^i_j}^*_{ql} \\ &= z^i_j \sum_{h,l} \langle z^i_j | h\rangle \rho_{hl} \langle l | z^i_j\rangle \delta_{p(d+1)} \delta_{q(d+1)}\end{aligned}$$

so only the lower right element of this matrix is non-zero. Dividing by the trace gives

$$\left(\frac{A^i_j \rho {A^i_j}^\dagger}{\mathrm{Tr}\left(A^i_j \rho {A^i_j}^\dagger\right)}\right)_{pq} = \delta_{p(d+1)}\delta_{q(d+1)} = \langle p|d+1\rangle\langle d+1|q\rangle$$

so

$$\frac{A^i_j \rho {A^i_j}^\dagger}{\mathrm{Tr}\left(A^i_j \rho {A^i_j}^\dagger\right)} = |d+1\rangle\langle d+1|.$$

□

Lemma 4: Let $\mathcal{S}$ be an instance of QMOP and let $Q(\mathcal{S}) = \{S, \Omega, \mathcal{A}, \rho_0, \rho_g\}$ be the corresponding QOMDP. Then ρ_g is absorbing.

Proof: By Lemma 3, we know that for $j < d+2$, $\frac{A^i_j|d+1\rangle\langle d+1|{A^i_j}^\dagger}{\mathrm{Tr}\left(A^i_j|d+1\rangle\langle d+1|{A^i_j}^\dagger\right)} = \rho_g$. Here we show that $\mathrm{Tr}\left(A^i_{d+2}\rho_g {A^i_{d+2}}^\dagger\right) = 0$ so that the probability of applying A^i_{d+2} is zero.

$$\mathrm{Tr}\left(A^i_{d+2}|d+1\rangle\langle d+1|{A^i_{d+2}}^\dagger\right) = \sum_p \sum_{hl} {A^i_{d+2}}_{ph} \delta_{h(d+1)} \delta_{l(d+1)} {A^i_{d+2}}^*_{pl} = \sum_p {A^i_{d+2}}_{p(d+1)} {A^i_{d+2}}^*_{p(d+1)} = 0$$

because the $d+1$st column of A^i_{d+2} is all zeros by construction. Therefore, ρ_g is absorbing. □

Now we are ready to show that $Q(\mathcal{S})$ is a goal QOMDP. All that remains is to show that the actions are actually superoperators.

Theorem 5: Let $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ be an instance of QMOP with Kraus matrices of dimension d. Then $Q(\mathcal{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g\rangle$ is a goal QOMDP.

Proof: We showed in Lemma 4 that ρ_g is absorbing so we must show that the actions are indeed superoperators. Let A^i_j be the jth Kraus matrix of action A^i. If $j < d+2$ then

$$\left({A^i_j}^\dagger A^i_j\right)_{pq} = \sum_h {A^i_j}^\dagger_{ph} {A^i_j}_{hq} = \sum_h {A^i_j}^*_{hp} {A^i_j}_{hq} = \sqrt{z^i_j}^* \langle p|z^i_j\rangle \sqrt{z^i_j}\langle z^i_j|q\rangle = z^i_j \langle p|z^i_j\rangle\langle z^i_j|q\rangle$$

where we have used that $\sqrt{z^i_j}^* = \sqrt{z^i_j}$ because $\sqrt{z^i_j}$ is real. Thus for $j < d+2$

$${A^i_j}^\dagger A^i_j = z^i_j |z^i_j\rangle\langle z^i_j|.$$

Now

$$\sum_{j=1}^{d+2} {A^i_j}^\dagger A^i_j = {A^i_{d+2}}^\dagger A^i_{d+2} + \sum_{j=1}^{d+1} z^i_j |z^i_j\rangle\langle z^i_j| = {A^i_{d+2}}^\dagger A^i_{d+2} + Z^i = \mathbb{I}_{d+1}.$$

Therefore $\{A^i_j\}$ is a set of Kraus matrices. □

Now we want to show that the probability of not reaching a goal state after taking actions $\{A^{i_1}, ..., A^{i_n}\}$ is the same as the probability of observing the sequence $\{i_1, ..., i_n\}$. However, before we can do that, we must take a short detour to show that the fact that the goal-state reachability problem is defined for state-dependent policies does not give it any advantage. Technically, a policy for a QOMDP is not time-dependent (specifying a sequence of actions) but state-dependent (specifying a *conditional* sequence of actions). QMOP is essentially time-dependent so this could be a potential problem. However, we have designed our goal QOMDP in such a way that, regardless of the policy and for any n, there is at most one non-goal state reachable after n time steps.

Lemma 6: Let $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ with Kraus matrices of dimension d be an instance of QMOP and let $Q(\mathcal{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ be the corresponding goal QOMDP. Let $\pi : S \to \mathcal{A}$ be any policy and let ρ_n be the state on the nth time step of following π. There is always at most one state $\sigma_n \neq \rho_g$ such that $\Pr(\sigma_n | \pi, n) > 0$.

Proof: We proceed by induction on n.

Base Case (n = 1): After 1 time step, we have applied a single action, $\pi(\rho_0)$. Lemma 3 gives us that there is only a single possible state besides ρ_g after the application of this action.

Induction Step: Assume that there are only two possible choices for ρ_{n-1}: σ_{n-1} and ρ_g. If $\rho_{n-1} = \rho_g$, then $\rho_n = \rho_g$ regardless of $\pi(\rho_g)$. If $\rho_{n-1} = \sigma_{n-1}$, action $\pi(\sigma_{n-1}) = A^{i_n}$ is taken. By Lemma 3 there is only a single possible state besides ρ_g after the application of A^{i_n}. □

Thus in a goal QOMDP created from a QMOP instance, the state-dependent policy π can be considered a "sequence of actions" by looking at the actions it will apply to each possible non-goal state in order.

Definition 5 (Policy Path): Let $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ with Kraus matrices of dimension d be a QMOP instance and let $Q(\mathcal{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ be the corresponding goal QOMDP. For any policy π let σ_k be the non-goal state with non-zero probability after k time steps of following π if it exists. Otherwise let $\sigma_k = \rho_g$. Choose $\sigma_0 = 0$. The set $\{\sigma_k\}$ is the *policy path* for policy π. By Lemma 6, this set is unique so this is well defined.

We have one more technical problem we need to address before we can look at how states evolve under policies in a goal QOMDP. When we created the goal QOMDP, we added a dimension to the problem so that we could have a defined goal state. We need to show that we can consider only the upper left $d \times d$ matrices when looking at evolution probabilities.

Lemma 7: Let $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ with Kraus matrices of dimension d be a QMOP instance and let $Q(\mathcal{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ be the corresponding goal QOMDP. Let M be any $d+1 \times d+1$ matrix and $d(M)$ be the upper left $d \times d$ matrix in which the $d+1$st column and row of M have been removed. Then for any action $A^i \in \mathcal{A}$,

$$A^i_{d+2} M {A^i_{d+2}}^\dagger = K_i d(M) K_i \oplus 0.$$

Proof: We consider the multiplication element-wise:

$$\left(A^i_{d+2} M {A^i_{d+2}}^\dagger\right)_{pq} = \sum_{h,l=1}^{d+1} {A^i_{d+2}}_{ph} M_{hl} {A^i_{d+2}}^\dagger_{lq} = \sum_{h,l=1}^{d} {A^i_{d+2}}_{ph} M_{hl} {A^i_{d+2}}^*_{ql}$$

where we have used that the $d+1$st column of A^i_{d+2} is zero to limit the sum. Additionally, if $p = d+1$ or

$q = d+1$, the sum is zero because the $d+1$st row of A^i_{d+2} is zero. Assume that $p < d+1$ and $q < d+1$. Then

$$\sum_{h,l=1}^{d} A^i_{d+2ph} M_{hl} {A^i_{d+2ql}}^* = \sum_{h,l=1}^{d} K_{iph} M_{hl} K^\dagger_{ilq} = \left(Kd(M)K^\dagger\right)_{ql}.$$

Thus

$$A^i_{d+2} M {A^i_{d+2}}^\dagger = K_i d(M) K_i^\dagger \oplus 0.$$

□

We are now ready to show that any path that does not terminate in the goal state in the goal QOMDP corresponds to some possible path through the superoperator in the QMOP instance.

Lemma 8: Let $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ with Kraus matrices of dimension d be a QMOP instance and let $Q(\mathcal{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ be the corresponding goal QOMDP. Let π be any policy for Q and let $\{\sigma_k\}$ be the policy path for π. Assume $\pi(\sigma_{k-1}) = A^{i_k}$. Then

$$\sigma_k = \frac{K_{i_k} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_k} \oplus 0}{\mathrm{Tr}\left(K_{i_k} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_k}\right)}.$$

Proof: We proceed by induction on k.

Base Case (k = 1): If $k = 1$ then we probabilistically apply either some $A^{i_1}_l$ with $l < d+2$ or $A^{i_1}_{d+2}$. In the first case, Lemma 3 gives us that the state becomes ρ_g. Therefore, σ_1 is the result of applying $A^{i_1}_{d+2}$ so

$$\sigma_1 = \frac{A^{i_1}_{d+2} \rho_0 {A^{i_1}_{d+2}}^\dagger}{\mathrm{Tr}\left(A^{i_1}_{d+2} \rho_0 {A^{i_1}_{d+2}}^\dagger\right)} = \frac{K_{i_1} d(\rho_0) K^\dagger_{i_1} \oplus 0}{\mathrm{Tr}\left(K_{i_1} d(\rho_0) K^\dagger_{i_1} \oplus 0\right)} = \frac{K_{i_1} d(\rho_0) K^\dagger_{i_1} \oplus 0}{\mathrm{Tr}\left(K_{i_1} d(\rho_0) K^\dagger_{i_1}\right)}$$

using Lemma 7 and the fact that $\mathrm{Tr}(A \oplus 0) = \mathrm{Tr}(A)$.

Induction Step: On time step k, we have $\rho_{k-1} = \sigma_{k-1}$ or $\rho_{k-1} = \rho_g$ by Lemma 6. If $\rho_{k-1} = \rho_g$ then $\rho_k = \rho_g$ by Lemma 4. Therefore, σ_k occurs only if $\rho_{k-1} = \sigma_{k-1}$. In this case we apply action A^{i_k}. If we apply $A^{i_k}_j$ with $j < d+2$, ρ_k is the goal state by Lemma 3. Therefore, we transition to σ_k exactly when $\rho_{k-1} = \sigma_{k-1}$ and we apply action $A^{i_k}_{d+2}$. By induction

$$\sigma_{k-1} = \frac{K_{i_{k-1}} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_{k-1}} \oplus 0}{\mathrm{Tr}\left(K_{i_{k-1}} ... K_{i_1} d(\rho_0) K^\dagger_1 ... K^\dagger_{i_{k-1}}\right)}.$$

Note that

$$d(\sigma_{k-1}) = \frac{K_{i_{k-1}} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_{k-1}}}{\mathrm{Tr}\left(K_{i_{k-1}} ... K_{i_1} d(\rho_0) K^\dagger_1 ... K^\dagger_{i_{k-1}}\right)}.$$

Then

$$\sigma_k = \frac{A^{i_k}_{d+2} \sigma_{k-1} A^{i_k}_{d+2}}{\mathrm{Tr}\left(A^{i_k}_{d+2} \sigma_{k-1} {A^{i_k}_{d+2}}^\dagger\right)} = \frac{K_{i_k} d(\sigma_{k-1}) K^\dagger_{i_k} \oplus 0}{\mathrm{Tr}\left(K_{i_k} d(\sigma_{k-1}) K^\dagger_{i_k}\right)}$$

using Lemma 7. Expanding this out, we have

$$\begin{aligned}\sigma_k &= \frac{K_{i_k}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_k}^\dagger \oplus 0}{\mathrm{Tr}\left(K_{i_{k-1}}...K_{i_1}d(\rho_0)K_1...K_{i_{k-1}}\right)}\frac{1}{\mathrm{Tr}\left(\frac{K_{i_k}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_k}^\dagger \oplus 0}{\mathrm{Tr}\left(K_{i_{k-1}}...K_{i_1}d(\rho_0)K_1...K_{i_{k-1}}\right)}\right)} \\ &= \frac{K_{i_k}...K_{i_1}d(\rho_0)K_{i_1}...K_{i_k} \oplus 0}{\mathrm{Tr}\left(K_{i_k}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_k}^\dagger\right)}.\end{aligned}$$

□

Now that we know how the state evolves, we can see that the probability the system is not in the goal state after taking actions $\left\{A^{i_1},...,A^{i_n}\right\}$ should correspond to the probability of observing measurements $\{i_1,...,i_n\}$ in the original QMOP instance.

Lemma 9: Let $\mathcal{S} = \{K_1,...,K_\mathcal{K}\}$ with Kraus matrices of dimension d be a QMOP instance and let $Q(\mathcal{S}) = \langle S,\Omega,\mathcal{A},\rho_0,\rho_g\rangle$ be the corresponding goal QOMDP. Let π be any policy and $\{\sigma_n\}$ be the state path for π. Assume $\pi(\sigma_{n-1}) = A^{i_n}$. The probability that ρ_n is not ρ_g is

$$\Pr\left(\rho_n \neq \rho_g\right) = \mathrm{Tr}\left(K_{i_n}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_n}^\dagger\right).$$

Proof: Firstly consider the probability that ρ_n is not ρ_g given that $\rho_{n-1} \neq \rho_g$. By Lemma 6, if $\rho_{n-1} \neq \rho_g$ then $\rho_{n-1} = \sigma_{n-1}$. By Lemma 8,

$$\sigma_{n-1} = \frac{K_{i_{n-1}}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_{n-1}}^\dagger \oplus 0}{\mathrm{Tr}\left(K_{i_{n-1}}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_{n-1}}^\dagger\right)}$$

so

$$d(\sigma_{n-1}) = \frac{K_{i_{n-1}}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_{n-1}}^\dagger}{\mathrm{Tr}\left(K_{i_{n-1}}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_{n-1}}^\dagger\right)}.$$

If $A_j^{i_n}$ for $j < d+2$ is applied then ρ_n will be ρ_g. Thus the probability that ρ_n is *not* ρ_g is the probability that $A_{d+2}^{i_n}$ is applied,

$$\begin{aligned}\Pr\left(\rho_n \neq \rho_g \,|\rho_{n-1} \neq \rho_g\right) &= \mathrm{Tr}\left(A_{d+2}^{i_n}\sigma_{n-1}{A_{d+2}^{i_n}}^\dagger\right) \\ &= \mathrm{Tr}\left(K_{i_n}d(\sigma_{n-1})K_{i_n}^\dagger \oplus 0\right) \\ &= \mathrm{Tr}\left(K_{i_n}d(\sigma_{n-1})K_{i_n}^\dagger\right) \\ &= \frac{\mathrm{Tr}\left(K_{i_n}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_n}^\dagger\right)}{\mathrm{Tr}\left(K_{i_{n-1}}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_{i-1}}^\dagger\right)}.\end{aligned}$$

Note that $\Pr(\rho_n \neq \rho_g \,|\rho_{n-1} = \rho_g) = 0$ by Lemma 4. The total probability that ρ_n is not ρ_g is

$$\begin{aligned}
\Pr(\rho_n \neq \rho_g) &= \Pr(\rho_n \neq \rho_g \cap \rho_{n-1} \neq \rho_g) + \Pr(\rho_n \neq \rho_g \cap \rho_{n-1} = \rho_g) \\
&= \Pr(\rho_n \neq \rho_g | \rho_{n-1} \neq \rho_g) \Pr(\rho_{n-1} \neq \rho_g) + \Pr(\rho_n \neq \rho_g | \rho_{n-1} = \rho_g) \Pr(\rho_{n-1} = \rho_g) \\
&= \Pr(\rho_n \neq \rho_g | \rho_{n-1} \neq \rho_g) \Pr(\rho_{n-1} \neq \rho_g | \rho_{n-2} \neq \rho_g) ... \Pr(\rho_1 \neq \rho_g | \rho_0 \neq \rho_g) \\
&= \prod_{k=1}^{n} \frac{\mathrm{Tr}\left(K_{i_k} ... K_{i_1} d(\rho_0) K_{i_1}^\dagger ... K_{i_k}^\dagger\right)}{\mathrm{Tr}\left(K_{i_{k-1}} ... K_{i_1} d(\rho_0) K_{i_1}^\dagger ... K_{i_{k-1}}^\dagger\right)} \\
&= \mathrm{Tr}\left(K_{i_n} ... K_{i_1} d(\rho_0) K_{i_1}^\dagger ... K_{i_n}^\dagger\right).
\end{aligned}$$

□

Since the probability that we observe the sequence of measurements $\{i_1, ..., i_n\}$ is the same as the probability that the sequence of actions $\{A^{i_1}, ..., A^{i_n}\}$ does not reach the goal state, we can solve QMOP by solving an instance of goal-state reachability for a QOMDP. Since QMOP is known to be undecidable, this proves that goal-state reachability is also undecidable for QOMDPs.

Theorem 10 (Undecidability of Goal-State Reachability for QOMDPs): The goal-state reachability problem for QOMDPs is undecidable.

Proof: We show we can reduce the quantum measurement occurrence problem to goal-state reachability for QOMDPs. Since QMOP is undecidable, this implies that goal-state reachability for QOMDPs is undecidable.

Let $\mathcal{S} = \{K_1, ..., K_\mathcal{K}\}$ be an instance of QMOP with Kraus matrices of dimension d and let $Q(\mathcal{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ be the corresponding goal QOMDP. By Theorem 5, $Q(\mathcal{S})$ is a goal QOMDP. We show that there is a policy that can reach ρ_g from ρ_0 with probability 1 if and only if there is some finite sequence $\{i_1, ..., i_n\}$ such that $K_{i_1}^\dagger ... K_{i_n}^\dagger K_{i_n} ... K_{i_1} = 0$.

Firstly assume there is some sequence $\{i_1, ..., i_n\}$ such that $K_{i_1}^\dagger ... K_{i_n}^\dagger K_{i_n} ... K_{i_1} = 0$. Consider the time-dependent policy that takes action A^{i_k} in after k time steps no matter the state. By Lemma 9, the probability that this policy is not in the goal state after n time steps is

$$\Pr(\rho_n \neq \rho_g) = \mathrm{Tr}(K_{i_n} ... K_{i_1} d(\rho_0) K_{i_1}^\dagger ... K_{i_n}^\dagger) = \mathrm{Tr}(K_{i_1}^\dagger ... K_{i_n}^\dagger K_{i_n} ... K_{i_1} d(\rho_0)) = \mathrm{Tr}(0) = 0.$$

Therefore this policy reaches the goal state with probability 1 after n time steps. As we have said, time cannot help goal decision processes since nothing changes with time. Therefore, there is also a purely state dependent policy (namely the one that assigns A^{i_k} to σ_k where σ_k is the kth state reached when following π) that can reach the goal state with probability 1.

Now assume there is some policy π that reaches the goal state with probability 1 after n time steps. Let $\{\sigma_k\}$ be the policy path and assume $\pi(\sigma_{k-1}) = A^{i_k}$. By Lemma 9, the probability that the state at time step n is not ρ_g is

$$\Pr(\rho_n \neq \rho_g | \pi) = \mathrm{Tr}(K_{i_1} ... K_{i_n} d(\rho_0) K_{i_1}^\dagger ... K_{i_n}^\dagger) = \mathrm{Tr}(K_{i_1}^\dagger ... K_{i_n}^\dagger K_{i_1} ... K_{i_n} d(\rho_0)).$$

Since π reaches the goal state with probability 1 after n time steps, we must have that this quantity is zero. By construction $d(\rho_0)$ is full rank so for the trace to be zero we must have

$$K_{i_1}^\dagger ... K_{i_n}^\dagger K_{i_1} ... K_{i_n} = 0.$$

Thus we can reduce the quantum measurement occurrence problem to the goal-state reachability problem for QOMDPs and the goal-state reachability problem is undecidable for QOMDPs. □

4.2 Decidability of Goal-State Reachability for POMDPs

The goal-state reachability problem for POMDPs is decidable. This is a known result [6] but we reproduce the proof here because it is interesting to note the differences in classical probability that lead to the decidability of the problem.

At a high level, the goal-state reachability problem is decidable for POMDPs because stochastic transition matrices have strictly non-negative elements. Since we are interested in a probability 1 event, we can treat probabilities as binary: either greater than zero or equal to zero. This gives us a belief space with $2^{|S|}$ states rather than a continuous one and we can show that the goal-state reachability problem is decidable for finite state spaces. We begin by formalizing the binary probability transformation.

Definition 6 (Binary Probability MDP): Given a goal POMDP $P = \left\langle S, A, \Omega, T, O, \vec{b}_0, g \right\rangle$, let $M(P) = \left\langle B, A, \tau, \vec{b}_0, \vec{b}_g \right\rangle$ be the corresponding goal belief MDP with τ^{ao} defined according to Equation 3. Throughout this section, we assume WLOG that g is the $|S|$th state in P so $\left(\vec{b}_g\right)_i = \delta_{i|S|}$. The *binary probability MDP* is an MDP $D(P) = \left\langle \mathbb{Z}_2^{|S|}, A, Z, \vec{z}_0, \vec{z}_g \right\rangle$ where $(\vec{z}_g)_i = \delta_{i|S|}$ and $(\vec{z}_0)_i = 1$ if and only if $\left(\vec{b}_0\right)_i > 0$. The transition function Z for action a non-deterministically applies the function Z^{ao} to $\vec{z}$. For $\vec{z} \in \mathbb{Z}_2^{|S|}$, the result of Z^{ao} acting on $\vec{z}$ is

$$Z^{ao}(\vec{z})_i = \begin{cases} 1 & \text{if } (\tau^{ao}\vec{z})_i > 0 \\ 0 & \text{if } (\tau^{ao}\vec{z})_i = 0. \end{cases}$$

Let

$$P_a^o(\vec{z}) = \begin{cases} 1 & \text{if } \tau^{ao}\vec{z} \neq \vec{0} \\ 0 & \text{else.} \end{cases}$$

If action a is taken in state $\vec{z}$, Z^{ao} is applied with probability

$$\Pr\left(Z^{ao}|a, \vec{z}\right) = \begin{cases} \frac{1}{\sum_{o' \in \Omega} P_a^{o'}(\vec{z})} & \text{if } P_o^a(\vec{z}) > 0 \\ 0 & \text{else.} \end{cases}$$

Note that the vector of all zeros is unreachable so the state space is really size $2^{|S|} - 1$.

We first show that we can keep track of the sign of the belief state just using the binary probability MDP. This lemma uses the fact that classical probability involves strictly non-negative numbers, which is not true of quantum evolution.

Lemma 11: Let $P = \left\langle S, A, \Omega, T, O, \vec{b}_0, g \right\rangle$ be a goal-state POMDP and let $D(P) = \left\langle \mathbb{Z}_2^{|S|}, A, Z, \vec{z}_0, \vec{z}_g \right\rangle$ be the associated binary probability MDP. Assume we have $\vec{z}$ and $\vec{b}$ where $\vec{z}_i = 0$ if and only if $\vec{b}_i = 0$. Let

$$\vec{z}^{ao} = Z^{ao}(\vec{z})$$

and

$$\vec{b}^{ao} = \frac{\tau^{ao}\vec{b}}{\left|\tau^{ao}\vec{b}\right|_1}.$$

Then $\vec{z}_i^{ao} = 0$ if and only if $\vec{b}_i^{ao} = 0$. Moreover, $P_a^o(\vec{z}) = 0$ if and only if $\left|\tau^{ao}\vec{b}\right|_1 = 0$.

Proof: Using the definition of Z^{ao},

$$\vec{z}_i^{ao} = Z^{ao}(\vec{z})_i = \begin{cases} 1 & \text{if } (\tau^{ao}\vec{z})_i > 0 \\ 0 & \text{else.} \end{cases}$$

Let $N = \left|\tau^{ao}\vec{b}\right|_1$. Then

$$\vec{b}_i^{ao} = \frac{1}{N}\sum_{j=1}^{|S|} \tau_{ij}^{ao}\vec{b}_j. \tag{12}$$

Firstly assume $\vec{b}_i^{ao} = 0$. Since $\tau_{ij}^{ao} \geq 0$ and $\vec{b}_j \geq 0$, we must have that every term in the sum in Equation 12 is zero individually[3]. Therefore, for all j, either $\tau_{ij}^{ao} = 0$ or $\vec{b}_j = 0$. If $\vec{b}_j = 0$ then $\vec{z}_j = 0$ so $\tau_{ij}^{ao}\vec{z}_j = 0$. If $\tau_{ij}^{ao} = 0$ then clearly $\tau_{ij}^{ao}\vec{z}_j = 0$. Therefore

$$0 = \sum_{j=1}^{|S|} \tau_{ij}^{ao}\vec{z}_j = \left(\tau_{ij}^{ao}\vec{z}\right)_i = \vec{z}_i^{ao}.$$

Now assume $\vec{b}_i^{ao} > 0$. Then there must be at least one term in the sum in Equation 12 with $\tau_{ik}^{ao}\vec{b}_k > 0$. In this case, we must have both $\tau_{ik}^{ao} > 0$ and $\vec{b}_k > 0$. If $\vec{b}_k > 0$ then $\vec{z}_k > 0$. Therefore[4]

$$\vec{z}_i^{ao} = (\tau^{ao}\vec{z})_i = \sum_{j=1}^{|S|} \tau_{ij}^{ao}\vec{z}_j = \sum_{j\neq k} \tau_{ij}^{ao}\vec{z}_j + \tau_{ik}^{ao}\vec{z}_k > 0.$$

Now assume $\left|\tau^{ao}\vec{b}\right|_1 = 0$. This is true only if $\tau_{ij}^{ao}\vec{b}_j = 0$ for all j. Thus by the same reasoning as above $\tau_{ij}^{ao}\vec{z}_j = 0$ for all j so $\tau^{ao}\vec{z} = \vec{0}$ and $P_a^o(\vec{z}) = 0$.

Now let $\left|\tau^{ao}\vec{b}\right|_1 > 0$. Then there is some k with $\tau_{ik}^{ao}\vec{z}_k > 0$ by the same reasoning as above. Therefore $\tau^{ao}\vec{z} \neq \vec{0}$ so $P_a^o(\vec{z}) = 1$. □

We now show that we can reach the goal in the binary probability MDP with probability 1 if and only if we could reach the goal in the original POMDP with probability 1. Because this is a long proof, we do each direction in a separate lemma.

Lemma 12: Let $P = \left\langle S, A, \Omega, T, O, \vec{b}_0, g\right\rangle$ be a goal POMDP and let $D(P) = \left\langle \mathbb{Z}_2^{|S|}, A, Z, \vec{z}_0, \vec{z}_g\right\rangle$ be the corresponding binary probability MDP. If there is a policy π^D that reaches the goal with probability 1 in a finite number of steps in $D(M)$ then there is a policy that reaches the goal in a finite number of steps with probability 1 in the belief MDP $M(P) = \left\langle B, A, \tau, \vec{b}_0, \vec{b}_g\right\rangle$.

Proof: For $\vec{b} \in B$ define $z(\vec{b})$ to be the single state $\vec{z} \in \mathbb{Z}_2^n$ with $\vec{z}_i = 0$ if and only if $\vec{b}_i = 0$. Let π be the policy for $M(P)$ with $\pi(\vec{b}) = \pi^D(z(\vec{b}))$. Let $\vec{b}^0, \vec{b}^1, ..., \vec{b}^n$ be some branch of length $n+1$ that can be created by following policy π with observations $\{o_{i_1}, ..., o_{i_n}\}$. Then

$$\vec{b}^{k+1} = \frac{\tau^{\pi(\vec{b}^k)o_{i_k}}\vec{b}^k}{\left|\tau^{\pi(\vec{b}^k)o_{i_k}}\vec{b}^k\right|_1} = \frac{\tau^{\pi^D(z(\vec{b}^k))o_{i_k}}\vec{b}^k}{\left|\tau^{\pi^D(z(\vec{b}^k))o_{i_k}}\vec{b}^k\right|_1}.$$

[3]This holds because we know probabilities are strictly non-negative. A similar analysis in the quantum case would fail at this step.

[4]Again we use that probability is only positive so having a single positive number in a sum means the sum is greater than zero. If this was a quantum analysis, guaranteeing positivity here would be more difficult.

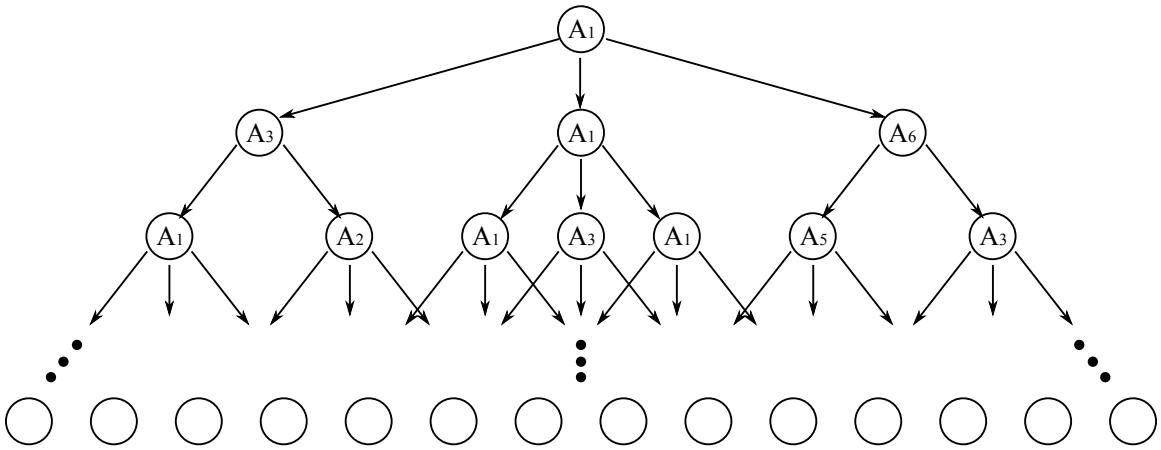

Figure 5: A policy in an MDP creates a tree (this is not to be confused with the policy tree of a POMDP). Here, we take action A_1 in the starting state, which can transition us non-deterministically to three other possible states. The policy specifies an action of A_3 for the state on the left, A_1 for the state in the middle and A_6 for the state on the right. Taking these actions transition these states non-deterministically. So this tree eventually encapsulates all states that can be reached with non-zero probability from the starting state under a particular policy. The goal can be reached with probability 1 if there is some depth below which every node is the goal state.

Define $a_k = \pi^D(z(\vec{b}^k))$. Consider the set of states $\vec{z}^0, \vec{z}^1, ..., \vec{z}^n$ with $\vec{z}^{k+1} = Z^{\pi^D(\vec{z}^k)o_{i_k}}\left(\vec{z}^k\right)$. We show by induction that $\vec{z}^k = z(\vec{b}^k)$.

Base Case (k = 0): We have that $\vec{z}^0 = z(\vec{b}^0)$ by definition.

Induction Step: Assume that $\vec{z}^k = z(\vec{b}^k)$. Then

$$\vec{z}^{k+1} = Z^{\pi^D(\vec{z}^k)o_{i_k}}(\vec{z}^k) = Z^{\pi^D(z(\vec{b}^k))o_{i_k}}(\vec{z}^k) = Z^{a_k o_{i_k}}(\vec{z}^k)$$

by induction. Now

$$\vec{b}^{k+1} = \frac{\tau^{a_{i_k} o_{i_k}} \vec{b}^k}{\left|\tau^{a_{i_k} o_{i_k}} \vec{b}^k\right|_1}.$$

Therefore $\vec{z}^{k+1} = z(\vec{b}^{k+1})$ by Lemma 11.

We must also show that the sequence $\vec{z}^0, \vec{z}^1, ..., \vec{z}^n$ has non-zero probability of occurring while following π^D. We must have that $P_{a_k}^{o_{i_k}} > 0$ for all k. We know that $\vec{b}^0, \vec{b}^1, ..., \vec{b}^n$ can be created by following π so the probability of $\vec{b}^0, \vec{b}^1, ..., \vec{b}^n$ is greater than zero. Therefore, we must have that

$$\Pr(o|a_k, \vec{b}^k) = \left|\tau^{a_k o_{i_k}} \vec{b}^k\right|_1 > 0$$

so Lemma 11 gives us that $P_{a_k}^{o_{i_k}} > 0$. Thus $\{\vec{z}^0, ..., \vec{z}^n\}$ is a possible branch of policy π^D. Since this policy reaches the goal state with probability 1 after n time steps, we have that $\vec{z}^n = \vec{z}_g$. Therefore, since $\vec{z}^n = z(\vec{b}^n)$, we must have $\vec{b}_i^n = 0$ if $i \neq |S|$ and only $\vec{b}_{|S|}^n > 0$. Since $|\vec{b}^n|_1 = 1$, we have $\vec{b}_{|S|}^n = 1$. Thus $\vec{b}^n = \vec{b}_g$ and π also reaches the goal state with non-zero probability after n time steps.

□

Lemma 13: Let $P = \left\langle S, A, \Omega, T, O, \vec{b}_0, g \right\rangle$ be a goal POMDP and let $D(P) = \left\langle \mathbb{Z}_2^{|S|}, A, Z, \vec{z}_0, \vec{z}_g \right\rangle$ be the corresponding binary probability MDP. If there is a policy π that reaches the goal with probability 1 in a finite number of steps in the belief state MDP $B(M) = \left\langle B, A, \tau, \vec{b}_0, \vec{b}_g \right\rangle$ then there is a policy that reaches the goal in a finite number of steps with probability 1 in $D(P)$.

Proof: MDP policies create trees of states and action choices as shown in Figure 5. Consider the tree π_T formed by π. Nodes at depth n or greater are guaranteed to be $\vec{b}_g$. For $\vec{z} \in \mathbb{Z}_2^{|S|}$, we let $b(\vec{z})$ be the first state $\vec{b} \in \pi_T$ reached when traversing π_T in depth-first order with $\vec{b}_i = 0$ if and only if $\vec{z}_i = 0$. If no such state is found in π_T, we set $b(\vec{z}) = \vec{b}_g$. We define a policy π^D for $D(P)$ by $\pi^D(\vec{z}) = \pi(b(\vec{z}))$. Let $\vec{z}^0, \vec{z}^1, ..., \vec{z}^n$ be a branch that can be created by following policy π^D for n time steps. Define $a_k = \pi^D(\vec{z}^k)$ and define i_k as the smallest number such that $\vec{z}^{k+1} = Z^{a_k o_{i_k}}(\vec{z}^k)$ (some such $Z^{a_k o_{i_k}}$ exists since $\vec{z}^0, ..., \vec{z}^n$ is a branch of π^D). Now consider $b(\vec{z}^k)$. We show by induction that this state is at least at level k of π_T.

Base Case (k = 0): We know that $\vec{b}_i^0 = 0$ if and only if $\vec{z}_i^0 = 0$ so $b(\vec{z}^0)$ is at least at level 0 of π_T.

Induction Step: Assume that $\vec{z}^k$ is at least at level k of π_T. Then

$$\vec{z}^{k+1} = Z^{a_k o_{i_k}}\left(\vec{z}^k\right).$$

Therefore by Lemma 11,

$$\vec{b}' = \frac{\tau^{a_k o_{i_k}} b(\vec{z}^k)}{|\tau^{a_k o_{i_k}} b(\vec{z}^k)|_1}$$

has entry i 0 if and only if $\vec{z}_i^{k+1} = 0$. Now $P_{o_k}^{a_k}(\vec{z}^k) \neq 0$ only if $|\tau^{a_k o_{i_k}} b(\vec{z}^k)|_1 \neq 0$ also by Lemma 11. Since $\vec{z}^1, ..., \vec{z}^n$ is a branch of π^D, we must have $P_{o_k}^{a_k} > 0$. Therefore $|\tau^{a_k o_{i_k}} b(\vec{z}^k)|_1 > 0$. Now $a_k = \pi(b(\vec{z}^k))$ so $\vec{b}'$ is a child of $b(\vec{z}^k)$ in π_T. Since, by induction, the level of $b(\vec{z}^k)$ is at least k, the level of $\vec{b}'$ is at least $k+1$. Now $\vec{b} = b(\vec{z}^{k+1})$ is the first state on the depth-first traversal with $\vec{b}_i = 0$ if and only if $\vec{z}_i^{k+1} = 0$ so level of $b(\vec{z}^{k+1})$ is at least the level of $\vec{b}'$. Therefore $b(\vec{z}^{k+1})$ has level at least $k+1$.

Thus the level of $b(\vec{z}^n)$ is at least n. We have $b(\vec{z}^n) = \vec{b}_g$ since π reaches the goal state in no more than n steps. Since $b(\vec{z}^n)_i = \delta_{i|S|}$, we have that $\vec{z}^n = \vec{z}_g$. Therefore π^D is a policy for $D(P)$ that reaches the goal with probability 1. □

We have now reduced goal-state reachability for POMDPs to goal-state reachability for finite state MDPs. We briefly show that this is a decidable problem.

Theorem 14 (Decidability of Goal-State Reachability for POMDPs): The goal-state reachability problem for POMDPs is decidable.

Proof: We showed in Lemmas 12 and 13 that goal-state reachability for POMDPs can be reduced to goal-state reachability for a finite state MDP. Therefore, there are only $O(|A||S|)$ possible policies (remember that for goal decision processes, we need only consider time independent policies). Given a policy π, we can evaluate it by creating a directed graph G in which we connect state s_i to state s_j if $\tau(s_i, \pi(s_i), s_j) > 0$. The policy π reaches the goal from the starting state with probability 1 if the goal is reachable from the starting state in G and no cycle is reachable. Since the graph has finitely many nodes, we can clearly decide this problem. Thus goal-state reachability is decidable for POMDPs. □

4.3 Other Complexity Separations

Although the goal-state reachability problem is the only complexity separation we have formally proved, we conjecture that there are a number of similar problems that are undecidable for QOMDPs while

decidable for POMDPs.

For instance, the finite policy value problem, in which we decide whether a goal QOMDP has a finite value (or an infinitely negative value) is very likely undecidable. This is close to the goal-state reachability problem, but allows a small "leak". For instance, if we have a probability of 0.9 of transitioning to the goal on every time step, then the value is finite even though we will never reach the goal with probability 1. Eisert et al. show that the small-probability quantum measurement occurrence problem is undecidable so a similar reduction to the one employed here will likely give the desired result for QOMDPs. It is unclear, however, whether this problem is decidable for POMDPs.

The zero-reward policy problem is also another likely candidate for complexity separation. In this problem, we still have a goal QOMDP(POMDP) but states other than the goal state are allowed to have zero reward. The problem is to decide whether the path to the goal state is zero reward. This is known to be decidable for POMDPs, but seems unlikely to be so for QOMDPs.

5 Future Work

We outlined in Section 4.3 a set of decision problems we expect might show similar complexity results. There are also a number of other unanswered questions about QOMDPs.

The biggest lack in this paper was that of a "true" POMDP problem. In the complexity analysis, we were only able to give an interesting result for a problem on goal decision processes, which ignore the reward function. Understanding the class of reward functions for a QOMDP, perhaps with an eye towards making them useful to some quantum application, would be an interesting direction for future research.

We also proved complexity results, but did not consider algorithms for solving any of the problems we posed beyond a very simple PSPACE algorithm for policy existence. Is there a quantum analog to Bellman's equation? Again, this requires a better understanding of the quantum reward function than we achieved in this work.

Another striking difference between a POMDP belief space and a QOMDP is that there is no inherent reason why we should know the starting state in the QOMDP. In our current formulation, the QOMDP is observable - the observations tell us exactly the current state. If we were not given the starting state or the superoperator's observations were somehow hidden from us, this may lead to a very different class of problems.

Lastly, because POMDPs are both so difficult and so relevant to planning, there are many algorithms for approximating them and many fewer lower bounds proved for complexity. Some of these algorithms focus on Monte Carlo sampling techniques, for which it seems likely a quantum computer could provide an efficiency gain.

References

[1] Robert B. Griffiths. Quantum Channels, Kraus Operators, POVMs. Quantum Computation and Quantum Information Theory Course Notes, Carnegie Mellon University, Spring 2010.

[2] J. Eisert and M. P. Mueller and C. Gogolin. Quantum Measurement Occurrence is Undecidable. *Physical Review Letters*, 108, 2012.

[3] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and Acting in Partially Observable Stochastic Domains. *Artificial Intelligence*, 101:99–134, 1998.

[4] O. Madani, S. Hanks, and A. Codon. On the Undecidability of Probabilistic Planning and Partially Observable Markov Decision Problems. In *Association for the Advancement of Artificial Intelligence*, 1999.

[5] Christos H. Papadimitriou and John N. Tsitsiklis. The Complexity of Markov Decision Processes. *Mathematics of Operations Research*, 12(3):441–450, August 1987.

[6] Jussi Rintanen. Complexity of Planning with Partial Observability. In *International Conference on Automated Planning and Scheduling*, pages 345–354, 2004.

[7] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Hall, New Jersey, second edition, 2003. Chapter 17.