Quantum POMDPs

Jennifer Barry,[1] Daniel T. Barry,[2] and Scott Aaronson[3]
[1] *Rethink Robotics*[*]
[2] *Denbar Robotics*[†]
[3] *MIT CSAIL*[‡]

We present quantum observable Markov decision processes (QOMDPs), the quantum analogues of partially observable Markov decision processes (POMDPs). In a QOMDP, an agent's state is represented as a quantum state 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 goal POMDPs and undecidable for goal QOMDPs.

I. INTRODUCTION

Partially observable Markov decision processes (POMDPs) are a world model commonly used in artificial intelligence [1–5]. POMDPs model an agent acting in a world of discrete states. The agent is not told its state, but it can take actions and receive observations about the world. The actions it takes are non-deterministic; before taking an action, the agent knows only the probability distribution of its next state given its current state. Similarly, an observation does not give the agent direct knowledge of its current state, but the agent knows the probability of receiving a given observation in each possible state. The agent is rewarded for its actual, hidden state at each time step, but, although it knows the reward model, it is not told the reward it received. POMDPs are often used to model robots, because robot sensors and actuators give them a very limited understanding of their environment.

As we will discuss further in Section II, we can maximize future expected reward in a POMDP by maintaining a probability distribution, known as a belief state, over the agent's current state. By carefully updating this belief state after every action and observation, we can ensure that it reflects the correct probability that the agent is in each world 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.

In this paper, we introduce and study "quantum observable Markov decision processes" (QOMDPs). A QOMDP is similar in spirit to a POMDP, but allows the belief state to be a quantum state (superposition or mixed state) rather than a simple probability distribution. We represent the action and observation process jointly as a superoperator. POMDPs are then just the special case of QOMDPs where the quantum state is always diagonal in some fixed basis.

Although QOMDPs are the quantum analogue of POMDPs, they have different computability properties. Our main result, in this paper, is that there exists a decision problem (namely, goal state reachability) that is computable for POMDPs but uncomputable for QOMDPs.

One motivation for studying QOMDPs is simply that they're the natural quantum generalizations of POMDPs, which are central objects of study in AI. Moreover, as we show here, QOMDPs have *different* computability properties than POMDPs, so the generalization is not an empty one. Beyond this conceptual motivation, though, QOMDPs might also find applications in quantum control and quantum fault-tolerance. For example, the general problem of controlling a noisy quantum system, given a discrete "library" of noisy gates and measurements, in order to manipulate the system to a desired end state, can be formulated as a QOMDP. Indeed, the very fact that POMDPs have turned out to be such a useful abstraction for modeling *classical* robots, suggests that QOMDPs would likewise be useful for modeling control systems that operate at the quantum scale. At any rate, this seems like sufficient reason to investigate the complexity and computability properties of QOMDPs, yet we know of no previous work in that direction. This paper represents a first step.

II. PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES (POMDPS)

For completeness, in this section we give an overview of Markov decision processes and POMDPs.

A. Fully Observable Case

We begin by defining fully observable Markov decision processes (MDPs). This will facilitate our discussion of POMDPs because POMDPs can be reduced to continuous-state MDPs. For more details, see Russell and Norvig, Chapter 17 [3].

[*] jbarry@csail.mit.edu
[†] dbarry@denbarrobotics.com
[‡] aaronson@csail.mit.edu

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 $T(s_i, a, s_j) : S \times A \times S \rightarrow [0, 1]$ giving the probability that taking action a in state s_i results in state s_j, a reward function $R(s_i, a) : S \times A \rightarrow \mathbb{R}$ 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. 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.

The solution to an MDP is a policy. A *policy* $\pi(s_i, t) : S \times \mathbb{Z}^+ \rightarrow A$ 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:

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

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 D, a starting state s, horizon h, and value V, whether there is a policy of horizon h that achieves value at least V for s in D.

For MDPs, we will evaluate the infinite horizon case. In this case, we will drop the time argument from the policy since it does not 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

$$V^*(s_i) = \max_{a \in A} \left[R(s_i, a) + \gamma \sum_{s_j \in S} T(s_i, a, s_j) V^*(s_j) \right] . \tag{2}$$

Equation 2 is called the *Bellman equation*, and there is a unique solution for V^* [3]. Note that V^* is non-infinite if $\gamma < 1$. When the input size is polynomial in $|S|$ and $|A|$, finding an ϵ-optimal policy for an MDP can be done in polynomial time [3].

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 D and starting state s, whether there exists a policy that can reach the goal state from s in a finite number of steps with probability 1.

Note that Definition 2 is only about reaching the goal in a finite number of steps. We are not considering the problem of reaching the goal with probability 1 in an infinite number of steps here, although we will discuss it briefly in Section IV C.

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 solely upon the current state.

B. Partially Observable Case

A partially observable Markov decision process (POMDP) generalizes an MDP to the case where the world is not fully observable. We follow the work of Kaelbling et al. [1] 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 $\langle S, A, \Omega, T, R, O, \vec{b}_0, \gamma \rangle$ where S is a set of states, A is a set of actions, Ω is a set of observations, $T(s_i, a, s_j) : S \times A \times S \rightarrow [0, 1]$ is the probability of transitioning to state s_j given that action a was taken in state s_i, $R(s_i, a) : S \times A \rightarrow \mathbb{R}$ is the reward for taking action a in state s_i, $O(s_j, a, o) : S \times A \times \Omega \rightarrow [0, 1]$ 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 state before the transition and T, and receives an observation according to its hidden state after the transition and O. As with MDPs, the goal is to maximize future expected reward.

POMDPs induce a *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 $\sum_i \vec{b}_i = 1$. If the agent is in belief state $\vec{b}$, takes action a, and receives observation o the new belief state is

$$\begin{aligned} \vec{b'}_i &= \Pr(s_i | o, a, \vec{b}) \\ &= \frac{\Pr\big(o | s_i, a, \vec{b}\big) \Pr\big(s_i | a, \vec{b}\big)}{\Pr\big(o | a, \vec{b}\big)} \\ &= \frac{O(s_i, a, o) \sum_j T(s_j, a, s_i) \vec{b}_j}{\Pr\big(o | a, \vec{b}\big)} . \end{aligned} \tag{3}$$

This is the belief update equation. $\Pr(o|a,\vec{b}) = \sum_k O(s_k,a,o)\sum_j T(s_j,a,s_k)\vec{b}_j$ is independent of i and usually just computed afterwards as a normalizing factor that causes $\vec{b}'$ to sum to 1. We define the matrix

$$(\tau^{ao})_{ij} = O(s_i,a,o)T(s_j,a,s_i). \tag{4}$$

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

$$\vec{b}' = \frac{\tau^{ao}\vec{b}}{\left|\tau^{ao}\vec{b}\right|_1} \tag{5}$$

where $|\vec{v}|_1 = \sum_i \vec{v}_i$ is the L_1-norm. The probability of transitioning from belief state $\vec{b}$ to belief state $\vec{b}'$ when taking action a is

$$\tau(\vec{b},a,\vec{b}') = \sum_{o\in\Omega}\Pr(\vec{b}'|a,\vec{b},o)\Pr(o|a,\vec{b}) \tag{6}$$

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

$$r(\vec{b},a) = \sum_i \vec{b}_i R(s_i,a). \tag{7}$$

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. It was shown in 1987 that the policy existence problem for POMDPs is PSPACE-hard [6]. If the horizon is polynomial in the size of the input, the policy existence problem is in PSPACE [1]. The policy existence problem for POMDPs in the infinite horizon case, however, is undecidable [7].

A *goal POMDP* is a tuple $P = \langle S,A,\Omega,T,O,\vec{b}_0,g\rangle$ 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 $\vec{b}_0$.

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

$$\begin{aligned}\Pr(o|a,\vec{b}_g) &= \sum_j O(s_j,a,o)\sum_i T(s_i,a,s_j)(\vec{b}_g)_i \\ &= \sum_j O(s_j,a,o)T(g,a,s_j) \\ &= O(g,a,o) \\ &= \delta_{oo_{|\Omega|}}.\end{aligned}$$

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

$$\begin{aligned}\vec{b'}_j &= \Pr(s_j|o_{|\Omega|},a,\vec{b}_g) \\ &= \frac{O(s_j,a,o_{|\Omega|})\sum_i T(s_i,a,s_j)\vec{b}_i}{\Pr(o_{|\Omega|}|a,\vec{b}_g)} \\ &= O(s_j,a,o_{|\Omega|})T(g,a,s_j) \\ &= \delta_{gs_j}.\end{aligned}$$

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

III. QUANTUM OBSERVABLE MARKOV DECISION PROCESSES (QOMDPS)

A quantum observable Markov decision process (QOMDP) generalizes a POMDP by using quantum states rather than belief states. In a QOMDP, an agent can apply a set of possible operations to a d-dimensional quantum state. The operations each have $\mathcal{K}$ possible outcomes. At each time step, the agent receives an observation corresponding to the outcome of the previous operation and can choose another operation to apply. The reward the agent receives is the expected value of some operator in the agent's current quantum state.

A. QOMDP Formulation

A QOMDP uses superoperators to express both actions and observations. A quantum superoperator $\mathbf{S} = \{K_1,...,K_\mathcal{K}\}$ acting on states of dimension d is defined by $\mathcal{K}$ $d\times d$ Kraus matrices [8] [9]. A set of matrices $\{K_1,...,K_\mathcal{K}\}$ of dimension d is a set of Kraus matrices if and only if

$$\sum_{i=1}^{\mathcal{K}} K_i^\dagger K_i = \mathbb{I}_d. \tag{8}$$

If $\mathbf{S}$ operates on a density matrix ρ, there are $\mathcal{K}$ possible next states for ρ. Specifically the next state is

$$\rho_i' \to \frac{K_i \rho K_i^\dagger}{\mathrm{Tr}(K_i \rho K_i^\dagger)} \tag{9}$$

with probability

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

The superoperator returns observation i if the i^{th} Kraus matrix was applied.

We can now define the quantum observable Markov decision process (QOMDP). A QOMDP is a tuple $\langle S, \Omega, \mathcal{A}, \mathcal{R}, \gamma, \rho_0 \rangle$ 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 $A^a = \{A_1^a, ..., A_{|\Omega|}^a\}$ 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 so taking action a in state ρ returns observation o_i with probability

$$\Pr(o_i|\rho, a) = \mathrm{Tr}\left(A_i^a \rho A_i^{a\dagger}\right). \tag{11}$$

  If o_i is observed after taking action a in state ρ, the next state is

$$N(\rho, a, o_i) = \frac{A_i^a \rho A_i^{a\dagger}}{\mathrm{Tr}\left(A_i^a \rho A_i^{a\dagger}\right)}. \tag{12}$$

- $\mathcal{R} = \{R_1, ..., R_{|\mathcal{A}|}\}$ is a set of operators. The reward associated with taking action a in state ρ is the expected value of operator R_a on ρ,

$$R(\rho, a) = \mathrm{Tr}(\rho R_a). \tag{13}$$

- $\gamma \in [0, 1)$ 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 agent's goal is to maximize its 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 received.

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

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

Let π_h be the policy at time h. Then

$$\begin{aligned} &V^{\pi_h}(\rho_0) = R(\rho_0, \pi_h(\rho_0)) + \\ &\quad \gamma \sum_{i=1}^{|\Omega|} \Pr(o_i|\rho_0, \pi_h(\rho_0)) V^{\pi_{h-1}}(N(\rho_0, \pi_h(\rho_0), o_i)) \end{aligned} \tag{14}$$

where $\Pr(o_i|\rho_0, \pi_h(\rho_0))$, $N(\rho_0, \pi_h(\rho_0), o_i)$, and $R(\rho_0, \pi_h(\rho_0))$ are defined by equations 11, 12, and 13 respectively. The Bellman equation (equation 2) still holds using these definitions.

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 $A_j^i \in A^i$ if $\mathrm{Tr}(A_j^i \rho_g A_j^{i\dagger}) > 0$ then

$$\frac{A_j^i \rho_g A_j^{i\dagger}}{\mathrm{Tr}(A_j^i \rho_g A_j^{i\dagger})} = \rho_g.$$

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

B. QOMDP Policy Existence Complexity

As we can always simulate classical evolution with a quantum system, the definition of QOMDPs contains POMDPs. Therefore we immediately find that the policy existence problem for QOMDPs in the infinite horizon case is undecidable. We also find that the polynomial horizon case is PSPACE-hard. 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 $\langle S, \Omega, \mathcal{A}, \mathcal{R}, \gamma, \rho_0 \rangle$ 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 $O((|\mathcal{A}||\Omega|)^h)$ possible states giving us $O(h\,|\mathcal{A}|\,(|\mathcal{A}||\Omega|)^h)$ 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 using only polynomial space. □

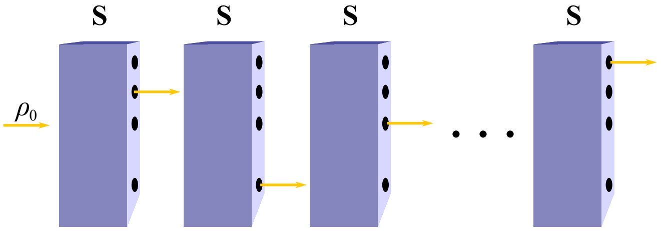

FIG. 1. The quantum measurement occurrence problem. The starting state ρ_0 is fed into the superoperator $\mathbf{S}$. The output is then fed iteratively back into $\mathbf{S}$. The question is whether there is some finite sequence of observations that can never occur.

IV. A COMPUTABILITY 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 (Definition 2) is decidable for goal POMDPs but undecidable for goal QOMDPs.

A. 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. [10] 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 $\mathbf{S} = \{K_1, ..., K_{\mathcal{K}}\}$, whether there is some finite sequence $\{i_1, ..., i_n\}$ such that $K_{i_1}^\dagger ... K_{i_n}^\dagger K_{i_n} ... K_{i_1} = 0$.

The setting for this problem is shown in Figure 1. We assume that the system starts in state ρ_0. This state is fed into $\mathbf{S}$. We then take the output of $\mathbf{S}$ acting on ρ_0 and feed that again into $\mathbf{S}$ and iterate. QMOP is equivalent to asking whether there is some finite sequence of observations $\{i_1, ..., i_n\}$ that can never occur 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 to provide 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 [10]. □

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 2.

Definition 4 (QMOP Goal QOMDP): Given an instance of QMOP with superoperator $\mathbf{S} = \{K_1, ..., K_{\mathcal{K}}\}$ and Kraus matrices of dimension d, we create a goal QOMDP $Q(\mathbf{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ 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.
- $\mathcal{A} = \{A^1, ..., A^{\mathcal{K}}\}$ is a set of $\mathcal{K}$ superoperators each with $d+2$ Kraus matrices $A^i = \{A^i_1, ..., A^i_{d+2}\}$ each of dimension $d+1 \times d+1$. We set

$$A^i_{d+2} = K_i \oplus 0 = \begin{bmatrix} & & 0 \\ & K_i & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \tag{15}$$

the ith Kraus matrix from the QMOP superoperator with the $d+1$st column and row all zeros. Additionally, let

$$Z^i = \mathbb{I}_{d+1} - {A^i_{d+2}}^\dagger A^i_{d+2} \tag{16}$$

$$= \left(\sum_{j \neq i} K_j^\dagger K_j \right) \oplus 1 \tag{17}$$

$$= \begin{bmatrix} & & 0 \\ \sum_{j \neq i} K_j^\dagger K_j & & 0 \\ & & \vdots \\ 0 \;\; 0 & \dots & 1 \end{bmatrix}. \tag{18}$$

Now $(K_j^\dagger K_j)^\dagger = K_j^\dagger K_j$ and the sum of Hermitian matrices is Hermitian so Z^i is Hermitian. Moreover, $K_j^\dagger K_j$ is positive semidefinite, and positive semidefinite matrices are closed under positive addition, so Z^i is positive semidefinite as well. Let an orthonormal eigendecomposition of Z^i be

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

Since Z^i is a positive semidefinite Hermitian matrix, z_j^i is nonnegative and real so $\sqrt{z_j^i}$ 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 $\sqrt{z_j^i}\langle z_j^i|$:

$$\left(A^i_{j<d+2}\right)_{pq} = \sqrt{z_j^i}\langle z_j^i|q\rangle \delta_{p(d+1)},$$

$$A^i_{j<d+2} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \\ & \sqrt{z_j^i}\langle z_j^i| & \end{bmatrix}.$$

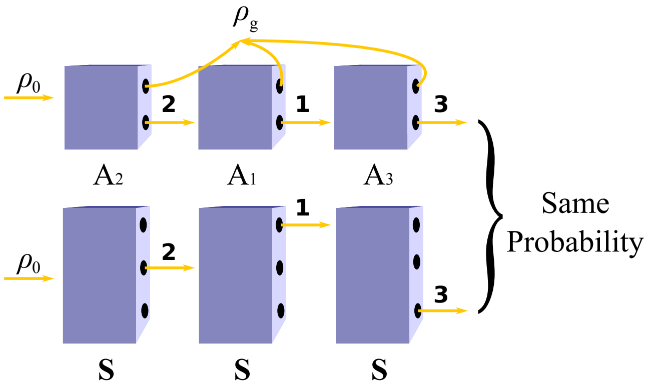

FIG. 2. A goal QOMDP for a QMOP instance with superoperator $\mathbf{S} = \{K_1, K_2, K_3\}$ with 3 possible outcomes. We create 3 actions to correspond to the 3 outputs of the superoperator. Each action A_i has two possible outcomes: either the state transitions according to K_i from $\mathbf{S}$ or 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 $\mathbf{S}$. 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.

(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 $Q(\mathbf{S})$ is shown in Figure 2. 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 transition 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.

Our proof will proceed as follows: 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 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 occur. 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 formalize the sketch we just gave. Before we can do anything else, we must show that $Q(\mathbf{S})$ 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 $\mathbf{S} = \{K_1, ..., K_{\mathcal{K}}\}$ with Kraus matrices of dimension d be the superoperator from an instance of QMOP and let $Q(\mathbf{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g\rangle$ be the corresponding goal QOMDP. For any density matrix $\rho \in S$, if A^i_j is the j^{th} Kraus matrix of the i^{th} action of $Q(\mathbf{S})$ and $j < d+2$ then

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

Proof: Consider

$$(A^i_j \rho A^{i\dagger}_j)_{pq} = \sum_{h,l} A^i_{j\,ph} \rho_{hl} A^{i\dagger}_{j\,lq} \tag{19}$$

$$= \sum_{h,l} A^i_{j\,ph} \rho_{hl} A^{i\,*}_{j\,ql} \tag{20}$$

$$= z^i_j \sum_{h,l} \langle z^i_j|h\rangle \rho_{hl} \langle l|z^i_j\rangle \delta_{p(d+1)}\delta_{q(d+1)} \tag{21}$$

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

$$\frac{A^i_j \rho A^{i\dagger}_j}{\mathrm{Tr}(A^i_j \rho A^{i\dagger}_j)} = |d+1\rangle\langle d+1|. \tag{22}$$

□

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

Proof: By Lemma 3, we know that for $j < d+2$, we have

$$\frac{A^i_j |d+1\rangle\langle d+1| A^{i\dagger}_j}{\mathrm{Tr}(A^i_j |d+1\rangle\langle d+1| A^{i\dagger}_j)} = \rho_g.$$

Here we show that $\mathrm{Tr}(A^i_{d+2}\rho_g A^{i\dagger}_{d+2}) = 0$ so that the probability of applying A^i_{d+2} is 0. We have:

$$\mathrm{Tr}\left(A^i_{d+2}|d+1\rangle\langle d+1|A^{i\dagger}_{d+2}\right) \tag{23}$$

$$= \sum_p \sum_{hl} A^i_{d+2\,ph} \delta_{h(d+1)}\delta_{l(d+1)} A^{i\,*}_{d+2\,pl} \tag{24}$$

$$= \sum_p A^i_{d+2\,p(d+1)} A^{i\,*}_{d+2\,p(d+1)} = 0 \tag{25}$$

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

Now we are ready to show that $Q(\mathbf{S})$ is a goal QOMDP.

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

Proof: We showed in Lemma 4 that ρ_g is absorbing, so all that remains to show is that the actions are superoperators. Let A^i_j be the j^{th} Kraus matrix of action A^i. If $j < d+2$ then

$$\begin{aligned}(A^{i\dagger}_j A^i_j)_{pq} &= \sum_h A^{i\dagger}_{j\,ph} A^i_{j\,hq} && (26)\\ &= \sum_h A^{i*}_{j\,hp} A^i_{j\,hq} && (27)\\ &= \sqrt{z^i_j}^* \langle p|z^i_j\rangle \sqrt{z^i_j}\langle z^i_j|q\rangle && (28)\\ &= z^i_j \langle p|z^i_j\rangle\langle z^i_j|q\rangle && (29)\end{aligned}$$

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

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

Now

$$\begin{aligned}\sum_{j=1}^{d+2} A^{i\dagger}_j A^i_j &= A^{i\ \dagger}_{d+2} A^i_{d+2} + \sum_{j=1}^{d+1} z^i_j |z^i_j\rangle\langle z^i_j| && (30)\\ &= A^{i\ \dagger}_{d+2} A^i_{d+2} + Z^i && (31)\\ &= \mathbb{I}_{d+1}. && (32)\end{aligned}$$

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 but state-dependent. The QMOP problem is essentially time-dependent: we want to know about a specific sequence of observations over time. A QOMDP, however, is state-dependent: the choice of action depends not upon the number of time steps, but upon the current state. When reducing a QMOP problem to a QOMDP problem, we need to ensure that the observations received in the QOMDP are dependent on time in the same way that they are in the QMOP instance. We will be able to do this because we have designed the QOMDP to which we reduce a QMOP instance such that after n time steps there is at most one possible non-goal state. The existence of such a state and the exact state that is reachable depends upon the policy chosen, but regardless of the policy, there will be at most one. This fact, which we will prove in the following lemma, allows us to consider the policy for these QOMDPs as time-dependent: the action we choose at time step n is the action the state-dependent policy chooses for the only non-goal state we could possibly reach at time n.

Lemma 6: Let $\mathbf{S} = \{K_1, ..., K_{\mathcal{K}}\}$ with Kraus matrices of dimension d be the superoperator from an instance of QMOP and let $Q(\mathbf{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ be the corresponding goal QOMDP. Let $\pi : S \rightarrow \mathcal{A}$ be any policy for $Q(\mathbf{S})$. 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: Let ρ_n be the state on the n^{th} time step and let ρ_{n-1} be the state on the $(n-1)^{\text{st}}$ time 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 $\mathbf{S} = \{K_1, ..., K_{\mathcal{K}}\}$ with Kraus matrices of dimension d be the superoperator from a QMOP instance and let $Q(\mathbf{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ be the corresponding goal QOMDP. For any policy π let σ_k be the non-goal state with nonzero probability after k time steps of following π if it exists. Otherwise let $\sigma_k = \rho_g$. Choose $\sigma_0 = \rho_0$. 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 Hilbert space 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 $\mathbf{S} = \{K_1, ..., K_{\mathcal{K}}\}$ with Kraus matrices of dimension d be the superoperator from a QMOP instance and let $Q(\mathbf{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ 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 $(d+1)^{\text{st}}$ 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:

$$({A^i_{d+2}} M {A^i_{d+2}}^\dagger)_{pq} = \sum_{h,l=1}^{d+1} {A^i_{d+2}}_{ph} M_{hl} {A^i_{d+2}}^\dagger_{lq} \quad (33)$$

$$= \sum_{h,l=1}^{d} {A^i_{d+2}}_{ph} M_{hl} {A^i_{d+2}}^*_{ql} \quad (34)$$

where we have used that the $(d+1)^{\text{st}}$ column of A^i_{d+2} is 0 to limit the sum. Additionally, if $p = d+1$ or $q = d+1$, the sum is 0 because the $(d+1)^{\text{st}}$ row of A^i_{d+2} is 0. Assume that $p < d+1$ and $q < d+1$. Then

$$\sum_{h,l=1}^{d} {A^i_{d+2}}_{ph} M_{hl} {A^i_{d+2}}^*_{ql}$$

$$= \sum_{h,l=1}^{d} K_{iph} M_{hl} K^\dagger_{i lq} = \left(K d(M) K^\dagger\right)_{ql}. \quad (35)$$

Thus

$$A^i_{d+2} M {A^i_{d+2}}^\dagger = K_i d(M) K_i^\dagger \oplus 0. \quad (36)$$

□

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 $\mathbf{S} = \{K_1, ..., K_\mathcal{K}\}$ with Kraus matrices of dimension d be the superoperator from a QMOP instance and let $Q(\mathbf{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ 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

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

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

$$\sigma_1 = \frac{A^{i_1}_{d+2} \rho_0 {A^{i_1}_{d+2}}^\dagger}{\text{Tr}(A^{i_1}_{d+2} \rho_0 {A^{i_1}_{d+2}}^\dagger)} \quad (37)$$

$$= \frac{K_{i_1} d(\rho_0) K^\dagger_{i_1} \oplus 0}{\text{Tr}(K_{i_1} d(\rho_0) K^\dagger_{i_1} \oplus 0)} \quad (38)$$

$$= \frac{K_{i_1} d(\rho_0) K^\dagger_{i_1} \oplus 0}{\text{Tr}(K_{i_1} d(\rho_0) K^\dagger_{i_1})} \quad (39)$$

using Lemma 7 for Equation 38 and the fact that $\text{Tr}(A \oplus 0) = \text{Tr}(A)$ for Equation 39.

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

$$\sigma_{k-1} = \frac{K_{i_{k-1}} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_{k-1}} \oplus 0}{\text{Tr}(K_{i_{k-1}} ... K_{i_1} d(\rho_0) K^\dagger_1 ... K^\dagger_{i_{k-1}})}. \quad (40)$$

Note that

$$d(\sigma_{k-1}) = \frac{K_{i_{k-1}} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_{k-1}}}{\text{Tr}(K_{i_{k-1}} ... K_{i_1} d(\rho_0) K^\dagger_1 ... K^\dagger_{i_{k-1}})}. \quad (41)$$

Then

$$\sigma_k = \frac{A^{i_k}_{d+2} \sigma_{k-1} A^{i_k}_{d+2}}{\text{Tr}(A^{i_k}_{d+2} \sigma_{k-1} {A^{i_k}_{d+2}}^\dagger)} = \frac{K_{i_k} d(\sigma_{k-1}) K^\dagger_{i_k} \oplus 0}{\text{Tr}(K_{i_k} d(\sigma_{k-1}) K^\dagger_{i_k})} \quad (42)$$

using Lemma 7. Using Equation 41 for $d(\sigma_{k-1})$, we have

$$K_{i_k} d(\sigma_{k-1}) K^\dagger_{i_k} = \frac{K_{i_k} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_k}}{\text{Tr}\left(K_{i_{k-1}} ... K_{i_1} d(\rho_0) K_1 ... K_{i_{k-1}}\right)}, \quad (43)$$

and

$$\text{Tr}(K_{i_k} d(\sigma_{k-1}) K^\dagger_{i_k})$$

$$= \text{Tr}\left(\frac{K_{i_k} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_k}}{\text{Tr}\left(K_{i_{k-1}} ... K_{i_1} d(\rho_0) K_1 ... K_{i_{k-1}}\right)}\right) \quad (44)$$

$$= \frac{\text{Tr}\left(K_{i_k} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_k}\right)}{\text{Tr}\left(K_{i_{k-1}} ... K_{i_1} d(\rho_0) K_1 ... K_{i_{k-1}}\right)}, \quad (45)$$

Substituting equations 43 and 45 for the numerator and denominator of equation 42 respectively, and canceling the traces, we find

$$\sigma_k = \frac{K_{i_k} ... K_{i_1} d(\rho_0) K_{i_1} ... K_{i_k} \oplus 0}{\text{Tr}(K_{i_k} ... K_{i_1} d(\rho_0) K^\dagger_{i_1} ... K^\dagger_{i_k})}. \quad (46)$$

□

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

Lemma 9: Let $\mathbf{S} = \{K_1, ..., K_\mathcal{K}\}$ with Kraus matrices of dimension d be the superoperator from a QMOP instance and let $Q(\mathbf{S}) = \langle S, \Omega, \mathcal{A}, \rho_0, \rho_g \rangle$ be the corresponding goal QOMDP. Let π be any policy and $\{\sigma_k\}$

be the policy path for π. Assume $\pi(\sigma_{j-1}) = A^{i_j}$. The probability that ρ_n is not ρ_g is

$$\Pr(\rho_n \neq \rho_g) = \mathrm{Tr}(K_{i_n}...K_{i_1} d(\rho_0) K^\dagger_{i_1}...K^\dagger_{i_n}). \tag{47}$$

Proof: First 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,

$$\sigma_{n-1} = \frac{K_{i_{n-1}}...K_{i_1} d(\rho_0) K^\dagger_{i_1}...K^\dagger_{i_{n-1}} \oplus 0}{\mathrm{Tr}(K_{i_{n-1}}...K_{i_1} d(\rho_0) K^\dagger_{i_1}...K^\dagger_{i_{n-1}})} \tag{48}$$

so

$$d(\sigma_{n-1}) = \frac{K_{i_{n-1}}...K_{i_1} d(\rho_0) K^\dagger_{i_1}...K^\dagger_{i_{n-1}}}{\mathrm{Tr}(K_{i_{n-1}}...K_{i_1} d(\rho_0) K^\dagger_{i_1}...K^\dagger_{i_{n-1}})}. \tag{49}$$

If $A^{i_n}_j$ 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 $A^{i_n}_{d+2}$ is applied:

$$\begin{aligned} &\Pr(\rho_n \neq \rho_g | \rho_{n-1} \neq \rho_g) \\ &\quad = \mathrm{Tr}(A^{i_n}_{d+2} \sigma_{n-1} {A^{i_n}_{d+2}}^\dagger) \end{aligned} \tag{50}$$

$$= \mathrm{Tr}(K_{i_n} d(\sigma_{n-1}) K^\dagger_{i_n} \oplus 0) \tag{51}$$

$$= \mathrm{Tr}(K_{i_n} d(\sigma_{n-1}) K^\dagger_{i_n}) \tag{52}$$

$$= \frac{\mathrm{Tr}(K_{i_n}...K_{i_1} d(\rho_0) K^\dagger_{i_1}...K^\dagger_{i_n})}{\mathrm{Tr}(K_{i_{n-1}}...K_{i_1} d(\rho_0) K^\dagger_{i_1}...K^\dagger_{i_{i-1}})}. \tag{53}$$

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

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

□

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: As noted above, it suffices to show that we can reduce the quantum measurement occurrence problem (QMOP) to goal-state reachability for QOMDPs.

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

First assume there is some sequence $\{i_1, ..., i_n\}$ such that $K^\dagger_{i_1}...K^\dagger_{i_n} K_{i_n}...K_{i_1} = 0$. 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

$$\Pr(\rho_n \neq \rho_g) = \mathrm{Tr}(K_{i_n}...K_{i_1} d(\rho_0) K^\dagger_{i_1}...K^\dagger_{i_n}) \tag{54}$$

$$= \mathrm{Tr}(K^\dagger_{i_1}...K^\dagger_{i_n} K_{i_n}...K_{i_1} d(\rho_0)) \tag{55}$$

$$= \mathrm{Tr}(0) \tag{56}$$

$$= 0 \tag{57}$$

using that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for all matrices A and B. 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 k^{th} 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

$$\Pr(\rho_n \neq \rho_g | \pi) = \mathrm{Tr}(K_{i_1}...K_{i_n} d(\rho_0) K^\dagger_{i_1}...K^\dagger_{i_n}) \tag{58}$$

$$= \mathrm{Tr}(K^\dagger_{i_1}...K^\dagger_{i_n} K_{i_n}...K_{i_i} d(\rho_0)). \tag{59}$$

Since π reaches the goal state with probability 1 after n time steps, we must have that the above quantity is 0. By construction $d(\rho_0)$ is full rank, so for the trace to be 0 we must have

$$K^\dagger_{i_1}...K^\dagger_{i_n} K_{i_n}...K_{i_i} = 0. \tag{60}$$

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. □

B. Decidability of Goal-State Reachability for POMDPs

The goal-state reachability problem for POMDPs is decidable. This is a known result [11], but we reproduce the proof here, because it is interesting to see the differences between classical and quantum probability that lead to decidability for the former.

At a high level, the goal-state reachability problem is decidable for POMDPs because stochastic transition matrices have strictly nonnegative elements. Since we are interested in a probability 1 event, we can treat probabilities as binary: either positive or 0. 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.

Definition 6 (Binary Probability MDP): Given a goal POMDP $P = \langle S, A, \Omega, T, O, \vec{b}_0, g\rangle$, let $M(P) = \langle B, A, \tau, \vec{b}_0, \vec{b}_g\rangle$ be the corresponding goal belief MDP with τ^{ao} defined according to Equation 4. Throughout this section, we assume without loss of generality that g is the $|S|^{\text{th}}$ state in P so $\left(\vec{b}_g\right)_i = \delta_{i|S|}$. The *binary probability MDP* is an MDP $D(P) = \langle \mathbb{Z}_{\{0,1\}}^{|S|}, A, Z, \vec{z}_0, \vec{z}_g\rangle$ where $\left(\vec{z}_g\right)_i = \delta_{i|S|}$ and $\left(\vec{z}_0\right)_i = 1$ if and only if $(\vec{b}_0)_i > 0$. The transition function Z for action a non-deterministically applies the function Z^{ao} to $\vec{z}$. For $\vec{z} \in \mathbb{Z}_{\{0,1\}}^{|S|}$, the result of Z^{ao} acting on $\vec{z}$ is

$$Z^{ao}(\vec{z})_i = \begin{cases} 1 & \text{if } \left(\tau^{ao}\vec{z}\right)_i > 0 \\ 0 & \text{if } \left(\tau^{ao}\vec{z}\right)_i = 0. \end{cases} \tag{61}$$

Let

$$P_a^o(\vec{z}) = \begin{cases} 1 & \text{if } \tau^{ao}\vec{z} \neq \vec{0} \\ 0 & \text{else.} \end{cases} \tag{62}$$

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

$$\Pr\left(Z^{ao} | a, \vec{z}\right) = \begin{cases} \frac{1}{\sum_{o' \in \Omega} P_a^{o'}(\vec{z})} & \text{if } P_o^a(\vec{z}) > 0 \\ 0 & \text{else.} \end{cases} \tag{63}$$

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

We first show that we can keep track of whether each entry in the belief state is zero or not just using the binary probability MDP. This lemma uses the fact that classical probability involves nonnegative numbers only.

Lemma 11: Let $P = \langle S, A, \Omega, T, O, \vec{b}_0, g\rangle$ be a goal-state POMDP and let $D(P) = \langle \mathbb{Z}_{\{0,1\}}^{|S|}, A, Z, \vec{z}_0, \vec{z}_g\rangle$ 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} from Equation 61,

$$\vec{z}_i^{ao} = Z^{ao}(\vec{z})_i = \begin{cases} 1 & \text{if } \left(\tau^{ao}\vec{z}\right)_i > 0 \\ 0 & \text{else.} \end{cases} \tag{64}$$

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

$$\vec{b}_i^{ao} = \frac{1}{N}\sum_{j=1}^{|S|} \tau_{ij}^{ao}\vec{b}_j. \tag{65}$$

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 65 is 0 individually[12]. 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

$$0 = \sum_{j=1}^{|S|} \tau_{ij}^{ao}\vec{z}_j = \left(\tau^{ao}\vec{z}\right)_i = \vec{z}_i^{ao}. \tag{66}$$

Now assume $\vec{b}_i^{ao} > 0$. Then there must be at least one term in the sum in Equation 65 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

$$\vec{z}_i^{ao} = \left(\tau^{ao}\vec{z}\right)_i = \sum_{j=1}^{|S|} \tau_{ij}^{ao}\vec{z}_j = \sum_{j \neq k} \tau_{ij}^{ao}\vec{z}_j + \tau_{ik}^{ao}\vec{z}_k > 0. \tag{67}$$

Since $\vec{b}_i^{ao} \geq 0$ and $\vec{z}_i^{ao} > 0$, we have shown that $\vec{z}_i^{ao} = 0$ exactly when $\vec{b}_i^{ao} = 0$.

Now assume $\left|\tau^{ao}\vec{b}\right|_1 = 0$. This is true only if $\tau_{ij}^{ao}\vec{b}_j = 0$ for all i and j. Thus by the same reasoning as above $\tau_{ij}^{ao}\vec{z}_j = 0$ for all i and 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. We do each direction in a separate lemma.

Lemma 12: Let $P = \langle S, A, \Omega, T, O, \vec{b}_0, g\rangle$ be a goal POMDP and let $D(P) = \langle \mathbb{Z}_{\{0,1\}}^{|S|}, A, Z, \vec{z}_0, \vec{z}_g\rangle$ 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 $\vec{z} \in \mathbb{Z}^n_{\{0,1\}}$ 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 sequence of beliefs of length $n+1$ that can be created by following policy π with observations $\{o_{i_1}, ..., o_{i_n}\}$. Then

$$\vec{b}^{k+1} = \frac{\tau^{\pi(\vec{b}^k) o_{i_k}} \vec{b}^k}{\left|\tau^{\pi(\vec{b}^k) o_{i_k}} \vec{b}^k\right|_1} = \frac{\tau^{\pi^D(z(\vec{b}^k)) o_{i_k}} \vec{b}^k}{\left|\tau^{\pi^D(z(\vec{b}^k)) o_{i_k}} \vec{b}^k\right|_1}. \tag{68}$$

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 $\vec{z}^0 = z(\vec{b}^0)$ by definition.

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

$$\vec{z}^{k+1} = Z^{\pi^D(\vec{z}^k) o_{i_k}}(\vec{z}^k) = Z^{\pi^D(z(\vec{b}^k)) o_{i_k}}(\vec{z}^k) = Z^{a_k o_{i_k}}(\vec{z}^k) \tag{69}$$

by induction. Now

$$\vec{b}^{k+1} = \frac{\tau^{a_{i_k} o_{i_k}} \vec{b}^k}{\left|\tau^{a_{i_k} o_{i_k}} \vec{b}^k\right|_1}. \tag{70}$$

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 nonzero probability of occurring while following π^D. We must have that $P^{o_{i_k}}_{a_k} > 0$ 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 0. Therefore, we must have

$$\Pr(o | a_k, \vec{b}^k) = \left|\tau^{a_k o_{i_k}} \vec{b}^k\right|_1 > 0 \tag{71}$$

for all k, so Lemma 11 gives us that $P^{o_{i_k}}_{a_k} > 0$ for all k. Thus $\{\vec{z}^0, ..., \vec{z}^n\}$ is a possible sequence of states seen while following policy π^D in the MDP $D(P)$. Since π^D reaches the goal state with probability 1 after n time steps, we have $\vec{z}^n = \vec{z}_g$. Therefore, since $\vec{z}^n = z(\vec{b}^n)$, we must have $\vec{b}^n_i = 0$ for all $i \neq |S|$, and only $\vec{b}^n_{|S|} > 0$. Since $|\vec{b}^n|_1 = 1$, we have $\vec{b}^n_{|S|} = 1$. Thus $\vec{b}^n = \vec{b}_g$ and π also reaches the goal state with nonzero probability after n time steps.

□

Lemma 13: Let $P = \langle S, A, \Omega, T, O, \vec{b}_0, g\rangle$ be a goal POMDP and let $D(P) = \langle \mathbb{Z}^{|S|}_{\{0,1\}}, A, Z, \vec{z}_0, \vec{z}_g\rangle$ 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 $B(M) = \langle B, A, \tau, \vec{b}_0, \vec{b}_g\rangle$ then there is a policy that reaches the goal in a finite number of steps with probability 1 in $D(P)$.

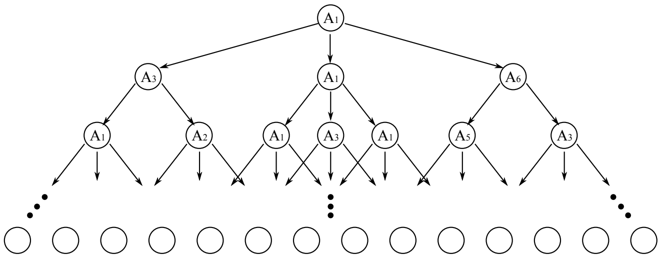

FIG. 3. A policy in an MDP creates a tree. Here, we take action A_1 in the starting state, which can transition us nondeterministically to three other possible states. The policy specifies an action of A_3 for the state on the left, A_1 for the state in the middle and A_6 for the state on the right. Taking these actions transition these states nondeterministically. So this tree eventually encapsulates all states that can be reached with nonzero 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.

Proof: MDP policies create trees of states and action choices as shown in Figure 3. Consider the tree π_T formed by π. Nodes at depth n or greater are guaranteed to be $\vec{b}_g$. For $\vec{z} \in \mathbb{Z}^{|S|}_{\{0,1\}}$, we let $b(\vec{z})$ be the deepest state in π_T for which $\vec{b}_i = 0$ if and only if $\vec{z}_i = 0$. If there are multiple states for which this is true at the same level, we choose the leftmost one. 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 any sequence of states that can be created by following policy π^D in $D(P)$ 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 $\vec{z}^0, ..., \vec{z}^n$ can be created by following 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 $\vec{b}^0_i = 0$ if and only if $\vec{z}^0_i = 0$ 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

$$\vec{z}^{k+1} = Z^{a_k o_{i_k}}\left(\vec{z}^k\right). \tag{72}$$

Therefore by Lemma 11,

$$\vec{b}' = \frac{\tau^{a_k o_{i_k}} b(\vec{z}^k)}{|\tau^{a_k o_{i_k}} b(\vec{z}^k)|_1} \tag{73}$$

has entry i 0 if and only if $\vec{z}^{k+1}_i = 0$. Now $P^{a_k}_{o_k}(\vec{z}^k) \neq 0$ 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 $P^{a_k}_{o_k} > 0$. 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 deepest state in the tree 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 at most 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 in at most n steps. □

We have now reduced goal-state reachability for POMDPs to goal-state reachability for finite-state MDPs. We briefly show that the latter is decidable.

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 in a finite number of steps with probability 1 if the goal is reachable from the starting state in G and no cycle is reachable. The number of nodes in the graph is at most the number of states in the MDP so we can clearly decide this problem. Thus goal-state reachability is decidable for POMDPs. □

C. Other Computability Separations

Although we looked only at goal-state reachability here, we conjecture that there are other similar problems that are undecidable for QOMDPs despite being decidable for POMDPs.

For instance, the zero-reward policy problem is a likely candidate for computability 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.

V. FUTURE WORK

We were only able to give an interesting computability result for a problem about goal decision processes, which ignore the reward function. It would be a great to prove a result about QOMDPs that made nontrivial use of the reward function.

We also proved computability results, but did not consider algorithms for solving any of the problems we posed beyond a very simple PSPACE algorithm for policy existence. Are there quantum analogues of POMDP algorithms or even MDP ones?

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. 0844626 and 1122374, as well as an NSF Waterman Award.

[1] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, Artificial Intelligence **101**, 99 (1998).

[2] J. Pineau, G. Gordon, and S. Thrun, in *International joint conference on artificial intelligence*, Vol. 18 (2003) pp. 1025–1032.

[3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. (Pearson Hall, New Jersey, 2003) chapter 17.

[4] M. T. J. Spaan and N. Vlassis, Journal of Artificial Intelligence Research **24**, 195 (2005).

[5] T. Smith and R. Simmons, in *Uncertainty in Artificial Intelligence* (2004) pp. 520–527.

[6] C. H. Papadimitriou and J. N. Tsitsiklis, Mathematics of Operations Research **12**, 441 (1987).

[7] O. Madani, S. Hanks, and A. Condon, in *Association for the Advancement of Artificial Intelligence* (1999).

[8] 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.

[9] R. B. Griffiths, "Quantum Channels, Kraus Operators, POVMs," Quantum Computation and Quantum Information Theory Course Notes, Carnegie Mellon University (2010).

[10] J. Eisert, M. P. Mueller, and C. Gogolin, Physical Review Letters **108** (2012).

[11] J. Rintanen, in *International Conference on Automated Planning and Scheduling* (2004) pp. 345–354.

[12] This holds because probabilities are nonnegative. A similar analysis in the quantum case would fail at this step.