PHYSICAL REVIEW A **90**, 032311 (2014)

Quantum partially observable Markov decision processes

Jennifer Barry,[1,*] Daniel T. Barry,[2,†] and Scott Aaronson[3,‡]
[1]*Rethink Robotics, Boston, Massachusetts 02210, USA*
[2]*Denbar Robotics, Sunnyvale, California 94085, USA*
[3]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
(Received 12 June 2014; published 9 September 2014)

We present quantum observable Markov decision processes (QOMDPs), the quantum analogs of partially observable Markov decision processes (POMDPs). In a QOMDP, an agent is acting in a world where the 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.

DOI: 10.1103/PhysRevA.90.032311 PACS number(s): 03.67.Ac, 89.20.Ff

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 world is always in exactly one state, but the agent is not told this state. Instead, it can take actions and receive observations about the world. The actions an agent takes are nondeterministic; before taking an action, the agent knows only the probability distribution of its next state given the current state. Similarly, an observation does not give the agent direct knowledge of the current world state, but the agent knows the probability of receiving a given observation in each possible state. The agent is rewarded for the actual, unknown world 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 Sec. II, an agent can maximize future expected reward in a POMDP by maintaining a probability distribution, known as a belief state, over the world's current state. By carefully updating this belief state after every action and observation, the agent can ensure that its belief state reflects the correct probability that the world is in each possible state. The agent can make decisions using only its belief about the state without ever needing to reason more directly about the actual world 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 analog 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 are the natural quantum generalizations of POMDPs, which are central objects of study in artificial intelligence. 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.

Recently, related work has been reported by Ying and Ying [6]. They considered quantum Markov decision processes (MDPs) and proved undecidability results for them that are very closely related to our results. In particular, these authors show that the finite-horizon reachability problem for quantum MDPs is undecidable, and they also do so via a reduction from the matrix mortality problem. Ying and Ying also prove EXP-hardness and uncomputability for the infinite-horizon case (depending on whether one is interested in reachability with probability 1 or with probability $p < 1$, respectively). On the other hand, they give an algorithm that decides, given a quantum MDP and an invariant subspace B, whether or not there exists a policy that reaches B with probability 1 regardless of the initial state, and they prove several other results about invariant subspaces in MDPs. These results nicely extend and complement ours as well as previous work by the same group [7].

One possible advantage of the present work is that, rather than considering (fully observable) MDPs, we consider POMDPs. The latter seem to us like a more natural starting point than MDPs for a quantum treatment, because there

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

is never "full observability" in quantum mechanics. Many results, including the undecidability results mentioned above, can be translated between the MDP and POMDP settings, by the simple expedient of considering "memoryful" MDP policies, that is, policies that remember the initial state, as well as all actions performed so far and all measurement outcomes obtained. Such knowledge is tantamount to knowing the system's *current* quantum state ρ. However, because we consider POMDPs, which by definition can take actions that depend on ρ, we never even need to deal with the issue of memory. A second advantage of this work is that we explicitly compare the quantum against the classical case (something not done in [6]), showing why the same problem is undecidable in the former case but decidable in the latter.

Finally, we mention that there has been other work that sought to model quantum agents in dynamic and uncertain environments [8,9], though without formal computability and uncomputability results.

II. PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

For completeness, in this section we give an overview of Markov decision processes and partially observable Markov decision processes (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 [3].

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. At each time step, the world is in exactly one known state, and the agent chooses to take a single action, which transitions the world to a new state according to T. The objective is for the agent to act in such a way as to maximize future expected reward.

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

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

The *solution* to an MDP of horizon h is the *optimal 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 1. The optimal policy over an infinite horizon is the one inducing the value function

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

Equation (2) is called the *Bellman equation*, and there is a unique solution for V^* [3]. Note that V^* is noninfinite 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.

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 the state of the world 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, $T(s_i,a,s_j) : S \times A \times S \to [0,1]$ 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 world state is "hidden," meaning that the agent does not know the world state, but the dynamics of the world behave according to the actual underlying state. At each time step, the agent chooses an action, the world transitions to a new state according to its current hidden state and T, and the agent receives an observation according to the world 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 world states. For $s_i \in S$,

$\vec{b}_i$ is the probability that the world is in state s_i. Since $\vec{b}$ is a probability distribution, $0 \leqslant \vec{b}_i \leqslant 1$ and $\sum_i \vec{b}_i = 1$. If the agent has belief state $\vec{b}$, takes action a, and receives observation o the agent's new belief state is

$$\vec{b'}_i = \Pr(s_i|o,a,\vec{b}) = \frac{\Pr(o|s_i,a,\vec{b})\Pr(s_i|a,\vec{b})}{\Pr(o|a,\vec{b})} = \frac{O(s_i,a,o)\sum_j T(s_j,a,s_i)\vec{b}_j}{\Pr(o|a,\vec{b})}. \quad (3)$$

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

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

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

$$\vec{b}' = \frac{\tau^{ao}\vec{b}}{|\tau^{ao}\vec{b}|_1}, \quad (5)$$

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

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

where

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

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

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

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 problem is that the state space of the belief state MDP 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 [10]. 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 [11].

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 $a \in A$, $T(g,a,g) = 1$. 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. First the probability of observing o after taking action a is

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

Therefore, if the agent has belief $\vec{b}_g$, regardless of the action taken, the agent sees observation $o_{|\Omega|}$. Assume the agent takes action a and sees observation $o_{|\Omega|}$. The next belief state is

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

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

III. QUANTUM OBSERVABLE MARKOV DECISION PROCESSES

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 system. 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 system'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 [12,13]. A set of matrices $\{K_1,...,K_\mathcal{K}\}$ of dimension d is a set of Kraus matrices if and only if

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

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

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

with probability

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

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

We can now define the quantum observable Markov decision process (QOMDP).

Definition 3 (QOMDP). A QOMDP is a tuple $\langle S,\Omega,\mathcal{A},\mathcal{R},\gamma,\rho_0\rangle$ where we have the following.

(1) S is a Hilbert space. We allow pure and mixed quantum states so we will represent states in S as density matrices.

(2) $\Omega = \{o_1,...,o_{|\Omega|}\}$ is a set of possible observations.

(3) $\mathcal{A} = \{A^1,...,A^{|\mathcal{A}|}\}$ is a set of superoperators. Each superoperator $A^a = \{A^a_1,...,A^a_{|\Omega|}\}$ has $|\Omega|$ Kraus matrices. Note that each superoperator returns the same set of possible observations; if this is not true in reality, some of the Kraus matrices 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

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

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

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

(4) $\mathcal{R} = \{R_1,...,R_{|\mathcal{A}|}\}$ 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:

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

(5) $\gamma \in [0,1)$ is a discount factor.

(6) $\rho_0 \in S$ is the starting state.

Like an MDP or POMDP, a QOMDP represents a world in which an agent chooses actions at discrete time steps and receives observations. The world modeled by the QOMDP is a quantum system that begins in ρ_0, the starting state of the QOMDP. At each time step, the agent chooses a superoperator from the set $\mathcal{A}$, whereupon the corresponding operation is done on the system and the agent receives an observation from the set Ω in accordance with the laws of quantum mechanics. The agent also receives a reward according to the state of the system after the operation and $\mathcal{R}$. As in an MDP or POMDP, the agent knows the entire QOMDP model *a priori* and its goal is to use this information to maximize its future expected reward.

A QOMDP is fully observable in the same sense that the belief state MDP for a POMDP is fully observable. Just as the agent in a POMDP always knows its belief state, the agent in a QOMDP always knows the current quantum superposition or mixed state of the system. In a POMDP, the agent can update its belief state when it takes an action and receives an observation using Eq. (5). Similarly, in a QOMDP, the agent can keep track of the quantum state using Eq. (12) each time it takes an action and receives an observation. Note that a QOMDP is much more analogous to the belief state MDP of a POMDP than to the POMDP itself. In a POMDP, the system is always in one actual underlying world state that is simply unknown to the agent; in a QOMDP, the system can be in a superposition state for which no underlying "real" state exists.

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

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

Let π_h be the policy at time h. Then

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

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 Eqs. (11), (12), and (13) respectively. The Bellman equation [Eq. (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 $A^i_j \in A^i$ if $\mathrm{Tr}(A^i_j \rho_g A^{i\dagger}_j) > 0$,

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

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. Papadimitriou and Tsitsiklis [10] showed that polynomial horizon POMDPs are in PSPACE and the proof still holds for QOMDPs with the appropriate substitution for the calculations of the probability of an observation given a quantum state and action [Eq. (11)], N [Eq. (12)], and R [Eq. (13)], all of which can clearly be done in PSPACE when the horizon is polynomial. ■

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.* [14] to it.

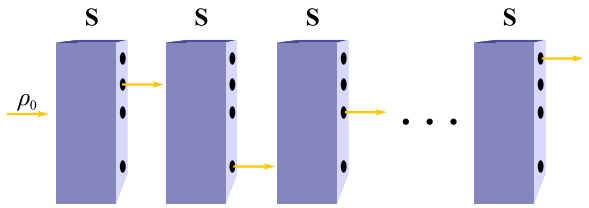

FIG. 1. (Color online) The quantum measurement occurrence problem. The starting state ρ_0 is fed into the superoperator **S**. The output is then fed iteratively back into **S**. The question is whether there is some finite sequence of observations that can never occur.

Definition 4 (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 Fig. 1. We assume that the system starts in state ρ_0. This state is fed into **S**. We then take the output of **S** acting on ρ_0 and feed that again into **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 4 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.* [14]. ■

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 Fig. 2.

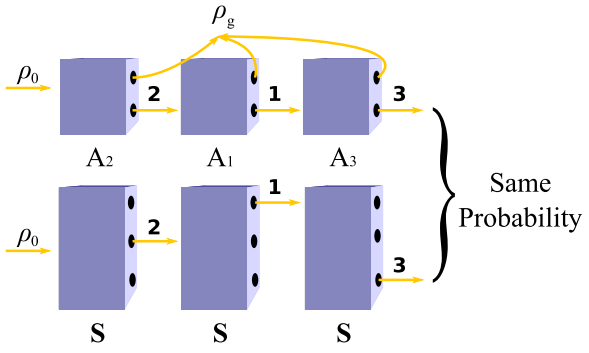

FIG. 2. (Color online) A goal QOMDP for a QMOP instance with superoperator $\mathbf{S} = \{K_1, K_2, K_3\}$ with three possible outcomes. We create three actions to correspond to the three outputs of the superoperator. Each action A_i has two possible outcomes: either the system transitions according 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.

Definition 5 (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.

(1) S is $(d+1)$-dimensional Hilbert space.

(2) $\Omega = \{o_1, o_2, ..., o_{d+2}\}$ 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.

(3) $\mathcal{A} = \{A^1,...,A^\mathcal{K}\}$ is a set of $\mathcal{K}$ superoperators each with $d+2$ Kraus matrices $A^i = \{A_1^i,...,A_{d+2}^i\}$ each of dimension $d+1 \times d+1$. We set

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

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

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

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

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

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_j^i for $j < d+2$ be the $d+1 \times d+1$ matrix in which the first d rows are all zeros and the bottom row is $\sqrt{z_j^i}\langle z_j^i|$:

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

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

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

(4) ρ_0 is the maximally mixed state $\rho_{0ij} = \frac{1}{d+1}\delta_{ij}$.

(5) ρ_g is the state $|d+1\rangle\langle d+1|$.

The intuition behind the definition of $Q(\mathbf{S})$ is shown in Fig. 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 action A^i really only provides two possibilities.

(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 the system transitions to the goal state is the same as the probability that it does not evolve according to first K_{i_1} then K_{i_2}, etc. Therefore, the system transitions to the goal state with probability 1 if and only if it is impossible for it 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, the agent 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. Therefore 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. First, we prove that $A^i_{j<d+2}$ transitions all density matrices to the goal state. Second, we show that ρ_g has zero probability of evolving according to A^i_{d+2}.

Lemma 1. 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 jth Kraus matrix of the ith action of $Q(\mathbf{S})$ 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

$$\left(A^i_j\rho {A^i_j}^\dagger\right)_{pq}=\sum_{h,l}{A^i_j}_{ph}\rho_{hl}{A^i_j}^\dagger_{lq} \tag{19}$$

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

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

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

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

■

Lemma 2. 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 1, we know that for $j<d+2$ we have

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

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

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

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

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

since the $(d+1)$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 3. 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 2 that ρ_g is absorbing, so all that remains to show is that the actions are superoperators. Let A^i_j be the jth Kraus matrix of action A^i. If $j<d+2$ then

$$\left({A^i_j}^\dagger A^i_j\right)_{pq}=\sum_h{A^i_j}^\dagger_{ph}{A^i_j}_{hq} \tag{26}$$

$$=\sum_h{A^i_j}^*_{hp}{A^i_j}_{hq} \tag{27}$$

$$=\sqrt{z^i_j}^*\langle p|z^i_j\rangle\sqrt{z^i_j}\langle z^i_j|q\rangle \tag{28}$$

$$=z^i_j\langle p|z^i_j\rangle\langle z^i_j|q\rangle, \tag{29}$$

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

$$\sum_{j=1}^{d+2}{A^i_j}^\dagger A^i_j={A^i_{d+2}}^\dagger A^i_{d+2}+\sum_{j=1}^{d+1}z^i_j|z^i_j\rangle\langle z^i_j| \tag{30}$$

$$={A^i_{d+2}}^\dagger A^i_{d+2}+Z^i \tag{31}$$

$$=\mathbb{I}_{d+1}. \tag{32}$$

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 policy, 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 nongoal state for the system. 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 the time-dependent policy chooses at time step n is the action the state-dependent policy chooses for the only nongoal state the system could possibly reach at time n.

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. Let $\pi : S \to \mathcal{A}$ 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 one time step, the agent has taken a single action, $\pi(\rho_0)$. Lemma 1 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 nth time step and let ρ_{n-1} be the state on the $(n-1)$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 $\rho_{n-1} = \sigma_{n-1}$, the agent takes action $\pi(\sigma_{n-1}) = A^{i_n}$. By Lemma 1 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 nongoal state in order.

Definition 6 (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 nongoal 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 sequence $\{\sigma_k\}$ is the *policy path* for policy π. By Lemma 3, this sequence 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 4. 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 let $d(M)$ 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 in terms of elements:

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

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

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

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

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

Thus

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

■

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 5. 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

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

Proof. We proceed by induction on k.

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

$$\sigma_1 = \frac{A^{i_1}_{d+2} \rho_0 {A^{i_1}_{d+2}}^\dagger}{\mathrm{Tr}\left(A^{i_1}_{d+2} \rho_0 {A^{i_1}_{d+2}}^\dagger\right)} \tag{37}$$

$$= \frac{K_{i_1} d(\rho_0) K^\dagger_{i_1} \oplus 0}{\mathrm{Tr}\left(K_{i_1} d(\rho_0) K^\dagger_{i_1} \oplus 0\right)} \tag{38}$$

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

using Lemma 4 for Eq. (38) and the fact that $\mathrm{Tr}(A \oplus 0) = \mathrm{Tr}(A)$ for Eq. (39).

Induction step. On time step k, we have $\rho_{k-1} = \sigma_{k-1}$ or $\rho_{k-1} = \rho_g$ by Lemma 3. If $\rho_{k-1} = \rho_g$ then $\rho_k = \rho_g$ by Lemma 2. Therefore, σ_k occurs only if $\rho_{k-1} = \sigma_{k-1}$. In this case the agent takes action A^{i_k}. If $A_j^{i_k}$ is applied to the system with $j < d+2$, ρ_k is the goal state by Lemma 1. Therefore, the system transitions to σ_k exactly when $\rho_{k-1} = \sigma_{k-1}$ and $A_{d+2}^{i_k}$ is applied. By induction

$$\sigma_{k-1} = \frac{K_{i_{k-1}} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_{k-1}}^\dagger \oplus 0}{\mathrm{Tr}\left(K_{i_{k-1}} \ldots K_{i_1} d(\rho_0) K_1^\dagger \ldots K_{i_{k-1}}^\dagger\right)}. \tag{40}$$

Note that

$$d(\sigma_{k-1}) = \frac{K_{i_{k-1}} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_{k-1}}^\dagger}{\mathrm{Tr}\left(K_{i_{k-1}} \ldots K_{i_1} d(\rho_0) K_1^\dagger \ldots K_{i_{k-1}}^\dagger\right)}. \tag{41}$$

Then

$$\sigma_k = \frac{A_{d+2}^{i_k} \sigma_{k-1} A_{d+2}^{i_k}}{\mathrm{Tr}\left(A_{d+2}^{i_k} \sigma_{k-1} {A_{d+2}^{i_k}}^\dagger\right)} = \frac{K_{i_k} d(\sigma_{k-1}) K_{i_k}^\dagger \oplus 0}{\mathrm{Tr}\left(K_{i_k} d(\sigma_{k-1}) K_{i_k}^\dagger\right)} \tag{42}$$

using Lemma 4. Using Eq. (41) for $d(\sigma_{k-1})$, we have

$$K_{i_k} d(\sigma_{k-1}) K_{i_k}^\dagger = \frac{K_{i_k} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_k}^\dagger}{\mathrm{Tr}\left(K_{i_{k-1}} \ldots K_{i_1} d(\rho_0) K_1 \ldots K_{i_{k-1}}\right)}, \tag{43}$$

and

$$\begin{aligned} &\mathrm{Tr}\left(K_{i_k} d(\sigma_{k-1}) K_{i_k}^\dagger\right) \\ &\quad = \mathrm{Tr}\left(\frac{K_{i_k} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_k}^\dagger}{\mathrm{Tr}\left(K_{i_{k-1}} \ldots K_{i_1} d(\rho_0) K_1 \ldots K_{i_{k-1}}\right)}\right) \qquad (44) \\ &\quad = \frac{\mathrm{Tr}\left(K_{i_k} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_k}^\dagger\right)}{\mathrm{Tr}\left(K_{i_{k-1}} \ldots K_{i_1} d(\rho_0) K_1 \ldots K_{i_{k-1}}\right)}, \qquad (45) \end{aligned}$$

Substituting Eqs. (43) and (45) for the numerator and denominator of Eq. (42), respectively, and canceling the traces, we find

$$\sigma_k = \frac{K_{i_k} \ldots K_{i_1} d(\rho_0) K_{i_1} \ldots K_{i_k} \oplus 0}{\mathrm{Tr}\left(K_{i_k} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_k}^\dagger\right)}. \tag{46}$$

■

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 $\{A^{i_1}, \ldots, A^{i_n}\}$ should correspond to the probability of observing measurements $\{i_1, \ldots, i_n\}$ in the original QMOP instance.

Lemma 6. Let $\mathbf{S} = \{K_1, \ldots, K_\mathcal{K}\}$ 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

$$\Pr(\rho_n \neq \rho_g) = \mathrm{Tr}\left(K_{i_n} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_n}^\dagger\right). \tag{47}$$

Proof. First consider the probability that ρ_n is not ρ_g given that $\rho_{n-1} \neq \rho_g$. By Lemma 3, if $\rho_{n-1} \neq \rho_g$ then $\rho_{n-1} = \sigma_{n-1}$. By Lemma 5,

$$\sigma_{n-1} = \frac{K_{i_{n-1}} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_{n-1}}^\dagger \oplus 0}{\mathrm{Tr}\left(K_{i_{n-1}} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_{n-1}}^\dagger\right)}, \tag{48}$$

so

$$d(\sigma_{n-1}) = \frac{K_{i_{n-1}} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_{n-1}}^\dagger}{\mathrm{Tr}\left(K_{i_{n-1}} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_{n-1}}^\dagger\right)}. \tag{49}$$

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:

$$\begin{aligned} &\Pr(\rho_n \neq \rho_g | \rho_{n-1} \neq \rho_g) \\ &\quad = \mathrm{Tr}\left(A_{d+2}^{i_n} \sigma_{n-1} {A_{d+2}^{i_n}}^\dagger\right) \qquad (50) \\ &\quad = \mathrm{Tr}\left(K_{i_n} d(\sigma_{n-1}) K_{i_n}^\dagger \oplus 0\right) \qquad (51) \\ &\quad = \mathrm{Tr}\left(K_{i_n} d(\sigma_{n-1}) K_{i_n}^\dagger\right) \qquad (52) \\ &\quad = \frac{\mathrm{Tr}\left(K_{i_n} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_n}^\dagger\right)}{\mathrm{Tr}\left(K_{i_{n-1}} \ldots K_{i_1} d(\rho_0) K_{i_1}^\dagger \ldots K_{i_{n-1}}^\dagger\right)}. \qquad (53) \end{aligned}$$

Note that $\Pr(\rho_n \neq \rho_g | \rho_{n-1} = \rho_g) = 0$ by Lemma 2. The total probability that ρ_n is not ρ_g is

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

■

Since the probability that the agent observes the sequence of measurements $\{i_1, \ldots, i_n\}$ is the same as the probability that the sequence of actions $\{A^{i_1}, \ldots, A^{i_n}\}$ 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 4 (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 $\mathbf{S} = \{K_1, \ldots, K_\mathcal{K}\}$ 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 3, $Q(\mathbf{S})$ 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 $\{i_1, \ldots, i_n\}$ such that $K_{i_1}^\dagger \ldots K_{i_n}^\dagger K_{i_n} \ldots K_{i_1} = 0$.

First assume there is some sequence $\{i_1, \ldots, i_n\}$ such that $K_{i_1}^\dagger \ldots K_{i_n}^\dagger K_{i_n} \ldots K_{i_1} = 0$. Consider the time-dependent policy that takes action A^{i_k} in after k time steps no matter the state.

By Lemma 6, the probability that this policy is not in the goal state after n time steps is

$$\Pr(\rho_n \neq \rho_g) = \mathrm{Tr}\left(K_{i_n}...K_{i_1}d(\rho_0)K_{i_1}^\dagger...K_{i_n}^\dagger\right) \quad (54)$$

$$= \mathrm{Tr}\left(K_{i_1}^\dagger...K_{i_n}^\dagger K_{i_n}...K_{i_1}d(\rho_0)\right) \quad (55)$$

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

$$= 0 \quad (57)$$

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 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 6, the probability that the state at time step n is not ρ_g is

$$\Pr(\rho_n \neq \rho_g|\pi) = \mathrm{Tr}\left(K_{i_1}...K_{i_n}d(\rho_0)K_{i_1}^\dagger...K_{i_n}^\dagger\right) \quad (58)$$

$$= \mathrm{Tr}\left(K_{i_1}^\dagger...K_{i_n}^\dagger K_{i_n}...K_{i_i}d(\rho_0)\right). \quad (59)$$

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

$$K_{i_1}^\dagger...K_{i_n}^\dagger K_{i_n}...K_{i_i} = 0. \quad (60)$$

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 [15], 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 non-negative elements. Since we are interested in a probability 1 event, we can treat probabilities as binary: either positive or 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.

Definition 7 (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 Eq. (4). Throughout this section, we assume without loss of generality that g is the $|S|$th state in P so $(\vec{b}_g)_i = \delta_{i|S|}$. 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 $(\vec{z}_g)_i = \delta_{i|S|}$ and $(\vec{z}_0)_i = 1$ if and only if $(\vec{b}_0)_i > 0$. The transition function Z for action a nondeterministically 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

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

Let

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

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

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

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 non-negative numbers only.

Lemma 7. 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

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

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 $|\tau^{ao}\vec{b}|_1 = 0$.

Proof. Using the definition of Z^{ao} from Eq. (61),

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

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

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

First assume $\vec{b}_i^{ao} = 0$. Since $\tau_{ij}^{ao} \geqslant 0$ and $\vec{b}_j \geqslant 0$, we must have that every term in the sum in Eq. (65) is zero individually [16]. 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

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

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

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

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

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

Now let $|\tau^{ao}\vec{b}|_1 > 0$. Then there is some k with $\tau^{ao}_{ik}\vec{z}_k > 0$ by the same reasoning as above. Therefore $\tau^{ao}\vec{z} \neq \vec{0}$ so $P^o_a(\vec{z}) = 1$. ■

We now show that the agent can reach the goal in the binary probability MDP with probability 1 if and only if it could reach the goal in the original POMDP with probability 1. We do each direction in a separate lemma.

Lemma 8. 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^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 $M(P) = \langle B, A, \tau, \vec{b}_0, \vec{b}_g\rangle$.

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

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

Therefore $\vec{z}^{k+1} = z(\vec{b}^{k+1})$ by Lemma 7.

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 zero. Therefore, we must have

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

for all k, so Lemma 7 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 9. 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 $D(P)$.

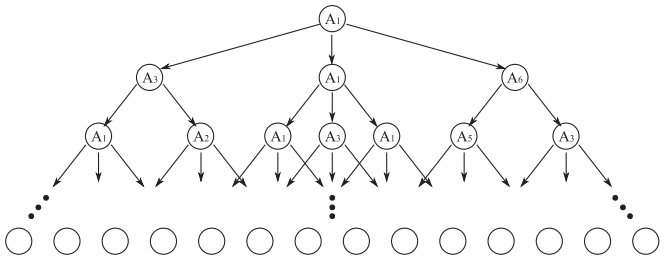

FIG. 3. A policy in an MDP creates a tree. Here, the agent takes action A_1 in the starting state, which can transition the world state nondeterministically to three other possible states. The policy specifies an action of A_3 for the state on the left, A_1 for the state in the middle, and A_6 for the state on the right. Taking these actions transitions these states nondeterministically. 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 Fig. 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 zero of π_T.

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

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

Therefore by Lemma 7

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

has entry i zero 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 7. 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 $\vec{z}^{k+1}_i = 0$ so the level of $b(\vec{z}^{k+1})$ is at least the level of $\vec{b}'$. Therefore the level of $b(\vec{z}^{k+1})$ is 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 5 (Decidability of goal-state reachability for POMDPs). The goal-state reachability problem for POMDPs is decidable.

Proof. We showed in Lemmas 8 and 9 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 of great interest 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 analogs of POMDP algorithms or even MDP ones?

ACKNOWLEDGMENTS

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

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

[2] J. Pineau, G. Gordon, and S. Thrun, in *Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico 2003*, edited by G. Gottlob and T. Walsh (Morgan Kauffman, San Francisco, California), pp. 1025–1032.

[3] S. Russell and P. Norvig, in *Artificial Intelligence: A Modern Approach*, 2nd ed. (Prentice-Hall, Englewood Cliffs, NJ, 2003), Chap. 17, pp. 613–648.

[4] M. T. J. Spaan and N. Vlassis, J. Artif. Intell. Res. **24**, 195 (2005).

[5] T. Smith and R. Simmons, in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence Banff, Canada* (AUAI Press, Arlington, Virginia, 2004), pp. 520–527.

[6] S. Ying and M. Ying, arXiv:1406.6146v2.

[7] S. Ying, N. Yu, and M. S. Ying, in *Proceedings of the 24th International Conference on Concurrency Theory, Buenos Aires, Argentina, 2013*, edited by P. R. D'Argenio and H. Melgratti (Springer, Heidelberg, 2013), pp. 334–338.

[8] C. J. Combes, M. Tiersch, G. Milburn, H. Briegel, and C. Caves, arXiv:1405.5656.

[9] M. Tiersch, E. Ganahl, and H. Briegel, arXiv:1407.1535.

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

[11] O. Madani, S. Hanks, and A. Condon, in *Proceedings of the 16th National Conference on Artificial Intelligence* (AAAI Press, Menlo Park, California, 1999), pp. 541–548.

[12] 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.

[13] M. Neilson and I. Chuang, *Quantum Computation and Quantum Information*, 10th ed. (Cambridge University Press, Cambridge, UK, 2011).

[14] J. Eisert, M. P. Müller, and C. Gogolin, Phys. Rev. Lett. **108**, 260501 (2012).

[15] J. Rintanen, in *Proceedings of the 14th International Conference on Automated Planning and Scheduling* (AAAI Press, Menlo Park, California, 2004), pp. 345–354.

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