Sampling-based Algorithms for Continuous-time POMDPs

Pratik Chaudhari* Sertac Karaman* David Hsu† Emilio Frazzoli*

***Abstract*— This paper focuses on a continuous-time, continuous-space formulation of the stochastic optimal control problem with nonlinear dynamics and observation noise. We lay the mathematical foundations to construct, via incremental sampling, an approximating sequence of discrete-time finite-state partially observable Markov decision processes (POMDPs), such that the behavior of successive approximations converges to the behavior of the original continuous system in an appropriate sense. We also show that the optimal cost function and control policies for these POMDP approximations converge almost surely to their counterparts for the underlying continuous system in the limit. We demonstrate this approach on two popular continuous-time problems, viz., the Linear-Quadratic-Gaussian (LQG) control problem and the light-dark domain problem.**

I. Introduction

Uncertainty, whether it arises from unmodeled dynamics or from imprecise sensors, forms a significant part of most systems. Control of such systems, many of which have continuous-time dynamics, necessitates a formulation that explicitly accounts for this uncertainty. Stochastic differential equations (SDEs) have been a popular approach to address different aspects of problems in control of general, continuous-time systems with uncertainty [1]. It is thus tempting to formulate the stochastic optimal control problem for robotic systems as control of SDEs. However, although they have long been a focus of control theory, closed form, analytical solutions for such models are hard to come by; solutions for only a few special cases, e.g, linear dynamics with Gaussian noise, can be computed easily [2].

On the other hand, more recent literature in the context of artificial intelligence and robotics, focuses on discrete-time and discrete-state models using partially-observable Markov decision processes (POMDPs). In this formulation, the robot and its environment are expressed by a finite number of states. The dynamics is governed by stochastic transitions which depend on the particular action chosen. However, the state of the underlying Markov decision process (MDP) is not directly observable to the robot; only its noisy observations are available. The problem of stochastic control for this model is often formulated as optimizing the expected value of some performance metric of this discrete system, where the expectation is taken over all possible realizations of noise.

Similar to the continuous-time case, solving discrete POMDPs is computationally challenging, it is in fact PSPACE-hard [3]. Despite this, there are a number of general-purpose algorithms that have been demonstrated to work on challenging examples. An enabling idea behind these algorithms has been the notion of "belief space", which is defined as the space of all probability distributions over the set of states. The problem is then reformulated with the dynamics consisting of controlled stochastic transitions in the belief space. A particularly successful class of algorithms, often called *point-based methods*, search for an optimal policy by sampling only the reachable belief space [4], [5]. These approaches are tailored to solve discrete-time discrete-space POMDPs; algorithmic tools for continuous-time continuous-space formulations of the problem have received relatively little attention so far. Only recently, point-based solvers have been adapted to continuous state-spaces by simulating the continuous-time system using particle-based methods [6] whereas continuous observation spaces have been studied in [7].

In this paper, we consider a continuous-time continuous-space system described by a set of stochastic differential equations. We propose a two-stage approach. First, we generate a sequence of discrete-time discrete-space POMDPs that approximate, in some suitable sense, the original continuous system. We then solve these POMDP approximations using an existing algorithm. We show that the resulting cost function and controllers converge to the optimal cost function and controller for the original continuous system in the limit. Inspired by recent advances in sampling-based optimal motion-planning [8], these POMDP approximations are constructed incrementally in a computationally-efficient manner using random sampling. We also demonstrate the proposed approach on the Linear-Quadratic-Gaussian (LQG) control problem and the light-dark domain problem.

The paper is organized as follows. After some preliminaries in Section II, the general problem is formulated in Section III. An algorithm for constructing POMDP approximations is outlined in Section IV its convergence properties are analyzed in Section V. Section VI discusses computational experiments with conclusions and directions for future work provided in Section VII.

II. Preliminaries

We introduce some notation and the Markov chain approximation method for constructing discrete-Markov chain approximations for continuous-time processes in this section.

A. Markov Chains

A *Markov chain* (MC) is denoted by the tuple $M = (S, P, z_0)$, where S is a finite set of states and $P : S \times S \to [0, 1]$ is a function that denotes the transition probabilities.

*The authors are with the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology. Email: pratikac@mit.edu, sertac@mit.edu, frazzoli@mit.edu

†The author is with the School of Computing at the National University of Singapore. Email: dyhsu@comp.nus.edu.sg

For convenience, we denote by $P(z \mid z')$, the probability that the next state is z given that the current state is z'. The random process denoted by $\{\xi_i;\ i \in \mathbb{N}\}$ is the (discrete-time) trajectory of the Markov chain M starting from z_0.

B. Markov Decision Processes

A *Markov Decision Process* (MDP) is a tuple $M = (S, U, P, z_0)$ where S is a finite set of states, U is a finite set of controls, $P : S \times U \times S \to [0, 1]$ is a transition probability function and $z_0 \in S$ is the initial state. The dynamics of the process has the Markov property, i.e., a control action $u \in U$ at a state $z \in S$ results in a new state $z' \in S$ with a probability that we denote as $P(z' \mid z, u)$.

A *policy* for a MDP is a mapping $\pi : S \to U$ that assigns a control action to each state. The *cost function* corresponding to a policy π for T time units is defined to be:

$$J_\pi(\bar{z}) = \mathbb{E}\left[\sum_{k=1}^{T} \gamma^k l\big(z(k), \pi(z(k))\big) + L\big(z(T)\big) \mid z_0 = \bar{z}\right],$$

where $\gamma < 1$ is the discount factor, $l : S \times U \to \mathbb{R}_{>0}$ is the running cost, and $L : S \to \mathbb{R}_{>0}$ is the terminal cost. Let Π denote the (finite) set of all policies. An *optimal policy* is a policy $\pi^* \in \Pi$ such that $J_{\pi^*}(z) = \min_{\pi \in \Pi} J_\pi(z)$ for all $z \in S$. Denote the optimal cost function by $J_{\pi^*}(\cdot)$.

C. Partially-observable Markov Decision Processes

A (discrete-time discrete-state) *partially-observable Markov decision process* (POMDP) is a tuple $M = (S, U, O, P, Q, b_0)$ such that S is a finite set of states, U is a finite set of controls, O is a finite set of observations, $P : S \times U \times S \to [0, 1]$ are the transition probabilities, $Q : S \times O \to [0, 1]$ are the observation probabilities, and $b_o : S \to [0, 1]$ is the initial distribution of states. A control action $u \in U$ at a state z results in $z' \in S$ with probability $P(z' \mid z, u)$. However, the process is only partially observable, an observation $o \in O$ is observed at $z \in S$ with probability $Q(z, o)$, denoted as $Q(z \mid o)$.

A *belief* is a probability mass function over the set of states $b : S \to [0, 1]$. Starting from an initial state z_0, drawn from a distribution b_0, applying a sequence of control actions, $u(k)$, where $k \in \{1, 2, \ldots, T\}$ results in a sequence of observations which we denote by $o(k)$. A distribution of possible states for each time step, computed using Bayes law, is called the *belief trajectory*, simply denoted by $b(k)$. Let B denote the (infinite) set of all beliefs. A *policy* is a function $\pi : B \to U$ that assigns a control to each belief. Under a policy π, the control $\pi(b)$ is executed when the current belief is $b \in B$. The *cost function* can be written as

$$J_\pi(\bar{b}) = \mathbb{E}\left[\sum_{k=1}^{T} \gamma^k l(b(k), \pi(b(k))) + L(b(T)) \mid b_0 = \bar{b}\right],$$

where γ is the discount factor, $l : B \times U \to \mathbb{R}_{>0}$ is the running cost, and $L : B \times U \to \mathbb{R}_{>0}$ is the terminal cost. Let Π denote the (infinite) set of all policies. An *optimal policy* is a policy $\pi^* \in \Pi$ such that $J_{\pi^*}(b) = \inf_{\pi \in \Pi} J_\pi(b)$ for all $b \in B$. The function J_{π^*} is the optimal cost function.

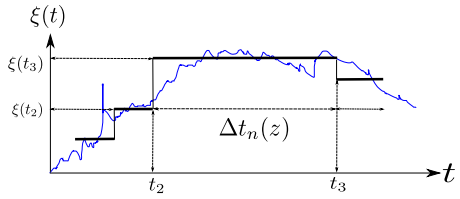


Fig. 1: Blue : Trajectory of the original stochastic system, Bold black : interpolated trajectory $\xi(t)$, $\Delta t = t_3 - t_2$ is the holding time of state $\xi(t_3)$.

D. The Markov Chain Approximation Method

In this paper, we are interested in continuous-time continuous-state processes. A widely-accepted continuous-time analogue of a Markov chain is the following stochastic differential equation (SDE):

$$dx(t) = f(x(t))\, dt \ + \ F(x(t))\, dw(t), \quad x(0) = x_0 \tag{1}$$

where $\{w(t) : t \in \mathbb{R}_{\geq 0}\}$ is the standard k-dimensional Wiener process, $x(t) \in \mathcal{S} \subset \mathbb{R}^d$ is the state, $x_0 \in \mathcal{S}$ is the initial state while $f : \mathbb{R}^d \to \mathbb{R}^d$ is the drift vector and $F : \mathbb{R}^d \to \mathbb{R}^{d \times k}$ is the diffusion matrix. The solution to this SDE is a stochastic process $\{x(t) : t \in \mathbb{R}_{\geq 0}\}$ that satisfies,

$$x(t) = x(0) + \int_0^t f(x(\tau))\, d\tau \ + \ \int_0^t F(x(\tau))\, dw(\tau), \tag{2}$$

where the last term is the usual Itô integral. The Markov chain approximation method, proposed by Kushner (see, e.g., [9]), provides a set of conditions under which a sequence of (discrete) Markov chains, approximate the original continuous process described by the SDE above.

For a Markov chain, $M = (S, P, z_0)$, let $\Delta t : S \to \mathbb{R}_{>0}$ be a function that assigns to each state, a positive real number $\Delta t(z)$ called the *holding time*. Let the continuous-time interpolation, $\xi(t)$, of the discrete Markov chain trajectory $\{\xi_i;\ i \in \mathbb{N}\}$, be given by $\xi(t) = \xi_i$ for all $[t_i, t_{i+1})$, where $t_i = \sum_{j=1}^{i} \Delta t(\xi_j)$. Roughly, the Markov chain spends a time $\Delta t(z)$ at state $\xi_i = z$ before making a transition. Fig. 1 shows an example interpolated trajectory. Similarly, we can also define the corresponding interpolated belief trajectory of a POMDP as $b(t) = b(t_i)$ for all $[t_i, t_{i+1})$.

Let $\{M_n = (S_n, P_n, z_{0,n}) : n \in \mathbb{N}\}$ be a sequence of Markov chains, Δt_n be a sequence of holding times and $\{\xi_i^n : i \in \mathbb{Z}_{\geq 0}\}$ be trajectories of M_n. The sequence M_n along with the sequence Δt_n is said to be *locally consistent* [9] with the original system described by Eqn. (1) if the following conditions are satisfied for all $z \in S$.

$$\circ \quad \lim_{n \to \infty} \Delta t_n(z) = 0, \tag{3}$$

$$\circ \quad \lim_{n \to \infty} \frac{\mathbb{E}[\xi_{i+1}^n - \xi_i^n \mid \xi_i^n = z]}{\Delta t_n(z)} = f(z), \tag{4}$$

$$\circ \quad \lim_{n \to \infty} \frac{\mathrm{Cov}[\xi_{i+1}^n - \xi_i^n \mid \xi_i^n = z]}{\Delta t_n(z)} = F(z)F(z)^T. \tag{5}$$

where $\mathrm{Cov}(x) = \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T]$. Local consistency implies that interpolated trajectories of successive Markov chains converge in distribution to trajectories of the stochastic differential equation given by Eqn. (1). This statement is made precise by the following theorem.

Theorem 1 (Thm. 10.4.1 in [9]) *If $\{M_n : n \in \mathbb{Z}_{\geq 0}\}$ is a sequence of Markov chains and Δt_n is a sequence of holding times satisfying local consistency conditions, then the sequence of trajectories $\xi_n(\cdot)$ has a subsequence that converges in distribution to $x(\cdot)$ that satisfies Eqn. (2).*

Let us note two recent approaches based on this method that use sampling-based techniques inspired from motion planning literature to create discrete MDPs [10] and Hidden Markov Models (HMMs) [11]. Roughly speaking, our approach merges these two ideas to create efficient POMDPs.

III. Problem Formulation and Approach

We formulate the continuous-time, continuous-state partially observed stochastic control problem in this section.

Problem 2 *Consider the stochastic dynamical system:*

$$\begin{aligned} dx(t) &= f(x(t), u(t))\ dt + F(x(t))\ dw(t) \\ dy(t) &= g(x(t))\ dt + G(x(t))\ dv(t) \end{aligned} \tag{6}$$

where $\{w(t) : t \in \mathbb{R}_{\geq 0}\}$ and $\{v(t) : t \in \mathbb{R}_{\geq 0}\}$ are independent k-dimensional and l-dimensional standard Wiener processes, $x(t) \in \mathcal{S} \subset \mathbb{R}^d$, $u(t) \in \mathcal{U} \subset \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$, $f : \mathcal{S} \times \mathcal{U} \to \mathbb{R}^d$, $F : \mathcal{S} \to \mathbb{R}^{d \times k}$, $g : \mathcal{S} \to \mathbb{R}^p$, and $G : \mathcal{S} \to \mathbb{R}^{p \times l}$. Find a control $\pi(t) \in \mathcal{U}$ which is a non-anticipating functional of the observation process, such that,

$$J_\pi = \mathbb{E}\Big[\int_0^T l(x(t), \pi(\{y(\tau) :\ 0 \leq \tau \leq t\}), t)\ dt + L(x(T)) \ \Big|\ x(0) = x_0 \Big]. \tag{7}$$

is minimized. x_0 is a random variable with distribution b_0. The terminal time T (finite or infinite) is defined as the exit time from a compact set K, i.e., $T = \inf\{t :\ x(t) \notin K^o\}$.

We tacitly assume that the sets $\mathcal{S}, \mathcal{U}$ are bounded and functions f, F, g, G, l and L are continuous and bounded on bounded intervals to guarantee existence and uniqueness of solutions (see [1]). Note that the observation process $y(t)$ given above is equivalent to the more popular version given as $y(t) = g'(x(t)) + G'(x(t))\ \tilde{v}$ where $\tilde{v}$ is white Gaussian noise [1]. The conditional expectation in Eqn. (7) will be clear from the context and is henceforth dropped. The following lemma shows that the above cost function also admits a belief space representation.

Lemma 3 *Prob. 2 is equivalent to minimizing*

$$J'_{\pi'} = \mathbb{E}\Big[\int_0^T l'(b(t), \pi(b(t)), t)\ dt + L'(b(T)) \ \Big|\ b(0) = b_0 \Big].$$

for some functions $l'(\cdot)$ and $L'(\cdot)$.

Proof: Let $u(t) = \pi(\{y(\tau) : \tau \leq t\})$. Using the Law of Iterated Expectations and Fubini's Thm.,

$$J_\pi = \mathbb{E}\left[\int_0^T \mathbb{E}[l(x, u, t) \,|\, \mathcal{F}_t^y]\ dt \right] + \mathbb{E}\Big[\ \mathbb{E}[L(x(T) \,|\, \mathcal{F}_T^y]\ \Big]$$

Note that since $b(t)$ is a sufficient statistic, i.e., it contains all information needed for control in the POMDP problem [12], we can calculate a new policy $\pi' : B \to U$ from $\pi(\cdot)$. This is equivalent to $J'_{\pi'}$ with

$$\begin{aligned} l'(\cdot, \cdot, \cdot) &= \mathbb{E}[l(x, u, t) \,|\, \mathcal{F}_t^y] = \int_{\mathcal{S}} l(x, u, t)\ b(x(t))\ dx(t), \\ L'(\cdot) &= \mathbb{E}[L(x(T)) \,|\, \mathcal{F}_T^y] = \int_{\mathcal{S}} L(x(T)) b(x(T))\ dx(T). \end{aligned}$$

■

Our approach to solving Prob. 2 can be briefly summarized as follows. We first generate a discrete-time discrete-space POMDP that approximates the continuous-time continuous-space stochastic system in Eqn. (6). We then use an existing POMDP solver, SARSOP [5], to obtain a policy for the POMDP approximation. The following section describes the construction of discrete POMDP approximations using sampling-based methods.

IV. Constructing POMDP Approximations

A. Primitive procedures

A few preliminary procedures required are as follows.

1) Sampling: For $x \in \mathcal{S} \subset \mathbb{R}^d$, the `SampleState` procedure returns states sampled independently and uniformly from $\mathcal{S}$. `SampleControl` samples control inputs uniformly randomly from the set of admissible controls, $\mathcal{U}$.

2) Neighboring states: The procedure $\texttt{Near}(z, S)$ returns all states within a distance of $r = \gamma_s\ (\log n/n)^{1/d}$ from z,

$$Z_{near} = \Big\{ z_k \in S,\ z_k : \|z_k - z\|_2 \leq \gamma_s\ (\log n/n)^{1/d} \Big\}$$

where $n = |S|$, $d = \dim(S)$ and $\gamma_s > 0$ is a constant calculated in Thm. 4 of [11].

3) Holding Time: Given $z \in S$ and $u \in U$, the $\texttt{ComputeHoldingTime}(z, u, S)$ procedure returns the holding time computed as $\Delta t(z, u) = \frac{r^2}{\|F(z)F^T(z)\|_2 + r\|f(z,u)\|_2}$, where r is as given in the procedure $\texttt{Near}(z, S)$.

4) Transition Probabilities: Let $|U| = m$, i.e., $u_1, \ldots u_m \in U \subset \mathcal{U}$. For every one of these controls, the $\texttt{ComputeTransProb}(z, u, \delta)$ procedure uses local consistency conditions to calculate transition probabilities as,

$$\begin{aligned} \mathbb{E}[\Delta\xi(z, u)] &= f(z, u)\ \delta \\ \mathrm{Cov}[\xi_{i+1}^n - \xi_i^n \,|\, \xi_i^n = z, u_i^n = u] &= F(z)F^T(z)\ \delta. \end{aligned}$$

Let the transition probabilities be $p_k = \mathbb{P}(z_k \,|\, z, u)$ for all $z_k \in Z_{near}$ with $\mathbb{P}(z' \,|\, z, u) = 0$ if $z' \notin Z_{near}$. These conditions are a set of linear equations for probabilities p_k. They can also be obtained using a small-time approximation [11].

5) Observation Probabilities: Given a state $z \in S$, the procedure $\texttt{ComputeObsProb}(z)$ returns

$$Q(z' \,|\, z) = \mathbb{P}(z' \,|\, o) = \eta\ N(g(z'), g(z), G(z)\ G^T(z)),$$

where $o = g(z)$ and $z' \in Z_{near}(z)$. $N(x, \mu, \Sigma)$ denotes the probability density of a normal random variable with mean μ and variance Σ calculated at x and η is a normalizing constant. Note that (i) this procedure can be modified suitably for cases where observation noise is not Gaussian and, (ii) we assume that the set of states and observations are same.

6) Connect State: `ConnectState` computes transition and observation probabilities for a given state $z \in S$.

B. Algorithm

The "batch construction" in Alg. 1 takes a set of n sampled states and m sampled controls to construct a discrete model of Prob. 2. On the other hand, Alg. 3 incrementally refines

Algorithm 1: Batch POMDP

1 $U_0 = \emptyset$, $S_0 = \emptyset$;
2 **for** $k \leq m$ **do**
3 $\quad u \leftarrow$ `SampleControl`;
4 $\quad U_k \leftarrow \{u\} \cup U_{k-1}$;
5 **for** $k \leq n$ **do**
6 $\quad z \leftarrow$ `SampleState`;
7 $\quad S_k \leftarrow \{z\} \cup S_{k-1}$;
8 **for** $z \in S_n$ **do**
9 $\quad$ **for** $u \in U_m$ **do**
10 $\quad\quad \Delta t_n(z,u) \leftarrow$ `ComputeHoldingTime`(z,u,S_n);
11 $\delta_n \leftarrow \min_{z \in S_n, u \in U_m} \Delta t_n(z,u)$;
12 **for** $z \in S_n$ **do**
13 $\quad$ `ConnectState`$(z, S_n, U_m, P_n, Q_n, \delta_n)$;
14 **return** $(S_n, U_m, P_n, Q_n, \delta_n)$;

Algorithm 2: `ConnectState`(z, S, U, P, Q, δ)

1 **for** $u \in U_m$ **do**
2 $\quad P(\cdot \,|\, z,u) \leftarrow$ `ComputeTransProb`(z,u,δ);
3 $\quad Q(\cdot \,|\, z) \leftarrow$ `ComputeObsProb`(z);

Algorithm 3: Incremental construction of POMDP

1 $z \leftarrow$ `SampleState`;
2 $S_{n+1} \leftarrow \{z\} \cup S_n$;
3 $U \leftarrow U_m$;
4 **if** $\lfloor c\,(\log(n+1) - \log n) \rfloor > 1$ **then**
5 $\quad u \leftarrow$ `SampleControl`;
6 $\quad U \leftarrow \{u\} \cup U_m$;
7 `ConnectState`$(z, S_{n+1}, U, P_n, Q_n, \delta_n)$;
8 **if** $\min_{u \in U} \Delta t_{n+1}(z_{n+1}, u) \leq \delta_n$ **then**
9 $\quad \delta_{n+1} = \delta_n / 2$;
10 $\quad$ **for** $z \in S_{n+1}$ **do**
11 $\quad\quad$ `ConnectState`$(z, S_{n+1}, U, P_{n+1}, Q_{n+1}, \delta_{n+1})$;
12 **else**
13 $\quad \delta_{n+1} \leftarrow \delta_n$;
14 $\quad$ **for** $z' \in Z_{\text{near}}(z_{n+1})$ **do**
15 $\quad\quad$ `ConnectState`$(z', S_{n+1}, U, P_{n+1}, Q_{n+1}, \delta_{n+1})$;
16 $U_{m+1} \leftarrow U$;
17 **return** $(S_{n+1}, U_{m+1}, P_{n+1}, Q_{n+1}, \delta_{n+1})$;

the POMDP created by the batch construction. In other words, it creates a new POMDP M_{n+1} from M_n by sampling an addition state z_{n+1} and control u_{m+1}. New control inputs are sampled to ensure that $|U| = m = \mathcal{O}(\log n)$. Transition probabilities of all states using the new control u_{m+1} need to be recalculated. However, it is can be shown that by recalculating probabilities only in the set Z_{near} (Lines 14–15), every state $z \in S_n$ will have transition probabilities using the new control after finitely many iterations (see Thm. 5 of [11]). The equalized holding time δ_n (Alg. 1, Line 11) is refined incrementally as $\delta_{n+1} = \delta_n/2$ and all transition probabilities are recalculated every time we add a new state that has $\Delta t(z_{n+1}) \leq \delta_n$ (Lines 8–11). The amortized complexity of Alg. 3 can be shown to be $\mathcal{O}((\log n)^2)$ per iteration [11].

Given a POMDP approximation M_n created using the above algorithms, we obtain the optimal cost function using SARSOP. An equivalent discrete cost function J_n that approximates Eqn. (7), e.g., $J = \int_0^\infty e^{-2\alpha t} l(x,u) dt$ is,

$$J_n \sim \sum_{k=0}^{\infty} e^{-2\alpha k \delta_n}\ l(x,u)\ \delta_n = \sum_{k=0}^{\infty} \gamma_n^k\ l'(x,u)$$

where $\gamma_n = e^{-2\alpha\delta_n}$, $l'(x,u) = l(x,u)\delta_n$ and $\alpha > 0$.

V. Analysis

In this section, we first prove that interpolated belief trajectories of a POMDP approximation, $b_n(\cdot)$, converge weakly to belief trajectories of the original system, $b(\cdot)$. Weak convergence will then imply that the cost function calculated on M_n converges to the optimal cost function almost surely. A technical construction known as "relaxed controls" will then be used to prove that the control policies also converge with probability one. The analysis in this section follows the analysis for the fully observed stochastic control problem in [9]. However, there are technical differences due to the fact that we are working with convergence in function spaces. For the sake of brevity, we only sketch important proofs.

A. Convergence of belief trajectories

Recall that from Thm. 1, trajectories of the Markov chain, i.e., $x_n(\cdot)$ converge in distribution to state trajectories of the original system, i.e., $x(\cdot)$. The belief of approximate POMDP is $b_n(t) = \mathbb{P}(x_n(t) \,|\, \mathcal{F}_t^{y,n})$ while the belief of the original stochastic system is $b(t) = \mathbb{P}(x(t) \,|\, \mathcal{F}_t^{y})$ where $\mathcal{F}_t^{y,n}$ denotes the filtration of n discrete observations. We shall use the notation $b_n(\cdot) \Rightarrow b(\cdot)$ to denote weak convergence.

Definition 4 *A probability measure P is tight if for each $\epsilon > 0$, there exists a compact set K such that $P(K) > 1 - \epsilon$. A sequence of measures P_n is tight if for every ϵ, there exists a compact set K such that $P_n(K) > 1 - \epsilon$ for all $n \in \mathbb{N}$.*

Tightness roughly means that we can always find a set K that contains most of the measure. It is essential to claim weak convergence of a sequence of measures in Thm. 7.

Lemma 5 *The sequence $b_n(\cdot)$ is tight.*

Proof: Given $b_n(\cdot)$ we will prove the conditions of Prokhorov's theorem to claim tightness. Let $\lambda_n(A,\omega) = \mathbb{P}_n(x_n(t) \in A \,|\, \mathcal{F}_t^y)$. Since $x_n(\cdot)$ is tight (see Thm. 10.4.1 in [9]), we have $\mathbb{P}_n(x_n(t) \in K_\epsilon) > 1 - \epsilon$ for all n. Thus,

$$\begin{aligned}
1 - \epsilon &< \mathbb{E}_n[\lambda(K_\epsilon)] \\
&= \int \lambda_n(K_\epsilon) \Big(\mathbf{1}_{\{\lambda_n(K_\epsilon) > 1 - 1/k\}} + \mathbf{1}_{\{\lambda_n(K_\epsilon) \leq 1 - 1/k\}} \Big) d\mathbb{P}_n \\
&< \frac{1}{k}\ \mathbb{P}_n\left(\lambda_n(K_\epsilon) > 1 - 1/k\right) + (1 - 1/k) \\
&\implies \mathbb{P}_n\left(\lambda_n(K_\epsilon) > 1 - 1/k\right) \geq 1 - \epsilon'/2^k \\
\implies \mathbb{P}_n\left(\lambda_n(K_\epsilon) > 1 - 1/k \quad \forall\ k\ \right) &\geq 1 - \epsilon'
\end{aligned}$$

where $\epsilon = \frac{\epsilon'}{k\ 2^k}$. This proves that the sequence of measures λ_n is tight by Prokhorov's theorem [13]. ∎

Theorem 6 (Thm. 2.1 in [14]) *Let X_n, Y_n be two random variables taking values in a Polish space S. Suppose (X_n, Y_n) defined on the probability space $(\Omega_n, \mathcal{F}_n, P_n)$ converges in distribution to (X, Y) defined on the space $(\Omega, \mathcal{F}, P)$. Suppose a measure Q_n exists such that (i) P_n be absolutely continuous with respect to Q_n for each n and, (ii) (X_n, Y_n) become independent under Q_n. If a corresponding distribution Q exists for (X, Y) and if Q_n converges weakly to Q, for every bounded continuous function $F : S \to \mathbb{R}$, $F(X_n)$ and $F(X)$ converge in distribution, i.e.,*

$$\mathbb{E}_{P^n}[F(X_n) \,|\, Y_n] \Rightarrow \mathbb{E}_P[F(X) \,|\, Y]$$

Furthermore, using the above result for random variables $x(\cdot)$ and $y(\cdot)$ given by Eqn. (6) we have,

$$\mathbb{E}_{P^n}[F(x_n(\cdot)) \,|\, \mathcal{F}_t^y] \Rightarrow \mathbb{E}_P[F(x(\cdot)) \,|\, \mathcal{F}_t^y].$$

We however require something stronger because the belief trajectory $b_n(\cdot)$ is calculated using sampled observations, i.e., observations in the set O_n, which generate the filtration $\mathcal{F}_t^{y,n}$. This is proved in the following theorem.

Theorem 7 ([15]) *Assume that the conditions of Thm. 6 are true. If the process $y(t)$ described by Eqn. (6) is approximated by a process $y_n(t)$ such that increments in $y_n(t)$ are Gaussian with zero mean, then,*

$$\mathbb{E}_{P^n}[F(x_n(\cdot)) \,|\, \mathcal{F}_t^{y,n}] \Rightarrow \mathbb{E}_P[F(x(\cdot)) \,|\, \mathcal{F}_t^y],$$

i.e., belief trajectories $b_n(\cdot)$ converge in distribution to $b(\cdot)$.

Proof: From Eqn. (6), corresponding to state $\xi_i = z$, the increments in observations are $\Delta y_i = G(z)\Delta v_i$ which is a Gaussian with zero mean and variance $G(z)G(z)^T$. ■

B. Relaxed Controls

Relaxed controls is a theoretical framework which roughly, compactifies the control space to ensure that control policies of the discrete POMDPs also converge [9]. Given a compact control space $\mathcal{U}$, let $\mathcal{B}(\mathcal{U})$ denote the σ-algebra of its subsets. A relaxed control is then a Borel measure $m(\cdot)$ such that $m(\mathcal{U} \times [0, T]) = t$ for all $t \geq 0$. The derivative $m_t(\cdot)$ is defined as $m_t(A) = \lim_{\delta \to 0} \frac{m(A \times [t-\delta, t])}{\delta}$. It can be shown that any relaxed control can be approximated arbitrarily well by an ordinary control.

As an example, consider the system $\dot{x}(t) = b(x, u)$, written using relaxed controls as $\dot{x}(t) = \int_{\mathcal{U}} b(x(t), \alpha)\ m_t(d\alpha)$. If the optimal control is non-unique with values ± 1, the corresponding relaxed control is given by $m_t(\cdot)$ which takes those values with equal probability, i.e., $m(A \times [0, T])$ is total control corresponding to the set $A \subset \mathcal{U}$ during the interval $[0, T]$. The solution of Eqn. (6) using control $m(\cdot)$ is,

$$x(t) = x(0) + \int_0^T \int_{\mathcal{U}} f(x, \alpha) m(d\alpha, ds) + \int_0^T F(x) dw \quad (8)$$

C. Convergence of cost function

Lemma 8 *The cost function J_n converges to J almost surely where, $J_n = \mathbb{E}\left[L(b_n(T), T) + \int_0^T l(b_n(t), t)\ dt\right]$ and $J = \mathbb{E}\left[L(b(T), T) + \int_0^T l(b(t), t)\ dt\right]$*

Proof: Thm. 7 proved that $b_n(t) \Rightarrow b(t)$. Define a continuous bounded function $f : \mathcal{B} \to \mathbb{R}$ as $f(b_n) = L(b_n(T), T) + \int_0^T l(b_n(t), t) dt$. By the Mapping Thm. [13] for a weakly convergent sequence b_n, we have that $f(b_n) \Rightarrow f(b)$. Since $f(b_n)$ and $f(b)$ converge in distribution, all moments converge almost surely, in particular, $\mathbb{E}[f(b_n)] \to \mathbb{E}[f(b)]$ almost surely. If the terminal time is infinite or an exit time from some compact set, T_n needs to be continuous under the measure induced by b_n to get $T_n \Rightarrow T$ (and $T_n \to T$ almost surely using Skorohod embedding). The above lemma still remains valid (see Thm. 9.4.3 in [9]). ■

Let the cost function using relaxed controls be given by,

$$W(b, m) := \mathbb{E}\left[\int_0^T \int_{\mathcal{U}} l'(b(s), \alpha, s)\ m(d\alpha\ ds) + L'(b(T))\right] \quad (9)$$

It can be shown that relaxed controls are continuous, i.e., if (b, m) be a solution such that it is ϵ away from the optimal cost, there exists a relaxed control m' such that $|W(b, m') - W(b, m)| \leq \delta$. Also, using Lem. 5 and Thm. 10.4.1 in [9], we can prove that any sequence $\{x_n, b_n, m_n, T_n\}$ contains a subsequence that converges to $\{x, b, m, T\}$ weakly if $x_n \Rightarrow x$.

The following theorem is the main result of this paper. It proves that the cost function approximation as calculated on the approximate Markov chain converges almost surely to the cost function of the original stochastic system. Since $m_n \Rightarrow m$ weakly, it also means that the relaxed controls converge in an almost sure sense in the Skorohod topology.

Theorem 9 (The Convergence Theorem) *Let $V_m(b)$ be the optimal cost function of Prob. 2 using relaxed controls, similarly let $V_{m_n}(b_n)$ be the optimal cost function as calculated on POMDP approximation M_n using relaxed controls. If we have $\{x_n, b_n, m_n, T_n\} \Rightarrow \{x, b, m, T\}$, then, $W(b_n, m_n) \to W(b, m) \geq V_m(b)$ almost surely. Also,*

$$\liminf_n\ V_{m_n}(b_n) \geq V_m(b)$$
$$\limsup_n\ V_{m_n}(b_n) \leq V_m(b)$$

Proof: (Sketch) Note that $W(b, m) \geq V_m(b)$ by definition. Skorohod representation of weak convergence gives,

$$\int_{\mathcal{U}} l'(b_n(s), \alpha, s)\ m_n(d\alpha\ ds) + L'(b_n(T)) \to \int_{\mathcal{U}} l'(b(s), \alpha, s)\ m(d\alpha\ ds) + L'(b(T))$$

almost surely, thereby giving $W(b_n, m_n) \to W(b, m) \geq V_m(b)$ almost surely using $T_n \xrightarrow{a.s.} T$. Using the same argument along with Fatou's lemma, we have, $\liminf_n W(b_n, m_n) \geq W(b, m)$; thereby giving $\liminf_n V_{m_n}(b_n) \geq V_m(b)$. Let $m_n^\epsilon(\cdot)$ be an adaptation of an "almost optimal" control m^ϵ to M_n. We have,

$$V_m(b_n) \leq W(b_n, m_n^\epsilon) \to W(b, m^\epsilon) \leq \epsilon + V_m(b)$$

thereby giving, $\limsup_n\ V_{m_n}(b_n) \leq V_m(b)$. ■

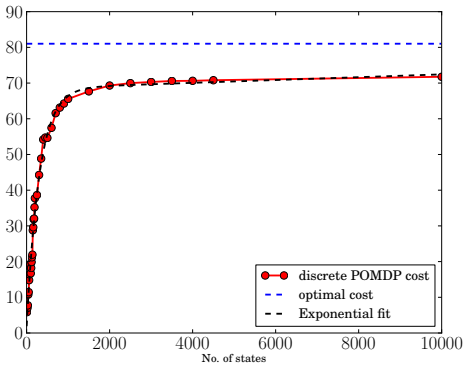


Fig. 2: Convergence of cost for 1-dimensional LQG problem

VI. Experiments

This section discusses simulation experiments where POMDP approximations of a continuous-time stochastic control problem are solved using SARSOP.

A. Linear Quadratic Gaussian (LQG)

If the dynamics and observations are linear and corrupted by white Gaussian noise, it turns out that we can solve Prob. 2 exactly. Consider a 1-dimensional linear system,

$$dx = -x\ dt + u\ dt + F\ dw$$
$$y(t) = x(t) + G\ \tilde{v}$$

where $\tilde{v}$ denotes unit variance white Gaussian noise. The objective then is to minimize a cost function of the form $J = \mathbb{E}\left[\int_0^5 (x^2 + u^2)\ dt\right]$. We use Alg. 3 to construct a discrete POMDP for the above dynamics $x \in [-1, 1]$ and $u \in [-1, 1]$ and $2 \log n$ uniformly sampled controls where n is the number of states in the POMDP. The discount factor in SARSOP is set to 0.99 to approximate a non-discounted cost function. Fig. 2 shows the cost of the policy obtained by solving discrete POMDPs with different number of states corresponding to the same continuous-time LQG problem with an optimal cost of 81.02. Note that the convergence slows down as the number of samples increases and as the problems grow larger, it becomes increasingly harder to search for an optimal policy for the discrete POMDP.

B. Light-dark domain

In this section, we test the proposed approach on a popular problem known as "light-dark domain" [16] which has been previously solved using techniques like belief-space planner [16] and convex optimization [17]. In this problem, a robot with noisy dynamics has to localize its position before entering a pre-defined goal region to obtain reward. There are regions in the state-space with beacons, i.e., light regions where highly accurate observations can be obtained while all other parts of the state-space are dark regions with large observation noise. Let the system be,

$$dx(t) = u(t)\ dt + F\ dw$$
$$y(t) = x(t) + G(x(t))\ \tilde{v}. \tag{10}$$

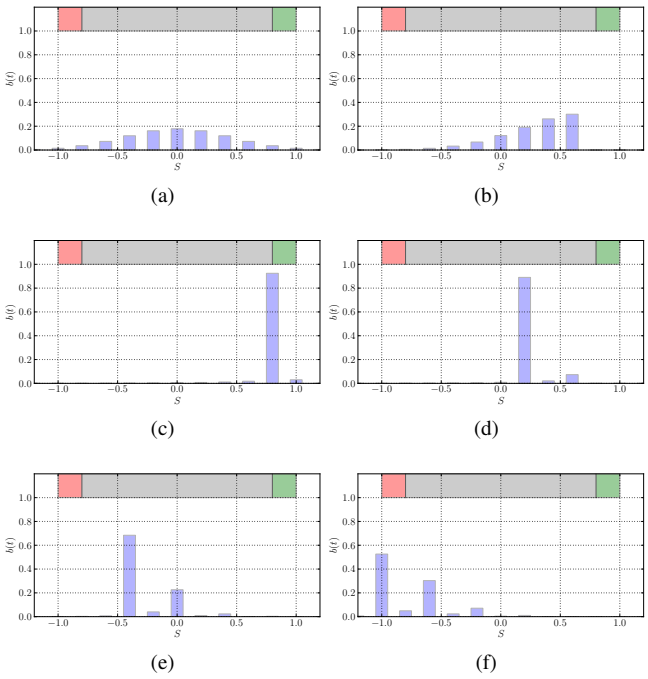


Fig. 3: An example policy calculated on a POMDP with 20 states for 1-dimensional Light-dark domain. Red denotes the goal region while the system has access to accurate observations in the green region. Blue rectangles denote the belief $b_n(t)$. These six figures show the belief at 6 different instants of policy execution.

where $x, y \in [-2, 2]$ while the observation noise is

$$G(x) = \begin{cases} \epsilon & : \ |x - b_1| < e_1 \\ 1/\epsilon & : \ \text{otherwise.} \end{cases}$$

Note that our formulation is non-convex and is a harder problem than a quadratic gradient in $G(x)$ considered in [16], [17]. A gradient ensures that even greedy policies can solve the problem whereas in our case, SARSOP is not aware of any good policy until it explicitly samples the light region.

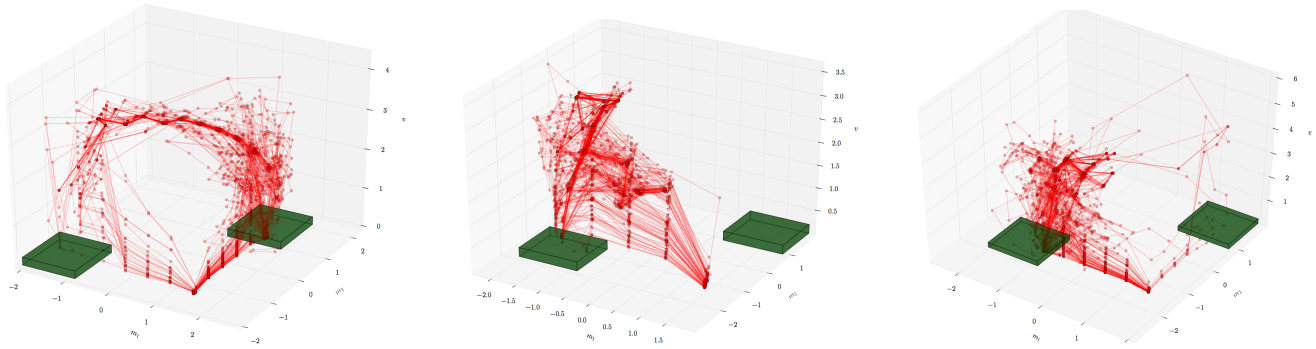
Define the goal region as $g = \{x : |x - g_1| < e_1, x \in \mathcal{S}\}$. The robot has $5 \log n$ actions $u \in [-1, 1]$ along with a terminal action called u_{goal} to claim the reward. This also determines the terminal time T. It gets a reward of 1000 if it reaches the goal region and a penalty of -1000 if it terminates at any other state. The reward function is

$$J = \mathbb{E}\left[-\sum_{k=0}^{T} \gamma^k l(x_k, u_k) + \gamma^T R(x(T))\right],$$

where $l(x_k, u_k) = \|u_k\ u_k^T\|_2 \delta_n$ is a quadratic cost.

1) Single Beacon: We will consider a 1-dimensional example with a single beacon first. An example policy calculated for $g_1 = -0.9, b_1 = 0.9, e_1 = 0.1$ is shown in Fig. 3. It is seen that the belief trajectory first travels to the light region to localize itself after which it proceeds to the goal to obtain the reward of 721 ± 40.3.

2) Two beacons: We demonstrate two aspects of our approach using the next example with two beacons, (i) incrementality of SARSOP and (ii) incremental refinement of POMDP approximations to get a better policy. Consider the dynamics in Eqn. (10) in two dimensions with two beacons



(a) 150 states, 100 sec, Reward: 212.1 $\pm$ 99.7 (b) 75 states, 200 sec, Reward: 377.9 $\pm$ 48.7 (c) 150 states, 200 sec, Reward: 428.6 $\pm$ 55.2

Fig. 4: This figure shows 1000 simulated belief trajectories for policies obtained for different discrete POMDPs. The XY plane represents the mean of the belief in the X, Y dimensions while the Z axis shows the 2-norm of the variance. The robot starts with a high variance before it localizes in the green light region to reach the goal region at (1.5, -1.5) with low variance.

placed at $b_1 = (1.4, 1.4)$ and $b_2 = (-1.4, -1.4)$, both of width $(1.2, 1.2)$ (shown in green). The initial position of the robot is at $(-1.5, -0.5)$ which means that it is closer to b_2 than b_1. The goal is located at $(1.5, -1.5)$ with a width of $(0.2, 0.2)$. Fig. 4 shows 1000 simulated belief trajectories for POMDP approximations with different number of states.

Fig.s 4a and 4c show that for the same discrete POMDP, SARSOP quickly finds a policy which goes through the light region but if given more computational resources, it finds a policy that goes through the light region closest to the starting point, thereby yielding a larger reward. Roughly, in Fig. 4b, the set of α-vectors calculated on the sparse POMDP is not accurate enough to ensure that system uncertainty is reduced by going into the light region. For a larger POMDP in Fig. 4c with the same computational resources, a more refined set of α-vectors results in much better reduction of uncertainty and eventually a larger reward.

VII. Conclusions

We modeled a continuous-time, continuous-state stochastic system by a pair of stochastic differential equations and incrementally constructed a sequence of discrete POMDP approximations of this system via random sampling. Belief trajectories of discrete POMDPs thus created can be shown to converge in distribution to belief trajectories of the original continuous system. We have also shown that the optimal cost function and control policies for these POMDP approximations converge almost surely to their counterparts for the underlying continuous system in the limit.

Our result lays the mathematical foundation for an incremental approach to optimal control of continuous-time, continuous-state stochastic systems. In our current implementation, each POMDP approximation is solved independently using an existing discrete POMDP solver. The challenge lies in building upon these results to obtain the solution incrementally, i.e., taking a coarse POMDP model of the continuous system as input along with a coarse policy and incrementally refining both of them.

Acknowledgements

This work is supported in part by the Army Research Office MURI grant W911NF-11-1-0046.

References

[1] B.K. Øksendal. *Stochastic Differential Equations: An introduction with applications*. Springer Verlag, 2003.

[2] C.D. Charalambous and R.J. Elliott. Certain nonlinear partially observable stochastic optimal control problems with explicit control laws equivalent to LEQG/LQG problems. *IEEE Transactions on Automatic Control*, 42(4):482–497, 1997.

[3] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of operations research*, 1987.

[4] J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *Proc. of International Joint Conference on Artificial Intelligence*, pages 1025–1032, 2003.

[5] H. Kurniawati, D. Hsu, and W.S. Lee. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Proc. of Robotics: Science and Systems*, 2008.

[6] H. Bai, D. Hsu, W. Lee, and V. Ngo. Monte carlo value iteration for continuous-state pomdps. *Proceedings of the International Workshop on the Algorithmic Foundations of Robotics*, pages 175–191, 2011.

[7] J. Hoey and P. Poupart. Solving POMDPs with continuous or large discrete observation spaces. In *International Joint Conference on Artificial Intelligence*, volume 19, page 1332, 2005.

[8] S. Karaman and E. Frazzoli. Sampling-based algorithms for optimal motion planning. *International Journal of Robotics Research*, 30(7):846–894, 2011.

[9] H. J. Kushner and P. Dupuis. *Numerical methods for stochastic control problems in continuous time*. Springer Verlag, 2001.

[10] V. Huynh, S. Karaman, and E. Frazzoli. An Incremental Sampling-based Algorithm for Stochastic Optimal Control. In *Proc. of the IEEE Conference on Robotics and Automation*, 2012.

[11] P. Chaudhari, S. Karaman, and E Frazzoli. Sampling-based algorithm for filtering using Makov chain approximations. In *Proc. of the IEEE Conference on Decision and Control*, 2012.

[12] R. E. Mortensen. Stochastic Optimal Control with Noisy Observations. *International Journal of Control*, 4(5):455–464, 1966.

[13] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons Inc., 1999.

[14] E.M. Goggin. Convergence in distribution of conditional expectations. *The Annals of Probability*, 22(2):1097–1114, 1994.

[15] E.M. Goggin. Convergence of filters with applications to the Kalman-Bucy case. *IEEE Transactions on Information Theory*, 1992.

[16] A. Perez, R. Platt, G.D. Konidaris, L.P. Kaelbling, and T. Lozano-Perez. LQR-RRT*: Optimal Sampling-Based Motion Planning with Automatically Derived Extension Heuristics. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2012.

[17] Robert Platt Jr. and Russ Tedrake. Non-Gaussian belief space planning as a convex program. In *Proc. of the IEEE Conference on Robotics and Automation*, 2012.