Belief space planning assuming maximum likelihood observations

Robert Platt Jr., Russ Tedrake, Leslie Kaelbling, Tomas Lozano-Perez
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{rplatt,russt,lpk,tlp}@csail.mit.edu

***Abstract*— We cast the partially observable control problem as a fully observable underactuated stochastic control problem in belief space and apply standard planning and control techniques. One of the difficulties of belief space planning is modeling the stochastic dynamics resulting from unknown future observations. The core of our proposal is to define deterministic belief-system dynamics based on an assumption that the maximum likelihood observation (calculated just prior to the observation) is always obtained. The stochastic effects of future observations are modelled as Gaussian noise. Given this model of the dynamics, two planning and control methods are applied. In the first, linear quadratic regulation (LQR) is applied to generate policies in the belief space. This approach is shown to be optimal for linear-Gaussian systems. In the second, a planner is used to find locally optimal plans in the belief space. We propose a replanning approach that is shown to converge to the belief space goal in a finite number of replanning steps. These approaches are characterized in the context of a simple nonlinear manipulation problem where a planar robot simultaneously locates and grasps an object.**

I. Introduction

Control problems in partially observable environments are important to robotics because all robots ultimately perceive the world through limited and imperfect sensors. In the context of robotic manipulation, tactile and range sensors mounted on the manipulator near the contact locations can provide a tremendous advantage in precision over remote sensing [1, 2]. However, co-locating the sensors with the contacts this way complicates planning and control because it forces the system to trade off sensing and acting. It essentially requires the system to solve a difficult instance of the partially observable control problem, often modeled as a partially observable Markov decision process (POMDP). Unfortunately this problem has been shown to be PSPACE complete, even for a finite planning horizon, discrete states, actions, and observations [3].

One solution to the partially observable control problem is to form plans in the "belief space" of the manipulator - the space of all possible distributions over the state space. The controller then selects actions based not only on the current most-likely state of the robot, but more generically on the information available to the robot. A hallmark of belief-space planning is the ability to generate information-gathering actions. However, planning in the belief space is challenging for a number of reasons. Even coarse finite-dimensional approximations of the belief state distributions require planning in dimensions that are much larger than the original state space. Furthermore, the resulting belief state dynamics are nonlinear, underactuated (number of control inputs is smaller than the dimension of the belief space), and stochastic (transitions depend on observations which have not yet been made).

A number of powerful tools exist for planning and control of high-dimensional non-linear underactuated systems with Gaussian process noise. In order to apply these tools, this paper defines nominal belief space dynamics based on an assumption that all future observations will obtain their maximum likelihood values (this assumption is also made in [4]). During execution, the system tracks the true belief based on the observations actually obtained. Departures from the nominal belief dynamics caused by these unexpected observations are treated as Gaussian process noise. As a result, it is possible to apply standard control and planning techniques. In particular, we use linear quadratic regulation (LQR) to calculate belief space policies based on a local linearization of the belief space dynamics. However, in spite of this linearization, the resulting belief space policy is shown to be optimal for underlying linear-Gaussian systems and produces reasonable policies within a local region about the linearization point. When large, more "jarring" observations cause system belief to leave the locally stabilized region, we propose replanning from the new belief state. We analyze this replanning approach and demonstrate that, under certain conditions, it is guaranteed to ultimately reach a goal region in belief space in a finite number of replanning steps. The method is explored in simulation.

A. Related Work

For discrete systems, the optimal value function is convex over the space of all possible multinomial distributions over the state space (this "belief" space is a simplex). For a finite planning horizon, the value function is piecewise linear with a maximum number of facets doubly exponential in the number of observations and the planning horizon [5]. The 'most-likely state' approximation to POMDPs assumes that the world state that is most likely in the current belief state is in fact true, then takes the optimal action for the state in the MDP that underlies the POMDP. This approximation takes uncertainty in the future evolution of the system seriously, but ignores uncertainty about the current state. More sophisticated versions of this approximation include Q-MDP [6] and FIB [7]. A fundamental failing of these approaches is that they assume

that there is no uncertainty either in the current state (or, in the next states), and so the system is never driven to take actions for the explicit purpose of reducing uncertainty. An approach related to the current work is the belief roadmap [8]. This method ranks paths through a probabilistic roadmap defined in the underlying state space in terms of the change in covariance over the path. Perhaps the main distinction between this and our current work is that our formulation allows a variety of planning and control techniques to be used. As a result, certain types of trajectories can be found more efficiently and it becomes possible to find policies over the belief space in addition to trajectories. An approximation method closely related to our current proposal is the nominal belief-state observation (NBO) approach that plans based on an expectation of nominal (maximum likelihood) observations [4]. The current work makes plans based on a similar assumption. However, we consider the problem in a more general context and analyze approaches to control and replanning in the belief space.

Our approach is also related to the 'determinize-and-replan' approximation to MDPs that assumes world dynamics are deterministic for the purposes of planning. It takes the first action, observes the actual resulting state, and then replans. This approach, as embodied in FF-Replan, has been very successful (won the ICAPS06 planning competition) in a variety of contexts [9]. However, it cannot effectively 'hedge its bets' by taking into consideration outcomes that have low likelihood but very high or very low utility. In domains that have these outcomes, the replanning approach can be significantly suboptimal, but it remains efficient and highly effective when there are no terrible outcomes. Our approach can be viewed as 'determinize-and-replan', applied to POMDPs. It has the significant advantage over the most-likely state approach that it can and does explicitly plan to gain information. And, by replanning when surprising observations are obtained, it remains robust to unlikely outcomes.

II. Problem specification

We reformulate the underlying partially observable problem as a fully observable belief space problem with an associated cost function.

A. Underlying system

Consider the following partially observable control problem. Let $x_t \in X$ be the unobserved d-dimensional combined state of the robot and the environment at time t. Although the state is not observed directly, noisy observations, $z_t \in Z$, are available as a non-linear stochastic function of x_t:

$$z_t = g(x_t) + \omega, \tag{1}$$

where g is the deterministic component of the measurement function and ω is zero-mean Gaussian noise with possibly state-dependent covariance, W_t. This paper only considers the case where the underlying process dynamics are deterministic:

$$x_{t+1} = f(x_t, u_t). \tag{2}$$

Both g and f are required to be a differentiable functions of x_t and u_t. Although we expect our technique to extend to systems with noisy process dynamics, consideration is limited to deterministic systems to simplify the analysis.

B. Belief system

Although the underlying state of the system is not directly observed, we assume that the controller tracks a fixed parameterization of a probability distribution over state, $P(x)$. The parameters of the probability density function (pdf) of this distribution will be referred to as the "belief state" and can be tracked using Bayesian filtering as a function of control actions and observations:

$$P(x_{t+1}) = \eta P(z_{t+1}|x_{t+1}) \int_x P(x_{t+1}|x, u_t) P(x),$$

where η is a normalization constant. Notice that since the belief update is a function of the measured observations accrued during execution, it is impossible in general to predict ahead of time exactly how system belief will change until the observations are made.

For the rest of this paper, we will focus on Gaussian belief state dynamics where the extended Kalman filter belief update is used. On each update, the extended Kalman filter linearizes the process and observation dynamics (Equations 2 and 1) about the current mean of the belief distribution:

$$x_{t+1} \approx A_t(x_t - m_t) + f(m_t, u_t), \tag{3}$$

and

$$z_t \approx C_t(x_t - m_t) + g(f(m_t, u_t)) + \omega, \tag{4}$$

where m_t is the belief state mean, and $A_t = \frac{\partial f}{\partial x}(m_t, u_t)$, and $C_t = \frac{\partial g}{\partial x}(m_t)$, are Jacobian matrices linearized about the mean and action.

For the Gaussian belief system, the distribution is a Gaussian with mean, m_t, and covariance, Σ_t: $P(x) = \mathcal{N}(x|m_t, \Sigma_t)$. This belief state will be denoted by a vector, $b_t = (m_t^T, s_t^T)^T$, where $s = (s_1^T, \ldots, s_d^T)^T$ is a vector composed of the d columns of $\Sigma = (s_1, \ldots, s_d)$. If the system takes action, u_t, and makes observation, z_t, then the EKF belief state update is:

$$\Sigma_{t+1} = \Gamma_t - \Gamma_t C_t^T (C_t \Gamma_t C_t^T + W_t)^{-1} C_t \Gamma_t, \tag{5}$$

$$m_{t+1} = f_t + \Gamma_t C_t^T (C_t \Gamma_t C_t^T + W_t)^{-1} (z_{t+1} - g(f_t)), \tag{6}$$

where

$$\Gamma_t = A_t \Sigma_t A_t^T$$

and f_t denotes $f(m_t, u_t)$.

C. Cost function

In general, we are concerned with the problem of reaching a given region of state space with high certainty. For a Gaussian belief space, this corresponds to a cost function that is minimized at zero covariance. However, it may be more important to reduce covariance in some directions over others. Let $\{\hat{n}_1, \ldots, \hat{n}_k\}$ be a set of unit vectors pointing in k directions in which it is desired to minimize covariance and

let the relative importance of these directions be described by the weights, $w_1, \ldots, w_k$. We define a finite horizon quadratic cost function:

$$J(b_{\tau:T}, u_{\tau:T}) = \sum_{i=1}^{k} w_i \left(\hat{n}_i^T \Sigma_T \hat{n}_i\right)^2 + \sum_{t=\tau}^{T-1} \tilde{m}_t^T Q \tilde{m}_t + \tilde{u}_t^T R \tilde{u}_t, \tag{7}$$

where $\tilde{m}_t = m_t - \bar{m}_t$ and $\tilde{u}_t = u_t - \bar{u}_t$ are the mean and action relative to a desired state-action point or trajectory, $\bar{m}_t$ and $\bar{u}_t$, and Σ_T is the covariance matrix at the end of the planning horizon. The first summation penalizes departures from zero covariance along the specified directions. The second summation over the planning horizon penalizes departures from a desired mean and action or mean and action trajectory with positive definite cost matrices Q and R. While Equation 7 is quadratic, the first summation is not expressed in the standard form. However, its terms can be re-written in terms of s, the vector of the columns of Σ:

$$(\hat{n}_i^T \Sigma \hat{n}_i)^2 = s^T L_i s,$$

where the cost matrix, L_i, is a function of $\hat{n}_i$:

$$L_i = \begin{pmatrix} \hat{n}_i n_{i,1} \\ \vdots \\ \hat{n}_i n_{i,d} \end{pmatrix} \begin{pmatrix} \hat{n}_i n_{i,1} \\ \vdots \\ \hat{n}_i n_{i,d} \end{pmatrix}^T,$$

where $n_{i,j}$ is the j^{th} element of $\hat{n}_i$. As a result, we re-write the cost function as:

$$J(b_{\tau:T}, u_{\tau:T}) = s^T \Lambda s + \sum_{t=\tau}^{T-1} \tilde{m}_t^T Q \tilde{m}_t + \tilde{u}_t^T R \tilde{u}_t, \tag{8}$$

where

$$\Lambda = \sum_{i=1}^{K} w_i L_i.$$

Although not included in Equation 8, all plans in the belief space are required to satisfy certain final value constraints. First, a constraint on the final value mean of the belief system is specified: $m_T = \bar{m}_T$. B-LQR (Section IV) incorporates this constraint by augmenting the cost function (Equation 15). Direct transcription (Section V-A) incorporates this constraint directly. Second, direct transcription incorporates a constraint on the magnitude of the final covariance cost (first term of Equation 8): $s^T \Lambda s < \gamma$.

For a policy defined over the belief space,

$$u_t = \pi(b_t),$$

the expected cost-to-go from a belief state, b, at time t with a planning horizon T (the planning horizon could be infinite) is:

$$J^\pi(b_\tau) = \mathcal{E}_{z_{\tau:T-1}} \left\{ J(b_{\tau:T}, \pi(b_{\tau:T-1})) \right\}, \tag{9}$$

where the expectation is taken over future observations. The optimal policy minimizes expected cost: $\pi^*(b_t) = \arg\min_\pi J^\pi(b_t)$.

III. Simplified belief space dynamics

In order to apply standard control and planning techniques to the belief space problem, it is necessary to define the dynamics of the belief system. Given our choice to use the EKF belief update, Equations 5 and 6 should be used. Notice that Equation 6 depends on the observation, z_t. Since these observations are unknown ahead of time, it should be necessary to evaluate the expected value of seeing each observation and take the expectation over all observations. However, since it is difficult to evaluate this marginalization, we make a key simplifying assumption for the purposes of planning and control: we assume that future observations are normally distributed about a maximum likelihood. If action u_t is taken from belief state b_t, then the maximum likelihood observation is:

$$z_{ml} = \arg\max_z P(z|b_t, u_t). \tag{10}$$

Evaluating the maximum likelihood observation for the EKF using Equation 10, we have:

$$\begin{aligned} z_{ml} &= \arg\max_z \int P(z_{t+1}|x_{t+1}) P(x_{t+1}|b_t, u_t) dx_{t+1} \\ &= \arg\max_z \int \mathcal{N}(z_{t+1}|C_t(x_{t+1} - m_t) + g(f_t), W_t) \\ &\qquad \mathcal{N}(x_{t+1}|f_t, \Gamma_t) dx_{t+1} \\ &= \arg\max_z \mathcal{N}(z_{t+1}|C_t(f_t - m_t) + g(f_t), \\ &\qquad C_t \Gamma_t C_t^T + W_t) \\ &= C_t(f_t - m_t) + g(f_t) \\ &\approx g(f_t). \end{aligned} \tag{11}$$

Substituting into Equation 6, and restating Equation 5, the simplified dynamics are:

$$b_{t+1} = F(b_t, u_t), \tag{12}$$

where F evaluates to the b_{t+1} corresponding to m_{t+1} and Σ_{t+1},

$$m_{t+1} = f_t + v, \tag{13}$$

$$\Sigma_{t+1} = \Gamma_t - \Gamma_t C_t^T (C_t \Gamma_t C_t^T + W_t)^{-1} C_t \Gamma_t, \tag{14}$$

and v is Gaussian noise.

IV. LQR in belief space

Linear quadratic regulation (LQR) is an important and simple approach to optimal control [10]. While LQR is optimal only for linear-Gaussian systems, it is also used to stabilize non-linear systems in a local neighborhood about a nominal point or trajectory in state space. LQR in belief space (B-LQR) is the application of LQR to belief space control using the simplified belief system dynamics (Equations 13 and 14). Since the belief system dynamics are always non-linear, policies found by B-LQR can be expected to be only locally stable. However, as we show below, it turns out that B-LQR is optimal and equivalent to linear quadratic Gaussian (LQG) control for systems with linear-Gaussian process and observation

dynamics. Moreover, B-LQR produces qualitatively interesting policies for systems with non-linear dynamics.

We are interested in the finite horizon problem with a final state constraint on the mean of the belief system. The final state constraint is accommodated by adding an additional term to Equation 7 that assigns a large cost to departures from the desired final mean value:

$$\begin{aligned} J(b_{\tau:T}, u_{\tau:T}) &= \tilde{m}_T^T Q_{large} \tilde{m}_T + s^T \Lambda s \\ &\quad + \sum_{t=\tau}^{T} \tilde{m}_t^T Q \tilde{m}_t + \tilde{u}_t^T R \tilde{u}_t. \end{aligned} \quad (15)$$

Since B-LQR is operating in belief space, is necessary to linearize Equation 12 about a nominal belief and action, denoted $\bar{b}_t$ and $\bar{u}_t$:

$$\mathbb{A}_t = \frac{\partial F}{\partial b}(\bar{b}_t, \bar{u}_t)$$

and

$$\mathbb{B}_t = \frac{\partial F}{\partial u}(\bar{b}_t, \bar{u}_t).$$

Notice that the mean of the Gaussian belief system has the same nominal dynamics as the underlying system. As a result, $\mathbb{A}_t$ always has the form:

$$\mathbb{A}_t = \begin{pmatrix} A_t & \mathbf{0} \\ \frac{\partial s_{t+1}}{\partial m_t} & \frac{\partial s_{t+1}}{\partial s_t} \end{pmatrix},$$

where $A_t = \frac{\partial f}{\partial x}(\bar{m}_t, \bar{u}_t)$ (see Equation 3). Also, note that since the control input never directly affects covariance, $\mathbb{B}$ always has the form

$$\mathbb{B}_t = \begin{pmatrix} B_t \\ 0 \end{pmatrix},$$

where $B_t = \frac{\partial f}{\partial u}(\bar{m}_t, \bar{u}_t)$. Finally, since Equation 15 only assigns departures from the mean a recurring cost, the recurring cost matrix for belief state is

$$\mathbb{Q} = \begin{pmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Using the above, the Riccati equation for the discrete time finite horizon problem is:

$$\begin{aligned} \mathbb{S}_t &= \mathbb{Q} + \mathbb{A}_t^T \mathbb{S}_{t+1} \mathbb{A}_t \\ &\quad - \mathbb{A}_t^T \mathbb{S}_{t+1} \mathbb{B}_t (\mathbb{B}_t^T \mathbb{S}_{t+1} \mathbb{B}_t + R_t)^{-1} \mathbb{B}_t^T \mathbb{S}_{t+1} \mathbb{A}_t, \end{aligned} \quad (16)$$

where $\mathbb{S}_t$ is the expected cost-to-go matrix for linear-Gaussian systems, $J^{\pi^*}(b_t) = b_t^T \mathbb{S}_t b_t$. The optimal action for linear-Gaussian systems is:

$$u^* = -(\mathbb{B}_t^T \mathbb{S}_{t+1} \mathbb{B}_t + R_t)^{-1} \mathbb{B}_t^T \mathbb{S}_{t+1} \mathbb{A}_t b_t. \quad (17)$$

An important difference between the above and the usual LQR formulation is that cost is measured with respect to different points than those about which the system is linearized. In particular, the first term in Equation 8 measures cost with respect to zero covariance whereas $\mathbb{A}_t$ is the linearization about the current belief.

The following theorem demonstrates that B-LQR is equivalent to LQG control for linear-Gaussian systems.

Theorem 1: If the cost function is of the form of Equation 8, and the underlying process and observation dynamics (Equations 1 and 2) are linear, then B-LQR is optimal.

Proof: We show that under the conditions of the theorem, B-LQR is equivalent to LQG and is optimal as a result.

First, if the underlying observation dynamics are linear, then $\frac{\partial s_{t+1}}{\partial m_t} = 0$ and $\mathbb{A}_t$ is block diagonal. Also, note that if the cost is of the form in Equation 8, then the state cost, $\mathbb{Q}$, is also block diagonal with respect to mean and covariance (m and s). As a result of these two facts, the solution to the belief space Riccati equation (Equation 16) at time t, $\mathbb{S}_t$, always has the form,

$$\mathbb{S}_t = \begin{pmatrix} S_t & \mathbf{0} \\ \mathbf{0} & P_t \end{pmatrix},$$

where S_t is the solution of the Riccati equation in the underlying space,

$$\begin{aligned} S_t &= Q + A_t^T S_{t+1} A_t \\ &\quad - A_t^T S_{t+1} B_t (B_t^T S_{t+1} B_t + R_t)^{-1} B_t^T S_{t+1} A_t, \end{aligned}$$

and P_t is arbitrary. Substituting into Equation 17, we have:

$$u^* = -(B_t^T S_{t+1} B_t + R_t)^{-1} B_t^T S_{t+1} A_t m_t.$$

Since this is exactly the solution of the LQG formulation, and LQG is optimal for linear-Gaussian process and observation dynamics, we conclude the B-LQR is optimal. ■

V. Planning

Since B-LQR is based on a local linearization of the simplified belief dynamics, it is clear that planning methods that work directly with the non-linear dynamics can have better performance. A number of planning methods that are applied to underactuated systems are relevant to the belief space planning problem: rapidly expanding random trees (RRTs) [11], LQR-trees [12], and nonlinear optimization methods [13]. Presently, we focus on an approach based on nonlinear optimization known as direct transcription. After finding a nominal trajectory through belief space, the trajectory is stabilized using time varying B-LQR about the trajectory. After introducing the method, we propose a replanning strategy and demonstrate that it is guaranteed to reach an arbitrary belief objective in finite time.

A. Direct transcription

Direction transcription is an approach to transcribing the optimal control problem to a nonlinear optimization problem. Suppose we want to find a path from time 1 to T starting at b_1 that optimizes Equation 8. Direction transcription parameterizes the space of possible trajectories by a series of k segments. Let δ be a user-defined integer that defines the length of each segment in time steps. Then the number of segments is $k = \frac{T-t}{\delta}$. Let $b'_{1:k}$ and $u'_{1:k-1}$ be sets of belief state and action variables that parameterize the trajectory in terms of the segments. Segment i begins at time $i\delta$ and ends at

time $i\delta+\delta-1$. The cost function, Equation 8, is approximated in terms of these segments:

$$\begin{aligned} J(b_{1:T}, u_{1:T}) &\approx \hat{J}(b'_{1:k}, u'_{1:k}) \\ &= s^T \Lambda s + \sum_{j=1}^{k} \tilde{m}_i'^T Q \tilde{m}'_i + \tilde{u}_i'^T R \tilde{u}'_i, \end{aligned} \quad (18)$$

The belief state on the last time step of this segment can be found by integrating F over δ:

$$\phi(b'_i) = F(b'_i, u_i) + \sum_{t=i\delta}^{i\delta+\delta-1} F(b_{t+1}, u_i) - F(b_t, u_i). \quad (19)$$

It is now possible to define the nonlinear optimization problem that approximates the optimal control problem. We want assignments to the variables, $b'_{1:k}$ and $u'_{1:k}$, that minimize the approximate cost, $\hat{J}(b'_{1:k}, u'_{1:k})$ subject to constraints that require each segment to be dynamically consistent with its neighboring segments:

$$\begin{aligned} b'_2 &= \phi(b'_1, u'_1) \\ &\vdots \\ b'_k &= \phi(b'_{k-1}, u'_{k-1}). \end{aligned} \quad (20)$$

Since we have a final value constraint on the mean component of belief, an additional final value constraint is added:

$$b_k = b_T. \quad (21)$$

By approximating the optimal control problem using Equations 19 through 21, it is possible to apply any optimization method available. In the current work, we use sequential quadratic programming (SQP) to minimize the Lagrangian constructed from the costs and constraints in Equations 19- 21. At each iteration, this method takes a Newton step found by solving the Karush-Kuhn-Tucker (KKT) system of equations. Ultimately SQP converges to a local minimum. For more information, the reader is recommended to [13].

B. Replanning strategy

Algorithm 1

0. initialize b
1. π = plan found using direct transcription
2. while $m_t - \bar{m}_t < \theta$
3. $\quad u$ = time-varying-LQR(π, b)
4. $\quad b = EKF(b, u, z)$

Since the actual belief transition dynamics are stochastic, a mechanism must exist for handling divergences from the planned trajectory. One approach is to use time varying B-LQR about the planned path to stabilize the trajectory. In this case, the linearization of the belief system dynamics, $\mathbb{A}$ and $\mathbb{B}$, is made about the planned trajectory, $\bar{b}_{1:T}$ and $\bar{u}_{1:T-1}$. However, while this works for small departures from the planned trajectory, replanning is needed to handle larger divergences. We propose the basic replanning strategy outlined in Algorithm 1 (above). On the first time step, the algorithm solves for a belief space trajectory using direct transcription that satisfies the final value constraints on the mean. It is assumed that such a trajectory always exists and is found by direct transcription within the planning horizon, T. Second, the algorithm solves for a locally stable policy about the trajectory. Next, the system executes the policy until the mean component of belief departs from the planned trajectory by more than a given threshold, θ:

$$m_t - \bar{m}_t = \tilde{m}_t > \theta.$$

It is assumed that the threshold, θ, is chosen such that B-LQR is stable within the threshold. When the threshold is violated, the system replans using direct transcription and the process starts over. While we expect that it will be possible to characterize this threshold using sum-of-squares Lyapunov function verification techniques from the convex optimization literature, we leave a closer examination of this approach for future work.

C. Analysis of belief space covariance during replanning

Under Algorithm 1, the mean component of belief can be shown to reach the final value constraint in a finite number of replanning steps. Essentially, we show that each time the belief system deviates from the planned trajectory by more than θ, covariance decreases by a finite amount. Each time this occurs, it becomes more difficult for the belief state mean to exceed the threshold again. After a finite number of replanning steps, this becomes impossible and the system converges to the final value constraint. In the following, we shall denote the spectral norm of a matrix, A, by $\|A\|_2$. The spectral norm of a matrix evaluates to its largest singular value.

Lemma 1: Suppose that 1) the observation process noise has non-zero covariance in all directions, $\forall \hat{v} \in \{\mathbb{R}^d : \|\hat{v}\| = 1\}$, $W_t \hat{v} > W_{min} > 0$, 2) the underlying process dynamics are deterministic, 3) A_t has eigenvalues no greater than one, 4) the observations are bounded, $z_t < z_{max}$, 5) the planner always finds a belief trajectory that meets the final value constraints within the planning horizon, T. Then, the change in covariance, $\|\Sigma_c - \Sigma_{c+1}\|_2$, between time t_a when $\tilde{m}_a = 0$ and time t_b when $\tilde{m}_b > \theta$ (the replanning threshold) is no less than a constant function of W_{min}, z_{max}, θ, $\tau = b - a$, and Λ.

Proof: Since m changes by at least θ over τ timesteps, there is at least one timestep between t_a and t_b (time t_c) where $\Delta m_c = \|m_{c+1} - m_c\| > \frac{\theta}{\tau}$. This implies:

$$\begin{aligned} \frac{\theta}{\tau} &\leq |\Gamma_c C_c^T (C_c \Gamma_c C_c^T + W_c)^{-1} (z_{c+1} - \bar{z}_{c+1})| \\ &\leq \|\Gamma_c C_c^T (C_c \Gamma_c C_c^T + W_c)^{-1}\|_2 z_{max}. \end{aligned}$$

Since

$$\|(C_c \Gamma_c C_c^T + W_c)^{-\frac{1}{2}}\|_2 \leq \|W_{min}^{-\frac{1}{2}}\|_2,$$

and using the properties of the spectral norm, we have that:

$$\frac{\theta}{\tau} \leq \|\Gamma_c C_c^T (C_c \Gamma_c C_c^T + W_c)^{-\frac{1}{2}}\|_2 \|W_{min}^{-\frac{1}{2}}\|_2 z_{max}.$$

Dividing through by $\|W_{min}^{-\frac{1}{2}}\|_2 z_{max}$ and squaring the result, we have:

$$\left(\frac{\theta}{\tau\|W_{min}^{-\frac{1}{2}}\|_2 z_{max}}\right)^2 \leq \|\Gamma_c C_c^T (C_c \Gamma_c C_c^T + W_c)^{-1} C_c \Gamma_c\|_2.$$

Considering the covariance update equation (Equation 14), and considering that $\|A_c\|_2 \leq 1$, it must be that

$$\|\Sigma_c - \Sigma_{c+1}\|_2 \geq \left(\frac{\theta}{\tau\|W_{min}^{-\frac{1}{2}}\|_2 z_{max}}\right)^2$$

and the Lemma is true for any value of Λ. ■

Since Lemma 1 establishes that Σ decreases by a constant amount each time algorithm 1 replans, the following Theorem is able to conclude that ultimately Σ becomes so small that it is impossible for any observation to cause the mean of the belief state to change by more than θ over the planning horizon.

Theorem 2: Under the conditions of Lemma 1 and assuming that time varying B-LQR stabilizes the trajectory within the replanning threshold, θ, Algorithm 1 reaches a belief state that satisfies the final value conditions after a finite number of replanning steps.

Proof: Under Algorithm 1, the system does one of two things: 1) it follows a planned trajectory until time varying B-LQR stabilizes the system within the final value constraints, or 2) it exceeds the replanning threshold. In the second case, we have by Lemma 1 that Σ decreases by a constant and non-zero amount. For a small enough Σ, it is impossible for $\tilde{m}$ to exceed θ for a bounded z_{max} over the finite planning horizon, T, and the system must follow the planned trajectory until the final value constraints are met. ■

VI. Experiments

We explored the capabilities of our approach to belief space planning in two experimental domains: the light-dark domain and the planar grasping domain.

A. The light-dark domain

In the light-dark domain, a robot must localize its position in the plane before approaching the goal. The robot's ability to localize itself depends upon the amount of light present at its actual position. Light varies as a quadratic function of the horizontal coordinate. Depending upon the goal position, the initial robot position, and the configuration of the light, the robot may need to move away from its ultimate goal in order to localize itself. Figure 1 illustrates the configuration of the light-dark domain used in our experiments. The goal position is at the origin, marked by an X in the figure. The intensity in the figure illustrates the magnitude of the light over the plane. The robot's initial position is unknown.

The underlying state space is the plane, $x \in \mathbb{R}^2$. The robot is modeled as a first-order system such that robot velocity is determined by the control actions, $u \in \mathbb{R}^2$. The underlying system dynamics are linear with zero process noise, $f(x_t, u_t) = x_t + u$. The observation function is identity,

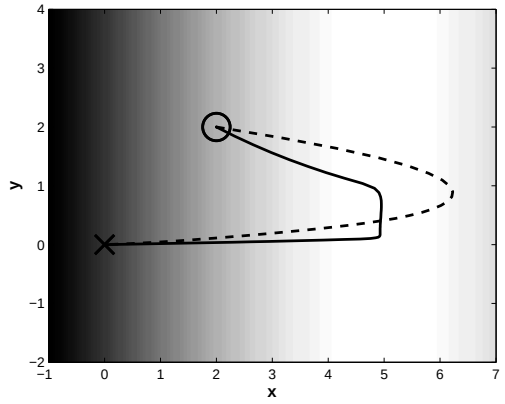

Fig. 1. Comparison of the mean of the planned belief state trajectory found by B-LQR (the dashed line) with the locally optimal trajectory found by direct transcription (the solid line).

$g(x_t) = x_t + \omega$, with zero-mean Gaussian observation noise a function of state, $\omega \sim \mathcal{N}(\cdot|0, w(x))$, where

$$w(x) = \frac{1}{2}(5 - x_x)^2$$

has a minimum when $x_x = 5$, where x_x is the first element of x. Belief state was modeled as an isotropic Gaussian pdf over the state space: $b = (m, s) \in \mathbb{R}^2 \times \mathbb{R}^+$. The cost function (Equation 8) used recurring state and action costs of $R = diag(0.5, 0.5)$ and $Q = diag(0.5, 0.5)$, and a final cost on covariance, $\Lambda = 200$. B-LQR had an additional large final cost on mean. Direct transcription used a final value constraint on mean, $m = (0, 0)$, instead. The true initial state was $x_1 = (2.5, 0)$ and the prior belief (initial belief state) was $b_1 = (2, 2, 5)$. The replanning threshold was $\theta = 0.1$. At each replanning step, direct transcription was initialized with a random trajectory. B-LQR linearized the belief system dynamics about $(0, 0, 0.5)$.

1) Results and discussion: Figure 1 shows solutions to the light-dark problem domain found by B-LQR (the dotted line) and by direct transcription (the solid line). The B-LQR trajectory shows the integrated policy assuming that the assumed observation dynamics were always obtained. The direct transcription trajectory shows the initial plan found before replanning. Most importantly, notice that even though the B-LQR trajectory is based on a poor linearization of the belief dynamics, it performs surprisingly well (compare with the locally optimal direct transcription trajectory). However, it is clear that B-LQR is sub-optimal because whereas the locally optimal trajectory lingers in the minimum-noise region at $x_x = 5$, the B-LQR trajectory overshoots past the minimum noise point to $x_x = 6$.

Figure 2 illustrates the behavior of the replanning performed by Algorithm 1. The dotted line shows the mean of the belief space trajectory that was found on the first planning step. The solid line shows the actual trajectory. Whereas the system expected that it began execution at $x = (2, 2)$, it

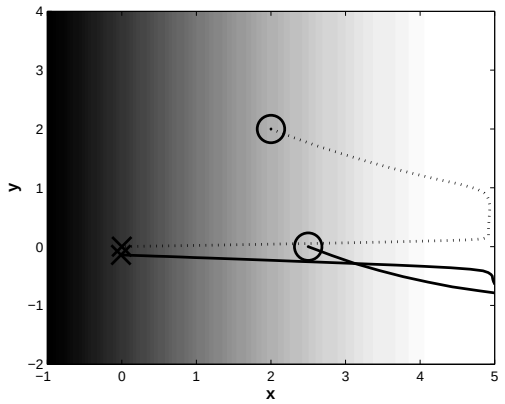

Fig. 2. Comparison of the true robot trajectory (solid line) and the mean of the belief trajectory that was initially planned by direct transcription (dotted line).

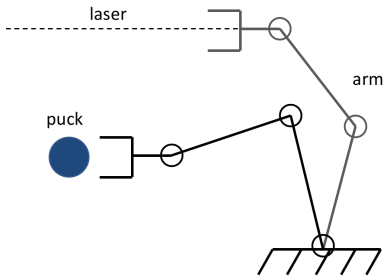

Fig. 3. Laser-grasp domain. A range-finding laser (the dashed line) points out from the robot end effector. The objective is to move the end-effector to a point just in front of the puck on the left (the manipulator approach configuration).

actually began execution at $x = (2.5, 0)$. As a result of this confusion, it initially took actions consistent with its initial belief. However, as it moved into the light, it quickly corrected its misperception. After reaching the point of maximum light intensity, the system subsequently followed a nearly straight line toward the goal.

B. Laser-grasp domain

In the laser-grasp domain, a planar robot manipulator must locate and approach a round puck as illustrated in Figure 3. The robot manipulator position is known, but the puck position is unknown. The robot locates the puck using a range-bound laser range finder that points out from the end-effector along a line. The end-effector always points horizontally as indicated. In order to solve the task, the manipulator must move back and forth in front of the puck so that the laser (the laser is always switched on) detects the puck location and then move to the grasp approach point. The robot is controlled by specifying Cartesian end-effector velocities.

1) Setup: The underlying state space, $x \in \mathbb{R}^2$, denotes the position of the manipulator relative to an "approach point" defined directly in front of the object. Although the end-effector position is assumed to be known completely, state is not observed directly because the object position is unknown. The control action, $u \in \mathbb{R}^2$, specifies the end-effector velocity:

$$f(x_t, u_t) = x_t + u_t.$$

In order to get a smooth measurement function that is not discontinuous at the puck edges, the puck was modeled using a symmetric squashing function about the origin. The result was roughly circular with a "radius" of approximately 0.65. The measurement gradient, C, was zero outside of that radius. State dependent noise was defined to be large when the laser scan line was outside the radius of the puck (modeling an unknown and noisy background). The noise function also incorporated a low-amplitude quadratic about $x_x = 5$ modeling a sensor with maximum measurement accuracy at 5 units away from the target. The belief space was modeled as a non-isotropic Gaussian in the plane: $b = (m, s) \in \mathbb{R}^2 \times \mathbb{R}^3$. The parameterization of covariance in $\mathbb{R}^3$ rather than $\mathbb{R}^4$ is a result of incorporating the symmetry of the covariance matrix into the representation. The cost function was of the form of Equation 8 with $Q = diag(10, 10, 10, 10, 10)$, $R = diag(10, 10)$, and $\Lambda = diag(10000, 0, 10000)$ (denoting no preference for covariance in one direction over another).

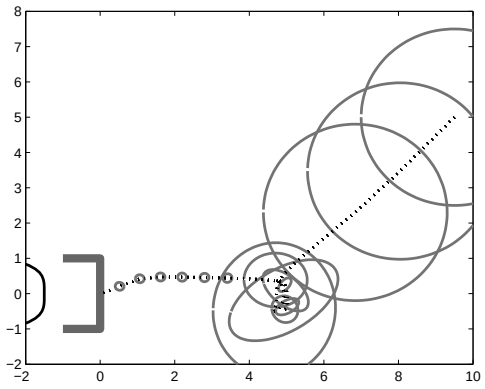

Fig. 4. Trajectory found using direct transcription for the laser-grasp domain. The dotted line denotes the mean of the belief state trajectory. The ellipses sample the covariance matrix at various points along the trajectory. The half circle on the left represents the puck. Just to the right of the puck, the end-effector is illustrated at the approach point.

2) Results and discussion: Figure 4 shows a representative plan found by direct transcription that moves the end-effector along the dotted line from its starting location in the upper right to the goal position in front of the puck (The geometry of the experiment roughly follows Figure 3). The initial belief state was $b_1 = (9, 5, 5, 0, 5)$. The ellipses in Figure 4 illustrate the planned trajectory of the Gaussian. First, notice that the system plans to move the laser in front of the puck so that it may begin to localize it. Corvariance does not change until the end-effector is actually in front of the puck. Also, notice that the plan lingers in front of the puck near the optimal sensor

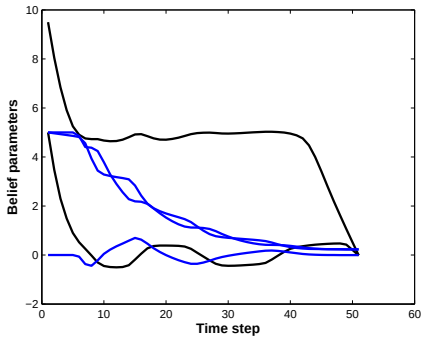

Fig. 5. Planned belief trajectory as a function of time step. The two black lines denote the mean of the belief. The three blue lines denote the elements of covariance. Notice that as the plan is "scanning" the puck, different elements of covariance change in alternation.

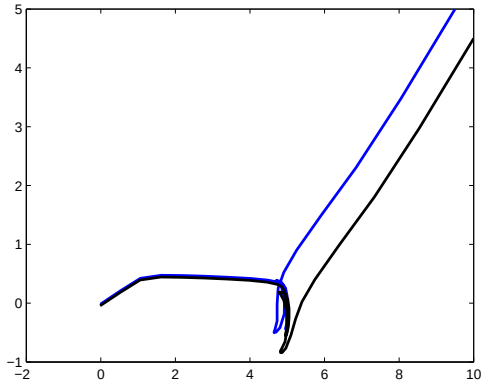

Fig. 6. Comparison between the initially planned trajectory (the blue line) and the actual trajectory (the black line).

range. During this time, the trajectory makes short jumps up and down, apparently "scanning" the puck. Finally, as time approaches the planning horizon, the end-effector moves off to the approach point. Figure 5 provides a more in-depth look at what is going on while the plan scans the puck. First, notice that the plan spends almost all its time in the "sweet spot" scanning the puck. Second, notice that different elements of covariance change during different phases of the scanning motion. This suggests that during scanning, the plan actively alternates between reducing covariance in different directions. The effect results from the fact that the dimension of the observation (a scalar scan depth) is one while the Gaussian belief is over a two-dimensional space. At a given point in time, it is only possible to minimize one dimension of covariance and the plan alternates between minimizing various different dimensions, resulting in the scanning behavior. Finally, Figure 6 illustrates the behavior the replanning strategy where time varying B-LQR stabilization was used. Initially, the mean of the system prior is at $m = (9, 5)$ but the true state is at $x = (10, 4.5)$. The incorrect belief persists until the system reaches a point in front of the puck. At this point, the incorrect prior is gradually corrected during the scanning process until the true state of the system finally reaches the origin.

VII. Conclusion

This paper explores the application of underactuated planning and control approaches to the belief space planning problem. Belief state dynamics are underactuated because the number of controlled dimensions (the parameters of a probability distribution) exceeds the number of independent control inputs. As a result, the dynamics are constrained in a way that can make planning difficult. Our contribution is to recast the belief space planning problem in such a way that conventional planning and control techniques are applicable. As a result, we are able to find belief space policies using linear quadratic regulation (LQR) and locally optimal belief space trajectories using direct transcription. We provide theoretical results characterizing the effectiveness of a plan-and-replan strategy. Finally, we show that the approach produces interesting and relevant behaviors on a simple grasp problem where it is necessary to acquire information before acting.

References

[1] C. Corcoran and R. Platt, "Tracking object pose and shape during robot manipulation based on tactile information," in *IEEE Int'l Conf. on Robotics and Automation*, vol. 2, 2010.

[2] A. Petrovskaya, O. Khatib, S. Thrun, and A. Ng, "Bayesian estimation for autonomous object manipulation based on tactile sensors," in *IEEE Int'l Conf. on Robotics and Automation*, 2006, pp. 707–714.

[3] C. Papadimitriou and J. Tsitsiklis, "The complexity of markov decision processes," *Mathematics of Operations Research*, vol. 12, no. 3, pp. 441–450, 1987.

[4] S. Miller, A. Harris, and E. Chong, "Coordinated guidance of autonomous uavs via nominal belief-state optimization," in *American Control Conference*, 2009, pp. 2811–2818.

[5] E. Sondik, "The optimal control of partially observable markov processes," Ph.D. dissertation, Stanford University, 1971.

[6] M. Littman, A. Cassandra, and L. Kaelbling, "Learning policies for partially observable environments: Scaling up," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

[7] M. Hauskrecht, "Value-function approximations for partially observable markov decision processes," *Journal of Artificial Intelligence Research*, vol. 13, pp. 33–94, 2000.

[8] S. Prentice and N. Roy, "The belief roadmap: Efficient planning in linear pomdps by factoring the covariance," in *12th International Symposium of Robotics Research*, 2008.

[9] S. Yoon and R. Fern, A. Givan, "FF-replan: A baseline for probabilistic planning," in *Proceedings of the International Conference on Automated Planning and Scheduling*, 2007.

[10] D. Bertsekas, *Dynamic Programming and Optimal Control: 3rd Edition*. Athena Scientific, 2007.

[11] S. LaValle and J. Kuffner, "Randomized kinodynamic planning," *International Journal of Robotics Research*, vol. 20, no. 5, pp. 378–400, 2001.

[12] R. Tedrake, "LQR-trees: Feedback motion planning on sparse randomized trees," in *Proceedings of Robotics: Science and Systems (RSS)*, 2009.

[13] J. Betts, *Practical methods for optimal control using nonlinear programming*. Siam, 2001.