Hybrid control trajectory optimization under uncertainty

Joni Pajarinen, Ville Kyrki, Michael Koval, Siddhartha Srinivasa, Jan Peters, and Gerhard Neumann

Abstract—Trajectory optimization is a fundamental problem in robotics. While optimization of continuous control trajectories is well developed, many applications require both discrete and continuous, i.e., hybrid, controls. Finding an optimal sequence of hybrid controls is challenging due to the exponential explosion of discrete control combinations. Our method, based on Differential Dynamic Programming (DDP), circumvents this problem by incorporating discrete actions inside DDP: we first optimize continuous mixtures of discrete actions, and, subsequently force the mixtures into fully discrete actions. Moreover, we show how our approach can be extended to partially observable Markov decision processes (POMDPs) for trajectory planning under uncertainty. We validate the approach in a car driving problem where the robot has to switch discrete gears and in a box pushing application where the robot can switch the side of the box to push. The pose and the friction parameters of the pushed box are initially unknown and only indirectly observable.

I. INTRODUCTION

Many control applications require both discrete and continuous, i.e., hybrid, controls. For example, consider a car with continuous acceleration and direction control but discrete gears [1]. Switching gears changes the dynamics of the car. Another example is pushing of a box [2]–[4] where the agent can select not only a continuous pushing direction and velocity but also a discrete side to push. Hybrid control is important also in other applications, for example, chemical engineering processes involving on-off valves [5].

Hybrid control has been an active research topic [1], [6]–[11]. In this paper, we investigate systems with non-linear dynamics and long sequences of hybrid controls and states (trajectories). We provide hybrid control trajectory planning methods for optimizing linear feedback controllers in systems with stochastic dynamics and partial observability which is a challenging but common setting in robotic applications. An example follows.

In the box pushing application that motivated this paper, we investigate the problem of a robot pushing an unknown object to a predefined location. The robot may not know in advance the physical properties of the object. An object may contain different kinds of materials which may introduce large errors to the initial guess of center of friction. Moreover, the robot can only make noisy observations about the current position and angle of the object, using, for example, a vision sensor. In order to accomplish its task, the robot needs to gather information about the object while pushing the object closer to the goal. After each pushing action, the robot can only indirectly infer the physical properties of the object from observed changes of the object pose. Moreover, dynamics of the object are uncertain due to modeling inaccuracies and changing surface properties. A natural model for this kind of problem is a partially observable Markov decision process (POMDP) which takes uncertainty in both world dynamics and observations into account. A POMDP also optimally balances long-term information gathering with greedy utility maximization. We treat the initially unknown physical variables, that describe how the object behaves when pushed, as part of the POMDP state space. Moreover, in the pushing application, the pushing velocity and direction are limited. In robotics, control limits are common; for example, joint motors have physical limits. We show how to add hard limits to DDP based POMDP trajectory optimization, which has so far only been shown for deterministic MDPs [12]. Below, we summarize the major contributions of this paper:

- **Hybrid Trajectory Optimization**: We introduce trajectory optimization with discrete and continuous actions using differential dynamic programming (DDP)
- **Extension to Partial Observable Environments**: We introduce the first POMDP trajectory optimization algorithm with hybrid controls
- **Hard Control Limits**: Inspired by hard-control limits for trajectory optimization [12], we introduce hard control limits for POMDP trajectory optimization.
- **Box-Pushing Application**: We use for the first time a POMDP formulation for pushing an unknown object in
a task specific way. Information gathering is required for performing the task.

II. RELATED WORK

In this paper, we consider hybrid control [1, 6–11, 13] for finite discrete time trajectory optimization [14] in systems with non-linear stochastic dynamics and noisy, partial, state information. We optimize time varying linear feedback controllers producing a trajectory consisting of states, covariances, and hybrid controls.

There are many methods for optimizing continuous control trajectories [14–17]. Discrete controls present an additional challenge since the number of combinations of discrete controls is exponential w.r.t. the planning horizon. Globally optimal hybrid control approaches have been proposed. The naïve approach of optimizing continuous controls for each combination of discrete controls and choosing the combination with lowest cost scales only to few time steps. [18] uses a tree of linear quadratic regulator (LQR) solutions for discrete action combinations and introduces a pruning technique for pruning the tree. In general, this approach scales only to a few time steps because of the exponential growth of the tree over time.

Approaches which scale to more complicated problems include approaches which try to find a local optimum by performing continuous control optimization and local discrete control improvements iteratively [11]. Mixed integer non-linear programming (MINLP) is a promising technique since hybrid control problems can be translated into MINLPs. However, in complex problems, MINLP solutions adapted to hybrid control problems can be restricted to only a few time step control horizons [9]. [1] shows how to apply “convexification” introduced in [8] for discrete MINLP variables. “Convexification” transforms discrete controls into weights replacing the original dynamics function with a convex mixture of dynamics functions. However, [1] considers only fully observable deterministic dynamics. Instead, we show how to apply a similar idea to differential dynamic programming (DDP) in order to get an algorithm for trajectory planning with stochastic dynamics and partial observations.

Instead of using hybrid control for optimizing trajectories, reinforcement learning approaches based on the options framework can compute high level discrete actions, also called options [19] or macro-actions, and execute a continuous control policy for each high level action.

RRTs. Rapidly-exploring random trees (RRTs) [20] are often used for trajectory initialization. [21] uses RRTs in hybrid control trajectory planning of fully observable systems. [22] uses also hybrid, that is, discrete and continuous actions, for pushing objects from an initial three dimensional configuration into another final configuration. Contrary to our work, [22] does not do trajectory optimization and does not take uncertainty into account.

POMDPs. Our proposed trajectory optimization approach plans under model, sensing, and actuation uncertainty. [23] and [24] investigate partial observations for computing switching times in switched systems by providing simple analytic examples of a few time steps. However, [23], [24] do not model state uncertainty. Previously, planning under uncertainty has been investigated in simulated robotic tasks in [25–28]. [28] presents a sampling based POMDP approach for continuous states and actions. [28] uses the POMDP approach for Bayesian reinforcement learning in a simulated pendulum swingup experiment.

[29], [30] use feedback based motion planning with uncertainty and partial observations. [29], [30] target navigation type of applications and use probabilistic roadmap planning to generate a graph where graph nodes represent locations.

One classic way of dealing with observation uncertainty is to assume maximum likelihood observations [15]. [17] uses shooting methods for POMDP trajectory optimization. The POMDP approach of [16] is based on iterated linear quadratic Gaussian (iLQG) control with covariance linearization. In this paper, we extend the fully observable DDP/iLQG [12] algorithm as well as the iLQG/DDP based POMDP approach of [16] to hybrid controls. We also show a straightforward way of using hard control limits with iLQG/DDP based POMDP.

Box pushing. When pushing an unknown object a robot needs to plan its motions and pushing actions while taking model [31], sensing, and actuation uncertainty into account. Therefore, we model the pushing task as a continuous POMDP and present a new POMDP approach for locally optimizing a policy. We base our box pushing simulation on the same quasi-static physics model [32] utilized in [2–4]. In our work, we model robot pushing as a continuous POMDP. [33] discretizes the state space. Discretizing an inherently continuous valued problem can increase the state space size exponentially w.r.t. the number of dimensions (also known as the state-space explosion problem)

III. PRELIMINARIES

In this section, we will first present the problem statement and subsequently discuss DDP algorithms, which we will then extend to hybrid control in the following sections.

A. Problem statement

In finite time-discrete hybrid sequential control, at each time step t, the agent executes continuous control ut and discrete control at out of N_a possibilities, incurs a cost of the form c(x, u, a) for intermediate time steps and traditionally c(x) for the final time step. subsequently, the world changes from state x_t to state x_{t+1} according to a dynamics function x_{t+1} = f(x_t, u_t, a_t). The goal is to minimize the total cost c(x_T) + \sum_{t=0}^{T-1} c(x_t, u_t, a_t) over a finite horizon T.

In a partially observable Markov decision process (POMDP), the agent does not directly observe the system state but instead makes an observation z_t at each time step t. In a POMDP, state dynamics and the observation function are stochastic. While the agent does not observe the current state directly, the agent can choose the control based on the

1 In reinforcement learning and AI research s often denotes the state, a the control or action, and rewards, corresponding to negative costs, are used for specifying the optimization goal.
belief $P(x_t)$, a probability distribution over states, which summarizes past controls and observations and is a sufficient statistic for optimal decision making.

Because solving POMDPs exactly is intractable even for small discrete problems, Gaussian belief space planning methods [16], [17], [34] assume a multi-variate Gaussian distribution for the state space, transitions, and observations with a Gaussian initial belief distribution $x_0 \sim \mathcal{N}(\bar{x}_0, \Sigma_0)$. We denote with $\bar{x}$ the mean and with $\Sigma$ the covariance of the belief. In a hybrid control Gaussian POMDP, the next state depends on a possibly non-linear function $f(x_t, u_t, a_t)$:

$$x_{t+1} = f(x_t, u_t, a_t) + m,$$

where $m \sim \mathcal{N}(0, M(x, u, a))$ is state and control specific multi-variate Gaussian noise. The observation $z_t$ at each time step is specified by a possibly non-linear function $h(x_t)$ with additive Gaussian noise:

$$z_t = h(x_t) + n,$$

where $n \sim \mathcal{N}(0, N(x))$ is state specific multi-variate Gaussian noise.

The immediate cost function is of the form $c(x, u, a, \Sigma)$ for intermediate time steps and $c(x)$ for the final time step. Many methods use the covariance as term in their cost function to penalize high uncertainty in the state estimates. The goal is to minimize the expected sum of costs over the finite horizon.

### B. Differential dynamic programming

DDP [35] is a widely used method for trajectory optimization with fast convergence [36], [37] and the ability to generate feedback controllers. In this section, we discuss DDP and iterative linear quadratic Gaussian (iLQG) [38], a special version of DDP which disregards second order dynamics derivatives and adds regularization and line search to deal with non-linear dynamics.

We will describe DDP here briefly. Please, see [12] for a more detailed description. DDP optimization starts from an initial nominal trajectory, a sequence of controls and states and then applies back and forward passes in succession. In the backwards pass, DDP quadratizes the value function around the nominal trajectory and uses dynamic programming to compute linear forward and feedback gains at each time step. In the forward pass, DDP uses the new policy to project a new trajectory of states and controls. The new trajectory is then used as nominal trajectory in the next backwards pass and so on. Details follow.

Denote the nominal trajectory with upper bars and the differences between the state, control, and covariance w.r.t. the nominal trajectory as $\Delta x = x - \bar{x}$, $\Delta u = u - \bar{u}$, respectively and denote a next time step with $\bar{t}$. DDP assumes that the value function is of the following form:

$$v(x) = V + \Delta x^T V_x + \Delta x^T V_{xx} \Delta x,$$

We now quadratize the one time step value difference w.r.t. states and controls, i.e.,

$$Q(\Delta x, \Delta u) = c_t(x + \Delta x, u + \Delta u) - c_t(x, u) + V_{t+1}(f(x + \Delta x, u + \Delta u)) - V_{t+1}(f(x, u)) \approx \frac{1}{2} \left[ \begin{array}{c} 1 \\ \Delta x \\ \Delta u \end{array} \right]^T \left[ \begin{array}{ccc} 0 & Q_x^T & Q_u^T \\ Q_x & Q_{xx} & Q_{xu} \\ Q_u & Q_{ux} & Q_{uu} \end{array} \right] \left[ \begin{array}{c} 1 \\ \Delta x \\ \Delta u \end{array} \right],$$

where

$$Q_x = c_x + f_x^T V_x'$$

$$Q_u = c_u + f_u^T V_u'$$

$$Q_{xx} = c_{xx} + f_x^T V_{xx} f_x + V_x' f_x^T$$

$$Q_{uu} = c_{uu} + f_u^T V_{uu} f_u + V_u' f_u^T$$

$$Q_{ux} = c_{ux} + f_u^T V_{ux} f_x + V_u' f_{ux}.$$

where we denoted derivatives w.r.t. the nominal trajectory as a function subindexed by the variables to derivate. We will use this convention through the whole paper. For example, $c_x = \frac{\partial c}{\partial x}(\bar{x}, \bar{u})$ denotes the derivative of the cost function $c(x, u)$ w.r.t. $x$ at the nominal trajectory point $\bar{x}, \bar{u}$. Note that iLQG ignores the last right hand side terms in Equations (7)-(9).

To minimize (4) w.r.t. $u$, we set the derivative of $Q(\Delta x, \Delta u)$ w.r.t. $u$ to zero and get

$$u = K(\Delta x) + k + \bar{u}$$

$$K = -Q^{-1}_{uu} Q_{ux}$$

where $K$ is the feedback gain and $k$ the forward gain. Plugging (10) into (4) yields the value function for the current time step:

$$\Delta V = \frac{1}{2} k^T Q_{uu} k + k^T Q_u$$

$$V_x = Q_x + K^T Q_u + K^T Q_{uu} k + Q_{xu}^T k$$

$$V_{xx} = \frac{1}{2} Q_{xx} + \frac{1}{2} K^T Q_{uu} K + Q_{xu}^T K$$

Recently, [12] introduced a version of DDP that allows efficient computation with hard control limits. Hard control limits require quadratic programming (QP) for computing the forward gains $k$:

$$k = \arg \min_{\Delta u} \frac{1}{2} \Delta u^T Q_{uu} \Delta u + \Delta u^T Q_u$$

$$u_{LB} \leq u + \Delta u \leq u_{UB},$$

where $u_{LB}$ and $u_{UB}$ are the lower and upper limits, respectively. For the feedback gain matrix $K$, the rows corresponding to clamped controls are set to zero. To solve QPs, [12] provides a gradient descent algorithm which allows initialization of the QP with a previously computed forward gain. Good initialization makes the approach computationally efficient. In the next Section, we will discuss how to extend the QP approach with equality constraints which allows the specification of action probabilities needed by our hybrid control approach.
IV. HYBRID TRAJECTORY OPTIMIZATION

The first problem with optimizing a hybrid control policy is that the discrete action choice depends on the combination of discrete actions at all time steps resulting in an exponential combinatorial explosion. The second problem is that in nonlinear problems we can adjust the approximation error due to linearization for continuous controls but not for discrete controls. In the continuous case, for example, iLQG adjusts to linearization for continuous controls but not for discrete linear problems we can adjust the approximation error due to combinatorial explosion. The second problem is that in non-discrete actions at all time steps resulting in an exponential amount of control change is not straightforward.

Below, we present three approaches for optimizing hybrid control DDP policies. The first two are simple greedy approaches which we provide for comparison. The third more powerful approach uses a continuous mixture of discrete actions driving the mixture during optimization into single selection using a special cost function.

A. Greedy discrete action choice

During the DDP back pass, at each time step, the greedy approach computes the expected value \( \mathbb{E}[x|u,a] \) at the nominal state and control for each discrete action separately and selects the action which yields minimum expected cost. In the greedy approach, there is no feedback control for discrete actions, only the fixed actions computed during the DDP back pass.

B. Interpolated discrete action choice

The second approach interpolates between nominal and the new discrete controls optimized by the greedy approach above. The interpolated approach computes first new actions identically to the greedy approach, and, then selects a fraction of \( \alpha \) of discrete controls which differ from the nominal ones. The selection is done evenly over the time steps. For example, for \( \alpha = 0.5 \) every other new differing discrete control would be chosen. This heuristic tries to allow for smooth scaling of the linearization error with discrete controls but there is no strong theoretical justification.

C. Mixture of discrete actions

We propose to use a discrete action mixture approach that assigns a continuous pseudo-probability to each discrete action. During optimization we drive the mixture into single component selection through a specialized cost function, discussed further down.

Our modified control \( \hat{u} \) is

\[
\hat{u} = \begin{bmatrix} u \\ p_a \end{bmatrix},
\]

where \( p_a \) contains the action probabilities. The dimensionality of the controls increases by the number of discrete actions. For simplicity, we assume here a single discrete control variable but our approach directly extends to several discrete control variables. For several discrete controls, the dimensionality would either increase by the sum of discrete control dimensions if one wants to minimize computational effort by treating discrete controls as independent, or, according to the product of discrete control dimensions if one wants to use a single large combination of discrete controls for better accuracy.

For hybrid controls, the dynamics model \( f(x, u, a) \) depends on both continuous controls \( u \) and discrete actions \( a \). The new dynamics model is a mixture of the original one:

\[
\hat{f}(x, \hat{u}) = \sum_a p_a f(x, u, a). \tag{17}
\]

Note that \( \hat{f} \) is essentially identical to the “convexified” dynamics function in [8] for fully observable MINLP hybrid control. However, [8] does not present a mixture model for immediate cost functions and our special cost function further down that forces the system into a bang-bang solution differs from the one in [8] because of the positive-definite Hessian for cost functions in DDP.

For hybrid controls, we define the cost function \( c(x, u, a) \) in the mixture model as

\[
\hat{c}(x, \hat{u}) = \sum_a \phi(p_a)c(x, u, a), \tag{18}
\]

where \( \phi(\cdot) \) is a smoothing function to make the Hessian of the cost function positive-definite. In the experiments, we used a pseudo-Huber smoothing function

\[
\phi(p) = \sqrt{p^2 + 0.01^2} - 0.01 \tag{19}
\]

which is close to linear but has a positive second derivative except.

To optimize the forward and feedback gains during dynamic programming, we add the following inequality and equality constraints for the probabilities to the quadratic program in Equation \( \text{(15)} \):

\[
0 \leq p \leq 1, \quad \sum_a p_a = 1. \tag{20}
\]

In order to utilize the efficient gradient descent method for quadratic programming from [12], we extend the algorithm to deal with the equality constraint \( \text{(20)} \). Shorty: 1) we subtract the mean from the search direction of probabilities guaranteeing that the equality constraint is satisfied, 2) we modify the Armijo line search step size dynamically so that the solution does not overstep probability equality constraints.

\( \hat{f}(x, u, p) \) can be seen as the expected dynamics of a partially stochastic policy. Such expected dynamics is often invalid, that is, there may not be any actual control values that would result in such dynamics. For example, in the box pushing application a mixture of discrete actions could correspond to pushing with several fingers at the same time although the robot may only have one finger. However, allowing for stochastic discrete actions in the beginning of optimization allows convergence to a good solution, even if we force the actions to become deterministic in the latter part of optimization. Our optimization procedure takes care of the major problems with sequential decision making.
with hybrid controls: the procedure allows to continuously decrease the approximation error due to linearization making local updates possible and is not subject to the combinatorial explosion of discrete action combinations. Next we discuss how to force a deterministic policy for discrete actions.

**Forcing deterministic discrete actions.** In the end, we want a valid deterministic policy for discrete actions. Therefore, we explicitly assign a cost to stochastic discrete actions that increases during optimization driving stochastic discrete controls into deterministic ones. Entropy would be a natural, widely used, cost measure. However, the Hessian matrix controls into deterministic ones. Entropy would be a natural, widely used, cost measure. However, the Hessian matrix of an entropy based cost function is not positive-definite (the second derivative is always negative). Instead, we use the following similarly shaped smoothed piece-wise cost function for stochastic discrete actions

\[
c_{ST}(x, u, p) = C_{ST} \sum_a \begin{cases} 
\phi(p_a) & \text{if } p_a < p_{th} \\
\phi \left( \frac{1 - p_a}{p_{th} - p_{th}} \right) & \text{if } p_a \geq p_{th}
\end{cases},
\]

where \( p_{th} = 1/N_a \) and \( C_{ST} \) is an adaptive constant. Note that while \( c_{ST}(x, u, p) \) is discontinuous at \( p_{th} \), derivatives can be computed below and above \( p_{th} \). When at \( p_{th} \), which corresponds to a uniform distribution, the cost achieves its maximum. The cost drives solutions away from \( p_{th} \) for increasing \( C_{ST} \). Tiny image on the right shows the cost function (solid) and its second derivative (dashed) for two discrete states and \( C_{ST} = 1 \). The x-axis shows the state probability.

**Practicalities.** During optimization we double \( C_{ST} \) everytime the cost decrease between DDP iterations is below a threshold. For a prespecified number of DDP iterations, we set \( C_{ST} \) to a maximum value when half of the iterations have elapsed. This allows for both smoothly increasing determinicity of discrete actions and then finally guaranteeing deterministic action selection.

Because the linear feedback control affects probabilities we normalize probabilities during the forward pass and select the most likely discrete action during evaluation.

V. HYBRID TRAJECTORY OPTIMIZATION FOR POMDPS

We are now ready to discuss hybrid control of POMDP trajectories. We start with the iLQG based approach for POMDPs [16] and then discuss how the iLQG POMDP approach can be extended with hybrid controls and hard control limits.

**A. DDP for POMDP trajectory optimization**

Instead of states, the POMDP approach of [16] operates on Gaussian beliefs. We will briefly discuss how to update the belief using an extended Kalman filter (EKF) and then how to utilize DDP for updating the control policy.

The EKF update equations (please, see [16] for details) for the belief \( \mathcal{N}(\hat{x}, \Sigma) \) are

\[
\begin{align*}
\hat{x}_{t+1} &= f(\hat{x}_t, u_t) + L_t(z_t - h(f(\hat{x}_t, u_t))) \\
\Sigma_{t+1} &= \Gamma_t - L_tH_t\Gamma_t,
\end{align*}
\]

where

\[
\begin{align*}
\Gamma_t &= A_t\Sigma_tA_t^T + M(\hat{x}, u) \\
L_t &= \Gamma_tH_t(\Gamma_t^{-1}H_t + N(f(\hat{x}_t, u_t)))^{-1} \\
A_t &= \partial f(\hat{x}, \bar{u})/\partial x(f(\hat{x}, \bar{u})).
\end{align*}
\]

We get

\[
\begin{align*}
\hat{x}_{t+1} &= f(\hat{x}_t, u_t) + w, \ w \sim \mathcal{N}(0, W(\hat{x}_t, u_t, \Sigma_t)) \\
\Sigma_{t+1} &= \Phi(\hat{x}_t, u_t, \Sigma_t),
\end{align*}
\]

where

\[
\begin{align*}
W(\hat{x}_t, u_t, \Sigma_t) &= L_tH_t\Gamma_t \\
\Phi(\hat{x}_t, u_t, \Sigma_t) &= \Gamma_t - L_tH_t\Gamma_t
\end{align*}
\]

The iLQG POMDP approach of [16] extends iLQG to POMDPs by using the Gaussian belief instead of the fully observable state. For the value function, iLQG POMDP linearizes the covariance in addition to quadraticizing states and controls. In more detail, replace in the equations in Section III-B the state \( x \) with the mean \( \hat{x} \) and add additional terms shortly discussed. The value function becomes

\[
\nu(\hat{x}, \Sigma) = V + \Delta\hat{x}^T V_{\hat{x}} + \Delta\hat{x}^T V_{\hat{x}\Sigma}\Delta\hat{x} + V^T_{\hat{x}\Sigma}\text{vec}[\Delta\Sigma],
\]

If we add covariance based terms to (5) and (6):

\[
\begin{align*}
Q_{\hat{x}} &= c_{\hat{x}} + f_{\hat{x}}^T V_{\hat{x}} + \Phi_{\hat{x}} V_{\Sigma} + \frac{1}{2} W_{\hat{x}\Sigma}\text{vec}[V_{\hat{x}\Sigma}] \\
Q_u &= c_u + f_u^T V_{u} + \Phi_u V_{\Sigma} + \frac{1}{2} W_u\text{vec}[V_{u\Sigma}],
\end{align*}
\]

we can directly compute the forward and feedback gains with Equation (11). Finally, we get the covariance based value function component

\[
V_{\Sigma} = c_{\Sigma} + \Phi_{\Sigma}^T V_{\Sigma} + \frac{1}{2} W_{\Sigma}\text{vec}[V_{\Sigma\Sigma}].
\]

Interestingly, we are not aware of a DDP POMDP method which would use the last terms in Equations (7)–(9) although in the POMDP case derivative computations involving covariances most often dominate dynamics derivative computations.

**B. Hybrid control for POMDP trajectory optimization**

In the partially observable case, the direct and indirect cost of uncertainty propagates through \( V_{\Sigma} \) into other value function components and through this the optimal policy has to take uncertainty into account. However, uncertainty is not directly present in Equation (12). In fact, Equation (12) is identical in the fully and partially observable cases and we can directly optimize controls using the quadratic program (QP) shown in Equation (15). Therefore, we can use hard limits and equality constraints on continuous controls in iLQG POMDP which is one technical insight in this paper.
A question this raises is whether we can also apply our proposed hybrid control approach to iLQG POMDP? Yes. Since our method of discrete action mixtures described further down in Section IV-C hides the action mixture inside the dynamics and cost functions, and, since the observation function does not directly depend on the controls, we can directly use the proposed hybrid control approach for POMDP optimization.

VI. EXPERIMENTS

We experimentally validate our hybrid control approach “Mixture”, described in Section IV-C in two different simulations: autonomous car driving and pushing of an unknown box. We are not aware of previous algorithms for trajectory planning under uncertainty which operate directly on both continuous and discrete actions. However, in the car driving and box pushing applications, we can reasonably map the hybrid controls directly to continuous controls and compare against continuous iLQG [12] and continuous iLQG POMDP [16], extended to support hard control limits, as described in Section IV-B. Note that in many hybrid control applications, for example, with discrete switches or on-off valves, mapping hybrid controls to continuous ones may not be possible. These applications require a method such as the one proposed in this paper. We also compare our approach with greedy action selection “Greedy” described in Section IV-A, and interpolated greedy action selection “Interpolate” described in Section IV-B. We used a time horizon of \( T = 500 \) and ran up to 400 optimization iterations for each method. We set the maximum for \( C_{ST} \) (please, see Section IV-C) to 1.28. \( C_{ST} \) starts from zero, increases to 0.01, and then doubles every time the cost difference is below the threshold of 0.01 for box pushing and 0.0001 for car driving.

A. Autonomous car driving

In autonomous car driving, the robot drives a nonholonomic car, switches discrete gears, accelerates, and tries to steer the car to zero position and pose. The dynamics are identical to the car parking dynamics in [12] with a few differences. We have discrete gear control with three gears: reverse, 1st gear, and 2nd gear. For the reverse gear, velocity \( v_{\text{CAR}} \in [-1, 0] \) and acceleration \( a_{\text{CAR}} \in [-1, 0] \). For the 1st gear, \( v_{\text{CAR}} \in [0, 1] \) and \( a_{\text{CAR}} \in [0, 1] \). For the 2nd gear, \( v_{\text{CAR}} \in [0, 4] \) and \( a_{\text{CAR}} \in [0, 0.5] \). The velocity limits mean that if the driver suddenly switches from the 2nd to reverse gear, the car effectively stops. For initial controls we used 0 wheel angle, 1st gear, \( a_{\text{CAR}} = 0.1 \). Non-optimized code on an Intel i7 CPU took 0.39, 0.80, 0.81, and 0.91 seconds per iteration for the “iLQG”, “Greedy”, “Interpolate”, and “Mixture” methods, respectively. Fig. 2 displays the costs and Fig. 3 shows the resulting trajectories and discrete controls.

B. Pushing an object

We will now describe the pushing task where the goal is to push an unknown object into a predefined goal-zone.

State. We define the state as

\[
\mathbf{x} = (x^C, u, x^{CF}, \mu_c, c),
\]

where \( x^C = (x^C, y^C) \) denotes the center coordinates and \( u \) the rotation angle of the object. \( x^{CF} = (x^{CF}, y^{CF}) \) denotes the center of friction (CF) coordinates relative w.r.t. \( x^C, \mu_c \) denotes the friction between the end effector and the object and \( c \) the friction coefficient describing the distribution of friction between object and supporting surface [32]. We assume that object edge locations w.r.t. \( x^C \) are fully observable.

Control action. When pushing we keep the speed of the robot hand constant while using sufficient force to move the robot hand. The control action is

\[
\mathbf{u} = (e, u^c, \alpha^p),
\]

where the control consists of the discrete edge \( e \) of the object to push, the continuous contact point location \( u^c \), \( 0 \leq u^c \leq 1 \), along the edge, and the pushing angle \( \alpha^p \) w.r.t. the normal unit vector at the contact point w.r.t. the edge. It is possible to parameterize \( e \) and \( u^c \) into a single continuous control. We do this for the continuous control version of iLQG. The continuous control version may not be able to jump easily from one discrete edge to another because of the dynamics discontinuity at the corners and because of potentially missing the box when box pose is uncertain: we place a cost on pushing close to corners depending on object pose uncertainty because the robot may otherwise miss the box. When pushing the box the finger of the robot can slide along the pushed edge. Combining sliding and non-sliding dynamics we get the pushing dynamics as described in [32].

Observations. At each time step the robot makes an observation about the pose of the object. The observation function in Equation 2 is then \( h(x_t) = (x^C, u) \).

Cost function. Our cost function rewards the robot when pushing the object into the predefined target pose, penalizes
the robot for final uncertainty, penalizes for the object to be close to obstacles, and penalizes for pushing too close to the edge corner relative to the rotation angle uncertainty.

In the box pushing experiment, the goal is to push the box to the target location at zero position. Position (1, 1) contains a soft obstacle with a cost. In the fully observable version the robot’s planned pushing location and angle correspond to the real ones. In the partially observable POMDP problem “Box POMDP” the controls are w.r.t. the planned pose and not the actual pose, and the robot may miss the box resulting in the box not moving. The initial standard deviation (SD) for the xy-coordinates is 0.01 and for the rotation angle 0.1. In “Box POMDP”, the friction parameters are known. In the “Box unknown” experiments, the coordinates of the center of friction are unknown and have initially a SD of 0.2 and in “Box all unknown” also the friction parameters $\mu_c$ and $c$ are not known and have initially a SD of 0.2. At each time step the robot gets a noisy observation about the box position with SD 0.0001 and angle of the box with SD 0.033. The SD of dynamics noise for xy-coordinates and rotation is 0.01. In the box pushing experiments the location of the center of friction was varied uniformly from the left bottom coordinates (0.2, 0.2) to the top right coordinates (0.8, 0.8). Friction parameters $\mu_c$ and $c$ had a mean of 1. The “iLQG”, “Greedy”, “Interpolate”, and “Mixture” methods, on the “Box”/“Box POMDP”/“Box unknown”/“Box all unknown” problems took 0.86/3.03/4.32/8.61, 1.36/3.53/4.55/8.74, 1.37/3.75/5.28/10.95, and 1.59/5.80/8.60/18.06 seconds per iteration, respectively. Fig. 2 shows mean costs. Fig. 4 shows high and low cost examples for both the “Box POMDP” and “Box unknown” problems for the continuous iLQG POMDP method [16] and our mixture method.

VII. CONCLUSIONS

We presented a novel DDP approach with linear feedback control for hybrid control of trajectories under uncertainty. The experiments indicate that our approach is useful, especially in POMDP problems. In the future, we plan on applying the approach to a real robot. We may use online replanning to improve the results further.

REFERENCES

[1] C. Kirches, S. Sager, H. G. Bock, and J. P. Schlöder, “Time-optimal control of automobile test drives with gear shifts,” Optimal Control Applications and Methods, vol. 31, no. 2, pp. 137–153, 2010.
[2] M. Dogar and S. Srinivasa, “Push-grasping with dexterous hands: Mechanics and a method,” in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2010.
[3] ——, “A planning framework for non-prehensile manipulation under clutter and uncertainty,” Autonomous Robots, vol. 33, no. 3, pp. 217–236, 2012.
[4] M. Koval, N. Pollard, and S. Srinivasa, “Pre- and post-contact policy decomposition for planar contact manipulation under uncertainty,” in Proceedings of Robotics: Science and Systems (R:SS), 2014.
[5] Y. Kawajiri and L. T. Biegler, “Large scale optimization strategies for zone configuration of simulated moving beds,” Computers & Chemical Engineering, vol. 32, no. 1, pp. 135–144, 2008.
[6] M. S. Branicky, V. S. Borkar, and S. K. Mitter, “A unified framework for hybrid control: Model and optimal control theory,” IEEE Transactions on Automatic Control, vol. 43, no. 1, pp. 31–45, 1998.
[7] A. Bemporad, G. Ferrari-Trecate, M. Morari, et al., “Observability and controllability of piecewise affine and hybrid systems,” IEEE Transactions on Automatic Control, vol. 45, no. 10, pp. 1864–1876, 2000.
[8] S. Sager, Numerical methods for mixed-integer optimal control problems, Der andere Verlag Tönning, Lübeck, Marburg, 2005.
[9] N. N. Namula and S. Bhardia, “A multiple model approach for predictive control of nonlinear hybrid systems,” Journal of process control, vol. 18, no. 2, pp. 131–148, 2008.
[10] V. Azemyakov, R. Galvan-Guerra, and M. Egerstedt, “Hybrid lq-optimization using dynamic programming,” in Proceedings of the American Control Conference. IEEE, 2009, pp. 3617–3623.
[11] F. Zhu and P. J. Antsaklis, “Optimal control of hybrid switched systems: A brief survey,” Discrete Event Dynamic Systems, vol. 25, no. 3, pp. 345–364, 2015.
[12] Y. Tassa, N. Mansard, and E. Todorov, “Control-limited differential dynamic programming,” in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2014, pp. 68–1175.
[13] P. Riedinger, F. Kratz, C. Jung, and C. Zanne, “Linear quadratic optimization for hybrid systems,” in IEEE Conference on Decision and Control, vol. 3. IEEE, 1999, pp. 3059–3064.
[14] O. Von Stryk and R. Bulirsch, “Direct and indirect methods for trajectory optimization,” Annals of operations research, vol. 37, no. 1, pp. 357–373, 1992.
[15] R. Platt Jr, R. Tedrake, L. Kaelbling, and T. Lozano-Perez, “Belief space planning assuming maximum likelihood observations,” in Robotics: Science and Systems (RSS), 2010.
[16] J. van den Berg, S. Patil, and R. Alterovitz, “Efficient Approximate Value Iteration for Continuous Gaussian POMDPs,” in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). AAAI Press, 2012.
[17] S. Patil, G. Kahn, M. Laskay, J. Schulman, K. Goldberg, and P. Abbeel, “Scaling up gaussian belief space planning through covariance-free trajectory optimization and automatic differentiation,” in Algorithmic Foundations of Robotics XI. Springer, 2015, pp. 515–533.
[18] B. Lincoln and B. Bernhardsson, “Lqr optimization of linear system switching,” Automatic Control, IEEE Transactions on, vol. 47, no. 10, pp. 1701–1705, 2002.
[19] C. Daniel, H. van Hoof, J. Peters, and G. Neumann, “Probabilistic inference for determining options in reinforcement learning,” in Proceedings of The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery. Springer, 2016. [Online]. Available: http://www.ausy.tu-darmstadt.de/uploads/ Site/EditorPublication/Daniel2016ECML.pdf
[20] S. LaValle and J. Kuffner, “Rapidly-exploring random trees: Progress and prospects,” Algorithmic and Computational Robotics: New Directions, pp. 293–308, 2001.
[21] M. S. Branicky and M. M. Curtiss, “Nonlinear and hybrid control via rrt,” in Proc. Intl. Symp. on Mathematical Theory of Networks and Systems, vol. 750, 2002.
[22] C. Zito, R. Stolkin, M. Kopicki, and J. L. Wyatt, “Two-level RRT planning for robotic push manipulation,” in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2012, pp. 678–685.
[23] M. Egerstedt, S.-i. Azuma, and Y. Wardi, “Optimal timing control of switched linear systems based on partial information,” Nonlinear Analysis: Theory, Methods & Applications, vol. 65, no. 9, pp. 1736–1750, 2006.
[24] S.-i. Azuma, M. Egerstedt, and Y. Wardi, “Output-based optimal timing control of switched systems,” in International Workshop on Hybrid Systems: Computation and Control. Springer, 2006, pp. 64–78.
[25] S. Ross, B. Chaib-draa, and J. Pineau, “Bayesian reinforcement learning in continuous POMDPs with application to robot navigation,” in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2008, pp. 2845–2851.
[26] P. Dallaire, C. Besse, S. Ross, and B. Chaib-draa, “Bayesian reinforcement learning in continuous POMDPs with Gaussian Processes,” in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2009, pp. 2604–2609.
[27] H. Bai, D. Hsu, and W. S. Lee, “Planning how to learn,” in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2013, pp. 2853–2859.
(a) iLQG (b) Greedy (c) Interpolate (d) Mixture

Fig. 3. Deterministic autonomous car experiment. The goal is to drive the car quickly to zero position. For each method, the optimized trajectories and the bottom row the selected gear at each time step. A low cost policy drives fast early and slows down late. Because of discontinuities, continuous iLQG has difficulty switching from 1st to 2nd gear and can not achieve maximum velocity. Because policy improvement starts in DDP from the last time step and proceeds to the first, greedy iLQG “Greedy” switches to 2nd gear not until at the last time steps. The proposed hybrid mixture method “Mixture” accelerates early and breaks late by switching to the reverse gear temporarily.

(a) iLQG (b) Mixture (c) iLQG (d) Mixture

Fig. 4. Stochastic box pushing experiment. The goal is to push the box to zero position. For (a) and (b) the box position and rotation are uncertain and partially observable, but for (c) and (d) also the center of friction is initially uncertain. For continuous iLQG and the mixture method, the top row shows the best performing policy and the bottom row the worst performing policy. In the plots, green rectangle denotes pushing finger, red dot actual center of friction, blue planned mean trajectory, and sampled trajectories (20 in each plot) are red. For three of the sampled trajectories and the planned mean trajectory we display four boxes and finger configurations, each, distributed evenly over time. In the worst case, continuous iLQG can not escape local optimums: the cost for potentially missing the box prevents switching pushing sides. In the POMDP problem, for a suitable center of friction, continuous iLQG computes a good policy but in the unknown POMDP problem, even for a suitable center of friction, continuous iLQG has some run away trajectories because it cannot adjust the edge to push as well as the mixture method.

[28] ——. "Integrated perception and planning in the continuous space: A POMDP approach," The International Journal of Robotics Research, vol. 33, no. 9, pp. 1288–1302, 2014.
[29] A.-A. Agha-Mohammadi, S. Chakravorty, and N. M. Amato, “FIRM: Sampling-based feedback motion planning under motion uncertainty and imperfect measurements,” The International Journal of Robotics Research, vol. 33, no. 2, pp. 268–304, 2014.
[30] A. Agha-mohammadi, S. Agarwal, A. Mahadevan, S. Chakravorty, D. Tomkains, J. Denny, and N. M. Amato, “Robust online belief space planning in changing environments: Application to physical mobile robots,” in Proc. of IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2014, pp. 149–156.
[31] M. Kopicki, S. Zurek, R. Stolkin, T. Moerwald, and J. L. Wyatt, “Learning modular and transferable forward models of the motions of push manipulated objects,” Autonomous Robots, pp. 1–22, 2016.
[32] K. M. Lynch, H. Mackawa, and K. Tanie, “Manipulation and active sensing by pushing using tactile feedback,” in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 1992, pp. 416–421.
[33] M. C. Koval, N. S. Pollard, and S. S. Srinivasa, “Pre-and post-contact policy decomposition for planar contact manipulation under uncertainty,” The International Journal of Robotics Research, vol. 35, no. 1-3, pp. 244–264, 2016.
[34] D. J. Webb, K. L. Crandall, and J. van den Berg, “Online Param-eter Estimation via Real-Time Replanning of Continuous Gaussian POMDPs,” in Proceedings of the IEEE International Conference on
[35] D. Mayne, “A second-order gradient method for determining optimal trajectories of non-linear discrete-time systems,” *International Journal of Control*, vol. 3, no. 1, pp. 85–95, 1966.

[36] D. Jacobson and D. Mayne, “Differential dynamic programming,” 1970.

[37] L.-z. Liao and C. A. Shoemaker, “Advantages of differential dynamic programming over newton’s method for discrete-time optimal control problems,” Cornell University, Tech. Rep., 1992.

[38] E. Todorov and W. Li, “A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems,” in *Proceedings of the American Control Conference*. IEEE, 2005, pp. 300–306.