Abstract
This paper presents a problem of model learning for the purpose of learning how to navigate a ball to a goal state in a circular maze environment with two degrees of freedom. The motion of the ball in the maze environment is influenced by several non-linear effects such as dry friction and contacts, which are difficult to model physically. We propose a semiparametric model to estimate the motion dynamics of the ball based on Gaussian Process Regression equipped with basis functions obtained from physics first principles. The accuracy of this semiparametric model is shown not only in estimation but also in prediction at n-steps ahead and its compared with standard algorithms for model learning. The learned model is then used in a trajectory optimization algorithm to compute ball trajectories. We propose the system presented in the paper as a benchmark problem for reinforcement and robot learning, for its interesting and challenging dynamics and its relative ease of reproducibility.
Semiparametrical Gaussian Processes Learning of Forward Dynamical Models for Navigating in a Circular Maze

Diego Romeres\textsuperscript{1}, Devesh K. Jha\textsuperscript{1}, Alberto DallaLibera\textsuperscript{2}, Bill Yerazunis\textsuperscript{1} and Daniel Nikovski\textsuperscript{1}

\textbf{Abstract}— This paper presents a problem of model learning for the purpose of learning how to navigate a ball to a goal state in a circular maze environment with two degrees of freedom. The motion of the ball in the maze environment is influenced by several non-linear effects such as dry friction and contacts, which are difficult to model physically. We propose a semiparametric model to estimate the motion dynamics of the ball based on Gaussian Process Regression equipped with basis functions obtained from physics first principles. The accuracy of this semiparametric model is shown not only in estimation but also in prediction at n-steps ahead and its compared with standard algorithms for model learning. The learned model is then used in a trajectory optimization algorithm to compute ball trajectories. We propose the system presented in the paper as a benchmark problem for reinforcement and robot learning, for its interesting and challenging dynamics and its relative ease of reproducibility.

I. INTRODUCTION

One challenge of learning in physical systems is that data are expensive to obtain, and system dynamics are affected by non-linear and discontinuous phenomena such as dry friction, contact, hysteresis, etc \cite{1}. It is, in general, very difficult to accurately model and/or learn the effect of these non-linear phenomena. Our work is motivated by learning the dynamics of physical systems that exhibit (highly) non-linear and discontinuous dynamics for the purpose of precise optimizing control. Deriving accurate models from first principles of physics could be extremely challenging, and in its turn, learning a model solely from data could be prohibitively expensive and inaccurate in sparsely sampled regions of state-space. For this reason, semiparametric models which combine physical analytical models and data-driven techniques might be advantageous \cite{2}, \cite{3}, \cite{4}.

This paper focuses on learning a dynamical model for the purpose of controlling the motion of a ball in a circular maze environment (see Fig. 1). The motion of the ball in the maze environment is significantly influenced by viscous and dry friction, contacts with the walls of the maze, and delays in sensing and actuation. Furthermore, the geometry of the maze constrains the motion of the ball to be discontinuous in the vicinity of each of the openings in each of the rings (see Fig. 1). As a result, it exhibits rich, highly non-linear dynamics. In this paper, we present results for model learning of the ball dynamics in a circular maze environment (CME), and initial results for trajectory optimization using the learned models in a model-based reinforcement learning (MBRL) setting.

\textbf{Related Work.} RL has seen explosive growth in recent years with several groundbreaking results in the areas of video games \cite{5}, \cite{6}, robotics \cite{7}, etc. RL algorithms can be broadly classified into model-free and model-based \cite{8}, \cite{9}. In the former class, policy is directly inferred using either a policy or a value-based method \cite{10}, \cite{11}. In the latter, a model for the forward dynamics of the system is learned, and is then used to compute a policy to solve the desired task \cite{12}, \cite{13}, \cite{14}. Model learning could result in better generalization across several different tasks. However, learning global dynamics for nonlinear systems is generally considered a challenging problem. The use of Gaussian processes for model learning followed by a suitable controller design has been studied in the literature. Some of the related works which are closely related to the approach presented in this paper can be found in \cite{12}, \cite{15}, \cite{16}, \cite{17}, \cite{18}, \cite{19}.

Apart from these, maze environments have been used to study RL algorithms \cite{20}, \cite{21}, \cite{22}. The proposed CME presents a more complex geometry, which results in significantly more convoluted motion dynamics of the ball, as will be presented in the later sections. Furthermore, the action space in the circular maze system (CMS) is continuous, whereas the action spaces considered in the earlier work with mazes \cite{20}, \cite{21}, \cite{22} were discrete. In \cite{23}, a deep RL-based end-to-end learning for the same CME is presented. The idea in this paper is to learn policies in simulation and transfer to the real system after minimum fine-tuning.

\textbf{Contributions.} Our contributions are the following. First, we present a semiparametric modeling technique \cite{24}, \cite{25}, \cite{26} which combines the physical model of the system with data-driven Gaussian Process Regression \cite{27} to learn the forward dynamics of the ball in a novel, real physical system.
To the best of our knowledge, semi-parametric Gaussian process regression has only been used to learn inverse dynamics previously [28], [26], [24], [29]. Second, continuous-time forward dynamics are learned, and not discrete-time state transition functions. We estimate the ball acceleration signal, and then integrate it forward in time to predict the state; we do not directly estimate the successor state. Third, performance of the modeling choices are shown in long-horizon predictive capabilities (known to be a hard task for the compounding error effect) and compared with several other techniques to highlight its accuracy. Rollout accuracy is fundamental to obtain successful optimal control algorithms that use known dynamics. Fourth, we illustrate the control performance in the CME using the learned models and the Iterative Linear Quadratic Gaussian (iLQG) algorithm [30], [31]. We make simplifying assumptions to decompose the full goal-directed navigation into a sequence of sub-goals in each of the rings. We show that using the learned models, we can control the ball between arbitrary points in the maze environment, thus also navigating to the goal ring. In the remainder of this paper, CMS refers to the full system comprising of the CME and the tip-tilt platform.

II. EXPERIMENTAL CONFIGURATION

The electromechanical tip-tilt system is composed of the maze, a maze attachment plate, two gimbals, two RC servos, a 2D laser tilt sensor, a base plate and Arduino-based control processor (see Fig. 2). Camera input is provided by an Intel RealSense camera on an overhead mount (see Figure 2). All the communication from the camera to the image processing system, the laser sensor and the servo motors for measuring the inclination and driving the tip-tilt platform is done using the Robot Operating System (ROS) [32] at 30Hz.

The maze itself is a commercially available toy purchased from Amazon.com (see Fig. 1). A 3D printed plate with 2 positioning setscrews and 3 spring clips fits tightly inside the base of the maze and provides attachment points for the servo link rods and the gimbals. The gimbals constrain the platform to have pure rotational motion around its origin \( o_m \). The latter, represents the Cartesian coordinates of the rotational fulcrum identified by the gimbals in the inertial frame \((o_b, x, y, z)\). W.l.o.g., we can assume that the origin of the inertial frame coincides with \( o_m \) (see Fig. 3). The configuration of the platform could be fully described by the rotation matrix between frame \((o_b, x, y, z)\) and the frame attached to the maze \((o_m, x', y', z')\); this can be described in a more compact fashion using the yaw, pitch and roll angles. The base maze is constrained by servo link rods, attached to the maze base through Traxxis ball joints; the lower end of each link rod is connected to a bellcrank arm of a hobbyist-grade HS-805BB RC servo motor with PWM-based control. The motors are positioned in perpendicular directions, so that the transformations between \((o_b, x, y, z)\) and \((o_m, x', y', z')\) is described only by \( \gamma \) and \( \beta \), the roll and pitch angles. The HS-805BB servos allow only position control. The control target of the motors, namely the input of our system, will be denoted by the vector \( u = (u^\beta, u^\gamma) \). The HS-805BB controller takes in the order of 200ms to move from an initial to a desired angle. Indirect measurements of \( \gamma \) and \( \beta \) are acquired using an inexpensive 5mW diode laser (Adafruit productid 1054), placed under the maze plate such that the direction of its optical axis is always equal to \( z' \). The laser projects downward onto the non-moving tilt position sensor, a 10mm \( \times \) 10mm position sensitive diode (a FirstSensor DL100-7-PCBA3 position sensitive photodiode). This sensor emits a pair of voltages corresponding to the x and y component of the impinging light. The map from the laser sensor readings to \( \gamma \) and \( \beta \) is learned by calibrating the laser readings with a precision (0.1°) clinometer.

The radius of the maze is 110mm and the diameter of the ball is 12.75mm. The width of each of the openings that connect rings is around 16mm. This suggests that a fairly accurate algorithm is needed in order to control the ball in the CME. The position of the ball is expressed in polar coordinates by \( r \) and \( \theta \), which represent the distance between \( o_m \) and the center of the ball, and the rotation around \( z' \) respectively. The position is measured using the camera and an off-the-shelf blob detection algorithm; angular velocity \( \dot{\theta} \) is estimated using a Kalman filter. The entire CMS except for the camera is connected to a computer via an Arduino Mega 2560. The entire mechanical system with the exception of the bearings and servo motors were designed in-house with OpenSCAD, with most of the non-catalog parts printed in InPLA polymer on an Lulzbot TAZ 3 printer.
III. PROBLEM FORMULATION

In this section, we describe the task in the proposed CMS and the assumptions considered in order to decompose the original problem into sub-problems with lower complexity. Throughout the paper, we assume that the relationship between observations and state is deterministic and known. The system is thus fully defined by the combination of the state $x_k$ and the control inputs $u_k$, and it evolves according to the dynamics $p(x_{k+1}|x_k, u_k)$ which are composed of the ball dynamics in the maze and the tip-tilt platform dynamics. The goal is to learn an accurate model of the ball dynamics which can be used in a controller, $\pi(u_k|x_k)$, based on a trajectory optimization algorithm which allows the CMS to choose an action $u_k$ given the state observation $x_k$ to drive a ball from an initial condition to the target state. As a first simplification, we assume that the ball dynamics are independent of the radial dynamics in each of the individual rings – i.e., we quantize the radius of the ball position into the 5 rings of the maze, denominated from 1 to 5 starting from the most outer ring toward the most inner one. The radius of a ring and therefore the radius of the ball position is approximated with $r_d$, the mean radius of the ring in which the ball is moving. As discussed in Section II, the actuators for the tip-tilt platform allow only position control, and they experience significant delay (around 200 [ms]) in attaining the desired angle. Thus, we include the orientation of the tip-tilt stage as part of the state for our dynamical system.

The desired angle. Thus, we include the orientation of the ball dynamics which can be composed of a sequence of micro goals. For example, from any initial position, the ball should reach one of the openings and then move to the next ring. This sequence of movements should be repeated until the ball reaches the goal ring. This decomposition is motivated by discontinuity in ball dynamics when the ball moves to an adjacent ring through one of the openings. Such a decomposition of long horizon planning problems, with sparse discontinuities, are common in the robot learning literature [33], [34], [35]. While the full long-horizon control problem could be challenging to solve using a MBRL algorithm, the sub-problems for each of the micro-goals could be solved by model learning followed by trajectory optimization.

IV. PROPOSED LEARNING APPROACH

In this section, we describe the proposed solution to learn how to navigate a ball in the CME. First, we learn a Gaussian process-based model for movement in each ring. A full motion model for the entire maze is obtained by using an additional discrete model for the movement between rings. Second, the learned models are used to compute a controller using iLQG to learn a local policy to navigate in the maze.

A. Model Learning

We consider discrete-time systems:

$$x_{k+1} = f(x_k, u_k) + e_k$$

where $x_k \in R^7$ denotes the state, $u_k \in R^2$ the actions, and $e_k$ is assumed to be a zero mean white Gaussian noise with diagonal covariance that represents the uncertainty about the state at the discrete time instant $k \in [1, \ldots, T]$. The transition function $f$ maps the state-action pair to the successor state, which is assumed to evolve smoothly over time. Given the quantization of the radius introduced in Section III, we do not use the measurements of the radius of the ball’s position. Even though this leads to information loss, preventing the possibility of modeling complex dynamics behaviours such as collisions, it reduces modeling complexity. The radial dynamics in the model will be incorporated in future works. The state $x$ for the CMS can be decomposed into two subsets: the ball state $x^{ball} = [r^d, \theta, \dot{\theta}]$, and the tip-tilt platform state $x^t = [\beta, \dot{\beta}, \gamma, \dot{\gamma}]$. This decomposition is convenient, because it is reasonable to assume that the tip-tilt platform dynamics are independent of the ball dynamics which is

$$p(x_{k+1}^d|x_k, u_k) = p(x_{k+1}^{ball}^d|x_k^d, u_k) + p(x_{k+1}^t|x_k^t, u_k)$$

In the proposed approach, we do not learn directly the static map (1), but we instead learn the evolution of $x^t$, which is a second order continuous-time system (subscript $t$ denotes continuous time). This is different from earlier work presented in [12], [13] where the discrete-time map was learned. For our system, this reduces to the estimation of three ordinary differential equations for the acceleration functions of the positional states $[\theta, \beta, \gamma]$. This function can then be integrated forward in time to make predictions on future states. In our opinion, this choice has the advantage that $x^t$ is described by the underlying physical system, and this knowledge can be exploited more efficiently in learning dynamic models. Notice also that in discrete time the evolution of $x^{ball}$ given $x^t$, since we need to know the input $u_k$ until the next observation to make future predictions. The in the continuous case, we have that $x^{ball}$ is independent from $u_k$ given $x^t$, $x^{ball}_t$ and $x^{ball}_{t+1}$, which is

$$p(x^{ball}_t|x_t, u_t) = p(x^{ball}_t|x_t, x^{ball}_{t+1})$$

The continuous time models are then discretized to compute the models used for control. In the following, we focus on learning acceleration models to be able to describe (1). First, we model the dynamics of the tip-tilt platform. Because of (2), the tip-tilt stage dynamics (1) can be learned independently from the ball dynamics. The dynamics of $\beta$ and $\gamma$ have been shown to be modelled effectively from physical quantities with a linear second order continuous-time differential equation [36], [37]. Therefore, the acceleration functions $\ddot{\beta}_t = f^\beta(\dot{\beta}_t, \beta_t, \gamma_t, u^\theta_{kt})$ and $\ddot{\gamma}_t = f^\gamma(\dot{\gamma}_t, \beta_t, u^\theta_{kt})$ have been estimated using standard linear system identification techniques [38] using Prediction Error Methods, resulting in two linear ARX models for each angle $\beta$ and $\gamma$. 
Next, we focus on modeling the more complex ball dynamics. Additionally, we define a new state vector $\bar{x}_t$ obtained by stacking together $x_t$ and $\bar{x}_t$:

$$\theta_t = f^{ball}(x_t, \bar{x}_t) + e_t = f^{ball}(\bar{x}_t) + e_t$$ (3)

Most of the models introduced in this section are equipped with a set of unknown hyperparameters that we estimate using the maximization of the marginal likelihood. Since all the variables are referred to time $t$, from now on we will neglect the dependencies on $t$.

1) Physics Inspired Model: We are interested in determining an analytical expression of the forward dynamics (3), and using it for data-driven modeling. It is noted that this model only describes the motion of the ball in a circular ring, and doesn’t consider the tip-tilt platform. As described in Section III, the radial dynamics of the ball is ignored, and the analytical model will be computed for a given radius $r$ that is assumed to be constant. Using a standard Lagrangian approach [39], [40], [41], we obtain the following equations of motion (we skip a detailed derivation for brevity):

$$\dot{\theta} = \beta \sin(\gamma) - \beta^2 \sin(\theta) \cos(\gamma)^2 \cos(\theta) - 2\beta \dot{\gamma} \cos(\gamma) \cos(\theta)^2$$

$$+ 2\dot{\beta} \gamma \cos(\gamma) + \frac{1}{2}\dot{\gamma}^2 \sin(2\theta) \sin(\beta) \sin(\theta)$$

$$- g \sin(\gamma) \cos(\beta) \cos(\theta)/r$$ (4)

Equation (4) reflects the complex dynamics of ball movement due to the circular geometry of the maze when compared to relatively simple geometries considered earlier in [33], [34], [35]. We refer to this model as the Physical model, “P”. While $P$ in equation (4) may be inaccurate for the CME due to making simplifying assumptions, it does provide a well-defined structure to the model that can be hard to identify in data-driven models in an efficient fashion. This model structure is used and interpret each component of the right-hand side of equation (4) as basis functions of the feature space of a Bayesian linear model as defined in [27, Chp. 2]:

$$\phi^\top(\bar{x}) = \begin{bmatrix} \beta \sin(\gamma), -\beta^2 \sin(\theta) \cos(\gamma)^2 \cos(\theta), 2\beta \dot{\gamma} \cos(\gamma), \\
-2\dot{\beta} \gamma \cos(\gamma) \cos(\theta)^2, 0.5\dot{\gamma}^2 \sin(2\theta) \sin(\beta) \sin(\theta)g/r, \\
-\sin(\gamma) \cos(\beta) \cos(\theta)g/r \end{bmatrix}$$ (5)

The expression of the forward dynamics (3) becomes

$$\dot{\theta} = \phi^\top(\bar{x})w + e$$ (6)

where $w \sim N(0, \Sigma_w)$ is a vector of weights modeled as a zero mean Gaussian prior with diagonal covariance matrix $\Sigma_w$. This model is denoted as Physics-Inspired model, “PI”.

2) Nonparametric Model: According to the Gaussian Process Regression framework [27], the forward dynamics can be postulated in a data driven fashion as:

$$\dot{\theta} = g(\bar{x}) + e$$ (7)

where $g(\bar{x})$ is a Gaussian process with zero mean and covariance function $k(\bar{x}, \bar{x}') := \mathbb{E}[g(\bar{x})g(\bar{x}')]$ defined through a radial basis function (RBF) kernel. This data-driven model class is known to have high flexibility in modeling functions and also high prediction performance, because it extrapolates the dynamics directly from the data without relying on any physical assumptions. However, this method can perform poorly in sparsely sampled regions of the state-space. This model is denoted as the Nonparametric model, “NP”.

3) Semiparametric Model: Semiparametric models attempt to combine the benefits of both the physical and data-driven model classes, which are good global behavior and accuracy, respectively. There can be several ways of combining the two methods and choosing the a priori distribution, as shown in [24], but one of the most effective has been shown to be:

$$\dot{\theta} = \phi^\top(\bar{x})w + g(\bar{x}) + e$$ (8)

where $\phi^\top(\bar{x})$ are the basis function suggested by the physics and defined as (5), $w$ is a vector of parameters with a zero mean Gaussian prior and a diagonal covariance matrix $\Sigma_w$ and $g(\bar{x})$ is a zero mean Gaussian process with RBF covariance function. The covariance of the whole process is $k(\bar{x}, \bar{x}') := \phi^\top(\bar{x})\Sigma_w\phi(\bar{x}) + \mathbb{E}[g(\bar{x})g(\bar{x}')]$. This class of models is also known as semi-parametrical models, and it has been shown to be a successful technique in model learning for robotic applications in the recent literature [24], [25], [26]. This model is denoted as Semiparametric model, “SP”.

B. Radial Movement

The model of the continuous movement of the ball in each of the rings was described in Section IV-A. Next, we model the discrete movement of the ball between rings. To learn this movement, we first learn the spatial clusters where there is a transition in $r^d$ in the collected trajectories of the ball. Thus, we obtain two clusters in the vicinity of each opening, based on the sign of $\Delta r^d$. The radial transition between rings is then approximated by the vector joining the centroids of the two clusters on the two sides of an opening, and the control action $u$ is approximated as the maximum signal whose projection on the horizontal plane is parallel to the direction of the radial transition vector.

C. Trajectory Optimization using iLQG

In this section, we briefly describe the iLQG algorithm which is used for trajectory optimization. The discrete time dynamics $x_{k+1} = f(x_k, u_k)$ and the cost function are used to compute local linear models and a quadratic cost function for the system along a trajectory. These linear models are then used to compute optimal control inputs and local gain matrices by iteratively solving the associated LQG problem. We skip the mathematical details for space limitations. For more details, interested readers are referred to [31].

V. EXPERIMENTAL RESULTS

In this section, we will provide experimental results for the models presented in Section IV-A, and we will show their efficacy in the real system, CMS, by learning a policy as described in Section IV-C.

Data collection consisted of around 50 minutes of operation on CMS, when control actions were applied at 30Hz, and measurements from the laser and camera were recorded. During data collection, control actions were generated as a sum of 50 sine and cosine waves with random sampled frequency between $[0, 1.5]$ Hz and shift phases in $[0, 2\pi]$. The data set consists of the control actions matrix $U \in$
Finally, the output vector \( Y \), given by the values of \( \tilde{\theta} \), is computed using an acausal filter for our supervised model learning algorithm. Note that \( \tilde{\theta} \) are not needed online while controlling CMS and an acausal filter is important in order to reduce the delay and improve the accuracy of the signal to be learned. Finally, the data collected are divided by rings, and split into two to obtain 4 training and testing data sets (ring 5 is the target, so it is not considered in the learning algorithm). The time of use of the CMS needed by our learning algorithm is less then 25 minutes. To limit the computational complexity of the algorithms, the 4 training data sets are uniformly downsampled with coefficient \( c_{DS} \), in order to obtain training sets of size \( N_{tr} \) less than 5000 points; the size of the test set is denoted by \( N_{test} \). During data collection, the ball naturally tends to stay more in the outermost ring, with the effect that the number of data per ring decreases going from ring 1 to ring 4 (see Table I). Each of the models described in Section IV-A and denoted by P, PI, NP, and SP, has been learned and tested with the data set for each ring and in the real CMS. The quality of the learned models is evaluated in terms of nRMSE in training and test data for each ring and each model in Table I. Model P represents only an approximation of \( \tilde{\theta} \) and, as expected, its prediction performance is significantly worse than the other methods. Similarly, PI performs poorly, however, it outperforms P, confirming the benefit of using the nonlinear behaviour of the state suggested by the physics as basis functions. The best methods are NP and SP, and we can notice that SP always outperforms NP. This shows the higher robustness of the semiparametric model due to its use of physical basis functions. Moreover, as the ring becomes smaller, the prediction performance deteriorates for all the models, and the gap in the performances of NP and SP between training and test increases when the amount of data available decreases, namely in ring 3 and 4. This agrees with theoretical expectations, and is likely due to the lower variability present in the data set. Finally, for comparison purposes, we added the performance of Random Forest regression (RF) [42], a standard machine learning method based on a different paradigm than GPR. Table I highlights that NP and SP exhibit better performance than RF. In Figure 4, we compare the learning curves of the two best performing models (SP and NP) as the number of training data increases for ring 1. SP outperforms NP in terms of nRMSE at each evaluated \( N_{tr} \), and achieves the same performance with half the data (note SP performance at \( N_{tr} = 2152 \) and NP at \( N_{tr} = 3874 \)).

After the evaluation of the prediction of acceleration, we are now interested on evaluating the n-steps-ahead prediction performance of the actual state variables \( \theta \) and \( \dot{\theta} \). This hard modelling task is fundamental for trajectory optimization control algorithms. Results are shown over a 40-step horizon, which is commensurate with the time needed for the ball to reach an opening, as we will see later in the experiments on the CME. From the test datasets, we randomly chose trajectories of length 40 steps that terminate in front of an opening (gate). The models are initialized with the initial conditions of these trajectories, and rolled out with the same control actions. In Figure 5, the results of all the models in ring 1 are shown, evaluating the absolute errors in position \( |\theta_k - \hat{\theta}_k| \) and velocity \( |\dot{\theta}_k - \hat{\dot{\theta}}_k| \) at each prediction step \([1, \ldots, 40]\), and reporting the mean over the 226 trajectories. This error index has been chosen for better interpretability of the control performance.

### Table I: nRMSE performance in training and in test data in all the rings for all the presented models and a Random Forest model.

| Ring | \( N_{tr} \) | \( c_{DS} \) | RF | PI | NP | SP | \( N_{test} \) | RF | PI | NP | SP |
|------|----------------|----------------|----|----|----|----|----------------|----|----|----|----|
| 1    | 3874           | 0.18           | 2.07 | 0.49 | 0.31 | 0.28 | 19368          | 0.49 | 1.97 | 0.48 | 0.38 | 0.35 |
| 2    | 4365           | 0.22           | 1.56 | 0.63 | 0.31 | 0.29 | 17459          | 0.59 | 1.57 | 0.63 | 0.45 | 0.43 |
| 3    | 4606           | 0.19           | 1.32 | 0.66 | 0.19 | 0.16 | 9213           | 0.53 | 1.35 | 0.65 | 0.42 | 0.40 |
| 4    | 2814           | 0.15           | 1.10 | 0.79 | 0.02 | 0.01 | 2815           | 0.67 | 1.08 | 0.81 | 0.71 | 0.65 |

### Table II: Rollout performance for all the rings evaluated as the average absolute error in \( \theta \) at 20 and 40 steps ahead.

| \( t \) | n=20 | n=40 | n=20 | n=40 | n=20 | n=40 | n=20 | n=40 |
|--------|------|------|------|------|------|------|------|------|
| 1      | 0.40 | 1.32 | 0.11 | 0.26 | 1.36 | 3.01 | 0.11 | 0.25 |
| 2      | 0.21 | 0.57 | 0.15 | 0.33 | 0.87 | 2.09 | 0.15 | 0.33 |
| 3      | 0.24 | 0.64 | 0.19 | 0.34 | 0.99 | 2.26 | 0.18 | 0.32 |
| 4      | 1.81 | 1.21 | 0.46 | 1.76 | 0.65 | 2.45 | 1.74 | 1.18 |
This model is denoted as Nonparametric discrete, “NPd”.

All models diverge rapidly from the ground truth, except for SP and NP, which achieve competitive performance. Model SP consistently outperforms all the others with an average error in position of 14 degrees after 40 steps. The confidence interval at 95% for model SP and NP is also shown, from which we can see that there are trajectories where contacts and bounces make the dynamics hard to predict. Finally, in Table II, the average of the absolute error $|\theta_k - 2\theta_k|$ of the same trajectories is computed at 20 and 40 steps ahead for all the models (except P which performs poorly) in all the rings. Consistently, model SP and NP outperform the others in all of the rings. The significantly higher performance of NP w.r.t. NPd shows the benefit of formulating the learning problem as a dynamical system, as in (3), and not as a static map, as in (1). The performance in ring 4 deteriorates significantly, even for model SP, because as explained before, the data collected in this ring are much less and of lower quality in comparison to the other rings. This problem will become clear while controlling the ball in CME.

In this paper, the problem of learning how to control a ball in a CME is presented. Our approach offers an initial solution to navigate the ball to the goal-state of the CME. The goal of reaching the final target state is decomposed into a sequence of micro-goals which consist of navigating in each ring, each of which is solved by a trajectory optimization algorithm using the learned models. Accurate semiparametric GP models are learned to describe the acceleration function of the dynamical system. While the current controller can navigate the ball to the goal state in the CME reasonably well, it is not robust to changes in initial conditions and other disturbances due to contacts. This is left as future work. Moreover, we are investigating on how to apply GP online learning techniques to improve the model while iteratively improving the policy and comparing our approach with a method composing the prior physics knowledge with a neural network framework. Apart from this, the proposed CMS could be used to study several other interesting problems. For example, how to leverage the models learned for a single ball in the environment to multiple balls or solving the navigation problems composing reusable skills.

The CMS is suitable to study several problems in RL and robot learning presenting several challenges such as nonlinear and fast dynamics, contacts and planning problem, thus it could possibly be used as a benchmark system for its ease of reproducibility as it is starting to be more and more required by the community.

VI. CONCLUSION AND DISCUSSION

In this paper, the problem of learning how to control a ball in a CME is presented. Our approach offers an initial solution to navigate the ball to the goal-state of the CME. The goal of reaching the final target state is decomposed into a sequence of micro-goals which consist of navigating in each ring, each of which is solved by a trajectory optimization algorithm using the learned models. Accurate semiparametric GP models are learned to describe the acceleration function of the dynamical system. While the current controller can navigate the ball to the goal state in the CME reasonably well, it is not robust to changes in initial conditions and other disturbances due to contacts. This is left as future work. Moreover, we are investigating on how to apply GP online learning techniques to improve the model while iteratively improving the policy and comparing our approach with a method composing the prior physics knowledge with a neural network framework.

Apart from this, the proposed CMS could be used to study several other interesting problems. For example, how to leverage the models learned for a single ball in the environment to multiple balls or solving the navigation problems composing reusable skills.

The CMS is suitable to study several problems in RL and robot learning presenting several challenges such as nonlinear and fast dynamics, contacts and planning problem, thus it could possibly be used as a benchmark system for its ease of reproducibility as it is starting to be more and more required by the community.
[47] G. Prando, D. Romeres, and A. Chiuso, “Online identification of time-varying systems: A bayesian approach,” in *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 3775–3780, Dec 2016.

[48] A. Gijsberts and G. Metta, “Incremental learning of robot dynamics using random features,” in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 951–956, 2011.

[49] A. Ajay, J. Wu, N. Fazeli, M. Bauza, L. P. Kaelbling, J. B. Tenenbaum, and A. Rodriguez, “Augmenting Physical Simulators with Stochastic Neural Networks: Case Study of Planar Pushing and Bouncing,” *ArXiv e-prints*, Aug. 2018.

[50] P. Bacon, J. Harb, and D. Precup, “The option-critic architecture,” *CoRR*, vol. abs/1609.05140, 2016.

[51] J. Andreas, D. Klein, and S. Levine, “Modular multitask reinforcement learning with policy sketches,” *CoRR*, vol. abs/1611.01796, 2016.

[52] H. Sahni, S. Kumar, F. Tejani, and C. L. Isbell, “Learning to compose skills,” *CoRR*, vol. abs/1711.11289, 2017.

[53] A. R. Mahmood, D. Korenkeych, B. J. Komer, and J. Bergstra, “Setting up a reinforcement learning task with a real-world robot,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4635–4640, IEEE, 2018.

[54] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, “Deep reinforcement learning that matters,” *CoRR*, vol. abs/1709.06560, 2017.

[55] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, “Openai gym,” *CoRR*, vol. abs/1606.01540, 2016.