Data-Efficient Learning for Complex and Real-Time Physical Problem Solving Using Augmented Simulation

Kei Ota, Devesh K. Jha, Diego Romeres, Jeroen van Baar, Kevin A. Smith, Takayuki Semitsu, Tomoaki Oiki, Alan Sullivan, Daniel Nikovski, and Joshua B. Tenenbaum

***Abstract*—Humans quickly solve tasks in novel systems with complex dynamics, without requiring much interaction. While deep reinforcement learning algorithms have achieved tremendous success in many complex tasks, these algorithms need a large number of samples to learn meaningful policies. In this letter, we present a task for navigating a marble to the center of a circular maze. While this system is very intuitive and easy for humans to solve, it can be very difficult and inefficient for standard reinforcement learning algorithms to learn meaningful policies. We present a model that learns to move a marble in the complex environment within minutes of interacting with the real system. Learning consists of initializing a physics engine with parameters estimated using data from the real system. The error in the physics engine is then corrected using Gaussian process regression, which is used to model the residual between real observations and physics engine simulations. The physics engine augmented with the residual model is then used to control the marble in the maze environment using a model-predictive feedback over a receding horizon. To the best of our knowledge, this is the first time that a hybrid model consisting of a full physics engine along with a statistical function approximator has been used to control a complex physical system in real-time using nonlinear model-predictive control (NMPC).**

***Index Terms*—Reinforcement learning, cognitive control architectures.**

I. Introduction

ARTIFICIAL Intelligence has long had the goal of designing robotic agents that can interact with the (complex) physical world in flexible, data-efficient and generalizable ways [1], [2]. Model-based control methods form plans based on predefined models of the world dynamics. However, although data-efficient, these systems require accurate dynamics models, which may not exist for complex tasks. Model-free methods on the other hand rely on reinforcement learning, where the agents simultaneously learn a model of the world dynamics and a control policy [3], [4]. However, although these methods can learn policies to solve tasks involving complex dynamics, training these policies is inefficient, as they require many samples. Furthermore, these method are typically not generalizable beyond the trained scenarios.

Our aim in this letter is to combine the best of both methodologies: our system uses nonlinear model predictive control with a predefined (inaccurate) model of dynamics at its core, but updates that model by learning residuals between predictions and real-world observations via physical parameter estimation and Gaussian process regression [5]. We take inspiration from cognitive science for this approach, as people can interact with and manipulate novel objects well with little or no prior experience [6]. Research suggests people have internal models of physics that are well calibrated to the world [7], [8], and that they use these models to learn how to use new objects to accomplish novel goals in just a handful of interactions [9]. Thus, we suggest that any agent that can perform flexible physical problem solving should have both prior knowledge of the dynamics of the world, as well as a way to augment those dynamics in a way that supports their interactions with the scene. Note that we do not suggest that this specific approach corresponds to the way that humans learn or reason about physics, but instead that we believe augmented simulation is key to human sample efficiency, and therefore should be important for robotic sample efficiency as well. Fig. 1 provides an idea of the proposed approach.

Our testbed for this problem is a circular maze environment (CME; see Fig. 1), in which the goal is to tip and tilt the maze so as to move a marble from an outer ring into an inner circle. This is an interesting domain for studying real-time control because it is intuitively easy to pick up for people — even children play with similar toys without prior experience with these mazes — and yet is a complex learning domain for artificial agents due to its constrained geometry, underactuated control, nonlinear dynamics, and long planning horizon with several discontinuities [10], [11]. Adding to this challenge, the CME

Manuscript received October 15, 2020; accepted February 17, 2021. Date of publication March 25, 2021; date of current version April 8, 2021. This letter was recommended for publication by Associate Editor E. Ugur and Editor T. Asfour upon evaluation of the reviewers' comments. *(Corresponding author: Kei Ota.)*

Kei Ota, Takayuki Semitsu, and Tomoaki Oiki are with the Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura 247-8501, Japan (e-mail: Ota.Kei@ds.MitsubishiElectric.co.jp; semitsu.takayuki@dw.mitsubishielectric.co.jp; Oiki.Tomohiro@ds.MitsubishiElectric.co.jp).

Devesh K. Jha, Diego Romeres, Jeroen van Baar, Alan Sullivan, and Daniel Nikovski are with the Data Analytics, Mitsubishi Electric Research Labs, Cambridge, MA 02139 USA (e-mail: devesh.jha@merl.com; diego.romeres@gmail.com; jeroen@merl.com; sullivan@merl.com; nikovski@merl.com).

Kevin A. Smith and Joshua B. Tenenbaum are with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: k2smith@mit.edu; jbt@mit.edu).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2021.3068887, provided by the authors.

Digital Object Identifier 10.1109/LRA.2021.3068887

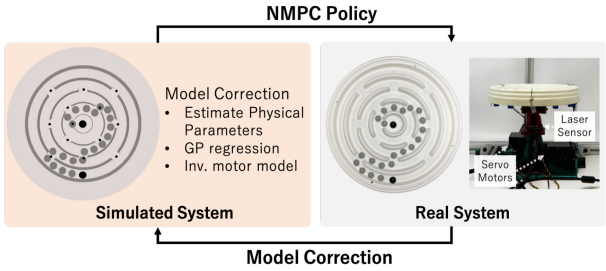

Fig. 1. We train a reinforcement learning agent that initializes a policy with a general purpose physics engine, then corrects its dynamics model using parameter estimation and residual learning. The agent uses this augmented model in a circular maze to drive a marble to the center.

is a system that is usually in motion, so planning and control must be done in real-time, or else the ball will continue to roll in possibly unintended ways.

The learning approach we present in this letter falls under the umbrella of Model-Based Reinforcement Learning (MBRL). In MBRL, a task-agnostic predictive model of the system dynamics is learned from exploration data. This model is then used to synthesize a controller which is used to perform the desired task using a suitable cost function. The model in our case is represented by a physics engine that roughly describes the CME with its physical properties. Additionally, we learn the residual between the actual system and the physics system using Gaussian process regression [5]. Such an augmented simulator – a combination of a physics engine and a statistical function approximator – allows us to efficiently learn models for physical systems while using minimal domain knowledge.

Contributions. Our main contributions are as follows:

- We present a novel framework where a hybrid model consisting of a full physics engine augmented with a machine learning model is used to control a complex physical system using NMPC in real time.
- We demonstrate that our proposed approach leads to sample-efficient learning in the CME: our agent learns to solve the maze within a couple of minutes of interaction.

We have released our code for the CME as it is a complex, low-dimensional system that can be used to study real-time physical control.[1]

II. Related Work

Our work is motivated by the recent advances in (deep) reinforcement learning to solve complex tasks in areas such as computer games [12] and robotics [3], [13]. While these algorithms have been very successful for solving simulated tasks, their applicability in real systems is sometimes questionable due to their relative sample inefficiency. This has motivated a lot of research in the area of transferring knowledge from a simulation environment to the real world [10], [14]–[16]. However, most of these techniques end up being very data intensive. Here we attempt to study complex physical puzzles using model-based agents in an attempt to learn to interact with the world in a sample-efficient manner.

[1]https://www.merl.com/research/license/CME

Recently the robotics community has seen a surge in interest in the use of general-purpose physics engines which can represent complex, multi-body dynamics [17]. These engines have been developed with the intention to allow real-time control of robotic systems while using them as an approximation of the physical world. However, these simulators still cannot model or represent the physical system accurately enough for control, and this has driven a lot of work in the area of sim-to-real transfer [18], [19]. The goal of these methods is to train an agent in simulation and then transfer them to the real system using minimum involvement of the real system during training. However, most of these approaches use a model-free learning approach and thus tend to be sample inefficient. In contrast, we propose a method that trains a MBRL sim-to-real agent and thus achieves very good sample efficiency.

The idea of using residual models for model correction, or hybrid learning models for control of physical systems during learning in physical systems has also been studied in the past [11], [20]–[23]. However, most of these studies use prior physics information in the form of differential equations, which requires domain expertise and thus the methods also become very domain specific. While we rely on some amount of domain expertise and assumptions, using a general purpose physics engine to represent the physical system will allow for more readily generalization across a wide range of systems.

A similar CME has been solved with MBRL and deep reinforcement learning, in [11] and [10], respectively. In [11], the analytical equations of motion of the CME have been derived to learn a semi-parametric GP model [24], [25] of the system, and then combined with an optimal controller. In [10], a sim-to-real approach has been proposed, where a policy to control the marble(s) is learned on a simulator from images, and then transferred to the real CME. However, the transfer learning still requires a large amount of data from the real CME.

While approaches that combine physical predictions and residuals have been used for control in the past [26], here we demonstrate that this combination can be used as part of a model-predictive controller (MPC) of a much more complex system in real-time. An important point to note here is that the work presented in [26] uses MPC in a discrete action space, whereas for the current system we have to use nonlinear model-predictive control (NMPC) that requires a solution to a nonlinear, continuous control problem in real-time (which requires non-trivial, compute-expensive optimization) [27]. Consequently, the present study deals with a more complicated learning and control problem that is relevant to a wide range of robotic systems.

III. Problem Formulation

We consider the problem of moving the marble to the center of the CME. Our goal is to study the sim-to-real problem in a model-based setting where an agent uses a physics engine as its initial knowledge of the environment's physics. Under these settings, we study and attempt to answer the following questions in the present paper.

1) What is needed in a model-based sim-to-real architecture for efficient learning in physical systems?

2) How can we design a sim-to-real agent that behaves and learns in a data-efficient manner?
3) How does the performance and learning of our agent compare against how humans learn to solve these tasks?

We use the CME as our test environment for the studies presented in this letter. However, our models and controller design are general-purpose and thus, we expect the proposed techniques could find generalized use in robotic systems. For the rest of the paper, we call the CME together with the tip-tilt platform the circular maze system (CMS). At this point, we would like to note that we make some simplifications for the CMS to model actuation delays and tackle discontinuities for controller design as we describe in the following text.

The goal of the learning agent is to learn an accurate model of the marble dynamics, that can be used in a controller, $\pi(\boldsymbol{u}_k|\boldsymbol{x}_k)$, in a model-predictive fashion which allows the CMS to choose an action $\boldsymbol{u}_k$ given the state observation $\boldsymbol{x}_k$ to drive a marble from an initial condition to the target state. We assume that the system is fully defined by the combination of the state $\boldsymbol{x}_k$ and the control inputs $\boldsymbol{u}_k$, and it evolves according to the dynamics $p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k, \boldsymbol{u}_k)$ which are composed of the marble dynamics in the maze and the tip-tilt platform dynamics.

As a simplification, we assume that the marble dynamics is independent of the radial dynamics in each of the individual rings, i.e., we quantize the radius of the marble position into the 4 rings of the maze. We include the orientation of the tip-tilt platform as part of the state for our dynamical system, obtaining a five-dimensional state representation for the system, i.e., $\boldsymbol{x} = (r_d, \beta, \gamma, \theta, \dot{\theta},)$. It can be noted that the radius r_d is a discrete variable, whereas the rest of the state variables are continuous. The terms β, γ represent the X and Y-orientation of the maze platform, respectively, and $\theta, \dot{\theta}$ represent the angular position and velocity of the marble, measured with respect to a fixed frame of reference. Since r_d is fixed for each ring of the CME, we remove r_d from the state representation of the CMS for the rest of the paper. Thus, the state is represented by a four-dimensional vector $\boldsymbol{x} = (\beta, \gamma, \theta, \dot{\theta})$. The angles β, γ are measured using a laser sensor that is mounted on the tip-tilt platform (see Figure 1) while the state of the ball could be observed from a camera mounted above the CMS. For more details, interested readers are referred to [11].

We assume that there is a discrete planner, which can return a sequence of gates that the marble can then follow to move to the center. Furthermore, from the human experiments we have observed that human subjects always try to bring the marble in front of the gate, and then tilt the CME to move it to the next ring. Therefore, we design a lower level controller to move the marble to the next ring when the marble is placed in front of the gate to the next ring. Thus, the task of the learned controller is to move the marble in a controlled way so that it can transition through the sequence of gates to reach the center of the CME. This makes our underlying control problem tractable by avoiding discontinuities in the marble movement (as the marble moves from one ring to the next).

Before describing our approach, we introduce additional nomenclature we will use in this letter. We represent the physics engine by f^{PE}, the residual dynamics model by f^{GP}, and the real system model by f^{real}, such that $f^{\text{real}}(\boldsymbol{x}_k, \boldsymbol{u}_k) \approx f^{\text{PE}}(\boldsymbol{x}_k, \boldsymbol{u}_k) + f^{\text{GP}}(\boldsymbol{x}_k, \boldsymbol{u}_k)$. We use MuJoCo [17] as the physics engine, however, we note that our approach is agnostic to the choice of physics engine. In the following sections, we describe how we design our sim-to-real agent in simulation, as well as on the real system.

IV. Approach

Our approach for designing the learning agent is inspired by human physical reasoning: people can solve novel manipulation tasks with a handful of trials. This is mainly because we rely on already-learned notions of physics. Following a similar principle, we design an agent whose notion of physics comes from a physics engine. The proposed approach is shown as a schematic in Fig. 2.

We want to design a sim-to-real agent, which can bridge the gap between the simulation environment and the real world in a principled fashion. The gap between the simulated environment and the real world can be attributed to mainly two factors. First, physics engines represent an approximation of the physics of the real systems, because they are designed based on limited laws of physics, domain knowledge, and convenient approximations often made for mathematical tractability. Second, there are additional errors due to system-level problems, such as observation noise and delays, actuation noise and delays, finite computation time to update controllers based on observations, etc.

Consequently, we train our agent by first estimating the parameters of the physics engine, and then compensate for the different system-level problems as the agent tries to interact with the real system. Finally, Gaussian process Regression is used to model the residual dynamics of the real system that cannot be described by the best estimated parameters of the physics engine. In the rest of this section, we describe the details of the physics engine for the CME, and provide our approach for correcting the physics engine as well as modeling other system-level issues with the CMS.

A. Physics Engine Model Description

As described earlier, we use MuJoCo as our physics engine, f^{PE}. Note that in our model we ignore the radial movement of the marble in each ring, and describe the state only with the angular position of the marble as described in Section III. Consequently, we restrict the physics engine to consider only the angular dynamics of the marble in each ring, i.e., the radius of the marble position is fixed. However, in order to study the performance of the agent in simulation, we also create a full model of the CME where the marble does not have the angular state constraint. Thus, we create two different physics engine models: $f^{\text{PE}}_{\text{red}}$ represents the reduced physics engine available to our RL model, and $f^{\text{PE}}_{\text{full}}$ uses the full internal state of the simulator. $f^{\text{PE}}_{\text{red}}$ differs from $f^{\text{PE}}_{\text{full}}$ in two key ways. In the forward dynamics of the $f^{\text{PE}}_{\text{red}}$ model, we set the location of the marble to be in the center of each ring because we cannot observe the accurate radial location of the marble in the real system, while this is tracked in $f^{\text{PE}}_{\text{full}}$. Additionally, because we cannot observe the spin of the ball in real experiments, we do not include it in $f^{\text{PE}}_{\text{red}}$, while it is included in $f^{\text{PE}}_{\text{full}}$. We use this $f^{\text{PE}}_{\text{full}}$ model for analyzing the behavior of our agent in the preliminary studies

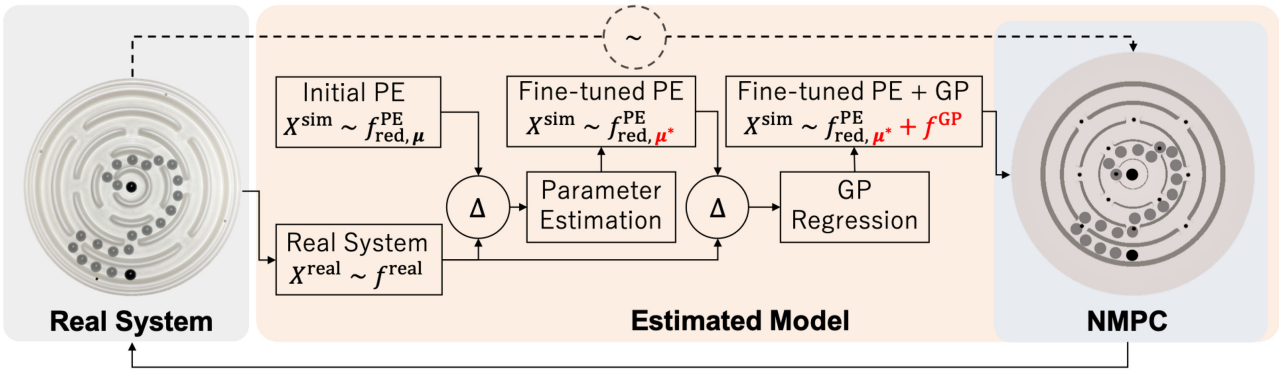

Fig. 2. The learning approach used in this letter to create a predictive model for the physics of the CME in the real system. We create a predictive model for the marble dynamics in the CME using a physics engine. We start with a MuJoCo-based physics engine (PE) with random initial parameters for dynamics, and estimate these parameters $\boldsymbol{\mu}^*$ from the residual error between simulated and real CME using CMA-ES. The remaining residual error between simulated and real CME is then compensated using Gaussian process (GP) regression during iterative learning. Finally, we use the augmented simulation model to control the real CME with NMPC policy.

Algorithm 1: Model Learning Procedure.

1: Collect N episodes in the real system using Alg. 2
2: Compute simulator trajectories as $f^{\text{PE}}_{\text{red},\boldsymbol{\mu}}(\boldsymbol{x}^{\text{real}}_k, \boldsymbol{u}^{\text{real}}_k)$, from the real system N episodes
3: *Estimate physical parameters* using CMA-ES
4: **while** Model performance not converged **do**
5: Collect N episodes in CMS using Alg. 2
6: Compute simulator trajectories $\boldsymbol{x}^{\text{sim}}_{k+1}$ for data in D
7: *Train residual GP model*
8: **end while**

Algorithm 2: Rollout an Episode Using NMPC.

1: Initialize time index $k \leftarrow 0$
2: Reset the real system by randomly placing the marble to outermost ring
3: **while** The marble does not reach innermost ring **and** not exceed time limit **do**
4: Set real state to simulator $\boldsymbol{x}^{\text{sim}}_k \leftarrow \boldsymbol{x}^{\text{real}}_k$
5: Compute trajectory $(X^{\text{sim}}, U^{\text{sim}})$ using NMPC
6: Apply initial action $\boldsymbol{u}^{\text{real}}_k = \boldsymbol{u}^{\text{sim}}_0$ to the real system
7: Store transition $D \leftarrow D \cup \{\boldsymbol{x}^{\text{real}}_k, \boldsymbol{u}^{\text{real}}_k, \boldsymbol{x}^{\text{real}}_{k+1}\}$
8: Increment time step $k \leftarrow k + 1$
9: **end while**

in simulation. This serves as an analog to the real system in the simulation studies we present in the paper. We call this set of experiments *sim-to-sim*. These experiments are done to determine whether the agent can successfully adapt its physics engine when initialized with an approximation of a more complicated environment.

B. Model Learning

We consider a discrete-time system:

$$\boldsymbol{x}_{k+1} = f(\boldsymbol{x}_k, \boldsymbol{u}_k) + \boldsymbol{e}_k, \tag{1}$$

where $\boldsymbol{x}_k \in \mathbb{R}^4$ denotes the state, $\boldsymbol{u}_k \in \mathbb{R}^2$ the actions, and $\boldsymbol{e}_k$ is assumed to be a zero mean white Gaussian noise with diagonal covariance, at the discrete time instant $k \in [1, ..., T]$.

In the proposed approach, the unknown dynamics f in Eq. 1 represents the CMS dynamics, f^{real}, and it is modeled as the sum of two components:

$$f^{\text{real}}(\boldsymbol{x}_k, \boldsymbol{u}_k) \approx f^{\text{PE}}_{\text{red}}(\boldsymbol{x}_k, \boldsymbol{u}_k) + f^{\text{GP}}(\boldsymbol{x}_k, \boldsymbol{u}_k), \tag{2}$$

where $f^{\text{PE}}_{\text{red}}$ denotes the physics engine model defined in the previous section, and f^{GP} denotes a Gaussian process model that learns the residual between real dynamics and simulator dynamics. We learn both the components $f^{\text{PE}}_{\text{red}}$ and f^{GP} to improve model accuracy. The approach is presented as psuedo-code in Algorithm 1 and described as follows.

1) Physical Parameter Estimation: We first estimate physical parameters of the real system. As measuring physical parameters directly in the real system is difficult, we estimate four friction parameters of MuJoCo by using CMA-ES [28]. More formally, we denote the physical parameters as $\boldsymbol{\mu} \in \mathbb{R}^4$, and the physics engine with the parameters as $f^{\text{PE}}_{\text{red},\boldsymbol{\mu}}$.

As described in Algorithm 1, we first collect multiple episodes with the real system using the NMPC controller described in Section IV-D. Then, CMA-ES is used to estimate the best friction parameters $\boldsymbol{\mu}^*$ that minimizes the difference between the movement of the marble in the real system and in simulation as:

$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu}} \frac{1}{\|D\|} \sum_{(\boldsymbol{x}^{\text{real}}_k, \boldsymbol{u}^{\text{real}}_k, \boldsymbol{x}^{\text{real}}_{k+1}) \in D} \|\boldsymbol{x}^{\text{real}}_{k+1} - f^{\text{PE}}_{\text{red},\boldsymbol{\mu}}(\boldsymbol{x}^{\text{real}}_k, \boldsymbol{u}^{\text{real}}_k)\|^2_{W_{\boldsymbol{\mu}}}, \tag{3}$$

where D represents the collected transitions in the real system, $W_{\boldsymbol{\mu}}$ is the weight matrix whose value is 1 only related to the angular position term of the marble θ_{k+1} in the state $\boldsymbol{x}_{k+1}$.

2) Residual Model Learning Using Gaussian Process: After estimating the physical parameters, a mismatch remains between the simulator and the real system because of the modeling limitations described in the beginning of this section. To get

a more accurate model, we train a Gaussian Process (GP) model via marginal likelihood maximization [5], with a standard linear kernel, to learn the residual between the two systems by minimizing the following objective:

$$L^{\mathrm{GP}} = \frac{1}{\|D\|} \sum_{(\boldsymbol{x}_k^{\mathrm{real}}, \boldsymbol{u}_k^{\mathrm{real}}, \boldsymbol{x}_{k+1}^{\mathrm{real}}) \in D} \| \left(\boldsymbol{x}_{k+1}^{\mathrm{real}} - f_{\mathrm{red},\boldsymbol{\mu}^*}^{\mathrm{PE}}(\boldsymbol{x}_k^{\mathrm{real}}, \boldsymbol{u}_k^{\mathrm{real}}) \right) - f^{\mathrm{GP}}(\boldsymbol{x}_k^{\mathrm{real}}, \boldsymbol{u}_k^{\mathrm{real}}) \|^2. \tag{4}$$

Note that after collecting the trajectories in the real system, we collect the simulator estimates of the next state $\boldsymbol{x}_{k+1}^{\mathrm{sim}}$ using the physics engine with the estimated physical parameters $\boldsymbol{\mu}^*$. This is done by resetting the state of the simulator to every state $\boldsymbol{x}_k^{\mathrm{real}}$ along the collected trajectory and applying the action $\boldsymbol{u}_k^{\mathrm{real}}$ to obtain the resulted next state $\boldsymbol{x}_{k+1}^{\mathrm{sim}} = f_{\mathrm{red},\boldsymbol{\mu}^*}^{\mathrm{PE}}(\boldsymbol{x}_k^{\mathrm{real}}, \boldsymbol{u}_k^{\mathrm{real}})$, and store the tuple $\{\boldsymbol{x}_k^{\mathrm{real}}, \boldsymbol{u}_k^{\mathrm{real}}, \boldsymbol{x}_{k+1}^{\mathrm{sim}}\}$. Thus, the GPs learn the input-output relationship: $f^{\mathrm{GP}}(\boldsymbol{x}_k^{\mathrm{real}}, \boldsymbol{u}_k^{\mathrm{real}}) = \boldsymbol{x}_{k+1}^{\mathrm{real}} - \boldsymbol{x}_{k+1}^{\mathrm{sim}}$. Two independent GP models are trained, one each for the position and velocity of the marble. We found GP models ideal for this system because of their accuracy in data prediction and data efficiency which is fundamental when working with real systems. However, other machine learning models could be adopted in different applications.

3) Modeling Motor Behavior: The tip-tilt platform in the CMS is actuated by hobby-grade servo motors which work in position control mode. These motors use a controller with a finite settling time which is longer than the control interval used in our experiments. This results in actuation delays for the action computed by any control algorithm, and the platform always has non-zero velocity. The physics engine, on the other hand, works in discrete time and thus the CME comes to a complete rest after completing a given action in a control interval. Consequently, there is a discrepancy between the simulation and the real system in the sense that the real system gets delayed actions. Such actuation delays are common in most (robotic) control systems and thus, needs to be considered during controller design for any application. To compensate for this problem, we learn an inverse model for motor actuation. This inverse model of the motor predicts the action to be sent to the motors for the tip-tilt platform to achieve a desired state $(\beta_{k+1}^{\mathrm{des}}, \gamma_{k+1}^{\mathrm{des}})$ given the current state (β_k, γ_k) at instant k. Thus, the control signals computed by the optimization process are passed through this function that generates the commands (u_x, u_y) for the servo motors. We represent this inverse motor model by f_{imm}. The motor model f_{imm} is learned using a standard autoregressive model with external input. This is learned by collecting motor response data by exciting the CMS using sinusoidal inputs for the motors before the model learning procedure in Algorithm 1.

C. Trajectory Optimization Using iLQR

We use the iterative LQR (iLQR) as the optimization algorithm for model-based control [29]. While there exist optimization solvers which can generate better optimal solutions for model-based control [30], we use iLQR as it provides a compute-efficient way of solving the optimization problem for designing the controller. Formally, we solve the following *trajectory optimization problem* to manipulate the controls $\boldsymbol{u}_k$ over a certain number of time steps $[T-1]$

$$\begin{aligned} \min_{\boldsymbol{x}_k, \boldsymbol{u}_k} & \sum_{k \in [T]} \ell(\boldsymbol{x}_k, \boldsymbol{u}_k) \\ \text{s.t. } & \boldsymbol{x}_{k+1} = f(\boldsymbol{x}_k, \boldsymbol{u}_k) \\ & \boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0. \end{aligned} \tag{5}$$

For the state cost, we use a quadratic cost function for the state error measured from the target state $\boldsymbol{x}_{\mathrm{target}}$ (which in the current case is the nearest gate for the marble), as represented by the following equation:

$$\ell(\boldsymbol{x}) = ||\boldsymbol{x} - \boldsymbol{x}_{\mathrm{target}}||_W^2, \tag{6}$$

where the matrix W represents weights used for different states. For the control cost, we penalize the control using a quadratic cost as well, given by the following equation:

$$\ell(\boldsymbol{u}) = \lambda_{\boldsymbol{u}} \|\boldsymbol{u}\|^2. \tag{7}$$

Other smoother versions of the cost function [29] did not change the behavior of the iLQR optimization. The discrete-time dynamics $\boldsymbol{x}_{k+1} = f(\boldsymbol{x}_k, \boldsymbol{u}_k)$ and the cost function are used to compute locally 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 LQR problem. For more details of iLQR, interested readers are referred to [29]. The solution to the trajectory optimization problem returns an optimal sequence of states and control inputs for the system to follow. We call this the reference trajectory for the system, denoted by $X^{\mathrm{ref}} \equiv \boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, and $U^{\mathrm{ref}} \equiv \boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{T-1}$. The matrix W used for the experiments is diagonal, $W = \mathrm{diag}(4, 4, 1, 0.4)$ and $\lambda_{\boldsymbol{u}} = 20$. These weights were tuned empirically only once at the beginning of learning.

D. Online Control Using Nonlinear Model-Predictive Control

While it is easy to control the movement of the marble in the simulation environment, controlling the movement of the marble in the real system is much more challenging. This is mainly due to complications such as static friction (which remains poorly modeled by the physics engine), or delays in actuation. As a result, the real system requires online model-based feedback control. While re-computing an entire new trajectory upon a new observation would be the optimal strategy, due to lack of computation time in the real system, we use a trajectory-tracking MPC controller. We use an iLQR-based NMPC controller to track the trajectory obtained from the trajectory optimization module to control the system in real-time. The controller uses the least-squares tracking cost function given by the following equation:

$$\ell_{\mathrm{tracking}}(\boldsymbol{x}) = \|\boldsymbol{x}_k - \boldsymbol{x}_k^{\mathrm{ref}}\|_Q^2, \tag{8}$$

where $\boldsymbol{x}_k$ is the system state at instant k, $\boldsymbol{x}_k^{\mathrm{ref}}$ is the reference state at instant k, and the matrix Q is a weight matrix. The matrix Q and the cost coefficient for control are kept the same

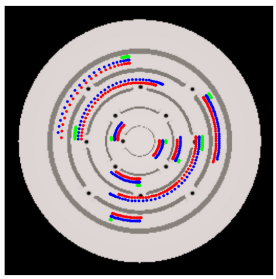

Fig. 3. Comparison of real trajectories (red), predicted trajectories (blue) using the estimated physical properties using CMA-ES, and trajectories using the default physical properties (green) in the sim-to-sim experiment. The trajectories are generated with a random policy from random initial points.

as during trajectory optimization. The system trajectory is rolled out forward in time from the observed state, and the objective in Eq. 8 is minimized to obtain the desired control signals.

We implement the control on both the real and the simulation environment at a control rate of 30 Hz. As a result, there is not enough time for the optimizer to converge to the optimal feedback solution. Thus, we warm-start the optimizer with a previously computed trajectory. Furthermore, the derivatives during the system linearization in the backward step of iLQR and the forward rollout of the iLQR are obtained using parallel computing in order to satisfy the time constraints to compute the feedback step.

V. Experiments

In this section we test how our proposed approach performs on the CMS, and how it compares to human performance.

A. Physical Property Estimation Using CMA-ES

We first demonstrate how physical parameter estimation works in two different environments; sim-to-sim and sim-to-real settings. For sim-to-sim setting, we regard the full model $f_{\text{full}}^{\text{PE}}$ as a real system because it contains full internal state that is difficult to observe in the real setup as described in Section IV-A. Also, we regard the reduced model $f_{\text{red}}^{\text{PE}}$, which has the same state that can be observed in the real system, as a simulator. For $f_{\text{red}}^{\text{PE}}$, we start with default values given by MuJoCo, and we set smaller friction parameters to $f_{\text{full}}^{\text{PE}}$ in the sim-to-sim setting, because we found the real maze board is much more slippery than what default MuJoCo's parameters would imply. For sim-to-real setting, we measure the difference between the real system and the reduced model $f_{\text{red}}^{\text{PE}}$.

To verify the performance of physical parameter estimation, we collected samples using the NMPC controller computed using current $f_{\text{red}}^{\text{PE}}$ models on both settings, which corresponds to line 1-3 of Algorithm 1, and found the objective defined in (3) converges only ~ 10 transitions for each ring. For sim-to-sim experiment, the RMSE of ball location θ in two dynamics becomes $\approx 2e-3$ [rad] (≈ 0.1 [deg]), which we conclude the CMA-ES produces accurate enough parameters. Figure 3 shows the real trajectories obtained by $f_{\text{full}}^{\text{PE}}$ (in red), simulated trajectories obtained by $f_{\text{red}}^{\text{PE}}$ with optimized friction parameters (in blue), and simulated trajectories before estimating friction parameters (in green). This qualitatively shows that the estimated friction parameters successfully bridge the gap between two different dynamics. Since tuning friction parameters for MuJoCo is not intuitive, it is evident that we can rely on CMA-ES to determine more optimal friction parameters instead. Similarly, we find that sim-to-real experiment, the RMSE of ball position θ between the physics engine and real system decreased to $\approx 9e-3$ [rad] after CMA-ES optimization. However, we believe this error still diverges in rollout and we still suffer from static friction. We also observed that CMA-ES optimization in the sim-to-real experiments quickly finds a local minima with very few samples, and further warm starting the optimization with more data results in another set of parameters for the physics engine with similar discrepancy between the physics engine and the real system. Thus, we perform the CMA-ES parameter estimation only once in the beginning and more finetuning to GP regression.

B. Control Performance on Real System

We found the *sim-to-sim* agent learns to perform well with just CMA-ES finetuning, and thus we skip further control results for the *sim-to-sim* agent, and only present results on the real system with additional residual learning for improved performance. While CMA-ES works well in the *sim-to-sim* transfer problem, if we want a robot to solve the CME, there will necessarily be differences between the internal model and real-world dynamics. We take inspiration from how people understand dynamics – they can both capture physical properties of items in the world, and also learn the dynamics of arbitrary objects and scenes. For this reason we augmented the CMA-ES model with machine learning data-driven models that can improve the model accuracy as more experience (data) is acquired. We opted for GP as data-driven models because of their high flexibility in describing data distribution and data efficiency [31].

The CMA-ES model is then iteratively improved with the GP residual model with data from 5 rollouts in each iteration. In the following text, 'CMA-ES' represents the CMA-ES model without any residual modeling, while 'CMA-ES + GP1' represents a model that has learned a residual model from 5 rollouts of the 'CMA-ES' model. Similarly, 'CMA-ES + GP2' and 'CMA-ES + GP3' learn the residual distribution from 10 experiments (5 with 'CMA-ES' and 5 with 'CMA-ES + GP1') and 15 experiments (5 each from 'CMA-ES', 'CMA-ES + GP1' and 'CMA-ES + GP2'), respectively. The trajectory optimization and tracking uses the mean prediction from the GP models.

Figure 4 shows the time spent in each ring averaged over 10 different rollouts at each iteration during training. As expected, models trained with a larger amount of data consistently improve the performance, i.e., spending less time in each ring. The improvement in performance can be seen especially in the outermost (Ring1; $F(3, 36) = 3.02$, $p = 0.042$) and innermost ring (Ring4; $F(3, 36) = 4.52$, $p = 0.009$).[2] The outermost ring has the largest radius and is more prone to oscillations, which the

[2]Due to extreme heteroscedasticity in the data we use White's corrected estimators in the ANOVA [32].

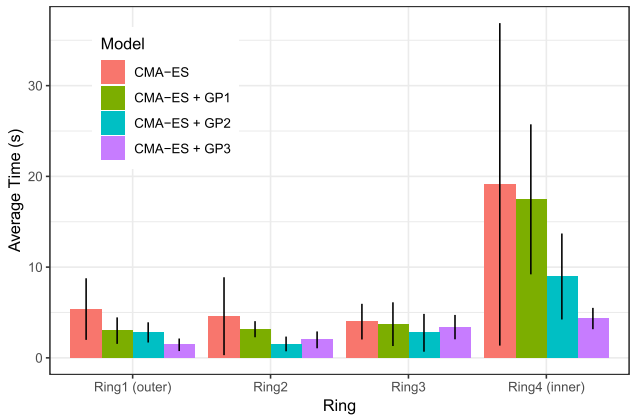

Fig. 4. Comparison of average time spent by the marble in each ring during learning and the corresponding standard deviation over 10 trials. This plot shows the improvement in the performance of the controller upon learning of the residual model. Note that the controller completely fails without CMA-ES initialization, and thus, those results are not included.

model learns to control. Similarly, in the innermost ring, static friction causes small actions to have larger effects.

C. Comparison With Human Performance

To compare our system's performance against human learning, we asked 15 participants to perform a similar CME task. These participants were other members of the Mitsubishi Electric Research Laboratories who were not involved in this project and were naive to the intent of the experiment. The particpants were instructed to solve the CME five consecutive times. A 2 DoF joystick was provided to control the two servo motors of the same experimental setup on which the learning algorithm was trained. To familiarize participants with the joystick control, they were given one minute to interact with the maze—without marble. Because people can adapt to even unnatural joystick mappings within minutes [33], we assumed that this familiarization would provide a reasonable control mapping for our participants, similar to how the model pre-learned the inverse motor model f_{imm} without learning ball dynamics. Since we found no reliable evidence of improvement throughout the trials (see below), we believe that any further motor control learning beyond this period was at most marginal. Three participants had prior experience solving the CME in the "convential" way by holding it with both hands.

Afterwards, the ball was placed at a random point in the outermost ring, and participants were asked to guide the ball to the center of the maze. They were asked to solve the CME five times, and we recorded how long they took for each solution and how much time the ball spent in each ring. Two participants were excluded from analysis because they could not solve the maze five times within the 15 minutes allotted to them.

Because people were given five maze attempts (and thus between zero and four prior chances to learn during each attempt), we compare human performance against the CMA-ES and CMA-ES+GP1 versions of our model that have comparable amounts of training.

We find that while there was a slight numerical decrease in participants' solution times over the course of the five trials, this did not reach statistical reliability ($\chi^2(1) = 1.63$, $p = 0.2$): participants spent an average of 110 seconds ($95\%\text{CI} : [66, 153]$) to solve the maze the first time, and 79 seconds ($95\%\text{CI} : [38, 120]$) to solve the maze the last time, and only 8 of 13 participants solved the maze faster on the last trial as compared to their first. This is similar to the learning pattern found in our model, where the solution time decreased from 33s using CMA-ES to 27s using CMA-ES+GP1, which was also not statistically reliable ($t(15) = 0.56$, $p = 0.58$).

TABLE I
AVERAGE TIME SPENT IN EACH RING [SEC]

	Human	CMA-ES + GP0/1
Ring 1 (outermost ring)	22.6	4.18
Ring 2	8.0	3.87
Ring 3	24.3	3.85
Ring 4 (innermost ring)	41.1	18.29

In addition, Table I shows the time that people and the model kept the ball in each ring. For statistical power we have averaged over all human attempts, and across CMA-ES and CMA-ES + GP1 to equate to human learning. In debriefing interviews, participants indicated that they found that solving the innermost ring was the most difficult, as indicated by spending more time in that ring than any others (all $ps < 0.05$ by Tukey HSD pairwise comparisons). This is likely because small movements will have the largest effect on the marble's radial position, requiring precise prediction and control. Similar to people, the model also spends the most time in the inner ring (all $ps < 0.002$ by Tukey HSD pairwise comparisons), suggesting that it shares similar prediction and control challenges to people. In contrast, a fully trained standard reinforcement learning algorithm – the soft actor-critic (SAC) [34] – learns a different type of control policy in simulation and spends the *least* amount of time in the innermost ring, since the marble has the shortest distance to travel (see Supplemental Materials for more detail).

VI. CONCLUSIONS AND FUTURE WORK

We take inspiration from cognitive science to build an agent that can plan its actions using an augmented simulator in order to learn to control its environment in a sample-efficient manner. We presented a learning method for navigating a marble in a complex circular maze environment. Learning consists of initializing a physics engine, where the physics parameters are initially estimated using the real system. The error in the physics engine is then compensated using a Gaussian process regression model which is used to model the residual dynamics. These models are used to control the marble in the maze environment using iLQR in a feedback MPC fashion. We showed that the proposed method can learn to solve the task of driving the marble to the center of the maze within a few minutes of interacting with the system, in contrast to traditional reinforcement systems that are data-hungry in simulation and cannot learn a good policy on a real robot.

To implement our approach on the CMS, we made some simplifications that are only applicable to the CMS, e.g., that the problem can be segmented into moving through the gates in the

rings. While this does limit the generality of the specific model used, most physical systems require some degree of domain knowledge to design an efficient and reliable control system. Nonetheless, we believe our approach is a step towards learning general-purpose, data-efficient controllers for complex robotic systems. One of the benefits of our approach is its flexibility: because it learns based off of a general-purpose physics engine, this approach should generalize well to other real-time physical control tasks. Furthermore, the separation of the dynamics and control policy should facilitate transfer learning. If the maze material or ball were changed (e.g., replacing it with a small die or coin), then the physical properties and residual model would need to be quickly relearned, but the control policy should be relatively similar. In future work, we plan to test the generality and transfer of this approach to different mazes and marbles. For more effective use of physics engines for these kind of problems, we would like to interface general-purpose robotics optimization software [35] to make it more useful for general-purpose robotics application.

REFERENCES

[1] K. Fang *et al.*, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *Int. J. Robot. Res.*, vol. 39, no. 2–3, pp. 202–216, 2020.

[2] M. Toussaint, K. R. Allen, K. A. Smith, and J. B. Tenenbaum, "Differentiable physics and stable modes for tool-use and manipulation planning," in *Proc. Robot.: Sci. Syst. XIV. Robot.: Sci. Syst. Found.*, 2018.

[3] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2016.

[4] A. Sanchez-Gonzalez *et al.*, "Graph networks as learnable physics engines for inference and control," in *Proc. 35th Int. Mach. Learn. Res., proc. Mach. Learn. Res.*, vol. 80, pp. 4470–4479, 2018. [Online]. Available: http://proceedings.mlr.press/v80/sanchez-gonzalez18a.html

[5] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[6] F. Osiurak and D. Heinke, "Looking for intoolligence: A unified framework for the cognitive study of human tool use and technology," *Amer. Psychol.*, vol. 73, no. 2, pp. 169–185, 2018.

[7] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proc. Nat. Acad. Sci.*, vol. 110, no. 45, pp. 18 327–18 332, 2013.

[8] K. A. Smith, P. W. Battaglia, and E. Vul, "Different physical intuitions exist between tasks, not domains," *Comput. Brain Behav.*, vol. 1, no. 2, pp. 101–118, 2018.

[9] K. R. Allen, K. A. Smith, and J. B. Tenenbaum, "Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning," *Proc. Nat. Acad. of Sci.*, vol. 117, no. 47, pp. 29 302–29 310, 2020.

[10] J. v. Baar, A. Sullivan, R. Corcodel, D. Jha, D. Romeres, and D. Nikovski, "Sim-to-real transfer learning using robustified controllers in robotic tasks involving complex dynamics," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2019, pp. 6001–6007.

[11] D. Romeres, D. K. Jha, A. DallaLibera, B. Yerazunis, and D. Nikovski, "Semiparametrical gaussian processes learning of forward dynamical models for navigating in a circular maze," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2019, pp. 3195–3202.

[12] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 1889–1897.

[14] S. James, A. J. Davison, and E. Johns, "Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task," in *Proc. 1st Annu. Conf. Robot Learn. PMLR*, vol. 78, pp. 334–343, 2017.

[15] F. Golemo, A. A. Taiga, A. Courville, and P.-Y. Oudeyer, "Sim-to-real transfer with neural-augmented robot simulation," in *Proc. Conf. Robot Learn.*, 2018, pp. 817–828.

[16] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 1–8.

[17] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 5026–5033.

[18] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 23–30.

[19] F. Ramos, R. Possas, and D. Fox, "Bayessim: Adaptive domain randomization via probabilistic inference for robotics simulators," in *Robotics: Sci. Syst.*, 2019. [Online]. Available: https://doi.org/10.15607/RSS.2019.XV.029

[20] L. Hewing, J. Kabzan, and M. N. Zeilinger, "Cautious model predictive control using gaussian process regression," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 6, pp. 2736–2743, Nov. 2020.

[21] M. Saveriano, Y. Yin, P. Falco, and D. Lee, "Data-efficient control policy search using residual dynamics learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 4709–4715.

[22] A. Ajay *et al.*, "Augmenting physical simulators with stochastic neural networks: Case study of planar pushing and bouncing," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 3066–3073.

[23] T. Wu and J. Movellan, "Semi-parametric gaussian process for robot system identification," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 725–731.

[24] D. Romeres, M. Zorzi, R. Camoriano, and A. Chiuso, "Online semi-parametric learning for inverse dynamics modeling," in *Proc. IEEE 55th Conf. Decis. Control*, 2016, pp. 2945–2950.

[25] D. Nguyen-Tuong and J. Peters, "Using model knowledge for learning inverse dynamics," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2010, pp. 2677–2682.

[26] A. Ajay *et al.*, "Combining physical simulators and object-based networks for control," in *Proc. Int. Conf. Robotics Automat. (ICRA)*, Montreal, QC, Canada, 2019, pp. 3217–3223, doi: 10.1109/ICRA.2019.8794358.

[27] M. Diehl, H. J. Ferreau, and N. Haverbeke, "Efficient numerical methods for nonlinear mpc and moving horizon estimation," *Nonlinear Model Predictive Control*. Berlin, Germany: Springer, 2009, pp. 391–417.

[28] N. Hansen, "The cma evolution strategy: a comparing review," *Towards A New Evolutionary Computation.* Berlin, Germany: Springer, 2006.

[29] Y. Tassa, T. Erez, and E. Todorov, "Synthesis and stabilization of complex behaviors through online trajectory optimization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 4906–4913.

[30] J. T. Betts, "Survey of numerical methods for trajectory optimization," *J. Guid., Control, Dyn.*, vol. 21, no. 2, pp. 193–207, 1998.

[31] A. Dalla Libera, D. Romeres, D. K. Jha, B. Yerazunis, and D. Nikovski, "Model-based reinforcement learning for physical systems without velocity and acceleration measurements," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 3548–3555, Apr. 2020.

[32] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica: J. Econometric Soc.*, vol. 48, pp. 817–838, 1980.

[33] O. Bock, S. Schneider, and J. Bloomberg, "Conditions for interference versus facilitation during sequential sensorimotor adaptation," *Exp. Brain Res.*, vol. 138, no. 3, pp. 359–365, 2001.

[34] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870. [Online]. Available: http://proceedings.mlr.press/v80/haarnoja18b.html

[35] R. Tedrake and the Drake Development Team, "Drake: Model-based design and verification for robotics," Accessed: Sep. 1, 2020. [Online]. Available: https://drake.mit.edu