2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Active Learning of Dynamics Using Prior Domain Knowledge in the Sampling Process

Kevin S. Miller, *Member, IEEE,* Adam J. Thorpe, *Member, IEEE,* Ufuk Topcu, *Fellow, IEEE*

Abstract— **We present an active learning algorithm for learning dynamics that leverages side information by explicitly incorporating prior domain knowledge into the sampling process. Our proposed algorithm guides the exploration toward regions that demonstrate high empirical discrepancy between the observed data and an imperfect prior model of the dynamics derived from side information. Through numerical experiments, we demonstrate that this strategy explores regions of high discrepancy and accelerates learning while simultaneously reducing model uncertainty. We rigorously prove that our active learning algorithm yields a consistent estimate of the underlying dynamics by providing an explicit rate of convergence for the maximum predictive variance. We demonstrate the efficacy of our approach on an under-actuated pendulum system and on the half-cheetah MuJoCo environment.**

I. Introduction

Model-based and data-driven methods typically represent two alternative approaches to stochastic optimal control. While purely data-driven control typically neglects prior domain knowledge (side information) to reduce bias, incorporating such knowledge into data-driven control can yield more accurate learned models of dynamical systems. For example, prior domain knowledge in the form of an imperfect physics-based model can be combined with data-driven modeling to more accurately approximate the true system dynamics. Despite the focus on side information in the learned model, there are foreseeable benefits to also leveraging side information to select informative data, and the question of how to select data according to available side information remains an open challenge.

We present an active dynamics learning method that utilizes side information to select sample data in regions where the observed discrepancy between prior domain knowledge and the observed data is highest. Our approach is based on the upper confidence bound (UCB) algorithm [1]–[3], known for its ability to balance exploration and exploitation in multi-armed bandit and Bayesian optimization settings. We specifically consider the Gaussian process (GP) setting [4], known as GP-UCB [5], which offers a principled approach to characterizing uncertainty—a key component utilized by the UCB algorithm. Our key innovation lies in incorporating side information *during the sampling phase*. We actively sample control inputs along trajectories that favor exploration in regions that demonstrate a higher discrepancy between our observed data and an imperfect prior model of the dynamics. This emphasizes sampling in regions of the state-action space where prior knowledge does not align with the data-driven estimate while avoiding redundant sampling in regions where our data-driven model aligns with the observed dynamics.

Our approach is enabled by two key elements: 1) actively learning dynamics in an episodic setting, and 2) incorporating side information.

Active sampling has been explored in the context of learning dynamical systems [6]–[9]. In the case of linear dynamics, active learning approaches can provide optimal or near-optimal sample complexity results [6]–[8]. However, these results often impose unrealistic assumptions–such as the ability to sample any state-action pair without regards to dynamic constraints. In the general case, sample complexity guarantees are more elusive. Optimistic planning approaches, such as OpAx [10] and H-UCRL [11] actively sample data in an episodic setting by selecting policies in an open-loop fashion to maximize information gain via an optimistic planner. We likewise consider an episodic setting, but we consider using prior domain knowledge to determine our exploration policy. Notably, GP-UCB has been used in a different context for robot kinematic calibration [9], in which the authors use active sampling to learn the correction to a prior, parametric model of the robot's kinematics. While they allow for arbitrary sampling of states and control inputs throughout a single sampling process, we restrict to the selection of control input sequences along trajectories in an episodic fashion. That is, we must account for planning action sequences under uncertain dynamics in an MPC-like framework.

Active sampling methods typically do not incorporate prior domain knowledge into the model or the sampling process. In particular, except for [9], the previously mentioned active learning methods (i.e., [6]–[8, 10]) do not utilize side information in the data-driven model of the system dynamics. The use of side information has been investigated for two-armed bandit problems in [12, 13], where the sampler has access to information of the reward. However, these results do not

This material is based upon work supported by the National Science Foundation under NSF Grant Number 2214939. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This material is based upon work supported by the Air Force Office of Scientific Research under award number AFOSR FA9550-22-1-0403 and the Air Force Research Laboratory under award number AFRL FA9550-19-1-0005. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Defense. K. S. Miller was supported by a Peter J. O'Donnell Jr. Postdoctoral Fellowship and NSF IFML Grant Number 2019844.

K. Miller, A. Thorpe, and U. Topcu are with the Oden Institute for Computational Engineering and Sciences at the University of Texas at Austin, Austin, TX, USA.

Email: {ksmiller,utopcu}@utexas.edu, adam.thorpe@austin.utexas.edu.

K. S. Miller and A. J. Thorpe contributed equally to this work.

979-8-3503-1632-2/24/$31.00 ©2024 IEEE

translate easily to the setting of dynamics learning. To our knowledge, our proposed active dynamics learning method is the first to leverage side information in the sampling process.

Our main contribution is an active dynamics learning algorithm based on GP-UCB that incorporates prior domain knowledge into both the sampling process and the learned model. Specifically, our approach incorporates a discrepancy term in the UCB sampling function that empirically models the difference between the data-driven portion of our model and the prior domain knowledge. Unlike existing approaches, we exploit prior domain knowledge via this discrepancy term to focus sampling in regions where the prior model is most misaligned with the true dynamics. In addition, we provide a proof that our approach yields a consistent estimator of the dynamics as more episodes are considered. It is often difficult to guarantee consistency for active sampling algorithms due to the non-i.i.d. nature of the sampling. We prove that as long as the trajectory generated during each episode explores state-action pairs that possess greater than average predictive variance under the GP model, our estimate converges to the true dynamics as the number of episodes increases. We demonstrate our approach on a simple pendulum system, and compare our approach against OpAx [10] and greedy variance-based sampling, which represents a pure exploration strategy. We then demonstrate that we can use our learned dynamics model for control on a high-dimensional half-cheetah MuJoCo environment.

The rest of the paper is outlined as follows. In Section II, we formally state the problem setting and provide preliminary background information to inform the discussion and introduction of our sampling method. We present our active sampling method in Section III and provide the consistency argument in Subsection III-A. In Section IV, we demonstrate our active sampling approach.

II. Preliminaries & Problem Statement

A. Problem Statement

Consider the following discrete-time dynamical system,

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

where $x_t \in \mathcal{X}$ is the state of the system at time t, $u_t \in \mathcal{U}$ is the control action applied to the system at time t, and $w_t \sim \mathcal{N}(0, \sigma^2 I)$ is an independent Gaussian noise term.

Given a cost function $c : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, we formulate the following stochastic optimal control problem, where the goal is to select a sequence of control inputs $u_0, \ldots, u_N$ to minimize the expected cumulative cost,

$$\min_{u_0,\ldots,u_N} \quad \mathbb{E}\left[\sum_{t=1}^{N} c(x_t, u_t)\right] \tag{2a}$$

$$\text{s.t.} \quad x_{t+1} = f(x_t, u_t) + w_t \tag{2b}$$

We presume that the system dynamics f in (1) are unknown, meaning the control problem (2) is intractable. However, we presume access to an *imperfect* model of the dynamics $p_0 : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$. Such side information may be available, for instance, if we have a coarse approximation of system parameters, a first-order estimate of the dynamics, or access to a low-fidelity model through a virtual simulation.

We consider the problem of sequentially computing a data-driven estimate μ_n of the system dynamics (1) by actively selecting the dataset $\mathcal{D}_n$ consisting of $n \in \mathbb{N}$ observed state transitions,

$$\mathcal{D}_n = \{(x^i, u^i, y^i)\}_{i=1}^{n}, \tag{3}$$

where x and u are in $\mathcal{X}$ and $\mathcal{U}$, respectively, and $y = f(x, u) + w$ with $w \sim \mathcal{N}(0, \sigma^2 I)$.

We study an episodic setting, where in each episode $\tau = 1, 2, \ldots, T$, we compute an exploratory policy $\pi_\tau : \mathcal{X} \to \mathcal{U}$ that we employ to collect new data. As we obtain new observations, we update the dataset, $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{(x', u', y')\}$ and subsequently update the data-driven estimate of the dynamics.

To that end, our goal is to develop an active sampling algorithm that defines an exploratory policy π_τ in each episode $\tau = 1, 2, \ldots$ to leverage (i) prior domain knowledge and (ii) information from prior episodes to efficiently learn a predictive model of the dynamics $f(x, u)$ in (1). Furthermore, we seek a *consistent* estimator, meaning $\mu_\tau(x, u) \to f(x, u)$ as $\tau \to \infty$.

Our key insight is that a non-zero mean GP can be used to model the residual between our observed data and a prior model derived from side information. Then, we modify the UCB algorithm to maximize the discrepancy between our data-driven model and the prior, and periodically accumulate observed data into our prior model at the end of each episode. This leads the active sampling process to sample more in areas where the prior model is empirically incorrect, and less in areas where our prior is closely aligned with the data.

B. Gaussian Processes

We use a Gaussian process [4] to estimate the dynamics. For notational simplicity, we denote $\mathcal{Z} = \mathcal{X} \times \mathcal{U}$ as the state-action space and $z_t = (x_t, u_t)$. A Gaussian process $f(z) \sim \mathcal{GP}(m(z), k(z, z'))$ is completely specified by a mean function $m(z)$ and a positive definite covariance function $k(z, z')$ [4],

$$m(z) = \mathbb{E}[f(z)] \tag{4}$$

$$k(z, z') = \mathbb{E}[(f(z) - m(z))(f(z') - m(z'))] \tag{5}$$

In the following, we presume that the covariance function k is given by the squared exponential, or RBF function, $k(z, z') = \exp(-\gamma \|z - z'\|^2)$, where $\gamma > 0$, and that the mean function $m(z)$ is non-zero [4, § 2.7]. The non-zero mean is key to our approach since we use this term to capture prior knowledge of the system dynamics. Critically, we will periodically update the prior to reflect the new data gathered during each episode.

Given a dataset $\mathcal{D}_n$ as in (3) consisting of n data points, the predictive mean μ_n and variance σ_n^2 can be evaluated at a given test point $z^* \in \mathcal{Z}$ as,

$$\mu_n(z^*) = (\boldsymbol{y} - \boldsymbol{m})^\top (G_n + \sigma^2 I)^{-1} \boldsymbol{k}_{z^*} + m(z^*) \tag{6}$$

$$\sigma_n^2(z^*) = k(z^*, z^*) - \boldsymbol{k}_{z^*}^\top (G_n + \sigma^2 I)^{-1} \boldsymbol{k}_{z^*} \tag{7}$$

where $\boldsymbol{y}$ and $\boldsymbol{m}$ are vectors, with the i^{th} elements given by $\boldsymbol{y}_i = y^i$ and $\boldsymbol{m}_i = m(x^i, u^i)$, $G_n = (g_{ij})$ is an $n \times n$ matrix with elements $g_{ij} = k(z^i, z^j)$, $\boldsymbol{k}_{z^*}$ is a vector where the i^{th} element is given by $k(z^i, z^*)$. As we collect new observations and augment the dataset $\mathcal{D}_n$, the predictive mean μ_n and variance σ_n^2 are recomputed using the new dataset. In practice, these equations can be solved efficiently using Cholesky factorization, see [4, Algorithm 2.1], and the Cholesky factors can be updated via rank-1 updates as new data becomes available.

C. Active Sampling in the GP Setting

We are concerned with fitting a GP model to a minimal dataset, i.e. sampling a limited number of data points that provide significant information about the true dynamics of the system, f. We utilize an adaptation of the GP-UCB algorithm [5] to guide the selection of actions along trajectories to regions of greatest mismatch (discrepancy) between the prior model and the true, underlying dynamics. The GP-UCB algorithm [5] for maximizing a function $g(z)$ for $z \in \mathcal{Z}$ uses the predictive mean μ_n and variance σ_n of a GP to decide the next sample point, and chooses points based on the following *acquisition function*,

$$A(z) = \mu_n(z) + \beta_n^{1/2}\sigma_n(z), \tag{8}$$

where μ_n is as in (6), σ_n is the square root of the predictive variance in (7), and $\beta \in \mathbb{R}_+$ is a positive real constant called the decay schedule.

The next sample point z_{n+1} is then chosen to maximize the acquisition function as

$$z_{n+1} = \arg\max_{z \in \mathcal{Z}} A_n(z). \tag{9}$$

Intuitively, the acquisition function A in (8) provides a tradeoff between sampling in areas where the predictive mean is large, and areas of high variance. This means that in practice, under correct choice of the decay schedule β_n, the UCB algorithm will trade off between "exploration" in areas of high uncertainty and "exploitation" by selecting points near the max of the predictive mean.

III. Active Sampling Using Side Information

Our key insight is to define an exploration policy $\pi_\tau : \mathcal{X} \to \mathcal{U}$ in episode τ using (9) to focus sampling on N state-action pairs, $\{(x^{\tau,n}, u^{\tau,n})\}_{n=1}^N$, in regions where the previous episode's learned model, $\mu_{\tau-1,N}$, is maximally different from the true system dynamics, f. As described in Section II-B, we iteratively update our GP model as

$$\mu_{\tau,n}(z) = (\boldsymbol{y} - \boldsymbol{m}_\tau)^\top (G_{\tau,n} + \sigma^2 I)^{-1} \boldsymbol{k}_z + m_\tau(z) \tag{10}$$

$$\sigma_{\tau,n}^2(z) = k(z,z) - \boldsymbol{k}_z^\top (G_{\tau,n} + \sigma^2 I)^{-1} \boldsymbol{k}_z, \tag{11}$$

where $G_{\tau,n} \in \mathbb{R}^{(\tau-1)N+n \times (\tau-1)N+n}$ is the kernel Gram matrix of all data points observed up to iteration n of episode τ, and $m_\tau : \mathcal{Z} \to \mathcal{X}$ is an *episode-dependent* prior term that we define to be

$$m_\tau(z) := \mu_{\tau-1,N}(z), \tag{12}$$

which is the previous episode's learned model. At $\tau = 1$, we define $m_1(z) = p_0(z)$, which is determined by the side information or prior domain knowledge available before sampling. Furthermore, we define $\mu_{\tau,0} \equiv \mu_{\tau-1,N}$ and $\sigma_{\tau,0} \equiv \sigma_{\tau-1,N}$ to accumulate observed data from the previous episode.

The main idea of the sampling procedure is that the exploration policy π_τ is computed to maximize an adapted GP-UCB acquisition function over action sequences along the trajectory predicted by the current model of the dynamics. We consider an MPC-like procedure to iteratively update and replan the action sequences at each iteration during the episode. For simplicity of notation, we consider the time horizon to be the same as the episode length, N.

We define the following *discrepancy-based* acquisition function,

$$\begin{aligned} A_{\tau,n}(u; x^{\tau,n}) = \big|\mu_{\tau,n}(x^{\tau,n}, u) - m_\tau(x^{\tau,n}, u)\big| \\ + \beta_{\tau,n}^{1/2}\sigma_{\tau,n}(x^{\tau,n}, u) + s_\tau(x^{\tau,n}, u). \end{aligned} \tag{13}$$

The first term of (13) models the *discrepancy* between our current model $\mu_{\tau,n}$ and the prior m_τ fixed at the beginning of the current episode. Maximizing this term encourages sampling actions where the current episode's prior appears to be most incorrect. As in (8), the middle term of (13) allows for the tradeoff between exploration of actions that possess high uncertainty (variance) and exploitation of actions that maximize the discrepancy. The last term s_τ of (13) is an additional term to allow for further side information to be accounted for in the exploration policy. For example, $s_\tau(z)$ could be defined to ensure the system avoids exploring states or actions known to be unsafe, e.g. as,

$$s_\tau(z) = \begin{cases} 0 & z \in \mathcal{S} \\ -\infty & z \notin \mathcal{S} \end{cases}, \tag{14}$$

where $\mathcal{S} \subset \mathcal{Z}$ is an episode-dependent safe set.

At each iteration n during an episode τ, we solve the following optimization problem,

$$\max_{u_0,\ldots,u_{N-1} \subset \mathcal{U}} \quad \sum_{t=0}^{N-1} A_{\tau,n}(u_t; x_t) \tag{15a}$$

$$\text{s.t.} \quad x_{t+1} = \mu_{\tau,n}(x_t, u_t) \tag{15b}$$

$$x_0 = x^{\tau,n} \tag{15c}$$

Intuitively, we plan a sequence of actions $u_0, \ldots, u_{N-1}$ from the current state $x^{\tau,n}$ that, given the current estimate of the dynamics $\mu_{\tau,n}$, will maximize the sum of the discrepancy-based acquisition function in (13) along the predicted trajectory. After solving (15), we execute the first control action in the planned sequence, accumulate a new data point in our dataset, and update the dynamics estimate $\mu_{\tau,n}$.

At the first iteration $n = 0$ during each episode τ, we set $x^{\tau,0}$ to be the state in $\mathcal{X}$ with the maximum predictive variance $\sigma_{\tau,n}^2$. Note that by definition of $m_\tau \equiv \mu_{\tau-1,N} \equiv \mu_{\tau,0}$, the discrepancy term in the first iteration is $\big|\mu_{\tau,0}(x^{\tau,n}, u) - m_\tau(x^{\tau,n}, u)\big| = 0$, and so at the first step, the problem in (15) reduces to selecting action sequences that maximize the sum of the variances along the trajectory (plus whatever other side information is included in the term s_τ).

At subsequent iterations, the discrepancy is no longer uniformly 0 throughout the state-action space, and the solutions of action sequences will incorporate how the observed dynamics (as represented by the updated models $\mu_{\tau,n}$) differ from the previous episode's learned model $\mu_{\tau-1,N}$. In this way, our method prioritizes the selection of actions that are likely to explore in regions of state-action space where our current data model m_τ is misaligned with the true dynamics f.

At each iteration n in episode τ, we solve (15) to compute the exploration policy. In practice, solving (15) can pose a challenge due to the presence of the square root of the variance $\sigma_{\tau,n}$ in the objective. Thus, we can use a sample-based MPC method to compute a sequence of control actions, for instance via Cross Entropy Maximization (CEM) [14].

A. Consistency Guarantee

Active sampling procedures should possess guarantees that they lead to consistent estimators of the underlying dynamics. It is usually the case that consistency arguments are made under the assumption that the observed data is given i.i.d., which is not the case in most active learning settings. We establish consistency of our GP model $\mu_{\tau,n}$ within our active sampling framework by leveraging the useful property that GPs are all-time calibrated statistical models [10, 11]. Our key insight is to leverage the decrease in predictive variance values of our GP model at key points along observed trajectories to control the overall decrease in predictive variance values over the state-action space $\mathcal{Z}$.

We begin by stating important definitions and assumptions that will allow us to conclude consistency of the GP mean $\mu_{\tau,N} \to f$ from an argument on the convergence of the predictive variance values $\sigma^2_{\tau,N} \to 0$. We make the following assumption and give a useful definition:

Assumption 1. *We assume that the coordinate-wise functions $[f(\cdot)]_\ell = f_\ell : \mathcal{Z} \to \mathbb{R}$ lie within a RKHS with kernel k and have bounded norm B. That is, $f \in \mathcal{H}^d_{k,B} = \{f : \|f_\ell\|_k \leq B, \ell = 1, \dots, d\}$.*

Definition 1 (All-time calibrated statistical model of f, [15])**.** *Let $z = (x, u)$ and $\mathcal{Z} := \mathcal{X} \times \mathcal{U}$. An all-time calibrated statistical model for the function f is a sequence $\{\mu_j, \sigma_j, \beta_j(\delta)\}_{j \geq 0}$ such that for all $z \in \mathcal{Z}$, $\ell \in \{1, \dots, d\}$, and $j \in \mathbb{N}$*

$$|[\mu_j(z)]_\ell - [f(z)]_\ell| \leq \beta_j(\delta)[\sigma_j(z)]_\ell, \tag{16}$$

with probability greater than or equal to $1 - \delta$. Here we denote the ℓ^{th} element of a vector $\mathbf{v} \in \mathbb{R}^d$ as $[\mathbf{v}]_\ell$. The scalar function, $\beta_n(\delta) \in \mathbb{R}_+$ quantifies the width of the $1-\delta$ confidence intervals. We assume without loss of generality that β_j monotonically increases with n, and that $[\sigma_j(z)]_\ell \leq \sigma_{max}$ for all $z \in \mathcal{Z}, j \geq 0$, and $\ell \in \{1, \dots, d\}$.

Then, assuming that the true dynamics f we wish to model is a bounded function (i.e., Assumption 1) within an RKHS of vector-valued functions, then we can conclude that the GP model is a well-calibrated statistical model for f:

Lemma 1 (Well-calibrated confidence intervals for RKHS, [15])**.** *Let $f \in \mathcal{H}^d_{k,B}$ and suppose that μ_j and σ_j are the posterior mean and variance of a GP with kernel k. There exists $\beta_j(\delta)$, for which the tuple $(\mu_j, \sigma_j, \beta_j(\delta))$ is an all-time-calibrated statistical model (Definition 1) of f .*

Note that while these definitions of the predictive variance are considered to be vector-valued, our GP model simply considers scalar variance values (i.e., the case that all entries of the vector-valued variance are equivalent). Thus, while our result could be stated in terms of bounding the norm of a vector-valued variance, $\|\boldsymbol{\sigma}^{\mathbf{2}}_{\tau,N}\|^2_2$, we will just consider the convergence of the scalar value $\sigma^2_{\tau,N}$.

To summarize, to show consistency of the underlying GP model $\mu_{\tau,N}(z) \to f(z)$ as $\tau \to \infty$, one simply needs to show that the variance of the GP model $\sigma^2_{\tau,N}$ vanishes as more episodes occur. We consider this guarantee under the following additional assumptions

Assumption 2. *The state-action space $\mathcal{Z}$ is finite; that is, $|\mathcal{Z}| = N_z < \infty$.*

Assumption 3. *The side information term s_τ is already captured in the definition of the domain of possible state-action pairs, $\mathcal{Z}$ (i.e., $s_\tau \equiv 0$ for our purposes).*

Assumption 4. *In each episode τ, there exists at least one index $n_* \in \{1, \dots, N\}$ at which the value z^{τ,n_*} in the sampled trajectory $\{z^{\tau,n} = (x^{\tau,n}, u^{\tau,n})\}_{n=1}^N$ satisfies*

$$\sigma^2_{\tau,n_*}(z^{\tau,n_*}) \geq \frac{1}{N_z} \sum_{z \in \mathcal{Z}} \sigma^2_{\tau,n_*}(z), \tag{17}$$

where N_z is the cardinality of the state-action space $\mathcal{Z}$.

Our results can be straightforwardly extended to bounded state-action spaces, but assuming that $\mathcal{Z}$ is finite (Assumption 2) simplifies the presentation of our proofs hereafter. The final assumption (Assumption 4) is a technical assumption that simply asserts that the trajectory in each episode will contain at least one state-action pair $z = (x, u)$ that has a relatively large variance value. We suggest that this is a rather mild condition which is reasonable to assume in various settings, such as in our case of discrepancy-based GP-UCB planning, since we take into account the variance values when identifying action sequences. A setting in which this assumption is obviously satisfied is when trajectories are allowed to start at arbitrary $z \in \mathcal{Z}$, such as those points with maximal variance.

Finally, we prove a useful Lemma that writes the iteration update in the predictive variance values in terms of the previous variance values.

Lemma 2. *Consider a GP utilizing kernel k with predictive mean* (6) *and variance* (7)*. Upon observation at $z^* \in \mathcal{Z}$, then the update to the predictive variance at the point $z \in \mathcal{Z}$ can* *be written as*

$$\sigma_{n+1}^2(z) = \sigma_n^2(z) - \frac{\operatorname{cov}_{\sigma^2}^2(z^*, z)}{\sigma_n^2(z^*) + \sigma^2}. \tag{18}$$

Proof. First, we write the definition of the predictive variance upon observing at z^*:

$$\begin{aligned}&\sigma_{n+1}^2(z) = k(z,z) \\ &- \left(\boldsymbol{k}_z \ k(z^*,z)\right) \begin{pmatrix} G_n + \sigma^2 I & \boldsymbol{k}_{z^*} \\ \boldsymbol{k}_{z^*}^T & k(z^*,z^*) + \sigma^2 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{k}_z \\ k(z^*,z) \end{pmatrix}\end{aligned} \tag{19}$$

$$=: k(z,z) - \left(\boldsymbol{k}_z \ k(z^*,z)\right) W_{n+1} \begin{pmatrix} \boldsymbol{k}_z \\ k(z^*,z) \end{pmatrix}. \tag{20}$$

Defining $W_n = (G_n + \sigma^2 I)^{-1}$ and $a^{-1} = k(z^*,z^*) + \sigma^2 - \boldsymbol{k}_{z^*}^T W_n \boldsymbol{k}_{z^*} = \sigma_n^2(z^*) + \sigma^2$, we can use the block matrix inversion formula to compute

$$W_{n+1} = \begin{pmatrix} W_n & 0 \\ 0 & 0 \end{pmatrix} + a \begin{pmatrix} W_n \boldsymbol{k}_{z^*} \boldsymbol{k}_{z^*}^T W_n & -W_n \boldsymbol{k}_{z^*} \\ -\boldsymbol{k}_{z^*}^T W_n & 1 \end{pmatrix} \tag{21}$$

$$= \begin{pmatrix} W_n & 0 \\ 0 & 0 \end{pmatrix} + a \begin{pmatrix} W_n \boldsymbol{k}_{z^*} \\ -1 \end{pmatrix} \begin{pmatrix} \boldsymbol{k}_{z^*}^T W_n & -1 \end{pmatrix}. \tag{22}$$

Now, we can compute the inner product

$$\begin{pmatrix} \boldsymbol{k}_{z^*}^T W_n & -1 \end{pmatrix} \begin{pmatrix} \boldsymbol{k}_z \\ k(z^*,z) \end{pmatrix} = -\left(k(z^*,z) - \boldsymbol{k}_{z^*}^T W_n \boldsymbol{k}_z\right) \tag{23}$$

$$= -\operatorname{cov}_{\sigma^2}(z^*, z), \tag{24}$$

which allows us to conclude

$$\sigma_{n+1}^2(z) = k(z,z) - \boldsymbol{k}_z^T W_n \boldsymbol{k}_z - a \left(\operatorname{cov}_{\sigma^2}(z^*,z)\right)^2 \tag{25}$$

$$= \sigma_n^2(z) - \frac{\operatorname{cov}_{\sigma^2}^2(z^*,z)}{\sigma_n^2(z^*) + \sigma^2}, \tag{26}$$

which concludes the proof. □

With the stated assumptions and Lemma 2, we now state our main theorem regarding the convergence of variance values resulting from our proposed active sampling method.

Theorem 1. *If for all $\tau \geq 1$ we have that Assumptions 2-4 are satisfied and that the regularization parameter is scaled as $\sigma^2 = ((\tau - 1)N + n)^{-2}$, then we have that sampling action sequences based on the finite-horizon planning using discrepancy-based GP-UCB acquisition function* (15) *gives convergence of the posterior variance values,*

$$\max_{z \in \mathcal{Z}} \ \sigma_{\tau,0}^2(z) \to 0, \tag{27}$$

as $\tau \to \infty$.

If Assumption 1 holds, then Lemma 1 implies consistency of the corresponding predictive mean $\mu_{\tau,0} \to f$ as $\tau \to \infty$.

Proof. Let $\tau \geq 2$ denote a given episode and let $n_* \in \{1, \dots, N\}$ be the iteration index given by Assumption 4. Furthermore, let $z_* := z^{\tau, n_*}$ be the corresponding state-action pair selected at the identified iteration. Due to the monotonic decreasing nature of the variance over iterations and episodes, we can write

$$\max_{z \in \mathcal{Z}} \ \sigma_{\tau+1,0}^2(z) \leq \sum_{z \in \mathcal{Z}} \ \sigma_{\tau,N}^2(z) \leq \sum_{z \in \mathcal{Z}} \ \sigma_{\tau,n_*+1}^2(z). \tag{28}$$

According to Lemma 2, the decrease in the variance for z_* at the identified iteration n_* can be written as

$$\sigma_{\tau,n_*+1}^2(z_*) = \sigma_{\tau,n_*}^2(z_*) - \frac{\sigma_{\tau,n_*}^4(z_*)}{\sigma^2 + \sigma_{\tau,n_*}^2(z_*)} \tag{29}$$

$$= \left(\frac{\sigma^2}{\sigma^2 + \sigma_{\tau,n_*}^2(z_*)}\right) \sigma_{\tau,n_*}^2(z_*) \tag{30}$$

$$\leq \left(\frac{((\tau-1)N)^{-2}}{((\tau-1)N)^{-2} + \sigma_{\tau,n_*}^2(z_*)}\right) \sigma_{\tau,n_*}^2(z_*) \tag{31}$$

$$\leq \left(\frac{N^{-2}}{N^{-2} + \sigma_{\tau,n_*}^2(z_*)}\right) \sigma_{\tau,n_*}^2(z_*), \tag{32}$$

where we have used the property that $\sigma^2 = ((\tau - 1)N + n)^{-2} \leq ((\tau - 1)N)^{-2}$ and that $\tau \geq 2$. Plugging this into (28) and using (17) from Assumption 4, we bound

$$\max_{z \in \mathcal{Z}} \ \sigma_{\tau+1,0}^2(z)$$

$$\leq \sum_{z \neq z_*} \sigma_{\tau,n_*}^2(z) + \left(\frac{N^{-2}}{N^{-2} + \sigma_{\tau,n_*}^2(z_*)}\right) \sigma_{\tau,n_*}^2(z_*) \tag{33}$$

$$\leq \sum_{z \in \mathcal{Z}} \sigma_{\tau,n_*}^2(z) - \left(\frac{\sigma_{\tau,n_*}^2(z_*)}{N^{-2} + \sigma_{\tau,n_*}^2(z_*)}\right) \sigma_{\tau,n_*}^2(z_*) \tag{34}$$

$$\leq \left(1 - \frac{1}{N_z}\left(\frac{\sigma_{\tau,n_*}^2(z_*)}{N^{-2} + \sigma_{\tau,n_*}^2(z_*)}\right)\right) \sum_{z \in \mathcal{Z}} \sigma_{\tau,n_*}^2(z) \tag{35}$$

$$\leq \left(1 - \frac{\sum_{z \in \mathcal{Z}} \sigma_{\tau,n_*}^2(z)}{N_z\left(N_z N^{-2} + \sum_{z \in \mathcal{Z}} \sigma_{\tau,n_*}^2(z)\right)}\right) \sum_{z \in \mathcal{Z}} \sigma_{\tau,n_*}^2(z) \tag{36}$$

$$\leq \left(1 - \frac{\sum_{z \in \mathcal{Z}} \sigma_{\tau,0}^2(z)}{N_z\left(N_z N^{-2} + \sum_{z \in \mathcal{Z}} \sigma_{\tau,0}^2(z)\right)}\right) \sum_{z \in \mathcal{Z}} \sigma_{\tau,0}^2(z), \tag{37}$$

where in the last line we have used the monotonic decreasing property of the predictive variance values over the iterations.

Defining $v_\tau = \frac{1}{N_z} \sum_{z \in \mathcal{Z}} \sigma_{\tau,0}^2(z)$, we see the recurrence relation

$$v_{\tau+1} - v_\tau \leq \frac{-v_\tau^2}{N_z(N_z N^{-2} + v_\tau)}, \tag{38}$$

with initial condition $v_1 = N_z$ (since $\sigma_{1,0}^2(z) = 1$ for all $z \in \mathcal{Z}$). Noting that the function $f(v) = -v^2/(a(b+v))$ is decreasing on $v \in (0, \infty)$, we can pass to a corresponding differential equation to serve as an upper bound for v_τ:

$$v'(t) = \frac{-v^2}{N_z(N_z N^{-2} + v)}, \qquad v(0) = 1, \tag{39}$$

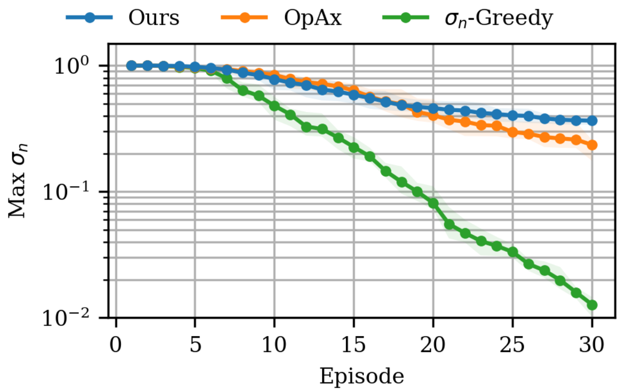

Fig. 1. Average reduction of maximum variance (uncertainty) over 8000 test points $\mathcal{T}$ spaced evenly over the entire state space. The shaded region shows the maximum and minimum values over 10 independent trials.

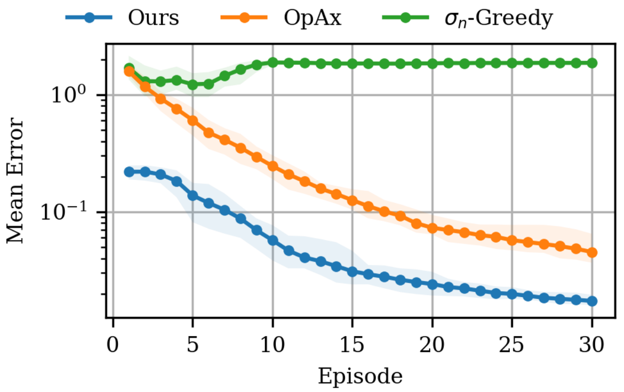

Fig. 2. Mean squared error (MSE) of the learned models over test points $\mathcal{T}$. The inclusion of side information leads to reduced error. The shaded region shows the maximum and minimum values over 10 independent trials.

for which we can solve for the "time" $t(\epsilon) > 0$ for which $v(t(\epsilon)) = \epsilon$ by separation of variables:

$$t(\epsilon) = -N_z \left(N_z N^{-2} \int_{N_z}^{\epsilon} \frac{1}{v^2} dv + \int_{N_z}^{\epsilon} \frac{1}{v} dv \right) \qquad (40)$$

$$= \frac{N_z(N_z - \epsilon)}{N^2 \epsilon} + N_z \log \left(\frac{N_z}{\epsilon} \right). \qquad (41)$$

That is to say, by episode $\tau(\epsilon) \geq t(\epsilon)$, we are ensured that

$$\max_{z \in \mathcal{Z}} \sigma^2_{\tau(\epsilon),0}(z) \leq \epsilon. \qquad (42)$$

With this explicit convergence rate, we can see that as $\tau \to \infty$ we then obtain that $\sigma^2_{\tau,0}(z) \to 0$ for all $z \in \mathcal{Z}$. □

The proof of Theorem 1 relies on the convenient fact that our discrepancy-based GP-UCB acquisition function reduces to maximum variance sampling in the first iteration of each episode. As such, the monotonicity of predictive variance in combination with the assumption that we can start trajectories at arbitrary $z \in \mathcal{Z}$ gives that we can bound the variance at each iteration in an episode by the variance of the value z_* given by Assumption 4.

IV. Numerical Results

For all experiments, we use a Gaussian (RBF) kernel $k(x, x') = \exp(-\gamma \|x - x'\|^2)$, $\gamma = 0.5$, and the regularization parameter is chosen to be $\sigma^2 = 1/n^2$, where n is the size of the dataset.

A. Estimating a Simple Pendulum System

We first consider the problem of estimating the system dynamics of the controlled pendulum system as in [16]. The equations of motion of the pendulum are given by $ml^2\ddot{\theta} + 3mgl\sin(\theta) = 3u$, where $g = 9.81$ is the acceleration due to gravity, the link mass is $m = 1$, and the link length is $l = 1$. The state of the system is given by the angle θ and angular velocity $\dot{\theta}$ of the pendulum, $x = [\theta, \dot{\theta}]^\top \in \mathbb{R}^2$, and the control input is the torque applied to the pendulum, $u \in \mathbb{R}$. For the benchmark, the angle θ is adjusted to be within the range $\theta \in [-\pi, \pi]$, the angular velocity is bounded such that $\dot{\theta} \in [-8, 8]$, and the control input is bounded to $u \in [-2, 2]$, meaning the system is under-actuated. We presume that the true dynamics are unknown but that we have access to an *imperfect* prior model of the system with mis-specified parameters $g = 9.0$, link mass $m = 0.5$, and link length $l = 2.0$.

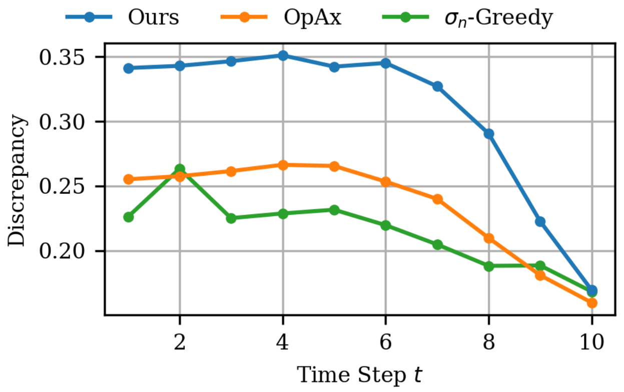

Fig. 3. Average discrepancy between the prior model and the true, underlying dynamics at the points visited by the algorithms during each episode.

We implement two baselines for comparison. First, we consider OpAx [10], which uses optimistic planning to identify action sequences that decrease predictive variance. At the beginning of each episode, OpAx computes an open-loop policy that maximizes the information gain during the episode using an optimistic estimate of the dynamics. It is called "optimistic" since it uses a planner that adds additional control variables to artificially steer plausible rollouts to states that maximize information gain [11]. Second, we consider a greedy variance-based exploration algorithm that myopically selects the control input at each iteration that has maximum variance. In other words, this σ_n-greedy policy is a purely exploration-based policy.

For all approaches, we use a GP model as in (4) and (5) and sample over 30 episodes with a time horizon of

$N = 10$. For our approach and OpAx, we use the improved cross-entropy MPC planner (iCEM) [14] to compute the exploration policy. We use a planning horizon of 10, with 10 iterations, 50 action sequence samples at each iteration, and an elite set size of 10, holding 5 elites between iterations. See [14] for more details. In each episode, we choose the initial condition x_0 to be the point in $\mathcal{X}$ with the highest variance according to the current GP estimate of the dynamics (11).

In order to compare our approach to OpAx and the σ_n-greedy policy, we compute the maximum uncertainty and mean prediction accuracy at a set of 8000 test points $\mathcal{T} = \{(x_j, u_j)\}_{j=1}^{8000}$ spaced uniformly in the state space. Figure 1 shows the reduction in maximum variance and Figure 2 shows the mean squared error (MSE) between the learned model $\mu_{\mathcal{T},N}$ as in (10) and the true underlying dynamics f. To demonstrate good performance, an active learning algorithm must balance exploration (reducing uncertainty or variance) while simultaneously improving accuracy.

We see in Figure 1 that our approach performs comparably to OpAx at reducing uncertainty, though we explore the state-action space at a slightly reduced rate. Nevertheless, we see in Figure 2 that our approach initially starts with a significantly reduced MSE due to the inclusion of side information via a non-zero mean prior. While it is possible to augment the GP model in OpAx to utilize such a prior, we note that the exploration procedure in OpAx does not take the prior into account during exploration. This is key because it means that even though both methods explore at a similar rate, our approach focuses its exploration in areas with the highest mismatch (discrepancy) between the prior and the observations, providing higher intrinsic value for the learning task. We can see this clearly in Figure 3, which shows the average discrepancy between the prior model p_0 and the true dynamics f of the visited states during each episode. This is an important result, since it means we favor regions with high model mismatch while avoiding redundant regions where our prior model aligns closely with the true dynamics. As expected, we see in Figure 1 that the σ_n-greedy policy performs the best at reducing the uncertainty of the GP-based estimate, but note that this does not correspond to a commensurate improvement in the approximation quality.

B. Control Performance Using Learned Dynamics

We next consider the problem of computing a control policy using the model learned via our proposed active learning approach. Our goal is to measure whether the learned model is sufficiently high-fidelity for the purpose of control. We consider the half-cheetah environment from the MuJoCo benchmark suite [17]. The half-cheetah system is specified by a collection of rigid links, joints, and actuators (see Figure 4). The state is in $\mathbb{R}^{18}$, consisting of the position and velocity of the various links, and the input space is in $\mathbb{R}^6$, representing the torques applied to each motor, bounded to be in $[-1, 1]$.

We consider solving an optimal control problem as in (2), where the objective is to maximize the travel speed to the right (the positive x direction), while minimizing control effort. We presume that the true system is unknown, but that we have access to side information in the form of an imperfect model with a torso that is 1.8 times as long as the actual system (Figure 4, right).

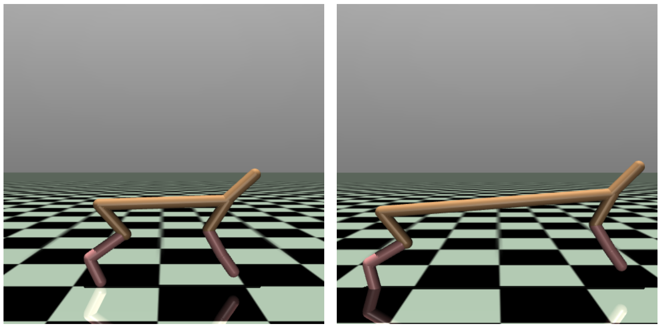

Fig. 4. The true half-cheetah system (left) and the imperfect half-cheetah system used as the bias for our algorithm (right).

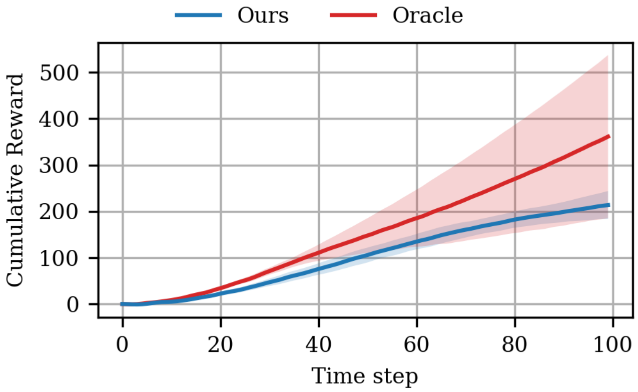

Fig. 5. Cumulative reward of our approach at a downstream control task compared to the oracle using the true dynamics.

As before, we use the iCEM planner [14] to compute the exploration policy. We use a planning horizon of 10, with 10 iterations, 100 action sequence samples at each iteration, and an elite set size of 10, holding 5 elites between iterations. Then, we fix the GP model of the dynamics and compute a separate policy to solve the optimal control problem as in (2) using MPC, using the mean predictor $\mu_{\mathcal{T},n}$ as in (10) for the predictive model of the dynamics in (2b). We similarly use iCEM to compute the control inputs, only using 200 action sequence samples, an elite set size of 50, and holding 15 elites per iteration.

Figure 5 shows the average cumulative reward of the system over 10 independent runs using the learned model as the predictive model. For comparison, we also computed the average cumulative reward using the actual system dynamics (oracle). We can see in Figure 5 that our approach yields a model that demonstrates good empirical performance at a downstream control task.

V. Conclusion

In this paper, we present an active learning method that incorporates prior domain knowledge in the sampling procedure as well as the learned model. Under reasonable assumptions, we prove that our active sampling method provides a consistent estimator of the dynamics. Through numerical experiments, we demonstrate that our active learning approach produces an empirical dynamics estimate with lower error than methods that neglect prior knowledge, while simultaneously prioritizing exploration in regions that demonstrate a higher discrepancy between the prior model and our data-driven estimate.

REFERENCES

[1] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 397–422, 2002.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, May 2002.

[3] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Annual Conference Computational Learning Theory*, 2008.

[4] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning.*, ser. Adaptive computation and machine learning. MIT Press, 2006.

[5] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[6] A. Wagenmaker and K. Jamieson, "Active learning for identification of linear dynamical systems," in *Proceedings of Thirty Third Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, J. Abernethy and S. Agarwal, Eds., vol. 125. PMLR, 09–12 Jul 2020, pp. 3487–3582.

[7] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Proceedings of the 31st Conference On Learning Theory*, ser. Proceedings of Machine Learning Research, S. Bubeck, V. Perchet, and P. Rigollet, Eds., vol. 75. PMLR, 06–09 Jul 2018, pp. 439–473.

[8] M. Simchowitz, R. Boczar, and B. Recht, "Learning linear dynamical systems with semi-parametric least squares," in *Proceedings of the Thirty-Second Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, Eds., vol. 99. PMLR, 25–28 Jun 2019, pp. 2714–2802.

[9] E. Daş and J. W. Burdick, "An active learning based robot kinematic calibration framework using Gaussian processes," 2023.

[10] B. Sukhija, L. Treven, C. Sancaktar, S. Blaes, S. Coros, and A. Krause, "Optimistic active exploration of dynamical systems," in *Advances in Neural Information Processing Systems 36*, 2023, Conference Paper, 37th Conference on Neural Information Processing Systems (NeurIPS 2023); Conference Location: New Orleans, LA, USA; Conference Date: December 10-16, 2023; Poster presentation on December 12, 2023.

[11] S. Curi, F. Berkenkamp, and A. Krause, "Efficient model-based reinforcement learning through optimistic policy search and planning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 14 156–14 170.

[12] A. Goldenshluger and A. Zeevi, "A note on performance limitations in bandit problems with side information," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1707–1713, 2011.

[13] C.-C. Wang, S. Kulkarni, and H. Poor, "Bandit problems with side observations," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 338–355, 2005.

[14] C. Pinneri, S. Sawant, S. Blaes, J. Achterhold, J. Stueckler, M. Rolinek, and G. Martius, "Sample-efficient cross-entropy method for real-time planning," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 1049–1065.

[15] J. Rothfuss, B. Sukhija, T. Birchler, P. Kassraie, and A. Krause, "Hallucinated adversarial control for conservative offline policy evaluation," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2023.

[16] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," 2016.

[17] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.