Exploration via Planning for Information about the Optimal Trajectory

Anonymous Author(s)
Affiliation
Address
email

Abstract

Many potential applications of reinforcement learning (RL) are stymied by the large numbers of samples required to learn an effective policy. This is especially true when applying RL to real-world control tasks, e.g. in the sciences or robotics, where executing a policy in the environment is costly. In popular RL algorithms, agents typically explore either by adding stochasticity to a reward-maximizing policy or by attempting to gather maximal information about environment dynamics without taking the given task into account. In this work, we develop a method that allows us to plan for exploration while taking both the task and the current knowledge about the dynamics into account. The key insight to our approach is to plan an action sequence that maximizes the expected information gain about the optimal trajectory for the task at hand. We demonstrate that our method learns strong policies with 2-200x fewer samples compared to 14 baselines on a diverse set of control tasks in both the open-loop and closed-loop control settings.

1 Introduction

The potential of reinforcement learning (RL) as a general-purpose method of learning solutions to sequential decision making problems is difficult to overstate. Ideally, RL could allow for agents that learn to accomplish all manner of tasks solely through a given reward function and the agent's experience; however, RL has so far broadly fallen short of this. Chief among the difficulties in realizing this potential is the fact that typical RL methods in continuous problems require very large numbers of samples to achieve a near-optimal policy. For many interesting applications of RL this problem is exacerbated by the fact that collecting samples from the true environment can incur huge costs. For example, expensive scientific experiments are needed for tokamak control [30, 12, 20] and design of molecules [62], and collecting experience on the road is both costly and runs the risk of car accidents for autonomous vehicles [60].

Though model-based methods like Deisenroth and Rasmussen [21], Chua et al. [14], and Curi et al. [15] are much more efficient than typical model-free methods, they do not explicitly reason about the prospective task-relevant information content of future observations. Moreover, these methods typically require thousands of timesteps of environment dynamics data to solve reasonably simple Markov decision processes (MDPs). Though progress has been made in using Thompson sampling [43], entropy bonuses [26], and upper confidence bounds (UCB) [15, 5] to more intelligently explore the state-action space of an MDP, these methods still do not explicitly reason about how information that the agent gathers will affect the estimated MDP solution. Furthermore, in continuous state-action settings these methods must make coarse approximations to be computationally tractable (e.g. bootstrapped Q networks to approxiate a posterior or one-step perturbations for approximate UCB).

Many methods in the vein of Pathak et al. [44] and Shyam et al. [56] explore directly to gather new information based on current uncertainty about environment dynamics, but they do not in general

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

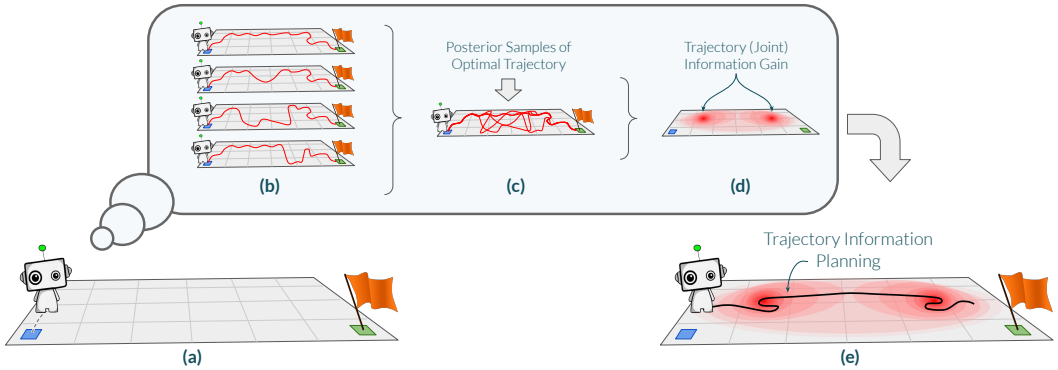

Figure 1: A schematic depiction of Trajectory Information Planning (TIP). Suppose the agent in (a) aims to determine where to explore next from its current state. To do so, in (b) the agent samples dynamics models $T' \sim P(T \mid D)$ from its current posterior and finds approximately optimal trajectories $\tau^* \sim P(\tau^* \mid T')$ for each sample. Then in (c) it pools these samples of posterior optimal trajectories τ^*. In (d) it constructs a function that gives the joint expected information gain about the optimal trajectory τ^* given a planned exploration trajectory (i.e. EIG_{τ^*} over the set of points visited). Finally, in (e) the agent can plan an action sequence which maximizes this joint expected information gain.

specialize the information that they aim to acquire to a particular task. We believe that an ideal exploration strategy for sample efficiency is able to take both a relevant task into account as well as uncertainty about the environment dynamics.

In this work, we start by showing how many methods can be cast as a Bayesian planning problem over a specific cost function. This framework helps illuminate the importance of the cost function and what impact it has on exploration. Viewed in this light, it becomes clear that many previous state-of-the-art methods rely on cost functions that either result in behavior that is too greedy—i.e. the policy simply tries to maximize returns during exploration—or too exploratory, i.e. the policy is incentivized to explore the environment dynamics and does not consider the task at hand. We therefore present a cost function that balances out these two extremes. In particular, our cost function captures the amount of information that would be gained about the *optimal trajectory*, if the agent were to explore by following a particular planned trajectory. As depicted in Figure 1, this involves the agent planning out what it would hypothetically do given different realizations of the dynamics and planning actions that are informative about those possibilities.

In summary, the contributions of this work are as follows: we develop a novel cost function for exploration that explicitly accounts for both the specific task as well as uncertainty about environment dynamics given past experience, a method for planning which applies the cost function to explore in MDPs with continuous states and actions, and a thorough empirical evaluation of our method across 5 closed-loop and 3 open-loop environments (with a focus on expensive RL tasks in plasma physics) compared against 14 baselines. We find that our proposed method is able to learn policies that perform as well as an agent with access to the ground truth dynamics using 2-200x fewer samples than comparison methods.

2 Related Work

Exploration in Reinforcement Learning The most common strategy for exploration in RL is to execute a greedy policy with some form of added stochasticity. The simplest strategy, ϵ-greedy exploration as used in Mnih et al. [39], simply takes the current action thought to be best with probability $1 - \epsilon$ and a random action with probability ϵ. Other methods use added Ornstein-Uhlenbeck action noise [36] to the greedy policy, or entropy bonuses [26] to the policy or value function objecties to add noise to a policy which is otherwise optimizing the RL objective.

Tabular RL is often solved by choosing actions based on upper confidence bounds on the value function [13, 35], but explicitly computing and optimizing these bounds in the continuous setting is substantially more challenging. Recent work [16] approximates this method by computing one-step confidence bounds on the dynamics and training a 'hallucinated' policy which chooses perturbations within these bound to maximize expected policy performance. Another recent work [5] uses anti-concentration inequalities to approximate upper confidence bounds in MDPs with discrete actions.

Thompson sampling (TS) [51], which samples a realization of the MDP from the posterior and acts optimally as if the realization was the true model, can be applied for exploration in a model-free manner as in [42] or in a model-based manner as in [57]. As the posterior over MDP dynamics or value functions can be high-dimensional and difficult to represent, the performance of TS can be hindered by approximation errors using both Gaussian processes and ensembles of neural networks. Curi et al. [15] recently investigated this and found that this was potentially due to an insufficiently expressive posterior over entire transition functions, implying that it may be quite difficult to solve tasks using sampled models. Similarly, the posterior over action-value functions in Osband et al. [42] is only roughly approximated by training a bootstrapped ensemble of neural networks.

There is also a rich literature of Bayesian methods for exploration, which are typically computationally expensive and hard to use, though they have attractive theoretical properties. These methods build upon the fundamental idea of the Bayes-adaptive MDP [49], which we detail in Section E.1 alongside a discussion of this literature.

Additionally, a broad set of methods explore to learn about the environment without a task in mind. This line of work is characterized by Pathak et al. [44], which synthesizes a task-agnostic reward function from model errors. Other techniques include MAX [56], which optimizes the information gain about the environment dynamics, Random Network Distillation [10], which forces the agent to learn about a random neural network across the state space, and Plan2Explore [55], which prospectively plans to find areas of novelty where the dynamics are uncertain.

Bayesian Experimental Design: BOED, BO, BAX, and BARL There is a large literature on Bayesian optimal experiment design (BOED) [11] which focuses on efficiently querying a process or function to get maximal information about some quantity of interest. When the quantity of interest is the location of a function optimum, related strategies have been proposed as the predictive entropy search family of Bayesian optimization (BO) algorithms [28, 29]. Recently, a flexible framework known as Bayesian algorithm execution (BAX) [40] has been proposed to efficiently estimate properties of expensive black-box functions; this framework gives a general procedure for sampling points which are informative about the future execution of a given algorithm that computes the property of interest, thereby allowing the function property to be estimated with far less data.

A subsequent related work [38], known as Bayesian Active Reinforcement Learning (BARL), uses ideas from BOED and BAX to sample points that are maximally informative about the optimal trajectory in an MDP. However, BARL relies on a setting the authors call Transition Query Reinforcement Learning (TQRL), which assumes that the environment dynamics can be iteratively queried at an arbitrary sequence of state-action pairs chosen by the agent. TQRL is thus a highly restrictive setting which is not suitable when data can only be accessed via a trajectory (rollout) of environment dynamics; it typically relies on an accurate environment simulator of sufficient expense to warrant its use. Even then, there will likely be differences between simulators and ground truth dynamics for complex systems. Thus, one would ideally like to collect data in real environments. However, this often requires leaving the TQRL setting, and instead collecting data via trajectories only.

In this paper, we aim to apply similar information-theoretic ideas, but extend them to the general MDP setting, as well as learning open loop model-based controllers. The typical method for learning to solve open-loop control problems was demonstrated successfully in Tesch et al. [59], where a value function was learned from action sequences to task success. Our method takes a model-based approach to this problem, using similar exploration strategies to Bayesian optimization but benefitting from the more substantial supervision that is typical in dynamics model learning.

3 Problem Setting

In this work we deal with finite-horizon discrete-time *Markov decision processes* (MDPs) which consist of a sextuple $\langle \mathcal{S}, \mathcal{A}, T, r, p_0, H \rangle$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, T is the transition function $T : \mathcal{S} \times \mathcal{A} \to P(\mathcal{S})$ (using the convention that $P(\mathcal{X})$ is the set of probability measures over $\mathcal{X}$), $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a reward function, $p_0(s)$ is a distribution over $\mathcal{S}$ of start states, and $H \in \mathbb{N}$ is the horizon. We always assume $\mathcal{S}, \mathcal{A}, p_0$, and H are known. We also assume the reward r is known, though our development of the method can easily be generalized to the case where r is unknown. Our primary function of interest is the transition function T, which we learn from data. We address both open and closed loop control settings. In the more common closed loop setting, our aim is to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes Objective (1). We discuss trajectories $\tau \sim p(\tau \mid \pi, T)$ where $\tau = [(s_0, a_0), \ldots, (s_{H-1}, a_{H-1}), s_H]$ generated by $s_0 \sim p_0$, $a_i = \pi(s_i)$,

and $s_{i+1} \sim T(s_i, a_i)$. We can write the return of a trajectory as $R(\tau) = \sum_{i=0}^{H-1} r(s_i, a_i, s_{i+1})$ for the states and actions s_i, a_i that make up τ. The MDP objective can then be written as

$$J_T(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi,T)} \left[R(\tau) \right]. \tag{1}$$

We aim to maximize this objective while minimizing the number of samples from the ground truth transition function T that are required to reach good performance. We denote the optimal policy as $\pi^* = \arg\max_\pi J_T(\pi)$, which we can assume to be deterministic [58] but not necessarily unique. We use τ^* to denote optimal trajectories, i.e. $\tau^* \sim p(\tau \mid \pi^*, T)$

Equivalently for the open-loop setting, we assume a fixed start state s_0 and aim to find an action sequence $a_0, \ldots, a_{H-1}$ that maximizes the sum of rewards in an episode. We will slightly abuse notation and write $\tau \sim p(\tau \mid a_{0:H-1})$ and $J_T(a_{0:H-1})$ with these actions fixed in place of a reactive policy and again use τ^* to refer to the trajectories generated by an optimal action sequence.

We assume in this work that applying planning algorithms like [47] to a dynamics function T will result in a trajectory that approximates τ^*. We will primarily focus on a Gaussian process (GP) model of the transition dynamics in order to take advantage of its expressive representation of uncertainty and grounded methods for sampling, conditioning, and joint inference. There is substantial prior work using GPs in RL—see Section E.2 for a discussion of this literature. Under this modeling choice, we assume that the dynamics are drawn from a GP prior (see Section A.3 for further details on our GP model).

4 Trajectory Information Planning

Our method consists of a generic framework for Bayesian or approximately Bayesian model-predictive control and a novel cost function for planning that allows us to explicitly plan to find the maximal amount of new information relevant to our task. In Section 4.1, we describe the MPC framework and highlight that many prior methods approximate this framework while using the greedy cost function corresponding to the future negative expected rewards or the pure exploration cost function corresponding to future information about the dynamics. Afterwards in Section 4.2, we derive our new cost function and describe how it is computed. The overall method is simply applying the planning framework with the new cost function.

4.1 Model-Predictive Control in Bayesian Model-Based RL

In this section, we give a formulation of Bayesian planning for control that generalizes ideas from methods such as PILCO [21] and PETS [14]. This formulation highlights these methods' inherently greedy nature and hints at a possible solution. The objective of Bayesian planning is simply to find the h-step action sequence that maximizes the expected future returns under model uncertainty. That is,

$$\underset{a_0,\ldots,a_{h-1}}{\operatorname{argmin}} \; \mathbb{E}_{T' \sim P(T|D), \tau_e \sim P(\tau|s_0=s, a_{0:h-1}, T')} \left[C(\tau_e) \right] \tag{2}$$

for some cost function C over trajectories and some start state s. If in the open-loop control setting, the agent simply executes the sequence of actions found. This procedure can also be extended to closed-loop control via model-predictive control (MPC). This procedure simply involves re-planning (2) at every state the agent visits and playing the first action from the optimal sequence. Concretely, the MPC policy for our Bayesian setting is as follows:

$$\pi_{\text{MPC}}(s) = \arg\min_{a_0} \min_{a_1,\ldots,a_{h-1}} \mathbb{E}_{T' \sim P(T|D), \tau_e \sim P(\tau|s_0=s, a_{0:h-1}, T')} \left[C(\tau_e) \right] \tag{3}$$

Whether we do open-loop control or closed-loop control via MPC, the cost function, C, is integral to how the agent will behave. Prior work has predominantly focused on two types of cost function:

$$\underbrace{C_g(\tau) = -R(\tau)}_{\text{Greedy Exploration}} \qquad \underbrace{C_e(\tau) = -\sum_{i=0}^{h} \mathbb{H}\left[T(s_i, a_i) \mid D \right]}_{\text{Task-Agnostic Exploration}} \tag{4}$$

Previous works such as Kamthe and Deisenroth [33] and PETS [14] use the greedy exploration cost function, C_g. This cost function simply plans trajectories that achieve high rewards over the next h

transitions on average. In works that focus on task-agnostic exploration such as Sekar et al. [55] and Shyam et al. [56], the cost function C_e (or similar) is used to encourage the agent to find areas of the state space in which the model is maximally uncertain.

The optimization problem in (3) is typically approximately solved in one of 3 ways: Deisenroth and Rasmussen [21] and Curi et al. [16] directly backpropagate through the estimated dynamics and reward functions to find policy parameters that would generate good actions, Janner et al. [31] use an actor-critic method trained via rollouts in the model alongside the data collected to find a policy, and Chua et al. [14] and Mehta et al. [38] use the cross-entropy method [17] to find action sequences which directly maximize the reward over the estimated dynamics. In this work we use a version of the last method given in Pinneri et al. [47] to directly find action sequences that optimize the cost function being used. We approximate the expectation by playing the actions on multiple samples from the posterior $P(T \mid D)$. Algorithm 1 gives a formal description of the method and Section A.5 provides further details.

Algorithm 1 Bayesian Model-Predictive Control with Cost Function C

Inputs: transition function episode query budget b, number of posterior function samples k, planning horizon h
Initialize $D \leftarrow \emptyset$.
for $i \in [1, \ldots, b]$ **do**
  Sample start state $s_0 \sim p_0$.
  for $t \in [0, \ldots, H-1]$ **do**
    Sample posterior functions $\{T'_\ell\}_{\ell=1}^k \sim P(T' \mid D)$.
    Approximately find $\arg\min_{a_0,\ldots,a_{h-1}} \sum_{\ell=1}^k \mathbb{E}_{\tau_\ell \sim p(\tau \mid T'_\ell, a_0, \ldots, a_{h-1})} [C(\tau_\ell)]$ via CEM.
    Execute action a_0 by sampling $s_{t+1} \sim T(s_t, a_0)$.
    Update dataset $D \leftarrow D \cup \{(s_t, a_0, s_{t+1}\}$.
  end for
end for
return π_g for the posterior $P(T' \mid D)$.

4.2 A Task-Specific Cost Function based on Trajectory Information

In this work, we aim to explore by choosing actions that maximize the conditional expected information gain (EIG) about the optimal trajectory τ^*. This is the same overall goal as that of Mehta et al. [38], where the EIG_{τ^*} acquisition function was introduced for this purpose. However, in this paper we generalize this acquisition function in order to allow for sequential information collection that accounts for the redundant information that could be collected between timesteps. As discussed at length in Osband et al. [43], it is essential to reason about how an action taken at the current timestep will affect the possibility of learning something useful in future timesteps. In other words, exploration must be *deep* and not greedy. Explicit examples are given in Osband et al. [43] where the time to find an ϵ-optimal policy in a tabular MDP is exponential in the state size unless exploration can be coordinated over large numbers of timesteps rather than being conducted independently at each action. As the EIG_{τ^*} acquisition function is only defined over a single state-action pair and mutual information is submodular, we cannot naively use the acquisition function as is (or sum it over many datapoints) to choose actions that lead to good long-term exploration. As an illustrative example, this is clear in navigation tasks, where the nearby points visited over trajectories will provide redundant information about the local environment.

We therefore give a cost function that generalizes EIG_{τ^*} by taking a set of points to query and computing the *joint* expected information gain from observing the set. Our cost function is non-Markovian in the state space of the MDP, but it is Markovian in the dataset, which represents a point in the belief space of the agent about the dynamics. Let $\tilde{\mathcal{S}} = \{x : x \subseteq \mathcal{S} \times \mathcal{A}, |x| < \infty\}$, the set of finite subsets of the set of all state-action pairs. Our cost function $C_{\tau^*} : \tilde{\mathcal{S}} \to \mathbb{R}$ is defined below to be the negative *joint expected information gain* about the optimal trajectory τ^* for a subset $S \in \tilde{\mathcal{S}}$. In particular, assuming an existing dataset D, a set of h query points $S = \{(s_i, a_i)\}_{i \in [h]}$, and a random set of next states $S' = \{s'_i \sim T(s_i, a_i), i \in [h]\}$,

$$C_{\tau^*}(S) = \mathbb{E}_{S' \sim p(S' \mid S, D)} \left[\mathbb{H}\left[\tau^* \mid D \cup S'\right]\right] - \mathbb{H}\left[\tau^* \mid D\right]. \tag{5}$$

This formulation of C_{τ^*} forces our method to handle the redundant information among queries—it is likely that there will be some information about τ^* that is duplicated between members of the set and

our method of computing this cost function must avoid 'double-counting' it. However, as written, this function relies on computing entropies on high-dimensional trajectories where the form of the joint distribution of the elements is unknown. To tractably estimate this quantity, we use the fact that $C_{\tau^*}(S) = -I(S', \tau^*) = -I(\tau^*, S')$ for the mutual information I. This allows us to exchange τ^* and our set of queries so that τ^* is giving information about the posterior predictive distribution of our set. In other words,

$$C_{\tau^*}(S) = \mathbb{E}_{\tau^* \sim p(\tau^* | D)} \left[\mathbb{H}\left[S' \mid D \cup \tau^*\right]\right] - \mathbb{H}\left[S' \mid D\right]. \tag{6}$$

In order to compute the right-hand term, we must take samples $\tau^*_{ij} \sim P(\tau^* \mid D), i = 1, \ldots, m, j = 1, \ldots, n$. To do this, we first sample m start states s_0^i from p_0 (we always set $m = 1$ in experiments but derive the procedure in general) and for each start state independently sample n posterior functions $T'_{ij} \sim P(T' \mid D)$ from our posterior over dynamics models. We then run a planning procedure using iCEM [47] on each of the posterior functions from s_0^i using T'_{ij} for T (using our assumption that planning can generate approximately optimal trajectories given ground-truth dynamics), giving our sampled τ^*_{ij}. Formally, we can approximate C_{τ^*} via Monte-Carlo as

$$C_{\tau^*}(S) \approx \frac{1}{mn} \left(\sum_{i \in [m]} \sum_{j \in [n]} \mathbb{H}[S' | D \cup \tau^*_{ij}] \right) - \mathbb{H}[S' \mid D]. \tag{7}$$

Assuming the dynamics are modelled with a Gaussian process, we can compute the joint Gaussian probability of the next states S' [48]. As the entropy of a multivariate Gaussian depends only on the log-determinant of the covariance, $\log |\Sigma|$, we can tractably compute the joint entropy of the model predictions $\mathbb{H}\left[S' \mid D\right]$ and optimize it with a zeroth order optimization algorithm. Finally, we must calculate the entropy $\mathbb{H}[S' | D \cup \tau^*_{ij}]$. For this, we follow a similar strategy as Neiswanger et al. [40]: since τ^*_i is a set of states output from the transition model, we can treat them as additional noiseless datapoints for our dynamics model and condition on them before computing the joint covariance matrix for S'. Given this newly generalized acquisition function, we can instantiate a method of planning in order to maximize future information gained. We give the concrete procedure for computing our acquisition function in Algorithm 2, noting that trajectories τ^*_{ij} do not depend on the query set S and can be cached for various values of S as long as the dataset D does not change.

Our ultimate procedure, which we name *Trajectory Information Planning* (TIP), is quite simple: run model-based RL using MPC as in Algorithm 1 but set the cost function to be $C_{\tau^*}(\tau)$ instead of C_g or C_e and compute this cost function using Algorithm 2. At test time, we simply return to planning with C_g as the cost function and greedily attempt to maximize returns over exploration. We can also formulate an open-loop variant of our method, oTIP, which simply involves planning once and then executing the entire action sequence.

Algorithm 2 Computation of C_{τ^*}

Inputs: dataset $D = \{(s_k, a_k, s'_k)\}$, query set S, number of start state samples n, number of posterior function samples m.
Sample n start states $\{s_0^{(i)}\}_{i=1}^n \sim p_0$.
for $i \in [n]$ **do**
  Sample m posterior functions $\{T'_j\}_{j=1}^m \sim P(T' \mid D)$.
  for $j \in [m]$ **do**
    Set $\pi^*_j \leftarrow \pi_{\text{MPC}}$ using C_g and a singleton posterior $P(T \mid D) = \delta(T'_j)$ as in (3).
    Compute τ^*_{ij} by executing π^*_j on T'_j starting from $s_0^{(i)}$.
  end for
end for
Compute the joint posterior covariance $\Sigma^{S'} \mid D$ across all points in S.
Compute the joint posterior covariances $\Sigma^{S'}_{ij} \mid D \cup \tau_{ij} \, \forall i \in [n], j \in [m]$ across all points in S.
return $\log |\Sigma^{S'}| - \frac{1}{nm} \sum_{i \in [n], j \in [m]} \log |\Sigma^{S'}_{ij}|$.

4.3 Computational Cost and Implementation Details

Though the TIP algorithm is designed for settings where samples are expensive, it is important to understand, both theoretically and practically, the computational cost of this method. For ease of

notation, we make the simplifying assumption that the planning algorithm used (in this case, iCEM from [47]) evaluates p action sequences consisting of h (the planning horizon) actions and that our current dataset is of size N. In order to efficiently sample functions from the posterior over dynamics functions, we use the method from Wilson et al. [61]. This reduces the naive complexity of querying these functions from $O(N^3)$ to a one-time $O(N)$ cost and then $O(1)$ for additional queries. As we derive in Section A.1, the computational complexity of one TIP planning iteration is $O\left(nm\left((N+H)^3 + ph\,(N+H)^2\right)\right)$. The two asymptotically expensive operations are (1) computing the Cholesky decompositions of the nm kernel matrices for datasets $D \cup \tau^*_{ij}$ and (2) solving the triangular systems using the cached Cholesky decompositions in order to compute the covariance matrices $\Sigma^{S'}_{ij} \mid D \cup \tau^*_{ij}$ for each of the p action sequences used by the planning algorithm.

However, our implementation choices mean that in practice these operations are not the most expensive step. The covariance matrix computations mentioned above as the theoretical bottleneck are implemented in JAX [8], allowing them to be compiled to much faster machine code and vectorized across large batches of queries. In fact, the most expensive operation in practice is planning on the sampled transition functions T'_i to sample optimal trajectories τ^*_{ij}. This is due to the fact that in practice p is large and we implemented the planner in NumPy [27] so it cannot be compiled together with the Tensorflow [1] code from Wilson et al. [61], which is used for predicting which states will be visited for the planner. We give further information on the implementation in Section A.

5 Experiments

The aim of our development of the TIP algorithm and the C_{τ^*} acquisition function for RL is to reduce the sample complexity of learning effective policies in continuous MDPs given limited access to expensive dynamics. In this section we demonstrate the effectiveness of TIP in quickly learning a good policy by comparing against a variety of state-of-the-art reinforcement learning algorithms and strong baselines (including some that use the TQRL setting from [38], which is also known as RL with a generative model in Kakade [32] and other works [6, 4]). In particular, we compare the average return across five evaluation episodes across five random seeds of each algorithm on 5 closed-loop control problems. We also assess the median amount of data taken by each algorithm to 'solve' the problem across five seeds, which we take to mean performing as well as an MPC controller using the ground truth dynamics.

We evaluate the open-loop variant of our method, oTIP, against three comparison methods on three control problems we see as suitable for open-loop control. In particular, to be suitable for open-loop control, the problem cannot be dynamically unstable (as Pendulum and Cartpole famously are) and must have a relatively short control horizon and fixed start state. Here too, we assess the average return as open-loop trials are conducted as well as the number of timesteps required to achieve 'solved' performance.

	Greedy (C_g) w/ Stochasticity	Task-agnostic (C_e) Exploration	Task-specific (C_τ) Exploration
TQRL	N/A	EIG$_\tau$	BARL
Rollout (Sum)	MPC, PETS, BPTT, PPO, SAC, TD3	sDIP	sTIP
Rollout (Joint)		DIP	TIP, TS, HUCRL

Figure 2: Our comparison methods can be broken down by the type of cost function used and how the methods do or do not handle sequential acquisition of information. As C_g is a sum, it naturally handles future timesteps jointly. For the other informational quantities, it is possible to upper-bound information acquired by summing each separate mutual information or to compute them jointly.

Comparison Methods We use several model-based and model-free comparison methods in this work. We compare to several published model-based methods. These include **PETS** [14], as implemented by Pineda et al. [46], which uses a probabilistic ensemble of neural networks and CEM over particle samples to do MPC. We also compare against three model-based techniques from the HUCRL [15] implementation: **HUCRL** itself, which relies on hallucinating dynamics perturbations as a way of realizing an upper confidence bound on the policy, model-based Thompson Sampling (**TS**), which samples from the posterior over models and chooses optimal actions for that sample, and a greedy model-based neural network method relying on backpropagation through time (**BPTT**). Our model-free comparison methods are Soft Actor-Critic (**SAC**) [26], which is an actor-critic method that uses an entropy bonus over the policy to encourage more exploration, **TD3** [23], which is a policy gradient method method

Environment	TIP	sTIP	DIP	MPC	PETS	SAC	TD3	PPO	HUCRL	TS	BARL	EIG_T
Pendulum	**21**	36	36	46	5.6k	7k	26k	14k	>50k	>50k	**21**	56
Cartpole	131	141	161	201	1.63k	32k	18k	>1M	>6k	>6k	**111**	121
β Tracking	**46**	76	276	76	330	12k	17k	39k	480	420	186	>1k
β + Rotation	**201**	>500	>500	>500	400	30k	>50k	>50k	>5k	>5k	>500	>1k
Reacher	**251**	>400	>1k	751	700	23k	13k	>100k	6.6k	4.5k	**251**	>1.5k

Table 1: **Sample Complexity:** Median number of samples across 5 seeds required to reach 'solved' performance, averaged across 5 trials. We determine 'solved' performance by running an MPC policy (similar to the one used for evaluation) on the ground truth dynamics to predict actions. We record $> n$ when the median run is unable to solve the problem by the end of training after collecting n datapoints. The methods in the rightmost section operate in the TQRL setting and therefore have more flexible access to the MDP dynamics for data collection. The full set of methods are shown in Section D.

that includes tricks such as a doubled critic function and delayed gradients to solve stability issues in learning, and Proximal Policy Optimization (**PPO**) [54], which accounts for the areas in policy space where the critic could be confident using the Fisher information. SAC, TD3, and PPO explore by maximizing long-term reward with stochasticity, though they are not explicitly Bayesian.

Finally, we compare against various ablations of the proposed method. These vary across two axes as described in Figure 2: the cost function they use and how they handle sequential queries. Besides these differences, they use the same GP model and iCEM planning algorithm with the same hyperparameters, so they are truly comparable methods. The three cost functions used are C_g, C_{τ^*}, and C_e. **DIP**, **oDIP**, and $\textbf{EIG}_T$ all use C_e but compute, respectively, the expected joint entropy of the action sequence, the sum of the pointwise entropies of the action sequence, and the individual pointwise entropies in the TQRL setting. **MPC** uses C_g and is very close to the method in [33]. Like TIP, **BARL** [38] and **sTIP** use C_{τ^*}. BARL operates in the TQRL setting and can therefore use the simpler EIG_{τ^*} acquisition function. sTIP investigates the use of $-\sum_{s_i,a_i \in S} \text{EIG}_{\tau^*}(s_i, a_i)$ as a cost function for planning. This computes individual information gains for each future observation without accounting for the information they may have in common and is therefore an overestimate of the joint information gain. In the open-loop setting we compare against **oDIP** and **oMPC**, the open-loop variants of DIP and MPC, and Bayesian optimization (**BO**) as implemented by Pedregosa et al. [45]. oDIP simply plans an action sequence to minimize the joint formulation of C_e and executes the actions found for each open-loop trial. oMPC does the same thing using C_g.

Control Problems Our closed loop control problems are the standard underactuated **Pendulum** swing-up task (Pendulum-v0 from Brockman et al. [9]) with 2D states and 1D actions, a **Cartpole** swing-up task with a sparse reward function, 2D s, and 1D actions, a 2-DOF robot arm control problem where the end effector is moved to a goal (**Reacher**-v2 from Brockman et al. [9]) with 10D states and 2D actions, a simplified β **Tracking** problem from plasma control [12, 37] with 4D states and 1D actions, and a more complicated problem in plasma control where β **+ Rotation** are tracked with 10D states and 2D actions. Our open loop control problems are a navigation problem with hazards (**Lava Path**) from [38] and two regulation problems with different **Nonlinear Gain** functions. The Lava Path problem has 4D states and 2D actions and the nonlinear regulation problems have 2D states and 2D actions.

Results As can be seen in Table 1, TIP is able to reach solved performance more quickly across the board than the model-based and model-free external baselines, often using a fraction or even orders of magnitude less data than other methods. For many of our ablation methods we see failures to solve some of the problems even though the model is demonstrated by TIP to be able to sufficiently predict the dynamics. This is especially apparent on the harder plasma control environment, β+Rotation, where TIP is the only method using our GP which is able to solve the problem. We believe that this is because the data acquired through exploration by the ablation methods is less useful for control than the data TIP collects. This is underscored by the second column of Figure 3, where it is clear that TIP achieves the lowest modeling error on the

Environment	oTIP	oMPC	oDIP	BO
Nonlinear Gain 1	**41**	91	51	210
Nonlinear Gain 2	**51**	61	>200	60
Lava Path	**41**	101	101	>2k

Table 2: **Open Loop Sample Complexity:** Median number of samples across 5 seeds required to reach 'solved' performance, averaged across 5 trials. We determine 'solved' performance by running an MPC policy (similar to the one used for evaluation) on the ground truth dynamics to predict actions. We record $> n$ when the median run is unable to solve the problem by the end of training after collecting n datapoints.

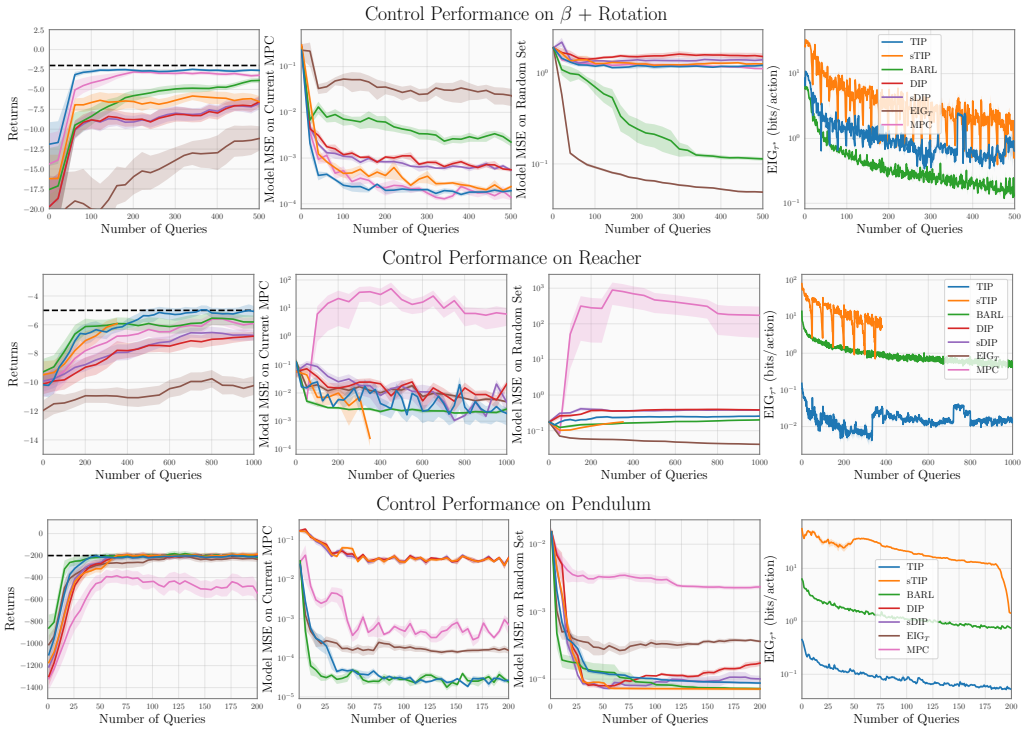

Figure 3: **Control and Modeling Details for TIP and Ablations.** Column 1: Learning curves for our ablation methods, all of which use the same planner and model. Column 2: Dynamics model accuracy on the points used by the planner to choose actions during MPC. Column 3: Dynamics model accuracy on a uniformly random test set in $\tilde{\mathcal{S}}$. Column 4: EIG_{τ^*} values normalized by the number of actions planned.

points actually needed during the execution of the policy but not on the uniform test set. In particular we find it interesting that TIP outperforms BARL on the β + Rotation environment, as BARL should in principle have a strictly stronger access to the problem and is optimizing the same quantity with fewer constraints. We hypothesis that this may be due to the fact that BARL optimizes the acquisition function EIG_{τ^*} by simply uniformly sampling a set of points and choosing the one that evaluates to the largest value. Our more sophisticated optimization algorithm and forced initialization at the start state distribution seems to allow us to collect more information in this case. This interpretation is bolstered by the fact that on the problems where TIP outperforms BARL, we see that TIP is actually collecting more information per action than BARL as evidenced by larger EIG values. We also see clearly that there is value in computing the C_{τ^*} function rather than summing over EIG_{τ^*} values, as TIP outperforms sTIP across the board. Additionally, there is clear evidence for the value of task-specific exploration as the task-agnostic exploration methods (EIG_T, DIP, sDIP) underperform both in terms of returns and model error on the trajectories visited.

For the open-loop experiments (Table 2), we see similarly strong performance from oTIP. As the model-based methods benefit hugely from the fact that they observe many model transitions for each open-loop trial, it is unsurprising that they are more sample-efficient than the model-free BO method. Within the model-based techniques, oTIP is the most sample efficient. We believe that this is for much the same reasons as in the closed-loop case—exciting evidence that the C_{τ^*} cost function can be applied in a variety of settings.

6 Conclusion

In this work, we presented and evaluated a cost function designed for intelligent, task-aware exploration. Using this cost function in model-predictive control allows agents to solve continuous MDPs with far less data than comparison methods in open- and closed-loop settings. Though the method is effective in data reduction, it is computationally expensive and relies on dynamics that are well-modeled by a GP. In future work, we aim to scale the method to higher-dimensional and more complex control environments. We also aim to apply this method in the real world. In particular, we aim to address similar plasma control problems in a small number of trials on a real tokamak.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

[2] Joseph Abbate, R Conlin, and E Kolemen. Data-driven profile prediction for diii-d. *Nuclear Fusion*, 61(4):046027, 2021.

[3] Jan Achterhold and Joerg Stueckler. Explore the context: Optimal data collection for context-conditional dynamics models. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3529–3537. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/achterhold21a.html.

[4] Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 67–83. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/agarwal20b.html.

[5] Jordan T. Ash, Cyril Zhang, Surbhi Goel, Akshay Krishnamurthy, and Sham M. Kakade. Anti-concentrated confidence bonuses for scalable exploration. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RXQ-FPbQYVn.

[6] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91 (3):325–349, 2013.

[7] Philip Ball, Jack Parker-Holder, Aldo Pacchiano, Krzysztof Choromanski, and Stephen Roberts. Ready policy one: World building through active learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 591–601. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/ball20a.html.

[8] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

[9] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[10] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1lJJnR5Ym.

[11] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[12] Ian Char, Youngseog Chung, Willie Neiswanger, Kirthevasan Kandasamy, Andrew O Nelson, Mark Boyer, Egemen Kolemen, and Jeff Schneider. Offline contextual bayesian optimization. *Advances in Neural Information Processing Systems*, 32:4627–4638, 2019.

[13] Richard Y. Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. UCB and infogain exploration via q-ensembles. *CoRR*, abs/1706.01502, 2017. URL http://arxiv.org/abs/1706.01502.

[14] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran

Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/3de568f8597b94bda53149c7d7f5958c-Paper.pdf.

[15] Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/a36b598abb934e4528412e5a2127b931-Abstract.html.

[16] Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. *Advances in Neural Information Processing Systems*, 33, 2020.

[17] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[18] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian q-learning. In *Aaai/iaai*, pages 761–768, 1998.

[19] Richard Dearden, Nir Friedman, and David Andre. Model-based bayesian exploration. *CoRR*, abs/1301.6690, 1999. URL http://arxiv.org/abs/1301.6690.

[20] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897): 414–419, 2022.

[21] Marc Peter Deisenroth and Carl Edward Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 465–472, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

[22] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

[23] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/fujimoto18a.html.

[24] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *CoRR*, abs/1609.04436, 2016. URL http://arxiv.org/abs/1609.04436.

[25] Arthur Guez, David Silver, and Peter Dayan. Efficient bayes-adaptive reinforcement learning using sample-based search. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1025–1033, Red Hook, NY, USA, 2012. Curran Associates Inc.

[26] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[27] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[28] Philipp Hennig and Christian J Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(6), 2012.

[29] José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, 27, 2014.

[30] David Humphreys, G Ambrosino, Peter de Vries, Federico Felici, Sun H Kim, Gary Jackson, A Kallenbach, Egemen Kolemen, J Lister, D Moreau, et al. Novel aspects of plasma control in iter. *Physics of Plasmas*, 22(2):021806, 2015.

[31] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

[33] Sanket Kamthe and Marc Deisenroth. Data-efficient reinforcement learning with probabilistic model predictive control. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1701–1710. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/kamthe18a.html.

[34] J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pages 513–520, 2009.

[35] Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International Conference on Machine Learning*, pages 6131–6141. PMLR, 2021.

[36] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[37] Viraj Mehta, Ian Char, Willie Neiswanger, Youngseog Chung, Andrew Nelson, Mark Boyer, Egemen Kolemen, and Jeff Schneider. Neural dynamical systems: Balancing structure and flexibility in physical prediction. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3735–3742. IEEE, 2021.

[38] Viraj Mehta, Biswajit Paria, Jeff Schneider, Willie Neiswanger, and Stefano Ermon. An experimental design perspective on model-based reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=0no8Motr-zO.

[39] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[40] Willie Neiswanger, Ke Alexander Wang, and Stefano Ermon. Bayesian algorithm execution: Estimating computable properties of black-box functions using mutual information. In *International Conference on Machine Learning*. PMLR, 2021.

[41] Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Byx83s09Km.

[42] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf.

[43] Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. URL http://jmlr.org/papers/v20/18-339.html.

[44] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[46] Luis Pineda, Brandon Amos, Amy Zhang, Nathan O. Lambert, and Roberto Calandra. Mbrl-lib: A modular library for model-based reinforcement learning. *Arxiv*, 2021. URL https://arxiv.org/abs/2104.10159.

[47] Cristina Pinneri, Shambhuraj Sawant, Sebastian Blaes, Jan Achterhold, Joerg Stueckler, Michal Rolinek, and Georg Martius. Sample-efficient cross-entropy method for real-time planning. *arXiv preprint arXiv:2008.06389*, 2020.

[48] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[49] Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-adaptive pomdps. In *NIPS*, pages 1225–1232, 2007.

[50] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/301ad0e3bd5cb1627a2044908a42fdc2-Paper.pdf.

[51] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[52] Ilya O Ryzhov and Warren B Powell. Information collection on a graph. *Operations Research*, 59(1):188–201, 2011.

[53] Ilya O Ryzhov, Martijn RK Mes, Warren B Powell, and Gerald van den Berg. Bayesian exploration for approximate dynamic programming. *Operations research*, 67(1):198–214, 2019.

[54] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[55] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pages 8583–8592. PMLR, 2020.

[56] Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5779–5788. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/shyam19a.html.

[57] Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pages 943–950, 2000.

[58] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1998. ISBN 0-262-19398-1. URL http://www.cs.ualberta.ca/%7Esutton/book/ebook/the-book.html.

[59] Matthew Tesch, Jeff Schneider, and Howie Choset. Expensive function optimization with stochastic binary outcomes. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1283–1291, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/tesch13.html.

[60] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7153–7162, 2020.

[61] James Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Efficiently sampling functions from gaussian process posteriors. In *International Conference on Machine Learning*, pages 10292–10302. PMLR, 2020.

[62] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Assumptions were fully stated for all derivations.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All code and instructions are included in supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All training details are specified in the paper or in the code.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We show error bars computed via the standard error of the mean performance between runs.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We detail computational cost (see Sec A.2).
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] All existing assets are cited.
   (b) Did you mention the license of the assets? [Yes] All licenses are properly attributed.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] New assets are included in supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] No human data was used.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No data with personally identifiable information was used.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]