Paper Discussion

Optimistic Active Exploration of Dynamical Systems

(Sukhija et al., NeurIPS 2023 [1])

Chung-En Tsai
chtsai@student.ethz.ch

Ben Bullinger
bben@student.ethz.ch

1 Introduction

This report presents the paper of Sukhija et al. [1], which introduces *Optimistic Active Exploration* (OpAX), an algorithm combining active learning and reinforcement learning to efficiently learn accurate models of dynamical systems. The paper shows that, under certain mild conditions, in Gaussian process dynamics the epistemic uncertainty under OpAX converges to zero, i.e. OpAX succeeds in learning the true underlying dynamics of the system. Moreover, the paper shows that OpAX outperforms task-specific model-based algorithms on unseen tasks and achieves zero-shot generalization performance on par with established active learning algorithms across multiple benchmarks.

1.1 Motivation

Most model-based reinforcement learning (RL) algorithms are designed to solve a single task by maximizing a specific cumulative reward [2, 3]. While sample-efficient, this approach often yields a dynamics model that is only accurate in task-relevant regions of the state-action space, limiting its utility for new tasks. An alternative paradigm is given by active exploration [4] or reward-free RL [5]. Here, learning a globally accurate dynamics model of the underlying dynamical system is prioritized. Such a model enables the solution of arbitrary downstream control tasks without requiring further environment interactions — a property which is desirable particularly in safety-critical or sample-limited environments. This paper addresses the central challenge in this paradigm: the design of policies that interact with the system to learn its dynamics as efficiently as possible.

1.2 Problem Description

Now we formalize our problem mathematically. Assume the environment is a dynamical system

$$s_{t+1} = f^\star(s_t, a_t) + w_t, \tag{1}$$

where $f^\star : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is an unknown dynamic, $\mathcal{S} \subseteq \mathbb{R}^d$ is a continuous state space, $\mathcal{A}$ is a continuous action space, and $w_1, w_2, \dots, w_T \sim \mathcal{N}(0, \sigma^2 I)$ are i.i.d. Gaussian noise. Our goal is to construct a model $\hat{f}$ that approximates $f^\star$ and performs well for arbitrary reward functions. We use boldface symbols for vectors and vector-valued functions.

We interact with the environment in an episodic setting; that is, in each episode, we deploy a carefully designed policy in the environment and observe its trajectory over T steps. We are *not* allowed to query the unknown dynamic $f^\star$ at an arbitrary state-action pairs, since doing so can be expensive or dangerous in practice. Therefore, after N episodes, our dataset consists of N trajectories from N policies, and a model $\hat{f}$ will be constructed based on this dataset. See Figure fig. 1 for an illustration.

The above formulation is motivated by real-world control tasks, though alternative formulations are possible. For example, reward-free RL [5] studies a similar problem in the tabular setting. We will compare these approaches in Section 3.

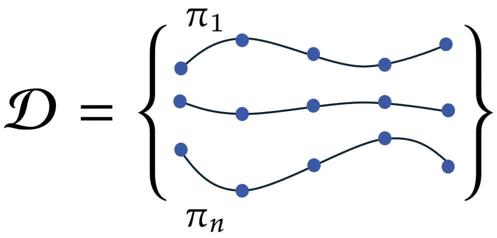


Figure 1: An illustration of the dataset

1.3 Main Contributions

The contributions of this paper are threefold.

1. This paper proposes an algorithm called OpAX, which stands for *Optimistic Active Exploration*, for solving the problem. In particular, OpAX learns a dynamical system with continuous state-action spaces. As we will see in Section 3, OpAX is a novel combination of techniques from both the active learning and RL literature.
2. From a theoretical perspective, OpAX achieves the *first* rigorous theoretical guarantees for a rich class of dynamics, and it is provably sample-efficient. Unlike reward-free RL, OpAX does not require any structural assumptions on the dynamics.
3. From a practical perspective, OpAX is computationally efficient and applicable to high-dimensional robotics tasks in real-world settings. Regarding zero-shot generalization, OpAX outperforms H-UCRL [3] across multiple tasks.

1.4 Organization

This report is organized as follows. In Section 2, we summarize the results of the paper. Then, in Section 3, we compare this paper with several related works to better position it and highlight its novelty. In Section 4, we identify a few weaknesses and limitations of the paper. Finally, in Section 5 and Section 6, we present our further theoretical and empirical investigations, respectively.

2 Summary of the Results

2.1 Well-Calibrated Assumption

The problem can be roughly divided into two phases:

- Exploration Phase: Collecting data by executing exploratory policies, and
- Estimation Phase: Constructing approximate models from the collected data.

This paper focuses on the exploration phase. To simplify the estimation phase, the paper assumes the existence of a well-calibrated model, which is used as a black box. We present a simplified version of the assumption below. Please refer to the original paper [1] for the complete version.

Assumption 1 (Well-calibrated assumption)**.** *Let $\delta \in (0,1)$. Given a dataset $\{(s_i, a_i, s_i')\}$ of size n, we can construct $(\mu_n, \sigma_n, \beta_n(\delta))$, where $\mu_n, \sigma_n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\beta_n(\delta) > 0$ such that with probability at least $1-\delta$, we have*

$$|f_j^\star(s,a) - \mu_{n,j}(s,a)| \le \beta_n(\delta)\sigma_{n,j}(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \forall j \in [d].$$

Intuitively, the well-calibrated assumption means that we can construct confidence intervals for $f^\star(s,a)$ for every state-action pair (s,a). See Figure 2 for an illustration. Furthermore, with probability at least $1-\delta$, *all* confidence intervals simultaneously cover the unknown dynamic $f^\star$. In other words, with probability at least $1-\delta$, there exists a function $\xi : \mathcal{S} \times \mathcal{A} \to [-1,1]^d$ such that

$$f_j^\star(s,a) = \mu_{n,j}(s,a) + \beta_n(\delta)\sigma_{n,j}(s,a)\xi_j(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \forall j \in [d].$$

Importantly, for every policy $\pi : \mathcal{S} \to \mathcal{A}$, it holds that

$$f_j^\star(s,\pi(s)) = \mu_{n,j}(s,\pi(s)) + \beta_n(\delta)\sigma_{n,j}(s,\pi(s))\eta_j(s), \quad \forall s \in \mathcal{S},\ \forall j \in [d]. \tag{2}$$

where $\eta(s) = \xi(s,\pi(s))$. The above equation is the key to understanding the optimism technique used in OpAX.

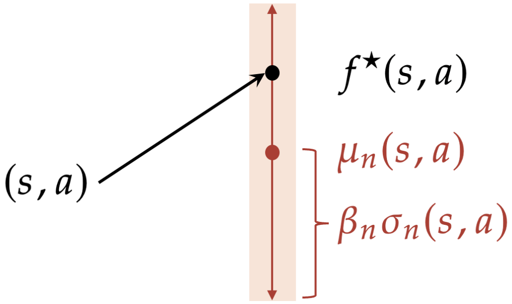


Figure 2: An illustration of the confidence interval in the well-calibration assumption

2.2 The OpAX Algorithm

This section introduces the *Optimistic Active Exploration* (OpAX) algorithm. The idea of the algorithm is to select policies that *maximize information gain* about the environment's dynamics.

2.2.1 Motivation of the OpAX Algorithm

To design such an algorithm, the authors overcome several challenges. Firstly, maximizing the expected information gain

$$\pi_n^* \in \operatorname*{argmax}_{\pi \in \Pi} \mathbb{E}_{\tau^\pi}\left[I\left(f_{\tau^\pi}^\star; \tau^\pi \mid \mathcal{D}_{1:n-1}\right)\right] \tag{3}$$

between the true dynamics $f_{\tau^\pi}^\star = (f^\star(s_0,a_0),\dots,f^\star(s_{T-1},a_{T-1}))$ along the trajectory τ^π and the data $\mathcal{D}_{1:n-1}$ from previous episodes is generally intractable. While the paper does not further discuss this intractability we provide a motivation in Section 5.1.

The intractability of the eq. (3) is overcome in the paper by establishing an upper bound (Lemma 1 in the paper):

Lemma 1 (Upper Bound on the Information Gain)**.** *For dynamics* $(s,a) \mapsto f^\star(s,a) + w$ *with Gaussian noise* $w \sim \mathcal{N}(0,\sigma^2 I)$ *and epistemic uncertainty* σ_{n-1} *after the* $(n-1)$*-th episode it holds for all* $n \geq 1$ *and dataset* $\mathcal{D}_{1:n-1}$ *that*

$$I\left(f_{\tau^\pi}^\star; \tau^\pi \mid \mathcal{D}_{1:n-1}\right) \leq \frac{1}{2}\sum_{t=0}^{T-1}\sum_{j=1}^{d}\log\left(1 + \frac{\sigma_{n-1,j}^2(s_t,a_t)}{\sigma^2}\right). \tag{4}$$

We notice that the reward function

$$r(s_t,a_t) = \frac{1}{2}\sum_{j=1}^{d}\log\left(1 + \frac{\sigma_{n-1,j}^2(s_t,a_t)}{\sigma^2}\right)$$

in the upper bound eq. (4) is now *time-separable*, contrary to the information objective in eq. (3) (cf. discussion in Section 5.1), meaning it does not depend on the history of state-action pairs preceding (s_t, a_t) in the trajectory τ^π. This makes the upper bound eq. (4) a tractable reinforcement learning objective. The optimization problem therefore becomes

$$\begin{aligned} \pi_n^* &= \underset{\pi \in \Pi}{\operatorname{argmax}} \ \mathbb{E}_{\tau^\pi} \left[\sum_{t=0}^{T-1} \sum_{j=1}^{d} \log \left(1 + \frac{\sigma_{n-1,j}^2(s_t, \pi(s_t))}{\sigma^2} \right) \right], \\ s_{t+1} &= f^*(s_t, \pi(s_t)) + w_t. \end{aligned} \tag{5}$$

But this is again intractable since the planning objective requires knowledge of the systems true dynamics $f^\star$. The paper notes that the naive solution of using the mean estimate μ_{n-1} is not desirable, e.g. since it is susceptible to model biases as shown by Chua et al. in [2].

2.2.2 Algorithm

Instead, the paper proposes to leverage the *optimism in the face of uncertainty* paradigm and the *hallucinated dynamics* introduced by Curi et al. [3]. It proposes to replace the true planning objective in eq. (5) by an optimistic planning objective based on the epistemic uncertainty σ_{n-1} and a new optimization parameter $\eta : \mathcal{S} \to [-1, 1]^d$ — the *hallucination policy*. By jointly optimizing over π and η the resulting optimization problem is given by

$$\begin{aligned} \pi_n &= \underset{\pi \in \Pi}{\operatorname{argmax}} \ \max_{\eta : \mathcal{S} \to [-1,1]^d} \ \mathbb{E}_{\tau^\pi} \left[\sum_{t=0}^{T-1} \sum_{j=1}^{d} \log \left(1 + \frac{\sigma_{n-1,j}^2(\hat{s}_t, \pi(\hat{s}_t))}{\sigma^2} \right) \right], \\ \hat{s}_{t+1} &= \mu_{n-1}(\hat{s}_t, \pi(\hat{s}_t)) + \beta_{n-1}(\delta)\sigma_{n-1}(\hat{s}_t, \pi(\hat{s}_t)) \cdot \eta(\hat{s}_t) + w_t. \end{aligned} \tag{6}$$

By the well-calibration assumption the true dynamics $f^\star$ then lie within the space of hallucinated dynamics with high probability (cf. fig. 3).

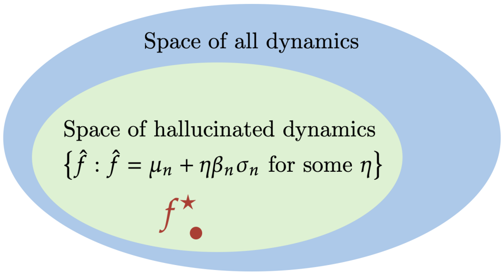

Space of all dynamics
Space of hallucinated dynamics
$\{\hat{f} : \hat{f} = \mu_n + \eta\beta_n\sigma_n$ for some $\eta\}$
$f^\star$

Figure 3: An illustration of the optimistic planning step in OpAX

The joint optimization in eq. (6) is the OpAX objective. The OpAX algorithm then simply consists of optimizing this objective, rolling out the resulting policy π_n, and updating the model, at each episode $n \in \{0, \dots, N\}$, using the well-calibrated assumption.

2.3 Main Theoretical Guarantee

In the paper, there are multiple theoretical guarantees for OpAX under different assumptions. Here we present the most concrete result, which applies when Gaussian processes are used as well-calibrated models. We will discuss this result further in Section 5.

Theorem 1 (Theorem 2 of Sukhija et al. [1]). *Assume the unknown dynamic $f^\star$ lies in a reproducible kernel Hilbert space $\mathcal{H}$ with kernel k and satisfies $\|f^\star\|_{\mathcal{H}} \leq B$. If we use Gaussian processes with kernel k as the well-calibrated models, then after N episodes, OpAX satisfies*

$$\max_{\pi} \mathbb{E} \left[\max_{0 \leq t \leq T-1} \|\sigma_N(s_t, a_t)\|^2 \middle| s_0 \sim \mu, \pi \right] \leq O \left(\beta_N(\delta) \sqrt{\gamma_N} \cdot \frac{T^{3/2}}{\sqrt{N}} \right),$$

with probability at least $1 - \delta$, where T is the number of horizon and γ_N is the maximum information gain of kernel k; see Theorem 2 of Sukhija et al. [1].

2.4 Experiments

This paper validates OpAX's effectiveness at reducing model uncertainty, generalizing to downstream tasks, and generalizing to high-dimensional problems. The algorithm is benchmarked against the random policy, other active exploration methods (the mean policy which follows the mean estimate μ_n, and PETS-AE, a trajectory sampling scheme introduced by Chua et al. [2]), H-UCRL, a task-specific method introduced by Curi et al. [3], and CEE-US, a state-of-the-art active exploration algorithm introduced by Sancaktar et al. [6].

The experiments demonstrate that OpAX significantly outperforms random exploration, generalizes better to new tasks than the specialized H-UCRL method, and generally performs on par with the other active exploration baselines, including CEE-US in complex manipulation tasks.

3 Related Work

This paper lies at the intersection of system identification, active learning, and model-based RL, so there are many related works that should be discussed. The original paper [1] has already addressed related works in system identification, active learning, and reward-free RL. In the following, we will first briefly mention the first two directions. Then, we will expand the discussion on reward-free RL based on the review-author discussion,[1]and discuss other related or follow-up works based on our own investigations.

Learning a dynamical system—also known as system identification [7] from a control perspective—has been extensively studied for the linear case, but few works focus on non-linear system identification [7]. This work provides a theoretically grounded algorithm for learning non-linear dynamics.

Inspired by the active learning literature [8, 4], OpAX is designed to maximize mutual information between the unknown dynamic and the collected data. Since most existing works study regression or classification tasks, active learning is less well understood in the context of RL.

Similar to our problem, reward-free RL [5] aims to learn a model that performs well for offline planning with arbitrary reward functions. In general, reward-free RL and the results presented in this paper are not directly comparable. This is because standard reward-free RL focuses on tabular Markov decision processes (MDP) [5, 9], which are inapplicable to our setting. When dealing with continuous state-action spaces, existing reward-free RL results typically impose structural assumptions on the transitions or value functions, such as linearity of the Q-function and the low-rank structure [10, 11, 12]. These assumptions are difficult to verify in practice. In contrast, the assumptions in this paper are inspired by the Gaussian process literature, making them more practically relevant.

Optimism in the face of uncertainty is a key principle in model-based RL [13, 14, 15, 3, 16] and multi-armed bandit [17, 18, 19, 20]. Regarding model-based RL, OpAX employs the same technique as H-UCRL [3], but unlike H-UCRL, OpAX does not maximize a specific reward function. The analyses of OpAX also borrow many ideas from those of H-UCRL. In comparison to the multi-armed bandit literature, where the unknown function is the objective, it is the dynamics in our setting. Consequently, long-term planning is required to achieve good performance. Additionally, the goal in the bandit setting is to find the maximizer, whereas our goal is to construct a model that approximates the unknown dynamics for every state-action pair. This makes our problem more challenging.

Both maximum entropy RL [21] and entropic regularized RL [22] use the entropy to encourage exploration. A natural idea to explore unknown dynamics is to deploy such exploratory policies.[2]Eysenbach and Levine [22] proved that entropic regularized RL works well when the reward functions are similar. However, in this direction, we are not aware of any work

[1] `https://openreview.net/forum?id=tLrkjK128n&noteId=D1SQLZodr8`

that provides guarantees for arbitrary reward functions. In contrast, OpAX achieves strong zero-shot performance for any bounded reward functions; see Theorem 3.

To conclude this section, we mention a few follow-up works. As et al. [23] incorporated safety constraints into OpAX. Sukhija et al. [24] proposed an algorithm similar to ε-greedy: With probability ε, instead of sampling an action uniformly at random, their algorithm selects the action that maximizes the model's uncertainty, similar to OpAX. Sukhija et al. [25] proposed a model-based RL algorithm that augments the original reward function with model's uncertainty.

4 Limitations and Potential Improvements

4.1 Weaknesses

Firstly, we believe the clarity of the theoretical results in this paper can be improved. For example, the big-O notation in Theorem 1 hides dependencies on the noise level σ, the norm bound B, and the probability δ. The guarantee also involves an unknown quantity $\beta_N(\delta)$, but no explicit bound for it is presented. This makes it difficult to understand the presented results or verify their correctness without checking the proofs. In Section 5, we will discuss and refine these results in detail.

Secondly, in the main guarantee of the paper (their Theorem 2), when the noise is σ-sub-Gaussian but not Gaussian, the guarantee of OpAX has an exponential dependence on the time horizon T. Such exponential dependence is undesirable, suggesting that OpAX may be inefficient for sub-Gaussian noise in general. The authors mentioned that this may be unavoidable without stronger assumptions, but they did not provide any justification.

Thirdly, Lemma 13 of the paper presents the zero-shot performance guarantee of OpAX. The lemma shows that a specific maximin policy achieves a good zero-shot performance guarantee (see Theorem 3). However, it is unclear whether this maximin policy can be computed efficiently. We will discuss this further in Section 5.

Furthermore, in the zero-shot generalization experiments in the paper, the aforementioned maximin policy is not used. This results in an inconsistency between the theoretical and empirical results.

4.2 Limitations

Firstly, this paper only considers settings where the noise is Gaussian or sub-Gaussian. Although Gaussian noise is important due to the central limit theorem, it would be interesting to consider more general classes of noise, such as noise with bounded variance or even adversarial noise. This could lead to more robust algorithms. Additionally, OpAX requires knowledge of the noise level σ. Can this requirement be removed?

Secondly, this paper only considers the finite-horizon episodic setting, and the guarantee scales polynomially with the time horizon T. In real-world applications, robots are often deployed for a long time. A natural direction for future work is to extend the results to the infinite-horizon discounted reward or the average reward settings.

Thirdly, the maximin policy in the zero-shot performance guarantee assumes knowledge of the reward at every state-action pair. However, knowing or designing a suitable reward function can be extremely challenging in practice. If the reward function is not known beforehand, can we still derive any theoretical guarantees for zero-shot performance of OpAX?

[2] These two directions were not discussed in the original paper, but we think they are also relevant.

5 Further Theoretical Investigations

5.1 A Motivation for the Intractability of the Information Gain Objective

The paper mentions that the information gain objective in eq. (3) is intractable but it does not expand on this claim. Given the reinforcement learning setting of the paper we believe that the following motivation is instructive: by the chain rule of mutual information the total reward of the information gain objective in eq. (3) decomposes as

$$I\left(f^\star_{\tau^\pi}; \tau^\pi \mid \mathcal{D}_{1:n-1}\right) = \sum_{t=0}^{T-1} I(f^\star_{\tau^\pi}; (s_t, a_t) \mid (s_0, a_0), \ldots, (s_{t-1}, a_{t-1}), \mathcal{D}_{1:n-1}), \tag{7}$$

and we notice that the summands on the RHS depend on the entire history $\{(s_\ell, a_\ell)\}_{\ell=0}^{t-1}$ preceding the current state-action pair (s_t, a_t). This means that the information gain objective is path-dependent (or *trajectory-dependent*), meaning that it is *not time-separable.*

We demonstrate this path-dependence in a trivial example: consider an arbitrary summand in the RHS of eq. (7)

$$I(f^\star_{\tau^\pi}; (s_t, a_t) \mid (s_0, a_0), \ldots, (s_{t-1}, a_{t-1}), \mathcal{D}_{1:n-1}).$$

If the state-action pair (s_t, a_t) is not part of the history $(s_0, a_0), \ldots, (s_{t-1}, a_{t-1})$, then some generally non-zero information gain is incurred. However, if (s_t, a_t) is part of the history, then we compute

$$\begin{aligned}
&I(f^\star_{\tau^\pi}; (s_t, a_t) \mid (s_0, a_0), \ldots, (s_{t-1}, a_{t-1}), \mathcal{D}_{1:n-1}) \\
&= H(f^\star_{\tau^\pi} \mid (s_0, a_0), \ldots, (s_{t-1}, a_{t-1}), \mathcal{D}_{1:n-1}) \\
&\quad - H(f^\star_{\tau^\pi} \mid (s_0, a_0), \ldots, (s_t, a_t), \mathcal{D}_{1:n-1}) \\
&= H(f^\star_{\tau^\pi} \mid (s_0, a_0), \ldots, (s_{t-1}, a_{t-1}), \mathcal{D}_{1:n-1}) \\
&\quad - H(f^\star_{\tau^\pi} \mid (s_0, a_0), \ldots, (s_{t-1}, a_{t-1}), \mathcal{D}_{1:n-1}) \\
&= 0,
\end{aligned}$$

i.e. a second visitation of the same state-action pair does not yield any information gain.

This path-dependence of the information gain objective is not compatible with standard optimal control methods requiring time-separability of the reward. Therefore, we can consider this objective to be intractable in reinforcement learning.

In general we note that the problem of maximizing the mutual information was shown to be NP-complete by Krause et al. [8].

5.2 An Explicit Version of Main Guarantee

This subsection discusses two issues of the main guarantee (Theorem 1). First, the dependencies of the upper bound on B, σ and δ are unclear. Second, the bound depends on the quantity $\beta_n(\delta)$, which characterizes the width of the confidence intervals constructed by the Gaussian processes. However, the authors did not provide an explicit upper bound for this quantity. In fact, such an upper bound for $\beta_N(\delta)$ exists [26]. The following theorem combines Theorem 2 of Sukhija et al. [1] and Lemma 3.6 of Rothfuss et al. [26], removing $\beta_N(\delta)$ and making the dependencies on δ, B, and δ explicit.

Theorem 2 (An explicit version of Theorem 1)**.** *Under the same assumptions of Theorem 1, after N episodes, OpAX satisfies*

$$\max_\pi \mathbb{E}\left[\max_{0\le t\le T-1} \|\sigma_N(s_t, a_t)\|^2 \middle| s_0 \sim \mu, \pi\right] \le O\left(\sigma^2 \left(\sigma\gamma_N + \sigma\sqrt{\gamma_N}\log\frac{d}{\delta} + B\sqrt{\gamma_N}\right) \cdot \frac{T^{3/2}}{\sqrt{N}}\right),$$

with probability at least $1-\delta$, where γ_N is the maximum information gain of kernel k; see Theorem 2 of Sukhija et al. [1].

Proof. By checking the proofs in Appendix A.4 of Sukhija et al. [1], we actually have

$$\max_{\pi} \mathbb{E}\left[\max_{0\le t\le T-1} \|\sigma_N(s_t,a_t)\|^2 \middle| s_0 \sim \mu, \pi\right] \le O\left(\frac{1}{\log(1+\sigma^{-2})}\beta_N(\delta)\sqrt{\gamma_N}\cdot\frac{T^{3/2}}{\sqrt{N}}\right)$$
$$\le O\left((1+\sigma^2)\beta_N(\delta)\sqrt{\gamma_N}\cdot\frac{T^{3/2}}{\sqrt{N}}\right),$$

where we use $\frac{1}{\log(1+x^{-1})} \le 1+x$ in the last inequality. Note the additional $1+\sigma^2$ term. By the proof of Lemma 3.6 of Rothfuss et al. [26],

$$\beta_N(\delta) = O\left(B + \sigma\left(\gamma_N + \log\frac{d}{\delta} + 1\right)\right). \tag{8}$$

The theorem follows by combining the above two inequalities. □

Remark 1. *Assume $\sqrt{\gamma_N} \ge \max\left(\frac{B}{\sigma}, \log\frac{d}{\delta}\right)$, which holds for common kernels [19] as $N \to \infty$. Then, the upper bound in Theorem 2 simplifies to*

$$O\left(\sigma^3\gamma_N\cdot\frac{T^{3/2}}{\sqrt{N}}\right),$$

which depends linearly on γ_N. This is different from the $\sqrt{\gamma_N}$ dependence stated in Theorem 1. We think the original statement may be misleading.

Remark 2. *Recall the well-calibrated assumption (Assumption 1): With probability at least $1-\delta$,*

$$|f^\star_j(s,a) - \mu_{N,j}(s,a)| \le \beta_N(\delta)\sigma_{N,j}(s,a), \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A},\ \forall j\in[d].$$

Suppose $\gamma_N = o(\sqrt{N})$, which holds when k is the linear or RBF kernels [19]. We can show that Theorem 1 implies $\sigma_{N,j}(s,a)\to 0$ for all reachable (s,a). However, the width of the confidence interval is actually $2\beta_N(\delta)\sigma_{N,j}(s,a)$. Thus, to show that our model is accurate, i.e., the confidence interval shrinks to a point, we need to show that $\beta_N(\delta)\sigma_{N,j}(s,a)\to 0$, not just that $\sigma_{N,j}(s,a)\to 0$. This extension is straightforward, so we omit it here.

5.3 A Simpler Proof of Zero-Shot Performance

Lemma 13 of Sukhija et al. [1] shows that OpAX achieves a rigorous zero-shot performance guarantee for arbitrary bounded reward functions. Their proof uses a contradiction argument, which we find difficult to follow. In the following, we first recall the statement and provide a simpler and more direct proof. We also make the dependencies on δ, B, and δ explicit.

Theorem 3 (Lemma 13 of Sukhija et al. [1])**.** *Consider the following RL problem with a* known *bounded reward $r : \mathcal{S}\to[0,1]$:*

$$\pi^\star \in \operatorname*{argmax}_{\pi} J_r(\pi, f^\star),$$

where $J_r(\pi,f)$ is the expected cumulative return following a dynamic f:

$$J_r(\pi,f) := \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t,a_t) \middle| s_0\sim\mu, s_{t+1} = f(s_t,\pi(s_t)) + w_t, \pi\right].$$

Under the same assumptions in Theorem 1, for $N \ge poly(d, T, \gamma_N, B, \varepsilon^{-1}, \sigma, \log(1/\delta))$ we have

$$J_r(\pi^\star, f^\star) - J_r(\hat{\pi}, f^\star) \le \varepsilon,$$

with probability at least $1-\delta$, where

$$\hat{\pi} \in \operatorname*{argmax}_{\pi} \min_{\eta:\mathcal{S}\to[-1,1]^d} J_r(\pi, f^\eta), \tag{9}$$

where f^η is the hallucinated dynamic:

$$f^\eta_j(s,a) := \mu_{N,j}(s,a) + \beta_N(\delta)\sigma_{N,j}(s,a)\eta_j(a), \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A},\ \forall j\in[d].$$

Proof. First, by Corollary 7 and the proof of Corollary 3 of Sukhija et al. [1], with probability at least $1-\delta$, we have

$$J_r(\pi, f^\star) \le \min_\eta J_r(\pi, f^\eta) + O\left(T \cdot \frac{\sqrt{d}\beta_N(\delta)}{\sigma^2} \cdot \sum_{t=0}^{T-1} \mathbb{E}\left[\|\sigma_N(s_t, a_t)\|_2\right]\right), \quad \forall \pi.$$

We can bound the inner sum by the Cauchy-Schwarz inequality:

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\|\sigma_N(s_t, a_t)\|_2\right] \le \sqrt{T \cdot \sum_{t=0}^{T-1} \mathbb{E}\left[\|\sigma_N(s_t, a_t)\|_2^2\right]} \le T \cdot \sqrt{\max_{0 \le t \le T-1} \mathbb{E}\left[\|\sigma_N(s_t, a_t)\|_2^2\right]}.$$

Furthermore, by Theorem 2, we have

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\|\sigma_N(s_t, a_t)\|_2^2\right] \le T \cdot \sqrt{\beta_N(\delta)\sqrt{\gamma_N} \cdot \frac{\sigma^2 \mathrm{poly}(T)}{\sqrt{N}}} = \mathrm{poly}(T) \cdot \frac{\sigma \beta_N(\delta)^{1/2} \gamma_N^{1/4}}{N^{1/4}}.$$

Therefore, we conclude that

$$J_r(\pi, f^\star) \le \min_\eta J_r(\pi, f^\eta) + O\left(\mathrm{poly}(T, d) \cdot \frac{\beta_N(\delta)^{3/2} \gamma_N^{1/4}}{\sigma N^{1/4}}\right).$$

The above upper bound shows that for every policy π, its expected cumulative reward can be upper bounded by its worst-case expected cumulative reward over $\mathcal{F}_N$ with an additional $O((\gamma_N/N)^{1/4})$ term. Therefore, to maximize $J_r(\pi, f^\star)$, one can try to maximize the upper bound, which gives

$$\max_\pi J_r(\pi, f^\star) \le \max_\pi \min_\eta J_r(\pi, f^\eta) + O\left(\mathrm{poly}(T, d) \cdot \frac{\beta_N(\delta)^{3/2} \gamma_N^{1/4}}{\sigma N^{1/4}}\right).$$

In other words,

$$J_r(\pi^\star, f^\star) \le \min_\eta J_r(\hat{\pi}, f^\eta) + O\left(\mathrm{poly}(T, d) \cdot \frac{\beta_N(\delta)^{3/2} \gamma_N^{1/4}}{\sigma N^{1/4}}\right).$$

By the well-calibrated assumption (Assumption 1 and (2)), there exists η such that $f^\star(\cdot, \hat{\pi}(\cdot)) = f^\eta(\cdot, \hat{\pi}(\cdot))$, so

$$\min_\eta J_r(\hat{\pi}, f^\eta) \le J_r(\hat{\pi}, f^\star).$$

Therefore, we have $J_r(\pi^\star, f^\star) - J_r(\hat{\pi}, f^\star) \le \varepsilon$ if

$$N \ge O\left(\mathrm{poly}(T, d) \cdot \frac{\beta_N(\delta)^6 \gamma_N}{\sigma^4 \varepsilon^4}\right) = O\left(\frac{\mathrm{poly}(T, d, B, \gamma_N, \sigma, \log(1/\delta))}{\varepsilon^4}\right),$$

where the last equality follows from (8). This completes the proof. □

5.4 Computing Maximin Policy in Zero-Shot Performance Guarantee

Theorem 3 shows that the maximin policy $\hat{\pi}$ achieves a rigorous zero-shot performance guarantee. However, whether this policy can be computed in a provably efficient manner remains unclear. We explore this question in this subsection.

The definition of $\hat{\pi}$ (9) can be written as

$$\hat{\pi} \in \operatorname*{argmax}_{\pi: \mathcal{S} \to \mathcal{A}} \min_{\eta: \mathcal{S} \to [-1,1]^d} \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, \pi(s_t)) \middle| s_0 \sim \mu\right],$$

following the dynamics

$$s_{t+1} = F(s_t, \pi(s_t), \eta(s_t)) + w_t, \quad w_t \sim \mathcal{N}(0, \sigma^2 I),$$

where $F : \mathcal{S} \times \mathcal{A} \times [-1,1]^d \to \mathcal{S}$ is defined as

$$F(s,a,b) = \mu_N(s,a) + \beta_N(\delta)\sigma_N(s,a)\eta(b).$$

Observe that this is a finite-horizon two-player zero-sum Markov game, where one agent plays the policy π and the other plays the policy η. The maximin policy $\hat{\pi}$ corresponds to the Nash policy of the first player. Since both state and action spaces are continuous, policy-based methods and function approximation techniques are generally required.

Unfortunately, most existing works in two-player zero-sum Markov game focus on the tabular setting [27, 28]. Although one can discretize the state-action spaces and apply existing algorithms, this approach is computationally inefficient. When the state space is continuous, Zhao et al. [29] considered the log-linear parameterization:

$$\pi_\theta(a|s) = \frac{\exp(\langle \theta, \phi_{s,a} \rangle)}{\sum_{a' \in \mathcal{A}} \exp(\langle \theta, \phi_{s,a'} \rangle)}.$$

Their algorithm is based on natural policy gradient, and it computes an ε-approximate Nash policy with a sample complexity of $O(\varepsilon^{-6})$. Their guarantee depends on several concentrability coefficients, which we are unable to estimate in our setting. We think that this problem is currently beyond our reach and require further investigation.

6 Further Empirical Investigations

6.1 Additional Baseline: Optimistic Random Exploration (ORX)

As summarized in Section 2.4 the empirical results in the paper support that OpAX performs on par with established active exploration methods. However, the experiments in the paper allow little insight into how the individual components of OpAX influence the algorithms performance. In particular, the following question remains unanswered:

Is optimizing the hallucination policy η critical to the performance of OpAX*?*

To investigate this question we isolate the influence of the optimization over η. We achieve this by comparing OpAX to a new baseline: *Optimistic Random Exploration* (ORX).

In ORX, a hallucination direction sequence $\{\eta_t\}_{t=0}^{T-1} \subseteq [-1,1]^d$ is sampled uniformly at random from $\mathcal{U}([-1,1]^d)$ for each candidate policy π, instead of jointly optimizing η with π as in OpAX. Comparing OpAX to ORX will thus shed light on the value of the optimization over η.

6.1.1 The ORX Control Problem

We define *Optimistic Random Exploration* (ORX) by modifying the OpAX optimal control problem [1, eq. (7)] as follows:

$$\pi_n^{\text{ORX}} = \underset{\pi \in \Pi}{\operatorname{argmax}}\ \mathbb{E}_{\{\eta_t\} \sim \mathcal{U}([-1,1]^d)^{\otimes T}, \tau^{\pi,\{\eta_t\}}} \left[\sum_{t=0}^{T-1} \sum_{j=1}^{d} \log\left(1 + \frac{\sigma_{n-1,j}^2(\hat{s}_t, \pi(\hat{s}_t))}{\sigma^2}\right)\right], \tag{10}$$

where $\{\eta_t\}$ denotes $\{\eta_t\}_{t=0}^{T-1}$. The trajectories $\tau^{\pi,\eta}$ are generated using the same hallucinated dynamics as in OpAX:

$$\hat{s}_{t+1} = \mu_{n-1}(\hat{s}_t, \pi(\hat{s}_t)) + \beta_{n-1}(\delta)\sigma_{n-1}(\hat{s}_t, \pi(\hat{s}_t)) \cdot \eta_t + w_t, \tag{11}$$

where σ_{n-1} is the epistemic uncertainty of the learned model, μ_{n-1} is the mean estimator, and $w_t \sim \mathcal{N}(0, \sigma^2 I)$.

Therefore, ORX does not optimize the reward w.r.t. the hallucination policy η but samples η uniformly from $[-1,1]^d$. The ORX objective in eq. (10) therefore encourages the policy π_n^{ORX} to achieve high reward in a manner which is robust to random sampling of the hallucination direction η.

6.1.2 Algorithmic Implementation of ORX

Our implementation extends the OpAX repository (`https://github.com/lasgroup/opax`) by Sukhija et al. [1]. We publish our code as a fork in: `https://github.com/BenBullinger/optimistic-random-exploration`.

We limit our experiments to a subset of the experiments carried out in the original paper. In particular we only consider the pendulum environment from the OpenAI gym benchmark suite [30].

Analogously to the OpAX implementation by Sukhija et al. [1] we use the *improved Cross-Entropy Method* (iCEM) by Pinneri et al. [31] to approximate the solution to eq. (10).

In each iCEM iteration:

1. A set of K candidate action sequences $\{a_t^{(k)}\}_{t=0}^{T-1}$, $k \in \{0, \dots, K-1\}$ (representing policies $\{\pi_k\}_{k=0}^{K-1}$) is sampled.

2. For each candidate action sequence $\{a_t^{(k)}\}_{t=0}^{T-1}$, a corresponding hallucination direction sequence $\{\eta_t^{(k)}\}_{t=0}^{T-1}$ is independently sampled uniformly from $\mathcal{U}([-1,1]^d)^{\otimes T}$.

3. Each pair $(\pi_k, \{\eta_t^{(k)}\}_{t=0}^{T-1})$ is evaluated by rolling out trajectories using the hallucinated dynamics in eq. (11) to obtain a score based on the objective in eq. (10).

4. A subset of elite policies $\{\pi_e\}$ is selected based on these scores.

5. The sampling distribution for the action sequences is updated based on the mean and variance of the elites $\{\pi_e\}$.
   Note: The hallucination direction sequences associated with the elites are not used to update any distribution, as the new $\{\eta_t^{(k)}\}_{t=0}^{T-1}$ are again drawn uniformly at random in the subsequent iteration.

This procedure optimizes the policy π_n^{ORX} to perform well under the influence of randomly chosen hallucination directions $\{\eta_t^{(k)}\}_{t=0}^{T-1}$ during its planning phase. Analogously to OpAX, the actual action executed in the environment at each step is the first action from the optimized policy π_n^{ORX}, with the hallucination component used only for planning.

6.1.3 Experimental Results for ORX

We simulate ORX in the Pendulum-v1 environment from the OpenAI gym benchmark suite [30] by extending the implementation in the OpAX repository.

We compare ORX to both OpAX and the random policy, and we plot the reduction in maximum epistemic uncertainty σ_n in fig. 4, and the total rewards for a trained task in fig. 5a and for a downstream task in fig. 5b.

We note that ORX reduces σ_n at a similar rate to OpAX, despite replacing the explicit maximization over the hallucination policy η by a random draw of a hallucination direction sequence from $\mathcal{U}([-1,1]^d)^{\otimes T}$.

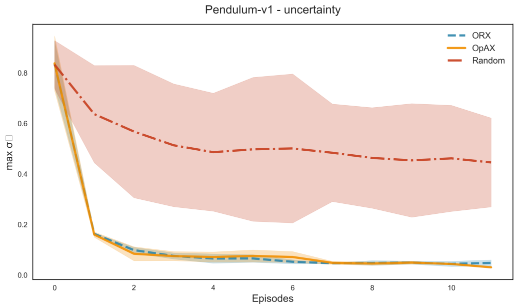
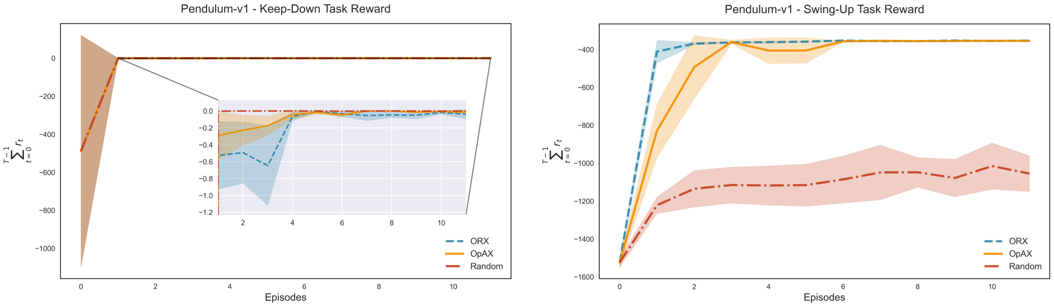


Figure 4: **Reduction in maximum epistemic uncertainty σ_n in reachable state-action space for the Pendulum-v1 environment**. We observe that ORX reduces σ_n at a similar rate to OpAX, despite selecting the hallucination policy η uniformly at random at each step. The plots are generated over three random seeds.



(a)

(b)

Figure 5: **Downstream performance for the Pendulum-v1 environment.** We apply ORX, OpAX, and the random policy to the task shown in **(a) Trained task (keep down).** The agent is rewarded for keeping the pendulum in its neutral position, and evaluate their zero-shot generalization through the downstream task shown in **(b) Zero-shot task (swing up).** The agent is rewarded for keeping the pendulum stable pointing upwards. We observe that ORX achieves *better* downstream performance than OpAX in this environment. The plots are generated over three random seeds.

6.1.4 Discussion of ORX

Given our initial experimental results for ORX we believe that it is worthwhile to consider the ORX algorithm in cases where $\dim(\mathcal{S}) \gg \dim(\mathcal{A})$. This could be the case for a robot with a complex camera-based state representation but simple motor controls. In such a scenario replacing an optimization over $\eta \in [-1, 1]^{\mathcal{S}}$ by a random draw could enable a significant speed up. However, we note that our experiments only consider the Pendulum-v1 environment where $\dim(\mathcal{S}) = 3$ and $\dim(\mathcal{A}) = 1$. It is possible that for environments with larger state and action spaces a significant difference in empirical performance between OpAX and ORX could arise.

References

[1] Bhavya Sukhija, Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic active exploration of dynamical systems. *Adv. Neural Inf. Syst.*, 2023.

[2] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Adv. Neural Inf. Process. Syst.*, 2018.

[3] Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. *Adv. Neural Inf. Process. Syst.*, 2020.

[4] Burr Settles. Active learning literature survey. 2009.

[5] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *Proc. 37th Int. Conf. Mach. Learn.*, 2020.

[6] Cansu Sancaktar, Sebastian Blaes, and Georg Martius. Curious exploration via structured world models yields zero-shot object manipulation. *Adv. Neural Inf. Process. Syst.*, 2022.

[7] Alessandro Chiuso and Gianluigi Pillonetto. System identification: A machine learning perspective. *Annu. Rev. Control Robot. Auton. Syst.*, 2019.

[8] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.*, 2008.

[9] Zihan Zhang, Simon Du, and Xiangyang Ji. Near optimal reward-free reinforcement learning. In *Proc. 38th Int. Conf. Mach. Learn.*, 2021.

[10] Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Adv. Neural Inf. Process. Syst.*, 2020.

[11] Shuang Qiu, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. On reward-free RL with kernel and neural function approximations: Single-agent MDP and Markov game. In *Proc. 38th Int. Conf. Mach. Learn.*, 2021.

[12] Jinglin Chen, Aditya Modi, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. On the statistical efficiency of reward-free exploration in non-linear RL. *Adv. Neural Inf. Process. Syst.*, 2022.

[13] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Mach. Learn.*, 2002.

[14] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 2002.

[15] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proc. 34th Int. Conf. Mach. Learn.*, 2017.

[16] Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms.* 2019.

[17] Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Stat.*, 1987.

[18] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proc. 22nd Conf. Learn. Theory*, 2009.

[19] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theory*, 2012.

[20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

[21] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *Proc. 36th Int. Conf. Mach. Learn.*, 2019.

[22] Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. In *Int. Conf. Learn. Representations*, 2022.

[23] Yarden As, Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Stelian Coros, and Andreas Krause. Actsafe: Active exploration with safety constraints for reinforcement learning. In *Int. Conf. Learn. Representations*, 2025.

[24] Bhavya Sukhija, Stelian Coros, Andreas Krause, Pieter Abbeel, and Carmelo Sferrazza. Maxinforl: Boosting exploration in reinforcement learning through information gain maximization. *arXiv preprint arXiv:2412.12098*, 2024.

[25] Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Florian Dorfler, Pieter Abbeel, and Andreas Krause. Optimism via intrinsic rewards: Scalable and principled exploration for model-based reinforcement learning. In *7th Robot Learn. Workshop*, 2025.

[26] Jonas Rothfuss, Bhavya Sukhija, Tobias Birchler, Parnian Kassraie, and Andreas Krause. Hallucinated adversarial control for conservative offline policy evaluation. In *Proc. 39th Conf.Uncertain. Artif. Intell.*, 2023.

[27] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Adv. Neural Inf. Process. Syst.*, 2020.

[28] Ahmet Alacaoglu, Luca Viano, Niao He, and Volkan Cevher. A natural actor-critic framework for zero-sum Markov games. In *Proc. 39th Int. Conf. Mach. Learn.*, 2022.

[29] Yulai Zhao, Yuandong Tian, Jason Lee, and Simon Du. Provably efficient policy optimization for two-player zero-sum markov games. In *Proc. 25th Int. Conf. Artif. Intell. Stat.*, 2022.

[30] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[31] Cristina Pinneri, Shambhuraj Sawant, Sebastian Blaes, Jan Achterhold, Joerg Stueckler, Michal Rolinek, and Georg Martius. Sample-efficient cross-entropy method for real-time planning. In *Proc. Conf. Robot. Learn.*, 2021.