OPTIMISM VIA INTRINSIC REWARDS: SCALABLE AND PRINCIPLED EXPLORATION FOR MODEL-BASED REINFORCEMENT LEARNING

Bhavya Sukhija[*,1], **Lenart Treven**[1], **Carmelo Sferrazza**[2],
Florian Dörfler[1], **Pieter Abbeel**[2], **Andreas Krause**[1]
ETH Zürich [1], UC Berkeley[2]
{sukhijab, trevenl, dorfler, krausea}@ethz.ch
{pabbeel, csferrazza}@berkeley.edu

ABSTRACT

We address the challenge of efficient exploration in model-based reinforcement learning (MBRL), where the system dynamics are unknown and the RL agent must learn directly from online interactions. We propose **O**ptimistic-**MBRL** (OMBRL), an approach based on the principle of optimism in the face of uncertainty. OMBRL learns an uncertainty-aware dynamics model and *greedily* maximizes a weighted sum of the extrinsic reward and the agent's epistemic uncertainty. Under common regularity assumptions on the system, we show that OMBRL has sublinear regret for nonlinear dynamics in the (*i*) finite-horizon, (*ii*) discounted infinite-horizon, and (*iii*) non-episodic setting. Additionally, OMBRL offers a flexible and scalable solution for principled exploration. We evaluate OMBRL on state-based and visual-control environments, where it displays favorable performance across all tasks and baselines. In hardware experiments on a dynamic RC car, OMBRL outperforms the state-of-the-art, illustrating the benefits of principled exploration for MBRL.

1 INTRODUCTION

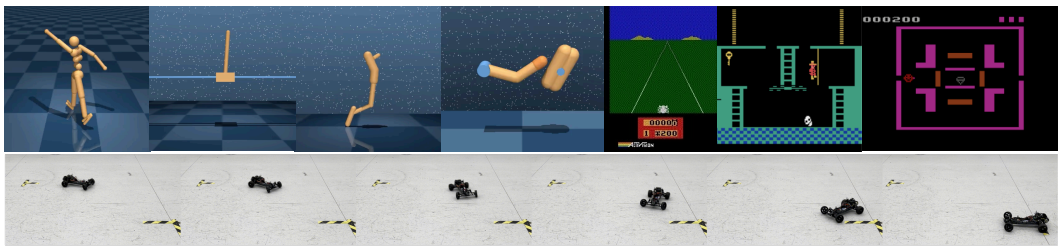

Figure 1: *Top:* We showcase scalability of the OMBRL on visual control tasks from DMC and Atari. *Bottom:* We evaluate OMBRL on a highly dynamic RC car where we learn to perform a complex parking maneuver in only 20 real-world episodes.

Reinforcement learning (RL) has been successfully applied to a variety of sequential-decision making problems like games (Silver et al., 2017), robotics (Brohan et al., 2023), mobile health interventions (Yom-Tov et al., 2017; Liao et al., 2020), and fine-tuning of large language models (Ouyang et al., 2022). RL offers a flexible learning paradigm, enabling agents to learn directly by interacting with their environment. However, this potential is often not fully realized in practice, as most widely used RL methods (Schulman et al., 2017) are highly sample-inefficient. This mostly rules out their direct application to real-world settings where data is scarce or expensive to acquire.

Model-based RL approaches (Moerland et al., 2023) offer a more sample-efficient alternative and have been successfully used for learning directly in the real-world Hansen et al. (2022); Wu et al. (2023); Rothfuss et al. (2024). However, these methods are mostly based on naive exploration strategies, such as Boltzmann exploration, which are provably sub-optimal Cesa-Bianchi et al. (2017) and often struggle in the presence of sparse rewards.

Several works study principled exploration approaches in RL (Even-Dar & Mansour (2001); Jaksch et al. (2010); Abbasi-Yadkori & Szepesvári (2011); Cohen et al. (2019); Dean et al. (2020); Kakade et al. (2020); Curi et al. (2020); Neu & Pike-Burke (2020); Eberhard et al. (2023); Wagenmaker et al. (2023); Sukhija et al. (2024c), see Section 3 and Section 7 for more details). In particular, optimism in the face of uncertainty is a celebrated exploration principle with strong theoretical guarantees for model-based RL (Brafman & Tennenholtz, 2002; Jaksch et al., 2010; Kakade et al., 2020; Curi et al., 2020; Moulin & Neu, 2023; Sukhija et al., 2024b). However, in practice, these algorithms are computationally prohibitive. As a result, naive exploration techniques remain dominant in real-world applications due to their simplicity. We address this gap between theory and practice and propose a simple yet principled method for exploration. Our approach combines the extrinsic reward from the environment with an intrinsic reward, the model epistemic uncertainty/disagreement. We show that *greedily* maximizing the weighted sum of extrinsic and intrinsic reward indeed results in optimistic exploration. Leveraging this key insight, we derive first-of-its-kind regret bounds for our approach. Our key contributions are summarized below.

Contributions

1. We propose OMBRL, a principled yet efficient exploration strategy for model-based RL. OMBRL is based on the principle of optimism in the face of uncertainty and jointly maximizes a weighted sum of the extrinsic reward and the agent's epistemic uncertainty/disagreement. Therefore, the agent selects policies that maximize rewards while also exploring less visited areas of the state space that yield high uncertainty.

2. We show that combining extrinsic rewards with the agent's epistemic uncertainty gives anytime high probability value-function bounds, which could be of independent interest to applications such as safe RL (Brunke et al., 2022) and offline RL (Levine et al., 2020). We leverage this key insight and show that OMBRL has sublinear regret for finite-horizon, discounted infinite-horizon, and nonepisodic settings with continuous state and action spaces. Our regret bounds are comparable to the ones derived by prior work (Kakade et al., 2020; Curi et al., 2020; Sukhija et al., 2024b), but our algorithm is considerably simpler and more scalable.

3. We validate OMBRL on standard deep RL benchmarks, showing that it outperforms several naive exploration baselines and scales effectively to high-dimensional tasks, such as visual control. We also demonstrate its real-world applicability by evaluating it on a dynamic RC car (see Figure 1), where it learns an agile parking maneuver in 20 trials, outperforming the state-of-the-art (Rothfuss et al., 2024) w.r.t. performance and sample efficiency. To the best of our knowledge, this is the first empirical demonstration of optimistic exploration in model-based RL for high-dimensional and real-world settings.

2 Problem Setting

We consider a discrete-time dynamical system of the form $\boldsymbol{x}_{t+1} = \boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{u}_t) + \boldsymbol{w}_t$, where $\boldsymbol{x}_t \in \mathcal{X} \subseteq \mathbb{R}^{d_{\boldsymbol{x}}}$ is the state, $\boldsymbol{u}_t \in \mathcal{U} \subseteq \mathbb{R}^{d_{\boldsymbol{u}}}$ the control input, and $\boldsymbol{w}_t \in \mathcal{W} \subseteq \mathbb{R}^{\boldsymbol{w}}$ the process noise[1]. The dynamics $\boldsymbol{f}^*$ are unknown.

Task In the finite-horizon RL setting (Puterman, 2014), we are given a reward function $r : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, and want to learn a policy that maximizes the following objective

$$J(\boldsymbol{\pi}^*) = \max_{\boldsymbol{\pi} \in \Pi} J(\boldsymbol{\pi}) = \max_{\boldsymbol{\pi} \in \Pi} \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{t=0}^{T-1} r(\boldsymbol{x}_t, \boldsymbol{u}_t) \right], \tag{1}$$

where action $\boldsymbol{u}_t$ follows policy $\boldsymbol{\pi}$, i.e., $\boldsymbol{u}_t \sim \boldsymbol{\pi}(\boldsymbol{x}_t)$. Moreover, we consider the episodic RL setting, with episodes $n \in \{1, \dots, N\}$, and study a model-based approach. Accordingly, at the beginning of episode n, we select and roll out a policy $\boldsymbol{\pi}_n$ for T steps on the true system. We then use the data collected from the rollouts to estimate the true dynamics $\boldsymbol{f}^*$. The goal is to find a policy that performs as well as $\boldsymbol{\pi}^*$, as quickly as possible. Therefore a natural performance metric in this context is the *cumulative regret* $R_N = \sum_{n=1}^{N} J(\boldsymbol{\pi}^*) - J(\boldsymbol{\pi}_n)$. In the following sections, we show that our proposed algorithm achieves sublinear regret. While in the main text for clarity we focus on the finite-horizon episodic setting, in Section 5, we show that our approach has sublinear regret also for

[1]For our theory, we assume the process noise to be known, but our algorithm can learn it from data.

1. γ-discounted infinite-horizon, episodic setting:

$$J_\gamma(\boldsymbol{\pi}^*) = \max_{\boldsymbol{\pi}\in\Pi} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r(\boldsymbol{x}_t, \boldsymbol{u}_t)\right] \tag{2}$$

2. Average reward, nonepisodic setting:

$$J_{\text{avg}}(\boldsymbol{\pi}^*) = \max_{\boldsymbol{\pi}\in\Pi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}_t, \boldsymbol{u}_t)\right] \tag{3}$$

3 Exploration Strategies in MBRL

In MBRL, we learn a model of the true dynamics $\boldsymbol{f}^*$ and use our learned model to select/update the next policy for data acquisition. Exploration algorithms for MBRL determine how the policy should be chosen given our learned model. Common strategies for this choice are (*i*) greedy planning, (*ii*) Thompson sampling, and (*iii*) optimistic exploration. We discuss these in detail below.

Let $J(\boldsymbol{\pi}, \boldsymbol{f})$ be the the expected returns under the policy $\boldsymbol{\pi}$ and dynamics $\boldsymbol{f}$, that is

$$J(\boldsymbol{\pi}, \boldsymbol{f}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}'_t, \boldsymbol{u}_t)\right]$$
$$\boldsymbol{x}'_{t+1} = \boldsymbol{f}(\boldsymbol{x}'_t, \boldsymbol{u}_t) + \boldsymbol{w}_t; \boldsymbol{x}'_0 = \boldsymbol{x}_0,$$

and $\boldsymbol{\mu}_n$ our estimate of the dynamics $\boldsymbol{f}^*$ at episode n.

Greedy planning The simplest selection strategy is to pick the policy $\boldsymbol{\pi}_n$ that maximizes the expected returns for our estimated dynamics $\boldsymbol{\mu}_n$.

$$\boldsymbol{\pi}_n^{\text{MEAN}} = \arg\max_{\boldsymbol{\pi}\in\Pi} J(\boldsymbol{\pi}, \boldsymbol{\mu}_n) \tag{4}$$

This strategy is greedy as it does not directly encourage exploration in areas where we have limited data or where our model has high uncertainty. Instead, it exploits our estimate $\boldsymbol{\mu}_n$ of the dynamics. This is the basis of methods such as those of Janner et al. (2019); Hafner et al. (2023), where exploration is induced using a stochastic policy that is optimized with an entropy bonus.

To incorporate epistemic uncertainty in our learned model and avoid overfitting to misestimated dynamics, Deisenroth & Rasmussen (2011); Chua et al. (2018); Rothfuss et al. (2024) learn a Bayesian model of $\boldsymbol{f}^*$: $p(\boldsymbol{f}|\mathcal{D}_{1:n})$. Here $\mathcal{D}_{1:n} = \cup_{i\leq n}\mathcal{D}_k$, and $\mathcal{D}_i = \{(\boldsymbol{x}_{t,i}, \boldsymbol{u}_{t,i}, \boldsymbol{x}_{t+1,i})\}_{t=0}^{T-1}$ is the data collected in episode i. The policy $\boldsymbol{\pi}_n$ is then selected as

$$\boldsymbol{\pi}_n^{\text{GREEDY}} = \arg\max_{\boldsymbol{\pi}\in\Pi} \mathbb{E}_{\boldsymbol{f}\sim p(\boldsymbol{f}|\mathcal{D}_{1:n})}[J(\boldsymbol{\pi}, \boldsymbol{f})]. \tag{5}$$

Curi et al. (2020) show that greedy planning may fail to perform well in practice, especially for difficult exploration problems (e.g., in context of action penalties).

Thompson Sampling In Thompson sampling (TS), we also learn a Bayesian model $p(\boldsymbol{f}|\mathcal{D}_{1:n})$ and pick policies by maximizing the reward under $\boldsymbol{f}$ sampled from the posterior

$$\boldsymbol{\pi}_n^{\text{TS}} = \arg\max_{\boldsymbol{\pi}\in\Pi} J(\boldsymbol{\pi}, \boldsymbol{f}),\ \ \boldsymbol{f} \sim p(\boldsymbol{f}|\mathcal{D}_{1:n}). \tag{6}$$

While TS encourages exploration in a theoretically grounded manner (Russo et al., 2018), in practice, it is often intractable to sample a function $\boldsymbol{f}$ from $p(\boldsymbol{f}|\mathcal{D}_{1:n})$.

Optimistic Exploration This strategy is based on the principle of optimism in the face of uncertainty. Optimistic exploration approaches maintain a set of *plausible dynamics models* $\mathcal{M}_n$ at each episode n, e.g., the set of functions that have a high probability w.r.t. a learned Bayesian model $p(\boldsymbol{f}|\mathcal{D}_{1:n})$. The policy is then selected according to

$$\boldsymbol{\pi}_n = \arg\max_{\boldsymbol{\pi}\in\Pi, \boldsymbol{f}\in\mathcal{M}_n} J(\boldsymbol{\pi}, \boldsymbol{f}) \tag{7}$$

There are several works that study optimistic exploration theoretically (Jaksch et al., 2010; Kakade et al., 2020; Curi et al., 2020; Treven et al., 2024; Sukhija et al., 2024b). However, optimizing $\boldsymbol{f}$ over $\mathcal{M}_n$, typically a difficult non-convex constraint, is often computationally prohibitive, restricting the application of these methods to fairly low-dimensional settings. The most efficient solvers of the optimization problem (7), to the best of our knowledge, are based on a reparametrization trick which introduces additional hallucinated controls (Curi et al., 2020). This increases the total control dimension from $d_{\boldsymbol{u}}$ to $d_{\boldsymbol{u}} + d_{\boldsymbol{x}}$, which is prohibitive in high-dimensional domains.

4 OMBRL: OPTIMISTIC-MBRL

We now present OMBRL, our approach for efficient optimistic exploration in MBRL, which alternates between two steps. First, given a dataset of transitions $\mathcal{D}_{1:n}$, we learn an uncertainty-aware model of the unknown dynamics $\boldsymbol{f}^*$. That is, after each episode n, we learn a mean estimate $\boldsymbol{\mu}_n$ of $\boldsymbol{f}^*$ and quantify our epistemic uncertainty $\boldsymbol{\sigma}_n$ over the estimate. Models such as Gaussian processes (GPs) (Rasmussen & Williams, 2005) can be directly used for this purpose. Bayesian deep learning approaches such as deep ensembles are also commonly used to quantify epistemic uncertainty or model disagreement in RL (Chua et al., 2018; Pathak et al., 2019; Curi et al., 2020; Sekar et al., 2020; Sukhija et al., 2024c). In the second step, we solve the following optimization problem for the policy $\boldsymbol{\pi}_n$

$$\begin{aligned}\boldsymbol{\pi}_n &= \arg\max_{\boldsymbol{\pi}\in\Pi} J_n(\boldsymbol{\pi}) \\ &:= \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}'_t, \boldsymbol{u}_t) + \lambda_n \|\boldsymbol{\sigma}_n(\boldsymbol{x}'_t, \boldsymbol{u}_t)\|\right] \\ \boldsymbol{x}'_{t+1} &= \boldsymbol{\mu}_n(\boldsymbol{x}'_t, \boldsymbol{u}_t) + \boldsymbol{w}_t,\end{aligned} \tag{8}$$

where λ_n is a positive constant which is used to trade off maximizing the extrinsic reward and model uncertainty (see Appendix A for how λ_n is defined in theory and Section 5.2 and Appendix C for how it is selected empirically). Note that in Equation (8), we use the mean dynamics for planning and only use the epistemic uncertainty as an additional *intrinsic* reward. Compared to the principled exploration strategies from Section 3, our approach does not require sampling from or maximizing over the dynamics. This makes OMBRL much simpler and more scalable. Moreover, OMBRL can be combined with any model-based algorithm such as those of Deisenroth & Rasmussen (2011); Janner et al. (2019); Hafner et al. (2023); Rothfuss et al. (2024). The only additional modification we make to these methods is that we add the epistemic uncertainty to the extrinsic reward. Also note that without the epistemic uncertainty reward, i.e., $\lambda_n = 0$, the agent follows the greedy strategy discussed in Section 3. Therefore, we use the model uncertainty to facilitate principled exploration for the agent.

The exploration objective in Equation (8) has also been studied by the control and deep RL community (Åström & Wittenmark, 1971; Chiuso et al., 2023; Grimaldi et al., 2024; Abeille & Lazaric, 2020; Sukhija et al., 2024a). We discuss their connection to our work in Section 7.

In the following, we show that by optimizing our objective in Equation (8), we are effectively maximizing an optimistic estimate of $J(\boldsymbol{\pi}^*)$, i.e., we are also performing optimistic exploration. Accordingly, our approach enjoys the same guarantees as other optimistic MBRL algorithms but is much simpler and computationally cheaper.

5 THEORETICAL RESULTS

For our analysis, we make some common assumptions on the underlying dynamics $\boldsymbol{f}^*$.

5.1 ASSUMPTIONS

Assumption 5.1 (Continuous closed-loop dynamics, bounded rewards, and Gaussian noise.)**.** The dynamics model $\boldsymbol{f}^*$ and all $\boldsymbol{\pi} \in \Pi$ are continuous. Furthermore, we assume that the reward is bounded, i.e., $r : \mathcal{X} \times \mathcal{U} \to [0, R_{\max}]$, and process noise is i.i.d. Gaussian[2] with variance σ^2, i.e., $\boldsymbol{w}_t \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

Next, we make an assumption that allows us to learn an uncertainty-aware model of $\boldsymbol{f}^*$ from data. More formally, we assume we learn a well-calibrated statistical model of $\boldsymbol{f}^*$ as defined in the following.

Definition 5.2 (Well-calibrated statistical model of $\boldsymbol{f}^*$, Rothfuss et al. (2023))**.** Let $\mathcal{Z} \overset{\text{def}}{=} \mathcal{X} \times \mathcal{U}$. A sequence of sets $\{\mathcal{M}_n(\delta)\}_{n\geq 0}$, where

$$\begin{aligned}\mathcal{M}_n(\delta) \overset{\text{def}}{=} \big\{ \boldsymbol{f} : \mathcal{Z} \to \mathbb{R}^{d_{\boldsymbol{x}}} \mid \forall \boldsymbol{z} \in \mathcal{Z}, \forall j \in \{1, \dots, d_{\boldsymbol{x}}\} : \\ |\mu_{n,j}(\boldsymbol{z}) - f_j(\boldsymbol{z})| \leq \beta_n(\delta)\sigma_{n,j}(\boldsymbol{z}) \big\},\end{aligned}$$

is an all-time well-calibrated statistical model of the function $\boldsymbol{f}^*$, if, with probability at least $1-\delta$, we have $\boldsymbol{f}^* \in \bigcap_{n\geq 0} \mathcal{M}_n(\delta)$. Here, f_j, $\mu_{n,j}$ and $\sigma_{n,j}$ denote the j-th element in the vector-valued

[2]For clarity of exposition we focus on the setting with Gaussian noise. In Appendix A, we also perform the analysis for the more general sub-Gaussian noise case.

functions $\boldsymbol{f}$, $\boldsymbol{\mu}_n$ and $\boldsymbol{\sigma}_n$ respectively, and $\beta_n(\delta) \in \mathbb{R}_{\geq 0}$ is a scalar function that depends on the confidence level $\delta \in (0, 1]$ and which is monotonically increasing in n.

While our theoretical guarantees can be extended to other classes of well-calibrated models, similar to Curi et al. (2020), here we focus on GPs where $\boldsymbol{\mu}_n$ and $\boldsymbol{\sigma}_n$ have a closed-form solution. Moreover, we assume that $\boldsymbol{f}^*$ resides in a Reproducing Kernel Hilbert Space (RKHS) of vector-valued functions and show that this is sufficient for us to obtain a well-calibrated model.

Assumption 5.3. We assume that the functions f_j^*, $j \in \{1, \dots, d_{\boldsymbol{x}}\}$ lie in a RKHS with kernel k and have a bounded norm B, that is $\boldsymbol{f}^* \in \mathcal{H}_{k,B}^{d_{\boldsymbol{x}}}$, with $\mathcal{H}_{k,B}^{d_{\boldsymbol{x}}} = \{\boldsymbol{f} \mid \|f_j\|_k \leq B, j = 1, \dots, d_{\boldsymbol{x}}\}$. Moreover, we assume that $k(\boldsymbol{z}, \boldsymbol{z}) \leq \sigma_{\max}$ for all $\boldsymbol{x} \in \mathcal{X}$.

Assumption 5.3 allows us to model $\boldsymbol{f}^*$ with GPs. The posterior mean $\boldsymbol{\mu}_n(\boldsymbol{z}) = [\mu_{n,j}(\boldsymbol{z})]_{j \leq d_{\boldsymbol{x}}}$ and epistemic uncertainty $\boldsymbol{\sigma}_n(\boldsymbol{z}) = [\sigma_{n,j}(\boldsymbol{z})]_{j \leq d_{\boldsymbol{x}}}$ can then be obtained using the following formula

$$\begin{aligned} \mu_{n,j}(\boldsymbol{z}) &= \boldsymbol{k}_n^\top(\boldsymbol{z})(\boldsymbol{K}_n + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}_{1:n}^j, \\ \sigma_{n,j}^2(\boldsymbol{z}) &= k(\boldsymbol{z}, \boldsymbol{z}) - \boldsymbol{k}_n^\top(\boldsymbol{z})(\boldsymbol{K}_n + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_n(\boldsymbol{z}), \end{aligned} \tag{9}$$

Here, $\boldsymbol{y}_{1:n}^j$ corresponds to the noisy measurements of f_j^*, i.e., the observed next state from the transitions dataset $\mathcal{D}_{1:n}$, $\boldsymbol{k}_n(\boldsymbol{z}) = [k(\boldsymbol{z}, \boldsymbol{z}_i)]_{\boldsymbol{z}_i \in \mathcal{D}_{1:n}}$, and $\boldsymbol{K}_n = [k(\boldsymbol{z}_i, \boldsymbol{z}_l)]_{\boldsymbol{z}_i, \boldsymbol{z}_l \in \mathcal{D}_{1:n}}$ is the data kernel matrix. The restriction on the kernel $k(\boldsymbol{z}, \boldsymbol{z}) \leq \sigma_{\max}$ implies boundedness of $\boldsymbol{f}^*$ and has also appeared in works studying the episodic setting for nonlinear dynamics (Mania et al., 2020; Kakade et al., 2020; Curi et al., 2020; Wagenmaker et al., 2023; Sukhija et al., 2024c). We can also define $\boldsymbol{f}^*$ such that $\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + \boldsymbol{f}^*(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1}) + \boldsymbol{w}_{k-1}$ in which case the boundedness of $\boldsymbol{f}^*$ captures many real-world systems.

Lemma 5.4 (Well calibrated confidence intervals for RKHS, Rothfuss et al. (2023))**.** *Let $\boldsymbol{f}^* \in \mathcal{H}_{k,B}^{d_{\boldsymbol{x}}}$. Suppose $\boldsymbol{\mu}_n$ and $\boldsymbol{\sigma}_n$ are the posterior mean and variance of a GP with kernel k, Equation (9). There exists $\beta_n(\delta)$, for which the tuple $(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \beta_n(\delta))$ is a well-calibrated statistical model of $\boldsymbol{f}^*$.*

In summary, in the RKHS setting, a GP is a well-calibrated model. Next, we present the following Proposition, which states that $J_n(\boldsymbol{\pi}_n)$ from Equation (8) is an optimistic estimate of $J(\boldsymbol{\pi}^*)$.

Proposition 5.5. *Let Assumption 5.1 and Assumption 5.3 hold. Then, there exists a $\lambda_n \in \Theta(\beta_n)$, such that we have $\forall n > 0$, $\boldsymbol{\pi} \in \Pi$, with probability at least $1 - \delta$, that $J(\boldsymbol{\pi}) \leq J_n(\boldsymbol{\pi})$. Moreover, we have $J(\boldsymbol{\pi}^*) \leq J_n(\boldsymbol{\pi}_n)$.*

Proposition 5.5 shows that for all policies $\boldsymbol{\pi} \in \Pi$, $J_n(\boldsymbol{\pi})$ gives an upperbound on the true return $J(\boldsymbol{\pi})$. This result is of independent interest and can be applied to settings beyond online RL such as safe RL (Brunke et al., 2022; As et al., 2024) and offline RL Levine et al. (2020); Yu et al. (2020); Rigter et al. (2022). The exact bound for λ_n is provided in Lemma A.1 in Appendix A.

Finally, we present our main theorem, which bounds the regret of OMBRL. Our bound depends on the *maximum information gain* of kernel k (Srinivas et al., 2012), defined as

$$\Gamma_N(k) = \max_{\mathcal{A} \subset \mathcal{X} \times \mathcal{U}; |\mathcal{A}| \leq N} \frac{1}{2} \log \left| \boldsymbol{I} + \sigma^{-2} \boldsymbol{K}_N \right|.$$

Γ_N is a measure of the complexity for learning $\boldsymbol{f}^*$ from N episodes and is sublinear for many kernels (e.g., $\mathcal{O}(\log^{d_x + d_u + 1}(N))$ for the exponential (RBF) kernel, $\mathcal{O}((d_x + d_u)\log(N))$ for the linear kernel). In Appendix A, we report the dependence of Γ_N on N in Table 1.

Theorem 5.6 (Finite horizon setting)**.** *Let Assumption 5.1 and Assumption 5.3 hold. Then we have $\forall N > 0$ with probability at least $1 - \delta$*

$$R_N \leq \mathcal{O}\left(\Gamma_N^{3/2} \sqrt{N}\right).$$

Theorem 5.6 guarantees sublinear regret for a rich class of RKHS functions. Accordingly, for many RKHS, our algorithm enjoys the same asymptotic guarantees as Kakade et al. (2020). Note that the regret bound from Kakade et al. (2020) is an order of $\sqrt{\Gamma_N}$ better. On the other hand, OMBRL is a much simpler and more scalable algorithm. In Appendix A, we show that OMBRL improves the regret bound from Curi et al. (2020) by a factor of Γ_N^T. Below, we also provide our regret bounds for the γ-discounted and the non-episodic setting.

In contrast to the finite-horizon case, where each episode has a fixed length T, in the γ-discounted case, we care about the infinite horizon. To still maintain the episodic nature of the problem, while also observing and learning the system for longer horizons, a crucial requirement for achieving sublinear regret, we let the length of episode n grow logarithmically with n, i.e., $T(n) \in \Theta(\log(n))$.

Theorem 5.7 (γ-discounted, infinite horizon setting)**.** *Let* $R_N = \sum_{n=1}^{N} J_\gamma(\boldsymbol{\pi}^*) - J_\gamma(\boldsymbol{\pi}_n)$. *Under the Assumption 5.1 and Assumption 5.3, we have for the* γ*-discounted infinite (Equation (2)) horizon setting* $\forall N > 0$ *that with probability at least* $1 - \delta$

$$R_N \leq \mathcal{O}\left(\Gamma^{3/2}_{N\log(N)}\sqrt{N}\right)$$

In Theorem 5.7, we show that even though we truncate each episode after $T(n)$ steps, OMBRL has sublinear regret w.r.t. the infinite horizon objective. Moreover, the regret for this setting follows the same structure as for the finite horizon case. To the best of our knowledge, we are the first to give a regret bound for optimistic model-based RL algorithms for the γ-discounted setting.

Finally, we give our regret bound for the non-episodic setting. In this setting, we cannot reset the agent and have to learn from a single trajectory. This is the most challenging and closest setting for learning directly in the real-world (Kakade, 2003), as resets are often prohibitive for many real-world applications (Sharma et al., 2021). Sukhija et al. (2024b) show that optimistic exploration methods have sublinear regret for the nonepisodic setting. However, their proposed algorithm is intractable in practice. We extend OMBRL to the nonepisodic case. In this setting, OMBRL also maximizes the reward together with the model epistemic uncertainty. However, unlike the episodic case, where we update our model and policy after every episode, for OMBRL we only update them once we have accumulated enough information, i.e., $\sum_{t=0}^{T_n-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}_{t,n}, \boldsymbol{\pi}_n(\boldsymbol{x}_{t,n}))\| > C$, for a positive constant C.

Theorem 5.8 (Informal statement; nonepisodic average reward case)**.** *Let* $R_N = \sum_{n=1}^{N} \mathbb{E}[J_{avg}(\boldsymbol{\pi}^*) - r(\boldsymbol{x}_n, \boldsymbol{\pi}_n(\boldsymbol{x}_n)]$. *Under the same assumptions as Sukhija et al. (2024b), we have for the average reward setting (Equation (3))* $\forall N > 0$ *that with probability at least* $1 - \delta$

$$R_N \leq \mathcal{O}\left(\Gamma^{3/2}_{N}\sqrt{N}\right).$$

In contrast to Sukhija et al. (2024b), OMBRL is much more tractable, and in Theorem 5.8 we show that OMBRL also has sublinear regret in the nonepisodic setting and therefore offers a theoretically strong and practical alternative for model-based exploration for this case.

In this section, we have shown that OMBRL offers a practical approach for exploration in MBRL and enjoys the same guarantees, i.e., sublinear regret for common kernels and RL settings, as other principled and often intractable/computationally prohibitive MBRL algorithms (Kakade et al., 2020; Curi et al., 2020; Sukhija et al., 2024b). We present additional theoretical results, for example, a sample complexity bound for pure intrinsic exploration algorithms such as Sekar et al. (2020); Buisson-Fenet et al. (2020); Sukhija et al. (2024c) and a regret bound for the sub-Gaussian noise setting in Appendix A. Our proofs are also provided in Appendix A.

5.2 Selecting λ_n in practice

The parameter λ_n controls the exploration-exploitation trade-off for OMBRL. In Appendix A we provide the theoretical bound for λ_n, however in practice, λ_n is treated as a hyperparameter. This is similar to other optimistic exploration and intrinsic exploration algorithms (Burda et al., 2018; Kakade et al., 2020; Curi et al., 2020), which also heuristically select the amount of exploration. Sukhija et al. (2024a) empirically study combining extrinsic and intrinsic rewards for model-free algorithms and propose an approach for automatically tuning the intrinsic reward coefficient, i.e., λ_n. We find their approach works well for our state-based and visual control tasks. Moreover, we describe their approach and how we choose λ_n for our experiments in Appendix C.

5.3 Application of OMBRL with GP dynamics

Finally, we empirically validate our theoretical findings for the GP case in Figure 2 and Figure 3, where we compare OMBRL to HUCRL (Curi et al., 2020), PETS (Chua et al., 2018), and greedy (mean) planning in the episodic setting. For the nonepisodic setting, we consider their nonepisodic counterparts as proposed in Sukhija et al. (2024b). We evaluate the algorithms on the Pendulum and MountainCar tasks from the OpenAI Gym benchmark (Brockman et al., 2016). From the experiments, we conclude that OMBRL performs the best across all baselines for both the episodic and the non-episodic setting. Moreover, while HUCRL and NeoRL, which explore according to Equation (7), perform better than other baselines, they are worse than OMBRL. We believe this is because of the practical challenges associated with solving the optimization problem in Equation (7).

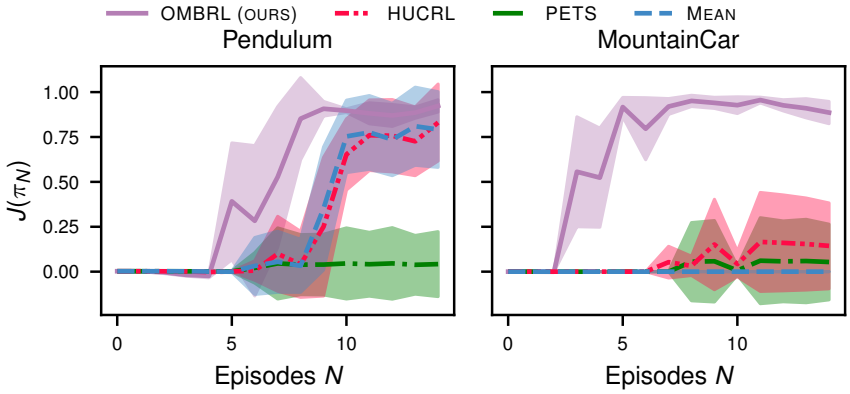

Figure 2: Learning curves for the episodic setting with GP dynamics. We report the median episode reward $J(\boldsymbol{\pi}_N)$ over an episode with 5 seeds and its standard deviation.

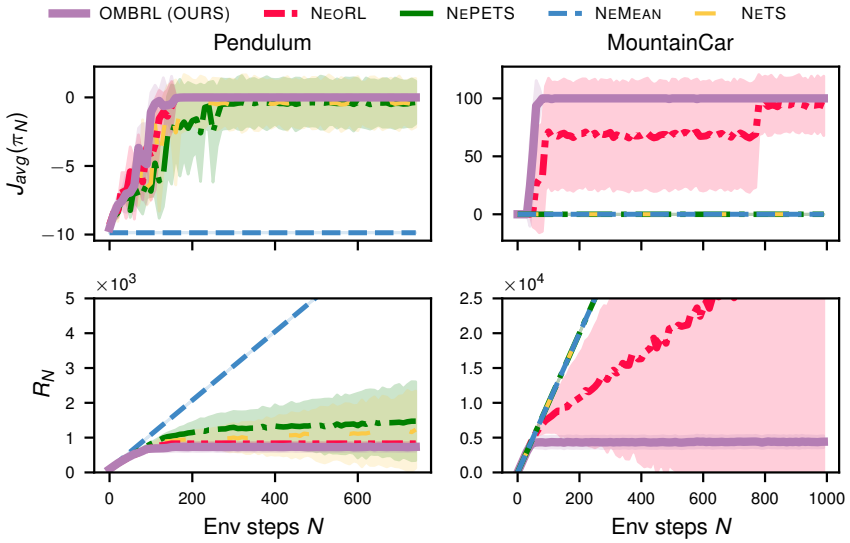

Figure 3: Learning curves for the nonepisodic setting with GP dynamics. Similar to Sukhija et al. (2024b), we report the average reward $J_{avg}(\boldsymbol{\pi}_N)$ and regret R_N. The curves are reported with 5 seeds and we plot the median return with its standard deviation.

6 EXPERIMENTS

In our experiments, we showcase the flexibility and scalability of OMBRL by combining it with three different model-based RL algorithms; (*i*) MBPO (Janner et al., 2019) for state-based tasks, DREAMER (Hafner et al., 2023) for visual control tasks, and SIMFSVGD (Rothfuss et al., 2024) for our hardware experiment on the RC car. We consider the DeepMind control (DMC) benchmark (Tassa et al., 2018) for the state-based and visual control tasks and test on environments with varying dimensionality[3]. We also evaluate on several environments from the Atari benchmark (Bellemare et al., 2013) for the visual control tasks. In all our experiments, we report the episodic returns using the median over 5 seeds along with its standard deviation. We provide additional experiment details in Appendix C.

State-based experiments We refer to the MBPO version of OMBRL as MBPO-OPTIMISTIC. The resulting algorithm operates similarly to Janner et al. (2019) and trains a policy from real and model-generated rollouts to maximize the extrinsic and intrinsic rewards. For the policy training, we use the soft actor-critic (SAC) algorithm (Haarnoja et al., 2018), and for the intrinsic reward

[3]including the humanoid from DMC: $d_{\boldsymbol{x}} = 67$, $d_{\boldsymbol{u}} = 21$

coefficient, λ_n, we use the auto-tuning approach from Sukhija et al. (2024a). We train an ensemble of dynamics models and use their disagreement to quantify the epistemic uncertainty. As baselines, we consider (*i*) MBPO-MEAN, which maximizes only the extrinsic reward, i.e., $\lambda_n = 0$, and (*ii*) MBPO-PETS, which is based on the PETS algorithm (Chua et al., 2018) maximizing the extrinsic rewards in expectation over the ensemble dynamics (see Equation (5)). We report the results on the left side of Figure 4. We conclude that across all tasks, MBPO-OPTIMISTIC performs the best. Particularly, in sparse reward tasks such as the Mountaincar and CartPole, MBPO-OPTIMISTIC successfully solves the task whereas the greedy baselines fail. MBPO-OPTIMISTIC also successfully scales to high dimensional problems such as the Quadruped and Humanoid environments. We provide additional experiments with MBPO-OPTIMISTIC in Appendix B, where we evaluate it on more environments and compare it with pure off-policy algorithms SAC and MaxInfoRL (Sukhija et al., 2024a).

Visual control experiments We investigate the scalability of OMBRL to challenging and high-dimensional problems by evaluating it on visual control tasks. We combine OMBRL with DREAMER (Hafner et al., 2023), an MBRL algorithm for visual control problems, and call the resulting algorithm DREAMER-OPTIMISTIC. We use the same approach as Sekar et al. (2020) for quantifying the epistemic uncertainty and for selecting the intrinsic reward coefficient, λ_n, we use the auto-tuning approach from Sukhija et al. (2024a). We report the results on the right side of Figure 4. Overall, DREAMER-OPTIMISTIC performs on-par with DREAMER on most tasks and outperforms it on the Finger-spin task from DMC and the Venture task from the Atari benchmark. Particularly, for Venture, a sparse reward task, Dreamer fails to achieve any reward.

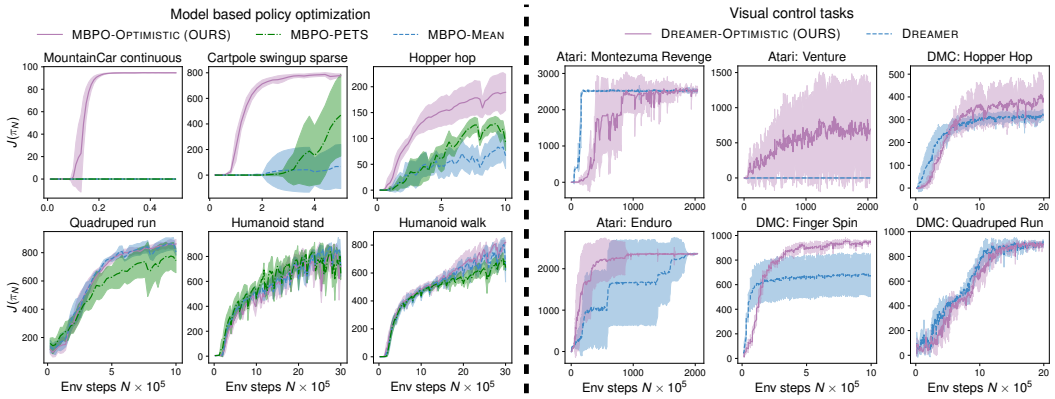

Figure 4: *Left:* Learning curves for the state-based tasks from DMC using MBPO as the base algorithm. Across all experiments, MBPO-OPTIMISTIC obtains the best performance compared to its greedy variants. MBPO-OPTIMISTIC also scales to high-dimensional tasks, specifically the humanoid environments from DMC. *Right:* Learning curves for the visual control tasks from DMC and Atari using DREAMER as the base algorithm. DREAMER-OPTIMISTIC either performs on-par or better than DREAMER in all our experiments. Particularly, in the Venture task from the Atari benchmark, where DREAMER fails to obtain any rewards.

Curi et al. (2020) study the sensitivity of greedy exploration algorithms w.r.t. the action penalties in the reward. Inspired by their experiments, we modify the reward for the CartPole and Finger spin environments by adding an action cost, $r_{\text{action}}(\boldsymbol{a}) = -K \|\boldsymbol{a}\|_2$, where K controls the penalty for large actions. Curi et al. (2020) show that even for small action costs, greedy exploration methods fail, converging to the sub-optimal solution of applying small actions. We observe a similar outcome in Figure 5 (left side), where DREAMER fails to solve the tasks for both the Finger spin and CartPole environments. On the other hand, DREAMER-OPTIMISTIC achieves much higher returns due to its optimistic exploration.

We provide additional experiments with DREAMER-OPTIMISTIC, including more environments and proprioceptive tasks in Appendix B. Even though DREAMER-OPTIMISTIC either performs on-par or better than DREAMER, in some cases it also spends more interactions exploring. This is particularly the case for the Finger turn hard and reacher hard environments (see Figure 8 in Appendix B). In essence, instead of auto-tuning λ_n as proposed by Sukhija et al. (2024a), we can select a smaller value for it to reduce the level of exploration. Overall, we believe investigating an instance-dependent schedule for λ_n is an intresting direction for future work.

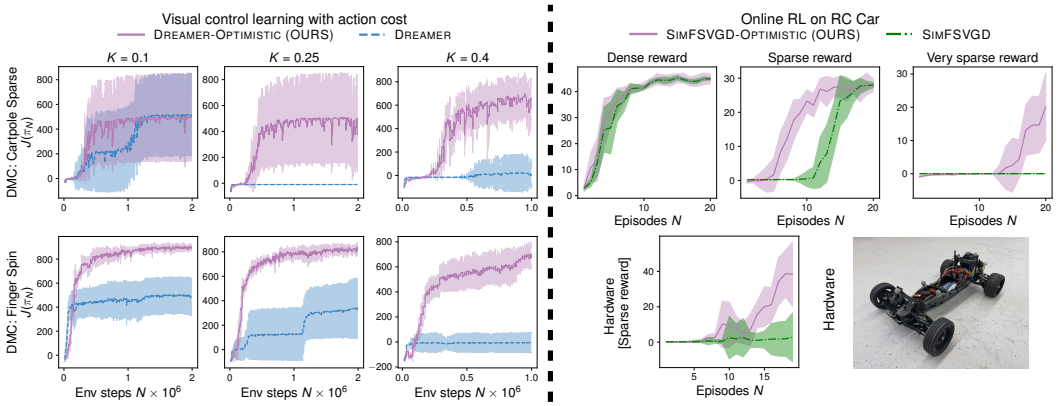

Figure 5: *Left:* Learning curves with action costs, where we compare DREAMER with DREAMER-OPTIMISTIC. DREAMER fails to explore sufficiently with action costs, whereas DREAMER-OPTIMISTIC is able to explore and obtain much higher performance. *Right:* Learning curves for our experiments with SIMFSVGD. *Top row*: We change the parameters of the reward function from Rothfuss et al. (2024), and make it sparse, starting from their dense reward. We observe that, as the reward gets sparser, SIMFSVGD drops in performance and SIMFSVGD-OPTIMISTIC outperforms it. *Bottom row*: We run the sparse reward configuration on hardware (depicted on the right side at the bottom), where we obtain similar results. As opposed to SIMFSVGD-OPTIMISTIC, SIMFSVGD fails to solve the task.

Hardware experiments Rothfuss et al. (2024) propose a novel approach for training deep Bayesian models that incorporates low-fidelity physical priors. Their approach significantly improves sample efficiency, which they illustrate in their hardware experiments on an RC car. Inspired by their experimental setup, we conduct a similar experiment on a highly dynamic RC car. The task is to perform a complex parking maneuver with drifting as depicted in Figure 1. Similar to Rothfuss et al. (2024), we use a simple bicycle model as the low-fidelity prior for SIMFSVGD. We use the same reward function structure and hyperparameters as Rothfuss et al. (2024). First, we evaluate our algorithm SIMFSVGD-OPTIMISTIC, a combination of OMBRL and SIMFSVGD, in simulation. The simulation is based on a realistic race car simulation from Kabzan et al. (2020). For the simulation experiments, we ablate different choices for the reward parameters, starting with the dense reward configuration from Rothfuss et al. (2024) and adapt parameters to obtain sparser rewards (see Appendix C for more detail). We report the results in the top row of Figure 5. We observe that, while for the dense reward setting, SIMFSVGD and SIMFSVGD-OPTIMISTIC perform similarly, for sparser rewards, SIMFSVGD-OPTIMISTIC outperforms SIMFSVGD. In particular, for the setting with very sparse rewards, SIMFSVGD completely fails to solve the task. We conduct our hardware experiments using the sparse reward configuration, and report the learning curve in the bottom row of Figure 5. In line with the simulation experiments, we also observe similar behavior on hardware. SIMFSVGD-OPTIMISTIC learns to solve the task, whereas SIMFSVGD completely fails. In fact, out of 5 attempts, SIMFSVGD worked only once, and otherwise converged to a local optimum of not moving from the starting position (see video in the supplementary material).

7 RELATED WORK

Deep model-based RL Model-based RL algorithms offer a sample-efficient solution for learning directly in the real world Hansen et al. (2022); Wu et al. (2023); Rothfuss et al. (2024). Most widely applied algorithms (Chua et al., 2018; Janner et al., 2019; Hafner et al., 2023; Hansen et al., 2023) commonly rely on naive exploration techniques such as Boltzmann exploration and differ primarily in the type of dynamics modeling and policy planners. Cesa-Bianchi et al. (2017) show that Boltzmann exploration is suboptimal even in the simplified setting of stochastic bandits. OMBRL is agnostic to the choice of modeling and planners, as we demonstrate in Section 5 and 6. Moreover, we focus on the problem of exploration for MBRL and propose a principled exploration approach. We derive regret bounds for our approach, showing that it is theoretically grounded. Furthermore, we illustrate the benefits of principled exploration in our hardware experiment, where the naive exploration baseline fails to obtain any meaningful exploration. To the best of our knowledge, we are the first to propose a

simple, flexible, scalable, and theoretically grounded approach for principled exploration and show its benefits directly in the real world.

Theoretical results for Model-based RL There are numerous works that study MBRL for linear dynamical systems theoretically (Abbasi-Yadkori & Szepesvári, 2011; Cohen et al., 2019; Simchowitz & Foster, 2020; Dean et al., 2020; Faradonbeh et al., 2020; Abeille & Lazaric, 2020; Treven et al., 2021), focusing primarily on the challenges of nonepisodic learning. In the nonlinear case, Kakade et al. (2020); Curi et al. (2020); Mania et al. (2020); Wagenmaker et al. (2023); Treven et al. (2024) analyze the finite-horizon episodic setting and provide regret bounds that are sublinear for many RKHS. Recently, Sukhija et al. (2024b) extended these results to the nonepisodic setting. Crucially, most of these algorithms are based on the principle of optimism in the face of uncertainty and require solving the problem in Equation (7). As highlighted in Section 1 and 3, solving these problems is often intractable or computationally expensive. Therefore, naive exploration techniques, such as Boltzmann exploration are more widely used. OMBRL addresses this drawback and proposes an alternative optimistic exploration method, which is much simpler and more scalable. Furthermore, it enjoys the same asymptotic guarantees as these methods and hence is also theoretically grounded.

Intrinsic exploration in RL Intrinsic rewards are often used as a surrogate objective for principled exploration in challenging tasks (see Aubret et al., 2023, for a comprehensive survey). Common choices of intrinsic rewards are model prediction error or "Curiosity" (Schmidhuber, 1991; Pathak et al., 2017; Burda et al., 2018), novelty of transitions/state-visitation counts (Stadie et al., 2015; Bellemare et al., 2016), diversity of skills/goals (Eysenbach et al., 2018; Sharma et al., 2019; Nair et al., 2018; Pong et al., 2019), empowerment (Klyubin et al., 2005; Salge et al., 2014), and information gain of the dynamics (Sekar et al., 2020; Mendonca et al., 2021; Sukhija et al., 2024c). However, these rewards are mostly used for pure exploration and rarely considered in combination with the extrinsic reward. We show that combining the model epistemic uncertainty, an intrinsic reward, with the extrinsic one, effectively performs optimistic exploration, thus, providing a theoretical grounding for our approach. There are a few works from bandits (Auer, 2002; Srinivas et al., 2012), data-driven control (Åström & Wittenmark, 1971; Chiuso et al., 2023; Grimaldi et al., 2024), and RL (Abeille & Lazaric, 2020; Sukhija et al., 2024a) that have also proposed maximizing extrinsic rewards jointly with epistemic uncertainty. The data-driven control community refers to this as the separation principle between model identification and control design (Åström & Wittenmark, 1971; Chiuso et al., 2023; Grimaldi et al., 2024). In RL, Abeille & Lazaric (2020) show duality between Equation (7) and Equation (8) for linear systems. For nonlinear systems and deep RL, Sukhija et al. (2024a) empirically study combining extrinsic and intrinsic rewards. However, compared to these works, we additionally ground this approach theoretically and provide regret bounds for nonlinear systems and common RL settings.

8 Conclusion

In this work, we propose OMBRL, which maximizes a weighted sum of the extrinsic reward and the agent's epistemic uncertainty. We show that OMBRL effectively performs optimistic exploration and provide regret bounds for it in a variety of settings, in particular for continuous state-action spaces and many common classes of RL problems, namely, finite-horizon, infinite horizon, episodic, and non-episodic RL. Compared to prior optimistic exploration methods, OMBRL is much simpler, more flexible, and scalable. We illustrate this in our experiments, where we combine OMBRL with different model-based RL algorithms, evaluate it on tasks of varying dimensionality, including visual control problems, and also illustrate the benefits of optimistic exploration on hardware. In all cases, OMBRL achieves the best performance.

A limitation of OMBRL is its tendency to over-explore in certain scenarios, which can lead to reduced sample efficiency compared to greedy approaches. This challenge could potentially be addressed by tuning the schedule for the parameter λ_n. In this work, we use the auto-tuning approach proposed in Sukhija et al. (2024a). However, we believe investigating other approaches for selecting λ_n is a promising direction for future work. Another promising direction would be deriving lower bounds on regret for OMBRL. As such bounds are not yet available for many model-based algorithms, including those by Kakade et al. (2020), Curi et al. (2020), and Sukhija et al. (2024b), this could provide valuable theoretical insights into the performance of these methods.

ACKNOWLEDGMENTS

This project was supported in part by the ONR Science of Autonomy Program N000142212121, the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40 180545, the Microsoft Swiss Joint Research Center, the SNSF Postdoc Mobility Fellowship 211086, and an ONR DURIP grant. Pieter Abbeel holds concurrent appointments as a Professor at UC Berkeley and as an Amazon Scholar. This paper describes work performed at UC Berkeley and is not associated with Amazon.

REFERENCES

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Conference on Learning Theory*, 2011.

Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. In *International Conference on Machine Learning*, 2020.

Yarden As, Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Stelian Coros, and Andreas Krause. Actsafe: Active exploration with safety constraints for reinforcement learning. *arXiv preprint arXiv:2410.09486*, 2024.

Karl Johan Åström and Björn Wittenmark. Problems of identification and control. *Journal of Mathematical analysis and applications*, 34(1):90–113, 1971.

Arthur Aubret, Laetitia Matignon, and Salima Hassas. An information-theoretic perspective on intrinsic motivation in reinforcement learning: A survey. *Entropy*, 2023.

P Auer. Finite-time analysis of the multiarmed bandit problem, 2002.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *NeurIPS*, 2016.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Felix Berkenkamp. *Safe exploration in reinforcement learning: Theory and applications in robotics*. PhD thesis, ETH Zurich, 2019.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *JMLR*, 2002.

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

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2023.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 2022.

Mona Buisson-Fenet, Friedrich Solowjow, and Sebastian Trimpe. Actively learning gaussian process dynamics. In *L4DC*. PMLR, 2020.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. *Advances in neural information processing systems*, 30, 2017.

Alessandro Chiuso, Marco Fabris, Valentina Breschi, and Simone Formentin. Harnessing uncertainty for a separation principle in direct data-driven predictive control. *arXiv preprint arXiv:2312.14788*, 2023.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *NeurIPS*, 2018.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, 2019.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. *NeurIPS*, 33:14156–14170, 2020.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *ICML*, pp. 465–472, 2011.

Onno Eberhard, Jakob Hollenstein, Cristina Pinneri, and Georg Martius. Pink noise is all you need: Colored noise exploration in deep reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Eyal Even-Dar and Yishay Mansour. Convergence of optimistic and incremental q-learning. *NeurIPS*, 14, 2001.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, 2020.

Riccardo Alessandro Grimaldi, Giacomo Baggio, Ruggero Carli, and Gianluigi Pillonetto. The bayesian separation principle for data-driven control. *arXiv preprint arXiv:2409.16717*, 2024.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nicklas Hansen, Yixin Lin, Hao Su, Xiaolong Wang, Vikash Kumar, and Aravind Rajeswaran. Modem: Accelerating visual model-based reinforcement learning with demonstrations. *arXiv preprint arXiv:2212.05698*, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *JMLR*, 2010.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *NeurIPS*, 32, 2019.

Juraj Kabzan, Miguel I Valls, Victor JF Reijgwart, Hubertus FC Hendrikx, Claas Ehmke, Manish Prajapat, Andreas Bühler, Nikhil Gosala, Mehak Gupta, Ramya Sivanesan, et al. Amz driverless: The full autonomous racing system. *Journal of Field Robotics*, 2020.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, pp. 267–274, 2002.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *NeurIPS*, 33:15312–15325, 2020.

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

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Hassan K Khalil. *Nonlinear control*, volume 406. Pearson New York, 2015.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *IEEE congress on evolutionary computation*, 2005.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2020.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and achieving goals via world models. *NeurIPS*, 2021.

Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023.

Antoine Moulin and Gergely Neu. Optimistic planning by regularized dynamic programming. In *International Conference on Machine Learning*, pp. 25337–25357. PMLR, 2023.

Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *NeurIPS*, 2018.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *NeurIPS*, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

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

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *ICML*, pp. 5062–5071. PMLR, 2019.

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 *Conference on Robot Learning*, pp. 1049–1065. PMLR, 2021.

Vitchyr H. Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning. *arXiv preprint arXiv:1903.03698*, 2019.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Marc Rigter, Bruno Lacerda, and Nick Hawes. Rambo-rl: Robust adversarial model-based offline reinforcement learning. *NeurIPS*, 35:16082–16097, 2022.

Jonas Rothfuss, Bhavya Sukhija, Tobias Birchler, Parnian Kassraie, and Andreas Krause. Hallucinated adversarial control for conservative offline policy evaluation. *UAI*, 2023.

Jonas Rothfuss, Bhavya Sukhija, Lenart Treven, Florian Dörfler, Stelian Coros, and Andreas Krause. Bridging the sim-to-real gap with bayesian inference. *IROS*, 2024.

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*, 2018.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. *Guided Self-Organization: Inception*, 2014.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pp. 222–227, 1991.

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

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*, 2020.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

Archit Sharma, Kelvin Xu, Nikhil Sardana, Abhishek Gupta, Karol Hausman, Sergey Levine, and Chelsea Finn. Autonomous reinforcement learning: Formalism and benchmarking. *arXiv preprint arXiv:2112.09605*, 2021.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 2017.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*. PMLR, 2020.

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

Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

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*, 2024a.

Bhavya Sukhija, Lenart Treven, Florian Dörfler, Stelian Coros, and Andreas Krause. Neorl: Efficient exploration for nonepisodic rl. *NeurIPS*, 2024b.

Bhavya Sukhija, Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic active exploration of dynamical systems. *NeurIPS*, 2024c.

Scott Sussex, Anastasiia Makarova, and Andreas Krause. Model-based causal bayesian optimization. In *ICLR*, May 2023.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Lenart Treven, Sebastian Curi, Mojmír Mutnỳ, and Andreas Krause. Learning stabilizing controllers for unstable linear quadratic regulators from a single trajectory. In *Learning for Dynamics and Control*, 2021.

Lenart Treven, Jonas Hübotter, Bhavya Sukhija, Florian Dörfler, and Andreas Krause. Efficient exploration in continuous-time model-based reinforcement learning. *NeurIPS*, 2024.

Andrew Wagenmaker, Guanya Shi, and Kevin Jamieson. Optimal exploration for model-based rl in nonlinear systems. *arXiv preprint arXiv:2306.09210*, 2023.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *CORL*. PMLR, 2023.

Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 2017.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *NeurIPS*, 2020.

A Proofs

A.1 Analysis for the finite horizon case

Lemma A.1. *Let Assumption 5.1 and Assumption 5.3 hold. Consider the following definitions*

$$
\begin{aligned}
&J(\boldsymbol{\pi}, \boldsymbol{f}^*) = \mathbb{E}_{\boldsymbol{f}^*}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\right], \\
&\textit{s.t., } \boldsymbol{x}_{t+1} = \boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}_0 = \boldsymbol{x}(0). \\
&J(\boldsymbol{\pi}, \boldsymbol{\mu}_n) = \mathbb{E}_{\boldsymbol{\mu}_n}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right], \\
&\textit{s.t., } \boldsymbol{x}'_{t+1} = \boldsymbol{\mu}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}'_0 = \boldsymbol{x}(0). \\
&\Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}^*) = \mathbb{E}_{\boldsymbol{f}^*}\left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\|\right], \\
&\textit{s.t., } \boldsymbol{x}_{t+1} = \boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}_0 = \boldsymbol{x}(0). \\
&\Sigma_n(\boldsymbol{\pi}, \boldsymbol{\mu}_n) = \mathbb{E}_{\boldsymbol{\mu}_n}\left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\|\right]. \\
&\textit{s.t., } \boldsymbol{x}'_{t+1} = \boldsymbol{\mu}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}'_0 = \boldsymbol{x}(0). \\
&\lambda_n = C_{\max} T \frac{(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma},
\end{aligned}
$$

where $C_{\max} = \max\{R_{\max}, \sigma_{\max}\}$. *Then we have for all* $n \geq 0$, $\boldsymbol{\pi} \in \Pi$ *with probability at least* $1-\delta$

$$
\begin{aligned}
|J(\boldsymbol{\pi}, \boldsymbol{f}^*) - J(\boldsymbol{\pi}, \boldsymbol{\mu}_n)| &\leq \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{\mu}_n) \\
|J(\boldsymbol{\pi}, \boldsymbol{f}^*) - J(\boldsymbol{\pi}, \boldsymbol{\mu}_n)| &\leq \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}^*)
\end{aligned}
$$

Proof. We give the proof for $|J(\boldsymbol{\pi}, \boldsymbol{f}^*) - J(\boldsymbol{\pi}, \boldsymbol{\mu}_n)| \leq \lambda_n(L_r, \boldsymbol{\mu}_n)\Sigma_n(\boldsymbol{\pi}, \boldsymbol{\mu}_n)$. The same argument holds for the second inequality. Let $J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x})$ denote the cost-to-go from state $\boldsymbol{x}$, step $t+1$ onwards under the dynamics $\boldsymbol{f}^*$. Following the Policy difference Lemma from (Kakade & Langford, 2002) and Sukhija et al. (2024c, Corollary 2.)

$$
\begin{aligned}
&J(\boldsymbol{\pi}, \boldsymbol{\mu}_n) - J(\boldsymbol{\pi}, \boldsymbol{f}^*) \\
&= \mathbb{E}_{\boldsymbol{\mu}_n}\left[\sum_{t=0}^{T-1} J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x}'_{t+1}) - J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \hat{\boldsymbol{x}}_{t+1})\right], \\
&\text{with } \hat{\boldsymbol{x}}_{t+1} = \boldsymbol{f}^*(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t,
\end{aligned}
$$

and $\boldsymbol{x}'_{t+1} = \boldsymbol{\mu}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t$.

Therefore,

$$\begin{aligned}
&|J(\boldsymbol{\pi}, \boldsymbol{\mu}_n) - J(\boldsymbol{\pi}, \boldsymbol{f}^*)| \\
&= \left| \mathbb{E}\left[\sum_{t=0}^{T-1} J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x}'_{t+1}) - J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \hat{\boldsymbol{x}}_{t+1}) \right] \right| \\
&\leq \sum_{t=0}^{T-1} \mathbb{E}\left[\left| \mathbb{E}_{\boldsymbol{w}_t}\left[J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x}'_{t+1}) - J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \hat{\boldsymbol{x}}_{t+1}) \right] \right| \right]
\end{aligned}$$

Next, we bound the last term using the derivation from Kakade et al. (2020). Let $C(\boldsymbol{x}) = J^2_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x})$.

$$\begin{aligned}
&\left| \mathbb{E}_{\boldsymbol{w}_t}\left[J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x}'_{t+1}) - J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \hat{\boldsymbol{x}}_{t+1}) \right] \right| \\
&\leq \sqrt{\max\left\{ \mathbb{E}_{\boldsymbol{w}_t}[C(\boldsymbol{x}'_{t+1})], \mathbb{E}_{\boldsymbol{w}_t}[C(\hat{\boldsymbol{x}}_{t+1})] \right\}} \\
&\times \min\left\{ \frac{\|\boldsymbol{f}^*(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) - \boldsymbol{\mu}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\|}{\sigma}, 1 \right\} \qquad \text{(Kakade et al., 2020, Lemma C.2.)} \\
&\leq R_{\max} T \frac{(1 + \sqrt{d_x})\beta_{n-1}(\delta)}{\sigma} \|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\|
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&|J(\boldsymbol{\pi}, \boldsymbol{\mu}_n) - J(\boldsymbol{\pi}, \boldsymbol{f}^*)| \\
&\leq \sum_{t=0}^{T-1} \mathbb{E}\left[\left| \mathbb{E}_{\boldsymbol{w}_t}\left[J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x}'_{t+1}) - J_{t+1}(\boldsymbol{\pi}, \boldsymbol{f}^*, \hat{\boldsymbol{x}}_{t+1}) \right] \right| \right] \\
&\leq \lambda_n \sum_{t=0}^{T-1} \mathbb{E}\left[\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\| \right].
\end{aligned}$$

□

Note that Proposition 5.5 follows directly from Lemma A.1.

Lemma A.2. *Let Assumption 5.1 and Assumption 5.3 hold and consider the simple regret at episode n, $r_n = J(\boldsymbol{\pi}^*, \boldsymbol{f}^*) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*)$. The following holds for all $n > 0$ with probability at least $1 - \delta$*

$$r_n \leq (2\lambda_n + \lambda_n^2)\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*)$$

Proof.

$$\begin{aligned}
r_n &= J(\boldsymbol{\pi}^*, \boldsymbol{f}^*) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*) \\
&\leq J(\boldsymbol{\pi}^*, \boldsymbol{\mu}_n) + \lambda_n \Sigma_n(\boldsymbol{\pi}^*, \boldsymbol{\mu}_n) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*) && \text{(Lemma A.1)} \\
&\leq J(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) + \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*) && \text{(Equation (8))} \\
&= J(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*) + \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) \\
&\leq \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*) + \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) && \text{(Lemma A.1)} \\
&= 2\lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*) + \lambda_n(\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) - \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*)) \\
&\leq (\lambda_n^2 + 2\lambda_n)\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*).
\end{aligned}$$

Here in the last inequality, we used the fact that $\|\boldsymbol{\sigma}(\cdot, \cdot)\|$ is bounded and positive, therefore, we can treat it similar to the reward (it is in fact an intrinsic reward) and use Lemma A.1. □

Proof of Theorem 5.6.

$$\begin{aligned}
R_N &= \sum_{n=1}^{N} r_n \\
&\leq \sum_{n=1}^{N} (\lambda_n^2 + 2\lambda_n)\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*) \\
&\leq (\lambda_N^2 + \lambda_N) \sum_{n=1}^{N} \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*)
\end{aligned}$$

$$= (\lambda_N^2 + 2\lambda_N) \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{f}^*} \left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\| \right]$$

$$\leq (\lambda_N^2 + 2\lambda_N)\sqrt{NT} \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{f}^*} \left[\sum_{t=0}^{T-1} \left\|\boldsymbol{\sigma}_n^2(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\right\| \right]$$

$$\leq C(\lambda_N^2 + 2\lambda_N) T\sqrt{N\Gamma_{NT}} \qquad \text{(Curi et al. (2020, Lemma 17))}$$

Finally, note that from Lemma A.1 we have $\lambda_N \propto T\beta_n$ and $\beta_n \propto \sqrt{\Gamma_n}$ (Chowdhury & Gopalan, 2017). Therefore, $R_N \leq \mathcal{O}(T^3 \Gamma_N^{3/2} \sqrt{N})$ □

Table 1: Maximum information gain bounds for common choice of kernels.

Kernel	$k(\boldsymbol{x}, \boldsymbol{x}')$	Γ_N
Linear	$\boldsymbol{x}^\top \boldsymbol{x}'$	$\mathcal{O}\left(d \log(N)\right)$
RBF	$e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2l^2}}$	$\mathcal{O}\left(\log^{d+1}(N)\right)$
Matèrn	$\frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\boldsymbol{x}'\|}{l}\right)^{\nu} B_\nu \left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\boldsymbol{x}'\|}{l}\right)$	$\mathcal{O}\left(N^{\frac{d}{2\nu+d}} \log^{\frac{2\nu}{2\nu+d}}(N)\right)$

In Table 1 we list rates of Γ_N for the most common choice of kernels.

A.2 Analysis for the discounted infinite horizon case

For the infinite horizon case, we first study the posterior variance $\boldsymbol{\sigma}_n$ in the feature space. Moreover, let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{u})$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{U}$.

For the ease of notation we denote $\boldsymbol{z}_{k,n} = (\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n))$. For $\boldsymbol{z}$ we define the kernel embedding $k_{\boldsymbol{z}} = k(\boldsymbol{z}, \cdot)$. The covariance matrix $\boldsymbol{V}_t : \mathcal{H} \to \mathcal{H}$ in the feature form is:

$$\boldsymbol{V}_t = \boldsymbol{I} + \frac{1}{\sigma^2} \sum_{i=1}^{t} k_{\boldsymbol{z}_i} k_{\boldsymbol{z}_i}^\top. \tag{10}$$

Note that we have $\boldsymbol{x}_{t+1} = \langle k_{\boldsymbol{z}_t}, \boldsymbol{f}^* \rangle_{\mathcal{H}} + \boldsymbol{w}_t$. With the design matrix $\boldsymbol{M}_t : \mathcal{H} \to \mathbb{R}^t$

$$\boldsymbol{M}_t = \begin{pmatrix} k_{\boldsymbol{z}_1} & k_{\boldsymbol{z}_2} & \cdots & k_{\boldsymbol{z}_t} \end{pmatrix} \tag{11}$$

we have $\boldsymbol{V}_t = \boldsymbol{I} + \frac{1}{\sigma^2} \boldsymbol{M}_t \boldsymbol{M}_t^\top$ and since $\boldsymbol{K}_t = \boldsymbol{M}_t^\top \boldsymbol{M}_t$ we have

$$\det(\boldsymbol{V}_t) = \det\left(\boldsymbol{I} + \frac{1}{\sigma^2} \boldsymbol{K}_t\right) \tag{12}$$

Corollary A.3 (Lower bound on the posterior log determinant)**.**

$$\log\left(|\boldsymbol{V}_n|\right) \geq \log\left(|\boldsymbol{V}_{n-1}|\right) + \log\left(1 + \sigma^{-2} \sum_{k=1}^{\widehat{T}_n} \|\boldsymbol{\sigma}_{n-1}(\boldsymbol{z}_{k,n})\|^2\right) \tag{13}$$

In particular, we have

$$\log\left(\frac{|\boldsymbol{V}_N|}{|\boldsymbol{V}_0|}\right) \geq \sum_{n=1}^{N} \log\left(1 + \sigma^{-2} \sum_{k=1}^{\widehat{T}_n} \|\boldsymbol{\sigma}_{n-1}(\boldsymbol{z}_{k,n})\|^2\right) \tag{14}$$

Proof.

$$\log\left(|\boldsymbol{V}_n|\right) = \log\left(|\boldsymbol{V}_{n-1}|\right) + \log\left(\left|\boldsymbol{I} + \sigma^{-2} \boldsymbol{V}_{n-1}^{-1/2} \sum_{k=1}^{\widehat{T}_n} \boldsymbol{k}_{\boldsymbol{z}_{k,n}} \boldsymbol{k}_{\boldsymbol{z}_{k,n}}^\top \boldsymbol{V}_{n-1}^{-1/2}\right|\right)$$

$$\geq \log\left(|\boldsymbol{V}_{n-1}|\right)$$

$$+\log\left(1+\operatorname{tr}\left(\sigma^{-2}\boldsymbol{V}_{n-1}^{-1/2}\sum_{k=1}^{\widehat{T}_n}\boldsymbol{k}_{\boldsymbol{z}_{k,n}}\boldsymbol{k}_{\boldsymbol{z}_{k,n}}^{\top}\boldsymbol{V}_{n-1}^{-1/2}\right)\right) \qquad \text{(see (*) below)}$$

$$=\log\left(|\boldsymbol{V}_{n-1}|\right)+\log\left(1+\sigma^{-2}\sum_{k=1}^{\widehat{T}_n}\left\|\boldsymbol{k}_{\boldsymbol{z}_{k,n}}\right\|^2_{\boldsymbol{V}_{n-1}^{-1}}\right)$$

$$=\log\left(|\boldsymbol{V}_{n-1}|\right)+\log\left(1+\sigma^{-2}\sum_{k=1}^{\widehat{T}_n}\left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{z}_{k,n})\right\|^2\right)$$

We prove (*) in the following, first let $\boldsymbol{m}_k=\sigma^{-1}\boldsymbol{V}_{n-1}^{-1/2}\boldsymbol{k}_{\boldsymbol{z}_{k,n}}$, then we have

$$\log\left(\left|\boldsymbol{I}+\sigma^{-2}\boldsymbol{V}_{n-1}^{-1/2}\sum_{k=1}^{\widehat{T}_n}\boldsymbol{k}_{\boldsymbol{z}_{k,n}}\boldsymbol{k}_{\boldsymbol{z}_{k,n}}^{\top}\boldsymbol{V}_{n-1}^{-1/2}\right|\right)$$

$$=\log\left(\left|\boldsymbol{I}+\sum_{k=1}^{\widehat{T}_n}\boldsymbol{m}_k\boldsymbol{m}_k^{\top}\right|\right).$$

The matrix $\boldsymbol{M}=\sum_{k=1}^{\widehat{T}_n}\boldsymbol{m}_k\boldsymbol{m}_k^{\top}$ by definition is positive semi-definite. Moreover, $|\boldsymbol{I}+\boldsymbol{M}|=\prod_{i\geq 1}(1+\alpha_i)$, where $\alpha_i\geq 0$ are the eigenvalues of $\boldsymbol{M}$. Furthermore, since $\alpha_i\geq 0$ and $\prod_{i\geq 1}(1+\alpha_i)=1+\sum_{i\geq 1}\alpha_i+\cdots+\prod_{i\geq 1}\alpha_i$, we get $\prod_{i\geq 1}(1+\alpha_i)\geq 1+\sum_{i\geq 1}\alpha_i$. Finally, since $\sum_{i\geq 1}\alpha_i=\operatorname{tr}(\boldsymbol{M})$, we get $|\boldsymbol{I}+\boldsymbol{M}|\geq 1+\operatorname{tr}(\boldsymbol{M})$. □

Corollary A.4 (Upper bound on the posterior log determinant)**.**

$$\log\left(|\boldsymbol{V}_n|\right)\leq\log\left(|\boldsymbol{V}_{n-1}|\right)+\sum_{k=1}^{\widehat{T}_n}\sum_{j=1}^{d_x}\log\left(1+\sigma^{-2}\sigma^2_{n-1,j}(\boldsymbol{z}_{k,n})\right)$$

Proof.

$$\log\left(|\boldsymbol{V}_n|\right)=\log\left(|\boldsymbol{V}_{n-1}|\right)+\log\left(|\boldsymbol{I}+\boldsymbol{M}|\right)$$
$$\leq\log\left(|\boldsymbol{V}_{n-1}|\right)+\log\left(|\operatorname{diag}\left(\boldsymbol{I}+\boldsymbol{M}\right)|\right) \qquad \text{(Hadamard's inequality for PSD matrices)}$$
$$=\log\left(|\boldsymbol{V}_{n-1}|\right)+\sum_{k=1}^{\widehat{T}_n}\sum_{j=1}^{d_x}\log\left(1+\sigma^{-2}\sigma^2_{n-1,j}(\boldsymbol{z}_{k,n})\right)$$

□

Corollary A.3 will be useful for the discounted horizon case, whereas Corollary A.4 will be applied for the nonepisodic setting.

Next, we show that OMBRL also performs optimism in the discounted horizon case.

Lemma A.5. *Let Assumption 5.1, and Assumption 5.3 hold. Consider the following definitions*

$$J_\gamma(\boldsymbol{\pi},\boldsymbol{f}^*)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(\boldsymbol{x}_t,\boldsymbol{\pi}(\boldsymbol{x}_t))\right]$$
$$\textit{s.t., } \boldsymbol{x}_{t+1}=\boldsymbol{f}^*(\boldsymbol{x}_t,\boldsymbol{\pi}(\boldsymbol{x}_t))+\boldsymbol{w}_t,\quad \boldsymbol{x}_0=\boldsymbol{x}(0),$$
$$J_\gamma(\boldsymbol{\pi},\boldsymbol{\mu}_n)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))\right]$$
$$\textit{s.t., } \boldsymbol{x}'_{t+1}=\boldsymbol{\mu}_n(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))+\boldsymbol{w}_t,\quad \boldsymbol{x}'_0=\boldsymbol{x}(0),$$
$$\Sigma_n^\gamma(\boldsymbol{\pi},\boldsymbol{f}^*)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\left\|\boldsymbol{\sigma}_n(\boldsymbol{x}_t,\boldsymbol{\pi}(\boldsymbol{x}_t))\right\|\right]$$
$$\textit{s.t., } \boldsymbol{x}_{t+1}=\boldsymbol{f}^*(\boldsymbol{x}_t,\boldsymbol{\pi}(\boldsymbol{x}_t))+\boldsymbol{w}_t,\quad \boldsymbol{x}_0=\boldsymbol{x}(0),$$

$$\Sigma_n^\gamma(\boldsymbol{\pi}, \boldsymbol{\mu}_n) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left\|\boldsymbol{\sigma}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|\right]$$
$$\text{s.t., } \boldsymbol{x}'_{t+1} = \boldsymbol{\mu}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}'_0 = \boldsymbol{x}(0),$$
$$\lambda_n = C_{\max} \frac{\gamma}{1-\gamma} \frac{(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma},$$

where $C_{\max} = \max\{R_{\max}, \sigma_{\max}\}$. *Then we have for all* $n \geq 0$, $\boldsymbol{\pi} \in \Pi$ *with probability at least* $1-\delta$

$$|J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*) - J_\gamma(\boldsymbol{\pi}, \boldsymbol{\mu}_n)| \leq \lambda_n \Sigma_n^\gamma(\boldsymbol{\pi}, \boldsymbol{\mu}_n)$$
$$|J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*) - J_\gamma(\boldsymbol{\pi}, \boldsymbol{\mu}_n)| \leq \lambda_n \Sigma_n^\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*)$$

Proof. We give the proof for $|J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*) - J_\gamma(\boldsymbol{\pi}, \boldsymbol{\mu}_n)| \leq \lambda_n(L_r, \boldsymbol{\mu}_n)\Sigma_n^\gamma(\boldsymbol{\pi}, \boldsymbol{\mu}_n)$. The same argument holds for the second inequality. We can extend the result from Sukhija et al. (2024a, Corollary 2.,) to the discounted case and get

$$\begin{aligned}
&J_\gamma(\boldsymbol{\pi}, \boldsymbol{\mu}_n) - J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*) \\
&= \mathbb{E}_{\boldsymbol{\mu}_n}\left[\sum_{t=0}^{\infty} \gamma^{t+1}(J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x}'_{t+1}) - J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*, \hat{\boldsymbol{x}}_{t+1}))\right], \\
&\text{with } \hat{\boldsymbol{x}}_{t+1} = \boldsymbol{f}^*(\boldsymbol{s}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t, \\
&\text{and } \boldsymbol{x}'_{t+1} = \boldsymbol{\mu}_n(\boldsymbol{s}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t.
\end{aligned}$$

Let $\beta_n \frac{1+\sqrt{d_x}}{\sigma} C(\boldsymbol{x}) = J_\gamma^2(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x})$. Note that $C(\boldsymbol{x}) \leq \lambda_n$ for all $\boldsymbol{c} \in \mathcal{X}$. Therefore, we have

$$\begin{aligned}
&|J_\gamma(\boldsymbol{\pi}, \boldsymbol{\mu}_n) - J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*)| \\
&= \left|\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t+1}(J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x}'_{t+1}) - J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*, \hat{\boldsymbol{x}}_{t+1}))\right]\right| \\
&\leq \sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}\left[\left|\mathbb{E}_{\boldsymbol{w}_t}\left[J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*, \boldsymbol{x}'_{t+1}) - J_\gamma(\boldsymbol{\pi}, \boldsymbol{f}^*, \hat{\boldsymbol{x}}_{t+1})\right]\right|\right] \\
&\leq \sum_{t=0}^{\infty} \gamma \mathbb{E}\left[\sqrt{\max\left\{\mathbb{E}_{\boldsymbol{w}_t}[C(\boldsymbol{x}'_{t+1})], \mathbb{E}_{\boldsymbol{w}_t}[C(\hat{\boldsymbol{x}}_{t+1})]\right\}}\right. \\
&\left.\times\ \gamma^t \min\left\{\frac{\|\boldsymbol{f}^*(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) - \boldsymbol{\mu}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\|}{\sigma}, 1\right\}\right] \qquad \text{(Kakade et al., 2020, Lemma C.2.)} \\
&\leq \lambda_n \sum_{t=0}^{\infty} \mathbb{E}\left[\gamma^t \left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|\right] \qquad \text{((Sukhija et al., 2024c, Corollary 3))}
\end{aligned}$$

□

Proof of Theorem 5.7. We start with bounding

$$\sum_{n=1}^{N} \sum_{t=0}^{\infty} \mathbb{E}_{\boldsymbol{w}_{1:t-1}}\left[\gamma^t \left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right] \tag{15}$$

To achieve this, we use a sampling strategy where we increase the horizon of rollouts with each episode n. In the discounted setting, this allows us to collect data at the tails of our rollouts, i.e., make observations with longer rollouts and thus approximate the true discounted value function asymptotically. Moreover, we set $T(n) = -\frac{\log(n)}{\log(\gamma)}$ (note that $\gamma < 1$ and therefore $T(n)$ is positive).

$$\begin{aligned}
&\sum_{n=1}^{N} \sum_{t=0}^{\infty} \mathbb{E}_{\boldsymbol{w}_{1:t-1}}\left[\gamma^t \left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right] \\
&= \sum_{n=1}^{N} \sum_{t=0}^{T(n)-1} \mathbb{E}_{\boldsymbol{w}_{1:t-1}}\left[\gamma^t \left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{n=1}^{N}\sum_{t=T(n)}^{\infty}\mathbb{E}_{\boldsymbol{w}_{1:t-1}}\left[\gamma^t\left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right]\\
&\leq\sum_{n=1}^{N}\sum_{t=0}^{T(n)-1}\mathbb{E}_{\boldsymbol{w}_{1:t-1}}\left[\gamma^t\left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right]\\
&+\sum_{n=1}^{N}\gamma^{T(n)}\frac{\sigma_{\max}^2}{1-\gamma}\\
&=\sum_{n=1}^{N}\sum_{t=0}^{T(n)-1}\mathbb{E}_{\boldsymbol{w}_{1:t-1}}\left[\gamma^t\left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right]\\
&+\sum_{n=1}^{N}n^{-1}\frac{\sigma_{\max}^2}{1-\gamma}\\
&=\sum_{n=1}^{N}\sum_{t=0}^{T(n)-1}\mathbb{E}_{\boldsymbol{w}_{1:t-1}}\left[\gamma^t\left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right]\\
&+\frac{C\sigma_{\max}^2}{1-\gamma}\log(N)
\end{aligned}
$$

Next, we bound the term

$$
s_n=\sum_{t=0}^{T(n)-1}\gamma^t\sigma^{-2}\left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2.
$$

Note that, $s_n\in\left[0,\frac{\sigma^{-2}d_x\sigma_{\max}^2}{1-\gamma}\right)$. Let $s_{\max}=\frac{\sigma^{-2}d_x\sigma_{\max}^2}{1-\gamma}$, we have $s_n\leq\frac{s_{\max}}{\log(1+s_{\max})}\log(1+s_n)$ (Srinivas et al., 2012). Define $C_\gamma=\frac{s_{\max}}{\log(1+s_{\max})}$. We have,

$$
\begin{aligned}
s_n&\leq C_\gamma\log\left(1+\sigma^{-2}\sum_{t=0}^{T(n)-1}\gamma^t\left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right)\\
&\leq C_\gamma\log\left(1+\sigma^{-2}\sum_{t=0}^{T(n)-1}\left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right)
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
&\sum_{n=1}^{N}s_n\\
&\leq C_\gamma\sum_{n=1}^{N}\log\left(1+\sigma^{-2}\sum_{t=0}^{T(n)-1}\left\|\boldsymbol{\sigma}_{n-1}(\boldsymbol{x}'_t,\boldsymbol{\pi}(\boldsymbol{x}'_t))\right\|^2\right)\\
&\leq C_\gamma\Gamma_{\sum_{n=1}^{N}T(n)} && \text{(Corollary A.3)}\\
&\leq C_\gamma\Gamma_{N\log(N)}
\end{aligned}
$$

$$
\begin{aligned}
R_N&=\sum_{n=1}^{N}r_n\\
&\leq\sum_{n=1}^{N}(\lambda_n^2+2\lambda_n)\Sigma_n^\gamma(\boldsymbol{\pi}_n,\boldsymbol{f}^*)\\
&\leq(\lambda_N^2+2\lambda_N)\sum_{n=1}^{N}\Sigma_n^\gamma(\boldsymbol{\pi}_n,\boldsymbol{f}^*)
\end{aligned}
$$

$$
\begin{aligned}
&= (\lambda_N^2 + 2\lambda_N)\sqrt{N}\sqrt{\sum_{n=1}^{N}(\Sigma_n^{\gamma}(\boldsymbol{\pi}_n, \boldsymbol{f}^*))^2} \\
&\leq (2\lambda_N + \lambda_N^2)\sqrt{N} \\
&\times \sqrt{\sum_{n=1}^{N}\mathbb{E}\left[\left(\sum_{t=0}^{\infty}\gamma^t \left\|\boldsymbol{\sigma}_n(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\right\|^2\right]\right)^2} \\
&\leq (2\lambda_N + \lambda_N^2)\sqrt{N} \\
&\times \sqrt{\sum_{n=1}^{N}\mathbb{E}\left[\left(\sum_{t=0}^{\infty}\gamma^t\right)\left(\sum_{t=0}^{\infty}\gamma^t \left\|\boldsymbol{\sigma}_n^2(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\right\|^2\right]\right)} \\
&= (2\lambda_N + \lambda_N^2)\sqrt{\frac{N}{1-\gamma}} \\
&\times \sqrt{\sum_{n=1}^{N}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t \left\|\boldsymbol{\sigma}_n^2(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\right\|^2\right]} \\
&\leq (2\lambda_N + \lambda_N^2)\sqrt{\frac{C_\gamma N \Gamma_{N\log(N)}}{1-\gamma} + \frac{C\sigma_{\max}^2 N\log(N)}{(1-\gamma)^2}}
\end{aligned}
$$

Since $\lambda_N \propto {}^{\beta_N}/_{1-\gamma}$, we get

$$R_N \leq \mathcal{O}\left(\Gamma_{N\log(N)}^{3/2}\sqrt{N}\right)$$

□

A.3 Analysis for the non-episodic RL case

In this section, we prove Theorem 5.8. First, we restate the bounded energy assumption from Sukhija et al. (2024b).

Definition A.6 ($\mathcal{K}_\infty$-functions)**.** The function $\xi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is of class $\mathcal{K}_\infty$, if it is continuous, strictly increasing, $\xi(0) = 0$ and $\xi(s) \to \infty$ for $s \to \infty$.

Assumption A.7 (Policies with bounded energy)**.** We assume there exists $\kappa, \xi \in \mathcal{K}_\infty$, positive constants K, C_u, C_l with $C_u > C_l$, and $\gamma \in (0,1)$ such that for each $\boldsymbol{\pi} \in \Pi$ we have,

Bounded energy: There exists a Lyapunov function $V^{\boldsymbol{\pi}} : \mathcal{X} \to [0, \infty)$ for which $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,

$$
\begin{aligned}
|V^{\boldsymbol{\pi}}(\boldsymbol{x}) - V^{\boldsymbol{\pi}}(\boldsymbol{x}')| &\leq \kappa(\|\boldsymbol{x} - \boldsymbol{x}'\|) && \text{(uniform continuity)} \\
C_l\xi(\|\boldsymbol{x}\|) \leq V^{\boldsymbol{\pi}}(\boldsymbol{x}) &\leq C_u\xi(\|\boldsymbol{x}\|) && \text{(positive definiteness)} \\
\mathbb{E}_{\boldsymbol{x}_+|\boldsymbol{x},\boldsymbol{\pi}}[V^{\boldsymbol{\pi}}(\boldsymbol{x}_+)] &\leq \gamma V^{\boldsymbol{\pi}}(\boldsymbol{x}) + K && \text{(drift condition)}
\end{aligned}
$$

where $\boldsymbol{x}_+ = \boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x})) + \boldsymbol{w}$.

Bounded norm of reward:

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \frac{r(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))}{1 + V^{\boldsymbol{\pi}}(\boldsymbol{x})} < \infty$$

Boundedness of the noise with respect to κ:

$$\mathbb{E}_{\boldsymbol{w}}\left[\kappa(\|\boldsymbol{w}\|)\right] < \infty,\ \mathbb{E}_{\boldsymbol{w}}\left[\kappa^2(\|\boldsymbol{w}\|)\right] < \infty$$

Sukhija et al. (2024b) argue that this assumption is often satisfied in practice. We refer the reader to Sukhija et al. (2024b) for further details. Next, we make an assumption on the underlying system $\boldsymbol{f}^*$.

Assumption A.8 (Continous closed-loop dynamics, and Gaussian noise.)**.** The dynamics model $\boldsymbol{f}^*$ and all $\boldsymbol{\pi} \in \Pi$ are continuous, and process noise is i.i.d. Gaussian with variance σ^2, i.e., $\boldsymbol{w}_t \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

An important quantity in the average reward setting is the bias

$$B(\boldsymbol{\pi}, \boldsymbol{x}_0) = \lim_{T\to\infty} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}_t, \boldsymbol{u}_t) - J_{\text{avg}}(\boldsymbol{\pi})\right]. \tag{16}$$

The Bellman equation for the average reward setting is given by

$$B(\boldsymbol{\pi}, \boldsymbol{x}) + J_{\text{avg}}(\boldsymbol{\pi}) = r(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x})) + \mathbb{E}_{\boldsymbol{x}_+}[B(\boldsymbol{\pi}, \boldsymbol{x}_+)|\boldsymbol{x}, \boldsymbol{\pi}] \tag{17}$$

Sukhija et al. (2024b) show that under Assumption A.7 and Assumption A.8 the average reward solution and the bias (c.f., Equation (3)) are bounded. Moreover, they show that with Assumption 5.3 the average reward and bias are bounded for all dynamics $\boldsymbol{f} \in \mathcal{M}_n \cap \mathcal{M}_0$.

Lemma A.9. *Let Assumption 5.3, Assumption A.7, and Assumption 5.3 hold. Consider the following definitions*

$$\begin{aligned}
&J_{avg}(\boldsymbol{\pi}, \boldsymbol{f}^*) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\right] \\
&\textit{s.t., } \boldsymbol{x}_{t+1} = \boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}_0 = \boldsymbol{x}(0), \\
&J_{avg}(\boldsymbol{\pi}, \boldsymbol{f}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right] \\
&\textit{s.t., } \boldsymbol{x}'_{t+1} = \boldsymbol{f}(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}'_0 = \boldsymbol{x}(0), \\
&\Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}^*) = \lim_{T\to\infty} \frac{1}{T-1}\mathbb{E}\left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\|\right] \\
&\textit{s.t., } \boldsymbol{x}_{t+1} = \boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}_0 = \boldsymbol{x}(0), \\
&\Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\|\right] \\
&\textit{s.t., } \boldsymbol{x}'_{t+1} = \boldsymbol{f}(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}'_0 = \boldsymbol{x}(0), \\
&\lambda_n = D_4(\boldsymbol{x}_0, \gamma, K)\beta_{n-1}(\delta),
\end{aligned}$$

and $D_4(\boldsymbol{x}_0, \gamma, K)$ is defined as in Sukhija et al. (2024b, Theorem 3.1), is independent of n and increases with $\|\boldsymbol{x}_0\|$, K and γ^{-1} (see Sukhija et al. (2024b) for the exact dependence). Then we have for all $n \geq 0$, $\boldsymbol{\pi} \in \Pi$, $\boldsymbol{f} \in \mathcal{M}_n \cap \mathcal{M}_0$ with probability at least $1-\delta$

$$\begin{aligned}
|J_{avg}(\boldsymbol{\pi}, \boldsymbol{f}^*) - J_{avg}(\boldsymbol{\pi}, \boldsymbol{f})| &\leq \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}) \\
|J_{avg}(\boldsymbol{\pi}, \boldsymbol{f}^*) - J_{avg}(\boldsymbol{\pi}, \boldsymbol{f})| &\leq \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}^*)
\end{aligned}$$

Proof.

$$\begin{aligned}
&|J_{\text{avg}}(\boldsymbol{\pi}, \boldsymbol{f}) - J_{\text{avg}}(\boldsymbol{\pi}, \boldsymbol{f}^*)| \\
&= \left|\lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\boldsymbol{f}}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) - J_{\text{avg}}(\boldsymbol{\pi}, \boldsymbol{f}^*)\right]\right| \\
&= \left|\lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\boldsymbol{f}}\left[\sum_{t=0}^{T-1} B(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) - B(\hat{\boldsymbol{x}}'_{t+1}, \boldsymbol{\pi}(\hat{\boldsymbol{x}}'_{t+1}))\right]\right| \\
&\leq \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{f})
\end{aligned} \tag{1}$$

In the second last equality, we used the Bellman equation for the average reward setting (Equation (17)), where $\hat{\boldsymbol{x}}'_{t+1}$ is the next state under the true dynamics $\boldsymbol{f}^*$.

For the last inequality, Sukhija et al. (2024b) bound the bias term with λ_n in Section A.3, on pages $23-24$.

We can use the same derivation to show that $|J_{\text{avg}}(\boldsymbol{\pi}, \boldsymbol{f}) - J_{\text{avg}}(\boldsymbol{\pi}, \boldsymbol{f}^*)| \leq \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}^*)$. □

OMBRL in the non-episodic setting operates similarly to NEORL (Sukhija et al., 2024b). In particular, we update our model and policy every T_n step, where T_n is defined as:

$$T_n = \max\left(\widehat{T_n}, \frac{\lceil \log(C_u/C_l) \rceil}{\log(1/\gamma)}\right), \tag{18}$$

$$\widehat{T_n} = \arg\max_{T \geq 1} T + 1 \tag{19}$$

$$\text{s.t.} \sum_{k=1}^{T} \sum_{j=1}^{d_x} \log\left(1 + \sigma^{-2}\sigma^2_{n-1,j}(\boldsymbol{z}_{k,n})\right) \leq \log(2). \tag{20}$$

Effectively, we update our model and policy only once we have accumulated more than one bit of information, i.e., $\sum_{k=1}^{T}\sum_{j=1}^{d_x} \log\left(1 + \sigma^{-2}\sigma^2_{n-1,j}(\boldsymbol{z}_{k,n})\right) > \log(2)$. With the updated model and model set $\mathcal{M}_n$, we select *any* dynamics in $\boldsymbol{f}_n \in \mathcal{M}_n \cap \mathcal{M}_0$ and pick the policy with

$$\boldsymbol{\pi}_n = \arg\max_{\boldsymbol{\pi}\in\Pi} J_{\text{avg}}(\boldsymbol{\pi}, \boldsymbol{f}_n) + \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}_n). \tag{21}$$

Note that (Sukhija et al., 2024b) require maximizing over the dynamics in $\mathcal{M}_n \cap \mathcal{M}_0$, whereas we do not. Moreover, while this optimization is generally intractable, for OMBRL, we can obtain $\boldsymbol{f}$ using the quadratic program described in Equation (22). However, in practice, we just pick the mean model $\boldsymbol{\mu}_n \in \mathcal{M}_n$. This practical modification is also made in Sukhija et al. (2024b) where they optimize over dynamics in $\mathcal{M}_n$ instead of $\mathcal{M}_n \cap \mathcal{M}_0$.

Theorem A.10 (Formal Theorem statement for informal Theorem 5.8)**.** *Define* $R_N = \sum_{n=1}^{N} \mathbb{E}[J_{avg}(\boldsymbol{\pi}^*) - r(\boldsymbol{x}_n, \boldsymbol{\pi}_n(\boldsymbol{x}_n)]$. *Let Assumption 5.3, Assumption A.7, and Assumption A.8 hold. Then we have for all* $N \geq 0$ *with probability at least* $1-\delta$

$$R_N \leq \mathcal{O}\left(\Gamma_N^{3/2}\sqrt{N}\right)$$

Proof. Let E_N denote the number of episodes after N interactions in the environment.

$$\begin{aligned}
R_N &= \mathbb{E}\left[\sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} J_{\text{avg}}(\boldsymbol{\pi}^*) - r(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n))\right] \\
&\leq \mathbb{E}\left[\sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} J_{\text{avg}}(\boldsymbol{\pi}^*, \boldsymbol{f}_n) + \lambda_n\Sigma_n(\boldsymbol{\pi}^*, \boldsymbol{f}_n) - r(\boldsymbol{z}_k^n)\right] && \text{(Lemma A.9)} \\
&\leq \mathbb{E}\left[\sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} J_{\text{avg}}(\boldsymbol{\pi}_n, \boldsymbol{f}_n) + \lambda_n\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}_n) - r(\boldsymbol{z}_k^n)\right] && \text{(Equation (21))} \\
&\leq \mathbb{E}\left[\sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} J_{\text{avg}}(\boldsymbol{\pi}_n, \boldsymbol{f}_n) - r(\boldsymbol{z}_k^n)\right] \\
&\quad + \lambda_N \mathbb{E}\left[\sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}_n)\right] \\
&\leq \mathcal{O}\left(\Gamma_N\sqrt{N}\right) + \mathbb{E}\left[\lambda_N \sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f})\right] && \text{(Theorem 3.1 Sukhija et al. (2024b))}
\end{aligned}$$

Next, we focus on $\mathbb{E}\left[\lambda_N \sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f})\right]$

$$\begin{aligned}
&\mathbb{E}\left[\sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f})\right] \\
&= \mathbb{E}\left[\sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\|\right] \\
&\quad + \mathbb{E}\left[\sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}) - \|\boldsymbol{\sigma}_n(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\|\right] \\
&\leq C\sqrt{\Gamma_N N} + \mathbb{E}\left[\sum_{n=1}^{E_N}\sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}) - \|\boldsymbol{\sigma}_n(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\|\right] \\
&\qquad \text{(Lemma A.1 Sukhija et al. (2024b))} \\
&\leq \sqrt{N\Gamma_N} + \mathcal{O}\left(\Gamma_N\sqrt{N}\right) && \text{(Sukhija et al. (2024b, Theorem 3.1) with reward } \boldsymbol{\sigma}_n)
\end{aligned}$$

Therefore

$$\mathbb{E}\left[\lambda_T \sum_{n=1}^{E_N} \sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f})\right] \leq \mathcal{O}\left(\lambda_N \Gamma_N \sqrt{N}\right)$$
$$\leq \mathcal{O}\left(\Gamma_N^{3/2} \sqrt{N}\right)$$

In conclusion,

$$R_N \leq \mathcal{O}\left(\Gamma_N^{3/2} \sqrt{N}\right)$$

□

A.4 Analysis for pure intrinsic exploration

In the following, we derive a sample complexity bound for a pure intrinsic exploration algorithm. Thereby showing convergence for methods such Buisson-Fenet et al. (2020).

Theorem A.11. *Let Assumption 5.1 and Assumption 5.3 hold. Consider* OMBRL *with extrinsic reward $r = 0$, i.e.,*

$$\boldsymbol{\pi}_n = \arg\max_{\boldsymbol{\pi} \in \Pi} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}'_t, \boldsymbol{u}_t)\|\right],$$
$$\boldsymbol{x}'_{t+1} = \boldsymbol{\mu}_n(\boldsymbol{x}'_t, \boldsymbol{u}_t) + \boldsymbol{w}_t.$$

Then we have $\forall N > 0$, with probability at least $1 - \delta$

$$\max_{\boldsymbol{\pi} \in \Pi} \mathbb{E}_{\boldsymbol{f}^*}\left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\|\right] \leq \mathcal{O}\left(\sqrt{\frac{\Gamma_N^3}{N}}\right).$$

Proof. Let $\Sigma_N^* = \max_{\boldsymbol{\pi}} \Sigma_N(\boldsymbol{\pi}, \boldsymbol{f}^*)$ and $\boldsymbol{\pi}_N^*$ the corresponding policy.

$$\begin{aligned}
\Sigma_N^* &\leq \frac{1}{N} \sum_{n=1}^{N} \Sigma_n^* && \text{(monotoncity of the variance)} \\
&\leq \frac{1}{N} \sum_{n=1}^{N} (1 + \lambda_n) \Sigma_n(\boldsymbol{\pi}_n^*, \boldsymbol{\mu}_n) && \text{(Lemma A.1)} \\
&\leq \frac{1}{N} \sum_{n=1}^{N} (1 + \lambda_n) \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) && (\boldsymbol{\pi}_n \text{ is the maximizer for mean dynamics } \boldsymbol{\mu}_n) \\
&\leq \sum_{n=1}^{N} (1 + \lambda_n)^2 \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*) && \text{(Lemma A.1)} \\
&\leq (1 + \lambda_N)^2 \frac{1}{N} \sum_{n=1}^{N} \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*) \\
&\leq (1 + \lambda_N)^2 \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \Sigma_n^2(\boldsymbol{\pi}_n, \boldsymbol{f}^*) \\
&\leq \mathcal{O}\left(\sqrt{\frac{\Gamma_N^3}{N}}\right)
\end{aligned}$$

□

Effectively, Theorem A.11 shows that pure intrinsic exploration reduces our model epistemic uncertainty with a rate of $\sqrt{\frac{\Gamma_N^3}{N}}$. To the best of our knowledge, we are the first to show this. Moreover, Sukhija et al. (2024c) derive a similar bound but their algorithm performs optimistic exploration from Equation (7) in addition to maximizing the intrinsic rewards. Our result shows that the optimistic exploration is not necessary for this setting.

A.5 Analysis for the finite horizon setting with sub-Gaussia noise

In the following, we analyse the regret for the setting where the process noise $\boldsymbol{w}$ is σ-sub Gaussian.

Assumption A.12. The dynamics model $\boldsymbol{f}^*$, reward r, and all $\boldsymbol{\pi} \in \Pi$ are L_f, L_r and $L_{\boldsymbol{\pi}}$ Lipschitz, respectively. Furthermore, we assume that process noise is i.i.d. σ-sub Gaussian.

We make the same assumptions as other works (Curi et al., 2020; Sussex et al., 2023) that study this setting. Moreover, Lipschitz continuity is a common assumption for nonlinear dynamics (Khalil, 2015) and is satisfied for many real-world systems.

Curi et al. (2020) provide a regret bound that depends exponentially on the horizon T, i.e., $R_N \in \mathcal{O}\left(\sqrt{\Gamma_N^T N}\right)$. They obtain an exponential dependence because when planning optimistically, i.e., solving Equation (7), they consider all plausible dynamics, including those that are not Lipschitz continuous for all n. Solving Equation (7) for only continuous dynamics is intractable. However, for OMBRL, as we do not maximize over the set of dynamics we can overcome this limitation.

Moreover, since $\boldsymbol{f}^*$ has bounded RKHS norm, i.e., $\|\boldsymbol{f}^*\|_k \leq B$ (Assumption 5.3). From Srinivas et al. (2012); Chowdhury & Gopalan (2017) follows that with probability $1-\delta$ we have for every n:

$$\|\boldsymbol{f}^* - \boldsymbol{\mu}_n\|_{k_n} \leq \beta_n.$$

For OMBRL, instead of planning with the mean, which in general might not be Lipschitz continuous for all n, we select a function $\boldsymbol{f}_n$ that not only approximates the $\boldsymbol{f}^*$ function well, i.e., $\|\boldsymbol{f}^* - \boldsymbol{f}_n\|_{k_n} \leq \beta_n$, but also its RKHS norm does not grow with n. To do that we propose to solve the following quadratic optimization problem:

$$\begin{aligned} \boldsymbol{f}_n = & \operatorname*{arg\,min}_{\boldsymbol{f} \in \text{span}(k(\boldsymbol{x}_1,\cdot),\ldots,k(\boldsymbol{x}_n,\cdot))} \|\boldsymbol{f} - \boldsymbol{\mu}_n\|_{k_n} \\ & \text{s.t. } \|\boldsymbol{f}\|_k \leq B \end{aligned} \tag{22}$$

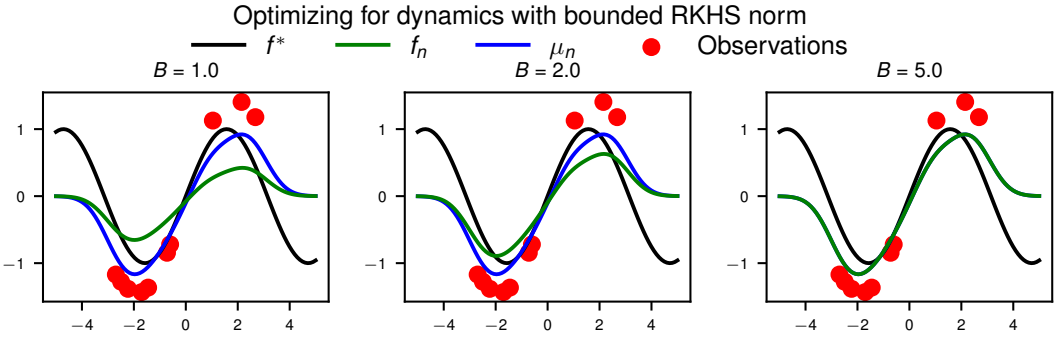

Figure 6: Solution to Equation (22) for different values for B. Effectively, for larger values for B, $\boldsymbol{\mu}_n$ and $\boldsymbol{f}_n$ coincide.

Theorem A.13. *The optimization problem Equation (22) is feasible and we have* $\|\boldsymbol{f}_n - \boldsymbol{\mu}_n\|_{k_n} \leq \beta_n$.

Proof. Consider the noise-free case, i.e., $\boldsymbol{w} = 0$, and let $\bar{\boldsymbol{\mu}}_n$ posterior mean for this setting. For the function $\bar{\boldsymbol{\mu}}_n$ holds that $\|\boldsymbol{f}^* - \bar{\boldsymbol{\mu}}_n\|_{k_n} \leq \beta_n$ (Corollary 3.11 of Kanagawa et al. (2018)) and $\|\bar{\boldsymbol{\mu}}_n\|_k \leq B$ (Theorem 3.5 of Kanagawa et al. (2018)). Since $\|\bar{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_n\|_{k_n} \leq \|\bar{\boldsymbol{\mu}}_n - \boldsymbol{f}^*\|_{k_n} + \|\boldsymbol{f}^* - \boldsymbol{\mu}_n\|_{k_n} \leq 2\beta_n$. By representer theorem, it also holds that $\bar{\boldsymbol{\mu}}_n \in \text{span}(k(\boldsymbol{z}_1,\cdot),\ldots,k(\boldsymbol{z}_n,\cdot))$. □

Let $\boldsymbol{\alpha}_n = (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{y} \in \mathbb{R}^n$ and reparametrize $\boldsymbol{f}(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})$. We have $\|\boldsymbol{f}\|_k^2 = \boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}$. We also have:

$$\|\boldsymbol{f} - \boldsymbol{\mu}_n\|_{k_n}^2 = (\boldsymbol{\alpha} - \boldsymbol{\alpha}_n)^\top \boldsymbol{K}\left(\boldsymbol{I} + \frac{1}{\sigma^2}\boldsymbol{K}\right)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_n)$$

Hence the optimization problem Equation (22) is equivalent to:

$$\begin{aligned} & \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_n)^\top \boldsymbol{K}\left(\boldsymbol{I} + \frac{1}{\sigma^2}\boldsymbol{K}\right)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_n) \\ & \text{s.t. } \boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha} \leq B^2 \end{aligned}$$

This is a quadratic program that can be solved using any QP solver. The program finds the closest function to the posterior mean $\boldsymbol{\mu}_n$ that is Lipschitz continuous (see Figure 6). In particular, note that since $\|\boldsymbol{f}_n\|_k \leq B$, $\boldsymbol{f}_n$ has a Lipschitz constant L_B which is independent of n (Berkenkamp, 2019). From hereon, let $L_* = \max\{L_f, L_B\}$.

For the sub-Gaussian case, OMBRL follows the same strategy as Equation (8) but instead of using the mean dynamics $\boldsymbol{\mu}_n$, we plan with the dynamics $\boldsymbol{f}_n$ that are obtained from solving Equation (22).

$$\begin{aligned}\boldsymbol{\pi}_n &= \arg\max_{\boldsymbol{\pi}\in\Pi} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}'_t, \boldsymbol{u}_t) + \lambda_n \|\boldsymbol{\sigma}_n(\boldsymbol{x}'_t, \boldsymbol{u}_t)\|\right] \\ \boldsymbol{x}'_{t+1} &= \boldsymbol{f}_n(\boldsymbol{x}'_t, \boldsymbol{u}_t) + \boldsymbol{w}_t,\end{aligned} \tag{23}$$

Lemma A.14. *Let Assumption A.12 and Assumption 5.3 hold. Consider the following definitions*

$$\begin{aligned}
&J(\boldsymbol{\pi}, \boldsymbol{f}^*) = E\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\right] \\
&\textit{s.t., } \boldsymbol{x}_{t+1} = \boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}_0 = \boldsymbol{x}(0), \\
&J(\boldsymbol{\pi}, \boldsymbol{f}_n) = E\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right] \\
&\textit{s.t., } \boldsymbol{x}'_{t+1} = \boldsymbol{f}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}'_0 = \boldsymbol{x}(0), \\
&\Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}^*) = E\left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t))\|\right] \\
&\textit{s.t., } \boldsymbol{x}_{t+1} = \boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}_0 = \boldsymbol{x}(0), \\
&\Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}_n) = E\left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\|\right] \\
&\textit{s.t., } \boldsymbol{x}'_{t+1} = \boldsymbol{f}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) + \boldsymbol{w}_t, \quad \boldsymbol{x}'_0 = \boldsymbol{x}(0), \\
&\lambda_n = (1 + d_x) L_r (1 + L_{\boldsymbol{\pi}}) \bar{L}_*^{T-1} T \beta_n.
\end{aligned}$$

Then we have for all $n \geq 0$, $\boldsymbol{\pi} \in \Pi$ with probability at least $1 - \delta$

$$\begin{aligned}
|J(\boldsymbol{\pi}, \boldsymbol{f}^*) - J(\boldsymbol{\pi}, \boldsymbol{f}_n)| &\leq \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}_n) \\
|J(\boldsymbol{\pi}, \boldsymbol{f}^*) - J(\boldsymbol{\pi}, \boldsymbol{f}_n)| &\leq \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{f}^*)
\end{aligned}$$

Proof.

$$\begin{aligned}
&|J(\boldsymbol{\pi}, \boldsymbol{f}^*) - J(\boldsymbol{\pi}, \boldsymbol{f}_n)| \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) - r(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right] \\
&\leq L_r(1 + L_{\boldsymbol{\pi}}) \mathbb{E}\left[\sum_{t=0}^{T-1} \|\boldsymbol{x}_t - \boldsymbol{x}'_t\|\right]
\end{aligned}$$

Next we analyze $\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|$ for any t. Without loss of generality, assume $L_* \geq 1$ and define $\bar{L}_* = L_*(1 + L_{\boldsymbol{\pi}})$.

We show that

$$\begin{aligned}
&\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}'_{t+1}\right\| \\
&\leq (1 + \sqrt{d_x})\beta_n \left(\sum_{k=0}^{t} \bar{L}_*^{t-k} \|\boldsymbol{\sigma}_n(\boldsymbol{x}'_k, \boldsymbol{\pi}(\boldsymbol{x}'_k))\|\right).
\end{aligned}$$

Consider $t = 1$

$$\begin{aligned}
\|\boldsymbol{x}_1 - \boldsymbol{x}'_1\| &= \|\boldsymbol{f}^*(\boldsymbol{x}'_0, \boldsymbol{\pi}(\boldsymbol{x}'_0)) - \boldsymbol{f}_n(\boldsymbol{x}'_0, \boldsymbol{\pi}(\boldsymbol{x}'_0))\| \\
&\leq (1 + \sqrt{d_x})\beta_n \|\boldsymbol{\sigma}_n(\boldsymbol{x}'_0, \boldsymbol{\pi}(\boldsymbol{x}'_0))\|
\end{aligned}$$

Consider any $t > 1$,

$$\begin{aligned}
&\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}'_{t+1}\right\| \\
&= \|\boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) - \boldsymbol{f}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\| \\
&\leq \|\boldsymbol{f}^*(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t)) - \boldsymbol{f}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\| \\
&+ \|\boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) - \boldsymbol{f}^*(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\|
\end{aligned}$$

$$
\begin{aligned}
&\leq (1+\sqrt{d_x})\beta_n \left\|\boldsymbol{\sigma}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right\| + \bar{L}_* \left\|\boldsymbol{x}_t - \boldsymbol{x}'_t\right\| \\
&\leq (1+\sqrt{d_x})\beta_n \left(\|\boldsymbol{\sigma}_n(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\|\right) \\
&+ (1+\sqrt{d_x})\beta_n \left(\bar{L}_* \left(\sum_{k=0}^{t-1} \bar{L}_*^{t-1-k} \left\|\boldsymbol{\sigma}_n(\boldsymbol{x}'_k, \boldsymbol{\pi}(\boldsymbol{x}'_k))\right\|\right)\right) \\
&= (1+\sqrt{d_x})\beta_n \left(\sum_{k=0}^{t} \bar{L}_*^{t-k} \left\|\boldsymbol{\sigma}_n(\boldsymbol{x}'_k, \boldsymbol{\pi}(\boldsymbol{x}'_k))\right\|\right)
\end{aligned}
$$

In particular, since $\bar{L}_* \geq 1$, we have $\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}'_{t+1}\right\| \leq (1+\sqrt{d_x})\beta_n \bar{L}_*^t \left(\sum_{k=0}^{t-1} \left\|\boldsymbol{\sigma}_n(\boldsymbol{x}'_k, \boldsymbol{\pi}(\boldsymbol{x}'_k))\right\|\right)$.

In summary, we have

$$
\begin{aligned}
&|J(\boldsymbol{\pi}, \boldsymbol{f}^*) - J(\boldsymbol{\pi}, \boldsymbol{\mu}_n)| \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1} r(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_t)) - r(\boldsymbol{x}'_t, \boldsymbol{\pi}(\boldsymbol{x}'_t))\right] \\
&\leq L_r(1+L_{\boldsymbol{\pi}})\mathbb{E}\left[\sum_{t=0}^{T-1} \left\|\boldsymbol{x}_t - \boldsymbol{x}'_t\right\|\right] \\
&\leq L_r(1+L_{\boldsymbol{\pi}})(1+\sqrt{d_x}) \\
&\times \mathbb{E}\left[\sum_{t=0}^{T-1} \beta_n \bar{L}_*^{t-1} \left(\sum_{k=0}^{t-1} \left\|\boldsymbol{\sigma}_n(\boldsymbol{x}'_k, \boldsymbol{\pi}(\boldsymbol{x}'_k))\right\|\right)\right] \\
&\leq (1+d_x)L_r(1+L_{\boldsymbol{\pi}})\bar{L}_*^{T-1} T \beta_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{\mu}_n) \\
&= \lambda_n \Sigma_n(\boldsymbol{\pi}, \boldsymbol{\mu}_n).
\end{aligned}
$$

□

The main difference between our analysis and the analysis from Curi et al. (2020) is that for us $\lambda_n \propto \beta_n$ if we plan with $\boldsymbol{f}_n$.

Lemma A.15. *Let Assumption A.12 and Assumption 5.3 hold and consider the simple regret at episode n, $r_n = J(\boldsymbol{\pi}^*, \boldsymbol{f}^*) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*)$. The following holds for all $n > 0$ with probability at least $1-\delta$*

$$
r_n \leq (2\lambda_n + \lambda_n^2)\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*)
$$

Proof.

$$
\begin{aligned}
r_n &= J(\boldsymbol{\pi}^*, \boldsymbol{f}^*) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*) \\
&\leq J(\boldsymbol{\pi}^*, \boldsymbol{f}_n) + \lambda_n \Sigma_n(\boldsymbol{\pi}^*, \boldsymbol{f}_n) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*) && \text{(Lemma A.14)} \\
&\leq J(\boldsymbol{\pi}_n, \boldsymbol{f}_n) + \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}_n) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*) && \text{(Equation (23))} \\
&= J(\boldsymbol{\pi}_n, \boldsymbol{f}_n) - J(\boldsymbol{\pi}_n, \boldsymbol{f}^*) + \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}_n) \\
&\leq \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*) + \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}_n) && \text{(Lemma A.14)} \\
&= 2\lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*) + \lambda_n(\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}_n) - \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*)) \\
&\leq (\lambda_n^2 + 2\lambda_n)\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*).
\end{aligned}
$$

□

Theorem A.16 (Finite horizon setting sub-Gaussian case)**.** *Let Assumption A.12 and Assumption 5.3 hold. Then we have $\forall N > 0$ with probability at least $1-\delta$*

$$
R_N \leq \mathcal{O}\left(\Gamma_N^{3/2}\sqrt{N}\right).
$$

Proof. The proof is the same as for Theorem 5.6, since in Lemma A.15 we show that also for the sub-Gaussian case, OMBRL has the same regret dependence w.r.t. λ_n and $\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{f}^*)$. □

B ADDITIONAL EXPERIMENTS

In this section, we provide additional experiments.

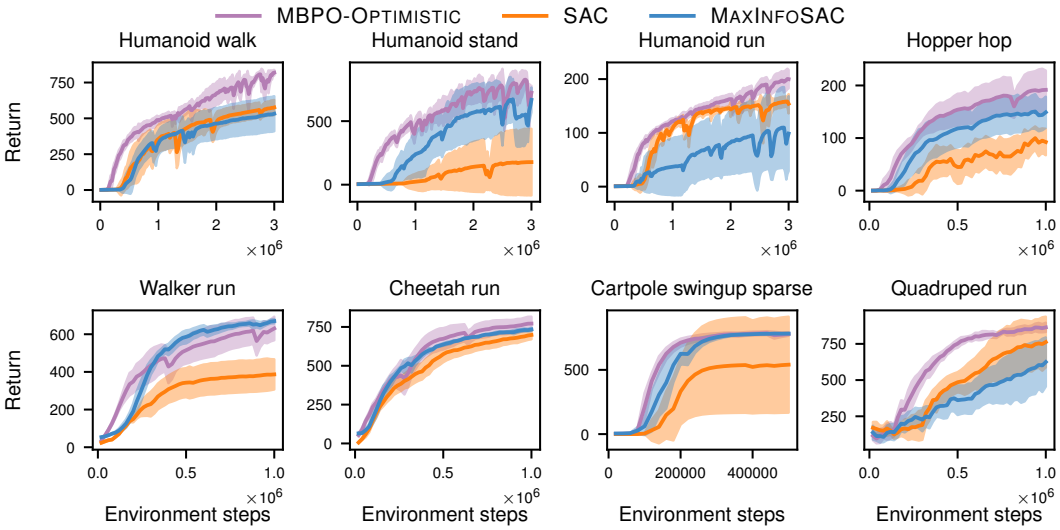

Figure 7: Comparison between MBPO-OPTIMSTIC and MAXINFOSAC and SAC. We observe that MBPO-OPTIMSTIC, being an MBRL algorithm, performs the best in terms of sample efficiency.

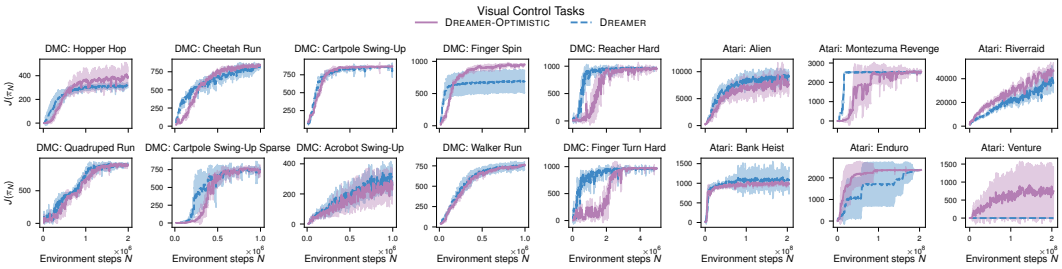

Figure 8: Learning curves for the visual control tasks from DMC and Atari using DREAMER as the base algorithm. DREAMER-OPTIMISTIC either performs on-par or better than DREAMER in all our experiments. Particularly, in the Venture task from the Atari benchmark, where DREAMER fails to obtain any rewards.

Experiments with MBPO In Figure 7 we compare MBPO-OPTIMISTIC with off-policy RL algorithms MAXINFOSAC (Sukhija et al., 2024a) and SAC (Haarnoja et al., 2018). From the figure, we conclude that MBPO-OPTIMISTIC performs the best in terms of sample-efficiency, particularly for the challenging/high-dimensional humanoid tasks. Moreover, between SAC and MAXINFOSAC, the latter achieves much better performance. We believe this is due to its intrinsic exploration reward.

Experiments with DREAMER In Figure 8 we compare DREAMER-OPTIMISTIC with DREAMER on additional environments. Overall, we observe that DREAMER-OPTIMISTIC performs either on par or better than DREAMER. However, for certain environments such as Reacher Hard or Finger Turn Hard, DREAMER is more sample-efficient. We believe this is because in these settings smaller values for λ_n would suffice for exploration. However, we use a constant value for λ_n across all environments and automatically update it using the approach proposed in Sukhija et al. (2024a). Investigating alternative strategies for λ_n, would generally benefit OMBRL methods. We think this is a promising direction for future work.

In Figure 9 and Figure 10 we compare DREAMER-OPTIMISTIC with DREAMER on proprioceptive tasks. In most environments, DREAMER-OPTIMISTIC performs on par. It performs better in the Finger Spin environment. However, when action costs are introduced (Figure 10), in line with our results in Section 6, DREAMER fails to obtain any meaningful rewards.

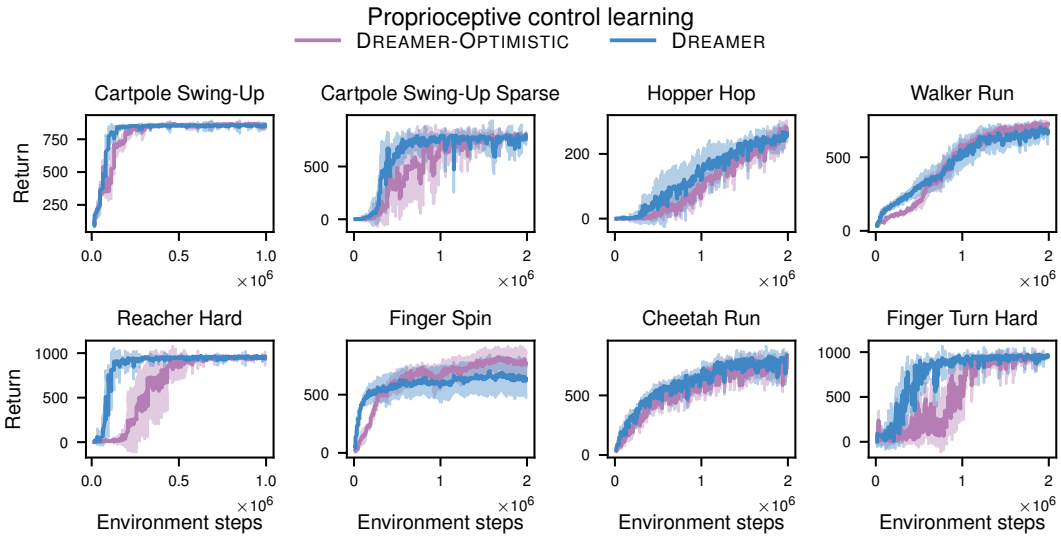

Figure 9: Experiments with DREAMER-OPTIMISTIC and DREAMER for proprioceptive tasks. DREAMER-OPTIMISTIC performs on par with DREAMER, obtaining slightly better performance on the Finger Spin task.

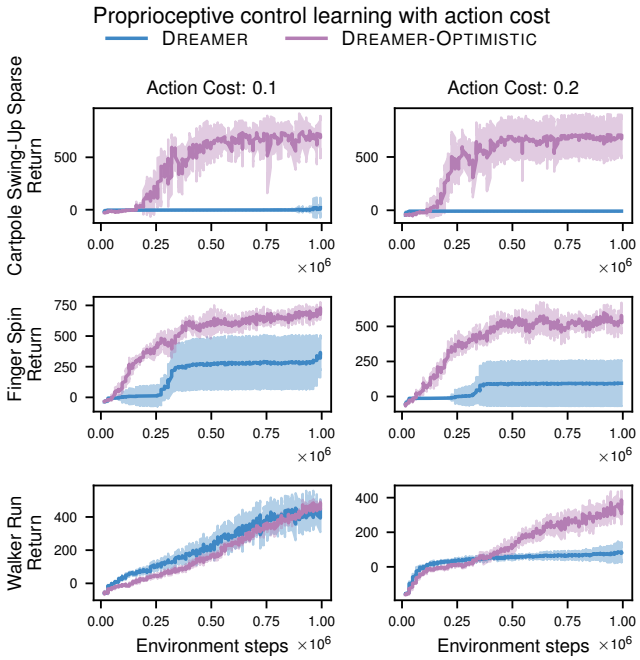

Figure 10: Experiments with DREAMER-OPTIMISTIC and DREAMER for proprioceptive tasks with action costs. DREAMER completely fails to solve the task, whereas DREAMER-OPTIMISTIC does not.

C EXPERIMENT DETAILS

In this section, we provide additional details for our experiments.

C.1 MBPO-OPTIMISTIC

For MBPO-OPTIMISTIC, we train an ensemble of forward dynamics models[4]. We use the disagreement between the ensembles to quantify model epistemic uncertainty, similar to Pathak et al. (2019); Curi et al. (2020); Sukhija et al. (2024c). For selecting λ_n, we use the auto-tuning approach from Sukhija et al. (2024a), where the intrinsic reward weight is optimized by minimizing the following loss with stochastic gradient descent

$$L(\lambda) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{1:n},\boldsymbol{u}\sim\boldsymbol{\pi}_n,\bar{\boldsymbol{u}}\sim\bar{\boldsymbol{\pi}}_n} \log(\lambda)(\boldsymbol{\sigma}_n(\boldsymbol{x},\boldsymbol{u}) - \boldsymbol{\sigma}_n(\boldsymbol{x},\bar{\boldsymbol{u}})). \tag{24}$$

Here $\bar{\boldsymbol{\pi}}_n$ is a target policy, which is updated using polyak updates of $\boldsymbol{\pi}_n$. This objective increases λ when the policy is under exploring compared to the target policy. Sukhija et al. (2024a) show that this strategy works across several model-free off-policy RL algorithms.

Besides using the model to quantify disagreement, we generate additional data by adding the transitions predicted by our learned model. In particular, for every policy update, we sample a batch of transitions from the data buffer $(\boldsymbol{x},\boldsymbol{u},\boldsymbol{x}') \sim \mathcal{D}_{1:n}$, and add $(\boldsymbol{x},\boldsymbol{u},\hat{\boldsymbol{x}}')$, transitions predicted by our mean model $\boldsymbol{\mu}_n$, to the batch. This allows us to combine true rollouts with model generated rollouts, as proposed in Janner et al. (2019). Since we can generate additional data through our learned model, we can efficiently increase our update-to-data ratio (UTD). For all our experiments with MBPO, with use an UTD of 5[5].

We use the same hyperparameters as Sukhija et al. (2024a) for all our state-based experiments.

C.2 DREAMER-OPTIMISTIC

We use DREAMERV3 As the base model. For quantifying the model epistemic uncertainty, we use the same approach as Sekar et al. (2020); Mendonca et al. (2021) and learn an ensemble of MLPs to model the latent dynamics[6]. The ensemble is only used for quantifying the model uncertainty/intrinsic reward. For the policy optimization, we use the DREAMER backbone, where the agent optimizes the policy using imagined rollouts. For selecting λ, we also use the objective in Equation (24). We found adding a regularize term $\alpha * |\lambda|$ to the objective worked better with DREAMER. We initialize λ with 2 and pick $\alpha = 0.001$. For the rest, we use the same hyperparemters as DREAMER[7].

C.3 SIMFSVGD-OPTIMISTIC

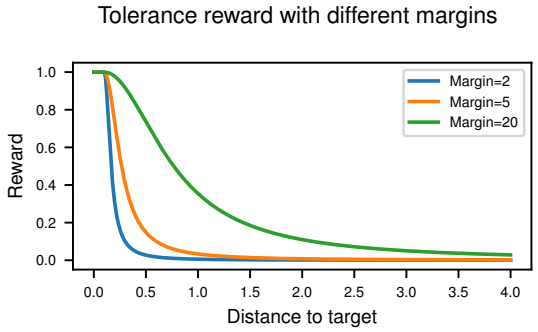

Figure 11: Tolerance reward function for different values of the margin. For larger margins, the agent receives rewards even if its further away from the target.

We use the same experiment setup, simulation prior, and hyperparameters as Rothfuss et al. (2024)[8]. The reward function in Rothfuss et al. (2024) is based on the tolerance reward from Tassa et al. (2018). The tolerance function, gives higher rewards when the agent is close to a desired state, i.e., in case of the RC car the target position. The "closesness" is quantified using a margin parameter for the reward

[4]For all tasks we use a $(256, 256)$ neural network architecture with 5 ensembles, except for the humanoid and quadruped tasks where we use $(512, 512)$.

[5]We did not tune the UTD and chose 5 to trade-off between computational cost and sample efficiency.

[6]For all tasks we use a $(512, 512)$ neural network architecture with 5 ensembles.

[7]We use the 12 million size model and the official DREAMERV3 implementation (https://github.com/danijar/dreamerv3/tree/main).

[8]official implementation: https://github.com/lasgroup/simulation_transfer

function. In Figure 11 we plot the reward for different margin parameters. As we decrease the margin, the reward becomes sparser. Rothfuss et al. (2024) use a margin of 20. In our simulation experiments, we show that SIMFSVGD performs worse than SIMFSVGD-OPTIMISTIC for smaller margins. For our hardware experiment, we use a margin of 5, for which SIMFSVGD fails to learn. For λ_n we found that a linearly decaying schedule worked the best. Therefore, we linearly interpolated from $\lambda_0 = 0.5$ and $\lambda_{10} = 0$. After the tenth episode, the agent greedily maximized the extrinsic reward.

C.4 GP EXPERIMENTS

For our GP experiments, we use the RBF kernel. The kernel parameters are updated online using maximum likelihood estimation (Rasmussen & Williams, 2005). For all the experiments, we use $\lambda_n = 10$ and for planning the iCEM optimizer (Pinneri et al., 2021). We use the same hyperparameters as Sukhija et al. (2024b)[9].

[9] official implementation: https://github.com/lasgroup/opax