
SOMBRL: Scalable and Optimistic Model-Based RL

Bhavya Sukhija
Department of Computer Science
ETH Zurich
sukhijab@ethz.ch

Lenart Treven
Department of Computer Science
ETH Zurich
trevenl@ethz.ch

Carmelo Sferrazza
Berkeley AI Research
UC Berkeley
csferrazza@berkeley.edu

Florian Dörfler
Department of Electrical Engineering
ETH Zurich
dorfler@ethz.ch

Pieter Abbeel
Berkeley AI Research
UC Berkeley
pabbeel@berkeley.edu

Andreas Krause
Department of Computer Science
ETH Zurich
krausea@ethz.ch

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 **Scalable and Optimistic MBRL (SOMBRL)**, an approach based on the principle of optimism in the face of uncertainty. SOMBRL learns an uncertainty-aware dynamics model and *greedily* maximizes a weighted sum of the extrinsic reward and the agent’s epistemic uncertainty. SOMBRL is compatible with any policy optimizers or planners, and under common regularity assumptions on the system, we show that SOMBRL has sublinear regret for nonlinear dynamics in the (i) finite-horizon, (ii) discounted infinite-horizon, and (iii) non-episodic settings. Additionally, SOMBRL offers a flexible and scalable solution for principled exploration. We evaluate SOMBRL on state-based and visual-control environments, where it displays strong performance across all tasks and baselines. We also evaluate SOMBRL on a dynamic RC car hardware and show SOMBRL outperforms the state-of-the-art, illustrating the benefits of principled exploration for MBRL.

1 Introduction

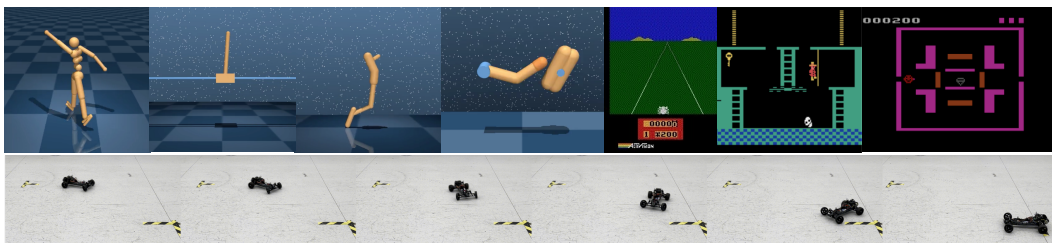


Figure 1: *Top*: We showcase scalability of SOMBRL on visual control tasks from DMC and Atari. *Bottom*: We evaluate SOMBRL on a highly dynamic RC car where we learn to perform a complex parking maneuver in only 20 real-world episodes.

Reinforcement learning (RL) has been successfully applied to a variety of sequential-decision making problems like games (Silver et al., 2017), robotics (Kober et al., 2013; Tang et al., 2025), 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.

Related Works 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); Foster & Rakhlin (2023); Wagenmaker et al. (2023); Sukhija et al. (2024c), see Section 3 and Section A 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, SOMBRL, for exploration that enjoys convergence guarantees across several RL settings. This is in contrast to prior works, which design and study algorithms only for specific settings, e.g., Curi et al. (2020); Kakade & Langford (2002) study the finite-horizon setting, Sukhija et al. (2024c) the unsupervised RL setting, and Sukhija et al. (2024b) the nonepisodic one. In addition to the theoretical guarantees, SOMBRL is also much more scalable and, unlike the aforementioned works, can be applied to high-dimensional settings such as visual control problems and real-world hardware platforms. We demonstrate this in our experiments (Section 6).

The differences between SOMBRL and prior work on exploration in MBRL are summarized in Table 1, and we provide a more detailed discussion of related works in Section A.

Contributions

1. We propose SOMBRL, a principled yet efficient exploration strategy for model-based RL. SOMBRL is based on the principle of optimism in the face of uncertainty, and *greedily* 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 where the uncertainty about the system is high.
2. We show, under common regularity assumptions on the dynamics, 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 SOMBRL 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 SOMBRL 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 *evaluate* SOMBRL *in the real-world* on a dynamic RC car (see Figure 1), where it learns an agile parking maneuver in only 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 deep RL for high-dimensional and real-world settings.

2 Problem Setting

We consider a discrete-time dynamical system of the form $\mathbf{x}_{t+1} = \mathbf{f}^*(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{w}_t$, where $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ is the state, $\mathbf{u}_t \in \mathcal{U} \subseteq \mathbb{R}^{d_u}$ the control input, and $\mathbf{w}_t \in \mathcal{W} \subseteq \mathbb{R}^w$ the process noise¹. The dynamics $\mathbf{f}^*$ are unknown.

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

Table 1: Comparison of SOMBRL and prior work on model-based RL (MBRL) in terms of sublinear regret/sample complexity for the kernelized settings and scalability. Kakade et al. (2020) give regret bound for the finite-horizon setting with Gaussian noise and Curi et al. (2020) for the sub-Gaussian noise case. We show that SOMBRL has sublinear regret for both settings (Theorem 5.4 and Theorem B.18)

	Finite Horizon	γ -discounted infinite-horizon	Nonepisodic	Unsupervised RL	Scalable + Practical
Greedy, e.g., Mean planning or Deisenroth & Rasmussen (2011); Chua et al. (2018)	✗	✗	✗	✗	✓
Curi et al. (2020); Kakade et al. (2020)	✓	✗	✗	✗	✗
Sukhija et al. (2024b)	✗	✗	✓	✗	✗
Sukhija et al. (2024c)	✗	✗	✗	✓	✗
SOMBRL (ours)	✓	✓	✓	✓	✓

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

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

where action $\mathbf{u}_t$ follows policy π , i.e., $\mathbf{u}_t \sim \pi(\mathbf{x}_t)$. 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 π_n for T steps on the true system. We then use the data collected from the rollouts to estimate the true dynamics $\mathbf{f}^*$. The goal is to find a policy that performs as well as π^* , as quickly as possible. Therefore a natural performance metric in this context is the *cumulative regret* $R_N = \sum_{n=1}^N J(\pi^*) - J(\pi_n)$. 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 γ -discounted infinite-horizon, episodic: and average reward, nonepisodic settings:

$$J_{\gamma}(\pi^*) = \max_{\pi \in \Pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{x}_t, \mathbf{u}_t) \right] \quad (2) \quad J_{\text{avg}}(\pi^*) = \max_{\pi \in \Pi} \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} r(\mathbf{x}_t, \mathbf{u}_t) \right] \quad (3)$$

3 Exploration Strategies in MBRL

In MBRL, we learn a model of the true dynamics $\mathbf{f}^*$ 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 $J(\pi, \mathbf{f})$ be the the expected returns under the policy π and dynamics $\mathbf{f}$, that is

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

and μ_n our mean estimate of the dynamics $\mathbf{f}^*$ at episode n .

Greedy planning The simplest selection strategy is to pick the policy π_n that maximizes the expected returns for our estimated dynamics μ_n .

$$\pi_n^{\text{MEAN}} = \arg \max_{\pi \in \Pi} J(\pi, \mu_n) \quad (4)$$

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 μ_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 $\mathbf{f}^*$: $p(\mathbf{f} | \mathcal{D}_{1:n})$. Here $\mathcal{D}_{1:n} = \cup_{i \leq n} \mathcal{D}_i$, and $\mathcal{D}_i = \{(\mathbf{x}_{t,i}, \mathbf{u}_{t,i}, \mathbf{x}_{t+1,i})\}_{t=0}^{T-1}$ is the data collected in episode i . The policy π_n is then selected as

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

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 $p(\mathbf{f}|\mathcal{D}_{1:n})$ and pick policies by maximizing the reward under $\mathbf{f}$ sampled from the posterior

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

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

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 $p(\mathbf{f}|\mathcal{D}_{1:n})$. The policy is then selected according to

$$\pi_n^{\text{OE}} = \arg \max_{\pi \in \Pi, \mathbf{f} \in \mathcal{M}_n} J(\pi, \mathbf{f}) \quad (7)$$

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 $\mathbf{f}$ 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_u to $d_u + d_x$, which is prohibitive in high-dimensional domains.

4 SOMBRL: Scalable and Optimistic MBRL

We now present SOMBRL, 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 $\mathbf{f}^*$. That is, after each episode n , we learn a mean estimate μ_n of $\mathbf{f}^*$ and quantify our epistemic uncertainty σ_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 π_n

$$\pi_n := \arg \max_{\pi \in \Pi} \mathbb{E}_{\pi} \left[\underbrace{\sum_{t=0}^{T-1} r(\mathbf{x}'_t, \mathbf{u}_t) + \lambda_n \|\sigma_n(\mathbf{x}'_t, \mathbf{u}_t)\|}_{J_n(\pi)} \right], \mathbf{x}'_{t+1} = \mu_n(\mathbf{x}'_t, \mathbf{u}_t) + \mathbf{w}_t, \quad (8)$$

where λ_n is a positive constant which is used to trade off maximizing the extrinsic reward and model uncertainty (see Section B for how λ_n is defined in theory and Section 5.2 and Section D 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 SOMBRL much simpler and more scalable. Moreover, SOMBRL can be combined with any model-based algorithm such as those of 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 and for $\lambda \rightarrow \infty$ the agent performs unsupervised exploration (Pathak et al., 2017; Sekar et al., 2020; Buisson-Fenet et al., 2020; Sukhija et al., 2024c). Therefore, we use the model uncertainty to facilitate principled exploration for the agent.

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

5 Theoretical Results

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

5.1 Assumptions

We first make continuity assumptions on the system. These assumptions are common in the control theory (Khalil, 2015) and reinforcement learning literature (Curi et al., 2020; Sussex et al., 2023; Sukhija et al., 2024c).

Assumption 5.1 (Continuous closed-loop dynamics, bounded rewards, and Gaussian noise.). The dynamics model $\mathbf{f}^*$ and all $\pi \in \Pi$ are continuous. Furthermore, we assume that the reward is bounded, i.e., $r : \mathcal{X} \times \mathcal{U} \rightarrow [0, R_{\max}]$, and process noise is i.i.d. Gaussian² with variance σ^2 , i.e., $\mathbf{w}_t \stackrel{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

SOMBRL learns an uncertainty-aware model of the true dynamics $\mathbf{f}^*$, i.e., a mean μ_n and uncertainty σ_n estimate. In our theoretical analysis, we focus on Gaussian Process (GP) dynamics models. For GPs, μ_n and σ_n have a closed-form solution, and our learned model is calibrated (c.f., Rothfuss et al. (2023) or Section B.1 for the definition of well-calibrated models or Kuleshov et al. (2018) on calibration for BNNs). Generally, our guarantees can be extended to broader classes of well-calibrated models, e.g., BNNs (similar to Curi et al. (2020)).

We assume that $\mathbf{f}^*$ resides in a Reproducing Kernel Hilbert Space (RKHS) of vector-valued functions.

Assumption 5.2. We assume that the functions $\mathbf{f}_j^*$, $j \in \{1, \dots, d_x\}$ lie in a RKHS with kernel k and have a bounded norm B , that is $\mathbf{f}^* \in \mathcal{H}_{k,B}^{d_x}$, with $\mathcal{H}_{k,B}^{d_x} = \{\mathbf{f} \mid \|\mathbf{f}_j\|_k \leq B, j = 1, \dots, d_x\}$. Moreover, we assume that $k(\mathbf{z}, \mathbf{z}) \leq \sigma_{\max}$ for all $\mathbf{x} \in \mathcal{X}$.

Theorem 5.2 ensures that we can model and learn $\mathbf{f}^*$ with GPs. This assumption is common in the Bayesian optimization (Srinivas et al., 2012; Chowdhury & Gopalan, 2017) and RL literature (Kakade et al., 2020; Curi et al., 2020). Moreover, GPs are nonparametric models and can learn very complex classes of nonlinear functions (Rasmussen & Williams, 2005).

The posterior mean $\mu_n(\mathbf{z}) = [\mu_{n,j}(\mathbf{z})]_{j \leq d_x}$ and epistemic uncertainty $\sigma_n(\mathbf{z}) = [\sigma_{n,j}(\mathbf{z})]_{j \leq d_x}$ can then be obtained using the following formula

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

Here, $\mathbf{y}_{1:n}^j$ corresponds to the noisy measurements of $\mathbf{f}_j^*$, i.e., the observed next state from the transitions dataset $\mathcal{D}_{1:n}$, $\mathbf{k}_n(\mathbf{z}) = [k(\mathbf{z}, \mathbf{z}_i)]_{\mathbf{z}_i \in \mathcal{D}_{1:n}}$, and $\mathbf{K}_n = [k(\mathbf{z}_i, \mathbf{z}_l)]_{\mathbf{z}_i, \mathbf{z}_l \in \mathcal{D}_{1:n}}$ is the data kernel matrix. The restriction on the kernel $k(\mathbf{z}, \mathbf{z}) \leq \sigma_{\max}$ implies boundedness of $\mathbf{f}^*$ 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 $\mathbf{f}^*$ such that $\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{f}^*(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}) + \mathbf{w}_{k-1}$ in which case the boundedness of $\mathbf{f}^*$ captures many real-world systems.

Our theoretical results depend on the *maximum information gain* of kernel k (Srinivas et al., 2012), defined as

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

Γ_N is a measure of the complexity for learning $\mathbf{f}^*$ from N episodes and is sublinear for many kernels (e.g., $\mathcal{O}(\log^{d_x+d_u+1}(N))$ for the squared exponential (RBF) kernel, $\mathcal{O}((d_x + d_u) \log(N))$ for the linear kernel). In Section B, we report the dependence of Γ_N on N in Table 2.

Next, we present the following Lemma, which states that $J_n(\pi_n)$ from Equation (8) is an optimistic estimate of $J(\pi^*)$.

Lemma 5.3. *Let Theorem 5.1 and Theorem 5.2 hold. Then, there exists a $\lambda_n \in \Theta(\sqrt{\Gamma_N})$, such that we have $\forall n > 0$, $\pi \in \Pi$, with probability at least $1 - \delta$, that $J(\pi) \leq J_n(\pi)$. Moreover, we have $J(\pi^*) \leq J_n(\pi_n)$.*

Proof Sketch: Crucially, we leverage the policy difference lemma (Kakade & Langford, 2002), which bounds the difference between the performance of two different policies with the advantage function. However, we study this lemma for a fixed policy but different dynamics, effectively obtaining a simulation lemma (Kearns & Singh, 2002) for our setting. Next, we bound the difference

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

in performance between the mean dynamics and the true one and show that this is proportional to the model epistemic uncertainty.

Theorem 5.3 shows that for all policies $\pi \in \Pi$, $J_n(\pi)$ gives an upper bound on the true return $J(\pi)$. 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 Theorem B.3 in Section B.

Finally, we present our main theorem, which bounds the regret of SOMBRL.

Theorem 5.4 (Finite horizon setting). *Let Theorem 5.1 and Theorem 5.2 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 Sketch: To bound the regret, we first prove that Equation (8) is an optimistic estimate of Equation (1). Then, we use Theorem 5.3 to bound the difference in performance for extrinsic rewards between the mean and true dynamics with the epistemic uncertainty of the collected rollout. Next, we analyse the intrinsic reward term and also show that it's bounded by the epistemic uncertainty of the collected rollout. This allows us to relate the cumulative regret R_N with the information gain Γ_N and obtain the final bound.

Theorem 5.4 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, SOMBRL is a much simpler and more scalable algorithm. In Section B, we show that SOMBRL improves the regret bound from Curi et al. (2020), for the sub-Gaussian noise case, by an exponential factor of Γ_N^T . Below, we also provide our regret bounds for the γ -discounted and the non-episodic setting³.

Theorem 5.5 (γ -discounted, infinite horizon setting). *Let $R_N = \sum_{n=1}^N J_\gamma(\pi^*) - J_\gamma(\pi_n)$. Under the Theorem 5.1 and Theorem 5.2, we have for the γ -discounted infinite (Equation (2)) horizon setting $\forall N > 0$ that with probability at least $1 - \delta$, $R_N \leq \mathcal{O}\left(\Gamma_N^{3/2} \sqrt{N}\right)$.*

Proof Sketch: The regret decomposition for this setting is similar to the finite-horizon case (Theorem 5.4). However, in contrast to the finite-horizon case, where each episode has a fixed length T , we care about the infinite horizon in the γ -discounted case. Therefore, to obtain sublinear regret for this setting, we require the agent to observe the system for longer horizons, i.e., the horizon $T(n) \rightarrow \infty$ for $n \rightarrow \infty$. We ensure this by picking $T(n) \in \Theta(\log(n))$ and show that this is sufficient to obtain sublinear regret.

In Theorem 5.5, we show that even though we truncate each episode after $T(n)$ steps, SOMBRL 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. As pointed out in Kakade (2003); Sharma et al. (2021), this is the most challenging and closest setting for learning directly in the real-world. Sukhija et al. (2024b) show that optimistic exploration methods have sublinear regret for the nonepisodic setting. However, their proposed algorithm is intractable in practice.

Theorem 5.6 (Informal statement; nonepisodic average reward case). *Let $R_N = \sum_{n=1}^N \mathbb{E}[J_{\text{avg}}(\pi^*) - r(x_n, \pi_n(x_n))]$. 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 $1 - \delta$, $R_N \leq \mathcal{O}\left(\Gamma_N^{3/2} \sqrt{N}\right)$.*

Proof Sketch: The regret analysis for this setting is, in spirit, similar to the finite horizon case. However, in the nonepisodic setting, we cannot reset the agent and have to learn from a single trajectory. Therefore, unlike the episodic case, where we update our model and policy after every episode, in the nonepisodic setting we have to decide when to update the agent. For SOMBRL we show that if we only update once we have accumulated enough information, i.e., $\sum_{t=0}^{T(n)-1} \|\sigma_n(x_{t,n}, \pi_n(x_{t,n}))\| > C$, for a positive constant C , then SOMBRL has sublinear regret. Intuitively, the model uncertainty measures how much information our agent has acquired since its last update at time step $T(n-1)$. We only update the agent once the information exceeds the threshold C .

³Both Kakade et al. (2020) and Curi et al. (2020) do not provide a regret bound for these settings.

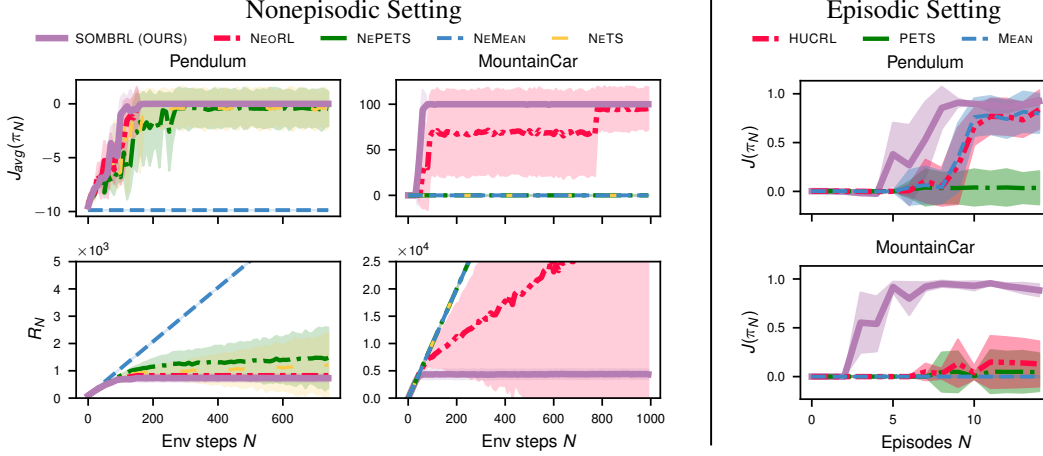


Figure 2: *Left:* Learning curves for the nonepisodic setting with GP dynamics. We report the average reward $J_{avg}(\pi_N)$ and regret R_N . The curves are reported with 5 seeds, and we plot the median return with its standard deviation. *Right:* Learning curves for the episodic setting with GP dynamics. We report the median episode reward $J(\pi_N)$ over an episode with 5 seeds and its standard deviation.

In contrast to Sukhija et al. (2024b), SOMBRL is much more tractable, and in Theorem 5.6 we show that SOMBRL 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 SOMBRL, which maximizes a combination of the extrinsic reward and the model epistemic uncertainty, enjoys sublinear regret for common kernels and RL settings. There are principled exploration algorithms designed individually for these settings, e.g., (Kakade et al., 2020; Curi et al., 2020) for the finite-horizon case and Sukhija et al. (2024b) for the non-episodic case. However, they are often intractable/computationally prohibitive. In contrast, SOMBRL works across the different RL settings while also being more practical and scalable.

We present additional theoretical results, for example, a sample complexity bound for unsupervised RL 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 Section B. Our detailed proofs are also provided in Section B.

5.2 Selecting λ_n in practice

The parameter λ_n controls the exploration-exploitation trade-off for SOMBRL. In Section B 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 Section D.

5.3 Application of SOMBRL with GP dynamics

Finally, we empirically validate our theoretical findings for the GP case in Figure 2, where we compare SOMBRL 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 SOMBRL performs the best across all baselines for both the episodic and the nonepisodic setting. Moreover, while HUCRL and NEORL, which explore according to Equation (7), perform better than other baselines, they are worse than SOMBRL. We believe this is because of the practical challenges associated with solving the optimization problem in Equation (7). Moreover, solving Equation (7) is also computationally more expensive. For instance, in our experiments, HUCRL requires roughly $3\times$ more compute time than OMBRL (see Section D.5).

6 Experiments

In our experiments, we showcase the flexibility and scalability of SOMBRL 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. Note that principled exploration methods such as Kakade et al. (2020); Curi et al. (2020) do not scale to the settings, such as visual control tasks, considered in this work. 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 dimensionality⁴. 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 Section D.

State-based experiments We refer to the MBPO version of SOMBRL 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 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 3. 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 Section C, 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 SOMBRL to challenging and high-dimensional problems by evaluating it on visual control tasks. We combine SOMBRL 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 3. 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. 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, $r_{\text{action}}(\mathbf{a}) = -K \|\mathbf{a}\|_2$, 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 4 (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 Section C.

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. We use the same experimental setup as Rothfuss et al. (2024).

First, we evaluate our algorithm SIMFSVGD-OPTIMISTIC, a combination of SOMBRL 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 Section D for more detail). We report the results in the top row of Figure 4. For the dense reward setting, SIMFSVGD and SIMFSVGD-OPTIMISTIC perform similarly, but for sparser rewards, SIMFSVGD-OPTIMISTIC outperforms SIMFSVGD. In particular, for the setting

⁴including the humanoid from DMC: $d_x = 67$, $d_u = 21$

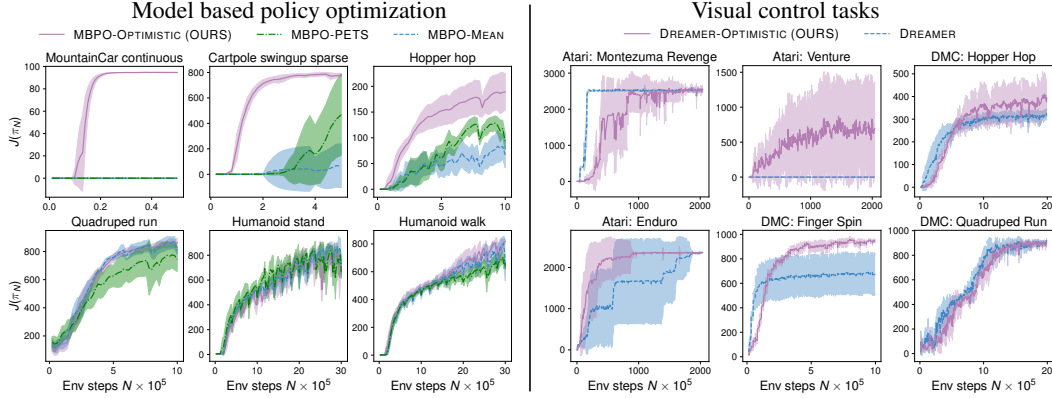


Figure 3: *Left*: Learning curves for the state-based tasks from DMC using MBPO as the base algorithm. Across all experiments, MBPO-OPTIMISTIC obtains the best performance compared to its greedy variants. MBPO-OPTIMISTIC also scales to high-dimensional tasks, specifically the humanoid environments from 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.

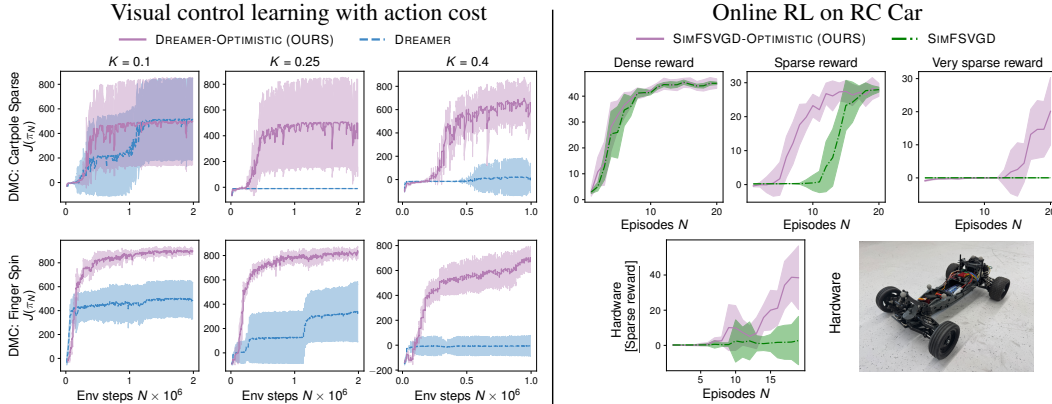


Figure 4: *Left*: Learning curves with action costs, where we compare DREAMER with DREAMER-OPTIMISTIC. DREAMER fails to explore sufficiently with action costs, whereas DREAMER-OPTIMISTIC is able to explore and obtain much higher performance. *Right*: Learning curves for our experiments with SIMFSVG. *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, SIMFSVG 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, SIMFSVG fails to solve the task.

with very sparse rewards, SIMFSVG 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 4. In line with the simulation experiments, we also observe similar behavior on hardware. SIMFSVGD-OPTIMISTIC learns to solve the task, whereas SIMFSVG completely fails. In fact, out of 5 attempts, SIMFSVG worked only once, and otherwise converged to a local optimum of not moving from the starting position⁵.

7 Conclusion

In this work, we propose SOMBRL, which maximizes a weighted sum of the extrinsic reward and the agent’s epistemic uncertainty. We show that SOMBRL effectively performs optimistic exploration and provide regret bounds for it in a variety of settings, in particular for continuous state-action spaces

⁵Video is available at <https://sukhijab.github.io/projects/sombtrl/>

and many common classes of RL problems, namely, finite-horizon, infinite-horizon, episodic, and non-episodic RL. To the best of our knowledge, we are the first to provide these theoretical guarantees, yielding a more flexible, scalable, and principled algorithm for exploration. We illustrate the strengths of SOMBRL in our experiments, where we combine SOMBRL 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, SOMBRL achieves the best performance, being significantly more scalable and computationally cheaper than prior optimistic exploration methods and stronger at exploration than SOTA deep RL baselines.

A limitation of SOMBRL is that it requires training a probabilistic model, e.g., deep ensembles, for quantifying epistemic uncertainty. We report the computational cost of our method in Section D.5 and show that the computational overhead is small compared to the training cost for the actor and critic.

Acknowledgments

We thank Scott Sussex for the insightful discussion and feedback on this work. This project has received funding from the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40 180545, the Microsoft Swiss Joint Research Center, and the SNSF Postdoc Mobility Fellowship 211086. Bhavya Sukhija was gratefully supported by ELSA (European Lighthouse on Secure and Safe AI) funded by the European Union under grant agreement No. 101070617.

References

- Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *Conference on Learning Theory*, 2011.
- Abeille, M. and Lazaric, A. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. In *International Conference on Machine Learning*, 2020.
- As, Y., Sukhija, B., Treven, L., Sferrazza, C., Coros, S., and Krause, A. Actsafe: Active exploration with safety constraints for reinforcement learning. *arXiv preprint arXiv:2410.09486*, 2024.
- Åström, K. J. and Wittenmark, B. Problems of identification and control. *Journal of Mathematical analysis and applications*, 34(1):90–113, 1971.
- Aubret, A., Matignon, L., and Hassas, S. An information-theoretic perspective on intrinsic motivation in reinforcement learning: A survey. *Entropy*, 2023.
- Auer, P. Finite-time analysis of the multiarmed bandit problem, 2002.
- Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. *NeurIPS*, 2016.
- Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.
- Berkenkamp, F. *Safe exploration in reinforcement learning: Theory and applications in robotics*. PhD thesis, ETH Zurich, 2019.
- Brafman, R. I. and Tenenbholz, M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *JMLR*, 2002.
- Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.
- Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., and Schoellig, A. P. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 2022.
- Buisson-Fenet, M., Solowjow, F., and Trimpe, S. Actively learning gaussian process dynamics. In *L4DC*. PMLR, 2020.
- Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.
- Cesa-Bianchi, N., Gentile, C., Lugosi, G., and Neu, G. Boltzmann exploration done right. *Advances in neural information processing systems*, 30, 2017.

- Chiuso, A., Fabris, M., Breschi, V., and Formentin, S. Harnessing uncertainty for a separation principle in direct data-driven predictive control. *arXiv preprint arXiv:2312.14788*, 2023.
- Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.
- Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *NeurIPS*, 2018.
- Cohen, A., Koren, T., and Mansour, Y. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, 2019.
- Curi, S., Berkenkamp, F., and Krause, A. Efficient model-based reinforcement learning through optimistic policy search and planning. *NeurIPS*, 33:14156–14170, 2020.
- Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.
- Deisenroth, M. and Rasmussen, C. E. Pilco: A model-based and data-efficient approach to policy search. In *ICML*, pp. 465–472, 2011.
- Even-Dar, E. and Mansour, Y. Convergence of optimistic and incremental q-learning. *NeurIPS*, 14, 2001.
- Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.
- Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, 2020.
- Foster, D. J. and Rakhlin, A. Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730*, 2023.
- Grimaldi, R. A., Baggio, G., Carli, R., and Pillonetto, G. The bayesian separation principle for data-driven control. *arXiv preprint arXiv:2409.16717*, 2024.
- Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.
- Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.
- Hansen, N., Lin, Y., Su, H., Wang, X., Kumar, V., and Rajeswaran, A. Modem: Accelerating visual model-based reinforcement learning with demonstrations. *arXiv preprint arXiv:2212.05698*, 2022.
- Hansen, N., Su, H., and Wang, X. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.
- Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *JMLR*, 2010.
- Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. *NeurIPS*, 32, 2019.
- Kabzan, J., Valls, M. I., Reijgwart, V. J., Hendrikx, H. F., Ehmke, C., Prajapat, M., Bühler, A., Gosala, N., Gupta, M., Sivanesan, R., et al. Amz driverless: The full autonomous racing system. *Journal of Field Robotics*, 2020.
- Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *ICML*, pp. 267–274, 2002.
- Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *NeurIPS*, 33:15312–15325, 2020.
- Kakade, S. M. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

- Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.
- Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 2002.
- Khalil, H. K. *Nonlinear control*, volume 406. Pearson New York, 2015.
- Klyubin, A. S., Polani, D., and Nehaniv, C. L. Empowerment: A universal agent-centric measure of control. In *IEEE congress on evolutionary computation*, 2005.
- Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 2013.
- Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *ICML*, pp. 2796–2804. PMLR, 2018.
- Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.
- Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2020.
- Mania, H., Jordan, M. I., and Recht, B. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.
- Mendonca, R., Rybkin, O., Daniilidis, K., Hafner, D., and Pathak, D. Discovering and achieving goals via world models. *NeurIPS*, 2021.
- Moerland, T. M., Broekens, J., Plaat, A., Jonker, C. M., et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023.
- Moulin, A. and Neu, G. Optimistic planning by regularized dynamic programming. In *International Conference on Machine Learning*, pp. 25337–25357. PMLR, 2023.
- Nair, A. V., Pong, V., Dalal, M., Bahl, S., Lin, S., and Levine, S. Visual reinforcement learning with imagined goals. *NeurIPS*, 2018.
- Neu, G. and Pike-Burke, C. A unifying view of optimism in episodic reinforcement learning. *NeurIPS*, 2020.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.
- Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.
- Pathak, D., Gandhi, D., and Gupta, A. Self-supervised exploration via disagreement. In *ICML*, pp. 5062–5071. PMLR, 2019.
- Pinneri, C., Sawant, S., Blaes, S., Achterhold, J., Stueckler, J., Rolinek, M., and Martius, G. Sample-efficient cross-entropy method for real-time planning. In *Conference on Robot Learning*, pp. 1049–1065. PMLR, 2021.
- Pong, V. H., Dalal, M., Lin, S., Nair, A., Bahl, S., and Levine, S. Skew-fit: State-covering self-supervised reinforcement learning. *arXiv preprint arXiv:1903.03698*, 2019.
- Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.
- Rigter, M., Lacerda, B., and Hawes, N. Rambo-rl: Robust adversarial model-based offline reinforcement learning. *NeurIPS*, 35:16082–16097, 2022.

- Rothfuss, J., Sukhija, B., Birchler, T., Kassraie, P., and Krause, A. Hallucinated adversarial control for conservative offline policy evaluation. *UAI*, 2023.
- Rothfuss, J., Sukhija, B., Treven, L., Dörfler, F., Coros, S., and Krause, A. Bridging the sim-to-real gap with bayesian inference. *IROS*, 2024.
- Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 2018.
- Salge, C., Glackin, C., and Polani, D. Empowerment—an introduction. *Guided Self-Organization: Inception*, 2014.
- Schmidhuber, J. 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.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. Planning to explore via self-supervised world models. In *International conference on machine learning*, 2020.
- Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.
- Sharma, A., Xu, K., Sardana, N., Gupta, A., Hausman, K., Levine, S., and Finn, C. Autonomous reinforcement learning: Formalism and benchmarking. *arXiv preprint arXiv:2112.09605*, 2021.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 2017.
- Simchowitz, M. and Foster, D. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*. PMLR, 2020.
- Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 2012.
- Stadie, B. C., Levine, S., and Abbeel, P. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.
- Sukhija, B., Coros, S., Krause, A., Abbeel, P., and Sferrazza, C. Maxinfo: Boosting exploration in reinforcement learning through information gain maximization. *arXiv preprint arXiv:2412.12098*, 2024a.
- Sukhija, B., Treven, L., Dörfler, F., Coros, S., and Krause, A. Neorl: Efficient exploration for nonepisodic rl. *NeurIPS*, 2024b.
- Sukhija, B., Treven, L., Sancaktar, C., Blaes, S., Coros, S., and Krause, A. Optimistic active exploration of dynamical systems. *NeurIPS*, 2024c.
- Sussex, S., Makarova, A., and Krause, A. Model-based causal bayesian optimization. In *ICLR*, May 2023.
- Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., and Stone, P. Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.
- Treven, L., Curi, S., Mutn̄, M., and Krause, A. Learning stabilizing controllers for unstable linear quadratic regulators from a single trajectory. In *Learning for Dynamics and Control*, 2021.
- Treven, L., Hübotter, J., Sukhija, B., Dörfler, F., and Krause, A. Efficient exploration in continuous-time model-based reinforcement learning. *NeurIPS*, 2024.

- Wagenmaker, A., Shi, G., and Jamieson, K. Optimal exploration for model-based rl in nonlinear systems. *arXiv preprint arXiv:2306.09210*, 2023.
- Wu, P., Escontrela, A., Hafner, D., Abbeel, P., and Goldberg, K. Daydreamer: World models for physical robot learning. In *CORL*. PMLR, 2023.
- Yom-Tov, E., Feraru, G., Kozdoba, M., Mannor, S., Tennenholtz, M., and Hochberg, I. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 2017.
- Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *NeurIPS*, 2020.

A 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. SOMBRL 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. SOMBRL 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 demonstrate SOMBRL’s scalability to high-dimensional settings such as visual control tasks, and additionally, ground this approach theoretically, providing regret bounds for nonlinear systems and common RL settings.

B Proofs

B.1 Definition of Well-calibrated Model

We define a well-calibrated statistical model of $\mathbf{f}^*$, which captures both the mean prediction μ_n and the uncertainty σ_n of our learned model.

Definition B.1 (Well-calibrated statistical model of $\mathbf{f}^*$, [Rothfuss et al. \(2023\)](#)). Let $\mathcal{Z} \stackrel{\text{def}}{=} \mathcal{X} \times \mathcal{U}$. A sequence of sets $\{\mathcal{M}_n(\delta)\}_{n \geq 0}$, where

$$\mathcal{M}_n(\delta) \stackrel{\text{def}}{=} \left\{ \mathbf{f} : \mathcal{Z} \rightarrow \mathbb{R}^{d_{\mathbf{x}}} \mid \forall \mathbf{z} \in \mathcal{Z}, \forall j \in \{1, \dots, d_{\mathbf{x}}\} : |\mu_{n,j}(\mathbf{z}) - f_j(\mathbf{z})| \leq \beta_n(\delta) \sigma_{n,j}(\mathbf{z}) \right\},$$

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

In the following, we show that Gaussian processes are well-calibrated models if Theorem 5.2 holds.

Lemma B.2 (Well calibrated confidence intervals for RKHS, [Rothfuss et al. \(2023\)](#)). Let $\mathbf{f}^* \in \mathcal{H}_{k,B}^{d_{\mathbf{x}}}$. Suppose μ_n and σ_n are the posterior mean and variance of a GP with kernel k , Equation (9). There exists $\beta_n(\delta)$, for which the tuple $(\mu_n, \sigma_n, \beta_n(\delta))$ is a well-calibrated statistical model of $\mathbf{f}^*$.

In summary, in the RKHS setting, a GP is a well-calibrated model.

B.2 Analysis for the finite horizon case

Lemma B.3. Let Theorem 5.1 and Theorem 5.2 hold. Consider the following definitions

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

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

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

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

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

Therefore,

$$|J(\pi, \mu_n) - J(\pi, \mathbf{f}^*)| = \left| \mathbb{E} \left[\sum_{t=0}^{T-1} J_{t+1}(\pi, \mathbf{f}^*, \mathbf{x}'_{t+1}) - J_{t+1}(\pi, \mathbf{f}^*, \hat{\mathbf{x}}_{t+1}) \right] \right|$$

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

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

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

Therefore, we have

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

□

Note that Theorem 5.3 follows directly from Theorem B.3.

Lemma B.4. *Let Theorem 5.1 and Theorem 5.2 hold and consider the simple regret at episode n , $r_n = J(\boldsymbol{\pi}^*, \mathbf{f}^*) - J(\boldsymbol{\pi}_n, \mathbf{f}^*)$. The following holds for all $n > 0$ with probability at least $1 - \delta$*

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

Proof.

$$\begin{aligned} r_n &= J(\boldsymbol{\pi}^*, \mathbf{f}^*) - J(\boldsymbol{\pi}_n, \mathbf{f}^*) \\ &\leq J(\boldsymbol{\pi}^*, \boldsymbol{\mu}_n) + \lambda_n \Sigma_n(\boldsymbol{\pi}^*, \boldsymbol{\mu}_n) - J(\boldsymbol{\pi}_n, \mathbf{f}^*) \quad (\text{Theorem B.3}) \\ &\leq J(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) + \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) - J(\boldsymbol{\pi}_n, \mathbf{f}^*) \quad (\text{Equation (8)}) \\ &= J(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) - J(\boldsymbol{\pi}_n, \mathbf{f}^*) + \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) \\ &\leq \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \mathbf{f}^*) + \lambda_n \Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) \quad (\text{Theorem B.3}) \\ &= 2\lambda_n \Sigma_n(\boldsymbol{\pi}_n, \mathbf{f}^*) + \lambda_n (\Sigma_n(\boldsymbol{\pi}_n, \boldsymbol{\mu}_n) - \Sigma_n(\boldsymbol{\pi}_n, \mathbf{f}^*)) \\ &\leq (\lambda_n^2 + 2\lambda_n) \Sigma_n(\boldsymbol{\pi}_n, \mathbf{f}^*). \end{aligned}$$

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 Theorem B.3. □

Proof of Theorem 5.4.

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

$$\begin{aligned}
&= (\lambda_N^2 + 2\lambda_N) \sum_{n=1}^N \mathbb{E}_{\mathbf{f}^*} \left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n(\mathbf{x}_t, \boldsymbol{\pi}(\mathbf{x}_t))\| \right] \\
&\leq (\lambda_N^2 + 2\lambda_N) \sqrt{NT} \sum_{n=1}^N \mathbb{E}_{\mathbf{f}^*} \left[\sum_{t=0}^{T-1} \|\boldsymbol{\sigma}_n^2(\mathbf{x}_t, \boldsymbol{\pi}(\mathbf{x}_t))\| \right] \\
&\leq C(\lambda_N^2 + 2\lambda_N) T \sqrt{N\Gamma_{NT}} \quad (\text{Curi et al. (2020, Lemma 17)})
\end{aligned}$$

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

Table 2: Maximum information gain bounds for common choice of kernels.

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

In Table 2 we list rates of Γ_N for the most common choice of kernels.

B.3 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 $\mathbf{z} = (\mathbf{x}, \mathbf{u})$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{U}$.

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

$$\mathbf{V}_t = \mathbf{I} + \frac{1}{\sigma^2} \sum_{i=1}^t k_{\mathbf{z}_i} k_{\mathbf{z}_i}^\top. \quad (10)$$

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

$$\mathbf{M}_t = (k_{\mathbf{z}_1} \quad k_{\mathbf{z}_2} \quad \cdots \quad k_{\mathbf{z}_t}) \quad (11)$$

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

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

Corollary B.5 (Lower bound on the posterior log determinant).

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

In particular, we have

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

Proof.

$$\begin{aligned}
\log(|\mathbf{V}_n|) &= \log(|\mathbf{V}_{n-1}|) + \log\left(\left|\mathbf{I} + \sigma^{-2} \mathbf{V}_{n-1}^{-1/2} \sum_{k=1}^{\hat{T}_n} \mathbf{k}_{\mathbf{z}_{k,n}} \mathbf{k}_{\mathbf{z}_{k,n}}^\top \mathbf{V}_{n-1}^{-1/2}\right|\right) \\
&\geq \log(|\mathbf{V}_{n-1}|)
\end{aligned}$$

$$\begin{aligned}
& + \log \left(1 + \text{tr} \left(\sigma^{-2} \mathbf{V}_{n-1}^{-1/2} \sum_{k=1}^{\hat{T}_n} \mathbf{k}_{\mathbf{z}_{k,n}} \mathbf{k}_{\mathbf{z}_{k,n}}^\top \mathbf{V}_{n-1}^{-1/2} \right) \right) \quad (\text{see } (*) \text{ below}) \\
& = \log(|\mathbf{V}_{n-1}|) + \log \left(1 + \sigma^{-2} \sum_{k=1}^{\hat{T}_n} \|\mathbf{k}_{\mathbf{z}_{k,n}}\|_{\mathbf{V}_{n-1}^{-1}}^2 \right) \\
& = \log(|\mathbf{V}_{n-1}|) + \log \left(1 + \sigma^{-2} \sum_{k=1}^{\hat{T}_n} \|\sigma_{n-1}(\mathbf{z}_{k,n})\|^2 \right)
\end{aligned}$$

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

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

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

Corollary B.6 (Upper bound on the posterior log determinant).

$$\log(|\mathbf{V}_n|) \leq \log(|\mathbf{V}_{n-1}|) + \sum_{k=1}^{\hat{T}_n} \sum_{j=1}^{d_x} \log(1 + \sigma^{-2} \sigma_{n-1,j}^2(\mathbf{z}_{k,n}))$$

Proof.

$$\begin{aligned}
\log(|\mathbf{V}_n|) &= \log(|\mathbf{V}_{n-1}|) + \log(|\mathbf{I} + \mathbf{M}|) \\
&\leq \log(|\mathbf{V}_{n-1}|) + \log(|\text{diag}(\mathbf{I} + \mathbf{M})|) \quad (\text{Hadamard's inequality for PSD matrices}) \\
&= \log(|\mathbf{V}_{n-1}|) + \sum_{k=1}^{\hat{T}_n} \sum_{j=1}^{d_x} \log(1 + \sigma^{-2} \sigma_{n-1,j}^2(\mathbf{z}_{k,n}))
\end{aligned}$$

$\square$

Theorem B.5 will be useful for the discounted horizon case, whereas Theorem B.6 will be applied for the nonepisodic setting.

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

Lemma B.7. *Let Theorem 5.1, and Theorem 5.2 hold. Consider the following definitions*

$$\begin{aligned}
J_\gamma(\pi, \mathbf{f}^*) &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{x}_t, \pi(\mathbf{x}_t)) \right] \quad \text{s.t., } \mathbf{x}_{t+1} = \mathbf{f}^*(\mathbf{x}_t, \pi(\mathbf{x}_t)) + \mathbf{w}_t, \quad \mathbf{x}_0 = \mathbf{x}(0), \\
J_\gamma(\pi, \mu_n) &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{x}'_t, \pi(\mathbf{x}'_t)) \right] \quad \text{s.t., } \mathbf{x}'_{t+1} = \mu_n(\mathbf{x}'_t, \pi(\mathbf{x}'_t)) + \mathbf{w}_t, \quad \mathbf{x}'_0 = \mathbf{x}(0), \\
\Sigma_n^\gamma(\pi, \mathbf{f}^*) &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(\mathbf{x}_t, \pi(\mathbf{x}_t))\| \right] \quad \text{s.t., } \mathbf{x}_{t+1} = \mathbf{f}^*(\mathbf{x}_t, \pi(\mathbf{x}_t)) + \mathbf{w}_t, \quad \mathbf{x}_0 = \mathbf{x}(0), \\
\Sigma_n^\gamma(\pi, \mu_n) &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(\mathbf{x}'_t, \pi(\mathbf{x}'_t))\| \right] \quad \text{s.t., } \mathbf{x}'_{t+1} = \mu_n(\mathbf{x}'_t, \pi(\mathbf{x}'_t)) + \mathbf{w}_t, \quad \mathbf{x}'_0 = \mathbf{x}(0), \\
\lambda_n &= C_{\max} \frac{\gamma}{1 - \gamma} \frac{(1 + \sqrt{d_x}) \beta_{n-1}(\delta)}{\sigma},
\end{aligned}$$

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

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

Proof. We give the proof for $|J_\gamma(\pi, \mathbf{f}^*) - J_\gamma(\pi, \mu_n)| \leq \lambda_n (L_r, \mu_n) \Sigma_n^\gamma(\pi, \mu_n)$. 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

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

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

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

□

Proof of Theorem 5.5. We start with bounding

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

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).

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

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

Next, we bound the term

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

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,

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

Finally, we have

$$\begin{aligned} \sum_{n=1}^N s_n &\leq C_\gamma \sum_{n=1}^N \log \left(1 + \sigma^{-2} \sum_{t=0}^{T(n)-1} \|\boldsymbol{\sigma}_{n-1}(\mathbf{x}'_t, \boldsymbol{\pi}(\mathbf{x}'_t))\|^2 \right) \\ &\leq C_\gamma \Gamma_{\sum_{n=1}^N T(n)} \quad (\text{Theorem B.5}) \\ &\leq C_\gamma \Gamma_N \log(N) \end{aligned}$$

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

Since $\lambda_N \propto \beta_N / (1-\gamma)$, we get

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

□

B.4 Analysis for the nonepisodic RL case

In this section, we prove Theorem 5.6. First, we restate the bounded energy assumption from Sukhija et al. (2024b).

Definition B.8 ($\mathcal{K}_\infty$ -functions). The function $\xi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ is of class $\mathcal{K}_\infty$, if it is continuous, strictly increasing, $\xi(0) = 0$ and $\xi(s) \rightarrow \infty$ for $s \rightarrow \infty$.

Assumption B.9 (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 $\pi \in \Pi$ we have,

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

$$|V^\pi(x) - V^\pi(x')| \leq \kappa(\|x - x'\|) \quad (\text{uniform continuity})$$

$$C_l \xi(\|x\|) \leq V^\pi(x) \leq C_u \xi(\|x\|) \quad (\text{positive definiteness})$$

$$\mathbb{E}_{x_+|x, \pi}[V^\pi(x_+)] \leq \gamma V^\pi(x) + K \quad (\text{drift condition})$$

where $x_+ = f^*(x, \pi(x)) + w$.

Bounded norm of reward:

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

Boundedness of the noise with respect to κ :

$$\mathbb{E}_w[\kappa(\|w\|)] < \infty, \quad \mathbb{E}_w[\kappa^2(\|w\|)] < \infty$$

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 f^* .

Assumption B.10 (Continuous closed-loop dynamics, and Gaussian noise.). The dynamics model f^* and all $\pi \in \Pi$ are continuous, and process noise is i.i.d. Gaussian with variance σ^2 , i.e., $w_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I)$.

An important quantity in the average reward setting is the bias

$$B(\pi, x_0) = \lim_{T \rightarrow \infty} \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} r(x_t, u_t) - J_{\text{avg}}(\pi) \right]. \quad (16)$$

The Bellman equation for the average reward setting is given by

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

Sukhija et al. (2024b) show that under Theorem B.9 and Theorem B.10 the average reward solution and the bias (c.f., Equation (3)) are bounded. Moreover, they show that with Theorem 5.2 the average reward and bias are bounded for all dynamics $f \in \mathcal{M}_n \cap \mathcal{M}_0$.

Lemma B.11. *Let Theorem 5.2, Theorem B.9, and Theorem 5.2 hold. Consider the following definitions*

$$\begin{aligned} J_{\text{avg}}(\pi, f^*) &= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{t=0}^{T-1} r(x_t, \pi(x_t)) \right] \text{ s.t., } x_{t+1} = f^*(x_t, \pi(x_t)) + w_t, \quad x_0 = x(0), \\ J_{\text{avg}}(\pi, f) &= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{t=0}^{T-1} r(x'_t, \pi(x'_t)) \right] \text{ s.t., } x'_{t+1} = f(x'_t, \pi(x'_t)) + w_t, \quad x'_0 = x(0), \\ \Sigma_n(\pi, f^*) &= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{t=0}^{T-1} \|\sigma_n(x_t, \pi(x_t))\| \right] \text{ s.t., } x_{t+1} = f^*(x_t, \pi(x_t)) + w_t, \quad x_0 = x(0), \\ \Sigma_n(\pi, f) &= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{t=0}^{T-1} \|\sigma_n(x'_t, \pi(x'_t))\| \right] \text{ s.t., } x'_{t+1} = f(x'_t, \pi(x'_t)) + w_t, \quad x'_0 = x(0), \\ \lambda_n &= D_4(x_0, \gamma, K) \beta_{n-1}(\delta), \end{aligned}$$

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

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

Proof.

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

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

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 $|J_{\text{avg}}(\pi, \mathbf{f}) - J_{\text{avg}}(\pi, \mathbf{f}^*)| \leq \lambda_n \Sigma_n(\pi, \mathbf{f}^*)$. $\square$

SOMBRL 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:

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

$$\widehat{T}_n = \arg \max_{T \geq 1} T + 1 \quad (19)$$

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

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

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

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 SOMBRL, we can obtain $\mathbf{f}$ using the quadratic program described in Equation (22). However, in practice, we just pick the mean model $\mu_n \in \mathcal{M}_n$. 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 B.12 (Formal Theorem statement for informal Theorem 5.6). *Define $R_N = \sum_{n=1}^N \mathbb{E}[J_{\text{avg}}(\pi^*) - r(\mathbf{x}_n, \pi_n(\mathbf{x}_n))]$. Let Theorem 5.2, Theorem B.9, and Theorem B.10 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.

$$\begin{aligned} R_N &= \mathbb{E} \left[\sum_{n=1}^{E_N} \sum_{k=0}^{T_n-1} J_{\text{avg}}(\pi^*) - r(\mathbf{x}_k^n, \pi_n(\mathbf{x}_k^n)) \right] \\ &\leq \mathbb{E} \left[\sum_{n=1}^{E_N} \sum_{k=0}^{T_n-1} J_{\text{avg}}(\pi^*, \mathbf{f}_n) + \lambda_n \Sigma_n(\pi^*, \mathbf{f}_n) - r(\mathbf{z}_k^n) \right] \end{aligned} \quad (\text{Theorem B.11})$$

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

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

$$\begin{aligned}
&\mathbb{E} \left[\sum_{n=1}^{E_N} \sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \mathbf{f}) \right] \\
&= \mathbb{E} \left[\sum_{n=1}^{E_N} \sum_{k=0}^{T_n-1} \|\boldsymbol{\sigma}_n(\mathbf{x}_t, \boldsymbol{\pi}(\mathbf{x}_t))\| \right] \\
&\quad + \mathbb{E} \left[\sum_{n=1}^{E_N} \sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \mathbf{f}) - \|\boldsymbol{\sigma}_n(\mathbf{x}_t, \boldsymbol{\pi}(\mathbf{x}_t))\| \right] \\
&\leq C\sqrt{\Gamma_N N} + \mathbb{E} \left[\sum_{n=1}^{E_N} \sum_{k=0}^{T_n-1} \Sigma_n(\boldsymbol{\pi}_n, \mathbf{f}) - \|\boldsymbol{\sigma}_n(\mathbf{x}_t, \boldsymbol{\pi}(\mathbf{x}_t))\| \right] \\
&\hspace{15em} \text{(Lemma A.1 Sukhija et al. (2024b))} \\
&\leq \sqrt{N\Gamma_N} + \mathcal{O}(\Gamma_N \sqrt{N}) && \text{(Sukhija et al. (2024b), Theorem 3.1) with reward } \boldsymbol{\sigma}_n
\end{aligned}$$

Therefore

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

In conclusion,

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

□

B.5 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 B.13. *Let Theorem 5.1 and Theorem 5.2 hold. Consider SOMBRL with extrinsic reward $r = 0$, i.e.,*

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

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

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

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

$$\begin{aligned}
\Sigma_N^* &\leq \frac{1}{N} \sum_{n=1}^N \Sigma_n^* && \text{(monotonicity of the variance)} \\
&\leq \frac{1}{N} \sum_{n=1}^N (1 + \lambda_n) \Sigma_n(\pi_n^*, \mu_n) && \text{(Theorem B.3)} \\
&\leq \frac{1}{N} \sum_{n=1}^N (1 + \lambda_n) \Sigma_n(\pi_n, \mu_n) && (\pi_n \text{ is the maximizer for mean dynamics } \mu_n) \\
&\leq \frac{1}{N} \sum_{n=1}^N (1 + \lambda_n)^2 \Sigma_n(\pi_n, \mathbf{f}^*) && \text{(Theorem B.3)} \\
&\leq (1 + \lambda_N)^2 \frac{1}{N} \sum_{n=1}^N \Sigma_n(\pi_n, \mathbf{f}^*) \\
&\leq (1 + \lambda_N)^2 \frac{1}{\sqrt{N}} \sum_{n=1}^N \Sigma_n^2(\pi_n, \mathbf{f}^*) \\
&\leq \mathcal{O} \left(\sqrt{\frac{\Gamma_N^3}{N}} \right)
\end{aligned}$$

□

Effectively, Theorem B.13 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.

B.6 Analysis for the finite horizon setting with sub-Gaussian noise

In the following, we analyse the regret for the setting where the process noise $\mathbf{w}$ is σ -sub Gaussian. **Assumption B.14.** The dynamics model $\mathbf{f}^*$, reward r , and all $\pi \in \Pi$ are L_f , L_r and L_π 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 SOMBRL, as we do not maximize over the set of dynamics we can overcome this limitation.

Moreover, since $\mathbf{f}^*$ has bounded RKHS norm, i.e., $\|\mathbf{f}^*\|_k \leq B$ (Theorem 5.2). From Srinivas et al. (2012); Chowdhury & Gopalan (2017) follows that with probability $1 - \delta$ we have for every n :

$$\|\mathbf{f}^* - \mu_n\|_{k_n} \leq \beta_n.$$

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

$$\begin{aligned}
\mathbf{f}_n &= \arg \min_{\mathbf{f} \in \text{span}(k(\mathbf{x}_1, \cdot), \dots, k(\mathbf{x}_n, \cdot))} \|\mathbf{f} - \mu_n\|_{k_n} \\
&\text{s.t. } \|\mathbf{f}\|_k \leq B
\end{aligned} \tag{22}$$

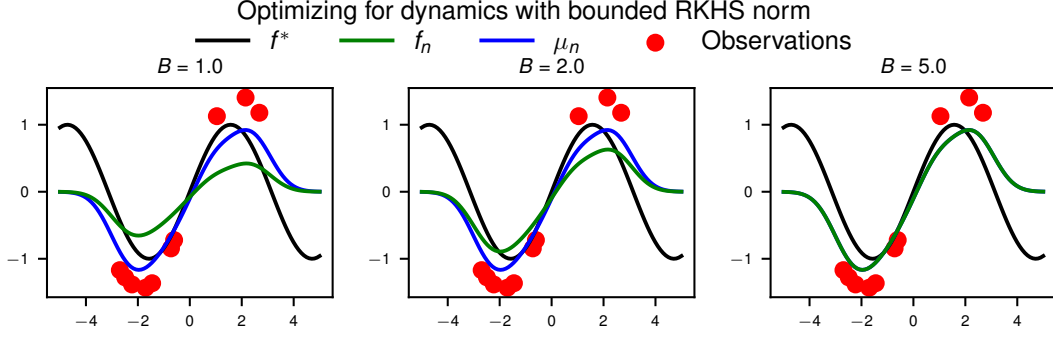


Figure 5: Solution to Equation (22) for different values for B . Effectively, for larger values for B , μ_n and f_n coincide.

Theorem B.15. *The optimization problem Equation (22) is feasible and we have $\|f_n - \mu_n\|_{k_n} \leq \beta_n$.*

Proof. Consider the noise-free case, i.e., $w = 0$, and let $\bar{\mu}_n$ posterior mean for this setting. For the function $\bar{\mu}_n$ holds that $\|f^* - \bar{\mu}_n\|_{k_n} \leq \beta_n$ (Corollary 3.11 of Kanagawa et al. (2018)) and $\|\bar{\mu}_n\|_k \leq B$ (Theorem 3.5 of Kanagawa et al. (2018)). Since $\|\mu_n - \mu_n\|_{k_n} \leq \|\bar{\mu}_n - f^*\|_{k_n} + \|f^* - \mu_n\|_{k_n} \leq 2\beta_n$. By representer theorem, it also holds that $\mu_n \in \text{span}(k(z_1, \cdot), \dots, k(z_n, \cdot))$. $\square$

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

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

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

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

This is a quadratic program that can be solved using any QP solver. The program finds the closest function to the posterior mean μ_n that is Lipschitz continuous (see Figure 5). In particular, note that since $\|f_n\|_k \leq B$, 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, SOMBRL follows the same strategy as Equation (8) but instead of using the mean dynamics μ_n , we plan with the dynamics f_n that are obtained from solving Equation (22).

$$\begin{aligned} \pi_n = \arg \max_{\pi \in \Pi} & \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} r(x'_t, u_t) + \lambda_n \|\sigma_n(x'_t, u_t)\| \right] \\ x'_{t+1} &= f_n(x'_t, u_t) + w_t, \end{aligned} \quad (23)$$

Lemma B.16. *Let Theorem B.14 and Theorem 5.2 hold. Consider the following definitions*

$$\begin{aligned} J(\pi, f^*) &= E \left[\sum_{t=0}^{T-1} r(x_t, \pi(x_t)) \right] \text{ s.t., } x_{t+1} = f^*(x_t, \pi(x_t)) + w_t, \quad x_0 = x(0), \\ J(\pi, f_n) &= E \left[\sum_{t=0}^{T-1} r(x'_t, \pi(x'_t)) \right] \text{ s.t., } x'_{t+1} = f_n(x'_t, \pi(x'_t)) + w_t, \quad x'_0 = x(0), \\ \Sigma_n(\pi, f^*) &= E \left[\sum_{t=0}^{T-1} \|\sigma_n(x_t, \pi(x_t))\| \right] \text{ s.t., } x_{t+1} = f^*(x_t, \pi(x_t)) + w_t, \quad x_0 = x(0), \\ \Sigma_n(\pi, f_n) &= E \left[\sum_{t=0}^{T-1} \|\sigma_n(x'_t, \pi(x'_t))\| \right] \text{ s.t., } x'_{t+1} = f_n(x'_t, \pi(x'_t)) + w_t, \quad x'_0 = x(0), \end{aligned}$$

$$\lambda_n = (1 + d_x)L_r(1 + L_\pi)\bar{L}_*^{T-1}T\beta_n.$$

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

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

Proof.

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

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

We show that

$$\|\mathbf{x}_{t+1} - \mathbf{x}'_{t+1}\| \leq (1 + \sqrt{d_x})\beta_n \left(\sum_{k=0}^t \bar{L}_*^{t-k} \|\sigma_n(\mathbf{x}'_k, \pi(\mathbf{x}'_k))\| \right).$$

Consider $t = 1$

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

Consider any $t > 1$,

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

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

In summary, we have

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

□

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 $\mathbf{f}_n$.

Lemma B.17. *Let Theorem B.14 and Theorem 5.2 hold and consider the simple regret at episode n , $r_n = J(\pi^*, \mathbf{f}^*) - J(\pi_n, \mathbf{f}^*)$. The following holds for all $n > 0$ with probability at least $1 - \delta$*

$$r_n \leq (2\lambda_n + \lambda_n^2) \Sigma_n(\pi_n, \mathbf{f}^*)$$

Proof.

$$\begin{aligned}
r_n &= J(\pi^*, \mathbf{f}^*) - J(\pi_n, \mathbf{f}^*) \\
&\leq J(\pi^*, \mathbf{f}_n) + \lambda_n \Sigma_n(\pi^*, \mathbf{f}_n) - J(\pi_n, \mathbf{f}^*) && \text{(Theorem B.16)} \\
&\leq J(\pi_n, \mathbf{f}_n) + \lambda_n \Sigma_n(\pi_n, \mathbf{f}_n) - J(\pi_n, \mathbf{f}^*) && \text{(Equation (23))} \\
&= J(\pi_n, \mathbf{f}_n) - J(\pi_n, \mathbf{f}^*) + \lambda_n \Sigma_n(\pi_n, \mathbf{f}_n) \\
&\leq \lambda_n \Sigma_n(\pi_n, \mathbf{f}^*) + \lambda_n \Sigma_n(\pi_n, \mathbf{f}_n) && \text{(Theorem B.16)} \\
&= 2\lambda_n \Sigma_n(\pi_n, \mathbf{f}^*) + \lambda_n (\Sigma_n(\pi_n, \mathbf{f}_n) - \Sigma_n(\pi_n, \mathbf{f}^*)) \\
&\leq (\lambda_n^2 + 2\lambda_n) \Sigma_n(\pi_n, \mathbf{f}^*).
\end{aligned}$$

□

Theorem B.18 (Finite horizon setting sub-Gaussian case). *Let Theorem B.14 and Theorem 5.2 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 5.4, since in Theorem B.17 we show that also for the sub-Gaussian case, SOMBRL has the same regret dependence w.r.t. λ_n and $\Sigma_n(\pi_n, \mathbf{f}^*)$. □

C Additional Experiments

In this section, we provide additional experiments.

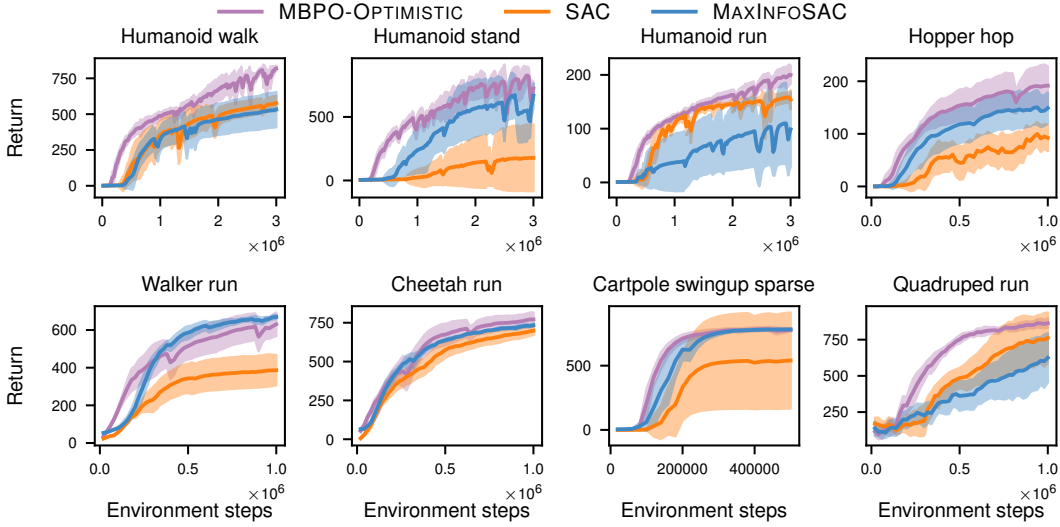


Figure 6: Comparison between MBPO-OPTIMISTIC and MAXINFOSAC and SAC. We observe that MBPO-OPTIMISTIC, being an MBRL algorithm, performs the best in terms of sample efficiency.

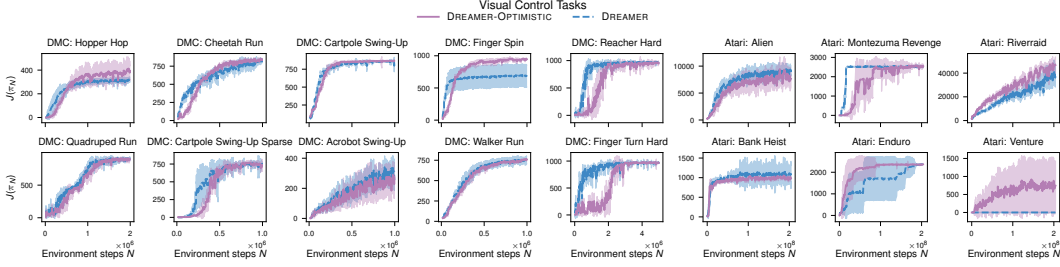


Figure 7: Learning curves for the visual control tasks from DMC and Atari using DREAMER as the base algorithm. DREAMER-OPTIMISTIC either performs on-par or better than DREAMER in all our experiments. Particularly, in the Venture task from the Atari benchmark, where DREAMER fails to obtain any rewards.

Experiments with MBPO In Figure 6 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 7 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 SOMBRL methods. We think this is a promising direction for future work.

In Figure 8 and Figure 9 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 9), in line with our results in Section 6, DREAMER fails to obtain any meaningful rewards.

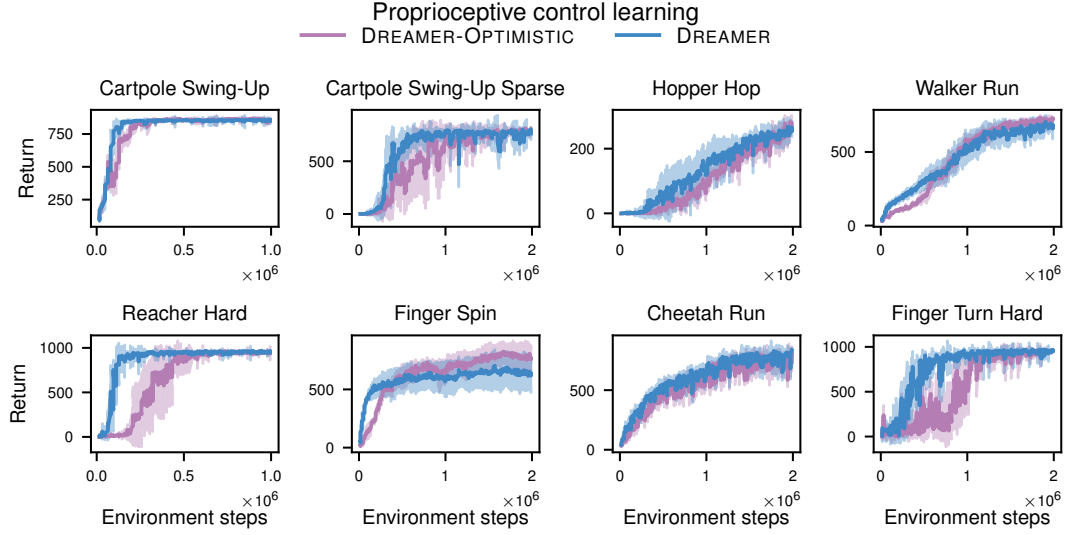


Figure 8: Experiments with DREAMER-OPTIMISTIC and DREAMER for proprioceptive tasks. DREAMER-OPTIMISTIC performs on par with DREAMER, obtaining slightly better performance on the Finger Spin task.

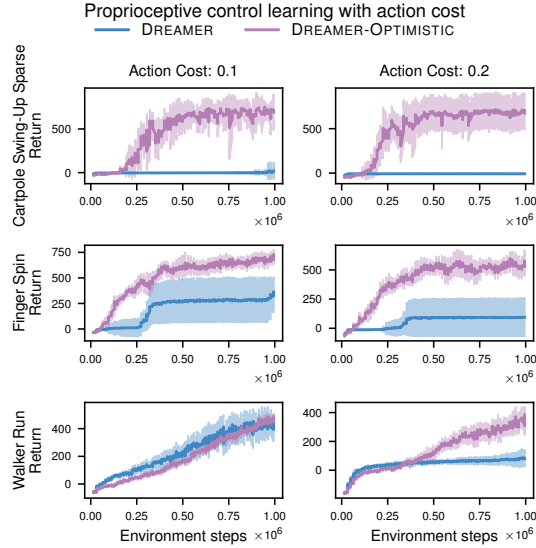


Figure 9: Experiments with DREAMER-OPTIMISTIC and DREAMER for proprioceptive tasks with action costs. DREAMER completely fails to solve the task, whereas DREAMER-OPTIMISTIC does not.

D Experiment Details

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

D.1 MBPO-OPTIMISTIC

For MBPO-OPTIMISTIC, we train an ensemble of forward dynamics models⁶. 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

$$L(\lambda) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{1:n}, \mathbf{u} \sim \pi_n, \bar{\mathbf{u}} \sim \bar{\pi}_n} \log(\lambda) (\sigma_n(\mathbf{x}, \mathbf{u}) - \sigma_n(\mathbf{x}, \bar{\mathbf{u}})). \quad (24)$$

Here $\bar{\pi}_n$ is a target policy, which is updated using polyak updates of π_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 $(\mathbf{x}, \mathbf{u}, \mathbf{x}') \sim \mathcal{D}_{1:n}$, and add $(\mathbf{x}, \mathbf{u}, \hat{\mathbf{x}}')$, transitions predicted by our mean model μ_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 5⁷.

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

D.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 dynamics⁸. 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 DREAMER⁹.

D.3 SIMFSVGD-OPTIMISTIC

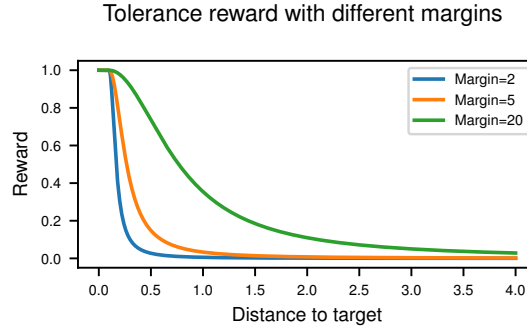


Figure 10: Tolerance reward function for different values of the margin. For larger margins, the agent receives rewards even if its further away from the target.

We use the same experiment setup, simulation prior, and hyperparameters as Rothfuss et al. (2024)¹⁰. The reward function in Rothfuss et al. (2024) is based on the tolerance reward from Tassa et al. (2018).

⁶For 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).

⁷We did not tune the UTD and chose 5 to trade-off between computational cost and sample efficiency.

⁸For all tasks we use a (512, 512) neural network architecture with 5 ensembles.

⁹We use the 12 million size model and the official DREAMERV3 implementation (<https://github.com/danijar/dreamerv3/tree/main>).

¹⁰official implementation: https://github.com/lasgroup/simulation_transfer

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 “closeness” is quantified using a margin parameter for the reward function. In Figure 10 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.

D.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)¹¹.

D.5 Computational Costs

Table 3: Computation cost comparison for SOMBRL with different base algorithms.

Algorithm	Training time
HUCRL (GPs)	90 +/- 3 min (Pendulum), 31.5 +/- 2.5 min (MountainCar)
SOMBRL (GPs)	30 +/- 0.6 min (Pendulum), 13.8 +/- 0.25 min (MountainCar)
MBPO-MEAN (Time per 100k steps, 1 ensemble, GPU: NVIDIA GeForce RTX 2080 Ti)	9.6 +/- 0.2 min
MBPO-OPTIMISTIC (Time per 100k steps, 5 ensembles, GPU: NVIDIA GeForce RTX 2080 Ti)	13.7 +/- 0.35 min
DREAMER (Time per 100k steps, GPU: NVIDIA GeForce RTX 4090)	42.24 +/- 0.95 min
DREAMER-OPTIMISTIC (Time per 100k steps, 5 ensembles, GPU: NVIDIA GeForce RTX 4090)	46.32 +/- 0.34 min

¹¹official implementation: <https://github.com/lasgroup/opax>