NeoRL: Efficient Exploration for Nonepisodic RL

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of nonepisodic reinforcement learning (RL) for nonlinear dynamical systems, where the system dynamics are unknown and the RL agent has to learn from a single trajectory, i.e., without resets. We propose ***N**onepisodic **O**ptimistic **RL** (NeoRL)*, an approach based on the principle of optimism in the face of uncertainty. NeoRL uses well-calibrated probabilistic models and plans optimistically w.r.t. the epistemic uncertainty about the unknown dynamics. Under continuity and bounded energy assumptions on the system, we provide a first-of-its-kind regret bound of $\mathcal{O}(\beta_T \sqrt{T\Gamma_T})$ for general nonlinear systems with Gaussian process dynamics. We compare NeoRL to other baselines on several deep RL environments and empirically demonstrate that NeoRL achieves the optimal average cost while incurring the least regret.

1 Introduction

In recent years, data-driven control approaches, such as reinforcement learning (RL), have demonstrated remarkable achievements. However, most RL algorithms are devised for an episodic setting, where during each episode, the agent interacts in the environment for a predetermined episode length or until a termination condition is met. After the episode, the agent is reset back to an initial state from where the next episode commences. Episodes prevent the system from blowing up, i.e., maintain stability, while also restricting exploration to states that are relevant to the task at hand. Moreover, resets ensure that the agent explores close to the initial states and does not end up at undesirable parts of the state space that exhibit low reward. In simulation, resetting is typically straightforward. However, if we wish to enable agents to learn by interacting with the real world, resets are often prohibitive since they typically involve manual intervention. Instead, agents should be able to learn autonomously (Sharma et al., 2021) i.e., from a single trajectory. While several works in the Deep RL community have addressed this challenge, (c.f., Section 5), the theoretical results for this setting are fairly limited. In particular, the setting has been extensively studied for finite state and action spaces (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002; Jaksch et al., 2010) and linear systems (Abbasi-Yadkori & Szepesvári, 2011; Simchowitz & Foster, 2020; Dean et al., 2020; Lale et al., 2020). However, the extension to nonlinear systems is much less understood. In our work, we address this gap and propose a practical RL algorithm that is grounded in theory. In particular, we make the following contributions.

Contributions

1. We propose, NeoRL, a novel model-based RL algorithm based on the principle of optimism in the face of uncertainty. NeoRL operates in a nonepisodic setting and picks average cost optimal policies optimistically w.r.t. to the model's epistemic uncertainty.
2. We show that when the dynamics lie in a reproducing kernel Hilbert space (RKHS) of kernel k, NeoRL exhibits a regret of $\mathcal{O}(\beta_T \sqrt{T\Gamma_T})$, where the regret, akin to prior work, is measured w.r.t to the optimal average cost under known dynamics, T is the number of environment steps

Submitted to the ICML 2024 Workshop on Foundations of Reinforcement Learning and Control. Do not distribute.

and Γ_T the maximum information gain of kernel k (Srinivas et al., 2012). Our regret bound is similar to the ones obtained in the episodic setting (Kakade et al., 2020; Curi et al., 2020; Sukhija et al., 2024; Treven et al., 2024) and Gaussian process (GP) bandit optimization (Srinivas et al., 2012; Chowdhury & Gopalan, 2017; Scarlett et al., 2017). To the best of our knowledge, we are the first to obtain regret bounds for the setting.

3. We evaluate NEORL on several RL benchmarks against common model-based RL baselines. Our experimental results demonstrate that NEORL consistently achieves sublinear regret, also when neural networks are employed instead of GPs for modeling dynamics. Moreover, in all our experiments, NEORL converges to the optimal average cost.

2 Problem Setting

We consider a discrete-time dynamical system with running costs c.

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}^*(\boldsymbol{x}_t, \boldsymbol{u}_t) + \boldsymbol{w}_t,\ (\boldsymbol{x}_t, \boldsymbol{u}_t) \in \mathcal{X} \times \mathcal{U},\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{1}$$

$$c(\boldsymbol{x}, \boldsymbol{u}) \in \mathbb{R}_{\geq 0} \tag{Running cost}$$

Here $\boldsymbol{x}_t \in \mathcal{X} \subseteq \mathbb{R}^{d_{\boldsymbol{x}}}$ is the state, $\boldsymbol{u}_t \in \mathcal{U} \subseteq \mathbb{R}^{d_{\boldsymbol{u}}}$ the control input, and $\boldsymbol{w}_t \in \mathcal{W} \subseteq \mathbb{R}^{\boldsymbol{w}}$ the process noise. The dynamics $\boldsymbol{f}^*$ are unknown and the cost c is assumed to be known.

Task In this work, we study the average cost RL problem (Puterman, 2014), i.e., we want to learn the solution to the following minimization problem

$$A(\boldsymbol{\pi}^*, \boldsymbol{x}_0) = \min_{\boldsymbol{\pi} \in \Pi} A(\boldsymbol{\pi}, \boldsymbol{x}_0) = \min_{\boldsymbol{\pi} \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t) \right]. \tag{2}$$

Moreover, we consider the nonepisodic RL setting where the system starts at an initial state $\boldsymbol{x}_0 \in \mathcal{X}$ but never resets back during learning, that is, we seek to learn from a single trajectory. After each step t in the environment, the RL system receives a transition tuple $(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{x}_{t+1})$ and updates its policy based on the data $\mathcal{D}_t$ collected thus far during learning. The average cost formulation is common for the nonepisodic setting (Jaksch et al., 2010; Abbasi-Yadkori & Szepesvári, 2011; Simchowitz & Foster, 2020), and the cumulative regret for the learning algorithm in this case is defined as

$$R_T = \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{u}_t | \boldsymbol{x}_0} [c(\boldsymbol{x}_t, \boldsymbol{u}_t) - A(\boldsymbol{\pi}^*, \boldsymbol{x}_0)]. \tag{3}$$

Studying the average cost criterion for general continuous state-action spaces is challenging even when the dynamics are known, since the average cost exists only for special classes of nonlinear systems (Arapostathis et al., 1993). In the following, we impose assumptions on the dynamics and policy class Π that enable our theoretical analysis.

2.1 Assumptions

Imposing continuity on $\boldsymbol{f}^*$ is quite common in the control theory (Khalil, 2015) and reinforcement learning literature (Curi et al., 2020; Sussex et al., 2023; Sukhija et al., 2024). To this end, for our analysis, we make the following assumption.

Assumption 2.1 (Continuity of $\boldsymbol{f}^*$ and $\boldsymbol{\pi}$)**.** The dynamics model $\boldsymbol{f}^*$ and all $\boldsymbol{\pi} \in \Pi$ are continuous.

Next, we make an assumption on the system's stochasticity.

Assumption 2.2 (Process noise distribution)**.** The process noise is i.i.d. Gaussian with variance σ^2, i.e., $\boldsymbol{w}_t \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

For simplicity, we focus on the homoscedastic setting. However, the analysis can be extended for the more general heteroscedastic case. In the following, we make assumptions on our policy class. To this end, we first introduce the class of $\mathcal{K}_\infty$ functions.

Definition 2.3 ($\mathcal{K}_\infty$-functions)**.** The function $\xi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is of class $\mathcal{K}_\infty$, if it is continuous, strictly increasing, $\xi(0) = 0$ and $\xi(s) \to \infty$ for $s \to \infty$.

Assumption 2.4 (Policies with bounded energy)**.** We assume there exists $\kappa, \xi \in \mathcal{K}_\infty$, positive constants K, C_u, C_l with $C_u > C_l$, and $\gamma \in (0, 1)$ such that for each $\boldsymbol{\pi} \in \Pi$ we have, *Bounded energy:* There exists a Lyapunov function $V^{\boldsymbol{\pi}} : \mathcal{X} \to [0, \infty)$ for which

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

Bounded norm of cost:

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

Boundedness of the noise with respect to κ:

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

The bounded energy assumption is introduced to ensure that the system does not end up in states from which it cannot recover. In particular, the Lyapunov function $V^{\boldsymbol{\pi}}$ can be viewed as an energy function for the dynamical system, and the drift condition above ensures that in expectation the energy at the next state $\boldsymbol{x}'$ is not increasing to ∞, that is, the system is not "blowing up". Other works that study learning nonlinear dynamics (Foster et al., 2020; Sattar & Oymak, 2022; Lale et al., 2021) in the nonepisodic setting also make stability assumptions such as global exponential stability for their analysis. In similar spirit, we make the bounded energy assumption for our policy class. The drift condition on the Lyapunov function is also used to study the ergodicity of Markov chains for continuous state spaces (Meyn & Tweedie, 2012; Hairer & Mattingly, 2011), which is crucial for our analysis of the infinite horizon behavior of the system. Moreover, for a very rich class of problems, the drift condition is satisfied. We highlight this in the corollary below.

Corollary 2.5. *Assume $\boldsymbol{f}^*$ is uniformly continuous and for all $\boldsymbol{\pi} \in \Pi$, $\boldsymbol{x} \in \mathcal{X}$, $\|\boldsymbol{\pi}(\boldsymbol{x})\| \leq u_{\max}$. Further assume, there exists $\boldsymbol{\pi}_s \in \Pi$ such that we have constants K, C_u, C_l with $C_u > C_l$, $\gamma \in (0, 1)$, $\kappa, \alpha \in \mathcal{K}_\infty$ and a Lyapunov function $V : \mathcal{X} \to [0, \infty)$ for which*

$$|V(\boldsymbol{x}) - V(\boldsymbol{x}')| \leq \kappa(\|\boldsymbol{x} - \boldsymbol{x}'\|)$$
$$C_l \xi(\|\boldsymbol{x}\|) \leq V(\boldsymbol{x}) \leq C_u \xi(\|\boldsymbol{x}\|)$$
$$\mathbb{E}_{\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{\pi}_s}[V(\boldsymbol{x}')] \leq \gamma V(\boldsymbol{x}) + K.$$

Then, V also satisfies the drift condition for all $\boldsymbol{\pi} \in \Pi$, i.e., is a Lyapunov function for all policies.

We prove this corollary in Appendix A. Intuitively, if the inputs are bounded, the energy inserted into the system by another policy is also bounded. Nearly all real-world systems have bounded inputs due to the physical limitations of actuators. For these systems, it suffices if only one policy in Π satisfies the drift condition. In Appendix A.4, we discuss an alternative set of assumptions on the costs, that relaxes the bounded energy requirement on the policy class Π.

The boundedness assumptions for the cost and the noise in Assumption 2.4 are satisfied for a rich class of cost and $\mathcal{K}_\infty$ functions.

Under these assumptions, we can show the existence of the average cost solution.

Theorem 2.6 (Existence of Average Cost Solution)**.** *Let Assumption 2.1 – 2.4 hold. Consider any $\boldsymbol{\pi} \in \Pi$ and let $P^{\boldsymbol{\pi}}$ denote its transition kernel, i.e., $P^{\boldsymbol{\pi}}(\boldsymbol{x}, \mathcal{A}) = \mathbb{P}(\boldsymbol{x}' \in \mathcal{A} | \boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))$. Then $P^{\boldsymbol{\pi}}$ admits a unique invariant measure $\bar{P}^{\boldsymbol{\pi}}$ and there exists $C_2, C_3 \in (0, \infty)$, $\lambda \in (0, 1)$ such that*

Average Cost*;*

$$A(\boldsymbol{\pi}) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t)\right] = \mathbb{E}_{\boldsymbol{x} \sim \bar{P}^{\boldsymbol{\pi}}}\left[c(\boldsymbol{x}, \boldsymbol{\pi}(x))\right]$$

Bias Cost*; Letting $B(\boldsymbol{\pi}, \boldsymbol{x}_0) = \lim_{T \to \infty} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t) - A(\boldsymbol{\pi})\right]$ denote the bias, we have*

$$|B(\boldsymbol{\pi}, \boldsymbol{x}_0)| = \left|\lim_{T \to \infty} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t) - A(\boldsymbol{\pi})\right]\right| \leq C_2(1 + V^{\boldsymbol{\pi}}(\boldsymbol{x}_0)) \frac{1}{1 - \lambda}$$

for all $\boldsymbol{x}_0 \in \mathcal{X}$.

Theorem 2.6 is a crucial result for our analysis since it implies that the average cost is bounded and *independent of the initial state* $\boldsymbol{x}_0$. Furthermore, it also shows that the bias is bounded. Similar to the discounted case, the average cost criterion satisfies the following Bellman equation (Puterman, 2014)

$$B(\boldsymbol{\pi}, \boldsymbol{x}) + A(\boldsymbol{\pi}) = c(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x})) + \mathbb{E}_{\boldsymbol{x}'}[B(\boldsymbol{\pi}, \boldsymbol{x}')|\boldsymbol{x}, \boldsymbol{\pi}] \tag{4}$$

Accordingly, the bias term plays an important role in the regret analysis (also notice its similarity to our regret term in Equation (3)).

Thus far, we have only made assumptions that make the average cost problem tractable. In the following, we make an assumption on the dynamics that allow us to learn it from data. We start with the definition of a well-calibrated statistical model of $\boldsymbol{f}^*$.

Definition 2.7 (Well-calibrated statistical model of $\boldsymbol{f}^*$, Rothfuss et al. (2023))**.** Let $\mathcal{Z} \stackrel{\text{def}}{=} \mathcal{X} \times \mathcal{U}$. An all-time well-calibrated statistical model of the function $\boldsymbol{f}^*$ is a sequence $\{\mathcal{M}_n(\delta)\}_{n \geq 0}$, where

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

if, with probability at least $1 - \delta$, we have $\boldsymbol{f}^* \in \bigcap_{n \geq 0} \mathcal{M}_n(\delta)$. Here, $\mu_{n,j}$ and $\sigma_{n,j}$ denote the j-th element in the vector-valued mean and standard deviation functions $\boldsymbol{\mu}_n$ and $\boldsymbol{\sigma}_n$ respectively, and $\beta_n(\delta) \in \mathbb{R}_{\geq 0}$ is a scalar function that depends on the confidence level $\delta \in (0, 1]$ and which is monotonically increasing in n.

Next, we assume that $\boldsymbol{f}^*$ resides in a Reproducing Kernel Hilbert Space (RKHS) of vector-valued functions and show that this is sufficient for us to obtain a well-calibrated model.

Assumption 2.8. We assume that the functions f^*_j, $j \in 1, \dots, d_x$ lie in a RKHS with kernel k and have a bounded norm B, that is $\boldsymbol{f}^* \in \mathcal{H}^{d_x}_{k,B}$, with $\mathcal{H}^{d_x}_{k,B} = \{\boldsymbol{f} \mid \|f_j\|_k \leq B, j = 1, \dots, d_x\}$. Moreover, we assume that $k(\boldsymbol{x}, \boldsymbol{x}) \leq \sigma_{\max}$ for all $\boldsymbol{x} \in \mathcal{X}$.

The mean and epistemic uncertainty of the vector-valued function $\boldsymbol{f}^*$ are denoted with $\boldsymbol{\mu}_n(\boldsymbol{z}) = [\mu_{n,j}(\boldsymbol{z})]_{j \leq d_x}$, and $\boldsymbol{\sigma}_n(\boldsymbol{z}) = [\sigma_{n,j}(\boldsymbol{z})]_{j \leq d_x}$ and have an analytical solution

$$\begin{aligned} \mu_{n,j}(\boldsymbol{z}) &= \bar{\mu}^j(\boldsymbol{z}) + \boldsymbol{k}_n^\top(\boldsymbol{z})(\boldsymbol{K}_n + \sigma^2 \boldsymbol{I})^{-1}(\boldsymbol{y}^j_{1:n} - \bar{\mu}^j_{1:n}), \\ \sigma^2_{n,j}(\boldsymbol{z}) &= k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}_n^\top(\boldsymbol{z})(\boldsymbol{K}_n + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_n(\boldsymbol{x}), \end{aligned} \tag{5}$$

Here, $\boldsymbol{y}^j_{1:n}$ corresponds to the noisy measurements of f^*_j, i.e., the observed next state from the transitions dataset $\mathcal{D}_{1:n}$, $\bar{\mu}^j(\boldsymbol{z})$ corresponds to the fixed mean function, e.g., $\bar{\mu}^j(\boldsymbol{z}) = \boldsymbol{x}$, $\bar{\mu}^j_{1:n}$ its values on the dataset, $\boldsymbol{k}_n = [k(\boldsymbol{z}, \boldsymbol{z}_i)]_{i \leq nT}, \boldsymbol{z}_i \in \mathcal{D}_{1:n}$, and $\boldsymbol{K}_n = [k(\boldsymbol{z}_i, \boldsymbol{z}_l)]_{i,l \leq nT}, \boldsymbol{z}_i, \boldsymbol{z}_l \in \mathcal{D}_{1:n}$ is the data kernel matrix. The restriction on the kernel $k(\boldsymbol{x}, \boldsymbol{x}) \leq \sigma_{\max}$ has also appeared in works studying the episodic setting for nonlinear systems (Mania et al., 2020; Kakade et al., 2020; Curi et al., 2020; Sukhija et al., 2024; Wagenmaker et al., 2023).

Lemma 2.9 (Well calibrated confidence intervals for RKHS, Rothfuss et al. (2023))**.** *Let* $\boldsymbol{f}^* \in \mathcal{H}^{d_x}_{k,B}$. *Suppose* $\boldsymbol{\mu}_n$ *and* $\boldsymbol{\sigma}_n$ *are the posterior mean and variance of a GP with kernel* k*, c.f., Equation* (5)*. There exists* $\beta_n(\delta)$*, for which the tuple* $(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \beta_n(\delta))$ *is a well-calibrated statistical model of* $\boldsymbol{f}^*$.

In summary, in the RKHS setting, a GP is a well-calibrated model. For more general models like BNNs, methods such as Kuleshov et al. (2018) can be used for calibration. Our results can also be extended beyond the RKHS setting to other classes of well-calibrated models similar to Curi et al. (2020).

3 NeoRL

In the following, we present our algorithm: **N**on**e**pisodic **O**ptimistic **RL** (NeoRL) for efficient nonepisodic exploration in continuous state-action spaces. NeoRL builds on recent advances in episodic RL (Kakade et al., 2020; Curi et al., 2020; Sukhija et al., 2024; Treven et al., 2024) and leverages the optimism in the face of uncertainty paradigm to pick policies that are optimistic w.r.t. the dynamics within our calibrated statistical model. Moreover, NeoRL suggests policies according to the following decision rule

$$(\boldsymbol{\pi}_n, \boldsymbol{f}_n) \stackrel{\text{def}}{=} \underset{\boldsymbol{\pi} \in \Pi, \ \boldsymbol{f} \in \mathcal{M}_{n-1} \cap \mathcal{M}_0}{\arg\min} A(\boldsymbol{\pi}, \boldsymbol{f}). \tag{6}$$

Algorithm 1 **NEORL:** NONEPISODIC OPTIMISTIC RL

Init: Aleatoric uncertainty σ, Probability δ, Statistical model $(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \beta_0(\delta))$, H_0

for $n = 1, \dots, N$ **do**

$\boldsymbol{\pi}_n = \arg\min_{\boldsymbol{\pi} \in \Pi} \min_{\boldsymbol{f} \in \mathcal{M}_{n-1} \cap \mathcal{M}_0} A(\boldsymbol{\pi}, \boldsymbol{f})$ ➤ Prepare policy

$H_n = 2H_{n-1}$ ➤ Set horizon

$\mathcal{D}_n \leftarrow \text{ROLLOUT}(\boldsymbol{\pi}_n)$ ➤ Collect measurements for horizon H_n

Update $(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \beta_n) \leftarrow \mathcal{D}_n$ ➤ Update model

end for

Here, $\boldsymbol{f}_n$ is a dynamical system such that the cost by controlling $\boldsymbol{f}_n$ with its optimal policy $\boldsymbol{\pi}_n$ is the lowest among all the plausible systems from $\mathcal{M}_{n-1} \cap \mathcal{M}_0$. Note, from Lemma 2.9 we have that $\boldsymbol{f}^* \in \mathcal{M}_{n-1} \cap \mathcal{M}_0$ (with high probability) and therefore the solution to Equation (6) gives an optimistic estimate for the average cost.

NEORL proceeds in the following manner. Similar to Jaksch et al. (2010), we bin the total time T the agent spends interacting in the environment into N "artificial" episodes. At each episode, we pick a policy according to Equation (6) and roll it out for H_n steps on the system. Next, we use the data collected during the rollout to update our statistical model. Finally, we double the horizon $H_{n+1} = 2H_n$, akin to Simchowitz & Foster (2020), and continue to the next episode *without resetting* the system back to the initial state $\boldsymbol{x}_0$. The algorithm is summarized in Algorithm 1.

3.1 Theoretical Results

In the following, we study the theoretical properties for NEORL and provide a first-of-its-kind bound on the cumulative regret for the average cost criterion for general nonlinear dynamical systems. Our bound depends on the *maximum information gain* of kernel k (Srinivas et al., 2012), defined as

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

Γ_T represents the complexity of learning $\boldsymbol{f}^*$ and is sublinear for a very rich class of kernels (Vakili et al., 2021). In Appendix A, we report the dependence of Γ_T on T in Table 1.

Theorem 3.1 (Cumulative Regret of NEORL)**.** *Let Assumption 2.1 – 2.8 hold, and define H_0 as the smallest integer such that*

$$H_0 > \frac{\log \left({}^{C_u}/_{C_l} \right)}{\log \left({}^{1}/_{\gamma} \right)}.$$

Then with probability at least $1 - \delta$, we have the following regret for NEORL

$$R_T \leq D_4(\boldsymbol{x}_0, K, \gamma) \beta_T \sqrt{T \Gamma_T} + D_5(\boldsymbol{x}_0, K, \gamma) \log_2 \left(\frac{T}{H_0} + 1 \right). \tag{7}$$

with $D_4(\boldsymbol{x}_0, K, \gamma)$, $D_5(\boldsymbol{x}_0, K, \gamma) \in (0, \infty)$ when $\|\boldsymbol{x}_0\| < \infty$, $K < \infty$, and $\gamma < 1$.

Theorem 3.1 gives sublinear regret for a rich class of RKHS functions. Moreover, it also gives a minimal horizon H_0 that we need to maintain before switching to the next policy. Even for the linear case, fast switching between stable controllers can destabilize the closed-loop system. We ensure this does not happen in our case by having a minimal horizon of H_0. Lastly, the regret bound depends on constants D_4 and D_5. The constants are finite when $\gamma < 1$, $K < \infty$ (bounded energy from Assumption 2.4 is satisfied), and $\|\boldsymbol{x}_0\| < \infty$. Theorem 3.1 can also be derived beyond the RKHS setting for a more general class of well-calibrated models. In this case, the maximum information gain is replaced by the model complexity from Curi et al. (2020) (c.f., Curi et al. (2020); Sukhija et al. (2024); Treven et al. (2024) for further detail).

3.2 Practical Modifications

For testing NEORL, we make three modifications that simplify its deployment in practice in terms of implementation and computation time. First, instead of doubling the horizon H_n we pick a fixed horizon H during the experiment. This makes the planning and training of the agent easier. Next,

Algorithm 2 Practical NeoRL:

Init: Aleatoric uncertainty σ, Probability δ, Statistical model $(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \beta_0(\delta))$
for $n = 1, \ldots, N$ **do**
  for $h = 1, \ldots, H$ **do**

$$\min_{\boldsymbol{u}_{0:H_{\text{MPC}}-1}, \boldsymbol{\eta}_{0:H_{\text{MPC}}-1}} \mathbb{E}\left[\sum_{h=0}^{H_{\text{MPC}}-1} c(\hat{\boldsymbol{x}}_h, \boldsymbol{u}_h)\right]; \boldsymbol{x}_0 = \boldsymbol{x}_h^n$$ ➤ Solve MPC problem

$$(\boldsymbol{x}_n^h, \boldsymbol{u}_0^*, \boldsymbol{x}_n^{h+1}) \leftarrow \text{ROLLOUT}(\boldsymbol{u}_0^*)$$ ➤ Collect transition

  end for
  Update $(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \beta_n) \leftarrow \mathcal{D}_n$
end for

we use a receding horizon controller, i.e., model predictive control (MPC) (García et al., 1989), instead of directly optimizing for the average cost in Equation (6). MPC is widely used to obtain a feedback controller for the infinite horizon setting. Moreover, while for linear systems, the Riccati equations (Anderson & Moore, 2007) provide an analytical solution to Equation (2), no such solution exists for the nonlinear case and MPC is commonly used as an approximation. Further, under additional assumptions on the cost and dynamics, MPC also obtains a policy with bounded average cost, which is crucial for the nonepisodic case (c.f., Assumption 2.4). We use the iCEM optimizer for planning (Pinneri et al., 2021). Finally, instead of optimizing over $\mathcal{M}_n \cap \mathcal{M}_0$, we optimize directly over $\mathcal{M}_n$. This allows us to use the reparameterization trick from Curi et al. (2020) and obtain a simple and tractable optimization problem. In summary, for each step t in the environment, we solve the following optimization problem

$$\min_{\boldsymbol{u}_{0:H_{\text{MPC}}-1}, \boldsymbol{\eta}_{0:H_{\text{MPC}}-1}} \mathbb{E}\left[\sum_{h=0}^{H_{\text{MPC}}-1} c(\hat{\boldsymbol{x}}_h, \boldsymbol{u}_h)\right], \tag{8}$$
$$\text{s.t. } \hat{\boldsymbol{x}}_{h+1} = \boldsymbol{\mu}_{n-1}(\hat{\boldsymbol{x}}_h, \boldsymbol{u}_h) + \beta_{n-1}(\delta)\boldsymbol{\sigma}_{n-1}(\hat{\boldsymbol{x}}_h, \boldsymbol{u}_h)\boldsymbol{\eta}_h + \boldsymbol{w}_h \text{ and } \hat{\boldsymbol{x}}_0 = \boldsymbol{x}_t.$$

Here H_{MPC} is the MPC horizon. We take the first input from the solution of the problem above, i.e., $\boldsymbol{u}_0^*$, and execute this in the system. We then repeat this procedure for H steps and then update our statistical model $\mathcal{M}_n$. The resulting optimization above considers a larger action space as it includes the hallucinated controls $\boldsymbol{\eta}$ (Curi et al., 2020) as additional input variables. Moreover, the final algorithm can be seen as a natural extension to H-UCRL (Curi et al., 2020) for the nonepisodic setting. We summarize the algorithm in Algorithm 2. Note while these modifications deviate from our theoretical analysis, empirically they work well for GP and BNN models, c.f., Section 4.

4 Experiments

We evaluate NeoRL on the Pendulum-v1 and MountainCar environment from the OpenAI gym benchmark suite (Brockman et al., 2016), Cartpole, Reacher, and Swimmer from the DeepMind control suite (Tassa et al., 2018), the racecar simulator from Kabzan et al. (2020), and a soft robotic arm from Tekinalp et al. (2024). The swimmer and the soft robotic arm are fairly high-dimensional systems – the swimmer has a 28-dimensional state and 5-dimensional action space, and the soft arm is represented by a 58-dimensional state and has a 12-dimensional action space. All environments are never reset during learning. Moreover, the Pendulum-v1, MountainCar, CartPole, and Reacher environments operate within a bounded domain and thus inherently satisfy Assumption 2.4. The swimmer, racecar, and soft arm can operate in an unbounded domain but have a cost function that penalizes the distance between the system's state $\boldsymbol{x}_t$ and a target state $\boldsymbol{x}^*$. Therefore, the cost encourages the system to move towards the target and remain within a bounded domain, as elaborated on further in Appendix A.4.

Baselines In this work, we focus on model-based RL (MBRL) algorithms due to their sample efficiency. To this end, we consider common techniques for planning with unknown dynamics, such as planning with the mean, trajectory sampling (Chua et al., 2018), and Thompson sampling (Osband & Van Roy, 2017). We adapt these three for our setting similar to as discussed in Section 3.2. For all experiments with probabilistic ensembles, we consider TS1 from Chua et al. (2018) for trajectory sampling, and for the GP experiment, we use distribution sampling from Chua et al. (2018). We

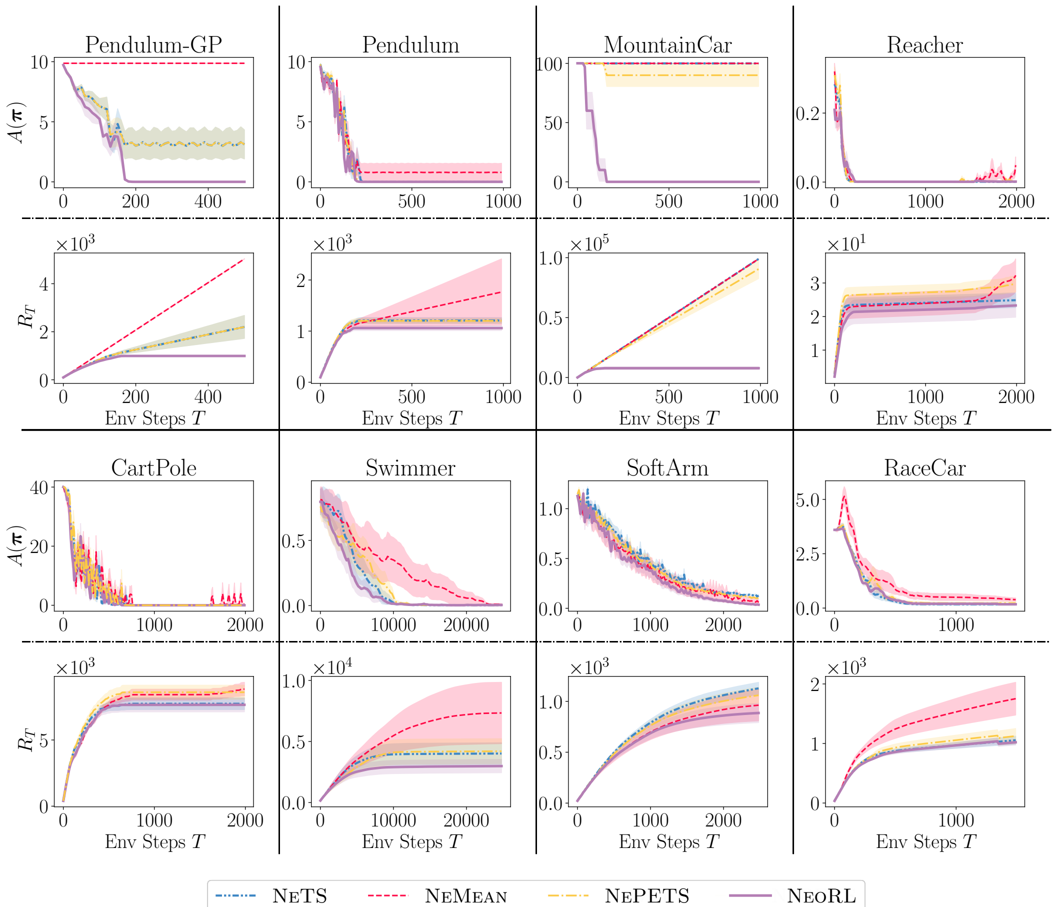

Figure 1: Average reward $A(\boldsymbol{\pi})$ and cumulative regret R_T over ten different seeds for all environments. We report the mean performance with one standard error as shaded regions. During all experiments, the environment is never reset. For all baselines, we model the dynamics with probabilistic ensembles, except in the Pendulum-GP experiment, where GPs are used instead. NEORL significantly outperforms all baselines and converges to the optimal average reward, $A(\boldsymbol{\pi}^*) = 0$, showing sublinear cumulative regret R_T for all environments.

call the three baselines NEMEAN (nonepisodic mean), NEPETS (nonepisodic PETS), and NETS (nonepisodic Thompson sampling). NEMEAN and NEPETS are greedy w.r.t. the current estimate of the dynamics, i.e., do not explicitly encourage exploration. In our experiments, we show that being greedy does not suffice to converge to the optimal average cost, that is, obtain sublinear regret.

Convergence to the optimal average cost In Figure 1 we report the normalized average cost and cumulative regret of NEORL, NEMEAN, NEPETS, and NETS. The normalized average cost is defined such that $A(\boldsymbol{\pi}^*) = 0$ for all environments. We observe that NEMEAN fails to converge to the optimal average cost for the Pendulum-v1 environment for both probabilistic ensembles and a GP model. It also fails to solve the MountainCar environment and is unstable for the Reacher and CartPole. In general, NEMEAN performs the worst among all methods. This is similar to the episodic case, where using the mean model often leads to the policy "overfitting" to the model inaccuracies (Chua et al., 2018). NEPETS performs better than the mean, however still significantly worse than NEORL. Even in the episodic setting, PETS tends to underexplore (Curi et al., 2020). We observe the same for the nonepisodic case, especially for the MountainCar task, which is a challenging RL environment with a sparse cost. Here NEPETS is also not able to achieve the optimal average cost and thus does not have sublinear cumulative regret. NETS performs similarly to NEPETS and is also not able to solve the MountainCar task.

NEORL performs the best among the baselines and converges to the optimal average cost achieving sublinear cumulative regret using only $\sim 10^3$ environment interactions. Moreover, this observation is consistent between different dynamics models (GPs and probabilistic ensembles) and environments.

Even in environments that are unbounded, i.e., Swimmer, SoftArm, and RaceCar, we observe that NEORL converges to the optimal average cost the fastest. We believe this is due to the MPC, which encourages the system to move closer to the target.

5 Related Work

Average cost RL for finite state-action spaces A significant amount of work studies the average cost/reward RL setting for finite-state action spaces. Moreover, seminal algorithms such as E^3 (Kearns & Singh, 2002) and R- $\max$ (Brafman & Tennenholtz, 2002) have established PAC bounds for the nonepisodic setting. These bounds are further improved for communicating MDPs by the UCRL2 (Jaksch et al., 2010) algorithm, which, similar to NEORL, is based on the optimism in the face of uncertainty paradigm and picks policies that are optimistic w.r.t. to the estimated dynamics. Their result is extended for weakly-communicating MDPs by REGAL (Bartlett & Tewari, 2012), similar results are derived for Thompson sampling based exploration (Ouyang et al., 2017), and for factored-MDP (Xu & Tewari, 2020). Albeit the significant amount of work for the finite case, progress for continuous state-action spaces has mostly been limited to linear dynamical systems.

Nonepisodic RL for linear systems There is a large body of work for nonepisodic learning with linear systems (Abbasi-Yadkori & Szepesvári, 2011; Cohen et al., 2019; Simchowitz & Foster, 2020; Dean et al., 2020; Lale et al., 2020; Faradonbeh et al., 2020; Abeille & Lazaric, 2020; Treven et al., 2021). For linear systems with quadratic costs, the average reward problem, also known as the linear quadratic-Gaussian (LQG), has a closed-form solution which is obtained via the Riccati equations (Anderson & Moore, 2007). Moreover, for LQG, stability and optimality are intertwined, making studying linear systems much easier than their nonlinear counterpart. For studying nonlinear systems, additional assumptions on their stability are usually made.

Nonepisodic RL beyond linear systems In the case of nonlinear systems, guarantees have mostly been established for the episodic setting (Mania et al., 2020; Kakade et al., 2020; Curi et al., 2020; Wagenmaker et al., 2023; Sukhija et al., 2024; Treven et al., 2024). Only a few works consider the nonepisodic/single-trajectory case. For instance, Foster et al. (2020); Sattar & Oymak (2022) study the problem of system identification of a closed-loop globally exponentially stable dynamical system from a single trajectory. Lale et al. (2021) study the nonepisodic setting for nonlinear systems with MPC. Moreover, they consider finite-order or exponentially fading NARX systems that lie in the RKHS of infinitely smooth functions, which they further approximate with random Fourier features (Rahimi & Recht, 2007) ϕ with feature size D. Further, they assume access to bounded persistently exciting inputs w.r.t. the feature matrix $\mathbf{\Phi}_t \mathbf{\Phi}_t^\intercal$. This assumption is generally tough to verify and common excitation strategies such as random exploration often don't perform well for nonlinear systems (Sukhija et al., 2024). Further, the algorithm acts greedily w.r.t. the estimated dynamics, akin to NEMEAN, and requires the feature size D to increase with the horizon T. They give a regret bound of $\mathcal{O}\left(T^{2/3}\right)$ where the regret is measured w.r.t. to the oracle MPC with access to the true dynamics. Lale et al. (2021) also assume exponential input-to-output stability of the system to avoid blow-up during exploration. Our work considers more general RKHS, does not require apriori knowledge of persistently exciting inputs, and gives a regret bound of $\mathcal{O}(\beta_T \sqrt{T \Gamma_T})$ w.r.t. the optimal average cost criterion. Moreover, our regret bound is similar to the ones obtained for nonlinear systems in the episodic case and Gaussian process bandits (Srinivas et al., 2012; Chowdhury & Gopalan, 2017; Scarlett et al., 2017). To the best of our knowledge, we are the first to give such a regret bound for nonlinear systems.

6 Conclusion

We propose, NEORL, a novel model-based RL algorithm for the nonepisodic setting with nonlinear dynamics and continuous state and action spaces. NEORL seeks for average-cost optimal policies and leverages the model's epistemic uncertainty to perform optimistic exploration. Similar to the episodic case (Kakade et al., 2020; Curi et al., 2020), we provide a regret bound for NEORL of $\mathcal{O}(\beta_T \sqrt{T \Gamma_T})$ for Gaussian process dynamics. To our knowledge, we are the first to obtain this result in the nonepisodic setting. We compare NEORL to other model-based RL methods on standard deep RL benchmarks. Our experiments demonstrate that NEORL, converges to the optimal average cost of $A(\boldsymbol{\pi}^*) = 0$ across all environments, suffering sublinear regret even when Bayesian neural networks are used to model the dynamics. Moreover, NEORL outperforms all our baselines across all environments requiring only $\sim 10^3$ samples for learning.

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.

Anderson, B. D. and Moore, J. B. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

Arapostathis, A., Borkar, V. S., Fernández-Gaucherand, E., Ghosh, M. K., and Marcus, S. I. Discrete-time controlled markov processes with average cost criterion: A survey. *SIAM Journal on Control and Optimization*, 1993.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*, 2012.

Brafman, R. I. and Tennenholtz, M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 2002.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *ICML*, 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.

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., Sarkar, T., and Rakhlin, A. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, 2020.

García, C. E., Prett, D. M., and Morari, M. Model predictive control: Theory and practice - a survey. *Automatica*, pp. 335–348, 1989.

Hairer, M. and Mattingly, J. C. Yet another look at harris' ergodic theorem for markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI: Centro Stefano Franscini, Ascona, May 2008*, pp. 109–117. Springer, 2011.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

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., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *NeurIPS*, 33:15312–15325, 2020.

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.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *ICML*, pp. 2796–2804. PMLR, 2018.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 2020.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. Model learning predictive control in nonlinear dynamical systems. In *Conference on Decision and Control (CDC)*. IEEE, 2021.

Mania, H., Jordan, M. I., and Recht, B. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, 2017.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown markov decision processes: A thompson sampling approach. *Advances in neural information processing systems*, 30, 2017.

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 *CORL*, Proceedings of Machine Learning Research, pp. 1049–1065, 2021.

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

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Rothfuss, J., Sukhija, B., Birchler, T., Kassraie, P., and Krause, A. Hallucinated adversarial control for conservative offline policy evaluation. *UAI*, 2023.

Sattar, Y. and Oymak, S. Non-asymptotic and accurate learning of nonlinear dynamical systems. *Journal of Machine Learning Research*, 2022.

Scarlett, J., Bogunovic, I., and Cevher, V. Lower bounds on regret for noisy Gaussian process bandit optimization. In *Conference on Learning Theory*, 2017.

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.

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.

Sukhija, B., Treven, L., Sancaktar, C., Blaes, S., Coros, S., and Krause, A. Optimistic active exploration of dynamical systems. *NeurIPS*, 2024.

Sussex, S., Makarova, A., and Krause, A. Model-based causal bayesian optimization. In *ICLR*, May 2023.

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.

Tekinalp, A., Kim, S. H., Bhosale, Y., Parthasarathy, T., Naughton, N., Albazroun, A., Joon, R., Cui, S., Nasiriziba, I., Stölzle, M., Shih, C.-H. C., and Gazzola, M. Gazzolalab/pyelastica: v0.3.2, 2024.

Treven, L., Curi, S., Mutnỳ, 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.

Vakili, S., Khezeli, K., and Picheny, V. On information gain and regret bounds in gaussian process bandits. In *AISTATS*, 2021.

Wagenmaker, A., Shi, G., and Jamieson, K. Optimal exploration for model-based rl in nonlinear systems. *arXiv preprint arXiv:2306.09210*, 2023.

Xu, Z. and Tewari, A. Reinforcement learning in factored mdps: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting. *Advances in Neural Information Processing Systems*, 2020.

Appendices

A Proofs

In this section, we prove Theorem 2.6 and Theorem 3.1. First, we start with the proof of Corollary 2.5.

Proof of Corollary 2.5. We first analyze the following term $\mathbb{E}_{\boldsymbol{w}}[V(\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))+\boldsymbol{w})-V(\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}_s(\boldsymbol{x}))+\boldsymbol{w})]$ for any $\boldsymbol{\pi}\in\Pi$.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}}[V(\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))+\boldsymbol{w})-V(\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}_s(\boldsymbol{x}))+\boldsymbol{w})] \\
&\quad\leq \mathbb{E}_{\boldsymbol{w}}[\kappa(\|\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))+\boldsymbol{w}-(\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}_s(\boldsymbol{x}))+\boldsymbol{w})\|)] && \text{(Uniform continuity of } V\text{)} \\
&\quad= \kappa(\|\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))-\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}_s(\boldsymbol{x}))\|) \\
&\quad\leq \kappa(\kappa_{\boldsymbol{f}^*}(\|\boldsymbol{\pi}(\boldsymbol{x})-\boldsymbol{\pi}_s(\boldsymbol{x})\|)) && \text{(Uniform continuity of } \boldsymbol{f}^*\text{)} \\
&\quad\leq \kappa(\kappa_{\boldsymbol{f}^*}(2u_{\max})). && \text{(Bounded inputs)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}'|\boldsymbol{\pi},\boldsymbol{x}}[V(\boldsymbol{x}')] &= \mathbb{E}_{\boldsymbol{w}}[V(\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))+\boldsymbol{w})] \\
&\leq \mathbb{E}_{\boldsymbol{w}}[V(\boldsymbol{f}^*(\boldsymbol{x},\boldsymbol{\pi}_s(\boldsymbol{x}))+\boldsymbol{w})]+\kappa(\kappa_{\boldsymbol{f}^*}(2u_{\max})) \\
&= \mathbb{E}_{\boldsymbol{x}'|\boldsymbol{\pi}_s,\boldsymbol{x}}[V(\boldsymbol{x}')]+\kappa(\kappa_{\boldsymbol{f}^*}(2u_{\max})) \\
&\leq \gamma V(\boldsymbol{x})+K+\kappa(\kappa_{\boldsymbol{f}^*}(2u_{\max})) \\
&= \gamma V(\boldsymbol{x})+\tilde{K} && (\tilde{K}=K+\kappa(\kappa_{\boldsymbol{f}^*}(2u_{\max})))
\end{aligned}$$

Hence, V satisfies the drift condition for $\boldsymbol{\pi}$. Furthermore, since V also satisfies positive definiteness by assumption, the bounded energy condition holds for all $\boldsymbol{\pi}\in\Pi$. □

A.1 Proof of Theorem 2.6

For proving Theorem 2.6, we invoke the results from (Hairer & Mattingly, 2011, Theorem 1.2 – 1.3). For this we require that the Markov chain induced by a policy $\boldsymbol{\pi}$ satisfies the drift condition. In our setting, this corresponds to Assumption 2.4. Next, we show that the chain satisfies the following minorisation condition.

Lemma A.1 (Minorisation condition)**.** *Consider the system in Equation (1) and let Assumption 2.1 – 2.4 hold. Let $P^{\boldsymbol{\pi}}$ denote the transition kernel for the policy $\boldsymbol{\pi}\in\Pi$, i.e., $P^{\boldsymbol{\pi}}(\boldsymbol{x},\mathcal{A})=\mathbb{P}(\boldsymbol{x}'\in\mathcal{A}|\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))$. Then, for all $\boldsymbol{\pi}\in\Pi$, exists a constant $\alpha\in(0,1)$ and a probability measure $\zeta(\cdot)$ s.t.,*

$$\inf_{\boldsymbol{x}\in\mathcal{C}} P^{\boldsymbol{\pi}}(\boldsymbol{x},\cdot)\geq\alpha\zeta(\cdot) \tag{9}$$

with $\mathcal{C}\stackrel{\text{def}}{=}\{\boldsymbol{x}\in\mathcal{X};V^{\boldsymbol{\pi}}(\boldsymbol{x})\leq R\}$ for some $R>{}^{2K}/_{1-\gamma}$

Proof. We prove it in 3 steps. First, we show that $\mathcal{C}$ is contained in a compact domain. From the Assumption 2.4 we pick the function $\xi\in\mathcal{K}_\infty$. Since $C_l\xi(0)=0,\lim_{s\to\infty}\xi(s)=+\infty$ and $C_l\xi$ is continuous, there exists M such that $C_l\xi(M)=R$. Then for $\|\boldsymbol{x}\|>M$ we have:

$$V^{\boldsymbol{\pi}}(\boldsymbol{x})\geq C_l\xi(\|\boldsymbol{x}\|)>\xi(M)=R.$$

Therefore we have: $\mathcal{C}\subseteq\mathcal{B}(\mathbf{0},M)\stackrel{\text{def}}{=}\{\boldsymbol{x}\mid\|\boldsymbol{x}-\mathbf{0}\|\leq M\}$. In the second step we show that $\boldsymbol{f}(\mathcal{C},\boldsymbol{\pi}(\mathcal{C}))$ is bounded, in particular we show that there exists $B>0$ such that: $\boldsymbol{f}(\mathcal{C},\boldsymbol{\pi}(\mathcal{C}))\subseteq\mathcal{B}(\mathbf{0},B)$. This is true since continuous image of compact set is compact and the observation:

$$\mathcal{C}\subseteq\mathcal{B}(\mathbf{0},M)\implies\boldsymbol{f}(\mathcal{C},\boldsymbol{\pi}(\mathcal{C}))\subseteq\boldsymbol{f}(\mathcal{B}(\mathbf{0},M),\boldsymbol{\pi}(\mathcal{B}(\mathbf{0},M))).$$

Since $\boldsymbol{f}(\mathcal{B}(\mathbf{0},M),\boldsymbol{\pi}(\mathcal{B}(\mathbf{0},M)))$ is compact there exists B such that $\boldsymbol{f}(\mathcal{C},\boldsymbol{\pi}(\mathcal{C}))\subseteq\mathcal{B}(\mathbf{0},B)$. In the last step we prove that $\alpha\stackrel{\text{def}}{=}2^{-d_{\boldsymbol{x}}}e^{-B^2/\sigma^2}$ and ζ with law of $\mathcal{N}\left(0,\frac{\sigma^2}{2}\right)$ satisfy condition of Lemma A.1. It is enough to show that $\forall\boldsymbol{\mu}\in\mathcal{B}(\mathbf{0},B),\forall\boldsymbol{x}\in\mathbb{R}^{d_{\boldsymbol{x}}}$ we have:

$$\alpha\frac{1}{(2\pi)^{\frac{d_x}{2}}\left(\frac{\sigma^2}{2}\right)^{\frac{d_{\boldsymbol{x}}}{2}}}e^{-\frac{\|\boldsymbol{x}\|^2}{\sigma^2}}\leq\frac{1}{(2\pi)^{\frac{d_x}{2}}(\sigma^2)^{\frac{d_{\boldsymbol{x}}}{2}}}e^{-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2}{2\sigma^2}}$$

which can be proven with simple algebraic manipulations. □

Through the minorisation condition and Assumption 2.4, we can prove the ergodicity of the closed-loop system for a given policy $\boldsymbol{\pi} \in \Pi$.

Theorem A.2 (Ergodicity of closed-loop system)**.** *Let Assumption 2.1 – 2.4, consider any probability measures ζ_1, ζ_2, and $\theta > 0$, define $P^{\boldsymbol{\pi}}\zeta$, $\|\varphi\|_{1+\theta V^{\boldsymbol{\pi}}}$, $\rho_\theta^{\boldsymbol{\pi}}$ as*

$$\begin{aligned}
(P^{\boldsymbol{\pi}}\zeta)(\mathcal{A}) &= \int_{\mathcal{X}} P^{\boldsymbol{\pi}}(\boldsymbol{x}, \mathcal{A})\zeta(d\boldsymbol{x}) \\
\|\varphi\|_{1+\theta V^{\boldsymbol{\pi}}} &= \sup_{\boldsymbol{x}\in\mathcal{X}} \frac{|\varphi(\boldsymbol{x})|}{1+\theta V^{\boldsymbol{\pi}}(\boldsymbol{x})} \\
\rho_\theta^{\boldsymbol{\pi}}(\zeta_1, \zeta_2) &= \sup_{\varphi:\|\varphi\|_{1+\theta V^{\boldsymbol{\pi}}}\le 1} \int_{\mathcal{X}} \varphi(\boldsymbol{x})(\zeta_1 - \zeta_2)(d\boldsymbol{x}) = \int_{\mathcal{X}} (1+\theta V^{\boldsymbol{\pi}}(\boldsymbol{x}))|\zeta_1 - \zeta_2|(d\boldsymbol{x}).
\end{aligned}$$

We have for all $\boldsymbol{\pi} \in \Pi$, that $P^{\boldsymbol{\pi}}$ admits a unique invariant measure $\bar{P}^{\boldsymbol{\pi}}$. Furthermore, there exist constants $C_1 > 0$, $\theta > 0$, $\lambda \in (0, 1)$ such that

$$\rho_\theta^{\boldsymbol{\pi}}(P^{\boldsymbol{\pi}}\zeta_1, P^{\boldsymbol{\pi}}\zeta_2) \le \lambda \rho_\theta^{\boldsymbol{\pi}}(\zeta_1, \zeta_2) \tag{1}$$

$$\left\|\mathbb{E}_{\boldsymbol{x}\sim(P^{\boldsymbol{\pi}})^t}[\varphi(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\bar{P}^{\boldsymbol{\pi}}}[\varphi(\boldsymbol{x})]\right\|_{1+V^{\boldsymbol{\pi}}} \le C_1\lambda^t \left\|\varphi - \mathbb{E}_{\boldsymbol{x}\sim\bar{P}^{\boldsymbol{\pi}}}[\varphi(\boldsymbol{x})]\right\|_{1+V^{\boldsymbol{\pi}}}. \tag{2}$$

holds for every measurable function $\varphi : \mathcal{X} \to \mathcal{R}$ with $\|\varphi\|_{1+V^{\boldsymbol{\pi}}} < \infty$. Here $(P^{\boldsymbol{\pi}})^t$ denotes the t-step transition kernel under the policy $\boldsymbol{\pi}$.

Moreover, $\theta = {}^{\alpha_0}/_K$, and

$$\lambda = \max\left\{1 - (\alpha - \alpha_0), \frac{2 + {}^R/_K\alpha_0\gamma_0}{2 + {}^R/_K\alpha_0}\right\} \tag{10}$$

for any $\alpha_0 \in (0, \alpha)$ and $\gamma_0 \in (\gamma + 2{}^K/_R, 1)$.

Proof. From Assumption 2.4, we have a value function for each policy that satisfies the drift condition. Furthermore, in Lemma A.1 we show that our system also satisfies the minorisation condition for all policies. Under these conditions, we can use the results from Hairer & Mattingly (2011, Theorem 1.2. – 1.3.). □

Note that $\|\cdot\|_{1+\theta V^{\boldsymbol{\pi}}}$ represents a family of equivalent norms for any $\theta > 0$. Now we prove Theorem 2.6.

Proof of Theorem 2.6. From Theorem A.2, we have

$$\rho_\theta^{\boldsymbol{\pi}}((P^{\boldsymbol{\pi}})^{t+1}, (P^{\boldsymbol{\pi}})^t) = \rho_\theta^{\boldsymbol{\pi}}(P^{\boldsymbol{\pi}}(P^{\boldsymbol{\pi}})^t, P^{\boldsymbol{\pi}}(P^{\boldsymbol{\pi}})^{t-1}) \le \lambda^t \rho_\theta^{\boldsymbol{\pi}}(P^{\boldsymbol{\pi}}\delta_{\boldsymbol{x}_0}, \delta_{\boldsymbol{x}_0}),$$

where $\delta_{\boldsymbol{x}_0}$ is the dirac measure. Therefore, $(P^{\boldsymbol{\pi}})^t$ is a Cauchy sequence. Furthermore, $\rho_\theta^{\boldsymbol{\pi}}$ is complete for the set of probability measures integrating V, thus $\rho_\theta^{\boldsymbol{\pi}}((P^{\boldsymbol{\pi}})^t, \bar{P}^{\boldsymbol{\pi}}) \to 0$ for $t \to \infty$ (c.f., Hairer & Mattingly (2011) for more details). In particular, we have for φ such that $\|\varphi\|_{1+\theta V^{\boldsymbol{\pi}}} \le 1$,

$$\lim_{t\to\infty} \int_{\mathcal{X}} \varphi(\boldsymbol{x})(P^{\boldsymbol{\pi}})^t(d\boldsymbol{x}) = \int_{\mathcal{X}} \varphi(\boldsymbol{x})\bar{P}^{\boldsymbol{\pi}}(d\boldsymbol{x}).$$

Note that since all $\|\cdot\|_{1+\theta V^{\boldsymbol{\pi}}}$ norms are equivalent for $\theta > 0$, if $\|c\|_{1+V^{\boldsymbol{\pi}}} \le C$ (Assumption 2.4), then $\|c\|_{1+\theta V^{\boldsymbol{\pi}}} \le C'$ for some $C' \in (0, \infty)$. Furthermore, note that $c(\cdot) \ge 0$. Therefore,

$$\begin{aligned}
\int_{\mathcal{X}} c(\boldsymbol{x})\bar{P}^{\boldsymbol{\pi}}(d\boldsymbol{x}) &= \lim_{t\to\infty} \int_{\mathcal{X}} c(\boldsymbol{x})(P^{\boldsymbol{\pi}})^t(d\boldsymbol{x}) \\
&\le C \lim_{t\to\infty} \int_{\mathcal{X}} (1 + V^{\boldsymbol{\pi}}(\boldsymbol{x}))(P^{\boldsymbol{\pi}})^t(d\boldsymbol{x}) \\
&= C + C \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{x}\sim(P^{\boldsymbol{\pi}})^t}[V^{\boldsymbol{\pi}}(\boldsymbol{x})] \\
&= C + C \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{x}\sim(P^{\boldsymbol{\pi}})^{t-1}}[\mathbb{E}_{\boldsymbol{x}'\sim(P^{\boldsymbol{\pi}})}[V^{\boldsymbol{\pi}}(\boldsymbol{x}')|\boldsymbol{x}]] \\
&\le C + C\left(\lim_{t\to\infty} \gamma\mathbb{E}_{\boldsymbol{x}\sim(P^{\boldsymbol{\pi}})^{t-1}}[V^{\boldsymbol{\pi}}(\boldsymbol{x})] + K\right) \qquad \text{(Assumption 2.4)}
\end{aligned}$$

$$
\begin{aligned}
&\leq C + C \lim_{t\to\infty} \gamma^t V^{\boldsymbol{\pi}}(\boldsymbol{x}_0) + K \frac{1-\gamma^t}{1-\gamma} \\
&= C\left(1 + K\frac{1}{1-\gamma}\right)
\end{aligned}
$$

In summary, we have $\mathbb{E}_{\boldsymbol{x}\sim \bar{P}^{\boldsymbol{\pi}}}\left[c(\boldsymbol{x})\right] \leq C\left(1 + K\frac{1}{1-\gamma}\right)$

Consider any $t > 0$, and note that from Theorem A.2 we have

$$
\begin{aligned}
\left\|\mathbb{E}_{\boldsymbol{x}\sim (P^{\boldsymbol{\pi}})^t}\left[c(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim \bar{P}^{\boldsymbol{\pi}}}\left[c(\boldsymbol{x})\right]\right\|_{1+V^{\boldsymbol{\pi}}} &= \sup_{\boldsymbol{x}_0\in\mathcal{X}} \frac{\left|\mathbb{E}_{\boldsymbol{x}\sim (P^{\boldsymbol{\pi}})^t}\left[c(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim \bar{P}^{\boldsymbol{\pi}}}\left[c(\boldsymbol{x})\right]\right|}{1+V^{\boldsymbol{\pi}}(\boldsymbol{x}_0)} \\
&\leq C_1\lambda^t \left\|c - \mathbb{E}_{\boldsymbol{x}\sim \bar{P}^{\boldsymbol{\pi}}}\left[c(\boldsymbol{x})\right]\right\|_{1+V^{\boldsymbol{\pi}}} \qquad \text{(Theorem A.2)} \\
&\leq C_1\lambda^t \left\|c\right\|_{1+V^{\boldsymbol{\pi}}} + C_1\lambda^t \mathbb{E}_{\boldsymbol{x}\sim \bar{P}^{\boldsymbol{\pi}}}\left[c(\boldsymbol{x})\right] \\
&= C_2\lambda^t,
\end{aligned}
$$

where $C_2 = C_1(\|c\|_{1+V^{\boldsymbol{\pi}}} + CK\frac{1}{1-\gamma})$.

Moreover, since the inequality holds for all $\boldsymbol{x}_0$, we have

$$
\frac{\left|\mathbb{E}_{\boldsymbol{x}\sim (P^{\boldsymbol{\pi}})^t}\left[c(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim \bar{P}^{\boldsymbol{\pi}}}\left[c(\boldsymbol{x})\right]\right|}{1+V^{\boldsymbol{\pi}}(\boldsymbol{x}_0)} \leq C_2\lambda^t.
$$

In summary,

$$
\left|\mathbb{E}_{\boldsymbol{x}\sim (P^{\boldsymbol{\pi}})^t}\left[c(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim \bar{P}^{\boldsymbol{\pi}}}\left[c(\boldsymbol{x})\right]\right| \leq C_2(1+V^{\boldsymbol{\pi}}(\boldsymbol{x}_0))\lambda^t.
$$

Consider any $T \geq 0$, and define with $\bar{c} = \mathbb{E}_{\boldsymbol{x}\sim \bar{P}^{\boldsymbol{\pi}}}\left[c(\boldsymbol{x}, \boldsymbol{\pi}(x))\right]$.

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t) - \bar{c}\right] &= \sum_{t=0}^{T-1} \mathbb{E}_{(P^{\boldsymbol{\pi}})^t}\left[c(\boldsymbol{x}_t, \boldsymbol{u}_t)\right] - \bar{c} \\
&\leq \sum_{t=0}^{T-1} \left|\mathbb{E}_{(P^{\boldsymbol{\pi}})^t}\left[c(\boldsymbol{x}_t, \boldsymbol{u}_t)\right] - \bar{c}\right| \\
&\leq C_2(1+V^{\boldsymbol{\pi}}(\boldsymbol{x}_0)) \sum_{t=0}^{T-1} \lambda^t \\
&= C_2(1+V^{\boldsymbol{\pi}}(\boldsymbol{x}_0)) \frac{1-\lambda^T}{1-\lambda}
\end{aligned}
$$

Hence, we have

$$
\lim_{T\to\infty} \left|\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t) - \bar{c}\right]\right| \leq C_2(1+V^{\boldsymbol{\pi}}(\boldsymbol{x}_0)) \frac{1}{1-\lambda},
$$

and for any $\boldsymbol{x}_0$ in a compact subset of $\mathcal{X}$

$$
\lim_{T\to\infty} \frac{1}{T} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t) - \bar{c}\right] = 0.
$$

Moreover,

$$
|B(\boldsymbol{\pi}, \boldsymbol{x}_0)| \leq C_2(1+V^{\boldsymbol{\pi}}(\boldsymbol{x}_0)) \frac{1}{1-\lambda}.
$$

$\square$

Another interesting, inequality that follows from the proof above is the difference in bias inequality.

$$
\left|\mathbb{E}_{\boldsymbol{x}_0\sim\zeta_1}[B(\boldsymbol{\pi}, \boldsymbol{x}_0)] - \mathbb{E}_{\boldsymbol{x}_0\sim\zeta_2}[B(\boldsymbol{\pi}, \boldsymbol{x}_0)]\right| \leq \frac{C_3}{1-\lambda} \int_{\mathcal{X}} (1+V^{\boldsymbol{\pi}}(\boldsymbol{x}))\left|\zeta_1 - \zeta_2\right|(d\boldsymbol{x})
$$

for all probability measures ζ_1, ζ_2. To show this holds, define $C' = \max_{\boldsymbol{\pi}\in\Pi} \|c(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))\|_{1+\theta V^{\boldsymbol{\pi}}}$. Furthermore, note that $C' < \infty$ from Assumption 2.4 and $\|{}^{c(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))}/{}_{C'}\|_{1+\theta V^{\boldsymbol{\pi}}} \leq 1$.

$$
\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{x}\sim(P^{\boldsymbol{\pi}})^t\zeta_1} c(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x})) - \mathbb{E}_{\boldsymbol{x}\sim(P^{\boldsymbol{\pi}})^t\zeta_2} c(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))\right| = \left|\int_{\mathcal{X}} c(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))((P^{\boldsymbol{\pi}})^t\zeta_1 - (P^{\boldsymbol{\pi}})^t\zeta_2)(d\boldsymbol{x})\right| \\
&= C' \left|\int_{\mathcal{X}} \frac{1}{C'} c(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))((P^{\boldsymbol{\pi}})^t\zeta_1 - (P^{\boldsymbol{\pi}})^t\zeta_2)(d\boldsymbol{x})\right| \\
&\leq C' \sup_{\varphi:\|\varphi\|_{1+\theta V^{\boldsymbol{\pi}}}\leq 1} \int_{\mathcal{X}} \varphi(\boldsymbol{x})((P^{\boldsymbol{\pi}})^t\zeta_1 - (P^{\boldsymbol{\pi}})^t\zeta_2)(d\boldsymbol{x}) = C'\rho_\theta^{\boldsymbol{\pi}}((P^{\boldsymbol{\pi}})^t\zeta_1, (P^{\boldsymbol{\pi}})^t\zeta_2) \\
&\leq C'\lambda\rho_\theta^{\boldsymbol{\pi}}((P^{\boldsymbol{\pi}})^{t-1}\zeta_1, (P^{\boldsymbol{\pi}})^{t-1}\zeta_2) \qquad \text{(Theorem A.2)} \\
&\leq C'\lambda^t\rho_\theta^{\boldsymbol{\pi}}(\zeta_1, \zeta_2).
\end{aligned}
$$

Also, note that there exists $C_\theta \in (0, \infty)$ such that $C_\theta \|\varphi\|_{1+\theta V^{\boldsymbol{\pi}}} \geq \|\varphi\|_{1+V^{\boldsymbol{\pi}}}$ due to the equivalence of the two norms.

$$
\begin{aligned}
\rho_\theta^{\boldsymbol{\pi}}(\zeta_1, \zeta_2) &= \sup_{\varphi:\|\varphi\|_{1+\theta V^{\boldsymbol{\pi}}}\leq 1} \int_{\mathcal{X}} \varphi(\boldsymbol{x})(\zeta_1 - \zeta_2)(d\boldsymbol{x}) \\
&\leq \sup_{\varphi:\|\varphi\|_{1+V^{\boldsymbol{\pi}}}\leq C_\theta} \int_{\mathcal{X}} \varphi(\boldsymbol{x})(\zeta_1 - \zeta_2)(d\boldsymbol{x}) \\
&= C_\theta \sup_{\varphi:\|\varphi\|_{1+V^{\boldsymbol{\pi}}}\leq 1} \int_{\mathcal{X}} \varphi(\boldsymbol{x})(\zeta_1 - \zeta_2)(d\boldsymbol{x}) \\
&= C_\theta\rho_1^{\boldsymbol{\pi}}(\zeta_1, \zeta_2)
\end{aligned}
$$

Therefore, for the bias we have

$$
\begin{aligned}
&|\mathbb{E}_{\boldsymbol{x}_0\sim\zeta_1}[B(\boldsymbol{\pi}, \boldsymbol{x}_0)] - \mathbb{E}_{\boldsymbol{x}_0\sim\zeta_2}[B(\boldsymbol{\pi}, \boldsymbol{x}_0)]| \\
&\leq \lim_{T\to\infty} \sum_{t=0}^{T-1} \left|\mathbb{E}_{\boldsymbol{x}\sim(P^{\boldsymbol{\pi}})^t\zeta_1} c(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x})) - \mathbb{E}_{\boldsymbol{x}\sim(P^{\boldsymbol{\pi}})^t\zeta_2} c(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))\right| \\
&\leq C'\rho_\theta^{\boldsymbol{\pi}}(\zeta_1, \zeta_2) \lim_{T\to\infty} \sum_{t=0}^{T-1} \lambda^t = \frac{C'}{1-\lambda}\rho_\theta^{\boldsymbol{\pi}}(\zeta_1, \zeta_2) \\
&\leq \frac{C'C_\theta}{1-\lambda}\rho_1^{\boldsymbol{\pi}}(\zeta_1, \zeta_2) = \frac{C'C_\theta}{1-\lambda}\int_{\mathcal{X}} (1 + V^{\boldsymbol{\pi}}(\boldsymbol{x}))\,|\zeta_1 - \zeta_2|\,(d\boldsymbol{x})
\end{aligned}
$$

Set $C_3 = C'C_\theta$.

A.2 Proof of bounded average cost for the optimistic system

In this section, we show that the results from Theorem 2.6 also transfer over to the optimistic dynamics.

Theorem A.3 (Existence of Average Cost Solution for the Optimistic System)**.** *Let Assumption 2.1 – 2.8 hold. Consider any $n > 0$ and let $\boldsymbol{\pi}_n, \boldsymbol{f}_n$ denote the solution to Equation (6), $P^{\boldsymbol{\pi},\boldsymbol{f}_n}$ its transition kernel. Then $P^{\boldsymbol{\pi},\boldsymbol{f}_n}$ admits a unique invariant measure $\bar{P}^{\boldsymbol{\pi}_n,\boldsymbol{f}_n}$ and there exists $C_2, C_3 \in (0, \infty)$, $\hat{\lambda} \in (0, 1)$ such that*

Average Cost*;*

$$
A(\boldsymbol{\pi}_n, \boldsymbol{f}_n) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\boldsymbol{\pi}_n,\boldsymbol{f}_n}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t)\right] = \mathbb{E}_{\boldsymbol{x}\sim\bar{P}^{\boldsymbol{\pi}_n,\boldsymbol{f}_n}}\left[c(\boldsymbol{x}, \boldsymbol{\pi}_n(x))\right]
$$

Bias Cost*;*

$$
|B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_0)| = \left|\lim_{T\to\infty} \mathbb{E}_{\boldsymbol{\pi}_n,\boldsymbol{f}_n}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t) - A(\boldsymbol{\pi}_n, \boldsymbol{f}_n)\right]\right| \leq C_2(1 + V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0))\frac{1}{1-\hat{\lambda}}
$$

for all $\boldsymbol{x}_0 \in \mathcal{X}$.

Difference in Bias*;*

$$|\mathbb{E}_{\boldsymbol{x}_0\sim\zeta_1}[B(\boldsymbol{\pi}_n,\boldsymbol{f}_n,\boldsymbol{x}_0)] - \mathbb{E}_{\boldsymbol{x}_0\sim\zeta_2}[B(\boldsymbol{\pi}_n,\boldsymbol{f}_n,\boldsymbol{x}_0)]| \leq \frac{C_3}{1-\hat{\lambda}}\int_{\mathcal{X}}(1+V^{\boldsymbol{\pi}}(\boldsymbol{x}))\,|\zeta_1-\zeta_2|\,(d\boldsymbol{x})$$

for all probability measures ζ_1,ζ_2.

Theorem A.3 shows that the optimistic dynamics $\boldsymbol{f}_n$ retain the boundedness property from the true dynamics $\boldsymbol{f}^*$ and give a well-defined solution w.r.t. average cost and the bias cost. To prove Theorem A.3 we show that the optimistic system also satisfies the drift and minorisation condition. Then we can invoke the result from Hairer & Mattingly (2011) similar to the proof of Theorem 2.6.

Lemma A.4 (Stability of optimistic system)**.** *Let Assumption 2.1 – 2.8 hold, then we have with probability at least $1-\delta$ for all $n\geq 0$, $\boldsymbol{\pi}\in\Pi$, $\boldsymbol{f}\in\mathcal{M}_n\cap\mathcal{M}_0$, that there exists a constant $\widehat{K}>0$;*

$$\mathbb{E}_{\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{f},\boldsymbol{\pi}}[V^{\boldsymbol{\pi}}(x')] \leq \gamma V^{\boldsymbol{\pi}}(x)+\widehat{K}.$$

Proof. Note, that $V^{\boldsymbol{\pi}}$ is uniformly continuous w.r.t. κ

$$|V^{\boldsymbol{\pi}}(\boldsymbol{x})-V^{\boldsymbol{\pi}}(\boldsymbol{x}')| \leq \kappa(\|\boldsymbol{x}-\boldsymbol{x}'\|).$$

Furthermore, since $\boldsymbol{f}\in\mathcal{M}_n\cap\mathcal{M}_0$ and therefore $\boldsymbol{f}\in\mathcal{M}_0$, we have that there exists some $\boldsymbol{\eta}\in[-1,1]^{dx}$ such that

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x})) = \boldsymbol{\mu}_0(\boldsymbol{x}.\boldsymbol{\pi}(\boldsymbol{x}))+\beta_0\boldsymbol{\sigma}_0(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))\boldsymbol{\eta}(\boldsymbol{x}).$$

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}}[V^{\boldsymbol{\pi}}(\boldsymbol{\mu}_0(\boldsymbol{x}.\boldsymbol{\pi}(\boldsymbol{x}))+\beta_0\boldsymbol{\sigma}_0(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))\boldsymbol{\eta}(\boldsymbol{x})+\boldsymbol{w})] - \mathbb{E}_{\boldsymbol{w}}[V^{\boldsymbol{\pi}}(\boldsymbol{f}^*(\boldsymbol{x}.\boldsymbol{\pi}(\boldsymbol{x}))+\boldsymbol{w})]\\
&\quad\leq \kappa\left(\|\boldsymbol{\mu}_0(\boldsymbol{x}.\boldsymbol{\pi}(\boldsymbol{x}))+\beta_0\boldsymbol{\sigma}_0(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))\boldsymbol{\eta}(\boldsymbol{x})-\boldsymbol{f}^*(\boldsymbol{x}.\boldsymbol{\pi}(\boldsymbol{x}))\|\right)\\
&\quad\leq \kappa\left(\|\boldsymbol{\mu}_0(\boldsymbol{x}.\boldsymbol{\pi}(\boldsymbol{x}))-\boldsymbol{f}^*(\boldsymbol{x}.\boldsymbol{\pi}(\boldsymbol{x}))\|+\|\beta_0\boldsymbol{\sigma}_0(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}))\boldsymbol{\eta}(\boldsymbol{x})\|\right)\\
&\quad\leq \kappa\left(\left(1+\sqrt{d_x}\right)\beta_0\sqrt{d_x}\sigma_{\max}\right). &&\text{(Assumption 2.8)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{f},\boldsymbol{\pi}}[V^{\boldsymbol{\pi}}(x')] &\leq \mathbb{E}_{\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{f}^*,\boldsymbol{\pi}}[V^{\boldsymbol{\pi}}(x')]+\kappa\left(\left(1+\sqrt{d_x}\right)\beta_0\sqrt{d_x}\sigma_{\max}\right)\\
&= \mathbb{E}_{\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{\pi}}[V^{\boldsymbol{\pi}}(x')]+\kappa\left(\left(1+\sqrt{d_x}\right)\beta_0\sqrt{d_x}\sigma_{\max}\right)\\
&\leq \gamma V^{\boldsymbol{\pi}}(x)+K+\kappa\left(\left(1+\sqrt{d_x}\right)\beta_0\sqrt{d_x}\sigma_{\max}\right). &&\text{(Assumption 2.4)}
\end{aligned}$$

Define $\widehat{K}=K+\kappa\left(\left(1+\sqrt{d_x}\right)\beta_0\sqrt{d_x}\sigma_{\max}\right)$. ☐

Lemma A.5 (Minorisation condition optimistic ystem)**.** *Consider the system*

$$\boldsymbol{x}'=\boldsymbol{f}(\boldsymbol{x}.\boldsymbol{\pi}(\boldsymbol{x}))+\boldsymbol{w}$$

for any $n\geq 0$, $\boldsymbol{\pi}\in\Pi$ and $\boldsymbol{f}\in\mathcal{M}_n\cap\mathcal{M}_0$. Let Assumption 2.1 – 2.8 hold. Let $P^{\boldsymbol{\pi},\boldsymbol{f}}$ denote the transition kernel for the policy $\boldsymbol{\pi}\in\Pi$ i.e., $P^{\boldsymbol{\pi},\boldsymbol{f}}(\boldsymbol{x},\mathcal{A})=\mathbb{P}(\boldsymbol{x}'\in\mathcal{A}|\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}),\boldsymbol{f})$. Then, there exists a constant $\hat{\alpha}\in(0,1)$ and a probability measure $\hat{\zeta}(\cdot)$ independent of n s.t.,

$$\inf_{\boldsymbol{x}\in\mathcal{C}} P^{\boldsymbol{\pi},\boldsymbol{f}}(\boldsymbol{x},\cdot)\geq\hat{\alpha}\hat{\zeta}(\cdot) \tag{11}$$

with $\mathcal{C}\stackrel{\text{def}}{=}\{\boldsymbol{x}\in\mathcal{X};V^{\boldsymbol{\pi}}(\boldsymbol{x})<\hat{R}\}$ for some $\hat{R}>{}^{2\widehat{K}}/_{1-\gamma}$

Proof. First, we show that $\mathcal{C}$ is contained in a compact domain. From the Assumption 2.4 we pick the function $\xi\in\mathcal{K}_\infty$. Since $C_l\xi(0)=0, \lim_{s\to\infty}\xi(s)=+\infty$ and $C_l\xi$ is continuous, there exists M such that $C_l\xi(M)=\hat{R}$. Then for $\|\boldsymbol{x}\|>M$ we have:

$$V^{\boldsymbol{\pi}}(\boldsymbol{x})\geq C_l\xi(\|\boldsymbol{x}\|)>\xi(M)=\hat{R}.$$

Therefore we have: $\mathcal{C} \subseteq \mathcal{B}(\mathbf{0}, M) \stackrel{\text{def}}{=} \{\boldsymbol{x} \mid \|\boldsymbol{x} - \mathbf{0}\| \leq M\}$. Since for any $\boldsymbol{x} \in \mathcal{C}$ we have $\|\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))\| \leq \|\boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))\| + \beta_0\sigma_{\max}$. Since $\boldsymbol{f}^*$ is continuous, there exists a B such that $\boldsymbol{f}^*(\mathcal{C}, \boldsymbol{\pi}(\mathcal{C})) \subset \mathcal{B}(\mathbf{0}, B)$. Therefore we have: $\boldsymbol{f}(\mathcal{C}, \boldsymbol{\pi}(\mathcal{C})) \subset \mathcal{B}(\mathbf{0}, B_1)$, where $B_1 = B + \beta_0\sigma_{\max}$. In the last step we prove that $\alpha \stackrel{\text{def}}{=} 2^{-d_{\boldsymbol{x}}} e^{-B_1^2/\sigma^2}$ and ζ with law of $\mathcal{N}\left(0, \frac{\sigma^2}{2}\right)$ satisfy condition of Lemma A.1. It is enough to show that $\forall \boldsymbol{\mu} \in \mathcal{B}(\mathbf{0}, B_1), \forall \boldsymbol{x} \in \mathbb{R}^{d_{\boldsymbol{x}}}$ we have:

$$\alpha \frac{1}{(2\pi)^{\frac{d_{\boldsymbol{x}}}{2}} \left(\frac{\sigma^2}{2}\right)^{\frac{d_{\boldsymbol{x}}}{2}}} e^{-\frac{\|\boldsymbol{x}\|^2}{\sigma^2}} \leq \frac{1}{(2\pi)^{\frac{d_{\boldsymbol{x}}}{2}} (\sigma^2)^{\frac{d_{\boldsymbol{x}}}{2}}} e^{-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2}{2\sigma^2}}$$

which can be proven with simple algebraic manipulations. □

Proof of Theorem A.3. As for the true system, the drift condition from Lemma A.4 and the minorisation condition from Lemma A.5 are sufficient to show ergodicity of the optimistic system (c.f., Theorem A.2 or Hairer & Mattingly (2011)). The rest of the proof is similar to Theorem 2.6. □

A.3 Proof of Theorem 3.1

Since NeoRL works in artificial episodes $n \in \{0, N-1\}$ of varying horizons H_n. We denote with $\boldsymbol{x}_k^n$ the state visited during episode n at time step $k \leq H_n$. Crucial, to our regret analysis is bounding the first and second moment of $V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)$ for all n, k. Given the nature of Assumption 2.4, this requires analyzing geometric series. Thus, we start with the following elementary result of geometric series.

Corollary A.6. *Consider the sequence $\{S_n\}_{n\geq 0}$ with $S_n \geq 0$ for all n. Let the following hold*

$$S_n \leq \rho S_{n-1} + C$$

for $\rho \in (0, 1)$ and $C > 0$. Then we have

$$S_n \leq \rho^n S_0 + C\frac{1}{1-\rho}.$$

Proof.

$$S_n \leq \rho S_{n-1} + C \leq \rho^2 S_{n-2} + C(1+\rho) \leq \rho^n S_0 + C\sum_{i=0}^{n} \rho^i \leq \rho^n S_0 + C\frac{1}{1-\rho}.$$

□

Lemma A.7. *Let Assumption 2.1 – 2.8 hold and let H_0 be the smallest integer such that*

$$H_0 > \frac{\log\left(C_u/C_l\right)}{\log\left(1/\gamma\right)}.$$

Moreover, define $\nu = \frac{C_u}{C_l}\gamma^{H_0}$. Note, by definition of H_0, $\nu < 1$. Then we have for all $k \in \{0, \ldots, H_n\}$ and $n > 0$

Bounded expectation over horizon

$$\mathbb{E}_{\boldsymbol{x}_k^n, \ldots, \boldsymbol{x}_1^0 | \boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)] \leq \gamma^k \mathbb{E}_{\boldsymbol{x}_0^n, \ldots, \boldsymbol{x}_1^0 | \boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n)] + K/(1-\gamma). \tag{12}$$

Bounded expectation over episodes

$$\mathbb{E}_{\boldsymbol{x}_0^n, \ldots, \boldsymbol{x}_1^0 | \boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n)] \leq \nu^n V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0) + \frac{C_u}{C_l} K/(1-\gamma)\frac{1}{1-\nu}. \tag{13}$$

Moreover, we have

$$\mathbb{E}_{\boldsymbol{x}_k^n, \ldots, \boldsymbol{x}_1^0 | \boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)] \leq D(\boldsymbol{x}_0, K, \gamma, \nu), \tag{14}$$

with $D(\boldsymbol{x}_0, K, \gamma, \nu) = V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0) + K/(1-\gamma)\left(\frac{C_u}{C_l}\frac{1}{1-\nu} + 1\right)$

Proof. We start with proving the first claim

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}_k^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)] &= \mathbb{E}_{\boldsymbol{x}_{k-1}^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[\mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)]] \\
&\leq \mathbb{E}_{\boldsymbol{x}_{k-1}^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[\gamma V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k-1}^n) + K] && \text{(Assumption 2.4)} \\
&= \gamma\mathbb{E}_{\boldsymbol{x}_{k-1}^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k-1}^n)] + K
\end{aligned}$$

We can apply Corollary A.6 to prove the claim. For the second claim, we note that for any $\boldsymbol{\pi}, \boldsymbol{\pi}'$ and $\boldsymbol{x} \in \mathcal{X}$ we have from Assumption 2.4

$$V^{\boldsymbol{\pi}}(\boldsymbol{x}) \leq C_u\alpha(\|\boldsymbol{x}\|) \leq \frac{C_u}{C_l}V^{\boldsymbol{\pi}'}(\boldsymbol{x}).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n)] \\
&\leq \frac{C_u}{C_l}\mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x}_0^n)] \\
&= \frac{C_u}{C_l}\mathbb{E}_{\boldsymbol{x}_{H_n}^{n-1},\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x}_{H_n}^{n-1})] && \text{(Since } \boldsymbol{x}_0^n = \boldsymbol{x}_{H_n}^{n-1}\text{)} \\
&\leq \left(\frac{C_u}{C_l}\gamma^{H_n}\right)\mathbb{E}_{\boldsymbol{x}_0^{n-1},\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x}_0^{n-1})] + \frac{C_u}{C_l}K/(1-\gamma) && \text{(Equation (12))}
\end{aligned}$$

For our choice of H_0, we have for all $n \geq 0$ that $\frac{C_u}{C_l}\gamma^{H_n} \leq \frac{C_u}{C_l}\gamma^{H_0} \leq \nu < 1$. From Corollary A.6, we get

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n)] &\leq \left(\frac{C_u}{C_l}\gamma^{H_n}\right)\mathbb{E}_{\boldsymbol{x}_0^{n-1},\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x}_0^{n-1})] + \frac{C_u}{C_l}K/(1-\gamma) \\
&\leq \nu\mathbb{E}_{\boldsymbol{x}_0^{n-1},\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x}_0^{n-1})] + \frac{C_u}{C_l}K/(1-\gamma) \\
&\leq \nu^n V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0) + \frac{C_u}{C_l}K/(1-\gamma)\frac{1}{1-\nu}. && \text{(Corollary A.6)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}_k^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)] &\leq \gamma^k\mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n)] + K/(1-\gamma) && \text{(Equation (12))} \\
&\leq \mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n)] + K/(1-\gamma) \\
&\leq \nu^n V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0) + \frac{C_u}{C_l}K/(1-\gamma)\frac{1}{1-\nu} + K/(1-\gamma) && \text{(Equation (13))} \\
&\leq V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0) + \frac{C_u}{C_l}K/(1-\gamma)\frac{1}{1-\nu} + K/(1-\gamma)
\end{aligned}$$

□

Lemma A.8. *Let Assumption 2.1 – 2.8 hold and let H_0 be the smallest integer such that*

$$H_0 > \frac{\log\left({}^{C_u}/C_l\right)}{\log\left({}^{1}/\gamma\right)}.$$

Moreover, define $\nu = \frac{C_u}{C_l}\gamma^{H_0}$. Note, by definition of H_0, $\nu < 1$.

Then we have for all $k \in \{0, \ldots, H_n\}$ and $n > 0$

Bounded second moment over horizon

$$\mathbb{E}_{\boldsymbol{x}_k^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n))^2\right] \leq \gamma^{2k}\mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n))^2\right] + \frac{D_2(\boldsymbol{x}_0, K, \gamma, \nu)}{1-\gamma^2} \tag{15}$$

with $D_2(\boldsymbol{x}_0, K, \gamma, \nu) = 2K\gamma D(\boldsymbol{x}_0, K, \gamma, \nu) + K^2 + C_{\boldsymbol{w}}$, *and* $C_{\boldsymbol{w}} = \mathbb{E}_{\boldsymbol{w}}\left[\kappa^2(\|w\|)\right] + 3(\mathbb{E}_{\boldsymbol{w}}\left[\kappa(\|w\|)\right])^2$.

Bounded second moment over episodes

$$\mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n)\right)^2\right] \leq \nu^{2n}\left(V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0)\right)^2 + \left(\frac{C_u}{C_l}\right)^2 \frac{D_2(\boldsymbol{x}_0,K,\gamma,\nu)}{1-\gamma^2}\frac{1}{1-\nu^2}. \tag{16}$$

Moreover, let $D_3(\boldsymbol{x}_0,K,\gamma,\nu) = \left(V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0)\right)^2 + D_2(\boldsymbol{x}_0,K,\gamma,\nu)\left(\left(\frac{C_u}{C_l}\right)^2\frac{1}{1-\gamma^2}\frac{1}{1-\nu^2}+\frac{1}{1-\gamma^2}\right)$.

$$\mathbb{E}_{\boldsymbol{x}_k^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right)^2\right] \leq D_3(\boldsymbol{x}_0,K,\gamma,\nu)$$

Proof. Note that,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right)^2\right] &= \left(\mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right]\right)^2 \\ &\quad + \mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n) - \mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right]\right)^2\right].\end{aligned}$$

We first bound the second term. Let $\bar{\boldsymbol{x}}_k^n = \boldsymbol{f}^*(\boldsymbol{x}_{k-1}^n, \boldsymbol{\pi}_n(\boldsymbol{x}_{k-1}^n))$, i.e., the next state in the absence of transition noise.

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n) - \mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right]\right)^2\right] \\ &\quad= \mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n) - V^{\boldsymbol{\pi}_n}(\bar{\boldsymbol{x}}_k^n) + V^{\boldsymbol{\pi}_n}(\bar{\boldsymbol{x}}_k^n) - \mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right]\right)^2\right] \\ &\quad= \mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n) - V^{\boldsymbol{\pi}_n}(\bar{\boldsymbol{x}}_k^n) + \mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[V^{\boldsymbol{\pi}_n}(\bar{\boldsymbol{x}}_k^n) - V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right]\right)^2\right] \\ &\quad\leq \mathbb{E}_{\boldsymbol{w}}\left[\left(\kappa(\|w\|) + \mathbb{E}_{\boldsymbol{w}}[\kappa(\|w\|)]\right)^2\right] && \text{(uniform continuity of } V^{\boldsymbol{\pi}_n}\text{)} \\ &\quad= \mathbb{E}_{\boldsymbol{w}}\left[\kappa^2(\|w\|)\right] + 3\left(\mathbb{E}_{\boldsymbol{w}}\left[\kappa(\|w\|)\right]\right)^2 \\ &\quad= C_{\boldsymbol{w}} && \text{(Assumption 2.4)}\end{aligned}$$

Therefore we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right)^2\right] &= \left(\mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right]\right)^2 + C_{\boldsymbol{w}} \\ &\leq \left(\gamma V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n) + K\right)^2 + C_{\boldsymbol{w}} \\ &= \gamma^2\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k-1}^n)\right)^2 + 2K\gamma V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k-1}^n) + K^2 + C_{\boldsymbol{w}}.\end{aligned}$$

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_k^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right)^2\right] \\ &= \mathbb{E}_{\boldsymbol{x}_{k-1}^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\mathbb{E}_{\boldsymbol{x}_k^n|\boldsymbol{x}_{k-1}^n}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right)^2\right]\right] \\ &\leq \gamma^2\mathbb{E}_{\boldsymbol{x}_{k-1}^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k-1}^n)\right)^2\right] + 2K\gamma\mathbb{E}_{\boldsymbol{x}_{k-1}^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k-1}^n)\right] + K^2 + C_{\boldsymbol{w}} \\ &\leq \gamma^2\mathbb{E}_{\boldsymbol{x}_{k-1}^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k-1}^n)\right)^2\right] + 2K\gamma D(\boldsymbol{x}_0,K,\gamma,\nu) + K^2 + C_{\boldsymbol{w}}. && \text{(Lemma A.7)}\end{aligned}$$

Let $D_2(\boldsymbol{x}_0,K,\gamma,\nu) = 2K\gamma D(\boldsymbol{x}_0,K,\gamma,\nu) + K^2 + C_{\boldsymbol{w}}$. Applying Corollary A.6 we get

$$\mathbb{E}_{\boldsymbol{x}_k^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n)\right)^2\right] \leq \gamma^{2k}\mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n)\right)^2\right] + \frac{D_2(\boldsymbol{x}_0,K,\gamma,\nu)}{1-\gamma^2}$$

Similar to the first moment, we leverage that $V^{\boldsymbol{\pi}_n}(\boldsymbol{x}) \leq \frac{C_u}{C_l}V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$, $\frac{C_u}{C_l}\gamma^{H_{n-1}} \leq \nu$, and get,

$$\mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n)\right)^2\right]$$

$$
\begin{aligned}
&\leq \left(\frac{C_u}{C_l}\right)^2 \mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[(V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x}_0^n))^2\right] \\
&= \left(\frac{C_u}{C_l}\right)^2 \mathbb{E}_{\boldsymbol{x}_{H_n}^{n-1},\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x}_{H_n}^{n-1})\right)^2\right] && \text{(Since } \boldsymbol{x}_0^n = \boldsymbol{x}_{H_n}^{n-1}\text{)} \\
&\leq \left(\frac{C_u}{C_l}\gamma^{H_n}\right)^2 \mathbb{E}_{\boldsymbol{x}_0^{n-1},\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x}_0^{n-1})\right)^2\right] + \left(\frac{C_u}{C_l}\right)^2 \frac{D_2(\boldsymbol{x}_0,K,\gamma,\nu)}{1-\gamma^2} && \text{(Equation (15))} \\
&\leq \nu^2 \mathbb{E}_{\boldsymbol{x}_0^{n-1},\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left(V^{\boldsymbol{\pi}_{n-1}}(\boldsymbol{x}_0^{n-1})\right)^2\right] + \left(\frac{C_u}{C_l}\right)^2 \frac{D_2(\boldsymbol{x}_0,K,\gamma,\nu)}{1-\gamma^2} \\
&\leq \nu^{2n}\left(V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0)\right)^2 + \left(\frac{C_u}{C_l}\right)^2 \frac{D_2(\boldsymbol{x}_0,K,\gamma,\nu)}{1-\gamma^2}\frac{1}{1-\nu^2} && \text{(Corollary A.6)}
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_k^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_k^n))^2\right] \\
&\leq \gamma^{2k}\mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n))^2\right] + \frac{D_2(\boldsymbol{x}_0,K,\gamma,\nu)}{1-\gamma^2} && \text{(Equation (15))} \\
&\leq \mathbb{E}_{\boldsymbol{x}_0^n,\ldots,\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_0^n))^2\right] + \frac{D_2(\boldsymbol{x}_0,K,\gamma,\nu)}{1-\gamma^2} \\
&\leq \nu^{2n}\left(V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0)\right)^2 + \left(\frac{C_u}{C_l}\right)^2 \frac{D_2(\boldsymbol{x}_0,K,\gamma,\nu)}{1-\gamma^2}\frac{1}{1-\nu^2} + \frac{D_2(\boldsymbol{x}_0,K,\gamma,\nu)}{1-\gamma^2} && \text{(Equation (16))} \\
&\leq \left(V^{\boldsymbol{\pi}_0}(\boldsymbol{x}_0)\right)^2 + D_2(\boldsymbol{x}_0,K,\gamma,\nu)\left(\left(\frac{C_u}{C_l}\right)^2 \frac{1}{1-\gamma^2}\frac{1}{1-\nu^2} + \frac{1}{1-\gamma^2}\right)
\end{aligned}
$$

□

Finally, we prove the regret bound of NeoRL.

Proof of Theorem 3.1. In the following, let $\hat{\boldsymbol{x}}_{k+1}^n = \boldsymbol{f}_n(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n)) + \boldsymbol{w}_k^n$ denote the state predicted under the optimistic dynamics and $\boldsymbol{x}_{k+1}^n = \boldsymbol{f}_n^*(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n)) + \boldsymbol{w}_k^n$ the true state.

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1} c(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n)) - A(\boldsymbol{\pi}^*)\right] \\
&\leq \mathbb{E}\left[\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1} c(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n)) - A(\boldsymbol{\pi}_n, \boldsymbol{f}_n)\right] && \text{(Optimism)} \\
&= \mathbb{E}\left[\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1} B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_k^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \hat{\boldsymbol{x}}_{k+1}^n)\right] && \text{(Bellman equation (Equation (4)))} \\
&= \mathbb{E}\left[\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1} B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_k^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_{k+1}^n) + B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_{k+1}^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \hat{\boldsymbol{x}}_{k+1}^n)\right] \\
&= \sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1} \mathbb{E}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_{k+1}^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \hat{\boldsymbol{x}}_{k+1}^n)\right] && \text{(A)} \\
&+ \sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1} \mathbb{E}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_k^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_{k+1}^n)\right] && \text{(B)}
\end{aligned}
$$

First, we study the term (A).

Proof for (A): Note that because $\boldsymbol{f}_n \in \mathcal{M}_n$, there exists a $\boldsymbol{\eta} \in [-1,1]^{d_x}$ such that $\hat{\boldsymbol{x}}_{k+1}^n = \boldsymbol{\mu}_n(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n)) + \beta_n \boldsymbol{\sigma}_n(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n))\boldsymbol{\eta}(\boldsymbol{x}_k^n) + \boldsymbol{w}_k^n$. Furthermore, $\boldsymbol{x}_{k+1}^n = \boldsymbol{f}^*(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n)) + \boldsymbol{w}_k^n$ and the transition noise is Gaussian. Let $\zeta_{2,k}^n$ and $\zeta_{1,k}^n$ denote the respective distributions of the two random variables, i.e., $\zeta_{1,k}^n \sim \mathcal{N}(\boldsymbol{f}^*(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n)), \sigma^2\mathbb{I})$ and $\zeta_{2,k}^n \sim \mathcal{N}(\boldsymbol{f}_n(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n)), \sigma^2\mathbb{I})$. Next, define $\bar{B} = \mathbb{E}_{\boldsymbol{x}\sim\zeta_{2,k}^n}[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x})]$, and consider the function $h(\boldsymbol{x}) = B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}) - \bar{B}$. Then we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}_k^n}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_{k+1}^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \hat{\boldsymbol{x}}_{k+1}^n)\right] \\
&\quad= \mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim\zeta_{2,k}^n}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x})\right] \\
&\quad= \mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}) - \bar{B}\right] - \mathbb{E}_{\boldsymbol{x}\sim\zeta_{2,k}^n}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}) - \bar{B}\right] \\
&\quad= \mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}[h(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\zeta_{2,k}^n}[h(\boldsymbol{x})].
\end{aligned}$$

Note that $\mathbb{E}_{\boldsymbol{x}\sim\zeta_{2,k}^n}[h(\boldsymbol{x})] = 0$ by the definition of h and thus,

$$\mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}[h(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\zeta_{2,k}^n}[h(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}[h(\boldsymbol{x})] \leq \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}[h^2(\boldsymbol{x})]}. \tag{17}$$

In the following, we bound the term above w.r.t. the Chi-squared distance

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}_k^n}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_{k+1}^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \hat{\boldsymbol{x}}_{k+1}^n)\right] = \mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}[h(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\zeta_{2,k}^n}[h(\boldsymbol{x})] \\
&= \int_{\mathcal{X}} h(\boldsymbol{x})\left(1 - \frac{\zeta_{2,k}^n}{\zeta_{1,k}^n}\right)\zeta_{1,k}^n(d\boldsymbol{x}) \leq \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}\left[h^2(\boldsymbol{x})\right]}\sqrt{d_{\mathcal{X}}(\zeta_{2,k}^n, \zeta_{1,k}^n)} \\
&\hfill ((\text{Kakade et al., 2020, Lemma C.2.,}))
\end{aligned}$$

With $d_{\mathcal{X}}(\zeta_{2,k}^n, \zeta_{1,k}^n)$ being the Chi-squared distance.

$$d_{\mathcal{X}}(\zeta_{2,k}^n, \zeta_{1,k}^n) = \int_{\mathcal{X}} \frac{\left(\zeta_{1,k}^n - \zeta_{2,k}^n\right)^2}{\zeta_{1,k}^n}(d\boldsymbol{x})$$

Since both bounds from Equation (17) and bound we got by applying (Kakade et al., 2020, Lemma C.2.,), we can apply minimum and have:

$$\mathbb{E}_{\boldsymbol{w}_k^n}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_{k+1}^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \hat{\boldsymbol{x}}_{k+1}^n)\right] \leq \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}\left[h^2(\boldsymbol{x})\right]}\sqrt{\min\left\{d_{\mathcal{X}}(\zeta_{2,k}^n, \zeta_{1,k}^n), 1\right\}}$$

Therefore, following Kakade et al. (2020, Lemma C.2.,) we get

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}_k^n}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_{k+1}^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \hat{\boldsymbol{x}}_{k+1}^n)\right] \\
&\quad\leq \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}\left[h^2(\boldsymbol{x})\right]}\min\left\{{}^1\!/\!_\sigma\left\|\boldsymbol{f}^*(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n)) - \boldsymbol{f}_n(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n))\right\|, 1\right\} \\
&\quad\leq \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}\left[h^2(\boldsymbol{x})\right]}(1 + \sqrt{d}_x)^{\beta_n/\sigma}\left\|\boldsymbol{\sigma}_n(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n))\right\|. \qquad ((\text{Sukhija et al., 2024, Cor. 3}))
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}_{\boldsymbol{x}_k^n,\dots\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\mathbb{E}_{\boldsymbol{w}_k^n}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}_{k+1}^n) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \hat{\boldsymbol{x}}_{k+1}^n)\right]\right] \\
&\leq \sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}_{\boldsymbol{x}_k^n,\dots\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\sqrt{\mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}\left[h^2(\boldsymbol{x})\right]}(1 + \sqrt{d}_x)^{\beta_n/\sigma}\left\|\boldsymbol{\sigma}_n(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n))\right\|\right] \\
&\leq \sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}(1 + \sqrt{d}_x)^{\beta_n/\sigma}\sqrt{\mathbb{E}_{\boldsymbol{x}_k^n,\dots\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}\left[h^2(\boldsymbol{x})\right]\right]\mathbb{E}_{\boldsymbol{x}_k^n,\dots\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left\|\boldsymbol{\sigma}_n(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n))\right\|^2\right]} \\
&\leq (1 + \sqrt{d}_x)^{\beta_T/\sigma}\sqrt{\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}_{\boldsymbol{x}_k^n,\dots\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\mathbb{E}_{\boldsymbol{x}\sim\zeta_{1,k}^n}\left[h^2(\boldsymbol{x})\right]\right]} \\
&\times\sqrt{\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}_{\boldsymbol{x}_k^n,\dots\boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\left\|\boldsymbol{\sigma}_n(\boldsymbol{x}_k^n, \boldsymbol{\pi}_n(\boldsymbol{x}_k^n))\right\|^2\right]}
\end{aligned}$$

Here, for the second and third inequality, we use Cauchy-Schwarz. Now we bound the two terms above individually.

First we bound $\mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{1,k}}\left[h^2(\boldsymbol{x})\right]$.

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{1,k}}\left[h^2(\boldsymbol{x})\right] &= \mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{1,k}}\left[(B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}) - \bar{B})^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{1,k}}\left[(B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{2,k}}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x})\right])^2\right] \\
&\le \left(\frac{C_2}{1-\hat{\lambda}}\right)^2 \mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{1,k}}\left[(2 + V^{\boldsymbol{\pi}_n}(\boldsymbol{x}) + \mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{2,k}}\left[V^{\boldsymbol{\pi}_n}(\boldsymbol{x})\right])^2\right] \qquad \text{(Theorem A.3)} \\
&\le \left(\frac{C_2}{1-\hat{\lambda}}\right)^2 \mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{1,k}}\left[(2 + V^{\boldsymbol{\pi}_n}(\boldsymbol{x}) + \gamma V^{\boldsymbol{\pi}_n}(\boldsymbol{x}^n_k) + \hat{K})^2\right] \qquad \text{(Lemma A.4)} \\
&\le \left(\frac{\sqrt{2}C_2}{1-\hat{\lambda}}\right)^2 \mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{1,k}}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}))^2 + (2 + \gamma V^{\boldsymbol{\pi}_n}(\boldsymbol{x}^n_k) + \hat{K})^2\right] \\
&\le \left(\frac{\sqrt{2}C_2}{1-\hat{\lambda}}\right)^2 \left(\mathbb{E}_{\boldsymbol{x}^n_{k+1}|\boldsymbol{x}^n_k}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k+1}))^2\right] + 2\gamma^2(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}^n_k))^2 + 2(2+\hat{K})^2\right)
\end{aligned}
$$

Furthermore, we have from Lemma A.8.

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}^n_k,\dots\boldsymbol{x}^0_1|\boldsymbol{x}_0}\left[\mathbb{E}_{\boldsymbol{x}^n_{k+1}|\boldsymbol{x}^n_k}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k+1}))^2\right] + 2\gamma^2(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}^n_k))^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}^n_{k+1},\dots\boldsymbol{x}^0_1|\boldsymbol{x}_0}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k+1}))^2\right] + 2\gamma^2\mathbb{E}_{\boldsymbol{x}^n_k,\dots\boldsymbol{x}^0_1|\boldsymbol{x}_0}\left[(V^{\boldsymbol{\pi}_n}(\boldsymbol{x}_{k+1}))^2\right] \le (1+2\gamma^2)D_3(\boldsymbol{x}_0, K, \gamma, \nu).
\end{aligned}
$$

In the end, we get

$$
\begin{aligned}
&\sqrt{\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}_{\boldsymbol{x}^n_k,\dots\boldsymbol{x}^0_1|\boldsymbol{x}_0}\left[\mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{1,k}}\left[h^2(\boldsymbol{x})\right]\right]} \\
&\le \left(\frac{\sqrt{2}C_2}{1-\hat{\lambda}}\right)\sqrt{\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}(1+2\gamma^2)D_3(\boldsymbol{x}_0, K, \gamma, \nu) + 2(2+\hat{K})^2} \\
&= \left(\frac{\sqrt{2}C_2}{1-\hat{\lambda}}\right)\sqrt{(1+2\gamma^2)D_3(\boldsymbol{x}_0, K, \gamma, \nu) + 2(2+\hat{K})^2}\sqrt{\sum_{n=0}^{N-1}H_n} \\
&= \left(\frac{\sqrt{2}C_2}{1-\hat{\lambda}}\right)\sqrt{(1+2\gamma^2)D_3(\boldsymbol{x}_0, K, \gamma, \nu) + 2(2+\hat{K})^2}\sqrt{T}.
\end{aligned}
$$

Next, we use the bound from Curi et al. (2020, Lemma 17.) for the second term.

$$
\sqrt{\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}_{\boldsymbol{x}^n_k,\dots\boldsymbol{x}^0_1|\boldsymbol{x}_0}\left[\|\boldsymbol{\sigma}_n(\boldsymbol{x}^n_k, \boldsymbol{\pi}_n(\boldsymbol{x}^n_k))\|^2\right]} \le C'\sqrt{\Gamma_T}
$$

Here Γ_T is the maximum information gain.

If we set $D_4(\boldsymbol{x}_0, K, \gamma) = \frac{C'(1+\sqrt{d_x})}{\sigma}\left(\frac{\sqrt{2}C_2}{1-\hat{\lambda}}\right)\sqrt{(1+2\gamma^2)D_3(\boldsymbol{x}_0, K, \gamma, \nu) + 2(2+\hat{K})^2}$, we have

$$
\begin{aligned}
&\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}_{\boldsymbol{x}^n_k,\dots\boldsymbol{x}^0_1|\boldsymbol{x}_0}\left[\mathbb{E}_{\boldsymbol{w}^n_k}\left[B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \boldsymbol{x}^n_{k+1}) - B(\boldsymbol{\pi}_n, \boldsymbol{f}_n, \hat{\boldsymbol{x}}^n_{k+1})\right]\right] \\
&\le (1+\sqrt{d_x})^{\beta_T/\sigma}\sqrt{\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}_{\boldsymbol{x}^n_k,\dots\boldsymbol{x}^0_1|\boldsymbol{x}_0}\left[\mathbb{E}_{\boldsymbol{x}\sim\zeta^n_{1,k}}\left[h^2(\boldsymbol{x})\right]\right]}
\end{aligned}
$$

$$
\begin{aligned}
&\times \sqrt{\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}_{\boldsymbol{x}_k^n,\dots \boldsymbol{x}_1^0|\boldsymbol{x}_0}\left[\|\boldsymbol{\sigma}_n(\boldsymbol{x}_k^n,\boldsymbol{\pi}_n(\boldsymbol{x}_k^n))\|^2\right]} \\
&\le (1+\sqrt{d}_x)^{\beta_T/\sigma}\left(\frac{\sqrt{2}C_2}{1-\hat{\lambda}}\right)\sqrt{(1+2\gamma^2)D_3(\boldsymbol{x}_0,K,\gamma,\nu)+2(2+\hat{K})^2}\sqrt{T}C'\sqrt{\Gamma_T} \\
&\le D_4(\boldsymbol{x}_0,K,\gamma)\beta_T\sqrt{T\Gamma_T}
\end{aligned}
$$

Proof for (B):

$$
\begin{aligned}
\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}\mathbb{E}\left[B(\boldsymbol{\pi},\boldsymbol{f}_n,\boldsymbol{x}_k^n)-B(\boldsymbol{\pi},\boldsymbol{f}_n,\boldsymbol{x}_{k+1}^n)\right] &= \sum_{n=0}^{N-1}\mathbb{E}\left[B(\boldsymbol{\pi},\boldsymbol{f}_n,\boldsymbol{x}_0^n)-B(\boldsymbol{\pi},\boldsymbol{f}_n,\boldsymbol{x}_{H_n}^n)\right] \\
&\le \frac{C_2}{1-\hat{\lambda}}\sum_{n=0}^{N-1}\left(2+\mathbb{E}\left[V^{\boldsymbol{\pi}}(\boldsymbol{x}_0^n)+V^{\boldsymbol{\pi}}(\boldsymbol{x}_{H_n}^n)\right]\right) && \text{(Theorem A.3)} \\
&\le \frac{2C_2}{1-\hat{\lambda}}\sum_{n=0}^{N-1}(1+D(\boldsymbol{x}_0,K,\gamma)) && \text{(Lemma A.7)} \\
&= \frac{2C_2}{1-\hat{\lambda}}(1+D(\boldsymbol{x}_0,K,\gamma))N \\
&= D_5(\boldsymbol{x}_0,K,\gamma)N.
\end{aligned}
$$

Here $D_5(\boldsymbol{x}_0,K,\gamma)=\frac{2C_2}{1-\hat{\lambda}}(1+D(\boldsymbol{x}_0,K,\gamma))$. Finally, for our choice, $H_n=H_0 2^n$, we get

$$
\sum_{n=0}^{N-1}H_n = H_0\sum_{n=0}^{N-1}2^n = H_0(2^N-1)=T.
$$

Therefore, $N=\log_2\left(\frac{T}{H_0}+1\right)$. To this end, we get for our regret

$$
\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{n=0}^{N-1}\sum_{k=0}^{H_n-1}c(\boldsymbol{x}_k^n,\boldsymbol{\pi}_n(\boldsymbol{x}_k^n))-A(\boldsymbol{\pi}^*)\right] \\
&\le D_4(\boldsymbol{x}_0,K,\gamma)\beta_T\sqrt{T\Gamma_T}+D_5(\boldsymbol{x}_0,K,\gamma)N \\
&\le D_4(\boldsymbol{x}_0,K,\gamma)\beta_T\sqrt{T\Gamma_T}+D_5(\boldsymbol{x}_0,K,\gamma)\log_2\left(\frac{T}{H_0}+1\right)
\end{aligned}
$$

□

This regret is sublinear for a very rich class of functions. We summarize bounds on Γ_T from Vakili et al. (2021) in Table 1. Furthermore, note that $D_4(\boldsymbol{x}_0,K,\gamma)\in(0,\infty)$ for all $\boldsymbol{x}_0\in\mathcal{X}$ with $\|\boldsymbol{x}_0\|<\infty$, $K<\infty$, $\gamma\in(0,1)$. The same holds for $D_5(\boldsymbol{x}_0,K,\gamma)$. Moreover, since $V^{\boldsymbol{\pi}}(\boldsymbol{x})$ is $\Theta(\zeta(\|\boldsymbol{x}\|))$, both D_4 and D_5 are $\Theta(\zeta(\|\boldsymbol{x}_0\|))$.

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

Kernel	$k(\boldsymbol{x},\boldsymbol{x}')$	Γ_T
Linear	$\boldsymbol{x}^\top\boldsymbol{x}'$	$\mathcal{O}(d\log(T))$
RBF	$e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2l^2}}$	$\mathcal{O}\left(\log^{d+1}(T)\right)$
Matèrn	$\frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\boldsymbol{x}'\|}{l}\right)^\nu B_\nu\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\boldsymbol{x}'\|}{l}\right)$	$\mathcal{O}\left(T^{\frac{d}{2\nu+d}}\log^{\frac{2\nu}{2\nu+d}}(T)\right)$

A.4 Relaxing Assumption 2.4

Our analysis assumes that Π consists only of policies with bounded energy. This assumption ensures that during the exploration our system remains stable. The average cost and stability are intertwined for the LQG case (Anderson & Moore, 2007). Moreover, a bounded average cost of a linear controller $\boldsymbol{\pi}(\boldsymbol{x}) = \boldsymbol{K}\boldsymbol{x}$ implies stability and vice-versa. This is not necessarily the case for nonlinear systems, i.e., stability implies a bounded average cost (c.f., Theorem 2.6) but not vice versa. An approach is to assume this link exists also for the nonlinear case.

Definition A.9 (Stable Policies)**.** We call $\Pi_S(\boldsymbol{f})$ the set of stable policies for the dynamics $\boldsymbol{f}$ if there exists positive constants C_u, C_l with $C_u > C_l$, $\zeta, \kappa \in \mathcal{K}_\infty$, $\gamma \in (0, 1)$ s.t., we have for all $\boldsymbol{\pi} \in \Pi_S(\boldsymbol{f})$,

Bounded energy; There exists a Lyapunov function $V^{\boldsymbol{\pi}} : \mathcal{X} \to [0, \infty)$, $K(\boldsymbol{\pi}, \boldsymbol{f}) < \infty$ for which

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

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

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

Bounded norm of cost;

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

Boundedness of noise with respect to κ

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

Assumption A.10 (Bounded average cost implies stability)**.** Consider any dynamics $\boldsymbol{f}$, let $\Pi_A(\boldsymbol{f})$ be the set of policies with bounded average cost for $\boldsymbol{f}$, i.e.,

$$\Pi_A(\boldsymbol{f}) = \{\boldsymbol{\pi} \in \Pi |\ A(\boldsymbol{\pi}, \boldsymbol{f}) < \infty\}. \tag{18}$$

We assume $\forall n \geq 0$, $\boldsymbol{f} \in \mathcal{M}_0 \cap \mathcal{M}_n$ that all policies $\boldsymbol{\pi} \in \Pi_A(\boldsymbol{f})$ are stable, i.e., $\boldsymbol{\pi} \in \Pi_S(\boldsymbol{f})$.

With Assumption A.10 we link the average cost criterion to the stability of our system. A natural consequence of this link is the following corollary.

Corollary A.11. *Let Assumption A.10 hold. Then the following two statements are equivalent for all $n \geq 0$, $\boldsymbol{f} \in \mathcal{M}_0 \cap \mathcal{M}_n$, and $\boldsymbol{\pi} \in \Pi$.*

1. $\boldsymbol{\pi} \in \Pi_A(\boldsymbol{f})$
2. $\boldsymbol{\pi} \in \Pi_S(\boldsymbol{f})$.

Proof. $1 \implies 2$ follows from Assumption A.10 and $2 \implies 1$ from Theorem 2.6. □

Assumption A.12 (Existence of a stable policy)**.** We assume $\Pi_S(\boldsymbol{f}^*) \neq \emptyset$.

Assumption A.12 assumes that there is at least one stable policy in Π. This is in contrast to Assumption 2.4, which assumes that all policies in Π are stable. We can relax this requirement because of Assumption A.10.

In the following, we show that $\boldsymbol{\pi}_n \in \Pi_S(\boldsymbol{f}_n)$ and that this implies $\boldsymbol{\pi}_n \in \Pi_S(\boldsymbol{f}^*)$. In summary, when doing optimistic planning, we inherently pick stable policies for the true system.

Lemma A.13. *Let Assumption 2.1 – 2.2, 2.8, A.10, and Assumption A.12 hold. Let $\boldsymbol{\pi}_n, \boldsymbol{f}_n$ denote the solution to Equation (6). Then we have with probability at least $1 - \delta$, $\boldsymbol{\pi}_n \in \Pi_S(\boldsymbol{f}^*)$.*

Proof. Since $\Pi_S(\boldsymbol{f}^*)$ is nonempty, from Corollary A.11, we must have a policy $\boldsymbol{\pi} \in \Pi_A(\boldsymbol{f}^*)$, and thus $A(\boldsymbol{\pi}) < \infty$. This implies that $A(\boldsymbol{\pi}^*) < \infty$. Since, Equation (6) is an optimistic estimate of $A(\boldsymbol{\pi}^*)$, we have $A(\boldsymbol{\pi}_n, \boldsymbol{f}_n) \leq A(\boldsymbol{\pi}^*) < \infty$. Thus, $\boldsymbol{\pi}_n \in \Pi_A(\boldsymbol{f}_n)$. Again from Corollary A.11, we have $\boldsymbol{\pi}_n \in \Pi_S(\boldsymbol{f}_n)$ and there exists a Lyapunov function $V^{\boldsymbol{\pi}_n}$ and $K(\boldsymbol{\pi}_n, \boldsymbol{f}_n)$ such that

$$\mathbb{E}_{\boldsymbol{x}'|\boldsymbol{f}_n,\boldsymbol{\pi}_n,\boldsymbol{x}}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}')] \leq \gamma V^{\boldsymbol{\pi}_n}(\boldsymbol{x}) + K(\boldsymbol{\pi}_n, \boldsymbol{f}_n)$$

Furthermore, due to the uniform continuity of $V^{\boldsymbol{\pi}_n}$ we have

$$\mathbb{E}_{\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{f}^*,\boldsymbol{\pi}_n}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}')] \leq \mathbb{E}_{\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{f}_n,\boldsymbol{\pi}_n}[V^{\boldsymbol{\pi}_n}(\boldsymbol{x}')] + \kappa\left(\left(1 + \sqrt{d_x}\right)\beta_0\sqrt{d_x}\sigma_{\max}\right) \quad \text{(c.f, Lemma A.4)}$$

$$\leq \gamma V^{\boldsymbol{\pi}_n}(\boldsymbol{x}) + \kappa \left(\left(1 + \sqrt{d_x}\right) \beta_0 \sqrt{d_x} \sigma_{\max} \right) + K(\boldsymbol{\pi}_n, \boldsymbol{f}_n)$$

In summary, we have $\boldsymbol{\pi}_n \in \Pi_S(\boldsymbol{f}^*)$ with $K(\boldsymbol{\pi}_n, \boldsymbol{f}^*) = \kappa \left(\left(1 + \sqrt{d_x}\right) \beta_0 \sqrt{d_x} \sigma_{\max} \right) + K(\boldsymbol{\pi}_n, \boldsymbol{f}_n)$. ☐

Lemma A.13 shows that Equation (6) returns policies that are stable for the true system and therefore with probability at least $1-\delta$ is optimizing over $\Pi_S(\boldsymbol{f}^*)$. Thus, even in cases where Π has policies that do not satisfy Assumption 2.4, these policies are not considered by NeoRL. NeoRL automatically optimizes over $\Pi_S(\boldsymbol{f}^*)$ and the rest of the guarantees follow with $K = \max_{\boldsymbol{\pi} \in \Pi_S(\boldsymbol{f}^*)} K(\boldsymbol{\pi}, \boldsymbol{f}^*)$.

B Experimental Details

In the following, we provide all hyperparameters used in our experiments in Table 2 and the cost function for the environments in Table 3. For NEORL, we use $\beta_n = 2$ for all the experiments, except for the Swimmer and the SoftArm environment where we use $\beta_n = 1$.

Table 2: Hyperparameters for results in Section 4.

Environment	iCEM parameters					Model training parameters						
	Number of samples	Number of elites	Optimizer steps	Horizon	Particles	Number of ensembles	Network architecture	Learning rate	Batch size	Number of epochs	H	Action Repeat
Pendulum-GP	500	50	10	20	5	-	-	0.01	64	-	10	1
Pendulum	500	50	10	20	5	10	256×2	0.001	64	50	10	1
MountainCar	1000	100	5	50	5	10	256×2	0.001	64	50	10	2
Reacher	1000	100	10	50	5	10	256×2	0.001	64	50	10	2
CartPole	1000	100	10	50	5	10	256×2	0.001	64	50	10	2
Swimmer	500	50	10	30	5	10	256×4	0.00005	64	100	200	4
SoftArm	500	50	10	20	5	10	256×4	0.00005	64	50	20	1
RaceCar	1000	100	10	50	5	10	256×2	0.001	64	50	10	1

Table 3: Cost function for the environments presented in Section 4.

Environment	**Cost** $c(\boldsymbol{x}_t, \boldsymbol{u}_t)$
Pendulum	$\theta_t^2 + 0.1\dot{\theta}_t + 0.1u_t^2$
MountainCar	$0.1u_t^2 - 100(1\{\boldsymbol{x}_t \in \boldsymbol{x}_{\text{goal}}\})$
Reacher	$\lVert \boldsymbol{x}_t - \boldsymbol{x}_{\text{target}} \rVert + 0.1 \lVert u_t \rVert$
CartPole	$\left\lVert \boldsymbol{x}_t^{\text{pos}} - \boldsymbol{x}_{\text{target}}^{\text{pos}} \right\rVert^2 + 10(\cos(\theta_t) - 1)^2 + 0.2 \lVert u_t \rVert^2$
Swimmer	$\lVert \boldsymbol{x}_t - \boldsymbol{x}_{\text{target}} \rVert$
SoftArm	$\lVert \boldsymbol{x}_t - \boldsymbol{x}_{\text{target}} \rVert$
RaceCar	$\lVert \boldsymbol{x}_t - \boldsymbol{x}_{\text{target}} \rVert$