Sampling-Based Safe Reinforcement Learning

Luca Vignola[*] ETH Zurich
Bruce D. Lee ETH Zurich
Manish Prajapat ETH Zurich
Manuel Wendl ETH Zurich
Melanie Zeilinger ETH Zurich
Andreas Krause ETH Zurich
Yarden As ETH Zurich

Abstract

Safe exploration remains a fundamental challenge in reinforcement learning (RL), limiting the deployment of RL agents in the real world. We propose *Sampling-Based Safe Reinforcement Learning* (SBSRL), a model-based RL algorithm that maintains safety throughout the learning process by enforcing constraints jointly across a *finite* set of dynamics samples. This formulation approximates an intractable worst-case optimization over uncertain dynamics and enables practical safety guarantees in continuous domains. We further introduce an exploration strategy based on constraining epistemic uncertainty, eliminating the need for explicit exploration bonuses. Under regularity conditions, we derive high-probability guarantees of safety throughout learning and a finite-time sample complexity bound for recovering a near-optimal policy. Empirically, SBSRL achieves safe and efficient exploration both in simulation and in real robotic hardware, and readily extends to practical deep-ensemble implementations that scale to high-dimensional continuous control problems.

1 Introduction

Reinforcement learning (RL) is a framework for sequential decision-making that has achieved remarkable success across a wide range of application domains [1, 2, 3, 4, 5, 6]. Despite this progress, training RL policies *during deployment* remains challenging due to two fundamental barriers: *sample efficiency* and *safety*. Standard RL algorithms typically require a large number of interactions with the environment, which is often impractical or costly in physical systems. Moreover, ensuring safety during training is itself a central challenge: because of epistemic uncertainty about the system dynamics, the consequences of any action are uncertain; in this sense, every interaction with the environment is inherently exploratory, and may cause catastrophic failures. Together, these challenges limit the applicability of RL largely to settings that can be accurately simulated or where high-quality demonstration data is available.

A key challenge in *safe exploration* [7] is to satisfy constraints despite uncertainty, while simultaneously visiting uncertain regions of the state-action space to gather informative data and improve performance. Model-based RL provides a natural framework for addressing this tension by explicitly modeling uncertainty in the learned environment dynamics. Existing approaches [8, 9] typically use such uncertainty estimates to pessimistically tighten constraints to guarantee safety throughout learning, and to add uncertainty-driven bonuses to encourage exploration. However, these methods often **(i)** introduce excessive conservatism through their constraint tightening, and **(ii)** require careful tuning of exploration bonuses to ensure both convergence and effective exploration.

To address these challenges, we propose *Sampling-Based Safe Reinforcement Learning* (SBSRL), a model-based algorithm for safe and sample-efficient exploration. SBSRL learns an uncertainty-

[*]Corresponding author: lvignola@ethz.ch

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

aware dynamics model, leveraging its *epistemic* uncertainty to drive exploration while also maintaining safety at all times. Sampling-based methods provide flexible uncertainty estimates, scale to expressive model classes [10] and admit natural parallelization. SBSRL leverages these properties to obtain a tractable approximation to the robust optimization problem at the core of safe exploration: finding a policy that satisfies safety constraints under *worst-case* dynamics among a set of plausible models of the environment (see Equation (6)). Rather than solving this worst-case problem explicitly, SBSRL enforces safety on each sample of the dynamics, as we illustrate in Figure 1. In addition, rather than adding an exploration bonus to the objective, SBSRL encodes exploration as a constraint that requires the agent to collect sufficient information along its trajectories. As a result, exploration is triggered only when a greedy policy is not informative enough, yielding a simpler and more robust strategy for safe exploration.

Our contribution.

- We propose SBSRL, a novel model-based RL algorithm for safe exploration in continuous state-action spaces. SBSRL enforces safety using a finite number of dynamics samples, and promotes exploration via a constraint on the epistemic uncertainty.
- We show that when the dynamics are modeled using a Gaussian process, SBSRL guarantees safety with high probability. In addition, we provide a sample-complexity bound showing that SBSRL obtains near-optimal policies in a finite number of episodes.
- We empirically validate SBSRL both in simulation and on hardware. With Gaussian process dynamics models, consistent with our theory, SBSRL achieves safe and efficient exploration. We also introduce a scalable neural ensemble variant that enforces constraints jointly across multiple dynamics and evaluate it in SafetyGym and RWRL [11, 12]. Finally, we show that these principles carry over to real-world hardware, enabling safe online learning in practice.

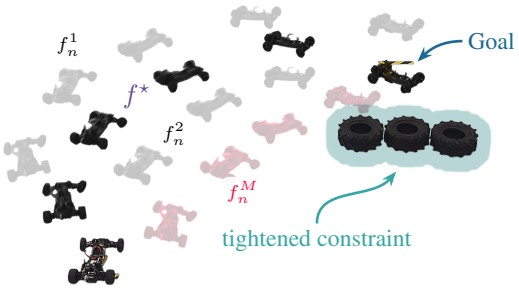

Figure 1: We illustrate a safe exploration task in which a car with unknown dynamics must learn to reach a goal while avoiding obstacles (tires). At each episode n, SBSRL maintains a set of M dynamics samples $f_n^1, \dots, f_n^M$ and enforces tightened safety constraints jointly across their corresponding model-generated rollouts, denoted by transparent cars. For instance, the depicted policy would lead to constraint violations under the plausible model f_n^M. SBSRL updates its policy by simulating it in dynamics $f_n^1, \dots, f_n^M$, thereby ensuring safety under the true system dynamics $f^\star$. As more episodes are collected and uncertainty shrinks, the sampling-based approximation becomes progressively less conservative, enabling SBSRL to discover higher-reward trajectories and ultimately reach the goal.

2 Related Work

Safety in CMDPs. Safety in RL has been modeled in various ways [13, 14, 15]. Among these, constrained Markov decision processes (CMDPs) [16] provide a natural framework, as they enjoy many classical results from planning in MDPs, and can be used to formulate different notions of safety [17, 18, 19]. Many works address learning and planning in CMDPs, both in discrete [20, 21, 22] and continuous [23, 24, 25, 26] state-action spaces. However, these algorithms ensure safety only at convergence, and may violate constraints during learning. This problem of safely learning, i.e., safe exploration, has been tackled in tabular CMDPs [27, 28, 29, 30, 31, 32], where strong guarantees are available, but these approaches do not scale well to continuous problems. Extensions to continuous state-action spaces have been explored in the model predictive control (MPC) literature [33, 34, 35, 36, 37]. However, these techniques rely on online planning, requiring the solution of an optimization problem at each step and thus incurring significant computational cost.

Sampling-based algorithms. Sampling-based methods are popular in robotics for their efficiency and parallelizability [38, 39, 40], but they remain largely restricted to unconstrained settings. Notable exceptions include scenario optimization [41, 42], which requires sampling from the true data-generating process, conformal prediction [43], which typically yields post-hoc guarantees; and dynamics sampling approaches within MPC [37, 44]. Closely related to our work, Prajapat et al.

[44] sample functions from GP dynamics models and enforce constraints jointly across all sampled dynamics within an MPC framework. We apply this principle—to the best of our knowledge, for the first time—to continuous-domain CMDPs. Existing CMDP methods [24, 8] typically tackle epistemic uncertainty via worst-case optimization over plausible models. However, this optimization is generally intractable and approximated via heuristics, sacrificing the original theoretical guarantees. We close this gap by jointly solving the CMDP across all sampled dynamics and establishing theoretical guarantees for the resulting algorithm.

Epistemic exploration and optimality. In online learning, purely greedy approaches under a nominal model may explore inadequately, motivating optimism-based methods [45, 46, 47]. In constrained settings, however, optimism alone is not sufficient, since the feasible set may fail to expand enough to include the optimal solution [48, 49]. In the CMDP setting, As et al. [8] propose ActSafe, an explore-then-commit strategy [50], with simple regret optimality guarantees. However, such pure exploration approaches prioritize information gathering and may sacrifice reward performance in the early stages of learning. More recently, Wendl et al. [9] develop an algorithm based on intrinsic rewards [51] that achieves sublinear cumulative regret. While theoretically grounded, intrinsic reward methods are often sensitive in practice to the choice of hyperparameters, potentially leading to over-exploration or premature exploitation. In contrast, our approach encourages exploration through a constraint on the epistemic uncertainty, allowing the policy to remain greedy whenever the constraint is satisfied. Akin to our work Prajapat et al. [52] adopts an exploration constraint formulation, however in an MPC framework, whereas we establish simple regret guarantees in the CMDP setting.

3 Problem Setting

3.1 Constrained Markov Decision Processes

We consider a discrete-time finite-horizon CMDP defined by the tuple $\langle \mathcal{X}, \mathcal{A}, T, p, r, c, d, \rho_0 \rangle$. The sets $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ denote the state and action spaces, respectively. The episode horizon is T. The reward and cost functions are assumed to be bounded and given by $r : \mathbb{R}^{d_x} \times \mathbb{R}^{d_a} \to [0, R_{\max}]$ and $c : \mathbb{R}^{d_x} \times \mathbb{R}^{d_a} \to [0, C_{\max}]$, respectively. The initial state distribution is denoted as ρ_0 and the transition probability $p(\cdot \mid x, a)$ is induced by the stochastic dynamics

$$x_{t+1} = f^\star(x_t, a_t) + w_t, \quad t = 0, \dots, T-1, \tag{1}$$

where $f^\star$ is an unknown function and $\{w_t\}_{t=0}^{T-1}$ is a sequence of noise variables taking values in $\mathcal{W} \subseteq \mathbb{R}^{d_x}$. The agent selects actions according to a policy $\pi : \mathcal{X} \to \mathcal{A}$ from a class Π. While finite-horizon MDPs require time-dependent policies [53], we omit the dependence on t for notational brevity.

Given a policy π and a transition function f, the noisy dynamics (1) induces a trajectory distribution from which we define the expected reward return $J_r(\pi, f)$ and cost return $J_c(\pi, f)$, where the return of an arbitrary function $g : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is given by

$$J_g(\pi, f) \coloneqq \mathbb{E}_{\tau_\pi^f} \left[\sum_{t=0}^{T-1} g(x_t, a_t) \right]. \tag{2}$$

Here, the subscript τ_π^f in the expectation denotes a trajectory rolled out according to the policy $a_t = \pi(x_t)$ and the dynamics $x_{t+1} = f(x_t, a_t) + w_t$. The expectation is taken over the initial state $x_0 \sim \rho_0$ and the process noise sequence $\{w_t\}_{t=0}^{T-1}$. The goal of the learning agent is to find a policy that maximizes the expected reward return subject to the expected cost remaining below a threshold d, both evaluated under the true, unknown dynamics $f^\star$:

$$\pi^\star \in \arg\max_{\pi \in \Pi} J_r(\pi, f^\star) \quad \text{s.t.} \quad J_c(\pi, f^\star) \leq d. \tag{3}$$

3.2 Task

Since $f^\star$ is unknown, the agent interacts with the system over multiple episodes to acquire information. During episode n, the agent executes policy π_n and collects a trajectory of length T consisting of transitions $\left\{(x_t^n, a_t^n, x_{t+1}^n)\right\}_{t=0}^{T-1}$. Importantly, beyond the final objective in (3), we require safety *during learning*: the policy π_n selected *at each episode* must satisfy the constraint despite not knowing the true dynamics, i.e.,

$$J_c(\pi_n, f^\star) \leq d \quad \forall n \in \{0, 1, \dots\}. \tag{4}$$

Our goal is to design an algorithm that collects data to progressively reduce uncertainty about the system until (3) can be solved up to arbitrary tolerances. The algorithm is designed to provably **(i)** satisfy the constraint on the expected cost return at every episode, **(ii)** terminate after a *finite* number of episodes, and **(iii)** achieve a near-optimal return upon termination. To this end, we make the following assumptions, which are standard in the safe MBRL literature (see, e.g., Assumptions 4.1-4.5 in [8] and Assumptions 5.1-5.2 in [51]).

3.3 Assumptions

Assumption 1 (Gaussian noise)**.** *The process noise is Gaussian i.i.d. with* $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2 I)$.

This assumption can be relaxed to sub-Gaussian noise as in Curi et al. [47] by assuming instead Lipschitz continuous costs and rewards, true dynamics, and policies. The resulting sample complexity would suffer an exponential-in-horizon amplification governed by the Lipschitz constants. Instead, we assume Gaussian noise to invoke Kakade et al.'s "Difference of Gaussians" simulation lemma [54] and obtain tighter sample complexity bounds.

Prior knowledge. In safe exploration, absences of prior knowledge makes safety generally ill-posed, since no policy can be certified as safe without structural assumptions on the unknown dynamics. This requirement is standard in the safe learning literature, where it is typically enforced by assuming access to an initially calibrated model and a corresponding set of safe policies [8, 9]. We therefore require a prior model class that captures admissible deviations from a nominal dynamics estimate.

Assumption 2 (RKHS with bounded norm)**.** *Let* $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{A}$. *We assume access to a known prior mean function* $\mu : \mathcal{Z} \to \mathbb{R}^{d_x}$ *and kernel* $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ *such that, for each component* $f_j^\star$ *of the true dynamics* $f^\star = (f_1^\star, \dots, f_{d_x}^\star)$, *we have that* $f_j^\star - \mu_j$ *is an element of the RKHS* $\mathcal{H}_k$ *induced by the kernel* k *and has a known bound* $B > 0$ *on the RKHS norm, i.e.* $f_j^\star - \mu_j \in \mathcal{H}_k$ *and* $\|f_j^\star - \mu_j\|_k \leq B,\ \forall\, j \in [d_x]$. *Moreover, the kernel satisfies* $k(z,z) \leq \sigma_{\max}\ \forall\, z \in \mathcal{Z}$.

Assumption 2 formalizes that the prior model μ is accurate up to a bounded error measured in RKHS norm. We further require that the prior model class induced by (μ, k, B) admits at least one policy that is uniformly safe over a class of initial admissible dynamics, defined as $\mathcal{F}_0^B \coloneqq \{f \mid |f_j(z) - \mu_j(z)| \leq B\sqrt{k(z,z)}\ \forall\, j \in [d_x], z \in \mathcal{Z}\}$. Note that, by definition of the RKHS norm, every dynamics function satisfying Assumption 2 must also lie in $\mathcal{F}_0^B$.

Assumption 3 (Feasible safe initialization)**.** *There exists a constant* $\Delta > 0$ *such that the set* $\Pi_{\textit{prior}} \coloneqq \{\pi \mid\ J_c(\pi, f) \leq d - \Delta\ \forall f \in \mathcal{F}_0^B\}$ *is non-empty.*

Together, Assumptions 2 and 3 encode that the prior knowledge is both sufficiently expressive to contain the true dynamics and sufficiently informative to certify at least one safe policy. In practice, the prior mean μ, and its corresponding safe initialization, can be obtained in different ways, such as from an offline dataset $\mathcal{D}_0$ or a simulator, with B capturing the resulting model mismatch or sim-to-real gap. This makes the framework agnostic to the source of prior information, as long as a meaningful bound on the discrepancy is available.

To establish continuity of the learned dynamics model and ensure finite sample complexity guarantees, we require two additional regularity assumptions on the prior mean and kernel.

Assumption 4. *The prior mean* μ *and kernel* k *are Lipschitz continuous on* $\mathcal{Z}$ *with constants* L_μ, L_k.

To characterize the complexity of the function class $\mathcal{H}_k$, we introduce the notion of maximum information gain for a kernel k with N observations from the set $\mathcal{Z}$ as $\gamma_N(k) \coloneqq \max_{\mathcal{G} \subset \mathcal{Z},\, |\mathcal{G}| \leq N} \frac{1}{2} \log\det\big(I + \sigma_w^{-2} K_{\mathcal{G}}\big)$, where $K_{\mathcal{G}} \coloneqq [\,k(z,z')\,]_{z,z' \in \mathcal{G}}$. This quantity measures the maximum reduction in uncertainty achievable after N observations. To ensure that the exploration complexity remains finite, we further restrict attention to kernels whose information gain grows sufficiently slowly with N, as formalized in the following assumption.

Assumption 5. *The maximum information gain* $\gamma_N(k)$ *of the kernel* k *over the set* $\mathcal{Z}$ *satisfies* $\gamma_N^3(k) = o(N)$.

Assumptions 4 and 5 are mild and are satisfied by common kernel choices such as linear and squared exponential, as well as Matérn kernels with sufficiently large smoothness parameters, when composed with bounded and Lipschitz continuous feature maps (cf. Appendix A).

4 Background

Gaussian process dynamics models. For our analysis, we model the unknown dynamics using Gaussian processes (GPs) built around a known prior mean function $\mu : \mathcal{Z} \to \mathbb{R}^{d_x}$ and kernel k. At episode n, the GP is updated using all data collected in earlier episodes. For each episode $\ell \in \{0, \dots, n-1\}$ and stage $t \in \{0, \dots, T-1\}$, let $z_t^\ell := (x_t^\ell, a_t^\ell) \in \mathcal{Z} := \mathcal{X} \times \mathcal{A}$ and $y_t^\ell := x_{t+1}^\ell \in \mathcal{X}$. The dataset available at episode n is $\mathcal{D}_{0:n-1} = \bigcup_{\ell=0}^{n-1} \mathcal{D}_\ell, \mathcal{D}_\ell = \{(z_t^\ell, y_t^\ell)\}_{t=0}^{T-1}$. Each state coordinate can be modeled independently with a GP. For $j \in [d_x]$, define the output vector $\mathbf{y}_{n-1,j} := [\, y_{t,j}^\ell \,]_{(\ell,t)}$, the mean vector $\boldsymbol{\mu}_{n-1,j} := [\, \mu_j(z_t^\ell) \,]_{(\ell,t)}$, the kernel vector $\mathbf{k}_{n-1}(z) := [\, k(z_t^\ell, z) \,]_{(\ell,t)}$, and the Gram matrix $\mathbf{K}_{n-1} := [\, k(z_t^\ell, z_{t'}^{\ell'}) \,]_{(\ell,t),(\ell',t')}$, where (ℓ, t) ranges over $\{0, \dots, n-1\} \times \{0, \dots, T-1\}$. Given the prior mean estimate μ_j and the kernel k, the posterior mean and variance for component j are obtained via standard GP regression:

$$\begin{aligned} \mu_{n,j}(z) &= \mu_j(z) + \mathbf{k}_{n-1}(z)^\top (\mathbf{K}_{n-1} + \sigma_w^2 I)^{-1} (\mathbf{y}_{n-1,j} - \boldsymbol{\mu}_{n-1,j}) \\ \sigma_{n,j}^2(z) &= k(z,z) - \mathbf{k}_{n-1}(z)^\top (\mathbf{K}_{n-1} + \sigma_w^2 I)^{-1} \mathbf{k}_{n-1}(z). \end{aligned} \tag{5}$$

The vector-valued mean and standard deviation are denoted by $\mu_n(z) := (\mu_{n,1}(z), \dots, \mu_{n,d_x}(z))$ and $\sigma_n(z) := (\sigma_{n,1}(z), \dots, \sigma_{n,d_x}(z))$. We further define the scalar uncertainty measure $s_n : \mathcal{Z} \to \mathbb{R}$ with $s_n(z) := \|\sigma_n(z)\|$. The following Lemma then establishes that a GP model is *well-calibrated*, in the sense that the true dynamics lie within the confidence intervals induced by the posterior mean and variance with high probability.

Lemma 1 (Vector-valued analog of Theorem 2 of Chowdhury and Gopalan [55])**.** *Let Assumptions 1 and 2 hold. For any $\delta \in (0, 1]$ and $n \geq 0$, define $\beta_n(\delta) := B + \sigma_w \sqrt{2(\gamma_{nT}(k) + 1 + \ln(\frac{d_x}{\delta}))}$. Then with probability at least $1 - \delta$ it holds that for all $n = 0, 1, 2, ...$, for all $z \in \mathcal{Z}$ and for all $j \in [d_x]$ that $|\mu_{n,j}(z) - f_j^\star(z)| \leq \beta_n(\delta)\sigma_{n,j}(z)$.*

These confidence regions are typically used in the literature to conservatively enforce constraints [8]. In our approach, we approximate these regions via sampling.

Approximation via GP sampling. We define a vector-valued *sample* $\tilde{f} = (\tilde{f}_1, \dots, \tilde{f}_{d_x}) \sim GP(\mu, k)$ from the prior GP by concatenating $\tilde{f}_j \sim GP(\mu_j, k)$ for $j \in [d_x]$. The key idea underlying our approach is that, by drawing sufficiently many samples, we can ensure with high probability that at least one is uniformly close to the true dynamics. Drawing continuous function realizations $\tilde{f} \sim GP(\mu, k)$ is computationally intractable due to their infinite dimensional nature. However, since we only need to evaluate these functions at the discrete, finite state-action pairs visited by the agent, several methods can efficiently generate GP samples in practice (cf. Appendix A). We now formalize a sufficient number of samples to achieve at least one that is ζ-close. To do so, we first introduce the *small-ball probability*.

Definition 1 (Small-ball probability; Van Der Vaart and Van Zanten [56])**.** *Let $\tilde{f}_j \sim GP(0, k)$. The* small-ball probability *is defined by* $\Pr\left[\left\|\tilde{f}_j\right\|_\infty < \zeta\right] =: \exp(-\phi(\zeta))$ *for any $\zeta > 0$ where the probability is given by the* small-ball exponent *$\phi : \mathbb{R} \to \mathbb{R}$.*

The small-ball exponent can be bounded using properties of the kernel k. See Prajapat et al. [44] for bounds for common kernel classes.

Lemma 2 (Theorem 1 of Prajapat et al. [44])**.** *Let Assumption 2 hold. Consider any $\delta \in (0, 1]$ and $\zeta > 0$. Select M satisfying*

$$M \geq \frac{\log(\delta)}{\log(1 - \exp(-d_x(\frac{1}{2}B^2 + \phi(\zeta))))}.$$

Let $\tilde{f}^1, \dots, \tilde{f}^M \sim GP(\mu, k)$ be function samples from the prior GP. It holds with probability at least $1 - \delta$ that for some $m \in [M]$, $\left\|\tilde{f}_j^m - f_j^\star\right\|_\infty \leq \zeta \;\; \forall j \in [d_x]$.

See Appendix B for the proof of Lemma 2. By characterizing the number of samples required to achieve a ζ-close approximation to an arbitrary function in the RKHS, the above lemma provides a conservative *worst-case* bound on the number of GP samples required so that enforcing constraints on *sampled* dynamics implies that they also hold on the *true unknown dynamics*. While this bound

can be large, it is required to formally guarantee safety with high-probability for a very general class of problems. In practice, we observe that significantly fewer samples are sufficient to obtain safe policies. Next, we describe how this bound is used in our algorithm design.

5 Algorithm

Enforcing safety via sampling. Given a well-calibrated model (Lemma 1), satisfying the safety constraint in each episode n amounts to solving a worst-case optimization given by

$$\max_{f \in \mathcal{F}_n} J_c(\pi, f) \leq d, \tag{6}$$

where $\mathcal{F}_n \coloneqq \{f \colon |\mu_{n,j}(z) - f_j(z)| \leq \beta_n(\delta)\sigma_{n,j}(z)\ \forall j \in [d_x]\}$. Prior work in safe MBRL addresses this in practice via heuristics [8, 24] or pessimistic cost penalties [9]. We instead approximate this optimization over the class $\mathcal{F}_n$ with a finite set of M sampled models $\tilde{f}^1, \ldots, \tilde{f}^M$ drawn i.i.d. from the GP prior. For sufficiently large M (see Lemma 2), at least one sample is ζ-close to $f^\star$ with high probability, allowing us to ensure safety on the true system by enforcing constraints over the sampled models.

Truncating unlikely samples. Since these samples are drawn from the prior and not yet informed by data, we refine them online by truncation [37]. At every episode n, each sample is truncated to lie within the intersection of all previous GP confidence sets,

$$f^m_{n,j}(z) = \mathrm{clip}(f^m_{n-1,j}(z), \mu_{n,j}(z) - \beta_n(\delta)\sigma_{n,j}(z), \mu_{n,j}(z) + \beta_n(\delta)\sigma_{n,j}(z)), \tag{7}$$

for all $z \in \mathcal{Z}, j \in [d_x]$. For $n = 0$, we overload the notation and apply Equation 7 to $f^m_{-1,j} \coloneqq \tilde{f}^m_j$, $\mu_0(z) \coloneqq \mu(z)$, $\sigma_{0,j}(z) \coloneqq \sqrt{k(z,z)}$, and $\beta_0(\delta) \coloneqq B$, yielding $f^m_0 \in \mathcal{F}^B_0\ \forall m \in [M]$. For well-calibrated models (cf. Lemma 1), truncation can only reduce deviation from $f^\star$, thereby preserving the ζ-close sample. This idea is illustrated in Figure 2.

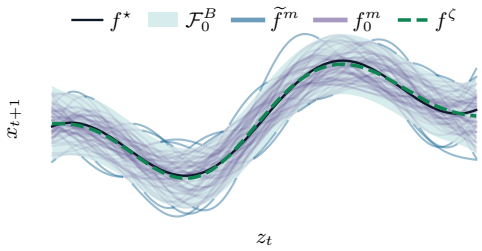

Figure 2: At the initial episode $n = 0$, we draw M samples $\tilde{f}^m$ from the GP prior. These are then clipped using the prior bound $\mathcal{F}_0^B$, yielding the truncated samples f_0^m. Finally, f^ζ denotes the ζ-close sample of Lemma 2.

Constraint tightening. To account for the gap between this sample and $f^\star$, we then tighten the constraint by a ζ-dependent constant and enforce it jointly across all samples:

$$J_c(\pi, f^m_n) \leq d - \Delta_\zeta \quad \forall m = 1, \ldots, M,$$

which guarantees safety under the true dynamics with high probability. Intuitively, the tightening factor Δ_ζ, defined in Algorithm 1, bounds the discrepancy between returns under $f^\star$ and a ζ-close model. This yields a natural tradeoff: increasing ζ reduces the number of samples M required by Lemma 2 but increases conservatism, while smaller ζ improves tightness at higher computational cost. We note that truncation is only one approach for refining GP samples with data. For other approaches, including posterior sampling, we refer the reader to Section A.

Enforcing exploration as a constraint. Prior safe MBRL methods often encourage exploration via intrinsic reward bonuses, where the agent maximizes $J_r(\pi, \mu_n) + \lambda_n^{\text{explore}} J_{s_n}(\pi, \mu_n)$ [9]. In practice, these approaches require careful tuning of $\lambda_n^{\text{explore}}$, and may either over-explore irrelevant regions or fail to explore sufficiently under model misspecification. In contrast, we decouple reward maximization and exploration by enforcing a constraint that the epistemic accumulated along trajectories under our nominal dynamics exceeds a threshold. Concretely, at episode n, we require the policy to satisfy $J_{s_n}(\pi_n, \mu_n) \geq d^n_\sigma$, where d^n_σ is a prescribed exploration threshold defined in Theorem 2. Importantly, this constraint becomes inactive when the greedy policy is already sufficiently informative, in which case no additional exploration is enforced. Moreover, infeasibility of the constraint provides a natural stopping criterion, certifying that remaining policies induce uniformly low epistemic uncertainty and linking the exploration threshold directly to the final suboptimality guarantee.

Sampling-Based Safe Reinforcement Learning (SBSRL). We now propose a model-based reinforcement learning algorithm that combines sampling-based constraint satisfaction with the

exploration constraint. In particular, as shown in Algorithm 1, at each episode n, the agent selects the policy π_n as

$$\pi_n = \arg\max_{\pi \in \Pi} J_r(\pi, \mu_n) \text{ s.t. } J_c(\pi, f_n^m) \leq d - \Delta_\zeta \; \forall m \in [M] \text{ and } J_{s_n}(\pi, \mu_n) \geq d_\sigma^n. \quad (8)$$

Algorithm 1 Sampling-Based Safe Reinforcement Learning (SBSRL)

Require: Confidence parameter δ, closeness threshold $\zeta \in (0, \frac{\sigma_w \Delta}{\sqrt{d_x} T^2 C_{\max}})$, exploration threshold sequence $\{d_\sigma^n\}_{n \geq 0}$, prior mean μ and kernel k

Dataset $\mathcal{D}_0 \leftarrow \varnothing$, episode $n \leftarrow 0$
Set constraint tightening $\Delta_\zeta \leftarrow \zeta \sqrt{d_x} \frac{T^2 C_{\max}}{\sigma_w}$
Set number of dynamics samples M to satisfy the condition of Lemma 2 under μ, k, δ, ζ
Draw M dynamics samples $\tilde{f}^1, \ldots, \tilde{f}^M \sim GP(\mu, k)$
while Problem (8) is feasible **do**
    Construct truncated dynamics samples $f_n^1, \ldots, f_n^M$ according to (7)
    Select π_n by solving Problem (8)
    $\mathcal{D}_{0:n} \leftarrow \mathcal{D}_{0:n-1} \cup$ rollout π_n
    Update model $\mathcal{GP}(\mu_{n+1}, \sigma_{n+1}) \leftarrow \mathcal{D}_{0:n}$
    $n \leftarrow n + 1$
end while
Select $\hat{\pi}$ by solving Problem (8) after removing the constraint on J_{s_n} (greedy policy).

In the following, we present our theoretical guarantees for SBSRL. We provide further details regarding our practical implementation of SBSRL in Section F.

6 Results

We now prove that with high probability, SBSRL satisfies the constraint at all episodes, terminates after a finite number of steps, and is nearly optimal upon termination. We defer proofs to the supplement.

Our first result shows that solving Problem (8) guarantees safety of the policy π_n at all episodes n with high probability. For this, define the set policies which satisfy the safety constraint of Problem (8) as

$$\Pi_{\text{safe}}^n := \{\pi \in \Pi \quad | \quad J_c(\pi, f_n^m) \leq d - \Delta_\zeta \quad \forall m = 1, \ldots, M\}.$$

Theorem 1 (Safety). *Suppose Assumptions 1-3 hold. Consider any $\delta \in (0, 1/2)$, $\zeta \in (0, \frac{\sigma_w \Delta}{\sqrt{d_x} T^2 C_{\max}})$, and sequence of positive numbers $\{d_\sigma^n\}_{n \geq 0}$. Consider running Algorithm 1 given these values. Then, it holds with probability at least $1 - 2\delta$ that for every episode n, any policy $\pi \in \Pi_{safe}^n$ satisfies $J_c(\pi, f^\star) \leq d$. In particular, this holds for the policy π_n selected during episode n and the policy $\hat{\pi}$ selected upon algorithm termination.*

Moreover, under the success event of Lemma 1, Assumption 3 ensures that the sampled safe set Π_{safe}^n is non-empty for every episode $n \geq 0$ (cf. Appendix A). Consequently, infeasibility can only occur due to the exploration constraint. This is studied in the following theorem, which bounds the number of episodes played by the algorithm before termination, leading to near-optimality guarantees. We next introduce the set of ξ-tightened safe policies under the true dynamics,

$$\Pi_\xi^\star := \{\pi \in \Pi \quad | \quad J_c(\pi, f^\star) \leq d - \xi\}.$$

In general, $\Pi_\xi^\star$ may be disconnected, with some regions that cannot be reached from the initial safe set Π_{prior} via safe policy updates. We therefore restrict our attention to a *connected* component $\Pi_\xi^{\star,c} \subseteq \Pi_\xi^\star$, defined as follows.

Definition 2. *We define $\Pi_\xi^{\star,c} \subseteq \Pi_\xi^\star$ as the subset of $\Pi_\xi^\star$ satisfying the following path-connectedness condition with respect to Π_{prior}: for every $\pi \in \Pi_\xi^{\star,c}$ there exists a continuous path $\rho : [0,1] \to \Pi_\xi^{\star,c}$ such that $\rho(0) = \pi$ and $\rho(1) \in \Pi_{prior}$. Continuity is with respect to the norm $\|\cdot\|_\infty$, defined as, $\|\pi - \pi'\|_\infty := \sup_{x \in \mathcal{X}} \|\pi(x) - \pi'(x)\|$, where $\|\cdot\|$ denotes the Euclidean norm in the action space.*

Similar notions of reachability are standard in the safe online learning literature [28, 57, 52].

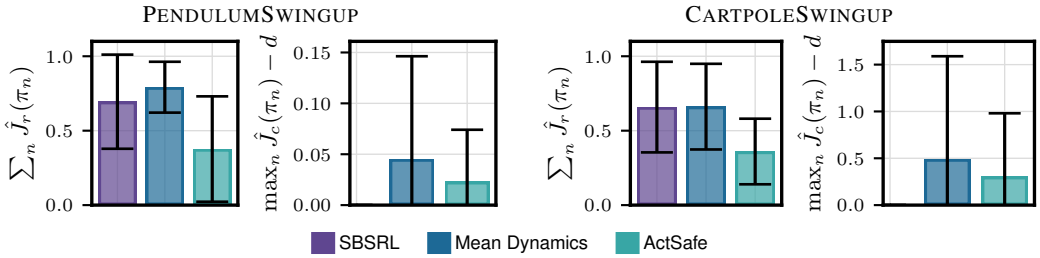

Figure 4: Evaluation of safety on PENDULUMSWINGUP and CARTPOLESWINGUP. We report mean and standard deviation of the normalized cumulative returns and maximum cost violations over five seeds. SBSRL maintains safety throughout training while achieving near-optimal performance.

Theorem 2 (Optimality). *Suppose Assumptions 1-5 hold. Fix some* $\delta \in (0, 1/2)$, $\zeta \in (0, \frac{\sigma_w \Delta}{\sqrt{d_x} T^2 C_{\max}})$, $\varepsilon > 0$ *and* $\xi > \varepsilon + \zeta \sqrt{d_x} \frac{T^2 C_{\max}}{\sigma_w}$. *Let* $d_\sigma^n = \frac{\varepsilon \sigma_w}{2 G_{max} T \beta_n(\delta)}$ *with* $G_{max} := \max\{C_{max}, R_{max}\}$, *and consider running Algorithm 1. There exists a problem-dependent constant* $C = C(d_x, \sigma_{\max}, G_{\max}, T, \delta)$ *such that if* $\bar{n}$ *is the smallest integer satisfying* $\bar{n} \geq C \frac{\gamma_{\bar{n}T}(k)^3}{\varepsilon^2}$, *then with probability at least* $1 - 2\delta$ *the following hold:*

1. *There exists* $n^\star \leq \bar{n}$ *such that* $J_{s_{n^\star}}(\pi, \mu_{n^\star}) < d_\sigma^{n^\star}$ $\forall \pi \in \Pi_{safe}^{n^\star}$, *i.e. Algorithm 1 terminates at iteration* $n^\star$.
2. *At termination, the policy* $\hat{\pi}$ *returned by Algorithm 1 satisfies*

$$\max_{\pi \in \Pi_\xi^{\star,c}} J_r(\pi, f^\star) - J_r(\hat{\pi}, f^\star) \leq \varepsilon. \tag{9}$$

The theorem guarantees near-optimality over the reachable component $\Pi_\xi^{\star,c}$ in finite time. In the definition of ξ, ε accounts for the fact that the dynamics can only be learned up to a non-zero error in finite time, while ζ captures the sampling approximation error and can be reduced by increasing the number of samples. At termination, the infeasibility of the exploration constraint implies uniformly low epistemic uncertainty over $\Pi_{\text{safe}}^{n^\star}$, allowing us to control the discrepancy between true, mean, and sampled dynamics and ensure that reward and cost are ε-accurate for all safe policies. In particular, this implies $\Pi_{\text{safe}}^{n^\star} \supseteq \Pi_\xi^{\star,c}$, as illustrated in Figure 3.

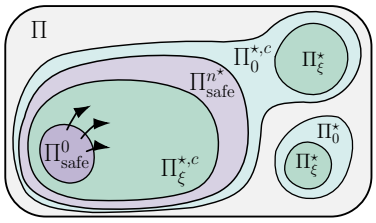

Figure 3: At termination, $\Pi_{\text{safe}}^{n^\star} \supseteq \Pi_\xi^{\star,c}$. The set $\Pi_\xi^\star$ may be disconnected, and only its reachable component $\Pi_\xi^{\star,c}$ can be identified via safe updates. The tightening ξ influences connectivity: smaller values can merge otherwise disconnected regions, as illustrated for $\xi = 0$.

7 Experiments

We empirically evaluate SBSRL in two complementary settings: **(i)** a GP-based instantiation that closely aligns with our theory, evaluated in a simulated environment; and **(ii)** a real-world safe RL setting, where SBSRL is implemented using deep ensembles. Together, these experiments support our theoretical findings and demonstrate that these principles can scale in practice to high-dimensional scenarios. Further extensive empirical evaluation, including comparisons of our deep variant against state-of-the-art safe exploration algorithms on SafetyGym [11] and RWRL [12], is deferred to the appendix, together with additional implementation details.

Safety and optimality with GPs. We first conduct an experiment where we use GPs to model the dynamics. We evaluate SBSRL on the PENDULUMSWINGUP and CARTPOLESWINGUP tasks from the Deepmind Control Suite [58], following the setup of [8]. As baselines, we consider ActSafe [8] and a baseline where the constraint is enforced only on the GP posterior mean, rather than on each sampled dynamics model as prescribed by Problem (8). Results in Figure 4 show that SBSRL achieves near-optimal performance while satisfying the safety constraint at every iteration n. In contrast, enforcing constraints on the mean dynamics leads to safety violations in both tasks, highlighting the importance of accounting for model uncertainty at the planning stage, and validating dynamics sampling as a way to ensure safety during learning.

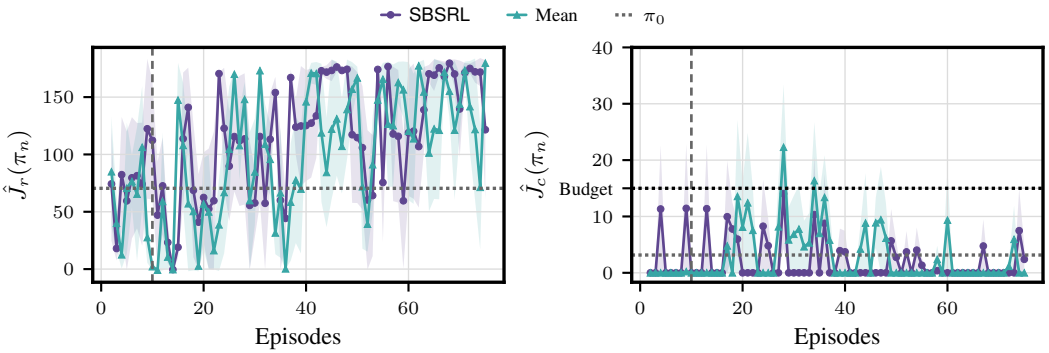

Figure 6: Safe offline-to-online on real-world hardware. We report mean and standard error over three seeds. The vertical line marks the transition from warm-up (offline prior π_0) to online learning. The cost represents the kinetic energy lost in collisions with the obstacle.

Exploration. We next validate the exploration strategy of SBSRL. Specifically, we consider the PENDULUMSWINGUP task and compare against two variants of ActSafe that perform pure exploration for the first $n = 3$ and $n = 5$ episodes, respectively, before switching to greedy exploitation. As shown in Figure 5, SBSRL shows superior exploration efficiency in this task, converging more rapidly to an optimal policy. This stems from the exploration constraint, which promotes exploration while still allowing reward maximization whenever this is sufficiently informative. We further illustrate how the choice of d_σ^0 can influence the behavior of SBSRL. From our theory, smaller values of d_σ^0 guarantee closer-to-optimal performance at termination but reduce sample efficiency, which helps explain why increasing its value may actually improve convergence in practice. It is also interesting to notice that such a constraint makes little difference in the first training episode, where the greedy policy is already sufficiently informative, consistently with our theoretical insights.

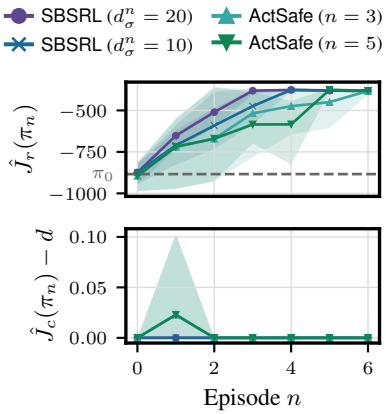

Figure 5: Evaluation of exploration via an exploration constraint in the PENDULUMSWINGUP task. We report mean and 95 percentile interval over five seeds.

Safe offline-to-online in the real-world. Finally, we demonstrate how SBSRL can be successfully deployed on *robotic hardware*, as presented in Figure 1 and detailed in Appendix F. We instantiate SBSRL using deep ensembles (see Appendix F) and test it on a highly-dynamic remote-controlled race car operating at 60 Hz. The combination of high-frequency control and delays in actuation and motion capture makes this high-dimensional environment particularly challenging. This challenge is partially mitigated by using a sliding window of the five past state measurements, resulting in an observation of 45 dimensions. Having no prior knowledge would make this task ill-posed (recall Assumption 3). Therefore, to calibrate our model, we initialize SBSRL offline using a dataset of real-world trajectories. We then deploy the algorithm online and compare it with a baseline that only enforces the constraint on the mean critic rather than across the ensemble. The results in Figure 6 show that SBSRL improves the rewards upon the prior policy and maintains safety at all times, while using the mean estimate incurs higher costs during training.

8 Conclusion

In this work, we propose SBSRL, a novel model-based reinforcement learning algorithm for safe exploration. SBSRL ensures safety throughout learning by enforcing constraints over a *finite* set of sampled dynamics models. When the dynamics are represented by Gaussian processes, this yields high-probability safety guarantees. Furthermore, we provide a simple regret bound that leverages a novel strategy for encouraging exploration in CMDPs, namely a constraint on the epistemic uncertainty which encourages exploration only when the greedy policy is insufficiently informative. We extensively validate SBSRL both in simulation and on real-world hardware, demonstrating that its deep RL variant can be scaled to high-dimensional continuous control tasks. Finally, while our

work provides a theoretically sound yet practical algorithm for safe exploration in CMDPs, several challenges remain. First, our work builds on the finite-horizon episodic setting, effectively requiring manual resets that are difficult to implement in practice. Extending this work to the non-episodic setting is an exciting direction for future work. In addition, this work builds on the classical expected accumulated cost formulation of constraints [16]. Extending this work to formulations that enforce state-wise safety with high-probability for high-dimensional non-linear dynamics and general policies (e.g., beyond MPC) is an important direction for future theoretical and empirical work.

Acknowledgments

We would like to thank Armin Lederer and Yunke Ao for insightful discussions during the development of this project. M.P. was supported by a ETH AI Center Doctoral Fellowship, and B.L. was supported by an ETH AI Center Postdoctoral Fellowship. Y.A. received funding from a grant of the Hasler foundation (grant no. 21039) and the ETH AI Center.

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei Rusu, Joel Veness, Marc Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015. (Cited on page 1)

[2] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 2017. (Cited on page 1)

[3] Jens Kober, J. Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 2013. (Cited on page 1)

[4] Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martin-Martin, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. (Cited on page 1)

[5] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, and Martin Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 2022. (Cited on page 1)

[6] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. (Cited on page 1)

[7] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. (Cited on page 1)

[8] Yarden As, Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Stelian Coros, and Andreas Krause. Actsafe: Active exploration with safety constraints for reinforcement learning. In *International Conference on Learning Representations*, 2025. (Cited on pages 1, 3, 4, 5, 6, 8, 16, 17, 18, 23, 25, 26, 27, and 28)

[9] Manuel Wendl, Yarden As, Manish Prajapat, Anton Pollak, Stelian Coros, and Andreas Krause. Safe exploration via policy priors. In *The Fourteenth International Conference on Learning Representations*, 2026. (Cited on pages 1, 3, 4, 6, 28, and 29)

[10] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *International Conference on Neural Information Processing Systems*, 2018. (Cited on pages 2, 25, 27, and 29)

[11] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. 2019. (Cited on pages 2, 8, 27, 28, and 29)

[12] Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. An empirical investigation of the challenges of real-world reinforcement learning, 2021. (Cited on pages 2, 8, 27, 28, and 29)

[13] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 2015. (Cited on page 2)

[14] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 2022. (Cited on page 2)

[15] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. (Cited on page 2)

[16] E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999. (Cited on pages 2 and 10)

[17] Nolan Wagener, Byron Boots, and Ching-An Cheng. Safe reinforcement learning using advantage-based intervention. In *International Conference on Machine Learning*, 2021. (Cited on page 2)

[18] Sebastian Curi, Armin Lederer, Sandra Hirche, and Andreas Krause. Safe reinforcement learning via confidence-based filters. In *Conference on Decision and Control*. IEEE, 2022. (Cited on page 2)

[19] Akifumi Wachi, Wataru Hashimoto, Xun Shen, and Kazumune Hashimoto. Safe exploration in reinforcement learning: A generalized formulation and algorithms. In *International Conference on Neural Information Processing Systems*, 2023. (Cited on page 2)

[20] Sharan Vaswani, Lin Yang, and Csaba Szepesvari. Near-optimal sample complexity bounds for constrained MDPs. In *International Conference on Neural Information Processing Systems*, 2022. (Cited on page 2)

[21] Dongsheng Ding, K. Zhang, Jiali Duan, Tamer Bacsar, and Mihailo R. Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. *ArXiv*, 2022. (Cited on page 2)

[22] Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained mdps. In *International Conference on Machine Learning*, 2024. (Cited on page 2)

[23] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, 2017. (Cited on page 2)

[24] Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause. Constrained policy optimization via bayesian world models. In *International Conference on Learning Representations*, 2022. (Cited on pages 2, 3, 6, and 28)

[25] Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, and Haitham Ammar. Saute rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*, 2022. (Cited on page 2)

[26] Weidong Huang, Jiaming Ji, Chunhe Xia, Borong Zhang, and Yaodong Yang. Safedreamer: Safe reinforcement learning with world models. In *International Conference on Learning Representations*, 2024. (Cited on page 2)

[27] Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810*, 2012. (Cited on page 2)

[28] Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite markov decision processes with gaussian processes. *International Conference on Neural Information Processing Systems*, 2016. (Cited on pages 2 and 7)

[29] Akifumi Wachi, Yanan Sui, Yisong Yue, and Masahiro Ono. Safe exploration and optimization of constrained mdps using gaussian processes. *AAAI*, 2018. (Cited on page 2)

[30] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, 2020. (Cited on page 2)

[31] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International conference on artificial intelligence and statistics*, 2021. (Cited on page 2)

[32] Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *International Conference on Neural Information Processing Systems*, 2022. (Cited on page 2)

[33] Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based model predictive control for safe exploration. *IEEE Conference on Decision and Control*, 2018. (Cited on page 2)

[34] Lukas Hewing, Juraj Kabzan, and Melanie N Zeilinger. Cautious model predictive control using gaussian process regression. *IEEE Transactions on Control Systems Technology*, 2019. (Cited on page 2)

[35] Kim Peter Wabersich and Melanie N Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*, 2021. (Cited on page 2)

[36] Manish Prajapat, Johannes Köhler, Matteo Turchetta, Andreas Krause, and Melanie N. Zeilinger. Safe guaranteed exploration for non-linear systems. *IEEE Transactions on Automatic Control*, 2025. (Cited on page 2)

[37] Manish Prajapat, Amon Lahr, Johannes Köhler, Andreas Krause, and Melanie N Zeilinger. Towards safe and tractable gaussian process-based mpc: Efficient sampling within a sequential quadratic programming framework. In *IEEE Conference on Decision and Control*. IEEE, 2024. (Cited on pages 2, 6, 15, and 16)

[38] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2017. (Cited on page 2)

[39] Thomas Lew, Lucas Janson, Riccardo Bonalli, and Marco Pavone. A simple and efficient sampling-based algorithm for general reachability analysis. In *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, 2022. (Cited on page 2)

[40] Grady Williams, Paul Drews, Brian Goldfain, James M. Rehg, and Evangelos A. Theodorou. Aggressive driving with model predictive path integral control. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016. (Cited on page 2)

[41] G. Calafiore and Marco Campi. The scenario approach to robust control design. *Automatic Control, IEEE Transactions on*, 2006. (Cited on page 2)

[42] Jonas Umlauft, Thomas Beckers, and Sandra Hirche. Scenario-based Optimal Control for Gaussian Process State Space Models. In *2018 European Control Conference (ECC)*, 2018. (Cited on pages 2 and 15)

[43] Lars Lindemann, Yiqi Zhao, Xinyi Yu, George J. Pappas, and Jyotirmoy V. Deshmukh. Formal verification and control with conformal prediction: Practical safety guarantees for autonomous systems. *IEEE Control Systems*, 2025. (Cited on page 2)

[44] Manish Prajapat, Johannes Köhler, Amon Lahr, Andreas Krause, and Melanie N Zeilinger. Finite-sample-based reachability for safe control with gaussian process dynamics. *arXiv preprint arXiv:2505.07594*, 2025. (Cited on pages 2, 3, 5, 15, 16, and 17)

[45] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *International Conference on Neural Information Processing Systems*, 2006. (Cited on page 3)

[46] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 2012. (Cited on pages 3 and 15)

[47] Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. In *International Conference on Neural Information Processing Systems*, 2020. (Cited on pages 3, 4, 21, and 25)

[48] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International conference on machine learning*, 2015. (Cited on page 3)

[49] Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, and Andreas Krause. Transductive active learning: Theory and applications. In *International Conference on Neural Information Processing Systems*, 2024. (Cited on page 3)

[50] Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 2016. (Cited on page 3)

[51] Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Florian Dörfler, Pieter Abbeel, and Andreas Krause. Sombrl: Scalable and optimistic model-based rl. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. (Cited on pages 3, 4, 17, and 18)

[52] Manish Prajapat, Johannes Köhler, Melanie N. Zeilinger, and Andreas Krause. Safe and near-optimal control with online dynamics learning. 2026. (Cited on pages 3 and 7)

[53] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. (Cited on page 3)

[54] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *International Conference on Neural Information Processing Systems*, 2020. (Cited on pages 4, 18, and 23)

[55] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, 2017. (Cited on page 5)

[56] Aad Van Der Vaart and Harry Van Zanten. Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12(6), 2011. (Cited on page 5)

[57] Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *International Conference on Neural Information Processing Systems*, 2017. (Cited on page 7)

[58] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. (Cited on page 8)

[59] T. Beckers and S. Hirche. Prediction with approximated gaussian process dynamical models. *IEEE Transactions on Automatic Control*, 2021. (Cited on page 15)

[60] Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023. (Cited on page 16)

[61] Bach Do, Nafeezat A. Ajenifuja, Taiwo A. Adebiyi, and Ruda Zhang. Sampling from gaussian processes: a tutorial and applications in global sensitivity analysis and optimization. *Structural and Multidisciplinary Optimization*, 2025. (Cited on page 16)

[62] Bhavya Sukhija, Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic active exploration of dynamical systems. In *International Conference on Neural Information Processing Systems*, 2023. (Cited on pages 17, 23, and 25)

[63] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. 2006. (Cited on page 20)

[64] Cristina Pinneri, Shambhuraj Sawant, Sebastian Blaes, Jan Achterhold, Joerg Stueckler, Michal Rolinek, and Georg Martius. Sample-efficient cross-entropy method for real-time planning. In *Conference on Robot Learning*, 2020. (Cited on page 25)

[65] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: model-based policy optimization. In *International Conference on Neural Information Processing Systems*, 2019. (Cited on pages 27 and 29)

[66] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *International Conference on Neural Information Processing Systems*, 2017. (Cited on page 27)

[67] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, 2018. (Cited on page 27)

[68] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, second edition, 2006. (Cited on page 28)

[69] Jingqi Li, David Fridovich-Keil, Somayeh Sojoudi, and Claire J. Tomlin. Augmented lagrangian method for instantaneously constrained reinforcement learning problems. *2021 60th IEEE Conference on Decision and Control (CDC)*, 2021. (Cited on page 28)

[70] Yarden As, Chengrui Qu, Benjamin Unger, Dongho Kang, Max van der Hart, Laixi Shi, Stelian Coros, Adam Wierman, and Andreas Krause. SPiDR: A simple approach for zero-shot safety in sim-to-real transfer. In *International Conference on Neural Information Processing Systems*, 2025. (Cited on page 28)

[71] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, 2020. (Cited on page 28)

A Discussion

We collect here several technical observations that clarify the scope of our assumptions and design choices. We first discuss the assumptions on the prior kernel.

A.1 Kernels and Unbounded Domains

We first note that the regularity assumptions on the kernel required by Assumptions 4 and 5 are not restrictive, and can hold despite the fact that the state-action space is unbounded.

In particular, suppose $k(z, z') = \tilde{k}(\phi(z), \phi(z'))$, where $\phi\colon \mathcal{Z} \to \tilde{\mathcal{Z}} \subset \mathbb{R}^{d_{\tilde{z}}}$ is a bounded and Lipschitz feature map , and $\tilde{k}$ is Lipschitz on $\tilde{\mathcal{Z}}$ (which holds for common kernels such as linear, squared exponential, and Matérn since $\tilde{\mathcal{Z}}$ is bounded). Then k is Lipschitz on $\mathcal{Z}$ (thus satisfying Assumption 4), even if $\mathcal{Z}$ is unbounded. Moreover, since the information gain of k on the inputs $\{z_i\}$ is equal to the information gain of $\tilde{k}$ on the transformed inputs $\{\phi(z_i)\}$, the maximum information gain $\gamma_N(k)$ over $\mathcal{Z}$ is upper bounded by the maximum information gain of $\gamma_N(\tilde{k})$ over the bounded domain $\tilde{Z}$.

Therefore, for common choices of $\tilde{k}$, such as the squared exponential or linear kernel, $\gamma_N(k)$ grows poly-logarithmically in N [46], and Assumption 5 is also satisfied. For Matérn kernels, the information gain grows polynomially in N with growth determined by a smoothness parameter, and satisfies Assumption 5 for sufficiently large choices of the smoothness parameter.

We next provide an overview of possible methods to sample a function realization from a GP in practice, followed by a discussion on alternative design choices for model updates.

A.2 Sampling from a GP

In practice, the sampling of continuous functions $\tilde{f} \sim GP(\mu, k)$ cannot be implemented directly due to the infinite-dimensional future space of GPs.

However, we are only interested in the function values at the locations in the state-action space reached by the agent across all episodes, i.e. $\{z_t^n\}_{t=0,\dots,T-1}^{n\geq 0} \subset \mathcal{Z}$. Therefore, for each scalar sample $\tilde{f}_j$, which represents a particular deterministic realization in function space, we are technically interested in its finite-dimensional marginals: $(\tilde{f}_j(z_t^n))_{t,n} \sim \mathcal{N}([\mu(z_t^n)]_{t,n}, [k(z_t^n, z_{t'}^{n'})]_{t,n,t',n'})$.

However, in a sequential control setting, the locations at later timesteps and episodes are not known in advance, preventing us from constructing this joint distribution upfront. Crucially, Property 1 in Beckers and Hirche [59] shows that this is equivalent to iteratively sampling each function value by conditioning the GP on the sample's own predictions at all previously visited locations, treating them as noise-free training points. This method, known as sequential pathwise conditioning or forward sampling, ensures the spatial consistency of the sampled function across episodes. This has been effectively applied in both scenario-based control [42] and MPC settings [37, 44].

For completeness, other methods can be used to draw approximate samples from a GP [60, 61]. For instance, one could construct a spectral function approximation of the GP (e.g., via Random Fourier Features) and draw a finite-dimensional weight vector from the prior. Although this yields an approximation rather than an exact draw, it defines a global function that can be trivially evaluated at any desired location, making it highly popular in practice due to its computational efficiency.

A.3 Updating the GP Samples

As explained in Section 5, we draw a finite set of samples from the prior GP, purposefully not conditioning the underlying function realization on the true data gathered during exploration. To embed the newly collected information, we instead only truncate our prior samples to lie within the confidence region established by Lemma 1. This is conceptually similar to Prajapat et al. [37], who enforce truncation by directly discarding all samples violating the confidence bounds.

A natural alternative to incorporate newly collected data is to occasionally discard the prior samples and draw new ones from the posterior GP. Since the posterior generally provides a tighter approximation of the true dynamics than the prior, resampling could potentially reduce the number of samples M required to obtain a ζ-close model via Lemma 2. However, repeated resampling introduces a more complex analytical structure, as the sampled models would no longer be coupled across episodes. Proving safety would require applying Lemma 2 at each resampling step and taking a union bound over all episodes to control the overall success probability. Remarkably, this still yields a valid high-probability safety guarantee, since the total number of episodes is bounded by Theorem 2, which still holds. We note that this alternative (sampling directly from the posterior GP) was employed in our practical GP experiments (detailed in Appendix F). We adopted this for empirical convenience and to ensure a fair, direct comparison with our baselines, performing our validation in the exact same setting as As et al. [8].

Finally, we remark that the episodic truncation is only necessary for our optimality result (specifically the second claim of Theorem 2); to guarantee safety, we merely need to satisfy the lower-bound on the number of samples in Lemma 2 (provided Assumption 3 holds, for which is sufficient to perform a truncation with respect to the prior bound B).

Finally, we clarify the feasibility properties of Algorithm 1.

A.4 Feasibility of the Safety Constraint

One may wonder whether the algorithm could terminate because the safety constraint becomes infeasible at some earlier episode, rather than due to the exploration constraint. Remarkably, under the success event of Lemma 1, Assumption 3 ensures that the sampled safe set Π^n_{safe} is non-empty for every episode $n \geq 0$. Consequently, under this event, infeasibility can only occur due to the exploration constraint.

More specifically, conditioning on the success event of Lemma 1, the truncation procedure is well-defined and the sequence of confidence sets containing the truncated samples is nested (i.e. non-increasing). Combined with the truncation defined for the initial episode $n = 0$, this implies that for all $n \geq 0$ and all $m \in [M]$ the samples f_n^m lie in the intersection of the successive confidence sets, and in particular remain within the initial truncation envelope $\mathcal{F}_0^B$. Thus, by Assumption 3 and the fact that Δ_ζ is chosen to satisfy $\Delta_\zeta \leq \Delta$, any policy $\pi \in \Pi_{\text{prior}}$ must also satisfy $J_c(\pi, f_n^m) \leq d - \Delta < d - \Delta_\zeta$ for $m = 1, \ldots, M$. This means that $\Pi_{\text{prior}} \subseteq \Pi^n_{\text{safe}}$ and thus Π^n_{safe} is non-empty for all $n \geq 0$ by Assumption 3.

Therefore, the safety constraint remains feasible throughout learning, and termination can only occur due to infeasibility of the exploration constraint.

B Technical Lemmas

In this section, we state and prove some auxiliary lemmas that will be useful in the proofs of our main results. We start with the proof of Lemma 2, which was already stated in Section 4.

Lemma 2 (Theorem 1 of Prajapat et al. [44])**.** *Let Assumption 2 hold. Consider any $\delta \in (0, 1]$ and $\zeta > 0$. Select M satisfying*

$$M \geq \frac{\log(\delta)}{\log(1 - \exp(-d_x(\frac{1}{2}B^2 + \phi(\zeta))))}.$$

Let $\tilde{f}^1, \ldots, \tilde{f}^M \sim GP(\mu, k)$ be function samples from the prior GP. It holds with probability at least $1-\delta$ that for some $m \in [M]$, $\left\| \tilde{f}_j^m - f_j^\star \right\|_\infty \leq \zeta \;\; \forall j \in [d_x]$.

Proof. Lemma 2 is a direct adaptation of Theorem 1 from [44] to our specific setting. To bridge the two results, we apply a centering transformation, defining the shifted quantities $\bar{f}^m := \tilde{f}^m - \mu$ and $\bar{f}^\star := f^\star - \mu$. Because $\tilde{f}^m \sim GP(\mu, k)$, the centered sample follows a zero-mean Gaussian process, $\bar{f}^m \sim GP(0, k)$. Similarly, under Assumption 2, the centered ground truth satisfies $\|\bar{f}_j^\star\|_k \leq B$ for all $j \in [d_x]$. Thus, the conditions for Theorem 1 of [44] are met, and establishing $\|\bar{f}_j^m - \bar{f}_j^\star\|_\infty \leq \zeta \;\; \forall j \in [d_x]$ directly implies $\|\tilde{f}_j^m - f_j^\star\|_\infty \leq \zeta \;\; \forall j \in [d_x]$.

Our formulation simplifies the original result in two key ways. First, by applying the theorem to the prior GP directly rather than on the posterior conditioned on data, we eliminate the reliance on data-dependent quantities, effectively simplifying their confidence parameter to $C_D = B^2/2$. Second, because Assumption 2 imposes a deterministic bound B on the RKHS norm of the mismatch, we bypass the need for the probabilistic union bound used in the original paper, allowing our sample complexity to scale with $\log(\delta)$ rather than $\log(\delta/2)$. Finally, we employ the vector-valued extension discussed in Remark 1 of [44]. To ensure the ∞-norm bound holds across all d_x independent state coordinates simultaneously, the marginal probabilities must be multiplied. This introduces the factor of d_x that multiplies the exponent $B^2/2 + \phi(\zeta)$ in Lemma 2. □

We next present a lemma that generalizes the notation of Lemma B.3 in [51], Lemma A.6 of [8] or Corollary 3 of [62]. Together with the subsequent corollary, it provides bounds on the discrepancy between returns under two different dynamics models, which are used in the proofs of safety and optimality.

Lemma 3. *Let Assumptions 1-2 hold. Consider any positive and bounded function $g \in [0, G_{\max}]$. Let f and f' be any pair of dynamics functions. Then, for all $n \geq 0$*

$$\left|J_g(\pi, f) - J_g(\pi, f')\right| \leq T G_{\max} \mathbb{E}_{\tau_\pi^f}\left[\sum_{t=0}^{T-1} \frac{\|f'(x_t, \pi(x_t)) - f(x_t, \pi(x_t))\|}{\sigma_w}\right]$$

where the expectation is rolled out over f, i.e. $x_{t+1} = f(x_t, \pi(x_t)) + w_t$. An analogous bound holds in terms of an expectation over f' with $x'_{t+1} = f'(x'_t, \pi(x'_t)) + w_t$:

$$\left|J_g(\pi, f) - J_g(\pi, f')\right| \leq T G_{\max} \mathbb{E}_{\tau_\pi^{f'}}\left[\sum_{t=0}^{T-1} \frac{\|f'(x'_t, \pi(x'_t)) - f(x'_t, \pi(x'_t))\|}{\sigma_w}\right].$$

Proof. We give proof for the first inequality. The same argument holds for the second inequality by swapping f and f'. Let $J_{g,t+1}(\pi, f, x)$ denote the cost-to-go from state x at step $t+1$ under some dynamics f and policy π:

$$J_{g,t+1}(\pi, f, x) = \mathbb{E}_{\tau_\pi^f}\left[\sum_{k=t+1}^{T-1} g(x_k, \pi(x_k))\right]$$

We can directly apply Corollary 2 of Sukhija et al. [62] to obtain

$$J_g(\pi, f') - J_g(\pi, f) = \mathbb{E}_{\tau_\pi^f}\left[\sum_{t=0}^{T-1} \Big(J_{g,t+1}(\pi, f', \tilde{x}_{t+1}) - J_{g,t+1}(\pi, f', x_{t+1})\Big)\right]$$

where the expectation is w.r.t. π under the dynamics f, and

$$\begin{aligned} x_{t+1} &= f(x_t, \pi(x_t)) + w_t, \\ \tilde{x}_{t+1} &= f'(x_t, \pi(x_t)) + w_t. \end{aligned}$$

Therefore,

$$\begin{aligned} |J_g(\pi, f') - J_g(\pi, f)| &= |\mathbb{E}_{\tau_\pi^f}\left[\sum_{t=0}^{T-1} \Big(J_{g,t+1}(\pi, f', \tilde{x}_{t+1}) - J_{g,t+1}(\pi, f', x_{t+1})\Big)\right]| \\ &\leq \sum_{t=0}^{T-1} \mathbb{E}_{\tau_\pi^f}\left[\left|\mathbb{E}_{w_t}\left[J_{g,t+1}(\pi, f', \tilde{x}_{t+1}) - J_{g,t+1}(\pi, f', x_{t+1}) \mid x_t\right]\right|\right] \end{aligned}$$

Crucially, since $w_t \sim \mathcal{N}(0, \sigma_w^2)$ (Assumption 1), we have that x_{t+1} and $\tilde{x}_{t+1}$ are also gaussian given x_t, so that we can leverage Lemma C.2 of [54]. Furthermore, $J_{g,t+1}(\pi, f, x) \in [0, TG_{\max}]$ for all π, f, x, and t. Therefore, given x_t,

$$
\begin{aligned}
&\left|\mathbb{E}_{w_t}\left[J_{g,t+1}(\pi, f', \tilde{x}_{t+1}) - J_{g,t+1}(\pi, f', x_{t+1}) \mid x_t\right]\right| \\
&\leq \max\left\{\sqrt{\mathbb{E}_{w_t}[J^2_{g,t+1}(\pi, f', \tilde{x}_{t+1}) \mid x_t]},\ \sqrt{\mathbb{E}_{w_t}[J^2_{g,t+1}(\pi, f', x_{t+1}) \mid x_t]}\right\} \\
&\times \min\left\{\frac{\|f'(x_t, \pi(x_t)) - f(x_t, \pi(x_t))\|}{\sigma_w}, 1\right\} \text{ ([54], Lemma C.2)} \\
&\leq TG_{\max} \min\left\{\frac{\|f'(x_t, \pi(x_t)) - f(x_t, \pi(x_t))\|}{\sigma_w},\ 1\right\}.
\end{aligned}
$$

Summing over t concludes the proof. □

Finally, in the case when one of the dynamics functions represents exactly the mean dynamics, i.e. $f' = \mu_n$, we can continue from the above lemma to give a bound in terms of the predicted uncertainty. This is formalized in the following corollary. We first define $\mathcal{F}_n$ to be the filtration that contains all the data observed through episode n (i.e. $\mathcal{D}_{0:n}$) and the the realizations of the samples obtained at the initial episode (i.e. $\tilde{f}^m\ \forall m \in [M]$). Then, overloading the definition of the return in Section 3, we can define

$$
J_{s_n}(\pi, f) \coloneqq \mathbb{E}_{\tau_\pi^f}\left[\sum_{t=0}^{T-1} s_n(x_t, \pi(x_t)) \mid \mathcal{F}_{n-1}\right],
$$

where we note that the random variable s_n is $\mathcal{F}_{n-1}$ measurable.

Corollary 1 (Adapted from Lemma B.3 of [51] or Lemma A.6 of [8])**.** *Let Assumptions 1-2 hold. Consider any positive and bounded function $g \in [0, G_{\max}]$ and fix any policy $\pi \in \Pi$. Consider the GP fit after episode n, characterized by (μ_n, σ_n). Let f be any dynamics function satisfying for all $x \in \mathcal{X}$ and all $j \in [d_x]$:*

$$
|\mu_{n,j}(x, \pi(x)) - f_j(x, \pi(x))| \leq \beta_n(\delta)\sigma_{n,j}(x, \pi(x)).
$$

where β_n is the one defined in Lemma 1. Then,

$$
\left|J_g(\pi, f) - J_g(\pi, \mu_n)\right| \leq \lambda_n^g \min\left\{J_{s_n}(\pi, f), J_{s_n}(\pi, \mu_n)\right\}
$$

where $\lambda_n^g \coloneqq G_{\max} T \frac{\beta_n(\delta)}{\sigma_w}$.

Proof. We give the proof for the first inequality, while the second inequality follows analogously after swapping f and μ_n.
By applying Lemma 3, we obtain

$$
\left|J_g(\pi, f) - J_g(\pi, \mu_n)\right| \leq TG_{\max}\, \mathbb{E}_{\tau_\pi^f}\left[\sum_{t=0}^{T-1} \frac{\|f(x_t, \pi(x_t)) - \mu_n(x_t, \pi(x_t))\|}{\sigma_w}\right]
$$

Then, by our assumption on f lying in the confidence interval defined by $(\mu_n, \sigma_n, \beta_n)$, we can bound:

$$
\|f(x_t, \pi(x_t)) - \mu_n(x_t, \pi(x_t))\| \leq \beta_n(\delta)\|\sigma_n(x_t, \pi(x_t))\|
$$

Combining this with the above bound finally proves the claim:

$$
\begin{aligned}
\left|J_g(\pi, f) - J_g(\pi, \mu_n)\right| &\leq TG_{\max}\, \mathbb{E}_{\tau_\pi^f}\left[\sum_{t=0}^{T-1} \frac{\|f(x_t, \pi(x_t)) - \mu_n(x_t, \pi(x_t))\|}{\sigma_w}\right] \\
&\leq G_{\max} T \frac{\beta_n(\delta)}{\sigma_w} \mathbb{E}_{\tau_\pi^f}\left[\sum_{t=0}^{T-1} s_n(x_t, \pi(x_t)) \mid \mathcal{F}_{n-1}\right] =: \lambda_n^g\, J_{s_n}(\pi, f).
\end{aligned}
$$

□

C Proof of Safety

Theorem 1 (Safety)**.** *Suppose Assumptions 1-3 hold. Consider any $\delta \in (0, 1/2)$, $\zeta \in (0, \frac{\sigma_w \Delta}{\sqrt{d_x} T^2 C_{\max}})$, and sequence of positive numbers $\{d_\sigma^n\}_{n \geq 0}$. Consider running Algorithm 1 given these values. Then, it holds with probability at least $1 - 2\delta$ that for every episode n, any policy $\pi \in \Pi_{safe}^n$ satisfies $J_c(\pi, f^\star) \leq d$. In particular, this holds for the policy π_n selected during episode n and the policy $\hat{\pi}$ selected upon algorithm termination.*

Proof. Lemma 3 allows to bound the difference in expected cost between the trajectory sampled from the true dynamics $f^\star$ and the one sampled from a truncated sample f_n^m, for any $m \in [M]$, obtained according to the *same policy π and the same initial distribution* x_0:

$$\left| J_c(\pi, f_n^m) - J_c(\pi, f^\star) \right| \leq T C_{\max} \, \mathbb{E}_{\tau_\pi^f} \left[\sum_{t=0}^{T-1} \frac{\| f^\star(x_t, \pi(x_t)) - f_n^m(x_t, \pi(x_t)) \|}{\sigma_w} \right] \tag{10}$$

Then, as described in Section 5, the number of samples M in Algorithm 1 is chosen large enough to ensure that with probability $1 - \delta$ there exists a sample $f_0^{m_\zeta}$ from the GP prior that is $\zeta-$ close to $f^\star$. Furthermore, under the success event of Lemma 1, which holds with probability $1 - \delta$, the truncation that we perform at every episode can only bring the samples closer to $f^\star$, preserving the existence of a ζ-close dynamics sample. Thus, by taking a union bound over the above events, we can state that with probability at least $1 - 2\delta$, there exists a ζ-close truncated sample $f_n^{m_\zeta}$ at any episode n:

$$\left\| f^\star(z) - f_n^{m_\zeta}(z) \right\| < \zeta \sqrt{d_x} \quad \forall z \in \mathcal{Z}$$

where the factor $\sqrt{d_x}$ arises from bounding the ℓ_2 norm above in terms of the ℓ_∞ norm of Lemma 2. Thus, conditioned on the existence of the ζ-close sample, we can substitute it in (10) to bound the difference in expected cost between the trajectory sampled from the true dynamics $f^\star$ and the ζ-close sample $f_n^{m_\zeta}$ under the same policy π and the same initial distribution x_0:

$$\begin{aligned} \left| J_c(\pi, f_n^{m_\zeta}) - J_c(\pi, f^\star) \right| &\leq T C_{\max} \, \mathbb{E}_{\tau_\pi^f} \left[\sum_{t=0}^{T-1} \frac{\| f^\star(x_t, \pi(x_t)) - f_n^{m_\zeta}(x_t, \pi(x_t)) \|}{\sigma_w} \right] \\ &\leq \zeta \sqrt{d_x} \frac{T^2 C_{\max}}{\sigma_w} =: \Delta_\zeta. \end{aligned}$$

This upper bound matches exactly the definition of the tightening Δ_ζ that we use when enforcing the constraints in Problem (8).

We are finally ready to show the desired safety bound. By definition, any policy $\pi \in \Pi_{\text{safe}}^n$ satisfies the tightened constraints for all truncated samples, i.e. $J_c(\pi, f_n^m) \leq d - \Delta_\zeta, \quad \forall m \in M$. However, as shown above, with probability at least $1 - 2\delta$ there exists a ζ-close sample $f_n^{m_\zeta}$ that satisfies $\left| J_c\left(\pi, f^\star\right) - J_c\left(\pi, f_n^{m_\zeta}\right) \right| \leq \Delta_\zeta$. Rearranging, we obtain

$$J_c\left(\pi, f^\star\right) \leq J_c\left(\pi, f_n^{m_\zeta}\right) + \Delta_\zeta \leq d - \Delta_\zeta + \Delta_\zeta = d, \tag{11}$$

which proves the claim.

□

In the following, we provide the proof of Theorem 2. For the sake of exposition, we separate the two claims of the theorem, proving the first claim (about finite-time termination of Algorithm 1) in Appendix D, and the second claim (about optimality of the policy returned by Algorithm 1) in Appendix E.

D Proof of Sample Complexity

Lemma 4. *Consider running Algorithm 1 in the setting of Theorem 2, until some finite episode $\bar{n}$. Then, with probability at least $1 - \delta$, we have:*

$$\sum_{n=0}^{\bar{n}-1} \mathbb{E}_{\tau_{\pi_n}^{f^\star}} \left[\sum_{t=0}^{T-1} s_n(x_t, \pi_n(x_t)) \mid \mathcal{F}_{n-1} \right] \leq \sum_{n=0}^{\bar{n}-1} \sum_{t=0}^{T-1} s_n(x_t^n, \pi_n(x_t^n)) + T \sqrt{2 \bar{n} d_x \sigma_{\max} \log(1/\delta)}, \tag{12}$$

where, for each episode $n \in \{0, \dots, \bar{n}-1\}$, the states x_t^n on the right-hand side denote the trajectory realized by the agent in the environment, while following the policy π_n under the true dynamics $f^\star$.

Proof. Let us define $Y_n \coloneqq \sum_{t=0}^{T-1} s_n(x_t^n, \pi_n(x_t^n))$, so that:

$$\mathbb{E}_{\tau_{\pi_n}^{f^\star}}\left[\sum_{t=0}^{T-1} s_n(x_t, \pi_n(x_t)) \mid \mathcal{F}_{n-1}\right] = \mathbb{E}_{\tau_{\pi_n}^{f^\star}}\left[Y_n \mid \mathcal{F}_{n-1}\right]. \tag{13}$$

Next, define

$$X_n \coloneqq \mathbb{E}_{\tau_{\pi_n}^{f^\star}}\left[Y_n \mid \mathcal{F}_{n-1}\right] - Y_n \tag{14}$$

Then, X_n is a *martingale difference sequence*, because:

1. $\mathbb{E}_{\tau_{\pi_n}^{f^\star}}\left[X_n \mid \mathcal{F}_{n-1}\right] = 0.$
2. $\mid X_n \mid \leq T\sqrt{d_x \sigma_{\max}}$ (because $Y_n \in \left[0,\, T\sqrt{d_x \sigma_{\max}}\right]$).

Therefore, we can apply Lemma A.7 of Cesa-Bianchi and Lugosi [63] [2](substituting $A_i \coloneqq -T\sqrt{d_x \sigma_{\max}}$ and $c_i \coloneqq 2T\sqrt{d_x \sigma_{\max}}$ in their notation) to get, with probability at least $1-\delta$:

$$\sum_{n=0}^{\bar{n}-1} X_n \leq T\sqrt{2\bar{n} d_x \sigma_{\max} \log(1/\delta)} \tag{15}$$

The claim follows by rearranging and using the definition of X_n, Y_n, and Equation (13). □

Next, we restate the first claim of Theorem 2 as follows.

Theorem 3 (Sample Complexity)**.** *Suppose Assumptions 1-5 hold. Fix some $\delta \in (0, 1/2)$, $\zeta \in (0, \frac{\sigma_w \Delta}{\sqrt{d_x} T^2 C_{\max}})$, and $\varepsilon > 0$. Let $d_\sigma^n = \frac{\varepsilon \sigma_w}{2 G_{max} T \beta_n(\delta)}$ with $G_{max} \coloneqq \max\{C_{max}, R_{max}\}$, and consider running Algorithm 1. There exists a problem-dependent constant $C = C(d_x, \sigma_{\max}, G_{\max}, T, \delta)$ such that, letting $\bar{n}$ be the smallest integer satisfying*

$$\bar{n} \;\geq\; C\,\frac{\gamma_{\bar{n}T}(k)^3}{\varepsilon^2}, \tag{16}$$

then, with probability at least $1-2\delta$ there exists $n^\star \leq \bar{n}$ such that $J_{s_{n^\star}}(\pi, \mu_{n^\star}) < d_\sigma^{n^\star}\ \forall \pi \in \Pi_{safe}^{n^\star}$, i.e. Algorithm 1 terminates at iteration $n^\star$.

Proof. We start by conditioning on the success event of Lemma 1, i.e. we assume well-calibration of the true dynamics $f^\star$. Under this event, that holds with probability $1-\delta$, and further noting that $s_n(\cdot) = \|\sigma_n(\cdot, \pi(\cdot))\|$ is bounded by $\sqrt{d_x \sigma_{\max}}$ thanks to Assumption 2, we can apply Corollary 1 to the scalar function $s_n(\cdot)$ and dynamics $f^\star$, for any $n \geq 0$. Thus, we can write for J_{s_n}:

$$\mid J_{s_n}(\pi_n, \mu_n) - J_{s_n}(\pi_n, f^\star) \mid \leq \lambda_n^s J_{s_n}(\pi_n, f^\star),$$

where $\lambda_n^s = \sqrt{d_x \sigma_{\max}} T \frac{\beta_n(\delta)}{\sigma_w}$. Rearranging the terms:

$$J_{s_n}(\pi_n, \mu_n) \leq (1 + \lambda_n^s)\, J_{s_n}(\pi_n, f^\star).$$

Notice that by our uncertainty constraint we know that before termination $J_{s_n}(\pi_n, \mu_n) \geq d_\sigma^n$. Furthermore, using the monotonicity of $\beta_n(\delta)$ we can bound $d_\sigma^n \geq d_\sigma^{\bar{n}}$ for all $n \leq \bar{n}$. By substituting

[2]Note that a variance-sensitive inequality (e.g. Freedman or Bernstein inequalities for martingales) would yield a sharper bound, but Azuma-Hoeffding suffices for our purposes and yields a cleaner proof.

this and taking the sum over the episodes from 1 to $\bar{n}$:

$$
\begin{aligned}
\bar{n} d_\sigma^{\bar{n}} \leq \sum_{n=0}^{\bar{n}-1} d_\sigma^n &\leq \sum_{n=0}^{\bar{n}-1} J_{s_n}(\pi_n, \mu_n) \\
&\leq \sum_{n=0}^{\bar{n}-1} (1+\lambda_n^s)\, J_{s_n}(\pi_n, f^\star) \\
&= \sum_{n=0}^{\bar{n}-1} (1+\lambda_n^s)\, \mathbb{E}_{\tau_{\pi_n}^{f^\star}} \left[\sum_{t=0}^{T-1} s_n(x_t, \pi_n(x_t)) \mid \mathcal{F}_{n-1} \right] \\
&\leq (1+\lambda_{\bar{n}}^s) \sum_{n=0}^{\bar{n}-1} \mathbb{E}_{\tau_{\pi_n}^{f^\star}} \left[\sum_{t=0}^{T-1} s_n(x_t, \pi_n(x_t)) \mid \mathcal{F}_{n-1} \right] \\
&\leq (1+\lambda_{\bar{n}}^s) \left(\sum_{n=0}^{\bar{n}-1} \sum_{t=0}^{T-1} s_n(x_t^n, \pi_n(x_t^n)) + T\sqrt{2\bar{n} d_x \sigma_{\max} \log(1/\delta)} \right) \quad \text{(Lemma 4)}, \\
&\leq (1+\lambda_{\bar{n}}^s) \left(\sqrt{\bar{n}T} \sqrt{\sum_{n=0}^{\bar{n}-1} \sum_{t=0}^{T-1} s_n(x_t^n, \pi_n(x_t^n))^2} + T\sqrt{2\bar{n} d_x \sigma_{\max} \log(1/\delta)} \right) \\
&\qquad \text{(Cauchy-Schwarz)}, \\
&\leq (1+\lambda_{\bar{n}}^s) \left(\sqrt{\bar{n}T} \sqrt{\frac{T d_x \sigma_{\max}}{\log(1+\sigma_w^{-2}\sigma_{\max})} \gamma_{\bar{n}T}} + T\sqrt{2\bar{n} d_x \sigma_{\max} \log(1/\delta)} \right) \\
&\qquad \text{(Curi et al. [47], Lemma 17)},
\end{aligned}
$$

where we used that $\lambda_n^s \leq \lambda_{\bar{n}}^s$ due to the monotonicity of $\beta_n(\delta)$. Note that the success event of Lemma 4 holds with probability probability at least $1-\delta$, so that taking a union bound with the success event of Lemma 1 we get that our final guarantee holds with probability at least $1-2\delta$. Thus, dividing both sides for $\sqrt{\bar{n}}$ and taking the square, we obtain:

$$
\begin{aligned}
\bar{n} &\leq \frac{T^2 (1+\lambda_{\bar{n}}^s)^2}{(d_\sigma^{\bar{n}})^2} \left(\sqrt{\frac{d_x \sigma_{\max}}{\log(1+\sigma_w^{-2}\sigma_{\max})} \gamma_{\bar{n}T}} + \sqrt{2 d_x \sigma_{\max} \log(1/\delta)} \right)^2 \\
&\leq \frac{4T^4 (1+\lambda_{\bar{n}}^s)^2 G_{\max}^2}{\varepsilon^2 \sigma_w^2} \beta_{\bar{n}}^2(\delta) \left(\sqrt{\frac{d_x \sigma_{\max}}{\log(1+\sigma_w^{-2}\sigma_{\max})} \gamma_{\bar{n}T}} + \sqrt{2 d_x \sigma_{\max} \log(1/\delta)} \right)^2 .
\end{aligned}
$$

Therefore, if there exists an integer $\bar{n}$ that violates the above inequality, the algorithm must have terminated in one of the previous rounds, which proves the claim. To show that such an integer exists, note that the calibration coefficient $\beta_n(\delta)$ is monotonically increasing in n. Therefore, a finite integer $\bar{n}$ violating the above inequality (i.e. satisfying Equation 16) can be found if $\beta_n(\delta)^4 \gamma_{nT}(k)$ grows sub-linearly with n (since asymptotically we can omit the low order terms in γ_{nTd_x}). By the definition of $\beta_n(\delta)$ from Lemma 1, the quantity $\beta_n^4(\delta)\gamma_{nT}(k)$ is of the same order as $\gamma_{nT}^3(k)$ for large n. Therefore, the existence of a finite $\bar{n}$ satisfying Equation (16) is guaranteed by Assumption 5. □

E Proof of Optimality

Lemma 5. *Let Assumptions 1-2 hold. Then, with probability at least $1-\delta$ we have that for all $n \geq 0$ and any truncated sample f_n^m:*

$$
|J_c(\pi, f^\star) - J_c(\pi, f_n^m)| \leq 2\lambda_n^c J_{s_n}(\pi, \mu_n). \tag{17}
$$

Moreover, the same holds for the reward function, i.e. replacing c and λ_n^c with r and λ_n^r respectively.

Proof. Under the success event of Lemma 1, which holds with probability $1-\delta$, and thanks to the truncation procedure described in Section 5, we can apply Corollary 1 to bound the distance between the costs attained along the mean dynamics and the ones attained along the trajectory rolled out using

any of the samples f_n^m, in terms of the uncertainty predicted along the mean dynamics. Thus, by Corollary 1 we immediately get

$$|J_c(\pi, f_n^m) - J_c(\pi, \mu_n)| \leq \lambda_n^c J_{s_n}(\pi, \mu_n). \tag{18}$$

We can then repeat the same argument for the true dynamics, i.e. swapping f_n^m with $f^\star$:

$$|J_c(\pi, f^\star) - J_c(\pi, \mu_n)| \leq \lambda_n^c J_{s_n}(\pi, \mu_n). \tag{19}$$

Finally, the claim follows by combining the two bounds above. □

We are finally ready to prove the second claim of Theorem 2, which can be restated as follows.

Theorem 4. *Suppose Assumptions 1-5 hold. Fix some $\delta \in (0, 1/2)$, $\zeta \in (0, \frac{\sigma_w \Delta}{\sqrt{d_x} T^2 C_{\max}})$, $\varepsilon > 0$ and $\xi > \varepsilon + \zeta\sqrt{d_x}\frac{T^2 C_{\max}}{\sigma_w}$. Let $d_\sigma^n = \frac{\varepsilon \sigma_w}{2G_{max} T \beta_n(\delta)}$ with $G_{max} := \max\{C_{max}, R_{max}\}$, and consider running Algorithm 1. Then, with probability at least $1 - 2\delta$, the policy $\hat{\pi}$ returned by Algorithm 1 satisfies*

$$J_r(\tilde{\pi}^\star, f^\star) - J_r(\hat{\pi}, f^\star) \leq \varepsilon, \tag{20}$$

where $\tilde{\pi}^\star$ is an approximation of the optimal policy defined by the following optimization problem:

$$\tilde{\pi}^\star = \arg\max_{\pi \in \Pi_\xi^{\star,c}} J_r(\pi, f^\star). \tag{21}$$

Proof. Let $n^\star$ denote the termination episode of Algorithm 1, which is guaranteed to be finite by Theorem 3. First, Lemma 5 allows us to calculate the difference between the expected cost along the real trajectory and the one generated from any truncated sample, in terms of the expected predicted uncertainty along the mean dynamics. Note that the bound holds jointly for all $n \geq 0$ with probability at least $1 - \delta$. Thus, conditioning on this event, we can write that for any policy $\pi \in \Pi$ at episode $n^\star$:

$$|J_c(\pi, f^\star) - J_c(\pi, f_{n^\star}^m)| \leq 2\lambda_{n^\star}^c J_{s_{n^\star}}(\pi, \mu_{n^\star}). \tag{22}$$

Note that the bound above holds for any policy $\pi \in \Pi$. Then, if the policy also belongs to the set of safe policies (i.e. $\pi \in \Pi_{\text{safe}}^{n^\star}$) we have that it must be $J_{s_{n^\star}}(\pi, \mu_{n^\star}) < d_\sigma^{n^\star}$, since we assume the algorithm has terminated due to the uncertainty constraint. However, recalling the definition of $d_{\sigma,n}$ and λ_n^c, we have that $d_\sigma^{n^\star} = \frac{\varepsilon \sigma_w}{2G_{\max} T \beta_{n^\star}(\delta)}$ and $\lambda_{n^\star}^c = C_{\max} T \frac{\beta_{n^\star}(\delta)}{\sigma_w}$. Thus, we can substitute these in (22) to get that for any policy $\pi \in \Pi_{\text{safe}}^{n^\star}$ and any truncated dynamics sample $f_{n^\star}^m$,

$$|J_c(\pi, f^\star) - J_c(\pi, f_{n^\star}^m)| \leq \varepsilon. \tag{23}$$

In the following, we will also use that for any policy $\pi \in \Pi_{\text{safe}}^{n^\star}$ we have,

$$|J_r(\pi, f^\star) - J_r(\pi, \mu_{n^\star})| \leq \frac{\varepsilon}{2}, \tag{24}$$

which can be obtained by repeating the same argument we used to obtain Equation 23, but applied directly to Equation 19 of Lemma 5, after substituting c with r.

We highlight that in general Equations 23-24 only hold for policies in $\Pi_{\text{safe}}^{n^\star}$, since Problem (8) being infeasible only implies that the uncertainty is lower than $d_\sigma^{n^\star}$ on the policies that satisfy the first constraint (i.e. that belong to $\Pi_{\text{safe}}^{n^\star}$). Indeed, upon termination there can still exist some $\pi \notin \Pi_{\text{safe}}^{n^\star}$ that satisfies the second constraint $J_{s_{n^\star}}(\pi, \mu_{n^\star}) \geq d_\sigma^{n^\star}$, so that we cannot conclude that (23) holds for these policies.

However, Lemma 7 shows that upon termination of the algorithm it holds $\Pi_\xi^{\star,c} \subseteq \Pi_{\text{safe}}^{n^\star}$. Thus, by the definition of $\tilde{\pi}^\star \in \Pi_\xi^{\star,c}$ and Lemma 7, we get that $\tilde{\pi}^\star \in \Pi_{\text{safe}}^{n^\star}$.

This implies that $\tilde{\pi}^\star$ is also a feasible solution of the problem obtained after removing the exploration constrain in Problem (8), and hence it achieves a lower reward than $\hat{\pi}$:

$$J_r(\tilde{\pi}^\star, \mu_{n^\star}) \leq J_r(\hat{\pi}, \mu_{n^\star}). \tag{25}$$

This allows to prove the following optimality bound that holds for the trajectory at the termination episode $n^\star$:

$$\begin{aligned} J_r\left(\tilde{\pi}^\star, f^\star\right) - J_r\left(\hat{\pi}, f^\star\right) &= J_r\left(\tilde{\pi}^\star, f^\star\right) - J_r\left(\tilde{\pi}^\star, \mu_{n^\star}\right) \\ &\quad + J_r\left(\tilde{\pi}^\star, \mu_{n^\star}\right) - J_r\left(\hat{\pi}, \mu_{n^\star}\right) \\ &\quad + J_r\left(\hat{\pi}, \mu_{n^\star}\right) - J_r\left(\hat{\pi}, f^\star\right) \\ &\leq \varepsilon, \end{aligned}$$

where we have used the bounds $|J_r\left(\tilde{\pi}^\star, f^\star\right) - J_r\left(\tilde{\pi}^\star, \mu_{n^\star}\right)| \leq \frac{\varepsilon}{2}$ and $|J_r\left(\hat{\pi}, \mu_{n^\star}\right) - J_r\left(\hat{\pi}, f^\star\right)| \leq \frac{\varepsilon}{2}$, which follow from Equation 24 combined with $\tilde{\pi}^\star, \hat{\pi} \in \Pi^{n^\star}_{\text{safe}}$. □

The above proof, together with that of Theorem 3 in Appendix D concludes the proof of Theorem 2. In the remainder of this section, we provide the lemmas used in the optimality proof.

Lemma 6 (adapted from Lemma A.3 of [8]). *Let Assumptions 1-4 hold, and consider*

$$\begin{aligned} x_{t+1} &= \mu_n(x_t, \pi(x_t)) + w_t \\ x'_{t+1} &= \mu_n(x'_t, \pi'(x'_t)) + w_t \end{aligned}$$

for two policies π, π' s.t $\|\pi' - \pi\|_\infty \leq \varepsilon_\pi$, assuming the same initial state distribution. Then,

$$|J_{s_n}(\pi', \mu_n) - J_{s_n}(\pi, \mu_n)| \leq \sqrt{\varepsilon_\pi} C_\pi$$

where $C^n_\pi = T\left(L_s + TL_{\mu_n}\frac{\sqrt{d_x \sigma_{\max} \varepsilon_\pi}}{\sigma_w}\right)$.

Proof. From Lemma 5 of [62] (Difference in Policy Performance) applied on the μ_n dynamics (instead of $f^\star$ as in the original notation) we have

$$J_{s_n}(\pi', \mu_n) - J_{s_n}(\pi, \mu_n) = \mathbb{E}_{\tau^{\mu_n}_{\pi'}}\left[\sum_{t=0}^{T-1} A_{s_n,t}(\pi, x'_t, \pi'(x'_t))\right]$$

where $A_{s_n,t}(\pi, x'_t, \pi'(x'_t)) = \mathbb{E}_{\tau^\pi}\left[s_n(x'_t, \pi'(x'_t)) + J_{s_n,t+1}(\pi, \mu_n, x'_{t+1}) - J_{s_n,t}(\pi, \mu_n, x'_t)\right]$. Notice that we can rewrite $J_{s_n,t}(\pi, \mu_n, x'_t) = s_n(x'_t, \pi(x'_t)) + \mathbb{E}_{\tau^\pi}\left[J_{s_n,t+1}(\pi, \mu_n, \tilde{x}_{t+1})\right]$ for $\tilde{x}_{t+1} = \mu_n(x'_t, \pi(x'_t)) + w_t$. We can thus substitute this in the above to get:

$$\begin{aligned} &J_{s_n}(\pi', \mu_n) - J_{s_n}(\pi, \mu_n) = \mathbb{E}_{\tau^{\mu_n}_{\pi'}}\left[\sum_{t=0}^{T-1} A_{s_n,t}(\pi, x'_t, \pi'(x'_t))\right] \\ &= \mathbb{E}_{\tau^{\mu_n}_{\pi'}}\left[\sum_{t=0}^{T-1}\left(s_n(x'_t, \pi'(x'_t)) - s_n(x'_t, \pi(x'_t))\right)\right] \\ &\quad + \mathbb{E}_{\tau^{\mu_n}_{\pi'}}\left[\sum_{t=0}^{T-1} \mathbb{E}_{x'_{t+1}|x'_t,\pi'(x'_t)}\left[J_{s_n,k+1}(\pi, \mu_n, x'_{t+1})\right] - \mathbb{E}_{\tilde{x}_{t+1}|x'_t,\pi(x'_t)}\left[J_{s_n,k+1}(\pi, \mu_n, \tilde{x}_{t+1})\right]\right] \\ &\overset{(i)}{\leq} \mathbb{E}_{\tau^{\mu_n}_{\pi'}}\left[\sum_{t=0}^{T-1} \min\left\{L_s\|\pi'(x'_t) - \pi(x'_t)\|^{1/2}, 2\sqrt{d_x\sigma_{\max}}\right\}\right] \\ &\quad + \mathbb{E}_{\tau^{\mu_n}_{\pi'}}\left[\sum_{t=0}^{T-1}\left\{\sqrt{\mathbb{E}_{x'_{t+1}|x'_t,\pi'(x'_t)}\left[J^2_{s_n,k+1}(\pi, \mu_n, x'_{t+1})\right]}\right.\right. \\ &\qquad \left.\left. \times \min\left\{\frac{\|\mu_n(x'_t, \pi'(x'_t)) - \mu_n(x'_t, \pi(x'_t))\|}{\sigma_w}, 1\right\}\right\}\right] \qquad \text{([54], Lemma C.2)} \\ &\overset{(ii)}{\leq} \mathbb{E}_{\tau^{\mu_n}_{\pi'}}\left[\sum_{t=0}^{T-1}\left\{\min\left\{L_s\|\pi'(x'_t) - \pi(x'_t)\|^{1/2}, 2\sqrt{d_x\sigma_{\max}}\right\}\right.\right. \\ &\qquad \left.\left. + T\sqrt{d_x\sigma_{\max}}\min\left\{\frac{L_{\mu_n}\|\pi'(x'_t) - \pi(x'_t)\|}{\sigma_w}, 1\right\}\right\}\right] \\ &\leq \mathbb{E}_{\tau^{\mu_n}_{\pi'}}\left[\sum_{t=0}^{T-1}\left\{L_s\|\pi'(x'_t) - \pi(x'_t)\|^{1/2} + T\sqrt{d_x\sigma_{\max}}\frac{L_{\mu_n}\|\pi'(x'_t) - \pi(x'_t)\|}{\sigma_w}\right\}\right] \end{aligned}$$

where in (i)-(ii) we used the Hölder bounds provided by Lemma 8. Finally, using that $\|\pi' - \pi\|_\infty \leq \varepsilon_\pi$ by assumption we get

$$J_{s_n}(\pi', \mu_n) - J_{s_n}(\pi, \mu_n) \leq \mathbb{E}_{\tau_{\pi'}^{\mu_n}}\left[\sum_{t=0}^{T-1}\left\{L_s\|\pi'(x'_t) - \pi(x'_t)\|^{1/2} + T\sqrt{d_x\sigma_{\max}}\frac{L_{\mu_n}\|\pi'(x'_t) - \pi(x'_t)\|}{\sigma_w}\right\}\right]$$
$$\leq \sqrt{\varepsilon_\pi}T\left(L_s + TL_{\mu_n}\frac{\sqrt{d_x\sigma_{\max}\varepsilon_\pi}}{\sigma_w}\right) =: \sqrt{\varepsilon_\pi}C_\pi^n.$$

□

Lemma 7. *Let Assumptions 1-5 hold. Fix $n^\star$ as in (16) and ε, ζ and ξ as in Theorem 4. Then, at episode $n^\star$, it holds $\Pi_\xi^{\star,c} \subseteq \Pi_{\text{safe}}^{n^\star}$.*

Proof. Assume, for the sake of contradiction, that $\Pi_\xi^{\star,c} \setminus \Pi_{\text{safe}}^{n^\star}$ is nonempty. First, we can choose $\varepsilon_\pi > 0$ sufficiently small to satisfy $2\sqrt{\varepsilon_\pi}C_\pi^{n^\star}\lambda_{n^\star}^c \leq \xi - \varepsilon - \Delta_\zeta$, where $C_\pi^{n^\star}$ is defined in Lemma 6 and constant (since we consider a fixed $n^\star$. Then, under the hypothesis that $\Pi_\xi^{\star,c} \setminus \Pi_{\text{safe}}^{n^\star}$ is nonempty and under the definition of path-connectedness of $\Pi_\xi^{\star,c}$, there exists $\pi_{\text{safe}} \in \Pi_{\text{safe}}^{n^\star}$ and $\pi' \in \Pi_\xi^{\star,c} \setminus \Pi_{\text{safe}}^{n^\star}$ such that $\|\pi' - \pi_{\text{safe}}\|_\infty \leq \varepsilon_\pi$. This is because by the definition of path-connectedness of $\Pi_\xi^{\star,c}$, we can connect a policy in $\Pi_\xi^{\star,c} \setminus \Pi_{\text{safe}}^{n^\star}$ to some policy in $\Pi_{\text{prior}} \subseteq \Pi_{\text{safe}}^{n^\star}$ with a continuous line ρ in the policy space, and then choose along this line π_{safe} and π' arbitrarily close and such that they lie right inside and right outside of $\Pi_{\text{safe}}^{n^\star}$ respectively.

For these policies we can then apply Lemma 6 to bound:

$$J_{s_{n^\star}}(\pi', \mu_{n^\star}) \leq J_{s_{n^\star}}(\pi_{\text{safe}}, \mu_{n^\star}) + \sqrt{\varepsilon_\pi}C_\pi.$$

This bound can be substituted in when evaluating (22) along the trajectories generated by π'(as remember that (22) holds $\forall \pi \in \Pi$):

$$|J_c(\pi', f^\star) - J_c(\pi', f_{n^\star}^m)| \leq 2\lambda_{n^\star}^c J_{s_{n^\star}}(\pi', \mu_{n^\star}) \tag{26}$$
$$\leq 2\lambda_{n^\star}^c\left(J_{s_{n^\star}}(\pi_{\text{safe}}, \mu_{n^\star}) + \sqrt{\varepsilon_\pi}C_\pi^{n^\star}\right) \tag{27}$$
$$\leq 2\lambda_{n^\star}^c\left(d_\sigma^{n^\star} + \sqrt{\varepsilon_\pi}C_\pi^{n^\star}\right) \tag{28}$$
$$\leq \varepsilon + 2\sqrt{\varepsilon_\pi}C_\pi^{n^\star}\lambda_{n^\star}^c \leq \xi - \Delta_\zeta. \tag{29}$$

From this it follows, for all $m \in [M]$:

$$J_c(\pi', f_{n^\star}^m) \leq J_c(\pi', f^\star) + \xi - \Delta_\zeta \leq d - \xi + \xi - \Delta_\zeta \leq d - \Delta_\zeta. \tag{30}$$

This means that $\pi' \in \Pi_{\text{safe}}^{n^\star}$, which is a contradiction, thus proving the statement. □

Lemma 8. *Under assumption 4, we have for all $z, z' \in \mathcal{Z}$ and $n \geq 0$ that $|s_n(z) - s_n(z')| \leq L_s\|z - z'\|^{1/2}$ and $|\mu_n(z) - \mu_n(z')| \leq L_{\mu_n}\|z - z'\|$, where $L_s := \sqrt{2d_xL_k}$ and $L_{\mu_n} := \sqrt{\sum_{j=1}^{d_x}\left(L_\mu + \sqrt{nT}L_k\frac{\|y_{n,j} - \mu_j\|}{\sigma_w^2}\right)^2}$.*

Proof. For s_n we can write:

$$\begin{aligned}
|s_n(z) - s_n(z')| &= | \, \|\sigma_n(z)\| - \|\sigma_n(z')\| \, | \\
&\overset{(i)}{\leq} \|\sigma_n(z) - \sigma_n(z')\| \\
&= \sqrt{\sum_{j=1}^{d_x} (\sigma_{n,j}(z) - \sigma_{n,j}(z'))^2} \\
&\overset{(ii)}{\leq} \sqrt{d_x}\sqrt{k(z,z) - k(z,z') + k(z',z') - k(z',z)} \\
&\overset{(iii)}{\leq} \sqrt{2d_x L_k}\|z - z'\|^{1/2}
\end{aligned}$$

where (i) follows from the reverse triangular inequality, (ii) follows by applying Lemma 12 of Curi et al. [47] componentwise to the scalars $\sigma_{n,j}$ (note that this does not rely on having a bounded domain), and (iii) follows from Assumption 4.

For the mean, we can write for each $j \in [d_x]$:

$$\begin{aligned}
|\mu_{n,j}(z) - \mu_{n,j}(z')| &\leq |\mu_j(z) - \mu_j(z')| + \|k_n(z) - k_n(z')\| \left\| \left(K_n + \sigma_w^2 I\right)^{-1} (y_{n,j} - \mu_j) \right\| \\
&\leq L_\mu \|z - z'\| + \frac{1}{\sigma_w^2} \|y_{n,j} - \mu_j\| \|k_n(z) - k_n(z')\| \\
&\leq L_\mu \|z - z'\| + \frac{1}{\sigma_w^2} \|y_{n,j} - \mu_j\| \sqrt{nT} \max_{i \in [nT]} | \, k(z_i, z) - k(z_i, z') \, | \\
&\leq L_\mu \|z - z'\| + \frac{1}{\sigma_w^2} \|y_{n,j} - \mu_j\| \sqrt{nT} L_k \|z - z'\| \\
&\leq \left(L_\mu + \sqrt{nT} L_k \frac{\|y_{n,j} - \mu_j\|}{\sigma_w^2} \right) \|z - z'\|
\end{aligned}$$

Therefore, $\|\mu_n(z) - \mu_n(z')\| \leq L_{\mu_n} \|z - z'\|$. □

F Experiment Details

In this section, we provide additional details about the experimental evaluation presented in Section 7.

F.1 GP Experiments

The code for reproducing our experiments is available at this repository, which includes further details on the implementation, experimental setup, and hyperparameters. All experiments were conducted using a single NVIDIA RTX 4090 GPU per run, along with 5 CPU cores and 40GB of RAM. Under this setup, each experiment completes in under 1.5 hours.

Implementation details. In the GP implementation, we approximate Problem (8) with the following unconstrained optimization problem

$$\arg\max_{\pi \in \Pi} \; J_r(\pi, \mu_n) \; - \; \lambda_c \sum_{m=1}^{M} \max\left(J_c(\pi, f_n^m) - d + \Delta_\zeta, \; 0\right) - \lambda_\sigma \max\left(d_\sigma^n - J_{s_n}(\pi, \mu_n), \; 0\right).$$

where λ_c and λ_σ are penalty coefficients used to discourage constraints violations. Note that unlike a proper primal-dual implementation, we fix these values in our experiments (as is done in [8]). We solve the resulting optimization problem using the iCEM [64] optimizer. Following [8], we roll out a sequence of actions $\{a_t\}_{t=0}^{T-1}$ on the learned GP model using the TS1 scheme of [10]. Moreover, we maintain M particles, and given the state (s_t^m, a_t^m) for the m-th particle, we determine the next state s_{t+1}^m by sampling from $\mathcal{N}\big(\mu_n(s_t^m, a_t^m), \sigma_n^2(s_t^m, a_t^m)\big)$. This differs from the pathwise conditioning procedure described in Appendix A, as it injects independent noise at each timestep; however, it is computationally more efficient, as it avoids conditioning on previously observed data. Therefore, for each action sequence $\{a_t\}_{t=0}^{T-1}$ we obtain M trajectories, which are used to evaluate $J_c(\pi, f_n^m)$ and enforce the safety constraint. Finally, the exploration constraint is implemented using a metric that combines epistemic and aleatoric uncertainty [62, 8], and following Algorithm 1 it is deactivated once it becomes infeasible (up to some threshold).

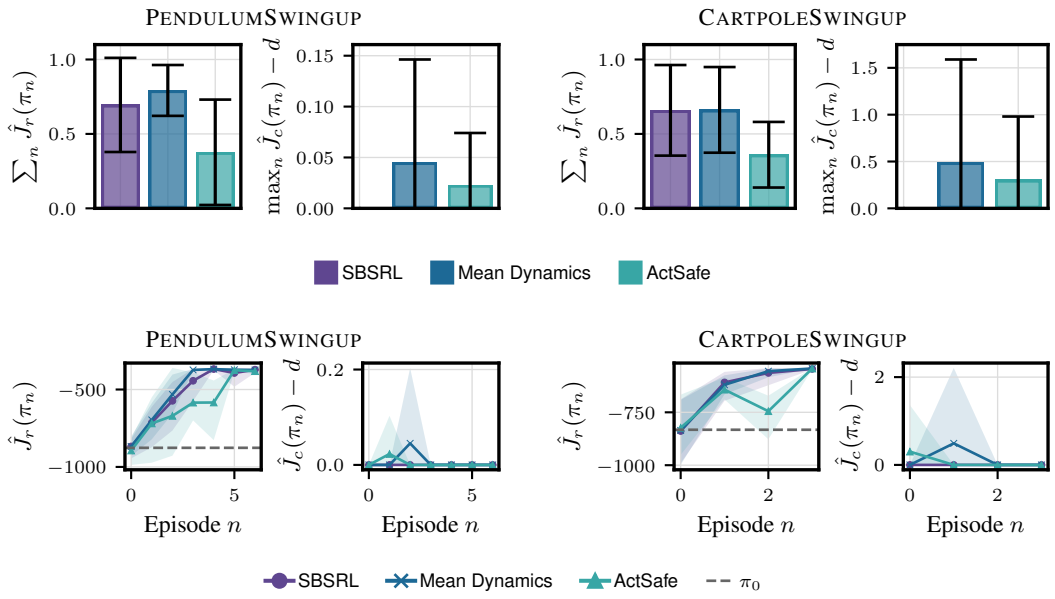

Figure 7: GP experiments demonstrating the effectiveness of using sampling to incentivize safety. The curves report mean and 95 percentile interval over five seeds, while the bar charts visualize mean and standard deviation of the normalized cumulative returns and maximum cost violations over five seeds.

Design choices. The sampling method described above differs from our theory, in that we resample the dynamics directly from the posterior and avoid projecting them onto the confidence bound, rather than maintaining a fixed set of prior samples refined via truncation as analyzed in Section 5. As discussed in Appendix A, our safety guarantee naturally extends to this resampling scheme via a union bound across episodes. We adopt this posterior-sampling approach to ensure a fair and direct comparison with our baselines, allowing us to faithfully replicate the experimental setup of As et al. [8]. As a second simplification, we keep the exploration threshold d_σ^n fixed across episodes rather than progressively tightening it. Although both variants are implemented in our code, we found empirically that this simplification preserves performance while improving training stability. Overall, these design choices simplify our implementation without altering the qualitative behavior predicted by our theoretical analysis. Finally, we fixed the number of samples to $M = 30$ across all experiments. As explained in Section 5, our approach induces a natural trade-off between computational cost and conservatism: since using more samples than prescribed by Lemma 2 does not compromise safety, one can in principle select the largest number of samples permitted by the available computational budget, and subsequently choose the constraint tightening accordingly. For instance, in our experiments we match the number of samples used by [8] to ensure comparable computational cost. Under this setup, we then empirically observe that smaller constraint tightenings than those required by ActSafe are sufficient to maintain safety.

Experimental setup. We evaluate SBSRL on the PENDULUM and CARTPOLE environments, focusing on the SWINGUP task. In CARTPOLE, following [8], we assume access to an initial offline dataset to ensure safe initialization (as required by Assumptions 2–3), whereas no such prior is required for PENDULUM. For completeness, Figure 7 complements Figure 4, including both cumulative rewards and training curves for the same safety experiment. Among the considered baselines, ActSafe includes an initial purely exploratory phase, during which rewards are not optimized, before transitioning to exploitation until the end of the simulation. In Figure 5 we report two possible choices of the transition episode.

Exploration constraint We provide an ablation study on the choice of d_σ^0 in Figure 8. While the theoretical definition of d_σ^n in Theorem 2 is difficult to estimate in practice, we find that SBSRL is not overly sensitive to its exact value. Choosing an intermediate value (neither too large nor too small) can improve performance; however, safety is not affected, since it is, at least in theory (ignoring

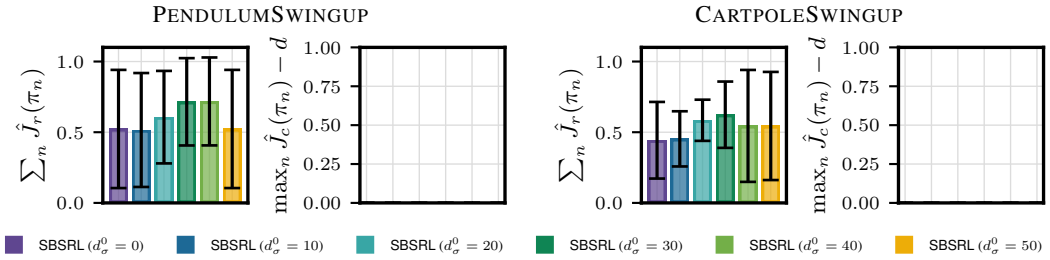

Figure 8: GP experiments ablating different values of d_σ^0 in the exploration constraint. The bar charts report the mean and standard deviation over five seeds for the normalized cumulative return and maximum cost violation.

approximation errors introduced by solving the constrained optimization problem), decoupled from exploration. Empirically, Figure 8 shows how setting d_σ^0 too large can deteriorate performance, since the constraint is deactivated early and the behavior quickly converges to that of $d_\sigma^0 = 0$. Conversely, choosing it very small can in principle lead to better asymptotic performance as suggested by Theorem 2, but it also increases sample complexity and thus can ultimately reduce the cumulative reward, as illustrated in the figure.

Rewards and constraints. We adopt reward and constraint definitions consistent with prior work on safe model-based reinforcement learning [8]. For both PENDULUM and CARTPOLE environments, let θ denote the pole angle, ω its angular velocity, and u the control input. We define the angular deviation from a desired target angle θ_{target} as $\Delta\theta = \theta - \theta_{\text{target}}$. In the PENDULUM environment, the objective penalizes deviations from the target configuration as well as large angular velocities and control inputs. The reward and cost are given by

$$r_{\text{Pendulum}} = -\left(\Delta\theta^2 + 0.1\,\omega^2 + 0.02\,u^2\right), \quad c_{\text{Pendulum}} = |\omega|, \quad d = 6.$$

For the CARTPOLE system, we additionally consider the cart position p and velocity v. The reward encourages stabilization of both the pole and the cart while penalizing control effort:

$$r_{\text{Cartpole}} = -\left(\Delta\theta^2 + p^2 + 0.1\left(v^2 + \omega^2\right)\right) - 0.01\,u^2,$$

while the cost captures violations of a position constraint on the cart,

$$c_{\text{Cartpole}} = |p|, \quad d = 1.5.$$

F.2 Hardware Experiment

While Gaussian Processes closely align with our theoretical framework, their cubic complexity in the number of data points limits scalability to high-dimensional control tasks. To address this, we also consider a scalable instantiation of SBSRL based on neural network dynamics models, enabling application to hardware, but also to standard continuous control benchmarks [11, 12]. The code for this implementation is available at this repository.

Implementation details. We instantiate SBSRL within a model-based actor-critic framework inspired by MBPO [65]. The system dynamics are modeled using a probabilistic ensemble of neural networks [66, 10], which enables multiple predictions of the dynamics for enforcing safety and provides an estimate of epistemic uncertainty [67] via the empirical standard deviation across ensemble members, which we exploit to implement the exploration constraint. To approximately solve Problem (8), we employ a mixture of real and model generated rollouts, and perform a fixed number of interleaved actor-critic updates, following standard practice in model-based RL [65]. In contrast to standard MBPO-style implementations, in order to approximate the sampling-based nature of our method we retain all ensemble predictions during rollouts, effectively simulating multiple dynamics realizations and enforcing safety constraints across them. To support this, we augment the cost critic to condition on the index m of the dynamics sample, thereby approximating

$$J_c(\pi, f_n^m) \approx \mathbb{E}_{\pi, x\sim\rho_0}\left[Q_c(x, \pi(x), m)\right].$$

Exploiting the structural similarity between J_c and J_{s_n}, we adopt the same approach to estimate the cumulative uncertainty along trajectories

$$J_{s_n}(\pi, f_n^m) \approx \mathbb{E}_{\pi, x\sim\rho_0}\left[Q_\sigma(x, \pi(x), m)\right].$$

This enables seamless integration of the exploration constraint within a standard actor-critic pipeline.

Design choices. Constraint satisfaction is enforced using an Augmented Lagrangian method [68], which provides strong empirical performance, consistent with prior work [69, 24], although other solvers could be used in principle. Furthermore, while the theoretical analysis prescribes a sample size M satisfying Lemma 2 for GPs, we found that fixing $M = 5$ across all experiments (both on hardware and in simulation) was sufficient for the implementation using deep neural networks to capture uncertainty while maintaining computational efficiency. Finally, note that in this implementation both reward and cost functions are unknown and therefore estimated via learned critics, as is standard in actor-critic methods.

Hardware setup. We follow the experiment setup of Wendl et al. [9] and evaluate SBSRL on a highly-dynamic remote-controlled race car operating at 60 Hz. States are observed through a motion capture systems that tracks the vehicle trajectory and sends real-time measurements to the controller. While the underlying system is characterized by a 7-dimensional state and a 2-dimensional action space, we concatenate past observations and control inputs using a sliding window. This mitigates delays arising from both actuation and motion capture, resulting in a 45-dimensional state-action representation. Such a representation is challenging for GP-based methods and motivates the use of deep ensembles. The task is to reach a target position while satisfying safety constraints, represented by obstacles that must be avoided by the vehicle (see Figure 1). Constraint violations are quantified using measurements from the motion capture system, and represent the kinetic energy dissipated during plastic collisions with the obstacles. The budget is set to $d = 15$, allowing for minor contacts with the tires while penalizing large collisions. In order to initialize our algorithm with a prior model that approximately satisfies Assumptions 2 and 3, we first collect 25K real-world transitions using a safe but suboptimal policy. This dataset is then used to empirically *calibrate* the model (and critics) of SBSRL offline. Finally, we deploy SBSRL online on the real system. At each episode, a policy is rolled-out on the real system to collect more data, which is added to the replay buffer to update the model before the next episode. We repeat the experiment for 75 episodes with three random seeds and report the mean and standard error of the cumulative rewards and costs after each iteration n in Figure 6. We compare SBSRL against a version of the same algorithm that does not use pessimism to enforce safety, i.e. it only enforces the constraint on the mean critic rather than across the ensemble. To ensure a fair comparison, both methods are initialized with the same offline prior, which includes the dynamics model and critics that were trained on the offline dataset.

G Additional Experiments

G.1 Experiments in Simulated Environments

We further evaluate SBSRL on standard continuous control benchmarks, including SafetyGym [11] and RWRL [12]. In these experiments, we follow the setup of [70, 9]; we refer the reader to these works for a detailed description of the tasks and environments. For these experiments we use a single NVIDIA RTX 4090 GPU per run, with 5 CPU cores and 40 GB of RAM. Multiple random seeds are executed in parallel via a Slurm cluster. This setup enables us to train our algorithm in less than an hour and a half for each experiment.

Safe offline-to-online. Ensuring safety when training neural network dynamics models from scratch is challenging without prior knowledge, as the initial model is typically poorly calibrated [8]. Our theoretical framework assumes access to a sufficiently good prior (cf. Assumptions 2 and 3), which is not directly satisfied in this setting. To address this, we adopt an offline-to-online training pipeline. However, note that our theory is agnostic to how such a prior is obtained, so that alternative approaches such as sim-to-real could also be considered. Specifically, we assume access to an initial dataset collected by a suboptimal (and potentially unsafe) behavior policy. We first run our algorithm in an offline setting (i.e. using only the offline data) to learn a (empirically) calibrated dynamics model and associated critics. These are then used to initialize the online phase, allowing the agent to begin learning from a reliable model. The only modification required to run SBSRL offline is to set $d_\sigma^n = 0$, thereby deactivating the exploration constraint. This choice reflects the offline RL setting, where the agent should remain close to the data distribution rather than be encouraged to explore. If one instead wishes to impose explicit exploration penalties during the offline phase, as suggested by Yu et al. [71], our framework can naturally accommodate this by reversing the exploration constraint. In our experiments, however, this was not necessary, allowing us to retain a single practical algorithm for the full offline-to-online pipeline. To fairly validate SBSRL accounting for the variability in the offline initialization, we generate multiple priors by running 5 independent offline training seeds on

the same dataset. For each resulting prior, we then run 5 independent online training seeds. This results in a total of 25 runs per experiment, all of which are reported in our evaluation.

Benchmarks and baselines. We evaluate our method on several tasks from the RWRL benchmark [12] and SafetyGym [11]. In Figure 9, we compare SBSRL against two baselines from the literature. First, we consider MBPO [65], which also serves as the base architecture for our implementation, as explained in the previous section. MBPO is a model-based actor-critic method that uses a neural network ensemble to model the dynamics and generates rollouts via the TS1 scheme of [10], randomly selecting a different ensemble member at each step. Notably, MBPO does not incorporate additional pessimism to enforce safety, which results in cost violations in some tasks. Second, we compare against SOOPER [9], a recent state-of-the-art safe RL algorithm. Its implementation is also based on an MBPO-like architecture, enabling a direct and fair comparison. In particular, we initialize both MBPO and SOOPER with a prior policy and backup critics that are constructed from the same offline run used to initialize SBSRL. Across all experiments, SBSRL maintains safety at all episodes throughout training while improving rewards over the safe baselines. The plots report the mean and 95% empirical percentile intervals over 25 runs, generated from 5 different offline priors as described above.

Constraint satisfaction through sampling. In Figure 10 we provide a comparison of SBSRL against a mean baseline that only uses the mean of the multiple critic predictions across ensemble members for planning, rather than accounting for each sampled dynamics model. The results show that SBSRL achieves comparable task performance while consistently satisfying the safety constraints across all runs. We report the mean and 95% empirical percentile intervals over all 25 runs, highlighting robustness of SBSRL to the offline prior. This explains the apparent high variance at initialization and provides a more informative measure of robustness than reporting standard errors alone.

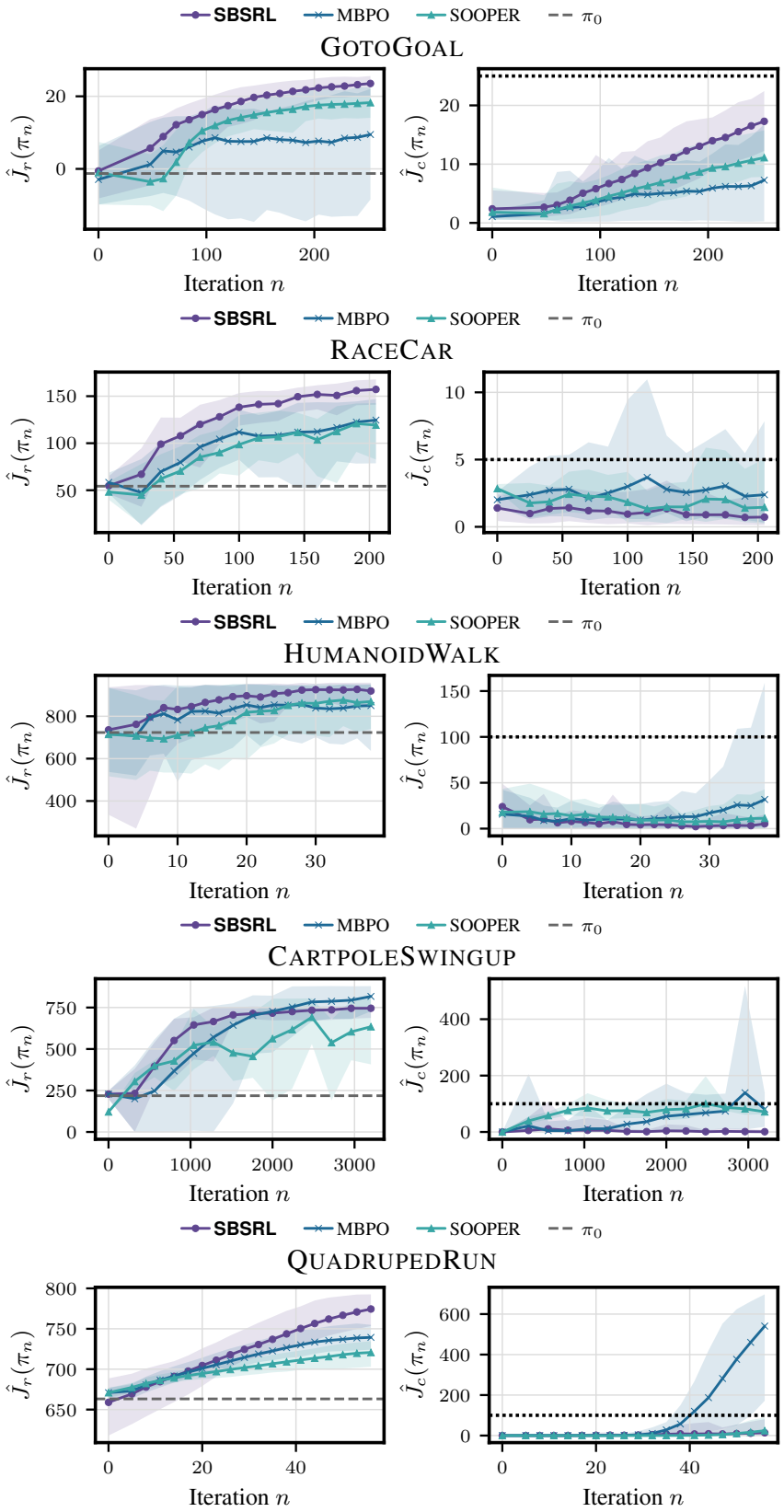

Figure 9: Learning curves comparing SBSRL against standard baselines. The performance is evaluated in the environments GOTOGOAL, RACECAR, HUMANOID, CARTPOLE and QUADRUPED . We report the mean and 95 percentile interval.

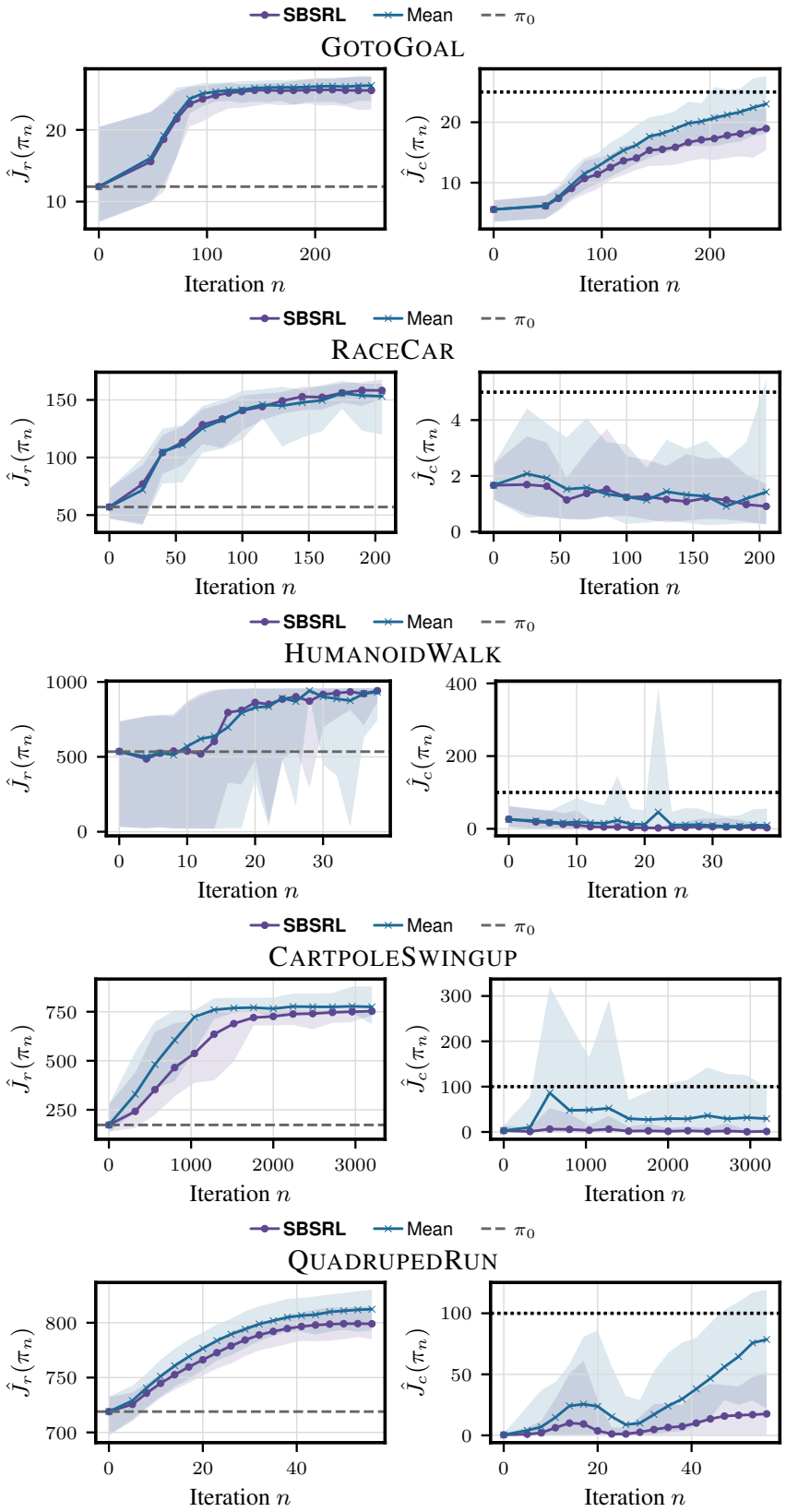

Figure 10: Learning curves comparing SBSRL against the mean baseline. The performance is evaluated in the environments GoToGoal, RaceCar, Humanoid, Cartpole and Quadruped. We report the mean and 95 percentile interval.