Probabilistic Safeguard for Reinforcement Learning Using Safety Index Guided Gaussian Process Models

Weiye Zhao* WEIYEZHA@ANDREW.CMU.EDU
Tairan He* TAIRANH@ANDREW.CMU.EDU
Changliu Liu CLIU6@ANDREW.CMU.EDU
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

Safety is one of the biggest concerns to applying reinforcement learning (RL) to the physical world. In its core part, it is challenging to ensure RL agents persistently satisfy a hard state constraint without white-box or black-box dynamics models. This paper presents an integrated model learning and safe control framework to safeguard any RL agent, where the environment dynamics are learned as Gaussian processes. The proposed theory provides (i) a novel method to construct an offline dataset for model learning that best achieves safety requirements; (ii) a design rule to construct the safety index to ensure the existence of safe control under control limits; (iii) a probablistic safety guarantee (i.e. probabilistic forward invariance) when the model is learned using the aforementioned dataset. Simulation results show that our framework achieves almost zero safety violation on various continuous control tasks.

Keywords: Safe control, Gaussian process, Dynamics learning

1. Introduction

While reinforcement learning (RL) has achieved impressive results in games like Atari (Zhao et al., 2019), Go (Silver et al., 2017) and Starcraft (Vinyals et al., 2019), the lack of safety guarantee limits the application of RL algorithms on real-world physical systems such as robotics (Wei et al., 2022). In its core part, it is critical to ensure that RL agents persistently satisfy a hard state constraint defined by a *safe set* (e.g., a set of non-colliding states) in many robotic applications (Zhao et al., 2021, 2020a,b). Though various constrained RL algorithms (He et al., 2023b; Achiam et al., 2017; Wachi et al., 2018; Yang et al., 2021; Zhao et al., 2023) have been introduced, the trial-and-error mechanism of these methods makes it hard to avoid safety violations during policy learning.

On the other hand, when the dynamics model of the system is accessible, energy-function-based safe control methods can achieve the safety guarantee, i.e., persistently satisfying the hard state constraint. These methods (Noren et al., 2021; Zhao et al., 2022b; Liu and Tomizuka, 2014; Gracia et al., 2013; He et al., 2023a) first synthesize an energy function such that the safe states have low energy, and then design a control law to satisfy the safe action constraints, i.e., to dissipate energy. Then these methods ensure *forward invariance* inside the safe set . However, their limitation is that they exploit either white-box dynamics models (e.g., analytic form) (Khatib, 1986; Ames et al., 2014; Liu and Tomizuka, 2014; Gracia et al., 2013) or black-box dynamics models (e.g.,

*. These authors contributed equally to this work.

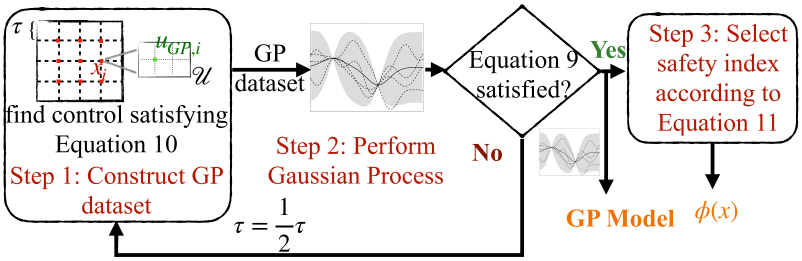
τ
$u_{GP,i}$
x_i
$\mathcal{U}$
find control satisfying Equation 10
Step 1: Construct GP dataset
GP dataset
Step 2: Perform Gaussian Process
Equation 9 satisfied?
Yes
No
Step 3: Select safety index according to Equation 11
$\tau = \frac{1}{2}\tau$
GP Model
$\phi(x)$

Figure 1: The flow chart that illustrate the offline state of UAISSA. We first select a proper state space discretization step size, construct the offline dynamics learning dataset for GP and design parameters of the safety index.

digital twin simulators) (Zhao et al., 2021), while these models are not easy to build in complex environments. Other related works are summarized in Appendix A.

Practically, compared to dynamics models (i.e., a full mapping from the current state and control to the next state), it is easier to obtain samples of the dynamic transitions in real-world applications (Huang et al., 2018; Caesar et al., 2020; Cheng et al., 2019b; Sun et al., 2023). This paper investigates approaches to utilize these transition samples to achieve safety guarantees under the energy-function-based safe control framework, while relaxing the requirements of white-box or black-box dynamics models. In our methods, we leverage Gaussian Process (GP) to learn a statistical dynamics model due to (i) GP's reliable estimate of uncertainty (Williams and Rasmussen, 2006); (ii) its well-established theory on uniform error bounds (Srinivas et al., 2009, 2012; Chowdhury and Gopalan, 2017; Kanagawa et al., 2018; Lederer et al., 2019). Instead of performing online model learning using online data, our dynamics model is learned based on an offline constructed dataset. When the dataset is constructed offline, we have the full control over the data distribution, which could result in (i) reliable convergence in model learning and (ii) good safety guarantees.

The main contribution of this paper is a theory to probabilistically safeguard robot policy learning using energy-function-based safe control with a GP dynamics model learned on an offline dataset. The overall pipeline of our method is shown in Figure 1. To achieve our goal of safeguarding RL agents with GP dynamics model, we first show how to construct the dataset for model learning and how to design the associated energy function (called *safety index*) so that there always exists a feasible safe control under control limits. Secondly, we show how to design a safeguard for arbitrary RL agents to guarantee forward invariance during policy learning. The method is evaluated on various challenging continuous control problems where the RL agents achieve almost zero constraint violation during policy learning. Additional results and discussions can be found in the appendix of the arxiv version https://arxiv.org/abs/2210.01041.

2. Problem Background

2.1. Notations

Dynamics Denote $x_t \in \mathcal{X} \subset \mathbb{R}^{n_x}$ as the robot state at time step t; $u_t \in \mathcal{U} \subset \mathbb{R}^{n_u}$ as the control input to the robot at time step t, and the control space $\mathcal{U}$ is bounded. And denote $\mathcal{W} := \mathcal{X} \times \mathcal{U}$, which is assumed to be compact. The system dynamics are defined as:

$$x_{t+1} = f(x_t, u_t), \tag{1}$$

where $f : \mathcal{W} \rightarrow \mathcal{X}$ is a function that maps the current robot state and control to the robot state in the next time step, and $f(\cdot)$ is L_f Lipschitz continuous with respect to the 1-norm. For simplicity, this paper considers deterministic dynamics (but is unknown).

Safety Specification The safety specification requires that the system state should be constrained in a closed subset in the state space, called the safe set $\mathcal{X}_S$. The safe set can be represented by the zero-sublevel set of a continuous and piecewise smooth function $\phi_0 : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$, i.e., $\mathcal{X}_S = \{x \mid \phi_0(x) \leq 0\}$. $\mathcal{X}_S$ and ϕ_0 are directly specified by users.

2.2. Preliminary

Gaussian Process A Gaussian process (GP) (Williams and Rasmussen, 2006) is a nonparametric regression method specified by its mean $\mu_g(z) = \mathbb{E}[g(z)]$ and covariance (kernel) functions $k(z, z') = \mathbb{E}[(g(z) - \mu(z))(g(z') - \mu(z'))]$. Given N finite measurements $y_N = [y(z_1), y(z_2), \cdots, y(z_N)]^T$ of the unknown function $g : \mathbb{R}^D \rightarrow \mathbb{R}$ subject to independent Gaussian noise $v \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$, the posterior mean $\mu(z_*)$ and variance $\sigma^2(z_*)$ are calculated as:

$$\mu(z_*) = k_*^T(z_*)(K + \sigma^2_{\text{noise}} I_N)^{-1} y_N \tag{2}$$

$$\sigma^2(z_*) = k(z_*, z_*) - k_*^T(z_*)(K + \sigma^2_{\text{noise}} I_N)^{-1} k_*(z_*), \tag{3}$$

where $K_{i,j} = k(z_i, z_j)$ and $k_*(z_*) = [k(z_1, z_*), k(z_2, z_*), \cdots, k(z_N, z_*)]^T$. In the following discussions, we assume observation is noise-free, i.e. $\sigma_{\text{noise}} = 0$. Note that noise reduction methods can be applied to eliminate σ_{noise} in practice (Kostelich and Schreiber, 1993). For most commonly used kernel functions, GP can approximate any continuous function on any compact subset of $\mathcal{Z}$ (Srinivas et al., 2012). In this paper, the dynamics are modeled using GP with the following definition.

Definition 1 (GP Dynamics Model) *The dynamics model of f in (1) is represented as a zero mean Gaussian process with a continuous covariance kernel $k(\cdot, \cdot)$ with Lipschitz constant L_k on the compact set $\mathcal{W}$, where L_k can be caluclated analytically for commonly-used kernels (Lederer et al., 2019). The posterior mean function and covariance matrix function of the GP model are denoted as $\mu_f(\cdot)$ and $\Sigma_f(\cdot)$, respectively.*

Safety Index To ensure system safety, all visited states should be inside $\mathcal{X}_S$. However, $\mathcal{X}_S$ may contain states that will inevitably go to the unsafe set no matter what control inputs we choose. Hence, we need to assign high energy values to those inevitably unsafe states, and ensure **forward invariance** in a subset of the safe set $\mathcal{X}_S$. Safe Set Algorithm (SSA) (Liu and Tomizuka, 2014) synthesizes the energy function as a continuous, piece-wise smooth scalar function $\phi : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$, named, safety index. And we denote its 0-sublevel set as $\mathcal{X}_S^D := \{x | \phi(x) \leq 0\}$. The general form of the safety index was proposed as $\phi = \phi_0^* + k_1 \dot{\phi}_0 + \cdots + k_n \phi_0^{(n)}$ where (i) the roots of $1 + k_1 s + \ldots + k_n s^n = 0$ are all negative real (to ensure zero-overshooting of the original safety constraints); (ii) the relative degree from $\phi_0^{(n)}$ to u is one (to avoid singularity); and (iii) ϕ_0^* defines the same zero sublevel set as ϕ_0 (to nonlinear shape the gradient of ϕ at the boundary of the safe set). It is shown in (Liu and Tomizuka, 2014) that choosing a control that decreases ϕ whenever ϕ is greater than or equal to 0 can ensure forward invariance inside $\mathcal{X}_S \cap \mathcal{X}_S^D$.

2.3. Problem Formulation

The core problem of this paper is to safeguard a nominal controller (i.e., an RL agent) such that all visited states are inside $\mathcal{X}_S$. In this paper, we are specifically interested in *degree two systems* (i.e., the relative degree from $\dot{\phi}_0$ to u is one), and the safety specification is defined as $\phi_0 = d_{min} - d$ where d denotes the safety status of the system, and the system becomes more unsafe when d decreases. For example, for collision avoidance, d can be designed to measure the relative distance between the robot and obstacles, which needs to be greater than some threshold d_{min}. Following the rules in (Liu and Tomizuka, 2014), we parameterize the safety index as $\phi = \sigma + d_{min}^n - d^n - k\dot{d}$, and $\sigma, n, k > 0$ are tunable parameters of the safety index. It is easy to verify that this design satisfies the three requirements discussed above.

The nominal control is an RL controller which aims to maximize cumulative discounted rewards in an infinite-horizon deterministic Markov decision process (MDP). An MDP is specified by a tuple $(\mathcal{X}, \mathcal{U}, \gamma, r, f)$, where $r : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$ is the reward function, $0 \leq \gamma < 1$ is the discount factor, and f is the deterministic system dynamics defined in (1), and we can access data samples of f. We then define the set of safe control as $\mathcal{U}_S^D(x) := \{u \in \mathcal{U} \mid \phi(f(x,u)) \leq \max\{\phi(x) - \eta, 0\}\}$, where η is a positive constant. Hence, the nominal controller can be safeguarded by projecting the nominal control u_t^r to $\mathcal{U}_S^D(x)$ by solving the following optimization:

$$\begin{aligned} &\min_{u_t \in \mathcal{U}} \|u_t - u_t^r\|^2 \\ &\text{s.t. } \phi(f(x_t, u_t)) \leq \max\{\phi(x_t) - \eta, 0\}. \end{aligned} \tag{4}$$

Since f is unknown, we need to first learn a statistical model of f and then solve (4). The well-established theories on uniform error bounds (Srinivas et al., 2009, 2012; Chowdhury and Gopalan, 2017; Kanagawa et al., 2018; Lederer et al., 2019) for GP allows us to a build a reliable statistical model for a given dataset.

Lemma 2 (Well-Calibrated Model) *For a dataset and $\delta \in (0, 1)$, there exists $\beta_f(\delta)$ that we can learn a GP model $\{\mu_f(x,u), \sigma_f(x,u)\}$ that satisfies: $\forall x \in \mathcal{X}, u \in \mathcal{U}, P\Big(||f(x,u) - \mu_f(x,u)||_1 \leq \beta_f \sigma_f(x,u)\Big) \geq 1 - \delta$, where β_f means $\beta_f(\delta)$ for simplicity and $\sigma_f(x,u) = \mathrm{Tr}(\Sigma_f^{\frac{1}{2}}(x,u))$.*

This lemma ensures that the confidence intervals of GP prediction cover the true dynamics function with high probability given an appropriate constant β_f. The expressions of β_f are discussed in (Srinivas et al., 2009, 2012; Chowdhury and Gopalan, 2017; Kanagawa et al., 2018; Lederer et al., 2019).

Challenges The challenges for solving (4) can be divided into two parts: (1) **offline synthesis stage**: how to generate a data set for model learning and safety index synthesis such that: (a) there is always a solution for (4) with the learnt dynamics under control limit; (b) safety is preserved under model mismatch. (2) **online computation stage**: how to efficiently solve (4) with learnt dynamics to find safe controls.

3. Offline Safety Index Synthesis

In this section, we introduce the theoretical results to tackle the aforementioned offline synthesis stage challenges. We first show that the deterministic constraint (4) can be verified via introducing an upper bound of safety index. Next, we introduce a theory that verifies the feasibility of the probabilistic constraint for all possible states (which is uncountably many) by verifying the feasibility of a similar problem for finitely many states (Proposition 5). Lastly, we discuss the criteria of dataset construction for model learning and the associated safety index design rule which ensure nonempty set of safe control for all possible system states (Theorem 7).

3.1. Preserving Safety with Learnt Model

As mentioned in Section 2.3, our ultimate goal is to solve (4), whereas it is intractable to directly solve (4) since f is unknown. On the other hand, we can learn a reliable statistical model of f via GP, i.e. $\{\mu_f(x,u), \sigma_f(x,u), \beta_f\}$ as stated in Lemma 2. Hence, as long as we can find a probabilistic upper bound of safety index $\phi(f(x,u))$, denoted as $\mathbf{U}_f(x,u)$ such that $\mathbf{U}_f(x,u) \geq \phi(f(x,u))$, the deterministic condition of (4) can be verified through a stricter condition, i.e.

$$\mathbf{U}_f(x,u) < \max(\phi(x) - \eta, 0). \tag{5}$$

In Lemma 10, we derive the probabilistic upper bound of safety index as

$$\mathbf{U}_f(x,u) := \phi(\mu_f(x,u)) + L_\phi \beta_f \sigma_f(x,u), \tag{6}$$

where L_ϕ is the Lipschitz constant of $\phi(\cdot)$ with respect to 1-norm. Lemma 10 shows that $\phi(f(x,u))$ is smaller than $\mathbf{U}_f(x,u)$ with probability at least $(1-\delta)$. The proof of this probabilistic upper bound is given in Appendix C.1.

Nonempty Set of Safe Control By introducing $\mathbf{U}_f(x,u)$, we have addressed the challenge (1.b). In the following two subsections, we will address challenge (1.a) by ensuring the existence of nonempty set of safe control for all possible states under control limit when solving (5), i.e.

$$\forall x \in \mathcal{X}, \exists u \in \mathcal{U}, \text{ s.t. } \mathbf{U}_f(x,u) < \max(\phi(x) - \eta, 0). \tag{7}$$

3.2. Infinite to Finite Conditions

Notice that verifying condition (7) on the continuous state space is still intractable. Therefore, we consider a discretization of the state space defined as follows.

Definition 3 (Discretization) *A τ-discretization $\mathcal{H}_\tau$ of a set $\mathcal{H}$ is defined as $\mathcal{H}_\tau := \{h_1, h_2, \ldots\}$ such that $\forall h \in \mathcal{H}, \exists h_i \in \mathcal{H}_\tau$ s.t. $||h_i - h||_1 \leq \tau$.*

Definition 4 (Data) *A dataset on a state space τ-discretization $\mathcal{X}_\tau$ is a collection of transition samples defined as $\mathcal{D}_\tau := \{(x_i, u_i, f(x_i, u_i))\}_{i=1}^{|\mathcal{X}_\tau|}$ where $x_i \in \mathcal{X}_\tau$.*

Given this discretization, if we ensure the existence of safe control for states in $\mathcal{X}_\tau$, together with the Lipschitz continuity and the bound on posterior variance of statistical models, then we can ensure the existence of safe control on the continuous state space $\mathcal{X}$.

Proposition 5 (Equivalence in Feasibility Conditions) *With the GP defined in Definition 1, the state-space τ_x-discretization $\mathcal{X}_{\tau_x}$ defined in Definition 3 and the dataset $\mathcal{D}_{\tau_x}$ defined in Definition 4, if the following condition holds:*

$$\forall (x_i, u_i), \mathbf{U}_f(x_i, u_i) < \max\{\phi(x_i) - \eta, 0\} - L_\phi L_f \tau_x - L_\phi \tau_x - 2L_\phi \beta_f \tilde{\sigma}_f\ , \tag{8}$$

where

$$\tilde{\sigma}_f = n_x\sqrt{2L_k\tau_x + 2|\mathcal{X}_{\tau_x}|L_k\tau_x\|K^{-1}\| \max_{w,w'\in\mathcal{W}} k(w,w')},$$

then it holds with probability $1-\delta$ *that*

$$\forall x \in \mathcal{X}, \exists u \in \mathcal{U},\ \textit{s.t.}\ \mathbf{U}_f(x,u) < \max(\phi(x)-\eta, 0).$$

The proof of Proposition 5 is given in Appendix C.6. Proposition 5 states that, in order to provide guarantee on the nonempty set of safe control in the whole continuous state space $\mathcal{X}$, it is sufficient to check a stricter condition (i.e., (8)) of nonempty set of safe control on the discretized state set $\mathcal{X}_{\tau_x}$. Note that the additional bounds on discretized states $\mathcal{X}_{\tau_x}$ (i.e. $L_\phi L_f\tau_x$, $L_\phi\tau_x$, $2L_\phi\beta_f\tilde{\sigma}_f$) of (8) become zero as the discretization constant τ goes to zero.

3.3. Dynamics Learning and Safety Index Design Theory

Synthesize Safe Index So far, we have shown that (8) implies (7) in a probabilistic way. Therefore, a theory that quantifies how to parameterize ϕ to make (8) hold is needed. To begin with, we first need to ensure there exists such a safety index to make (8) hold. Hence, an assumption is made:

Assumption 1 (Safe Control) *The state space is bounded, and the infimum of the supremum of* $\Delta\dot{d}$ *can achieve positive, i.e.,* $\inf_x \sup_u \Delta\dot{d}(x,u) > 0$.

Here $\Delta\dot{d}$ denotes the change of $\dot{d}$ at one time step. The necessity of Assumption 1 is summarized in Appendix C.2. Essentially, Assumption 1 enables a degree two system to dissipate energy (i.e., $\ddot{\phi} < 0$) at all states. Subsequently, the safety index design rule is summarized as follows:

Theorem 6 (Feasibility of Safety Index Design) *Denote* $d(\cdot)$ *and* $\dot{d}(\cdot)$ *as the mappings from* x *to* d *and* $\dot{d}$ *with Lipschitz constant* L_{d_x} *and* $L_{\dot{d}_x}$ *with respect to 1-norm. Under Assumption 1, if we (1) select a state-space* τ_x*-discretization* $\mathcal{X}_{\tau_x}$ *with step size such that*

$$\tau_x \leq \min\left\{1, \left[\frac{\inf_x \sup_u \Delta\dot{d}(x,u)}{2(L_{d_x}+L_{\dot{d}_x})\left(1+L_f+2\beta_f n_x\sqrt{2L_k}\sqrt{1+|\mathcal{X}_{\tau_x}|\|K^{-1}\| \max_{w,w'\in\mathcal{W}} k(w,w')}\right)}\right]^2\right\} \tag{9}$$

(2) construct the corresponding dataset $\left\{\left(x_i, u_{GP,i}, f(x_i, u_{GP,i})\right)\right\}_{i=1}^{|\mathcal{X}_{\tau_x}|}$ *on* $\mathcal{X}_{\tau_x}$ *by selecting* $u_{GP,i}$ *such that for any* $x_i \in \mathcal{X}_{\tau_x}$

$$\underbrace{\dot{d}(f(x_i, u_{GP,i}))}_{\dot{d}_{GP,i}} - \underbrace{\dot{d}(x_i)}_{\dot{d}_i} > \frac{\inf_x \sup_u \Delta\dot{d}(x,u)}{2}, \tag{10}$$

(3) choose the safety index parameters such that

$$\begin{cases} \sigma = 0, \\ n = 1, \\ k > \max_{x_i\in\mathcal{X}_{\tau_x}}\left\{\max\left\{1, \Upsilon_i\right\}\right\} \end{cases} \tag{11}$$

where we denote $d_{GP,i} = d(f(x_i, u_{GP,i}))$, $d_i = d(x_i)$, *and*

$$\Upsilon_i = \frac{\eta + d_i - d_{GP,i}}{\dot{d}_{GP,i} - \dot{d}_i - (L_{d_x} + L_{\dot{d}_x})\big(\tau_x - L_f\tau_x - 2\beta_f n_x \tilde{\sigma}_f\big)}$$
$$\tilde{\sigma}_f = n_x\sqrt{2L_k\tau_x + 2|\mathcal{X}_{\tau_x}|L_k\tau_x\|K^{-1}\| \max_{w,w'\in\mathcal{W}} k(w,w')},$$

then there always exists a safe control for any discretized state

$$\begin{aligned}&\forall x_i \in \mathcal{X}_\tau,\ \exists u \in \mathcal{U},\ \textit{s.t.}\\ &\mathbf{U}_f(f(x_i,u)) < \max\{\phi(x_i) - \eta, 0\} - L_\phi L_f \tau_x - L_\phi \tau_x - 2L_\phi\beta_f\tilde{\sigma}_f.\end{aligned} \tag{12}$$

The proof for Theorem 6 is summarized in Appendix C.9 . Theorem 6 states that, firstly, we select a proper discretization gap of state space such that it is small enough according to (9). Secondly, we construct an offline dataset such that the selected control for each discretized state can increase $\dot{d}$ by a certain volume according to (10). Lastly, by performing GP regression on the constructed dataset, the safety index designed according to (11) ensures the existence of probabilistic safe control for all discretized states to satisfy (12). Note that (12) is equivalent to (8), the following theorem is thus a direct consequence of Proposition 5 and Theorem 6.

Theorem 7 *Under the same assumptions of Theorem 6, by selecting state discretization step size according to* (9)*, constructing Gaussian process dataset according to* (10)*, and defining safety index according to* (11)*, then it holds with probability* $1 - \delta$ *that*

$$\begin{aligned}&\forall x \in \mathcal{X},\ \exists u,\ \textit{s.t.}\\ &\phi(f(x,u)) \leq \mathbf{U}_f(x,u) < \max(\phi(x) - \eta, 0).\end{aligned} \tag{13}$$

Remark 8 *It is worth noting that the system property* $\inf_x \sup_u \Delta\dot{d}(x,u) > 0$ *is crucial for establishing the nonempty set of safe control theorem as indicated in* (9) *and* (10)*. In practice, a lower bound of* $\inf_x \sup_u \Delta\dot{d}(x,u)$ *can be obtained via sampling the state space and control space, which is summarized in Appendix C.10.*

4. Uncertainty-Aware Implicit Safe Set Algorithm

In the previous section, we established theoretical results for safety index design to ensure a nonempty set of safe control with learned dynamics models. However, due to the non-control-affine nature of the GP dynamics model and the limitations of conventional QP-based projection methods, we employ a multi-directional line search approach to solve the black-box optimization problem in (4). In this section, we present a practical algorithm called Uncertainty-Aware Implicit Safe Set Algorithm (UAISSA) that builds upon the theoretical foundations discussed earlier and utilizes a sample-efficient black-box constrained optimization algorithm (Zhao et al., 2021). The details of UAISSA can be found in Algorithm 1 (see Appendix B). To have a better understanding on the *Offline Stage* of UAISSA, we summarize the procedure for constructing a valid safety index and the associated GP dynamics model in Figure 1. Firstly, we randomly select a step size τ, and perform τ-discretization of the state space. For each discretized state x_i, we use sampling (grid sampling or random sampling) to find a control $u_{GP,i}$ satisfying (10), which results in a dynamics learning dataset $\big\{\big(x_i, u_{GP,i}, f(x_i, u_{GP,i})\big)\big\}_{i=1}^{|\mathcal{X}_{\tau_x}|}$. Next, we learn a GP dynamics model from the constructed dataset. Together with the Lipschitz constants and well-calibrated GP dynamics model, we can then evaluate (9). If (9) does not hold, we will further shrink the discretization step size by half, and repeat the aforementioned procedures. If (9) holds, we will evaluate Υ_i for each x_i from the dataset, and select the parameters for the safety index $\phi(x)$ according to (11).

With the guarantee of nonempty set of safe control provided by Theorem 7, and the fact that ISSA can always find a suboptimal solution of (4) with finite iterations if the set of safe control is non-empty [Proposition 2, (Zhao et al., 2021)], the following theorem is thus a direct consequence of Theorem 1 from (Zhao et al., 2021).

Theorem 9 (Forward Invariance) *If the control system satisfies Assumption 1 and with the GP model and the safety index as specified in Theorem 6, then if $\phi(x_t) \leq 0$, Algorithm 1 guarantees $\phi(x_{t+1}) \leq 0$ with probability $1 - \delta$.*

5. Experiment

We evaluate UAISSA in two experiments: (i) Robot arm, where we apply Theorem 7 to ensure nonempty set of safe control for an unknown robotics manipulator system ; (ii) Safety Gym, where we apply UAISSA to safeguard unknown complex systems.

5.1. Robot Arm

We verify the correctness of our approach on a planar robotics manipulator with 2 degrees of freedom (2DOFs) (Zhao et al., 2022a). The robot has a four dimensional state space: $x = [\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2]$, where θ_i is the i-th joint angle in the world frame. We consider limited state space, i.e., $\theta_1 \in [0, \pi]$, $\theta_2 \in [0, 2\pi]$, $\forall i = 1, 2, \dot{\theta}_i \in [-0.1, 0.1]$. The system inputs are accelerations of the two joints, i.e. $[\ddot{\theta}_1, \ddot{\theta}_2]$. The system is simulated with $dt = 1$ms. The system is shown in Figure 2, where the robot is randomly exploring the environment and we need to safeguard the robot from colliding with the wall. The link length of the robot is 1 meter. The wall is 1 meter away from the robot base.

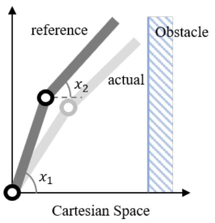

Figure 2: 2DOFs robot manipulator.

5.1.1. SAFETY INDEX DESIGN RUNNING EXAMPLE

To apply Theorem 7 to obtain the safety index parameters, we consider $L_f, L_{\Delta \dot{d}}, L_{d_x}, L_{\dot{d}_x}$ to be known. Firstly, we need to find the proper state space disretization step size τ_x. We start with $\tau_x = 0.5$, and construct a learning dataset where a safe control is sampled for each discretized state, such that (10) holds. dynamics data sample include input entry and output entry, where the input entry is a stack of state and sampled control ($[\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2, \ddot{\theta}_1, \ddot{\theta}_2]$), and output entry is the state at next time step. An example for data sample is: $\{[0.1, 0.5, -0.1, -0.1, 0.82, 0.36], [0.1, 0.49, -0.09, -0.09]\}$.

Then, we perform Gaussian Process to learn a well-calibrated dynamics model, where a uniform error bound theory (Lemma 16 in Appendix C.7) with $\delta = 1\%$ (i.e., 99% confidence interval) is applied to select β_f. With the learnt GP model, we check if (9) holds. If not, we further decrease τ_x by multiplying τ_x with 0.99 and repeat the process.

Finally, we find a discretization step of $\tau_x = 0.174$, resulting a dataset with 2516 samples. By setting $\eta = 0.05$, the safety index parameterization is obtained as: $\sigma = 0, n = 1, k = 2.54$

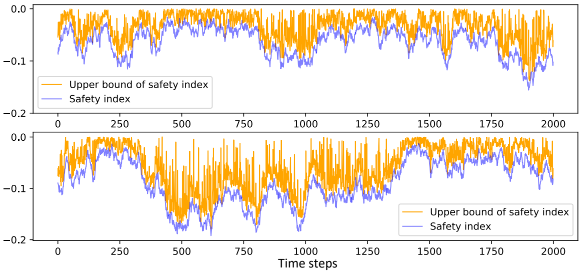

Figure 3: Evolutions of safety index and its upper bound with UAISSA over two runs.

according to (11). Intuitively, k reflects UAISSA reaction sensitivity to unsafe situations, e.g. larger k indicates safe control is more likely to be generated when the robot moves toward the obstacle.

5.1.2. ROBOT ARM RESULTS

This section numerically verifies that the synthesized safety index facilitates probabilistic forward invariance, by showing that 1) the upper bound $\mathbf{U}_f$ of the safety index is a true upper bound; 2) there is always a feasible control that satisfies the constraint in (4).

We simulate the system for 2000 time steps with the safety index parameters designed above, and the evolution of $\mathbf{U}_f(x,u)$ (orange curves) and $\phi(f(x,u))$ (blue curves) is summarized in Figure 3. Overall, by ensuring $\mathbf{U}_f(x,u) < \max(\phi(x) - \eta, 0)$, UAISSA ensures $\phi(f(x,u)) < \max(\phi(x) - \eta, 0)$ along the simulations. As shown in Figure 4, with the safety index synthesized using (11), the nonempty set of safe control for all possible states are guaranteed. Furthermore, We conduct an ablation study on different discretization gap τ. The results are summarized in Figure 7 (Appendix D.5), where the gap between the upper bound of the safety index and the safety index decreases with smaller discretization gaps. This result validates our theoretical results as smaller discretization gaps result in smaller error bounds of the safety index. In practice, we believe a smaller discretization gap is beneficial to the performance of robot controllers since more accurate estimates of $\mathbf{U}_f(x,u)$ alleviate the performance drop caused by conservative safeguards. However, note that smaller discretization gaps also result in large datasets which may be computationally expensive for GP. It is a trade-off between lower computational cost and better performance.

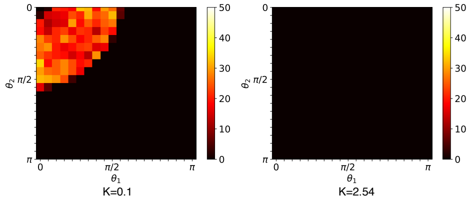

Figure 4: Distribution of states with infeasible safe controls when we optimize the safety index. Each grid in the graph corresponds to a position of joint angles (θ_1, θ_2). We sample 100 states at each position (with different velocities of joint angles). The color denotes how many states at this position has an empty set of safe control. The left shows that a randomly selected safety index ($k = 0.1$) results in empty set of safe control for many system states. The right shows that our synthesized safety index ($k = 2.54$) ensures that we can always find a feasible safe control.

5.2. Safety Gym

Scale to High-dimensional Environments One drawback of GP is that it scales very badly with the number of observations. To scale UAISSA to hign-dimensional environments, we propose to use deep Gaussian Process (Gal and Ghahramani, 2016) as an approximation of GP for dynamics learning. Note that the scalability comes with the price of losing theoretical safety guarantee because the uniform error bound of GP no longer holds when we use deep GP. Nevertheless, this section shows that UAISSA empirically achieves near zero-violation safety performance with deep GP.

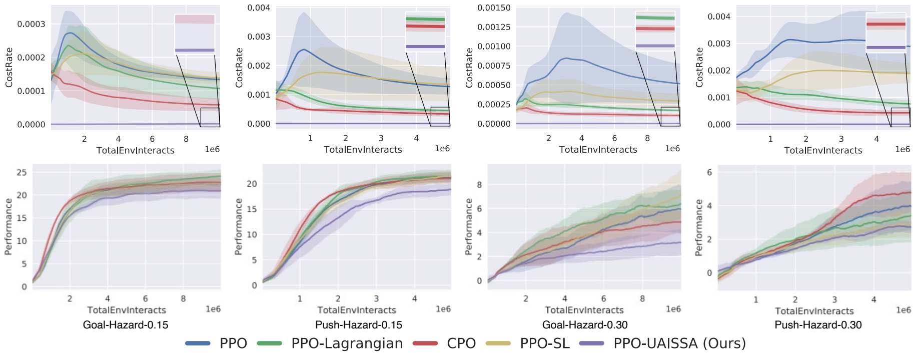

Figure 5: Cost rates and rewards of UAISSA and baselines on Safety Gym benchmarks with different tasks and sizes of the hazard over five random seeds.

Environment Setting To test how UAISSA safeguards RL policies in high-dimensional complex environments, we conduct experiments on the widely adopted benchmark of safe RL, Safety Gym (Ray et al., 2019). We evaluate UAISSA on two different control tasks (i.e., Goal-Hazard and Push-Hazard), where the environment settings are introduced in Appendix D.

Baseline Selection We choose PPO (Schulman et al., 2017) as the nomial RL algorithm and add UAISSA as a safeguard (namely PPO-UAISSA) on top of the nominal RL policy. We compare UAISSA with (i) PPO (Schulman et al., 2017) (stardard RL algorithm); (ii) PPO-Lagrangian (Chow et al., 2017) and CPO (Achiam et al., 2017) (safe RL algorithms); (iii) PPO-SL (Dalal et al., 2018) (RL with a different safeguard).

Policy Settings Detailed parameter settings are summarized in Table 2 (Appendix D). All the policies in our experiments use the default hyper-parameter settings hand-tuned by Safety Gym (Ray et al., 2019) except that we set the cost limit $= 0$ for PPO-Lagrangian and CPO since the goal is to achieve zero-violation performance.

Evaluation Results The evaluation results are shown in Figure 5, where PPO-UAISSA achieves near zero violation while gaining comparable rewards on both tasks. Note that the violations made by PPO-UAISSA are so few (nearly 1% of violations made by standard PPO), making it hard to observe in Figure 5. Such results align with our probabilistic safety guarantee given in Theorem 9. As for safe RL methods, both CPO and PPO-Lagrangian fail to achieve zero violation even with a cost limit of zero. PPO-SL proposed in (Dalal et al., 2018) also uses learned dynamics with an offline dataset, but PPO-SL failed to reduce safety violation due to (i) the assumption of linear cost functions is unrealistic in complex environments like MuJoCo (Todorov et al., 2012); (ii) the lack of quantification of the error bound from neural networks. More experiments details, comparison metrics and experimental results are summarized in Appendix D.6.

6. Conclusion

This paper presented a safe control framework with a learned dynamics model using Gaussian process. The proposed theory guarantees (i) the nonempty set of safe control for all states under control limits, and (ii) probabilistic forward invariance to the safe set. Simulation results on a robot arm and Safety Gym show near zero violation safety performance. One limitation of our work is that offline synthesis requires grid-based discretization of state space, which is computationally expensive for high-dimensional system. In the future work, we are going to investigate how to speed up offline synthesis, such as parallelization computation.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

Aaron D Ames, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs with application to adaptive cruise control. In *53rd IEEE Conference on Decision and Control*, pages 6271–6278. IEEE, 2014.

EF Beckenbach. On hölder's inequality. *Journal of Mathematical Analysis and Applications*, 15(1): 21–29, 1966.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30, 2017.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3387–3395, 2019a.

Yujiao Cheng, Weiye Zhao, Changliu Liu, and Masayoshi Tomizuka. Human motion prediction using semi-adaptable neural networks. In *2019 American Control Conference (ACC)*, pages 4884–4890. IEEE, 2019b.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

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

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *CoRR*, abs/1801.08757, 2018.

Sarah Dean, Stephen Tu, Nikolai Matni, and Benjamin Recht. Safely learning to control the constrained linear quadratic regulator. In *2019 American Control Conference (ACC)*, pages 5582–5588. IEEE, 2019.

James Ferlez, Mahmoud Elnaggar, Yasser Shoukry, and Cody Fleming. Shieldnn: A provably safe nn filter for unsafe nn controllers. *CoRR*, abs/2006.09564, 2020.

Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

Thomas Muirhead Flett. 2742. a mean value theorem. *The Mathematical Gazette*, 42(339):38–39, 1958.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Luis Gracia, Fabricio Garelli, and Antonio Sala. Reactive sliding-mode algorithm for collision avoidance in robotic systems. *IEEE Transactions on Control Systems Technology*, 21(6):2391–2399, 2013.

Suqin He, Weiye Zhao, Chuxiong Hu, Yu Zhu, and Changliu Liu. A hierarchical long short term safety framework for efficient robot manipulation under uncertainty. *Robotics and Computer-Integrated Manufacturing*, 82:102522, 2023a.

Tairan He, Weiye Zhao, and Changliu Liu. Autocost: Evolving intrinsic cost for zero-violation reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023b.

Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018.

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

Oussama Khatib. Real-time obstacle avoidance for manipulators and mobile robots. In *Autonomous robot vehicles*, pages 396–404. Springer, 1986.

Jason Kong, Mark Pfeiffer, Georg Schildbach, and Francesco Borrelli. Kinematic and dynamic vehicle models for autonomous driving control design. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 1094–1099. IEEE, 2015.

Eric J Kostelich and Thomas Schreiber. Noise reduction in chaotic time-series data: A survey of common methods. *Physical Review E*, 48(3):1752, 1993.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Uniform error bounds for gaussian process regression with application to safe control. *Advances in Neural Information Processing Systems*, 32, 2019.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Uniform error and posterior variance bounds for gaussian process regression with application to safe control. *arXiv preprint arXiv:2101.05328*, 2021.

Anjian Li, Somil Bansal, Georgios Giovanis, Varun Tolani, Claire Tomlin, and Mo Chen. Generating robust supervision for learning-based visual navigation using hamilton-jacobi reachability. In *Learning for Dynamics and Control*, pages 500–510. PMLR, 2020.

Changliu Liu and Masayoshi Tomizuka. Control in a safe set: Addressing safety in human-robot interactions. In *Dynamic Systems and Control Conference*, volume 46209, page V003T42A003. American Society of Mechanical Engineers, 2014.

Charles Noren, Weiye Zhao, and Changliu Liu. Safe adaptation with multiplicative uncertainties using robust safe set algorithm. *IFAC-PapersOnLine*, 54(20):360–365, 2021.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *CoRR*, abs/1910.01708, 2019.

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

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*, 550(7676):354–359, 2017.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

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*, 58(5):3250–3265, 2012.

Yifan Sun, Weiye Zhao, and Changliu Liu. Hybrid task constrained planner for robot manipulator in confined environment. *arXiv preprint arXiv:2304.09260*, 2023.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Akifumi Wachi, Yanan Sui, Yisong Yue, and Masahiro Ono. Safe exploration and optimization of constrained mdps using gaussian processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Tianhao Wei, Shucheng Kang, Weiye Zhao, and Changliu Liu. Persistently feasible robust safe control by safety index synthesis and convex semi-infinite programming. *IEEE Control Systems Letters*, 2022.

Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs TJ Spaan. Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence. AAAI Press, online*, 2021.

Wei-Ye Zhao, Xi-Ya Guan, Yang Liu, Xiaoming Zhao, and Jian Peng. Stochastic variance reduction for deep q-learning. *arXiv preprint arXiv:1905.08152*, 2019.

Wei-Ye Zhao, Suqin He, Chengtao Wen, and Changliu Liu. Contact-rich trajectory generation in confined environments using iterative convex optimization. In *Dynamic Systems and Control Conference*, volume 84287, page V002T31A002. American Society of Mechanical Engineers, 2020a.

Weiye Zhao, Liting Sun, Changliu Liu, and Masayoshi Tomizuka. Experimental evaluation of human motion prediction toward safe and efficient human robot collaboration. In *2020 American Control Conference (ACC)*, pages 4349–4354. IEEE, 2020b.

Weiye Zhao, Tairan He, and Changliu Liu. Model-free safe control for zero-violation reinforcement learning. In *5th Annual Conference on Robot Learning*, 2021.

Weiye Zhao, Suqin He, and Changliu Liu. Provably safe tolerance estimation for robot arms via sum-of-squares programming. *IEEE Control Systems Letters*, 6:3439–3444, 2022a.

Weiye Zhao, Tairan He, Tianhao Wei, Simin Liu, and Changliu Liu. Safety index synthesis via sum-of-squares programming. *arXiv preprint arXiv:2209.09134*, 2022b.

Weiye Zhao, Tairan He, Rui Chen, Tianhao Wei, and Changliu Liu. State-wise safe reinforcement learning: A survey. *arXiv preprint arXiv:2302.03122*, 2023.