2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Risk-Sensitive Inhibitory Control for Safe Reinforcement Learning

Armin Lederer[1], Erfaun Noorani[2], John S. Baras[2], Sandra Hirche[1]

Abstract— **Humans have the ability to deviate from their natural behavior when necessary, which is a cognitive process called response inhibition. Similar approaches have independently received increasing attention in recent years for ensuring the safety of control. Realized using control barrier functions or predictive safety filters, these approaches can effectively ensure the satisfaction of state constraints through an online adaptation of nominal control laws, e.g., obtained through reinforcement learning. While the focus of these realizations of inhibitory control has been on risk-neutral formulations, human studies have shown a tight link between response inhibition and risk attitude. Inspired by this insight, we propose a flexible, risk-sensitive method for inhibitory control. Our method is based on a risk-aware condition for value functions, which guarantees the satisfaction of state constraints. We propose a method for learning these value functions using common techniques from reinforcement learning and derive sufficient conditions for its success. By enforcing the derived safety conditions online using the learned value function, risk-sensitive inhibitory control is effectively achieved. The effectiveness of the developed control scheme is demonstrated in simulations.**

I. Introduction

Having a pause before responding is a mental technique that helps humans perceive, control, and manage our emotions. Human's ability to think before reacting, especially in difficult and complex situations, is a cognitive mechanism to keep our actions in check. This cognitive process is called inhibitory control, also known as response inhibition [1]. Response inhibition allows an individual to inhibit their prepotent (natural and habitual) responses in order to select a more appropriate (e.g. safer) behavior.

Independent from this foundation in psychology, response inhibition has become increasingly popular in learning-based control [2] and Reinforcement Learning (RL) [3] in recent years, where safety is a major concern [4]. The idea is to decouple optimality and safety by independently determining safe and optimal control laws. Before applying an optimal, but potentially unsafe control input to the real system, its safety is checked, such that a safe control input can be chosen instead [5]. Thereby, the prepotent optimal response is inhibited to guarantee the safety of the closed-loop system.

[1]A. Lederer and S. Hirche are with the Chair of Information-oriented Control (ITR), School of Computation, Information and Technology, Technical University of Munich, 80333 Munich, Germany. Emails: {armin.lederer, hirche}@tum.de.

[2]E. Noorani and J. Baras are with the Department of Electrical and Computer Engineering and the Institute for Systems Research (ISR) at the University of Maryland, College Park, MD, USA. Emails: {enoorani,baras}@umd.edu.

Research partially supported by ONR grant N00014-17-1-2622, by a grant from the Army Research Lab, by the Clark Foundation, and by the European Research Council (ERC) Consolidator Grant "Safe data-driven control for human-centric systems (CO-MAN)" under grant agreement number 864686.

The challenge of this approach lies in finding safe policies and efficient methods to determine the safety of a control input online. When the dynamics of the systems are known to exhibit a control-affine structure, control barrier functions (CBF) can be effectively employed to address this challenge [6]. Since their analytical derivation for more flexible classes of dynamical systems is difficult at best, techniques from model predictive control have become popular for computing safe backup strategies online [7], [8]. While such predictive safety filters provide a conceptionally flexible approach for realizing inhibitory control, they generally suffer from high computational complexity. This limitation can be mitigated by combining ideas from reachability analysis [9] or optimal control [10] with reinforcement learning techniques to learn safety conditions and safe control laws offline, such that resource-demanding computations can be avoided during the application of the inhibited control law.

While these approaches allow the seemingly straightforward realization of inhibitory control for ensuring the safety of real-world systems, they do not consider the risk of losing safety due to uncertainty arising from approximate system models and process noise. This is in strong contrast to humans, for which psychological studies have shown a critical link between response inhibition and an individual's risk attitude (willingness to take risk or not) [11]. When inhibitory control is implemented in technical systems through analytically derived safety conditions such as CBFs, this risk-sensitivity can be easily achieved by reformulating standard conditions using risk measures [12]. However, the extension to flexible approaches for constructing safety conditions, e.g., using RL techniques remains an open problem.

We address this problem of realizing inhibitory control with risk-awareness similar to humans for ensuring the safety of a wide class of systems via the following contributions:

- **Risk-sensitive safety conditions:** To ensure the probabilistic satisfaction of state constraints, we introduce cost functions allowing us to express safety via risk-sensitive conditions on the cumulative cost along system trajectories. These conditions reveal an intuitive relationship between risk-aversion and safety probability.
- **Safe policies and value functions through RL:** Based on these results, we develop an approach for determining safe policies and corresponding safety value functions using common techniques from reinforcement learning. The success of the proposed approach is shown to be guaranteed under weak assumptions relating to the controllability properties of the system dynamics.
- **Inhibitory control through safety filters:** By enforcing the satisfaction of the derived safety conditions with

the learned value function online, we obtain a risk-sensitive safety filter. Moreover, we prove it to inherit probabilistic safety guarantees from the safe policy obtained through RL.

The remainder of this paper is structured as follows. In Section II, the problem of rendering a given policy safe with respect to state constraints using safety filters is formalized. Our approach for realizing response inhibition in control using risk-sensitive safety filters is derived in Section III. In Section IV, the effectiveness of the proposed safety filter is demonstrated, before the paper is concluded in Section V.

II. Problem Statement

We consider a discrete-time dynamical system[1]

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{\omega}_k), \tag{1}$$

where $\boldsymbol{x}_k \in \mathbb{X} \subset \mathbb{R}^{d_x}$ are states, $\boldsymbol{u}_k \in \mathbb{U} \subset \mathbb{R}^{d_u}$ are control inputs, $\boldsymbol{\omega}_k \in \Omega \subset \mathbb{R}^{d_\omega}$, $\boldsymbol{\omega}_k \sim \rho(\boldsymbol{x}_k)$ is independent process noise drawn from a potentially state-dependent distribution $\rho(\boldsymbol{x}_k)$ with zero mean, and $\boldsymbol{f} : \mathbb{X} \times \mathbb{U} \times \Omega \to \mathbb{X}$ denotes an unknown, continuous transition function. We assume that a nominal, potentially unsafe policy $\boldsymbol{\pi}^* : \mathbb{X} \to \mathbb{U}$ is given, which can be obtained, e.g., using standard reinforcement learning techniques [3].

The goal is to render the nominal policy safe using inhibitory control of the form

$$\boldsymbol{\pi}^*_{\text{safe}}(\boldsymbol{x}) = \underset{\boldsymbol{u}\in\mathbb{U}}{\arg\min} \|\boldsymbol{\pi}^*(\boldsymbol{x}) - \boldsymbol{u}\| \tag{2a}$$

$$\text{such that } \boldsymbol{u} \text{ is safe.} \tag{2b}$$

In this response inhibition, our notion of safety follows the common principle of classifying the state space $\mathbb{X}$ into a safe region $\mathbb{X}_{\text{safe}} \subset \mathbb{X}$ and an unsafe region $\mathbb{X}_{\text{unsafe}} = \mathbb{X} \setminus \mathbb{X}_{\text{safe}}$. For example, the safe set $\mathbb{X}_{\text{safe}}$ can represent the joint angles for which self-collisions of a robotic manipulator are excluded. Due to the process noise $\boldsymbol{\omega}$ with a potentially unbounded probability distribution, it is generally not possible to deterministically ensure that the system never enters the unsafe state space $\mathbb{X}_{\text{unsafe}}$. Therefore, we define safety probabilistically through the following form of forward invariance.

Definition 1: A policy $\boldsymbol{\pi}(\cdot)$ is called δ-safe if there exists a subset $\mathbb{V} \subseteq \mathbb{X}_{\text{safe}}$ such that $\mathcal{P}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}) \in \mathbb{V}) \geq 1-\delta$ for all $\boldsymbol{x} \in \mathbb{V}$.

Since Definition 1 requires a form of forward invariance of $\mathbb{V}$, it immediately induces guarantees for all states along a K-step trajectories of the form

$$\mathcal{P}(\boldsymbol{x}_k \in \mathbb{V},\ \forall k = 1 \ldots, K) \geq (1-\delta)^K, \tag{3}$$

where $\boldsymbol{x}_k$ is defined through iterative application of (1). Hence, the considered notion of safety in this paper is stronger than merely requiring the next state to lie in the safe subset, i.e., $\mathcal{P}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}) \in \mathbb{X}_{\text{safe}}) \geq 1 - \delta$.

[1]Notation: Lower/upper case bold symbols denote vectors/matrices, blackboard bold letters denote sets, $\mathbb{R}_+/\mathbb{R}_{0,+}$ all real positive/non-negative numbers, $\|\cdot\|$ the Euclidean norm, $\mathbb{E}_x[\cdot]$ the expectation with respect to the distribution of x, and $\mathbb{P}(\cdot)$ the probability.

Based on the definition of δ-safety, we consider the problem of deriving a tractable safety condition (2b) for inhibitory control, which is guaranteed to be feasible for some risk-aversion as measured through δ. Since we assume the transition function $\boldsymbol{f}$ is unknown, solving this problem is generally impossible without any further assumptions. Therefore, we require the availability of a probabilistic model in the form of a distribution over functions as formalized in the following.

Assumption 1: A probability distribution $\mathcal{F}$ over potential dynamics $\boldsymbol{f}$ is known, i.e., $\boldsymbol{f} \sim \mathcal{F}$.

In practice, suitable distributions over functions $\mathcal{F}$ can be straightforwardly obtained using Bayes' theorem, e.g., through Gaussian process regression [13]. Moreover, approximate distributions can be learned using deep ensembles [14]. Therefore, this assumption is not restrictive in practice.

III. Risk-Sensitive Inhibitory Control

Even with the knowledge of $\mathcal{F}$, determining a safety condition (2b) is a challenging problem since we generally do not know which subset $\mathbb{V}$ is suitable for Definition 1. Here, we follow the ideas of [10] and employ RL techniques to define these subsets through a value function. For this purpose, we first show how state constraints can be expressed through risk-sensitive cost conditions in Section III-A. After deriving these safety conditions, in Section III-B, we address the problem of learning a separate, so-called backup policy whose pure focus lies on ensuring safety. Based on this policy, a risk-sensitive safety filter for realizing inhibitory control in reinforcement learning is finally presented in Section III-C.

A. State Constraints as Risk-Sensitive Cost Conditions

In order to express state constraints through risk-sensitive cost conditions, we define the expected cumulative cost for a policy $\boldsymbol{\pi}(\cdot)$ as

$$V_{\boldsymbol{\pi}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{f},\boldsymbol{\omega}}\left[\sum_{k=0}^{\infty} \gamma^k c(\boldsymbol{x}_k)\right], \tag{4}$$

where $c : \mathbb{R}^{d_x} \to \mathbb{R}_{0,+}$ denotes an immediate cost, $\gamma \in (0, 1)$ is a discount factor, and $\boldsymbol{x}_k$ is defined through the iterative application of (1) with $\boldsymbol{x}_0 = \boldsymbol{x}$ and $\boldsymbol{u}_k = \boldsymbol{\pi}(\boldsymbol{x}_k)$. If the immediate cost $c(\cdot)$ can be used as an indicator of the unsafe subset $\mathbb{X}_{\text{unsafe}}$, there exists a sub-level set of $V_{\boldsymbol{\pi}}(\cdot)$ contained in $\mathbb{X}_{\text{safe}}$, as guaranteed by the following lemma.

Lemma 1 ([10]): Assume there exists a constant $\hat{c} \in \mathbb{R}_+$, such that the cost $c : \mathbb{R}^{d_x} \to \mathbb{R}_{0,+}$ satisfies

$$c(\boldsymbol{x}) \geq \hat{c} \quad \forall \boldsymbol{x} \in \mathbb{X}_{\text{unsafe}}. \tag{5}$$

Then, there exists a constant $\bar{\xi} \in \mathbb{R}_+$, such that the intersection between the sub-level set $\mathbb{V}^{\bar{\xi}}_{\boldsymbol{\pi}} = \{\boldsymbol{x} \in \mathbb{X} : V_{\boldsymbol{\pi}}(\boldsymbol{x}) \leq \bar{\xi}\}$ and $\mathbb{X}_{\text{unssafe}}$ is empty, i.e., $\mathbb{V}^{\bar{\xi}}_{\boldsymbol{\pi}} \cap \mathbb{X}_{\text{unsafe}} = \emptyset$.

Based on this lemma, we can choose any sub-level set $\mathbb{V}^{\xi}_{\boldsymbol{\pi}}$ with $\xi \leq \bar{\xi}$ for showing δ-safety as introduced in Definition 1. As discussed in [10], the immediate cost $c(\cdot)$ for defining sub-level sets $\mathbb{V}^{\xi}_{\boldsymbol{\pi}}$ can be selected relatively freely, such that simple choices as the indicator function are applicable in principle. However, this choice does not provide informative gradients, which complicates the learning process. Therefore,

other cost functions such as rectified linear unit functions generally need to be considered, even though they can potentially lead to more conservative approximations of the safe set $\mathbb{X}_{\text{safe}}$. To obtain suitable values for $\bar{\xi}$, different approaches can be used. For example, potentially conservative closed-form expressions can be employed as shown in [10]. Moreover, optimal solutions can be found by formulating the search for $\bar{\xi}$ as a robust optimization problem, which can be solved numerically. Therefore, it only remains to derive conditions that ensure the state stays in $\mathbb{V}_{\boldsymbol{\pi}}^{\xi}$ after a transition. While this could be achieved using a probabilistic "worst case" consideration as shown in [10], this approach yields a computationally challenging min-max problem for unknown system dynamics. Therefore, we follow a fully probabilistic approach by introducing the risk operator [15]

$$\mathbb{R}_{\beta}[C] = \frac{1}{\beta} \log\left(\mathbb{E}\left[\exp\left(\beta C\right)\right]\right) \tag{6}$$

for an arbitrary random variable C and risk parameter $\beta \in \mathbb{R}_{+}$. This operator allows the derivation of a computationally efficient condition for ensuring δ-safety as shown in the following proposition.

Proposition 1: Consider a cost function $c(\cdot)$ satisfying (5). If there exist constants $\xi, \beta \in \mathbb{R}_{+}$ with $\xi < \bar{\xi}$ such that

$$\mathbb{R}_{\beta}[V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+})] \leq \xi, \qquad \forall \boldsymbol{x} \in \mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}} \tag{7}$$

holds for $\boldsymbol{x}^{+}=\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega})$, then, $\boldsymbol{\pi}(\cdot)$ is δ-safe on $\mathbb{V}_{\boldsymbol{\pi}}^{\xi}$ with

$$\delta = \exp\left(\beta\left(\xi - \bar{\xi}\right)\right). \tag{8}$$

Proof: Due to Lemma 1, we can bound the probability of leaving $\mathbb{X}_{\text{safe}}$ by the probability of leaving $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$. Therefore, it is sufficient to derive an upper bound for the probability

$$\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+}) \geq \bar{\xi}\right) = \mathbb{E}_{\boldsymbol{x}^{+}}\left[I_{\bar{\xi}}(V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+})\right], \tag{9}$$

where the indicator function $I_{\bar{\xi}} : \mathbb{R} \to \{0, 1\}$ is defined as

$$I_{\bar{\xi}}(V) = \begin{cases} 0 & \text{if } V \leq \bar{\xi} \\ 1 & \text{if } V > \bar{\xi}. \end{cases} \tag{10}$$

Note that $V_{\boldsymbol{\pi}}(\cdot)$ is a deterministic function, such that the expectation affects only the random variable $\boldsymbol{x}^{+}$ in (9). Moreover, β is positive, $\exp(0) = 1$ and the exponential function is strictly increasing and positive. Therefore, we can bound the indicator function through the exponential expression

$$I_{\bar{\xi}}(V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+})) \leq \exp\left(\beta\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+}) - \bar{\xi}\right)\right) \tag{11}$$

due to the positivity of β. By taking the expectation of both sides, this inequality immediately leads to

$$\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+}) \geq \bar{\xi}\right) \leq \mathbb{E}_{\boldsymbol{x}^{+}}\left[\exp\left(\beta V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+})\right)\right] \exp(-\beta\bar{\xi}). \tag{12}$$

Due to the definition of the risk operator in (6), we can simplify the right side of this inequality to obtain

$$\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+}) \geq \bar{\xi}\right) \leq \exp\left(\beta\left(\mathbb{R}_{\beta}[V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+})] - \bar{\xi}\right)\right). \tag{13}$$

Since $\mathbb{R}_{\beta}[V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+})] \leq \xi$ is ensured by (7), we have $\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^{+}) \geq \bar{\xi}\right) \leq \delta$ with δ defined in (8). ∎

This result provides a straightforward condition, which merely requires the evaluation of the risk operator and the computation of the cumulative cost, which is a problem commonly encountered in reinforcement learning. Moreover, it offers a simple expression for the probability of safety, such that it can easily be computed in practice.

Remark 1: Since the probability of a safety violation δ guaranteed by Proposition 1 only depends on three parameters, it allows an intuitive interpretation:

- The difference between ξ and $\bar{\xi}$ can be interpreted as a safety margin since it requires the dynamics to be contractive on the set $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}} \setminus \mathbb{V}_{\boldsymbol{\pi}}^{\xi}$ towards $\mathbb{V}_{\boldsymbol{\pi}}^{\xi}$. The larger this safety margin, the more contractive is the behavior at the boundary of $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ and consequently, it becomes more unlikely that the state reaches $\mathbb{X} \setminus \mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$.
- The parameter β reflects the risk-sensitivity of the safety condition (7). A large value of β corresponds to a high risk-aversion since it causes the tails of the noise distribution ρ and the function distribution $\mathcal{F}$ to have a larger effect on the left side of (7). In the extreme case of $\beta \to \infty$, this leads to (7) corresponding to a condition on the worst case realization of $\boldsymbol{\omega}_k$ and $\boldsymbol{f}(\cdot)$ [15]. This increasing risk-aversion with growing β is intuitively accompanied by an increase in the probability of safety.

B. Safe backup Policies via Reinforcement Learning

While Section III-A describes an approach for obtaining the probability of safety for a given policy, it does not address the problem of determining a safe policy. In this section, we show that this problem can be solved using standard reinforcement learning techniques through the following minimization problem

$$\boldsymbol{\pi}_{\text{safe}} = \underset{\boldsymbol{\pi} \in \Pi}{\arg\min}\, \mathbb{E}_{\boldsymbol{x}}\left[V_{\boldsymbol{\pi}}(\boldsymbol{x})\right]. \tag{14}$$

Even though this optimization problem does not involve the risk operator $\mathbb{R}_{\beta}[\cdot]$, its solution $\boldsymbol{\pi}_{\text{safe}}$ is guaranteed to satisfy the conditions of Proposition 1 under weak assumptions. This is demonstrated by the subsequent theorem. The proof follows after a discussion of the assumptions.

Theorem 1: Consider a cost function $c(\cdot)$ satisfying (5) and assume that there exist a policy $\tilde{\boldsymbol{\pi}}(\cdot)$ and constants $\theta_1, \theta_2 \in \mathbb{R}_{+}$ with $\theta_1 < {}^{1}/_{(1-\gamma)}$ such that

$$V_{\boldsymbol{\pi}}(\boldsymbol{x}) \leq \theta_1 c(\boldsymbol{x}) + \theta_2, \quad \forall \boldsymbol{x} \in \mathbb{X} \tag{15}$$

is satisfied. Moreover, assume there exist constants $\theta_3, \theta_4 \in \mathbb{R}_{0,+}$ such that

$$V_{\boldsymbol{\pi}}(\boldsymbol{x}) \geq \theta_3 c(\boldsymbol{x}) + \theta_4, \quad \forall \boldsymbol{x} \in \mathbb{X} \tag{16}$$

holds for all policies $\boldsymbol{\pi}(\cdot)$. If

$$\hat{c} > \frac{\theta_2}{\theta_3(\theta_1(\gamma - 1) + 1)} - \frac{\theta_4}{\theta_3} \tag{17}$$

holds, then, the policy (14) is δ^{*}-safe on $\mathbb{V}_{\xi^{*}}$ with $\delta^{*} = \exp\left(\beta^{*}\left(\xi^{*} - \bar{\xi}\right)\right)$, where

$$\beta^{*}, \xi^{*} = \underset{\beta \in \mathbb{R}_{+}, \xi \in \mathbb{R}_{+}}{\arg\min}\ \exp\left(\beta\left(\xi - \bar{\xi}\right)\right) \tag{18a}$$

$$\text{s.t. } \xi < \bar{\xi} \tag{18b}$$

$$(7) \text{ holds.} \tag{18c}$$

Discussion: While large values for θ_3 and θ_4 in (16) are generally beneficial for admitting larger values of $\hat{c}$ in (17), it is always possible to trivially choose $\theta_3 = 1$, $\theta_4 = 0$ due to non-negativity of $c(\cdot)$. Condition (15) essentially requires a sufficiently fast decay of the immediate costs $c(\boldsymbol{x}_k)$ along trajectories for some policy $\tilde{\boldsymbol{\pi}}(\cdot)$. This decay can be achieved if, e.g., variants of exponential controllability hold [16]. Since merely the existence of a policy $\tilde{\boldsymbol{\pi}}(\cdot)$ satisfying (15) is necessary, this admits the derivation of the constants θ_1 and θ_2 via properties such as exponential controllability [16]. Therefore, the assumptions of Theorem 1 are not restrictive in practice.

Note that the required lower bound (16) for all possible cost functions $V_{\boldsymbol{\pi}}(\cdot)$ is only necessary because of the offset θ_2, which leads to a lower bound for the admissible values of $\bar{\xi}$. Since the admissible value $\bar{\xi}$ depends directly on the cost function $V_{\boldsymbol{\pi}}(\cdot)$, it indirectly depends on the policy $\boldsymbol{\pi}(\cdot)$. Therefore, $V_{\tilde{\boldsymbol{\pi}}}(\cdot)$ and $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$ potentially admit different values for $\bar{\xi}$, such that general constraints cannot be posed on $\bar{\xi}$. This issue is resolved by (16), which establishes a direct relationship between $\hat{c}$ and $\bar{\xi}$ for all possible cost functions $V_{\boldsymbol{\pi}}(\cdot)$ and thereby leads to the lower bound (17). If no offset exists, i.e., $\theta_2 = \theta_4 = 0$, it can be easily seen that $\hat{c} > 0$ must be satisfied. This is the trivial lower bound for $\hat{c}$ due to the assumed non-negativity of immediate cost functions $c(\cdot)$. Therefore, the offset θ_2 is the only reason for the restriction of the admissible threshold $\hat{c}$.

Proof: In order to prove Theorem 1, we first show that a risk-neutral variant of condition (7) guarantees the existence of parameters ξ and β satisfying the requirements of Proposition 1.

Lemma 2: Assume that

$$\mathbb{E}_{\boldsymbol{x}^+}[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] \leq \tilde{\xi}, \qquad \forall \boldsymbol{x} \in \mathbb{V}_{\bar{\boldsymbol{\xi}}} \tag{19}$$

holds for some constant $\tilde{\xi} < \bar{\xi}$. Then, there exist constants $\beta \in \mathbb{R}_+$ and $\xi < \bar{\xi}$ such that (7) is satisfied.

Proof: By the Taylor series expansion of the exponential function, we have

$$\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] = \frac{1}{\beta}\log\left(1+\beta\mathbb{E}_{\boldsymbol{x}^+}\left[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)\right]+\frac{\beta^2}{2}\mathbb{E}_{\boldsymbol{x}^+}\left[V_{\boldsymbol{\pi}}^2(\boldsymbol{x}^+)\right]+\dots\right). \tag{20}$$

From the premise of the lemma, it follows that

$$\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] \leq \frac{1}{\beta}\log\left(1+\beta\tilde{\xi}+\frac{\beta^2}{2}\mathbb{E}_{\boldsymbol{x}^+}\left[V_{\boldsymbol{\pi}}^2(\boldsymbol{x}^+)\right]+\dots\right). \tag{21}$$

Since $\log(1+a) < a$ for $a \in \mathbb{R}_+$ and by noting the positivity of $V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)$ and the risk-aversion parameter β, we have

$$\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] < \tilde{\xi} + \beta\left(\frac{1}{2}\mathbb{E}_{\boldsymbol{x}^+}\left[V_{\boldsymbol{\pi}}^2(\boldsymbol{x}^+)\right]+\dots\right). \tag{22}$$

Since the second summand can be brought arbitrarily close to 0 by choosing a sufficiently small β, there exists a β such that the right side of (22) is smaller than $\bar{\xi}$, which concludes the proof. ∎

The key idea behind this result is that (7) converges to (19) for $\beta \to 0$. Therefore, it is sufficient to determine a policy $\boldsymbol{\pi}$, which satisfies the risk-neutral condition (19), for ensuring (7) with a suitably small value of $\beta \in \mathbb{R}_+$.

Although (19) is a risk-neutral condition, it exhibits an expectation with respect to the next state $\boldsymbol{x}^+$. Therefore, it does not directly enable the applicability of standard RL techniques and consequently, it does not coincide with the acquisition function considered in the definition of the safe policy (14). In order to overcome this issue, we exploit (15) to relate $\mathbb{E}_{\boldsymbol{x}^+}[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)]$ to $V_{\boldsymbol{\pi}}(\boldsymbol{x})$. This is achieved using the following lemma.

Lemma 3: Assume that there exist $\theta_1, \theta_2 \in \mathbb{R}_+$ with $\theta_1 < {}^1\!/\!_{(1-\gamma)}$ such that (15) is satisfied. Then, it holds that

$$\mathbb{E}_{\boldsymbol{x}^+}[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] - V_{\boldsymbol{\pi}}(\boldsymbol{x}) \leq \frac{\theta_1-\theta_1\gamma-1}{\theta_1\gamma}V_{\boldsymbol{\pi}}(\boldsymbol{x}) + \frac{\theta_2}{\gamma\theta_1}. \tag{23}$$

Proof: By solving Bellman's identity

$$V_{\boldsymbol{\pi}}(\boldsymbol{x}) = c(\boldsymbol{x}) + \gamma\mathbb{E}_{\boldsymbol{x}^+}\left[V_{\boldsymbol{\pi}}(\boldsymbol{x}')\right], \tag{24}$$

for $\mathbb{E}_{\boldsymbol{x}^+}[V_{\boldsymbol{\pi}}(\boldsymbol{x}')]$, we can express $\Delta V_{\boldsymbol{\pi}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}^+}[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] - V_{\boldsymbol{\pi}}(\boldsymbol{x})$ as

$$\Delta V_{\boldsymbol{\pi}}(\boldsymbol{x}) = \frac{1}{\gamma}(-c(\boldsymbol{x}) + (1-\gamma)V_{\boldsymbol{\pi}}(\boldsymbol{x})). \tag{25}$$

Due to (15), we have

$$c(\boldsymbol{x}) \geq \frac{V_{\boldsymbol{\pi}}(\boldsymbol{x}) - \theta_2}{\theta_1}, \tag{26}$$

which allows us to bound (25) by

$$\Delta V_{\boldsymbol{\pi}}(\boldsymbol{x}) \leq \frac{1}{\gamma}\left(-\frac{V_{\boldsymbol{\pi}}(\boldsymbol{x}) - \theta_2}{\theta_1} + (1-\gamma)V_{\boldsymbol{\pi}}(\boldsymbol{x})\right). \tag{27}$$

Rearranging the terms on the right side finally yields

$$\Delta V_{\boldsymbol{\pi}} \leq \frac{\theta_1-\theta_1\gamma-1}{\theta_1\gamma}V_{\boldsymbol{\pi}}(\boldsymbol{x}) + \frac{\theta_2}{\gamma\theta_1}, \tag{28}$$

where ${}^{(\theta_1-\theta_1\gamma-1)}\!/\!_{\theta_1\gamma}$ is guaranteed to be negative since $\theta_1 < {}^1\!/\!_{(1-\gamma)}$ is assumed. ∎

Lemma 3 ensures that the minimization of $V_{\boldsymbol{\pi}}(\boldsymbol{x})$ also reduces $\mathbb{E}_{\boldsymbol{x}^+}[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)]$. This directly allows proving Theorem 1 in combination with Lemma 2 as shown in the following.

Proof of Theorem 1: It is straightforward to see that optimizing with respect to the expectation over $\boldsymbol{x}$ yields identical policies $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ as the point-wise optimum $\pi_{\boldsymbol{x}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{\pi}\in\Pi} V_{\boldsymbol{\pi}}(\boldsymbol{x})$ for a given $\boldsymbol{x}$ and a continuous transition function $\boldsymbol{f}(\cdot,\cdot,\cdot)$. Due to optimality of $\boldsymbol{\pi}_{\boldsymbol{x}}(\cdot)$, we additionally have the inequality $V_{\boldsymbol{x}}(\boldsymbol{x}) \leq V_{\tilde{\boldsymbol{\pi}}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{X}$. Therefore, it follows from Lemma 3 that

$$\mathbb{E}[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x}^+)] \leq \frac{1}{\gamma}\left(1-\frac{1}{\theta_1}\right)V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x}) + \frac{\theta_2}{\gamma\theta_1}. \tag{29}$$

Since the right side of (29) is linear in $V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x})$, the maximum inside $\mathbb{V}_{\bar{\xi}}$ is achieved for $V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x}) = \bar{\xi}$. Therefore, we obtain the inequality

$$\bar{\xi} > \frac{1}{\gamma}\left(1-\frac{1}{\theta_1}\right)\bar{\xi} + \frac{\theta_2}{\gamma\theta_1} \tag{30}$$

since Lemma 2 requires $\mathbb{E}[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x}^+)] \leq \xi < \bar{\xi}$. Solving

Algorithm 1: Safe RL using Risk-Sensitive Filters

```
/* Solve (32)                                          */
while optimization not converged do
    Sample function f̂(·) ~ F
    Roll-out policy π*(·) on f̂(·)
    Update π*(·) using gathered system data
/* Solve (14)                                          */
while optimization not converged do
    Sample function f̂(·) ~ F
    Roll-out policy π_safe(·) on f̂(·)
    Update π_safe(·) using gathered system data
/* Safe roll-out via online optimization (33) */
Apply π*_safe(·) to unknown system f(·)
```

for $\bar{\xi}$ and noting that $\bar{\xi} = \theta_3 \hat{c} + \theta_4$ due to (16) yields

$$\theta_3 \hat{c} + \theta_4 > \frac{\theta_2}{\theta_1(\gamma - 1) + 1}. \tag{31}$$

It is straightforward to see that (17) guarantees the satisfaction of this inequality, such that Lemma 2 and Proposition 1 ensure that (18) is feasible and results in a probability $\delta^* < 1$. This immediately implies δ^*-safety of $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ and thereby concludes the proof. ∎

C. Risk-Sensitive Inhibitory Control for Safe Roll-outs

Based on the safe policy $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ obtained using (14), we propose a risk-sensitive inhibitory control strategy for enabling safe RL as outlined in Alg. 1. For this purpose, we first obtain an optimal, potentially unsafe policy by solving the optimization problem

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi} \in \Pi} \mathbb{E}_{\boldsymbol{f}, \boldsymbol{\omega}, \boldsymbol{x}_0} \left[\sum_{k=0}^{\infty} \gamma^k r(\boldsymbol{x}_k, \boldsymbol{\pi}(\boldsymbol{x}_k)) \right], \tag{32}$$

where $r : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_{0,+}$ denotes a reward function and $\boldsymbol{x}_k$ is defined through the iterative application of (1) with $\boldsymbol{x}_0 = \boldsymbol{x}$ and $\boldsymbol{u}_k = \boldsymbol{\pi}(\boldsymbol{x}_k)$. This problem can be solved using standard off-policy reinforcement learning algorithms such as soft actor-critic reinforcement learning [17]. Afterward, a safe backup policy $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ is computed by solving (14), which can be straightforwardly achieved using standard off-policy reinforcement learning techniques. Finally, we apply the policy to the true system (1). For this roll-out, we employ the risk-sensitive filter

$$\boldsymbol{\pi}^*_{\text{safe}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u} \in \mathbb{U}} \|\boldsymbol{\pi}^*(\boldsymbol{x}) - \boldsymbol{u}\| \tag{33a}$$

$$\text{s.t. } \mathbb{R}_\beta[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\omega}))] \leq \xi^* \tag{33b}$$

which makes use of the safe backup policy $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ through the cost function $V_{\boldsymbol{\pi}_{\text{safe}}}$ and minimally adjusts the policy $\boldsymbol{\pi}^*(\cdot)$ such that the safety condition (7) is satisfied.

Due to the safety filter (33), the state constraints $\mathbb{X}_{\text{safe}}$ can straightforwardly be considered in Alg. 1. In fact, δ-safety of $\boldsymbol{\pi}^*_{\text{safe}}(\cdot)$ is directly inherited from the safe backup policy $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ as shown in the following theorem.

Theorem 2: Consider a cost function $c(\cdot)$ satisfying (5) and a threshold $\hat{c}$, for which (17) holds. Moreover, assume that there exists a policy $\tilde{\boldsymbol{\pi}}(\cdot)$ satisfying (15) with $\theta_1 < {}^1/_{(1-\gamma)}$ for all $\boldsymbol{x} \in \mathbb{X}_{\text{safe}}$. Then, the safety filtered policy (33) is δ^*-safe on $\mathbb{V}^{\xi^*}_{\boldsymbol{\pi}_{\text{safe}}}$ with $\delta^* = \exp\left(\beta^* \left(\xi^* - \bar{\xi}\right)\right)$, where β^* and ξ^* are defined in (18).

Proof: Due to Theorem 1, $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ defined in (14) satisfies (33b). Thus, the optimization problem (33) is guaranteed to be feasible for all states $\boldsymbol{x} \in \mathbb{V}^{\xi^*}_{\boldsymbol{\pi}_{\text{safe}}}$ with the trivial solution $\boldsymbol{u} = \boldsymbol{\pi}_{\text{safe}}(\boldsymbol{x})$. Finally, δ^*-safety directly follows from Proposition 1. ∎

While this theorem employs the optimal parameters β^* and ξ^*, it immediately follows from the proof of Theorem 1 that for every value ξ with $\xi^* \leq \xi < \bar{\xi}$, there exists a $\beta \in \mathbb{R}_+$ satisfying (18b). Therefore, δ-safety on $\mathbb{V}_\xi \supset \mathbb{V}_{\xi^*}$ with $\delta > \delta^*$ can be straightforwardly ensured in practice by choosing a sufficiently large value $\xi < \bar{\xi}$ and a suitably small value $\beta \in \mathbb{R}_+$.

Remark 2: When β becomes larger, the control becomes more pessimistic, and therefore, the probability of safety generally increases. However, there exists a critical value at which the safety constraint (33b) becomes infeasible for all $\xi < \bar{\xi}$. That is, the control becomes too phobic to act. This resembles a well-known behavior in risk-sensitive control and RL commonly referred to as neurotic breakdown [18].

IV. Simulations

In this section, we evaluate the proposed risk-sensitive inhibitory control approach, described in Alg. 1, using the popular Mujoco Half-Cheetah environment [19]. The Half-Cheetah is a planar model of a large, cat-like robot with 6 actuated joints. The main goal is to maximize the robot's walking velocity with the least control effort possible, which is encoded in the default reward function. We consider the default model parameters for the Cheetah robot, but assume a body mass perturbed by a Gaussian distributed random variable with 0 mean and standard deviation 0.1. In order to obtain a challenging safety condition, we set optimality and safety in a direct conflict similar as in [10] by constraining the velocity to $v \leq v_{\text{crit}}$, $v_{\text{crit}} = 2$. As cost function for the computation of the safe policy (14), $c(\boldsymbol{x}) = v - \underline{v}$ is employed with threshold $\hat{c} = 2 - \underline{v}$, where $\underline{v} = -10$ denotes the considered minimum velocity of the Half-Cheetah robot. This cost function encourages the robot to run with a negative velocity, such that the distance to the safety threshold velocity v_{crit} is maximized. Note that the subtraction of $\underline{v}$ is necessary to ensure the non-negativity of the cost $c(\cdot)$ assumed in our derivations, but it merely causes a constant off-set in the cumulative cost $V_{\boldsymbol{\pi}}(\cdot)$.

The optimal and safe policies are obtained using the Soft-Actor Critic (SAC) algorithm [17] with 400 training iterations each with 1000 time steps and the hyper-parameters provided by [20]. For computing the expectations over dynamics $\boldsymbol{f}(\cdot)$ in (4) and (32), we randomly sample 10 body masses, such that we can use the corresponding sample environments to empirically approximate all necessary expected values. The risk-sensitive safety filter (33) is implemented using the cross-entropy method [21] with 5 iterations per time step and 10 particles. The safety constraints are considered in an augmented objective function using fixed Lagrange

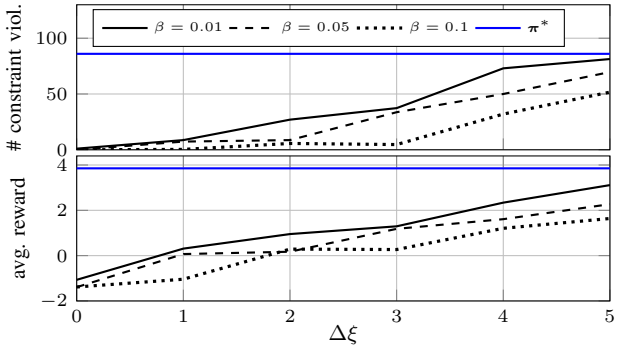

Fig. 1. Number of constraint violations and average rewards in dependency on the safety constraint threshold $\xi = 521 + \Delta\xi$ and the risk-sensitivity β. Reducing β and increasing ξ have a similar effect of admitting more risky behavior in the response inhibition, such that the number of constraint violations and the average reward increase.

multipliers, such that they are effectively enforced using soft constraints to allow recovery after constraint violations. The risk operator $\mathbb{R}_\beta[\cdot]$ is approximated through 100 sample environments. For each parameter combination (ξ, β), 100 time steps are simulated and 3 random seeds are averaged.

The resulting numbers of constraint violations and the average reward for different values of β and ξ are depicted in Fig. 1. We can observe that increasing ξ has exactly the expected effect of loosening the safety constraint by admitting higher velocities v, such that the probability of safety decreases and more constraint violations can be observed. At the same time, this allows a higher robot velocity, which in turn causes an increasing average reward. A similar effect can be observed with the risk parameter β due to the considered state-independent model uncertainty. When β is increased, the conservatism of the safety filter increases. This leads to a lower number of constraint violations, but the average reward also reduces. Therefore, the parameters ξ and β exhibit the impact on the probability of safety as discussed in Remark 1. Note that the risk-inhibition with the considered soft constraint formulation has a clearly visible effect on the average robot velocity, even when it does not manage to enforce the safety constraints. This can be observed in a comparison with the optimal policy $\boldsymbol{\pi}^*(\cdot)$, which achieves a significantly higher reward with a similar number of constraint violations for large values of ξ and small β. Therefore, the proposed risk-sensitive inhibitory control not only allows to reduce the number of constraint violations, but also the amount by which the constraint is violated.

V. Conclusion

Inspired by the psychological concept of inhibitory control, this paper proposes a risk-sensitive method for rendering arbitrary policies safe. This method is based on the introduction of cost functions, such that state constraints can be expressed in terms of value functions. We show that this formulation allows us to employ standard reinforcement learning techniques for obtaining policies that their only goal is to ensure safety. Based on the determined safe policies and corresponding value functions, a risk-sensitive safety constraint is employed to enforce the satisfaction of state constraints online. Thereby, risk-sensitive inhibitory control is realized and its effectiveness is demonstrated in simulations.

References

[1] J. T. Nigg, "On Inhibition/Disinhibition in Developmental Psychopathology: Views from Cognitive and Personality Psychology and a Working Inhibition Taxonomy," *Psychological Bulletin*, vol. 126, no. 2, pp. 220–246, 2000.

[2] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe Learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2017.

[4] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of Real-World Reinforcement Learning," in *ICML Workshop on Real-Life Reinforcement Learning*, 2019. [Online]. Available: http://arxiv.org/abs/1904.12901

[5] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe Reinforcement Learning via Shielding," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 2669–2678.

[6] A. Taylor, A. Singletary, Y. Yue, and A. Ames, "Learning for Safety-Critical Control with Control Barrier Functions," in *Learning for Dynamics & Control*, 2019, pp. 708–717.

[7] O. Bastani, "Safe Reinforcement Learning with Nonlinear Dynamics via Model Predictive Shielding," in *American Control Conference*, 2021, pp. 3488–3494.

[8] K. P. Wabersich, L. Hewing, A. Carron, and M. N. Zeilinger, "Probabilistic Model Predictive Safety Certification for Learning-Based Control," *IEEE Transactions on Automatic Control*, vol. 76, no. 1, pp. 176–188, 2021.

[9] K. C. Hsu, V. Rubies-Royo, C. J. Tomlin, and J. F. Fisac, "Safety and Liveness Guarantees through Reach-Avoid Reinforcement Learning," in *Robotics: Science and Systems*, 2021.

[10] S. Curi, A. Lederer, S. Hirche, and A. Krause, "Safe Reinforcement Learning via Confidence-Based Filters," in *IEEE Conference on Decision and Control*, 2022.

[11] L. Sherman, L. Steinberg, and J. Chein, "Connecting Brain Responsivity and Real-World Risk Taking: Strengths and Limitations of Current Methodological Approaches," *Developmental Cognitive Neuroscience*, vol. 33, pp. 27–41, 2018.

[12] M. Ahmadi, X. Xiong, and A. D. Ames, "Risk-Averse Control via CVaR Barrier Functions: Application to Bipedal Robot Locomotion," *IEEE Control Systems Letters*, vol. 6, pp. 878–883, 2022.

[13] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press, 2006.

[14] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems*, 2017, pp. 6405–6416.

[15] M. James, J. Baras, and R. Elliott, "Risk-Sensitive Control and Dynamic Games for Partially Observed Discrete-Time Nonlinear Systems," *IEEE Transactions on Automatic Control*, vol. 39, no. 4, pp. 780–792, 1994.

[16] V. Gaitsgory, L. Grüne, M. Höger, C. M. Kellett, and S. R. Weller, "Stabilization of Strictly Dissipative Discrete Time Systems with Discounted Optimal Control," *Automatica*, vol. 93, pp. 311–320, 2018.

[17] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in *International Conference on Machine Learning*, 2018, pp. 1861–1870.

[18] W. H. Fleming, "Risk Sensitive Stochastic Control and Differential Games," *Communications in Information and Systems*, vol. 6, no. 3, pp. 161–177, 2006.

[19] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.

[20] E. Liang, R. Liaw, P. Moritz, R. Nishihara, R. Fox, K. Goldberg, J. E. Gonzalez, M. I. Jordan, and I. Stoica, "RLlib: Abstractions for Distributed Reinforcement Learning," in *International Conference on Machine Learning*, 2018, pp. 4768–4780.

[21] Z. I. Botev, D. P. Kroese, R. Y. Rubinstein, and P. L'Ecuyer, "The cross-entropy method for optimization," in *Handbook of Statistics*. Elsevier, 2013, vol. 31, pp. 35–59.