Risk-Sensitive Safety Filters for Reinforcement Learning with Probabilistic Guarantees

Armin Lederer, *Member, IEEE*, Erfaun Noorani, *Member, IEEE*, John S. Baras, *Fellow, IEEE*, and Sandra Hirche, *Fellow, IEEE*

***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.**

***Index Terms*— Safe reinforcement learning, risk sensitive control, stochastic systems, safety filter, uncertain systems**

I. Introduction

Inhibitory control, also known as response inhibition, refers to the cognitive ability to suppress or override prepotent or habitual responses in favor of more appropriate (e.g. safer) actions [1]. For example, in industrial settings, employees with strong inhibitory control can adhere to safety protocols and refrain from engaging in risky behaviors that may lead to accidents or injuries. The ability to inhibit impulsive actions or responses can prevent accidents and mitigate risks.

Independent from this foundation in psychology, response inhibition has become increasingly popular in learning-based control and Reinforcement Learning (RL) [2] in recent years, where safety is a major concern [3]. The idea is to decouple optimality and safety in the design phase by separately determining optimal control laws and so-called safety filters [4], [5]. These two are subsequently combined online by monitoring the safety of the optimal, but potentially unsafe control input, such that it can be modified using the safety filter whenever necessary [6]. Thereby, the prepotent optimal response is inhibited to guarantee the safety of the closed-loop system. Compared to related approaches for safe reinforcement learning, e.g., primal-dual approaches [7], [8], this decoupled approach has advantages regarding modularity, ensures a stronger notion of safety, and does not require interactions with the real system before it can ensure safety.

The challenge of this approach lies in designing flexible safety filters that can effectively render nominal policies safe. A conceptionally simple approach for realizing this form of inhibitory control, sometimes referred to as shielding, formulates safety as a constraint in an optimal control problem which aims to minimize the adaptation of the nominal control input [9]. By solving the resulting optimization problem in a receding horizon fashion, such predictive safety filters can be flexibly applied to a wide range of dynamics, but the necessary online optimization can prevent their application in real-time critical control problems [10]. This weakness can be overcome by expressing safety through conditions on value functions, which can be obtained prior to the application of the safety filter. These value functions can be obtained using Hamilton-Jacobi reachability theory to maximize the safe set [11], [12], such that switching at the boundary of the safe set from the nominal to the reachability-induced controller guarantees safety. Since reachability-based methods commonly employ spatial discretization approaches to compute the value functions, they are generally restricted to problems with rather low-dimensional state spaces. Control barrier functions (CBFs) address this issue through a parametric description of the value-function [13], [14]. For simple problems, the value function can be manually designed using first principles, but classical techniques from control theory such as sum of squares optimization [15], scenario optimization [16], and convex relaxations [17] are applicable to automate the design

This work was partially supported by the European Research Council (ERC) Consolidator Grant "Safe data-driven control for human-centric systems (CO-MAN)" under grant agreement number 864686, ONR grant N00014-17-1-2622, by a grant from the Army Research Lab, by the Clark Foundation, and as a part of NCCR Automation, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40 225155).

Armin Lederer is with the Learning and Adaptive Systems Group, Institute for Machine Learning, Department of Computer Science, ETH Zurich, Switzerland (e-mail: armin.lederer@inf.ethz.ch).

Erfaun Noorani and John 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).

Sandra Hirche is with the Chair of Information-oriented Control (ITR), School of Computation, Information and Technology, Technical University of Munich, 80333 Munich, Germany (email: hirche@tum.de).

of CBFs using data. Tools from machine learning allow to further increase the flexibility and scalability of the automated design, e.g., via the supervised training of neural networks [18], [19], through imitation learning [20], or using RL with binary rewards [21] [22].

While these approaches allow the seemingly straightforward realization of inhibitory control for ensuring safety, they typically assume no uncertainty in the dynamics, which appears in real-word systems in the form of process noise and model inaccuracy. The process noise can be straightforwardly handled using expected value functions and model inaccuracy can be treated via probabilistic worst-case approximations [23], [24] for safety filters, but such approaches do not consider the *risk* of losing safety due to uncertainty. This is in strong contrast to human decision making, for which psychological studies have shown a critical link between response inhibition and an individual's risk attitude (willingness to take risk or not) [25]. The importance of this risk sensitivity is not limited to humans, but also plays a crucial role for inhibitory control in engineered systems. It can be easily achieved in principle by reformulating standard conditions using risk measures, e.g., Conditional-Value-at-Risk (CVaR) [26], when inhibitory control is implemented through analytically derived safety conditions such as CBFs. However, the extension to flexible and data-driven approaches for constructing safety conditions, e.g., using RL techniques remains an open problem.

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

- **Sufficient risk-sensitive safety conditions:** To ensure the probabilistic satisfaction of state constraints, we introduce cost functions allowing us to express sufficient safety constraints via risk-sensitive conditions for the cumulative cost along system trajectories. These conditions reveal an intuitive relationship between risk-aversion and safety probability.
- **Necessary safety conditions for cumulative costs:** Using a risk-seeking perspective, we derive an upper bound on the probability of systems remaining below a desired cumulative cost value. Thereby, this analysis provides a necessary condition for the safety of control laws.
- **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 RL. 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 during the operation of the controller, we obtain a risk-sensitive safety filter. Moreover, we prove it to inherit probabilistic safety guarantees from the safe policy obtained through RL.

These contributions form the foundation of our novel and comprehensive framework for inferring risk-sensitive safety filters using common techniques from RL. Note that this is a significant extension compared to our preliminary work [27], where we focused solely on the special case of safety filters for systems with stochastic process noise without a consideration of necessary safety conditions.

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 for environments with stochastic noise is derived in Section III. We present our method for learning risk-sensitive safety filters in settings with bounded disturbances in Section IV. In Section V, the effectiveness of the proposed safety filter is illustrated, before the paper is concluded in Section VI.

II. Problem Setting

A. Notation

We use lower/upper case symbols to denote vectors/matrices. Blackboard bold letters denote sets with indexes $+/+,0$ restricting the set to positive/non-negative numbers, e.g., $\mathbb{R}_+$ and $\mathbb{R}_{0,+}$ are the real, positive and real, non-negative numbers, respectively. If not specified differently, $\|\cdot\|$ denotes the Euclidean norm. Probabilities and expectations are denoted by $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$, respectively. Whenever necessary for clarity of exposition, an index is used to indicate the variable with respect to which the probabilities and expectations are computed, e.g., $\mathbb{E}_{\boldsymbol{x}}[\cdot]$ is the expectation with respect the random vector $\boldsymbol{x}$. Uniform and Gaussian distributions are denoted by $\mathcal{U}([a,b])$ and $\mathcal{N}(\mu,\sigma^2)$, respectively, where $a,b\in\mathbb{R}$ specify the support of the uniform distribution and $\mu,\sigma^2\in\mathbb{R}$ denote the mean and variance of the Gaussian distribution. The floor/ceil operators are denoted by $\lfloor\cdot\rfloor/\lceil\cdot\rceil$, and $\operatorname{arctan2}:\mathbb{R}\times\mathbb{R}\to(-\pi,\pi)$ is the standard 2-argument extension of the $\arctan$ function.

B. Problem Statement

We consider a discrete-time dynamical system

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

where $\boldsymbol{x}_k\in\mathbb{X}\subset\mathbb{R}^{d_x}$ are the system states, $\boldsymbol{u}_k\in\mathbb{U}\subset\mathbb{R}^{d_u}$ are control inputs, and $\boldsymbol{\omega}_k\in\mathbb{\Omega}\subset\mathbb{R}^{d_\omega}$ comprises process noise and disturbances. While we assume the true, continuous transition function $\boldsymbol{f}:\mathbb{X}\times\mathbb{U}\times\mathbb{\Omega}\to\mathbb{R}^{d_x}$ to be unknown, we require the knowledge 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 possible dynamics $\boldsymbol{f}$ is known, i.e., $\boldsymbol{f}\sim\mathcal{F}$.

This class of probabilistic models covers a wide range of practically found approaches. When the uncertainties are limited to certain parameters, i.e., we can write $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_k,\boldsymbol{u}_k,\boldsymbol{\omega}_k)$, we can implicitly define a distribution $\mathcal{F}$ by specifying distributions for the parameters $\boldsymbol{\theta}\in\mathbb{R}^{d_\theta}$. This type of system description can also be obtained automatically, e.g., when learning a model of the function $\boldsymbol{f}(\cdot,\cdot,\cdot)$ from data using linear Bayesian regression [28]. In addition to these parametric uncertainty descriptions, Assumption 1 is also satisfied when using non-parametric techniques for function inference such as Gaussian process regression [29]. Finally, approximate distributions $\mathcal{F}$

can be learned using neural network approaches, e.g., deep ensembles [30]. Therefore, this assumption holds for a wide range of scenarios.

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 RL techniques [31]. Our goal is to render the nominal policy $\boldsymbol{\pi}^*(\cdot)$ safe using inhibitory control of the form

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

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

In this response inhibition, our notion of safety follows the common principle of classifying the state space $\mathbb{X}$ into a compact 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 mere availability of a probabilistic model with potentially unbounded uncertainty, it is generally not possible to deterministically ensure that the system never enters the unsafe state space. Therefore, we define the following two notions of probabilistic safety, which follow from the straightforward extension of the concept of forward invariance [32].

Definition 1: A policy $\boldsymbol{\pi}(\cdot)$ is called 1-step δ-safe if there exists a subset $\mathbb{V} \subseteq \mathbb{X}_{\text{safe}}$ such that

$$\mathbb{P}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}) \in \mathbb{V}) \geq 1 - \delta \tag{3}$$

holds for all $\boldsymbol{x} \in \mathbb{V}$.

Definition 2: A policy $\boldsymbol{\pi}(\cdot)$ is called δ-safe if there exists a subset $\mathbb{V} \subseteq \mathbb{X}_{\text{safe}}$ such that

$$\mathbb{P}(\boldsymbol{x}_k \in \mathbb{V},\ \forall k \in \mathbb{R}_{0,+}) \geq 1 - \delta, \tag{4}$$

holds for all $\boldsymbol{x}_0 \in \mathbb{V}$ and the state sequence recursively defined through $\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{\pi}(\boldsymbol{x}_k), \boldsymbol{\omega}_k)$.

Note that 1-step δ-safety as introduced in Definition 1 is a significantly stronger requirement than merely demanding the next state to lie in the safe subset, i.e., $\mathcal{P}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}) \in \mathbb{X}_{\text{safe}}) \geq 1 - \delta$. A direct consequence of this fact is that the forward invariant set $\mathbb{V}$ is often not identical to $\mathbb{X}_{\text{safe}}$ since there are commonly states $\boldsymbol{x} \in \mathbb{X}_{\text{safe}}$ for which no control input $\boldsymbol{u}$ exists such that $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\omega})$ lies in $\mathbb{X}_{\text{safe}}$ with probability $1 - \delta$. This generally prevents the direct enforcement of safety via a constraint set $\mathbb{X}_{\text{safe}}$ on the next state $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\omega})$.

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

$$\mathbb{P}(\boldsymbol{x}_k \in \mathbb{V},\ \forall k = 1 \ldots, K) \geq (1 - \delta)^K, \tag{5}$$

where $\boldsymbol{x}_k$ is defined through iterative application of (1). However, 1-step δ-safety does not imply δ-safety as introduced in Definition 2 since it cannot ensure a constant probability for the system state $\boldsymbol{x}_k$ remaining in $\mathbb{V}$ with increasing horizon length K, i.e., for $K \to \infty$ the right side of (5) vanishes. Therefore, Definition 2 can be interpreted as an infinite horizon version of (1), and is consequently a stronger notion of safety.

In order to ensure either of these forms of safety using risk inhibition, we need additional assumptions on $\boldsymbol{\omega}_k$. In the remainder of this paper, we will distinguish the following two scenarios.

1) Stochastic Noise: In the first scenario, we assume that the noise is stochastic, such that (1) can be interpreted as a Markov Decision Process. This is formalized in the following assumption.

Assumption 2: The process noise $\boldsymbol{\omega}_k$ follows a known, potentially state-dependent, probability distribution with zero mean, i.e., $\boldsymbol{\omega}_k \sim \rho(\boldsymbol{x}_k)$.

Since we do not restrict the support of the noise distribution $\rho(\cdot)$, noise realizations can be unbounded. Hence, it is generally not possible to guarantee that the state $\boldsymbol{x}_k$ remains in any compact set for all $k \in \mathbb{R}_{0,+}$ with a positive probability as discussed in [23]. As this prevents any approach from guaranteeing δ-safety, we focus on 1-step δ-safety in this scenario. The response inhibition approach for ensuring this form of safety is derived in Section III.

2) Bounded Disturbance: In the second scenario, we do not make any assumption about the type of the disturbance $\boldsymbol{\omega}$, such that it can be deterministic, stochastic or adversarial. Instead, we only restrict the size of the disturbance as formally stated in the following assumption.

Assumption 3: An upper bound $\bar{\omega} \in \mathbb{R}_+$ of the norm of the disturbance $\boldsymbol{\omega}_k$ is known, i.e., $\|\boldsymbol{\omega}_k\| \leq \bar{\omega}$, $\forall k \in \mathbb{R}_{0,+}$.

Using the knowledge of an upper bound $\bar{\omega}$, the effect of the disturbance $\boldsymbol{\omega}_k$ on the transitions can be robustly bounded in contrast to the necessary probabilistic treatment in the stochastic noise scenario. Thereby, we are not limited to 1-step δ-safety in the bounded disturbance scenario. An approach for response inhibition with δ-safety guarantees is derived in Section IV.

III. Risk-Sensitive Response Inhibition in Stochastic Environments

Determining a condition (2b) for 1-step δ-safety with potentially unbounded process noise $\boldsymbol{\omega}_k$ 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 [23] and employ RL techniques to define these subsets through a value function. This allows us to derive sufficient conditions for 1-step δ-safety using a risk-averse perspective in Section III-A. By exploiting a risk-seeking view, necessary conditions for 1-step δ-safety are proven in Section III-B. Based on these conditions, in Section III-C, we address the problem of learning a separate, so-called backup policy whose pure focus lies on ensuring safety. Finally, a risk-sensitive safety filter employing the safe back-up policy is presented for realizing inhibitory control in RL in Section III-D.

A. Sufficient Safety Criteria via Risk-Averse Analysis

In order to derive sufficient conditions for the 1-step δ-safety of a policy $\boldsymbol{\pi}(\cdot)$, we need to find suitable sets $\mathbb{V} \subset \mathbb{X}_{\text{safe}}$ following Definition 1. For the eventual goal of inferring safety filters, it is important that these sets $\mathbb{V}$ exhibit a representation that admits a straightforward integration into optimization

algorithms. This desired property immediately motivates the use of an expected cumulative cost function

$$V_{\boldsymbol{\pi}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{f},\boldsymbol{\omega}}\left[\sum_{k=0}^{\infty}\gamma^k c(\boldsymbol{x}_k)\right] \tag{6}$$

for the definition of the set $\mathbb{V}$, where $\boldsymbol{x}_k$ is defined through the iterative application of (1) with $\boldsymbol{x}_0 = \boldsymbol{x}$ and $\boldsymbol{u}_k = \boldsymbol{\pi}(\boldsymbol{x}_k)$. Note that the immediate cost $c : \mathbb{R}^{d_x} \to \mathbb{R}_{0,+}$ and the discount factor $\gamma \in (0,1)$ in (6) are not related in any way to the computation of the nominal control law $\boldsymbol{\pi}^*(\cdot)$, but they are only used as the basis for our derivation of safety conditions. Therefore, the immediate cost $c(\cdot)$ can be thought of as an indicator of the unsafe subset $\mathbb{X}_{\text{unsafe}}$. This perspective straightforwardly allows to show the existence of a sub-level set of $V_{\boldsymbol{\pi}}(\cdot)$ contained in $\mathbb{X}_{\text{safe}}$, as guaranteed by the following lemma.

Lemma 1 ([23]): Assume there exists a constant $\hat{c} \in \mathbb{R}_+$, such that the cost $c : \mathbb{R}^{d_x} \to \mathbb{R}_{0,+}$ satisfies

$$c(\boldsymbol{x}) \geq \hat{c} \quad \forall \boldsymbol{x} \in \mathbb{X}_{\text{unsafe}}. \tag{7}$$

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

Based on this lemma, we can choose any sub-level set $\mathbb{V}_{\boldsymbol{\pi}}^{\xi}$ with $\xi \leq \bar{\xi}$ for showing 1-step δ-safety as introduced in Definition 1. In order to determine the necessary threshold value $\bar{\xi}$, different approaches can be pursued. As shown in [23], it follows directly from the proof of Lemma 1 that the search for $\bar{\xi}$ can be posed as a global optimization problem

$$\bar{\xi} = \gamma \min_{\boldsymbol{x} \in \mathbb{X}} V_{\boldsymbol{\pi}}(\boldsymbol{x}) + \hat{c}. \tag{8}$$

Since $\min_{\boldsymbol{x}\in\mathbb{X}} V_{\boldsymbol{\pi}}(\boldsymbol{x}) \geq 0$ is guaranteed, this leads to the simple threshold $\bar{\xi} \geq 0$. Due to the conservatism necessary for the proof of Lemma 1, these approaches for computing $\bar{\xi}$ can result in rather conservative values. Therefore, it can be better to directly base the computation of $\bar{\xi}$ on the condition $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}} \subset \mathbb{X}_{\text{safe}}$. This can be achieved, e.g., by defining $\bar{\xi}$ as the solution of a robust optimization problem of the form

$$\max \bar{\xi} \tag{9}$$

$$\text{s. t. } V_{\boldsymbol{\pi}}(\boldsymbol{x}) \geq \bar{\xi} \qquad \forall \boldsymbol{x} \in \mathbb{X}_{\text{unsafe}}. \tag{10}$$

As many efficient approaches for (approximately) solving such robust optimization problems have been proposed [33], [34], (9) can be effectively solved using numerical tools in practice. Therefore, finding a suitable value $\bar{\xi}$ for Lemma 1 is not an issue, such that we make the following assumption.

Assumption 4: The parameter $\bar{\xi}$ is chosen such that $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ is included in $\mathbb{X}_{\text{safe}}$, i.e., $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}} \subset \mathbb{X}_{\text{safe}}$.

Remark 1: The requirements posed on the immediate cost function $c(\cdot)$ in Lemma 1 are generally not very restrictive. The most straightforward choice fulfilling condition (7) is probably the indicator function, which is 1 if $\boldsymbol{x} \in \mathbb{X}_{\text{unsafe}}$ and 0 otherwise. While this choice is theoretically valid, it is not ideal for our goal of using standard RL algorithms for learning the expected cumulative cost function $V_{\boldsymbol{\pi}}(\cdot)$ due to a lack of informative gradients. Even though this issue can be avoided using other choices for the cost function $c(\cdot)$, e.g., with Rectified Linear Unit (ReLU) functions which can also satisfy (7), they can have a negative effect on the approximation quality of $\mathbb{X}_{\text{safe}}$. Therefore, a suitable trade-off between approximation quality and ease of learning must be found.

Remark 2: While there is no explicit constraint on the choice of γ in Lemma 1, it is clear from (6) that a sufficiently small value needs to be chosen to ensure a finite value of the cumulative cost function. Therefore, the choice of γ is important to ensure the well-definedness of (6).

Remark 3: In practice, no closed-form for (6) can be easily found, but a finite approximation of the infinite sum together with flexible function approximators allow an arbitrarily accurate approximation. It is straightforward to show that Lemma 1 remains valid for sufficiently small approximation error. Subsequent results are not directly affected by the approximation.

After a cost function $c(\cdot)$ and a corresponding threshold $\bar{\xi}$ have been determined, it only remains to derive conditions that ensure the state stays in $\mathbb{V}_{\boldsymbol{\pi}}^{\xi}$ for some $\xi \leq \bar{\xi}$ after a transition. While this could be achieved using a probabilistic "worst case" consideration as shown in [23], 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 [35]

$$\mathbb{R}_\beta[C] = \frac{1}{\beta}\log\left(\mathbb{E}\left[\exp\left(\beta C\right)\right]\right) \tag{11}$$

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 1-step δ-safety as shown in the following proposition.

Proposition 1: Consider a cost function $c(\cdot)$ satisfying (7), process noise $\boldsymbol{\omega}$ for which Assumption 2 holds, and a constant $\bar{\xi} \in \mathbb{R}_+$ satisfying Assumption 4. If there exist constants $\xi, \beta \in \mathbb{R}_+$ with $\xi < \bar{\xi}$ such that[1]

$$\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] \leq \xi, \qquad \forall \boldsymbol{x} \in \mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}} \tag{12}$$

holds for $\boldsymbol{x}^+ = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega})$, then, $\boldsymbol{\pi}(\cdot)$ is 1-step δ-safe on $\mathbb{V}_{\boldsymbol{\pi}}^{\xi}$ with

$$\delta = \exp\left(\beta\left(\xi - \bar{\xi}\right)\right). \tag{13}$$

Proof: Due to Lemma 1 and Assumption 4, 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

$$\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) > \bar{\xi}\right) = \mathbb{E}_{\boldsymbol{x}^+}\left[I_{\bar{\xi}}(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)\right], \tag{14}$$

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

$$I_{\bar{\xi}}(V) = \begin{cases} 0 & \text{if } V \leq \bar{\xi} \\ 1 & \text{if } V > \bar{\xi}. \end{cases} \tag{15}$$

Note that $V_{\boldsymbol{\pi}}(\cdot)$ is a deterministic function, such that the expectation affects only the random variable $\boldsymbol{x}^+$ in (23). Moreover, β is positive, $\exp(0) = 1$ and the exponential function

[1]The risk of the value function $V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)$ is computed with respect to the noise distribution ρ and the function distribution $\mathcal{F}$.

is strictly increasing and positive. Therefore, we can bound the indicator function through the exponential expression

$$I_{\bar{\xi}}(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)) \leq \exp\left(\beta\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) - \bar{\xi}\right)\right) \tag{16}$$

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

$$\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) > \bar{\xi}\right) \leq \mathbb{E}_{\boldsymbol{x}^+}\left[\exp\left(\beta V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)\right)\right]\exp(-\beta\bar{\xi}). \tag{17}$$

Due to the definition of the risk operator in (11), we can simplify the right side of this inequality to obtain

$$\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) > \bar{\xi}\right) \leq \exp\left(\beta\left(\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] - \bar{\xi}\right)\right). \tag{18}$$

Since $\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] \leq \xi$ is ensured by (12), we have $\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) > \bar{\xi}\right) \leq \delta$ with δ defined in (13). ■

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 RL. Moreover, it offers a simple expression for the probability of safety, such that it can easily be computed. 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 (12). 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 (12). In the extreme case of $\beta \to \infty$, this leads to (12) corresponding to a condition on the worst case realization of $\boldsymbol{\omega}_k$ and $\boldsymbol{f}(\cdot)$ [35]. This increasing risk-aversion with growing β is intuitively accompanied by an increase in the probability of safety.

Remark 4: When using Proposition 1 to provide safety guarantees for a policy, it is not possible to choose β and ξ independently. This is due to the coupling of these two parameters caused by (12), which can become violated when choosing a large β and small ξ simultaneously. Therefore, β and ξ need to be jointly determined, e.g., by maximizing (13) such that (12) is satisfied, which can be performed using robust optimization techniques. Note that this coupling via (12) is also how the choice of γ and $c(\cdot)$ impact the certifiable probability of safety. While the exact relation between these design choices and the probability of safety is non-trivial in general, the understanding of this coupling still allows the beneficial integration of prior knowledge about states $\boldsymbol{x}$ likely to transition to $\mathbb{X}_{\text{unsafe}}$ by assigning them high costs $c(\boldsymbol{x})$. Moreover, this coupling loses importance when considering the special case of policies minimizing the cumulative cost because this induces guarantees on the satisfaction of (12) as shown in Section III-C.

Remark 5: As discussed after Definition 2, Proposition 1 immediately implies that $\boldsymbol{x}_k$ remains in $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ for K time steps with probability of at least $(1-\delta)^K$ if it starts in $\mathbb{V}_{\boldsymbol{\pi}}^{\xi}$, i.e., $\boldsymbol{x}_0 \in \mathbb{V}_{\boldsymbol{\pi}}^{\xi}$. Therefore, the safety certificate provided by Proposition 1 is not limited to one time step, but straightforwardly extends to arbitrary long but finite time intervals. However, it is also clearly visible that δ-safety over infinitely long time intervals cannot be guaranteed for any positive value δ. This is a direct consequence of the stochastic noise, which requires us to treat every time step independently, such that unbounded realizations of the noise can drive the system outside the safe set $\mathbb{X}_{\text{safe}}$ at every time step without the possibility to exploit spatial correlations to improve safety guarantees.

Remark 6: The safety condition (12) can be interpreted as a variant of commonly used barrier function criteria for state constraints [36]. This can be easily seen by negating (12) and adding $V_{\boldsymbol{\pi}}(\boldsymbol{x})$ to both sides, which results in

$$V_{\boldsymbol{\pi}}(\boldsymbol{x}) - \mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] \geq V_{\boldsymbol{\pi}}(\boldsymbol{x}) - \xi. \tag{19}$$

Therefore, $B(\boldsymbol{x}) = \xi - V_{\boldsymbol{\pi}}(\boldsymbol{x})$ can be considered a zeroing barrier function [14] and (19) corresponds to the risk-averse version of the deterministic barrier condition

$$B(\boldsymbol{x}^+) - B(\boldsymbol{x}) \geq -\alpha(B(\boldsymbol{x})) \tag{20}$$

with extended class $\mathcal{K}_\infty$ function $\alpha : \mathbb{R} \to \mathbb{R}$ satisfying $\alpha(B) = B$ [14].

B. Necessary Safety Criteria via Risk-Seeking Analysis

In addition to deriving a sufficient condition for 1-step δ-safety, the expected cumulative value function (6) also admits the derivation of a necessary condition for the safety of its sub-level set $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$. This condition is obtained by taking a risk-seeking perspective, i.e., employing the risk operator $\mathbb{R}_\beta[\cdot]$ with $\beta < 0$, to determine an upper bound on the probability of staying in the sub-level set $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$. Thereby, we can specify, under which criteria $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ cannot be 1-step δ-safe as shown in the following proposition.

Proposition 2: Consider a cost function $c(\cdot)$ satisfying (7), process noise $\boldsymbol{\omega}$ for which Assumption 2 holds, and a constant $\bar{\xi} \in \mathbb{R}_+$ satisfying Assumption 4. If there exist constants $\beta \in \mathbb{R}_-$ and $\xi \in \mathbb{R}_+$ with $\xi > \bar{\xi}$ such that

$$\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] \geq \xi \tag{21}$$

holds for some state $\boldsymbol{x}$ and $\boldsymbol{x}^+ = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega})$, then, $\boldsymbol{\pi}(\cdot)$ is not 1-step δ-safe on $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ for all

$$\delta < 1 - \exp\left(\beta\left(\xi - \bar{\xi}\right)\right). \tag{22}$$

Proof: This proof is analogous to the proof of Proposition 1, but we start our analysis with the probability of staying in $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$. Therefore, we have

$$\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) \leq \bar{\xi}\right) = 1 - \mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) > \bar{\xi}\right) \tag{23}$$

$$= \mathbb{E}_{\boldsymbol{x}^+}\left[1 - I_{\bar{\xi}}(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)\right], \tag{24}$$

where we use the probability of the complementary event in the first line and the indicator funtion (15) in the second line. Note that β is negative, $\exp(0) = 1$ and the exponential function is strictly increasing and positive. Therefore, we can bound the indicator function through the exponential

expression

$$1 - I_{\bar{\xi}}(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)) \leq \exp\left(\beta\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) - \bar{\xi}\right)\right), \quad \beta < 0 \tag{25}$$

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

$$\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) \leq \bar{\xi}\right) \leq \mathbb{E}_{\boldsymbol{x}^+}\left[\exp\left(\beta V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)\right)\right]\exp(-\beta\bar{\xi}). \tag{26}$$

Due to the definition of the risk operator in (11), we can simplify the right side of this inequality to obtain

$$\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) \leq \bar{\xi}\right) \leq \exp\left(\beta\left(\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] - \bar{\xi}\right)\right). \tag{27}$$

Since $\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{x}^+)] \geq \xi$ is ensured by (12), we have $\mathbb{P}\left(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) \leq \bar{\xi}\right) \leq \exp\left(\beta\left(\xi - \bar{\xi}\right)\right)$. This implies that (22) must hold since at least one state $\boldsymbol{x}$ is mapped outside $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ with probability of at least $1 - \exp\left(\beta\left(\xi - \bar{\xi}\right)\right)$. ∎

This result is similar in its conditions to Proposition 1, but differs in essential points. Firstly, (21) is reversed compared to (12) in the sense that ξ must be a lower bound. Thereby, it allows us to lower-bound the probability of leaving the set. Moreover, a single state satisfying condition (21) is sufficient in contrast to Proposition 1, where all states in $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ have to satisfy (12). This is a natural simplification since a single state suffices to make the satisfaction of (12) impossible. Thereby, Proposition 2 provides a necessary condition for the safety on $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$: No combination of $\boldsymbol{x}$, β and ξ is allowed to exist such that (21) holds.

Remark 7: A policy does not have to satisfy the conditions of either Proposition 1 or 2. This is partially due to the fact that the approximation quality of the estimated safe set depends on the choice of the cost $c(\cdot)$, but more crucially it is a result of the approximations used in bounding the probability of safety based on the risk. Hence, both propositions do not allow a classification of all policies into certifiable 1-step δ-safe or not.

C. Inferring Safe Policies via Reinforcement Learning

While Propositions 1 and 2 allow to decide about the safety of a policy, they do not address the problem of determining a safe policy. In this section, we show that this problem can be solved by formulating it as the optimization problem

$$\boldsymbol{\pi}_{\text{safe}} = \underset{\boldsymbol{\pi}\in\Pi}{\arg\min}\, \mathbb{E}_{\boldsymbol{x}}\left[V_{\boldsymbol{\pi}}(\boldsymbol{x})\right]. \tag{28}$$

This optimization problem has the form of a standard reinforcement learning problem, such that any reinforcement learning algorithm can be used in principle to solve (28). While the choice of algorithms is not inherently restricted, the importance of the value function in our approach can render modern actor-critic approaches, e.g., SAC [37], beneficial as they directly infer a value function along with the policy. Thereby, the additional step of learning a model of the value function $V_{\boldsymbol{\pi}_{\text{safe}}}$ from roll-outs of the safe policy $\boldsymbol{\pi}_{\text{safe}}$ can be alleviated via a suitable choice of the reinforcement learning algorithm.

Even though (28) 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 (7) and process noise $\boldsymbol{\omega}$ for which Assumption 2 holds. 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

$$V_{\tilde{\boldsymbol{\pi}}}(\boldsymbol{x}) \leq \theta_1 c(\boldsymbol{x}) + \theta_2, \quad \forall \boldsymbol{x} \in \mathbb{X} \tag{29}$$

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

$$V_{\boldsymbol{\pi}}(\boldsymbol{x}) \geq \theta_3 c(\boldsymbol{x}) + \theta_4, \quad \forall \boldsymbol{x} \in \mathbb{X} \tag{30}$$

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

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

holds, then, the policy (28) is 1-step δ^*-safe on $\mathbb{V}_{\boldsymbol{\pi}_{\text{safe}}}^{\xi^*}$ with $\delta^* = \exp\left(\beta^*\left(\xi^* - \bar{\xi}\right)\right)$, where

$$\beta^*, \xi^* = \underset{\beta\in\mathbb{R}_+, \xi\in\mathbb{R}_+}{\arg\min}\ \exp\left(\beta\left(\xi - \bar{\xi}\right)\right) \tag{32a}$$

$$\text{s.t. } \xi < \bar{\xi} \tag{32b}$$

$$(12) \text{ holds.} \tag{32c}$$

Discussion: The core challenge overcome by Theorem 1 lies in establishing a connection between the optimization problem (28) suited for generic RL algorithms, and the risk-based decrease condition (12) for safety. This is achieved by exploiting properties of value functions to show an expected decrease and subsequently extending this guarantee to the risk condition (12). While large values for θ_3 and θ_4 in (30) are generally beneficial for admitting larger values of $\hat{c}$ in (31), it is always possible to trivially choose $\theta_3 = 1$, $\theta_4 = 0$ due to non-negativity of $c(\cdot)$. Condition (29) essentially requires a sufficiently slow increase of the immediate costs $c(\boldsymbol{x}_k)$ along trajectories for some policy $\tilde{\boldsymbol{\pi}}(\cdot)$. The additional requirement of $\theta_1 < {}^1\!/_{(1-\gamma)}$ merely requires that $V_{\tilde{\boldsymbol{\pi}}}(\boldsymbol{x}) < {}^{c(\boldsymbol{x})}\!/_{(1-\gamma)} + \theta_2$, which can be straightforwardly seen to be satisfied if the policy $\tilde{\boldsymbol{\pi}}(\cdot)$ performs better than maintaining the initial cost $c(\boldsymbol{x})$ for all times. Such a behavior can be shown to be achieved if, e.g., variants of exponential controllability hold, which admits the direct derivation of the constants θ_1 and θ_2 [38]. It is important to note that these properties do not need to be shown for the optimized policy $\boldsymbol{\pi}_{\text{safe}}$, but a simpler, e.g., parametric, policy can be used to prove these properties. Thus, (29) can be interpreted as the requirement for the existence of a safe policy considering the decrease of the value function it guarantees. Hence, (29) and (30) do not pose severe restrictions in practice, such that Theorem 1 is flexibly applicable in general.

Note that the required lower bound (30) for all possible cost functions $V_{\boldsymbol{\pi}}(\cdot)$ is only necessary because of the offset θ_2, which leads to a lower bound

$$\underline{\xi} = \frac{\theta_2}{\theta_1(\gamma-1)+1} \tag{33}$$

for the admissible values of $\bar{\xi}$ due to (30) and (31). Since the admissible value $\bar{\xi}$ depends directly on the cost function $V_{\boldsymbol{\pi}}(\cdot)$, it causes the challenge of an indirect dependence of $\bar{\xi}$ on the policy $\boldsymbol{\pi}(\cdot)$. Therefore, $V_{\tilde{\boldsymbol{\pi}}}(\cdot)$ and $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$ potentially admit different values for $\bar{\xi}$ as illustrated in Fig. 1. This potential ambiguity is resolved by (30), which establishes

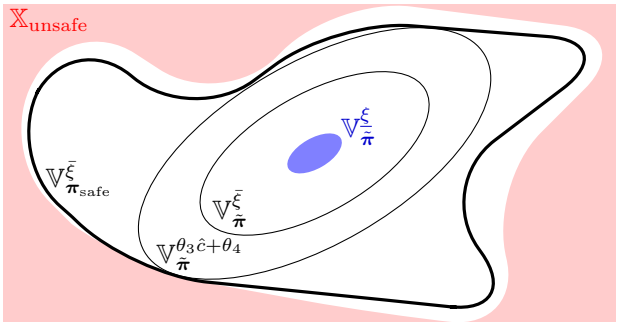

Fig. 1. The largest safe sub-level set $\mathbb{V}^{\bar{\xi}}_{\pi_{\text{safe}}}$ for the safe policy $\pi_{\text{safe}}(\cdot)$ typically enlarges the maximal sub-level set $\mathbb{V}^{\theta_3\hat{c}+\theta_4}_{\tilde{\pi}}$ contained in $\mathbb{X}_{\text{safe}}$ for an arbitrary policy $\tilde{\pi}(\cdot)$ satisfying (29). In particular, the set $\mathbb{V}^{\bar{\xi}}_{\pi_{\text{safe}}}$ is usually significantly larger than the sub-level set $\mathbb{V}^{\bar{\xi}}_{\tilde{\pi}}$ with the same threshold $\bar{\xi}$, which generally does not correspond to the largest safe sub-level set. Note that due to the offset θ_2, there exists a lower bound $\underline{\xi}$ for $\bar{\xi}$ below which the conditions of Proposition 1 cannot be satisfied anymore.

a direct relationship between $\hat{c}$ and $\bar{\xi}$ for all possible cost functions $V_\pi(\cdot)$ and thereby leads to the lower bound (31), which alleviates the need for Assumption 4. Thereby, (31) can be interpreted as the condition ensuring $\mathbb{V}^{\xi^*}_{\pi_{\text{safe}}} \subset \mathbb{X}_{\text{safe}}$. 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 (12) guarantees the existence of parameters ξ and β satisfying the requirements of Proposition 1.

Lemma 2: Assume that

$$\mathbb{E}_{\boldsymbol{x}^+}[V_\pi(\boldsymbol{x}^+)] \leq \tilde{\xi}, \qquad \forall \boldsymbol{x} \in \mathbb{V}^{\bar{\xi}}_{\pi} \tag{34}$$

holds for some constant $\tilde{\xi} < \bar{\xi}$. Then, there exist constants $\beta \in \mathbb{R}_+$ and $\xi < \bar{\xi}$ such that (12) is satisfied.

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

$$\mathbb{R}_\beta[V_\pi(\boldsymbol{x}^+)] = \frac{1}{\beta}\log\left(1+\beta\mathbb{E}_{\boldsymbol{x}^+}\left[V_\pi(\boldsymbol{x}^+)\right]+\frac{\beta^2}{2}\mathbb{E}_{\boldsymbol{x}^+}\left[V_\pi^2(\boldsymbol{x}^+)\right]+\dots\right). \tag{35}$$

From the premise of the lemma, it follows that

$$\mathbb{R}_\beta[V_\pi(\boldsymbol{x}^+)] \leq \frac{1}{\beta}\log\left(1+\beta\tilde{\xi}+\frac{\beta^2}{2}\mathbb{E}_{\boldsymbol{x}^+}\left[V_\pi^2(\boldsymbol{x}^+)\right]+\dots\right). \tag{36}$$

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

$$\mathbb{R}_\beta[V_\pi(\boldsymbol{x}^+)] < \tilde{\xi} + \beta\left(\frac{1}{2}\mathbb{E}_{\boldsymbol{x}^+}\left[V_\pi^2(\boldsymbol{x}^+)\right]+\dots\right). \tag{37}$$

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 (37) is smaller than $\bar{\xi}$, which concludes the proof. ∎

The key idea behind this result is that (12) converges to (34) for $\beta \to 0$. Therefore, it is sufficient to determine a policy π, which satisfies the risk-neutral condition (34), for ensuring (12) with a suitably small value of $\beta \in \mathbb{R}_+$.

Although (34) is a risk-neutral condition, it exhibits an expectation with respect to the next state $\boldsymbol{x}^+$. Hence, 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 (28). In order to overcome this issue, we exploit (29) to relate $\mathbb{E}_{\boldsymbol{x}^+}[V_\pi(\boldsymbol{x}^+)]$ to $V_\pi(\boldsymbol{x})$. 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 (29) is satisfied. Then, it holds that

$$\mathbb{E}_{\boldsymbol{x}^+}[V_\pi(\boldsymbol{x}^+)] - V_\pi(\boldsymbol{x}) \leq \frac{\theta_1-\theta_1\gamma-1}{\theta_1\gamma}V_\pi(\boldsymbol{x}) + \frac{\theta_2}{\gamma\theta_1}. \tag{38}$$

Proof: By solving Bellman's identity

$$V_\pi(\boldsymbol{x}) = c(\boldsymbol{x}) + \gamma\mathbb{E}_{\boldsymbol{x}^+}\left[V_\pi(\boldsymbol{x}')\right], \tag{39}$$

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

$$\Delta V_\pi(\boldsymbol{x}) = \frac{1}{\gamma}(-c(\boldsymbol{x}) + (1-\gamma)V_\pi(\boldsymbol{x})). \tag{40}$$

Due to (29), we have

$$c(\boldsymbol{x}) \geq \frac{V_\pi(\boldsymbol{x})-\theta_2}{\theta_1}, \tag{41}$$

which allows us to bound (40) by

$$\Delta V_\pi(\boldsymbol{x}) \leq \frac{1}{\gamma}\left(-\frac{V_\pi(\boldsymbol{x})-\theta_2}{\theta_1} + (1-\gamma)V_\pi(\boldsymbol{x})\right). \tag{42}$$

Rearranging the terms on the right side finally yields

$$\Delta V_\pi \leq \frac{\theta_1-\theta_1\gamma-1}{\theta_1\gamma}V_\pi(\boldsymbol{x}) + \frac{\theta_2}{\gamma\theta_1}, \tag{43}$$

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 $V_\pi(\boldsymbol{x})$ also reduces $\mathbb{E}_{\boldsymbol{x}^+}[V_\pi(\boldsymbol{x}^+)]$. 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 $\pi_{\text{safe}}(\cdot)$ as the point-wise optimum $\pi_{\boldsymbol{x}}(\boldsymbol{x}) = \arg\min_{\pi\in\Pi} V_\pi(\boldsymbol{x})$ for a given $\boldsymbol{x}$ and a continuous transition function $\boldsymbol{f}(\cdot,\cdot,\cdot)$. Due to optimality of $\pi_{\boldsymbol{x}}(\cdot)$, we additionally have the inequality $V_{\pi_{\boldsymbol{x}}}(\boldsymbol{x}) \leq V_{\tilde{\pi}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{X}$. Therefore, it follows from Lemma 3 that

$$\mathbb{E}[V_{\pi_{\text{safe}}}(\boldsymbol{x}^+)] \leq \frac{1}{\gamma}\left(1-\frac{1}{\theta_1}\right)V_{\pi_{\text{safe}}}(\boldsymbol{x}) + \frac{\theta_2}{\gamma\theta_1}. \tag{44}$$

Since the right side of (44) is linear in $V_{\pi_{\text{safe}}}(\boldsymbol{x})$, the maximum inside $\mathbb{V}_{\bar{\xi}}$ is achieved for $V_{\pi_{\text{safe}}}(\boldsymbol{x}) = \bar{\xi}$, such that Lemma 2 guarantees

$$\mathbb{R}_\beta[V_{\pi_{\text{safe}}}(\boldsymbol{x}^+)] \leq \frac{1}{\gamma}\left(1-\frac{1}{\theta_1}\right)\bar{\xi} + \frac{\theta_2}{\gamma\theta_1} \tag{45}$$

for sufficiently small $\beta \in \mathbb{R}_+$. Therefore, (12) holds if

$$\frac{1}{\gamma}\left(1-\frac{1}{\theta_1}\right)\bar{\xi} + \frac{\theta_2}{\gamma\theta_1} < \bar{\xi}. \tag{46}$$

Solving this inequality for $\bar{\xi}$, we obtain that

$$\frac{\theta_2}{\theta_1(\gamma-1)+1} < \bar{\xi} \tag{47}$$

needs to be satisfied. As (30) implies $\bar{\xi} \geq \theta_3\hat{c} + \theta_4$, we can adapt this requirement to a condition on $\hat{c}$ yielding

$$\theta_3\hat{c} + \theta_4 > \frac{\theta_2}{\theta_1(\gamma-1)+1}. \tag{48}$$

Due to (31), this inequality holds. Hence, (12) is satisfied for sufficiently small β, which ensures that (32) is feasible and results in a probability $\delta^* < 1$. Thus, Proposition 1 immediately implies 1-step δ^*-safety of $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ and thereby concludes the proof. ∎

Remark 8: In practice, we can generally only obtain an approximate solution to (28). However, this does not crucially affect our results as we only need to account for approximation errors via an additive term in (44), while other steps of the proof of Theorem 1 do not change. Therefore, it is straightforward to show that Theorem 1 remains valid, but requires a larger lower bound for $\hat{c}$ compared to (31) when using an approximate solution for (28).

D. Risk-Sensitive Response Inhibition via Safety Filters

Based on the safe policy $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ obtained using (28), we propose a risk-sensitive inhibitory control strategy for enabling safe RL. For this purpose, we first obtain an optimal, potentially unsafe policy by solving the optimization problem

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}\in\Pi} \mathbb{E}_{\boldsymbol{f},\boldsymbol{\omega},\boldsymbol{x}_0}\left[\sum_{k=0}^{\infty}\gamma^k r(\boldsymbol{x}_k, \boldsymbol{\pi}(\boldsymbol{x}_k))\right], \tag{49}$$

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 RL algorithms such as soft actor-critic RL [37]. Afterward, a safe backup policy $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ is computed by solving (28), which can be straightforwardly achieved using standard off-policy RL techniques. Finally, we apply the policy to the true system (1). For this roll-out, we employ the risk-sensitive filter

$$\boldsymbol{\pi}^*_{\text{safe}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}\in\mathbb{U}} \|\boldsymbol{\pi}^*(\boldsymbol{x}) - \boldsymbol{u}\| \tag{50a}$$

$$\text{s.t. } \mathbb{R}_\beta[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{\omega}))] \leq \xi^* \tag{50b}$$

which makes use of the safe backup policy $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ through the cost function $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$ and minimally adjusts the policy $\boldsymbol{\pi}^*(\cdot)$ such that the safety condition (12) is satisfied. Thereby, 1-step δ-safety $\boldsymbol{\pi}^*_{\text{safe}}(\cdot)$ 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 (7), a threshold $\hat{c}$ for which (31) holds, and process noise $\boldsymbol{\omega}$ satisfying Assumption 2. Moreover, assume that there exists a policy $\tilde{\boldsymbol{\pi}}(\cdot)$ satisfying (29) with $\theta_1 < {}^1/_{(1-\gamma)}$ for all $\boldsymbol{x} \in \mathbb{X}_{\text{safe}}$. Then, the safety filtered policy (50) is 1-step 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 (32).

Proof: Due to Theorem 1, $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ defined in (28) satisfies (50b). Thus, the optimization problem (50) 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 (32b). Therefore, δ-safety on $\mathbb{V}_\xi \subset \mathbb{V}_{\xi^*}$ with $\delta > \delta^*$ can be straightforwardly ensured in practice by choosing a sufficiently large value $\xi < \bar{\xi}$ and a suitably small value $\beta \in \mathbb{R}_+$. The specific values that are necessary depend strongly on the noise distribution ρ and function distribution $\mathcal{F}$. This can easily be seen from the fact that distributions with larger variance generally cause higher risks $\mathbb{R}_\beta[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{\omega}))]$, such that smaller values of β and larger values of ξ can be required.

To reach a desired probability $\delta_d \in \mathbb{R}_+$, it is possible to adaptively choose these parameters by solving a modified optimization problem

$$\pi^*_{\text{safe}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}\in\mathbb{U}} \min_{\beta\in\mathbb{R}_+,\xi\in\mathbb{R}_+} \|\boldsymbol{\pi}^*(\boldsymbol{x}) - \boldsymbol{u}\| + \lambda|\delta - \delta_d| \tag{51a}$$

$$\text{s.t. } \mathbb{R}_\beta[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{\omega}))] \leq \xi \tag{51b}$$

$$\delta = \exp\left(\beta\left(\xi - \bar{\xi}\right)\right) \tag{51c}$$

$$\xi < \bar{\xi}, \tag{51d}$$

where $\lambda \in \mathbb{R}^+$ is a Lagrange multiplier. By selecting a sufficiently large λ, the second term in (51a) acts as a soft constraint and ensures that the desired probability δ_d is reached whenever possible and minimizes the difference otherwise. Since Theorem 1 ensures the existence of β^*, ξ^* satisfying the constraints (51b)-(51d), this optimization problem is guaranteed to be feasible. Hence, it can provide an effective way to determine safe control inputs without the need to specify β and ξ a priori.

Remark 9: 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 (50b) 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 [39].

IV. Risk-Sensitive Response Inhibition under Bounded Disturbances

While Section III provides approaches for certifying and learning safe policies in environments with stochastic noise, these methods cannot be directly applied when merely a bound for the disturbances are available. Moreover, the previously presented results do not provide δ-safety guarantees. Therefore, we extend them and derive conditions for δ-safety in environments with bounded disturbances in Section IV-A. Based on these conditions, we show how δ-safe policies can be learned despite bounded disturbances in Section IV-B. Finally, a risk-sensitive safety filter employing the learned backup policy is proposed in Section IV-C.

A. Robustness of Risk-Averse Cost Conditions

Since we do not make any assumptions about the distribution of the bounded disturbance, we cannot compute expectations with respect to the disturbance distribution. Therefore,

we marginally adapt the definition of the expected cumulative cost (6) by defining it as

$$V_{\boldsymbol{\pi}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{f}}\left[\sum_{k=0}^{\infty} \gamma^k c(\boldsymbol{x}_k)\right], \tag{52}$$

where we consider the state sequence $\boldsymbol{x}_k$ to be generated by dynamics (1) with noise $\boldsymbol{\omega}_k = \mathbf{0}$ for all $k \in \mathbb{N}$. Note that we only ignore the presence of disturbances $\boldsymbol{\omega}$ by setting $\boldsymbol{\omega}_k = \mathbf{0}$ in the definition of the value function (52), but all guarantees in this section are derived for systems with bounded noise as stated in Assumption 3.

Despite this minor change, it can be directly seen that Lemma 1 still holds for (52), such that we can focus on the derivation of an analogous result to Proposition 1, but with the stronger notion of δ-safety. In order to achieve this, we require the well known concept of covering numbers [40]. While the computation of covering numbers for a given discretization $\tau \in \mathbb{R}_+$ is generally a complicated problem, they can be easily upper bounded in d_x-dimensional Euclidean spaces by [41]

$$N(\tau) = \left(\frac{\sqrt{d_x}\max_{\boldsymbol{x},\boldsymbol{x}' \in \mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}} \|\boldsymbol{x} - \boldsymbol{x}'\|}{2\tau}\right)^{d_x}. \tag{53}$$

Based on covering numbers, we can straightforwardly adapt and extend Proposition 1 to prove δ-safety for the bounded disturbance scenario as shown in the following proposition.

Proposition 3: Consider a cost function $c(\cdot)$ satisfying (7), disturbances satisfying Assumption 3, a constant $\bar{\xi} \in \mathbb{R}_+$ satisfying Assumption 4, and L_V-, $L_{f,x}$- and $L_{f,\omega}$-Lipschitz functions $V_{\boldsymbol{\pi}}(\cdot)$, $\boldsymbol{f}(\cdot, \boldsymbol{\pi}(\cdot), \boldsymbol{\omega})$ and $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \cdot)$, respectively. Assume there exist constants $\xi, \beta \in \mathbb{R}_+$ with $\xi + L_V L_{f,\omega}\bar{\omega} < \bar{\xi}$ such that (12) holds for $\boldsymbol{x}^+ = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \mathbf{0})$ and define

$$\Delta\xi = \bar{\xi} - \xi - L_V L_{f,\omega}\bar{\omega}, \tag{54}$$

$$\delta = \min_{\eta \in (0,1)} N\left(\frac{\eta \Delta\xi}{L_V L_{f,x}}\right) \exp\left(-\beta(1-\eta)\Delta\xi\right). \tag{55}$$

If $\delta < 1$, then, the system is safe for all $k \in \mathbb{N}$ on $\mathbb{V}_{\xi}^{\boldsymbol{\pi}}$ with probability $1 - \delta$.

Proof: We proof this proposition by showing safety probabilistically for a finite set of points and extending the guarantees to the whole set $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ through Lipschitz continuity. For this purpose, we define a grid $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N(\tau)}$ such that

$$\max_{\boldsymbol{x} \in \mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}} \min_{n=1,\dots,N(\tau)} \|\boldsymbol{x} - \boldsymbol{x}^{(n)}\| \leq \tau. \tag{56}$$

This ensures that any point x within the safe set $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ is not more than τ distance away from a test point on the grid. Lipschitz continuity of $V_{\boldsymbol{\pi}}(\cdot)$ and $\boldsymbol{f}(\cdot,\cdot,\cdot)$ implies that

$$\max_{\boldsymbol{x}: \|\boldsymbol{x} - \boldsymbol{x}^{(n)}\| \leq \tau} \mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \mathbf{0}))] \leq \mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{f}(\boldsymbol{x}^{(n)}, \boldsymbol{\pi}(\boldsymbol{x}^{(n)}), \mathbf{0}))] + L_V L_{f,x}\tau \tag{57}$$

due to the linearity of the risk operator with respect to constants. Similarly, Lipschitz continuity of $V_{\boldsymbol{\pi}}(\cdot)$ and $\boldsymbol{f}(\cdot,\cdot,\cdot)$ guarantees that

$$\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{f}(\boldsymbol{x}^{(n)}, \boldsymbol{\pi}(\boldsymbol{x}^{(n)}), \boldsymbol{\omega}))] \leq \mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{f}(\boldsymbol{x}^{(n)}, \boldsymbol{\pi}(\boldsymbol{x}^{(n)}), \mathbf{0}))] + L_V L_{f,\omega}\|\boldsymbol{\omega}\|. \tag{58}$$

Combining these two inequalities, exploiting (12) to bound $\mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{f}(\boldsymbol{x}^{(n)}, \boldsymbol{\pi}(\boldsymbol{x}^{(n)}), \mathbf{0}))]$ by ξ, and employing the assumed bound $\|\boldsymbol{\omega}\| \leq \bar{\omega}$, we obtain

$$\max_{\boldsymbol{x}: \|\boldsymbol{x} - \boldsymbol{x}^{(n)}\|} \mathbb{R}_\beta[V_{\boldsymbol{\pi}}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}))] \leq L_V L_{f,x}\tau + L_V L_{f,\omega}\bar{\omega} + \xi \tag{59}$$

with probability $1 - \exp(\beta(\xi + L_V L_{f,x}\tau + L_V L_{f,\omega}\bar{\omega} - \bar{\xi}))$ for each grid point $\boldsymbol{x}^{(n)}$ individually due to Proposition 1. Therefore, we choose $\tau = \eta(\bar{\xi} - \xi - L_V L_{f,\omega}\bar{\omega}) / L_V L_{f,x}$ for $\eta \in (0,1)$ to ensure $L_V L_{f,x}\tau + L_V L_{f,\omega}\bar{\omega} + \xi < \bar{\xi}$. Moreover, note that by over-approximating $\mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}$ using a box, it follows directly from [41] that the covering number $N(\tau)$ can be bounded by (53). Hence, the union bound over all $n = 1, \dots, N$ grid points yields

$$\mathbb{P}(V_{\boldsymbol{\pi}}(\boldsymbol{x}^+) \leq \bar{\xi},\ \forall \boldsymbol{x} \in \mathbb{V}_{\boldsymbol{\pi}}^{\bar{\xi}}) \geq 1 - \delta, \tag{60}$$

where δ is defined in (55). ■

In this theorem, we exploit Lipschitz continuity to quantify the worst-case effect of the bounded disturbance deterministically. This allows the derivation of the bound (56), such that we immediately obtain a tightened condition $\xi + \Delta\xi < \bar{\xi}$ which explicitly takes the disturbance bound $\bar{\omega}$ into account via (54). While this might seem like a more restrictive condition compared to Proposition 1, it is important to note that this tightening is already included in $\mathbb{R}[\cdot]$ in the stochastic noise setting. Therefore, the explicit appearance of $\Delta\xi$ is merely an artifact of our lack of knowledge about the disturbance distribution, but does not directly imply additional conservatism of Proposition 3. The tightness of our result becomes apparent when considering the special case of value functions $V_\pi(\boldsymbol{x}) = |\boldsymbol{c}^T\boldsymbol{x}|$ for some vector $\boldsymbol{c} \in \mathbb{R}^{d_x}$ and dynamics with additive disturbance $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}) = \tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))$. Then, we have

$$\max_{\|\boldsymbol{\omega}\| \leq \bar{\omega}} \mathbb{R}_\beta[|\boldsymbol{c}^T\tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x})) + \boldsymbol{c}^T\boldsymbol{\omega}|] = \mathbb{R}_\beta[|\boldsymbol{c}^T\tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}))|] + \bar{\omega}\|\boldsymbol{c}\|, \tag{61}$$

i.e., (59) becomes an equality. This similarly holds true for (57) when additionally considering linear dynamics. Therefore, we cannot expect to obtain any less conservative results without further and more restrictive assumptions on $V_{\boldsymbol{\pi}}(\cdot)$ and $\boldsymbol{f}(\cdot,\cdot,\cdot)$.

The transition from 1-step δ-safety to δ-safety is achieved using a common approach for ensuring safety with learned probabilistic models [42], [43], for which we exploit the deterministic treatment of the bounded disturbances. This allows us to avoid considering the safety probabilities for individual time steps and jointly lower bound the probability of the next state being in $\mathbb{V}_{\xi}^{\boldsymbol{\pi}}$ all states using (60). Therefore, when starting in the set $\mathbb{V}_{\xi}^{\boldsymbol{\pi}}$, we do not have to multiply the individual probabilities of each step along the path since they are all covered already by the joint probability.

Since the safety condition (12) is used both in Proposition 1 and Proposition 3, most of the interpretation of Proposition 1 directly transfers here. The main difference lies in the resulting probability of a safety violation δ, where $\Delta\xi$ plays a twofold role in (55). A large value of $\Delta\xi$ has not only a positive effect on the exponential term in (55), but it also reduces the covering

number $N(\cdot)$. Therefore, $\bar{\omega} \approx 0$ is generally beneficial for showing δ-safety, which is intuitive since $\bar{\omega} = 0$ resembles the case of no disturbance.

Remark 10: It straightforwardly follows from the proof of Proposition 3 that the satisfaction of (12) for $\boldsymbol{x}^+ = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \mathbf{0})$ with $\Delta\xi > 0$ implies the 1-step safety of the system on $\mathbb{V}_{\xi}^{\boldsymbol{\pi}}$ with probability $\delta = \exp(-\beta\Delta\xi)$. The additional conservatism caused by the factor $N\left(\eta\Delta\xi / L_V L_{f,x}\right)$ in (55) can be completely attributed to the stronger notion of safety guaranteed by Proposition 3, which can lead to the fact that a 1-step δ-safe controller under bounded disturbances is not necessarily δ-safe for infinitely many time steps. Therefore, the requirements for safety in the bounded disturbance scenario are generally not more restrictive than in the stochastic noise setting, but stronger notions of safety can be shown.

Remark 11: While we cannot expect to achieve a less conservative result under the considered assumptions as previously discussed, it is clear that a Lipschitz-based analysis can cause significant approximation errors compared to one based on more restrictive assumptions. Therefore, Proposition 3 is best suited for small disturbance bounds $\bar{\omega}$ and systems with minor nonlinearities in the dynamics $\boldsymbol{f}(\cdot,\cdot,\cdot)$. The derivation of improved results under further restrictions on $V_{\boldsymbol{\pi}}(\cdot)$ and $\boldsymbol{f}(\cdot,\cdot,\cdot)$ is left for future research.

Remark 12: Necessary results for safety analogous to Proposition 2 can also be straightforwardly obtained in the bounded disturbance setting by adapting the proof of Proposition 2 similarly as done for the proof of Proposition 3. Therefore, a detailed proof is omitted here for the sake of brevity.

B. Learning Safe Policies under Bounded Disturbances

Similar to Section III-C, merely an approach for certifying the δ-safety of a policy $\boldsymbol{\pi}(\cdot)$ is introduced in Section IV-A. In this section, we show that a suitable policy satisfying the conditions of Proposition 3 can be obtained by solving (28) using the expected cumulative cost function (52) defined for the disturbance-free dynamics. In order to achieve this, we require a bound on the spread of the distribution over functions $\mathcal{F}$ around its mean

$$\boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{f}}\left[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega})\right]. \tag{62}$$

This bound is formalized in the following assumption.

Assumption 5: The expected norm deviation of the stochastic model $\boldsymbol{f}(\cdot,\cdot,\cdot)$ from its mean is bounded by υ, i.e.,

$$\mathbb{E}_{\boldsymbol{f}}\left[\|\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}) - \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega})\|\right] \leq \upsilon \tag{63}$$

for all $\boldsymbol{\omega}$ satisfying Assumption 3.

When the distribution $\mathcal{F}$ is a Gaussian process, then $\mathbb{E}_{\boldsymbol{f}}\left[\|\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}) - \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega})\|_{\boldsymbol{\Sigma}}\right]$ is the mean of a Chi distribution. Therefore, we can straightforwardly bound the left side of (63) by

$$\mathbb{E}_{\boldsymbol{f}}\Big[\|\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}) - \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega})\|\Big] \leq \frac{\Gamma(\frac{1}{2}(d_x+1)}{\Gamma(\frac{1}{2}d_x)\underline{\lambda}(\boldsymbol{\Sigma}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega}))}, \tag{64}$$

where $\Gamma : \mathbb{R} :\to \mathbb{R}$ denotes the gamma function, $\boldsymbol{\Sigma} : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \times \mathbb{R}^{d_\omega} \to \mathbb{R}^{d_x \times d_x}$ outputs the covariance matrix of the Gaussian process, and $\underline{\lambda} : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$ returns the minimum eigenvalue of a matrix. Therefore, we can directly obtain the value of υ in (63) by taking the maximum with respect to $\boldsymbol{x}$ and ω, which underlines the non-restrictiveness of Assumption 5.

Using this assumption, we can show that $\boldsymbol{\pi}_{\text{safe}}$ satisfies the conditions of Proposition 3 under certain assumptions, which is demonstrated in the following theorem. The proof follows after a discussion of the assumptions.

Theorem 3: Consider a cost function $c(\cdot)$ satisfying (7), a stochastic model $\mathcal{F}$ satisfying Assumption 5, a disturbance satisfying Assumption 3, and $L_{f,x}$- and $L_{f,\omega}$-Lipschitz functions $\boldsymbol{f}(\cdot, \boldsymbol{\pi}(\cdot), \boldsymbol{\omega})$ and $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \cdot)$, respectively. Assume that $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$ is L_V-Lipschitz and upper bounded by $\bar{V}$ on $\mathbb{V}_{\boldsymbol{\pi}_{\text{safe}}}^{\bar{\xi}}$. Moreover, 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 (29) is satisfied. If there exists a $\eta \in (0,1)$, such that υ and $\bar{\omega}$ satisfy

$$\frac{\kappa}{2}\Delta c > 2L_V \upsilon \left(N\left(\frac{\eta\kappa\Delta c}{2L_V L_{f,x}}\right)\right)^{\frac{2\bar{V}}{(1-\eta)\kappa\Delta c}} + L_V L_{f,\omega}\bar{\omega}, \tag{65}$$

where

$$\kappa = \frac{\theta_3((\gamma-1)\theta_1+1)}{\gamma\theta_1}, \tag{66}$$

$$\Delta c = \hat{c} - \frac{\theta_2}{\theta_3(\theta_1(\gamma-1)+1)} - \frac{\theta_4}{\theta_3}, \tag{67}$$

and (31) holds for $\hat{c}$, then, the system is safe for all $k \in \mathbb{N}$ with probability $1-\delta \in (0,1)$.

Discussion: The assumption of a Lipschitz continuous optimal value function $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$ is often satisfied in practice and follows immediately from Liptschitz continuity of the dynamics $\boldsymbol{f}(\cdot,\cdot,\cdot)$, the policy parameterization $\boldsymbol{\pi}$ and the cost $c(\cdot)$. Moreover, an upper bound for $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$ on the compact set $\mathbb{V}_{\boldsymbol{\pi}_{\text{safe}}}^{\bar{\xi}}$ follows immediately from Lipschitz continuity. Therefore, these requirements are generally not very restrictive. In comparison to Theorem 1, condition (65) can be considered as a tightening of the constraint $\Delta c > 0$. Since Δc is merely the difference between (31) and $\hat{c}$, this can be interpreted as the requirement of a sufficiently large safety margin. Note that the factor κ introduced in Theorem 3 is only an artifact caused by mapping the condition on $V_{\boldsymbol{\pi}}(\cdot)$ in Proposition 3 to the step cost function $c(\cdot)$.

Similar to Proposition 3, a significant amount of the increased restrictiveness of the assumptions of Theorem 3 is caused by the consideration of δ-safety instead of 1-step δ-safety. In fact, we can simplify condition (65) to

$$\frac{\kappa}{2}\Delta c > 2L_V \upsilon \exp\left(\beta\bar{V}\right) + L_V L_f \bar{\omega}, \tag{68}$$

when we are only concerned with 1-step δ-safety. Analogous to Remark 10, this still causes a slightly tightened constraint due to the term $L_V L_f \bar{\omega}$. Additionally, we now have to consider the spread of the distribution over dynamics, which leads to the summand $2L_V \upsilon \exp\left(\beta\bar{V}\right)$. This simplification clearly illustrates that Theorem 3 can be seen as a tightened version of Theorem 1 to deal with the possibly adversarial disturbance $\boldsymbol{\omega}$.

Proof: In order to prove Theorem 3, we bound the risk $\mathbb{R}_\beta[V(\boldsymbol{x}^+)]$ in terms of the expected cost and the spread of the dynamics υ around their mean. This is shown in the following lemma.

Lemma 4: Consider a probabilistic model $\boldsymbol{f} \sim \mathcal{F}$, such that Assumption 5 holds. If a function $V(\cdot)$ is L_V-Lipschitz and upper bounded by $\bar{V}$, then, the risk is bounded by

$$\mathbb{R}_\beta[V(\boldsymbol{x}^+)] \leq \mathbb{E}[V(\boldsymbol{x}^+)] + 2L_V \upsilon \exp(\beta \bar{V}). \tag{69}$$

Proof: A Taylor expansion of the risk of the cumulative cost around $\mathbb{E}[V(\boldsymbol{x}^+)]$ yields

$$\mathbb{R}_\beta[V(\boldsymbol{x}^+)] \leq \tag{70}$$
$$\frac{1}{\beta} \log\big(\exp(\beta \mathbb{E}[V(\boldsymbol{x}^+)]) + \beta \exp(\beta \bar{V}) \mathbb{E}[|V(\boldsymbol{x}^+) - \mathbb{E}[V(\boldsymbol{x}^+)]|]\big).$$

Since $\log(1+a) \leq a$ for $a \in \mathbb{R}_+$ and $\mathbb{E}[V(\boldsymbol{x}^+)] \geq 0$, we can bound this expression by

$$\mathbb{R}_\beta[V(\boldsymbol{x}^+)] \leq \mathbb{E}\left[V(\boldsymbol{x}^+)\right] + \exp(\beta \bar{V}) \mathbb{E}\left[|V(\boldsymbol{x}^+) - \mathbb{E}[V(\boldsymbol{x}^+)]|\right]. \tag{71}$$

Due to Lispchitz continuity of $V(\cdot)$, we have

$$V(\boldsymbol{x}^+) \leq V(\boldsymbol{\mu}(\boldsymbol{x}, \pi(\boldsymbol{x}), \boldsymbol{\omega})) + L_V \|\boldsymbol{x}^+ - \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega})\|, \tag{72}$$

which we can analogously exploit to determine a lower bound for $V(\boldsymbol{x}^+)$. This immediately leads to

$$\mathbb{E}[|V(\boldsymbol{x}^+) - \mathbb{E}[V(\boldsymbol{x}^+)]|] \leq 2\mathbb{E}[L_V \|\boldsymbol{x}^+ - \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\omega})\|] \tag{73}$$
$$\leq 2L_V \upsilon, \tag{74}$$

where the second line follows from the assumed bound on the expectation. Substituting (73) into (71) concludes the proof. ∎

While we can use the continuity of the risk operator $\mathbb{R}_\beta[\cdot]$ with respect to β to transition from expectation to risk in the proof of Theorem 1, this approach cannot be used for the bounded disturbance case since this noise is not considered in the risk operator. Lemma 4 allows us to circumvent this problem and effectively substitutes Lemma 2. Thereby, it enables the following proof of Theorem 3.

Proof of Theorem 3: Due to Lemma 3, we have

$$\mathbb{E}[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x}^+)] \leq \frac{1}{\gamma}\left(1 - \frac{1}{\theta_1}\right)\bar{\xi} + \frac{\theta_2}{\gamma \theta_1} \tag{75}$$

for all $\boldsymbol{x} \in \mathbb{V}^{\bar{\xi}}_{\boldsymbol{\pi}_{\text{safe}}}$. Moreover, it follows from the proof of Theorem 1 and the definition of Δc in (67) that

$$\bar{\xi} = \frac{\theta_2}{(\theta_1(\gamma - 1) + 1)} + \theta_3 \Delta c. \tag{76}$$

Substituting this expression into (75) and employing Lemma 4, it follows that $\mathbb{R}_\beta[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x}^+)] \leq \xi$, where

$$\xi = \left(\frac{\theta_2}{(\theta_1(\gamma - 1) + 1)} + \frac{\theta_3(\theta_1 - 1)}{\gamma \theta_1} \Delta c\right) + 2L_V \upsilon \exp(\beta \bar{V}). \tag{77}$$

Therefore, we obtain

$$\Delta \xi \geq \kappa \Delta c - 2L_V \upsilon \exp(\beta \bar{V}) - L_V L_{f,\omega} \bar{\omega} \tag{78}$$

for $\Delta \xi$ defined in (54) and κ defined (66), where $\kappa > 0$ is guaranteed since $\theta_1 < \frac{1}{1-\gamma}$ holds by assumption. In order to ensure $\Delta \xi > 0$, we choose

$$\beta = \frac{2}{(1-\eta)\kappa \Delta c} \log\left(N\left(\frac{\eta \kappa \Delta c}{2 L_V L_{f,x}}\right)\right), \tag{79}$$

such that

$$2L_V \upsilon \exp(\beta \bar{V}) + L_V L_{f,\omega} \bar{\omega} < \frac{\kappa}{2} \Delta c \tag{80}$$

due to (65). Moreover, it follows from Proposition 3 that

$$\delta = N\left(\frac{\eta \kappa \Delta c}{2 L_V L_{f,x}}\right) \exp\left(-\frac{1}{2}\beta(1-\eta)\kappa \Delta c\right) < 1, \tag{81}$$

which concludes the proof. ∎

C. Response Inhibition using Robustified Safety Filters

Using the safe policy (28), we propose a robustified risk-sensitive approach for response inhibition to enable safe RL despite bounded process disturbance. Given a nominal, potentially unsafe policy $\boldsymbol{\pi}^*(\cdot)$ obtained, e.g., via (49), this approach determines control inputs $\boldsymbol{u}$ by filtering them through the optimization problem

$$\boldsymbol{\pi}^*_{\text{safe}}(\boldsymbol{x}) = \underset{\boldsymbol{u} \in \mathbb{U}}{\arg\min} \|\boldsymbol{\pi}^*(\boldsymbol{x}) - \boldsymbol{u}\| \tag{82a}$$
$$\text{s.t. } \mathbb{R}_\beta[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{0}))] \leq \xi^* - L_V L_{f,\omega} \bar{\omega}. \tag{82b}$$

The safety filter employs the backup policy $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ indirectly through the cost function $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$. In contrast to the safety filter (50) for the stochastic noise setting, (82) explicitly tightens the safety constraint using the term $L_V L_{f,\omega} \bar{\omega}$ to ensure a sufficient robustness against the disturbance $\boldsymbol{\omega}$. Thereby, it aims to minimize the modification of the nominal control law $\boldsymbol{\pi}^*(\cdot)$ without any knowledge of a distribution of the disturbance $\boldsymbol{\omega}$ to ensure δ-safety as shown in the following theorem.

Theorem 4: Consider a cost function $c(\cdot)$ satisfying (7), a stochastic model $\mathcal{F}$ satisfying Assumption 5, a disturbance satisfying Assumption 3, and $L_{f,x}$- and $L_{f,\omega}$-Lipschitz functions $\boldsymbol{f}(\cdot, \boldsymbol{\pi}(\cdot), \boldsymbol{\omega})$ and $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \cdot)$, respectively. Assume that $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$ is L_V-Lipschitz and upper bounded by $\bar{V}$ on $\mathbb{V}^{\bar{\xi}}_{\boldsymbol{\pi}_{\text{safe}}}$. Moreover, 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 (29) is satisfied. If there exists a $\eta \in (0,1)$, such that υ and $\bar{\omega}$ satisfy (65), then, the safety filtered policy (82) is δ-safe on $\mathbb{V}^{\xi}_{\boldsymbol{\pi}_{\text{safe}}}$.

Proof: Due to Theorem 3, $\boldsymbol{\pi}_{\text{safe}}(\cdot)$ defined in (28) satisfies (82b). Thus, the optimization problem (82) 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 3. ∎

Theorem 4 can be considered the analog to Theorem 2, such that most of the discussion after Theorem 2 applies in the bounded disturbance setting of this section too. Hence, it is also possible to use sub-optimal values for β and ξ, and it is straightforward to design an adaptive strategy for automatically finding suitable values of ξ and β similar to (51a). Consequently, Theorem 4 allows to effectively ensure safety using risk-sensitive response inhibition based on (82b).

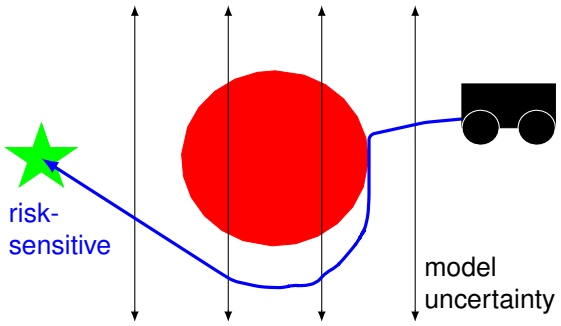

Fig. 2. Toy example illustrating the importance of risk-sensitivity in response inhibition. A vehicle has the goal of reaching its target position (green star), while avoiding a collision with an obstacle (red circle). Our proposed risk-sensitive safety filter restricts exhibits tightened safety constraints only in the directions affected by the uncertainty (black arrows), such that the vehicle can approach the obstacle arbitrarily close in the horizontal direction (blue curve).

V. Simulations

In order to demonstrate the flexible applicability and effectiveness of our risk-sensitive inhibitory control approach, we apply it to multiple problems in simulation. In Section V-A, we first consider the simplified maneuvering problem illustrated in Fig. 2 to illustrate the benefits of risk-sensitivity in response inhibition. We demonstrate the applicability of our proposed methods by applying them to the Half-Cheetah environment [44] and investigate the performance-safety trade-off for conflicting design goals in Section V-B.

A. Illustrative Example for Benefits of Risk-Sensitivity

We illustrate the proposed methodology for determining risk-averse safety filters by considering the example illustrated in Fig. 2, in which a wheeled vehicle needs to reach a target position while avoiding collisions with a circular obstacle. We describe the vehicle through the dynamics

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{\omega}_k) = \boldsymbol{x}_k + v_k \begin{bmatrix} \sin(q_k) \\ \cos(q_k) \end{bmatrix} \tag{83}$$

such that the state $\boldsymbol{x} \in \mathbb{R}^2$ corresponds to the position of a point mass moving in a two-dimensional space. The control input $\boldsymbol{u} = [v \; q]^T$ consists of the vehicle's orientation $q \in \mathbb{R}$ and its velocity $v \in \mathbb{R}_{0,+}$, which we require to be upper-bounded by $v \leq 0.1$. The desired movement to the target position $\boldsymbol{x}_g = [-2.5 \; 0]$ is encoded through a reward function

$$r(\boldsymbol{x}) = -\|\boldsymbol{x} - \boldsymbol{x}_g\|^2. \tag{84}$$

Moreover, we define the circular obstacle via an unsafe region

$$\mathbb{X}_{\text{unsafe}} = \{\boldsymbol{x} \in \mathbb{X} : \|\boldsymbol{x} - \boldsymbol{x}_c\| \leq R\}, \tag{85}$$

where $R = 1$ denotes the radius of the obstacle and $\boldsymbol{x}_c = [0 \; 0]$ is its center. Note that these dynamics are independent of disturbances and we consider a deterministic model knowledge, such that this setting constitutes a special case of the theoretical results for the bounded disturbance scenario.

Since there is no unknown dynamics component, we can manually solve the optimal control problem (49), which leads to the nominal optimal policy

$$\boldsymbol{\pi}^*(\boldsymbol{x}) = \begin{bmatrix} \min\{\bar{v}, \|\boldsymbol{x} - \boldsymbol{x}_g\|\} \\ \arctan2(x_{g,1} - x_1, x_{g,2} - x_2) \end{bmatrix}. \tag{86}$$

The behavior of the vehicle controlled by this policy is illustrated in Fig. 3 a). It can be easily seen that the nominal optimal policy is ignorant of the obstacle and many starting positions will lead to a collision with it. However, the target position is quickly reached from all starting positions.

In order to determine a policy respecting the safety constraints, we define a cost function

$$c(\boldsymbol{x}) = \max\left\{0, \tilde{R} - \|\boldsymbol{x} - \boldsymbol{x}_c\|\right\}, \tag{87}$$

where $\tilde{R} = 2 > R$ defines the maximum value of the cost function and ensures its positiveness. Therefore, safety is ensured if $c(\boldsymbol{x}) < \hat{c} = 1$. Analogously to the optimal policy, we can calculate a safe back-up policy and its value function (6) manually under the assumption of $\zeta = 0$. This yields

$$\boldsymbol{\pi}_{\text{safe}}(\boldsymbol{x}) = \begin{bmatrix} \bar{v} \\ \arctan2(x_1 - x_{c,1}, x_2 - x_{c,2}) \end{bmatrix} \tag{88}$$

$$V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x}) = \sum_{i=0}^{\lfloor (\tilde{R} - \|\boldsymbol{x} - \boldsymbol{x}_c\|)/\bar{v} \rfloor} \gamma^i \left(\tilde{R} - \|\boldsymbol{x} - \boldsymbol{x}_c\| - i\bar{v}\right), \tag{89}$$

where the sum is only over a finite number of time steps since $\tilde{R} - \|\boldsymbol{x} - \boldsymbol{x}_c\| - i\bar{v}$ will be negative after $\lfloor (\tilde{R} - \|\boldsymbol{x} - \boldsymbol{x}_c\|)/\bar{v} \rfloor$ steps. Therefore, this value function can be straightforwardly computed in practice. Applying the safe back-up policy (88) to the system (83) leads to the trajectories illustrated in Fig. 3 b). While the obstacle is avoided for all initial states starting outside of it, the back-up policy is neglecting the target position and yields a poor performance.

To overcome these weaknesses, we employ the safety filter (82). Since the value function $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$ depends only on the distance $\|\boldsymbol{x} - \boldsymbol{x}_c\|$, its sub-level sets are circles. Hence, $\hat{\xi} = 48.87$ can be easily computed by evaluating $V_{\boldsymbol{\pi}_{\text{safe}}}(\cdot)$ for any point $\boldsymbol{x}$ with $\|\boldsymbol{x} - \boldsymbol{x}_c\| = R$. Following Remark 6, this definition induces a barrier function

$$B(\boldsymbol{x}) = \bar{\xi} - V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x}) \eqsim R - \|\boldsymbol{x} - \boldsymbol{x}_c\|, \tag{90}$$

which is similar to the straightforward choice

$$B(\boldsymbol{x}) = R - \|\boldsymbol{x} - \boldsymbol{x}_c\| \tag{91}$$

when directly designing the barrier function. This essentially renders the safety filter (82) based on the value function (89) a CBF constrained optimization problem, which can be commonly found in literature [13], [36]. Therefore, the application of our safety filter with $\xi = 48.8$ to the dynamics (83) expectably leads to a closed-loop behavior with trajectories avoiding the obstacle and converging towards the target position as depicted in Fig. 3 c). Thereby, the strengths of both the nominal optimal policy and the safe back-up policy can be exploited. Note that the trajectories can only approach the obstacle so closely since there is no uncertainty and no disturbances, such that no cautiousness is required.

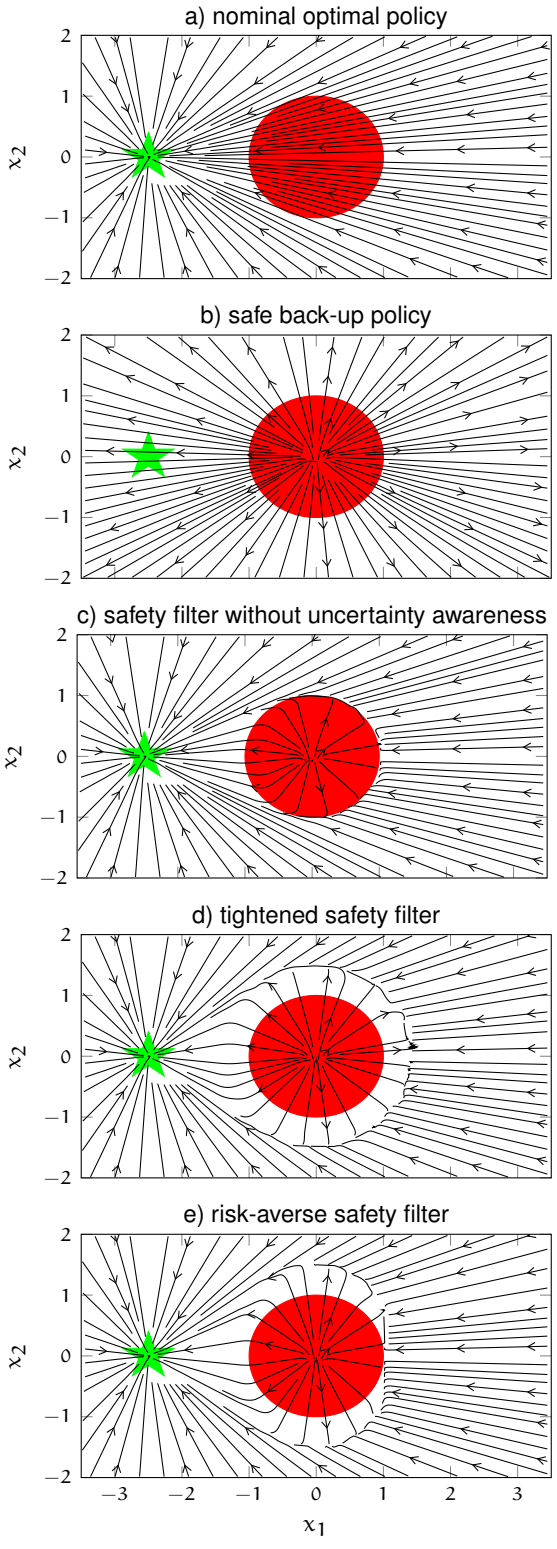

Fig. 3. Illustration of different policies for the toy example with a circular obstacle obstructing the direct path to a target position. While the nominal optimal policy leads to strong safety violations, the safe back-up policy does not achieve the task of moving the vehicle to the target position. By combining both of them through a safety filter, obstacle avoidance and task execution are achieved with low conservatism when the dynamics are known perfectly. Extending this behavior to probabilistic models through a tightening of constraints leads to a conservative behavior around the obstacle which ignores knowledge about the uncertainty distribution. In contrast, risk-aversion in the safety filter exploits this knowledge to only tighten the safety constraint in directions where it is necessary to account for the uncertainty.

This changes when we consider uncertain dynamics

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{\omega}_k) + \begin{bmatrix} 0 \\ d(\boldsymbol{x}_k, \boldsymbol{u}_k) \end{bmatrix} \tag{92}$$

in the safety filter (82), where we assume that $d(\boldsymbol{x}, \boldsymbol{u}) = (1 + v)\zeta$, $\zeta \sim \mathcal{U}([-0.5, 0.5])$, represents an unknown component that only affects the dynamics in the vertical direction. A common approach in the literature on CBFs to deal with such an uncertainty is to ensure robustness of the control law by tightening the safety constraints [45]. This essentially corresponds to choosing a smaller value $\xi = 15$, which results in the behavior shown in Fig. 3 d) when applying the safety filter to (83). The increased robustness keeps the trajectories farther away from the obstacle, but the direct tightening of the constraint is ignorant of the direction of the uncertainty. Therefore, the resulting safety filtered policy is conservative.

This conservatism can be avoided by making the safety filter risk-averse, which corresponds to choosing a comparatively large value $\beta = 10$. The resulting behavior of the closed-loop system can be seen in Fig. 3 e), where the trajectories only stay away from the obstacle in the vertical direction, while they approach it in the horizontal direction similarly as for the safety filter with the exact dynamics depicted in Fig. 3 c). This can be easily explained by the fact that uncertainty in $\boldsymbol{x}_{k+1}$ tangential to the level sets of the value function barely affects the value of the risk $\mathbb{R}_\beta[V_{\boldsymbol{\pi}_{\text{safe}}}(\boldsymbol{x}_{k+1})]$, such that independent actuation for every state in (83) allows the safety filter to push the system away from the obstacle only in the vertical direction. Therefore, the risk-awareness only tightens the safety constraint in the direction, in which the uncertainty acts and thereby, significantly reduces the overall conservatism of the resulting policy.

B. Safety-Performance Trade-off in Response Inhibition

While we can construct value functions or barrier functions for simple problems such as the vehicle control example in Section V-A by hand, a manual design is generally challenging or not even possible for complex dynamics and constraints. This does not pose a hurdle for our risk-sensitive response inhibition approach since we can infer suitable value functions using RL techniques. We demonstrate this by computing and applying the safety filter (50) to the popular Mujoco Half-Cheetah environment [44], which is illustrated in Fig. 4. 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 illustrate the trade-off between safety and performance realized by risk inhibition, we set optimality and safety in a direct conflict by constraining the velocity to $v \leq v_{\text{crit}}$, $v_{\text{crit}} = 2$. As cost function for the computation of the safe policy (28), $c(\boldsymbol{x}) = v - \underline{v}$ 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

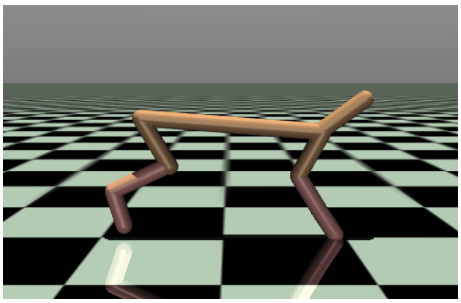

Fig. 4. Illustration of the Half-Cheetah model, which is a planar abstraction of a cat-like robot with 6 degrees of freedom.

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 [37] with 400 training iterations each with 1000 time steps and the hyper-parameters provided by [46]. For computing the expectations over dynamics $\boldsymbol{f}(\cdot)$ in (6) and (49), 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 (50) is implemented using the cross-entropy method [47] with 5 iterations per time step and 10 particles. The safety constraints are considered in an augmented objective function using fixed Lagrange 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 5 random seeds are averaged. Our response inhibition approach is compared to an adaptation of the model predictive controller in [48] to a predictive safety filter. We employ this safety filter with a horizon 10 over which the original state constraint $v \leq \bar{v}$ is enforced through a soft constraint with Lagrange multiplier 10^4. The optimization is executed using the cross-entropy method with 100 particles to account for the more difficult optimization problem.

The resulting numbers of constraint violations and the average reward for different values of β and ξ are depicted in Fig. 5. 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 Section III-A, such that they allow to naturally balance conservatism and the number of constraint violations. The predictive safety

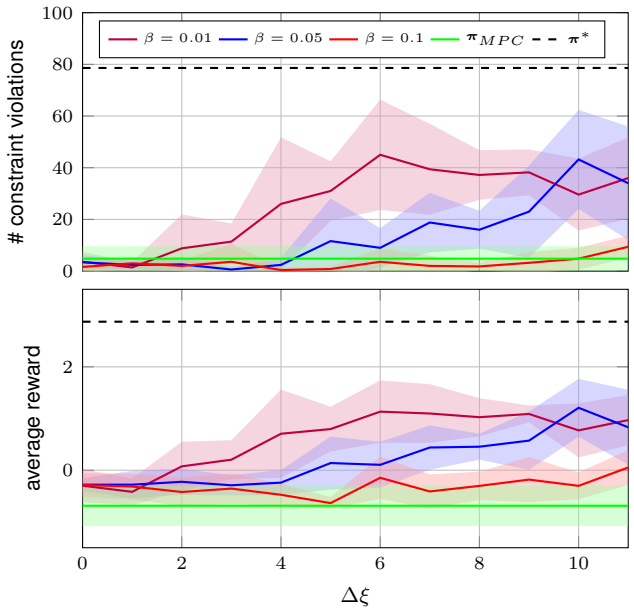

Fig. 5. Number of constraint violations and average rewards in dependency on the safety constraint threshold $\boldsymbol{\xi} = \mathbf{605} + \boldsymbol{\Delta\xi}$ and the risk-sensitivity $\boldsymbol{\beta}$. Shaded areas represent the standard deviation between the simulations. Reducing $\boldsymbol{\beta}$ and increasing $\boldsymbol{\xi}$ 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. The predictive safety filter $\boldsymbol{\pi}_{MPC}$ leads to more conservative behavior in comparison.

filter achieves a similar number of constraint violations as our approach for risk inhibition with large risk sensitivity β and safety threshold ξ, but it yields smaller average rewards. This observation can be attributed to the significantly more challenging optimization problem resulting from the longer prediction horizon, such that even the considerably larger number of particles used for optimization does not suffice to obtain a comparable trade-off between safety and performance as our proposed risk-inhibition approach. Combined with a lack of strong theoretical guarantees for such a predictive safety filter that avoids hard-to-design terminal constraints, this example clearly highlights the beneficial properties of the proposed method for risk inhibition.

VI. 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 RL 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] 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

[4] K.-C. Hsu, H. Hu, and J. F. Fisac, "The safety filter: A unified view of safety-critical control in autonomous systems," 2023.

[5] K. P. Wabersich, A. J. Taylor, J. J. Choi, K. Sreenath, C. J. Tomlin, A. D. Ames, and M. N. Zeilinger, "Data-driven safety filters: Hamilton-jacobi reachability, control barrier functions, and predictive methods for uncertain systems," *IEEE Control Systems Magazine*, vol. 43, no. 5, pp. 137–177, 2023.

[6] 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.

[7] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *Journal of Machine Learning Research*, vol. 18, pp. 1–51, 2018.

[8] S. Paternain, L. F. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 7555–7565.

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

[10] 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.

[11] 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.

[12] A. Lin and S. Bansal, "Generating formal safety assurances for high-dimensional reachability," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 10 525–10 531.

[13] A. Agrawal and K. Sreenath, "Discrete control barrier functions for safety-critical control of discrete systems with application to bipedal robot navigation," in *Robotics: Systems and Science*, 2017.

[14] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control Barrier Function Based Quadratic Programs for Safety Critical Systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3861–3876, 2017.

[15] C. Santoyo, M. Dutreix, and S. Coogan, "A barrier function approach to finite-time stochastic system verification and control," *Automatica*, vol. 125, p. 109439, 2021.

[16] T. Badings, A. Abate, N. Jansen, D. Parker, H. Poonawala, and M. Stoelinga, "Sampling-based robust control of autonomous systems with non-gaussian noise," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 9669–9678.

[17] R. Mazouz, K. Muvvala, A. Ratheesh Babu, L. Laurenti, and M. Lahijanian, "Safety guarantees for neural network dynamic systems via stochastic barrier functions," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9672–9686, 2022.

[18] M. Srinivasan, A. Dabholkar, S. Coogan, and P. A. Vela, "Synthesis of control barrier functions using a supervised machine learning approach," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 7139–7145.

[19] C. Dawson, Z. Qin, S. Gao, and C. Fan, "Safe nonlinear control using robust neural lyapunov-barrier functions," in *Proceedings of the 5th Conference on Robot Learning*, 2022, pp. 1724–1735.

[20] A. Robey, H. Hu, L. Lindemann, H. Zhang, D. V. Dimarogonas, S. Tu, and N. Matni, "Learning control barrier functions from expert demonstrations," in *2020 59th IEEE Conference on Decision and Control (CDC)*, 2020, pp. 3717–3724.

[21] H. Bharadhwaj, A. Kumar, N. Rhinehart, S. Levine, F. Shkurti, and A. Garg, "Conservative safety critics for exploration," in *Proceedings of the International Conference on Learning Representations*, 2021.

[22] B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg, "Recovery rl: Safe reinforcement learning with learned recovery zones," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4915–4922, 2021.

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

[24] K. Long, Y. Yi, J. Cortés, and N. Atanasov, "Safe and stable control synthesis for uncertain system models via distributionally robust optimization," in *Proceedings of the American Control Conference*, 2023, pp. 4651–4658.

[25] 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.

[26] 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.

[27] A. Lederer, E. Noorani, J. S. Baras, and S. Hirche, "Risk-sensitive inhibitory control for safe reinforcement learning," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 1040–1045.

[28] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer Science+Business Media, 2006.

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

[30] 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.

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

[32] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 2002.

[33] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton, New Jersey: Princeton University Press, 2009.

[34] G. Calafiore and M. Campi, "The scenario approach to robust control design," *IEEE Transactions on Automatic Control*, vol. 51, no. 5, pp. 742–753, 2006.

[35] 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.

[36] M. Ahmadi, A. Singletary, J. W. Burdick, and A. D. Ames, "Safe policy synthesis in multi-agent pomdps via discrete-time barrier functions," in *Proceedings of the IEEE Conference on Decision and Control*, 2019, pp. 4979–4803.

[37] 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.

[38] 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.

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

[40] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. New York, NY: Cambridge University Press, 2013, vol. 9781107057.

[41] A. Lederer, J. Umlauft, and S. Hirche, "Uniform Error and Posterior Variance Bounds for Gaussian Process Regression with Application to Safe Control," 2021. [Online]. Available: http://arxiv.org/abs/2101.05328

[42] F. Castañeda, J. J. Choi, B. Zhang, C. J. Tomlin, and K. Sreenath, "Pointwise Feasibility of Gaussian Process-based Safety-Critical Control under Model Uncertainty," 2021. [Online]. Available: http://arxiv.org/abs/2106.07108

[43] V. Dhiman, M. J. Khojasteh, M. Franceschetti, and N. Atanasov, "Control Barriers in Bayesian Learning of System Dynamics," *IEEE Transactions on Automatic Control*, pp. 1–16, 2021.

[44] 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.

[45] W. Shaw Cortez, D. Oetomo, C. Manzie, and P. Choong, "Control Barrier Functions for Mechanical Systems: Theory and Application to Robotic Grasping," *IEEE Transactions on Control Systems Technology*, vol. 29, no. 2, pp. 530–545, 2021.

[46] 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.

[47] 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.

[48] Z. Liu, H. Zhou, B. Chen, S. Zhong, M. Hebert, and D. Zhao, "Constrained model-based reinforcement learning with robust cross-entropy method," in *ICLR Workshop on Security and Safety in Machine Learning Systems*, 2021.

Armin Lederer received the B.Sc., M.Sc., and Dr. Ing. degree in electrical and computer engineering from the Technical University of Munich, Germany, in 2015, 2018, and 2023, respectively. Since 2023, he is a postdoctoral researcher at the Learning and Adaptive Systems Group in the Department of Computer Science at ETH Zurich, Switzerland. He was finalist for the European Systems & Control PhD Thesis Award 2024, and has received the Chorafas Foundation Award and the Rohde & Schwarz PhD Thesis Award. His current research interests include the stability and safety of data-driven control systems and machine learning in closed-loop systems.

Erfaun Noorani is a Technical Staff at MIT Lincoln Laboratory. Before joining the Laboratory, he was a Postdoctoral Associate within the Institute for Systems Research (ISR) at the University of Maryland, College Park, where he received his Ph.D. and M.S. as a Clark Doctoral Fellow with the Department of Electrical and Computer Engineering. His doctoral work, under the supervision of John Baras, focused on robust and risk-sensitive reinforcement learning. Erfaun obtained a B.Sc. degree in Electrical Engineering from Drexel University, Philadelphia, Pennsylvania.

John S. Baras received the Diploma degree in electrical and mechanical engineering from the National Technical University of Athens, Athens, Greece, in 1970, and the M.S. and Ph.D. degrees in applied mathematics from Harvard University, Cambridge, MA, USA, in 1971 and 1973, respectively. He is a Distinguished University Professor and holds the Lockheed Martin Chair in Systems Engineering, with the Department of Electrical and Computer Engineering and the Institute for Systems Research (ISR), at the University of Maryland College Park. From 1985 to 1991, he was the Founding Director of the ISR. Since 1992, he has been the Director of the Maryland Center for Hybrid Networks (HYNET), which he co-founded. His research interests include systems and control, optimization, communication networks, applied mathematics, machine learning, artificial intelligence, signal processing, robotics, computing systems, security, trust, systems biology, healthcare systems, model-based systems engineering. Dr. Baras is a Fellow of IEEE (Life), SIAM, AAAS, NAI, IFAC, AMS, AIAA, Member of the National Academy of Inventors and a Foreign Member of the Royal Swedish Academy of Engineering Sciences. Major honors include the 1980 George Axelby Award from the IEEE Control Systems Society, the 2006 Leonard Abraham Prize from the IEEE Communications Society, the 2017 IEEE Simon Ramo Medal, the 2017 AACC Richard E. Bellman Control Heritage Award, the 2018 AIAA Aerospace Communications Award. In 2016 he was inducted in the A. J. Clark School of Engineering Innovation Hall of Fame. In 2018 he was awarded a Doctorate Honoris Causa by his alma mater the National Technical University of Athens, Greece.

Sandra Hirche received the Dipl.-Ing degree in aeronautical engineering from the Technical University of Berlin, Berlin, Germany, in 2002, and the Dr. Ing. degree in electrical engineering from the Technical University of Munich, Munich, Germany, in 2005. From 2005 to 2007, she was awarded a Post-doctoral scholarship from the Japanese Society for the Promotion of Science at the Fujita Laboratory, Tokyo Institute of Technology, Tokyo, Japan. From 2008 to 2012, she was an Associate Professor with the Technical University of Munich. Since 2013, she has served as Technical University of Munich Liesel Beckmann Distinguished Professor and has been with the Chair of Information-Oriented Control, Department of Electrical and Computer Engineering, Technical University of Munich. She has authored or coauthored more than 150 papers in international journals, books, and refereed conferences. Her main research interests include cooperative, distributed, and networked control with applications in human–machine interaction, multirobot systems, and general robotics. Dr. Hirche has served on the editorial boards of the IEEE Transactions on Control of Network Systems, the IEEE Transactions on Control Systems Technology, and the IEEE Transactions on Haptics. She has received multiple awards such as the Rohde & Schwarz Award for her Ph.D. thesis, the IFAC World Congress Best Poster Award in 2005, and – together with students – the 2018 Outstanding Student Paper Award of the IEEE Conference on Decision and Control as well as Best Paper Awards from IEEE Worldhaptics and the IFAC Conference of Manoeuvring and Control of Marine Craft in 2009.