Probabilistic Safety Guarantee for Stochastic Control Systems Using Average Reward MDPs

Abstract

Safety in stochastic control systems, which are subject to random noise with a known probability distribution, aims to compute policies that satisfy predefined operational constraints with high confidence throughout the uncertain evolution of the state variables. The unpredictable evolution of state variables poses a significant challenge for meeting predefined constraints using various control methods. To address this, we present a new algorithm that computes safe policies to determine the safety level across a finite state set. This algorithm reduces the safety objective to the standard average reward Markov Decision Process (MDP) objective. This reduction enables us to use standard techniques, such as linear programs, to compute and analyze safe policies. We validate the proposed method numerically on the Double Integrator and the Inverted Pendulum systems. Results indicate that the average-reward MDPs solution is more comprehensive, converges faster, and offers higher quality compared to the minimum discounted-reward solution.

Keywords: Safety Critical Systems, Robotics, Average Reward MDPs, Stochastic Control.

1. Introduction

Safety-critical algorithms are a vital requirement for stochastic control systems deployed in fields such as autonomous robots, quadrotors, and self-driving cars to prevent injuries or financial losses (Dawood et al., 2024; Zhong et al., 2025; Soleimani et al., 2025). Traditionally, safety problems are often framed as a reach-avoid problem, where the agent is guaranteed to reach the intended goal while avoiding actions that may lead to undesirable states (Compton et al., 2024; Rabiee and Safari, 2023; Wabersich and Zeilinger, 2021; Alan et al., 2023). These approaches achieve safety by treating the domain's stochasticity using a worst-case adversarial, or robust, approach that can tractably compute safe policies by solving the Hamilton-Jacobi-Isaacs (HJI) equation (Chen et al., 2017; Margellos and Lygeros, 2011; Moon, 2022; Arnström and Teixeira, 2024). Although these standard approaches have been used successfully in some domains, they often struggle in domains with significant uncertainty (Dallas et al., 2025).

In this paper, we propose a new approach to computing safe policies in stochastic control by reducing the problem to the average-reward criterion in Markov Decision Process (MDP) (Puterman, 1990). MDPs represent a flexible framework that is used to model reinforcement learning problems. In particular, we make the following contributions in this paper. (1) We show that average-reward MDPs provide a direct and computationally efficient approach for calculating the probabilistic safety value function using linear programs. (2) Our average-reward-based approach uses standard tools to dispense with the need for artificial discount factors used in prior work Akametalu et al. (2023). Previously, discount factors were used to improve the computational complexity of computing the fixed point. (3) Our reduction makes it possible to compute states that are safe with high confidence, rather than just computing safe and unsafe states.

Although there is a rich literature on safety, the analysis of probabilistic safety remains insufficiently addressed, particularly when robust control methods are overly conservative or when worst-case scenarios cannot be precisely characterized. Overly conservative control policies significantly constrain the operational range of real-world stochastic systems, where high-risk events are infrequent. Moreover, MDPs commonly employ a discount factor in safety analyses; however, even average reward frameworks for reach-avoid problems do not guarantee probabilistic safety in the system's long-term behavior.

Our approach departs significantly from common approaches to safety in stochastic control Chen et al. (2017). Most existing literature on the subject guarantees safety by reducing the stochastic outcomes to adversarial noise. The adversarial approach makes it possible to rely on deterministic differential game theory tools to formulate and solve the safety problem (Margellos and Lygeros, 2011; Moon, 2022; Arnström and Teixeira, 2024). Central to this framework is the solution of the Hamilton-Jacobi-Isaacs (HJI) equation (Evans and Souganidis, 1984; Abu-Khalaf et al., 2006; Akametalu et al., 2023; Begzadić et al., 2025). For instance, Fisac et al. (2018) introduce an HJI-based safety framework as a variational inequality, guaranteeing constraint satisfaction while minimizing disruption to the learning process. This framework also incorporates a Bayesian mechanism to adaptively update the safety analysis as new data is collected.

In related work, Ávila and Junca (2021) also reduce infinite-horizon reach-avoid problems to the average-reward criterion for MDPs. Similarly to our approach, they probabilistically characterize reachability and reach-avoid solutions, inspired by the stochastic target problem (Soner and Touzi, 2002). However, our safety objective cannot be immediately cast as a reach-avoid problem. Our work is also related to Gao et al. (2023), which addresses the limitations of traditional MDP reachability to account for transient distributions explicitly.

The remainder of the paper is organized as follows. Section 2 introduces the preliminaries for safety in uncertain control. Section 3 presents the Average Reward MDP criterion for the probabilistic safety problem. Section 4 reviews related work in the HJI framework. Section 5 presents a numerical validation, and Section 6 concludes the paper.

2. Problem Statement

The analysis of safety-critical systems requires a formal method to guarantee the satisfaction of state and input constraints even when the system is subject to stochastic disturbances. This section presents the foundational framework for providing probabilistic safety guarantees in control systems affected by disturbances. We seek to identify a set of safe states that account for the system's uncertain response to control actions.

Notation: Vectors and matrices are denoted by bold lowercase and uppercase letters, respectively. Random variables are distinguished by the use of tildes. Subscripts generally indicate time. The symbols $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{N}$ denote the sets of real, non-negative real, and natural numbers, respectively. We use $\Delta_n \subseteq \mathbb{R}^n$ for $n \in \mathbb{N}$ to represent the probability simplex. To this end, we assume a probability space $(\Omega, \mathcal{B}, p)$ and we us $\mathbb{X}_{\mathcal{S}}$ to denote the set of all $\mathcal{S}$-valued random variables.

For finite and non-empty state set $\mathcal{S} = \{1, 2, 3, \ldots, S\}$ and action set $\mathcal{A} = \{1, 2, 3, \ldots, A\}$, denote the state and action variables at time $k \in \mathbb{N}$ as $s_k \in \mathcal{S}$ and $a_k \in \mathcal{A}$, respectively. The evolution of the state over time is influenced by an external disturbance d, which takes values in the set $\mathcal{D} = \{1, 2, 3, \ldots, D\}$. The disturbance function $d \colon \mathcal{S} \to \mathcal{D}$ depends on the current state. The system dynamics are described by a known difference equation, where the function $f \colon \mathcal{S} \times \mathcal{A} \times \mathcal{D} \to$

$\mathcal{S}$ maps s_k, a_k, and $d(s_k)$ to the next state s_{k+1}:

$$s_{k+1} = f(s_k, a_k, d(s_k)). \tag{1}$$

A random disturbance is represented by a random variable $\tilde{d}\colon \mathcal{S} \to \mathbb{X}_{\mathcal{D}}$ where $\mathbb{X}_{\mathcal{D}}$ is the set of all $\mathcal{D}$-valued, $\mathcal{B}$-measurable random variables. Let $\pi\colon \mathcal{S} \to \mathcal{A}$ denote a deterministic policy that prescribes a control action for each state, and define the set of all stationary deterministic policies as $\Pi_{\mathrm{SD}} := \mathcal{A}^{\mathcal{S}}$. For each policy $\pi \in \Pi_{\mathrm{SD}}$, we define a stochastic process $\tilde{x}_k\colon \mathcal{S} \to \mathbb{X}_{\mathcal{S}}$ where the system dynamics in (1) are reformulated as a stochastic process starting from a state $s \in \mathcal{S}$, and the probability of the state transition equals,

$$\mathbb{P}_s^\pi \left[\tilde{x}_{k+1} = f\big(\tilde{x}_k, \pi(\tilde{x}_k),\, \tilde{d}(\tilde{x}_k)\big),\, \forall k \in \mathbb{N}\right] = 1, \qquad \mathbb{P}_s^\pi \left[\tilde{x}_0 = s\right] = 1. \tag{2}$$

Here, $\mathbb{P}_s^\pi$ represents the probability measure over the system's future state trajectories, given that it starts in state s and follows a specific policy π. The random variables are defined in Equation (2) such as $\tilde{x}_k$ are measurable with respect to the standard filtration of the stochastic process (Jazwinski, 2007). After defining Equation (2), we now provide definitions of the constraint and safe sets.

Definition 1 *The constraint set $\mathcal{C} \subseteq \mathcal{S}$ represents admissible states.*

While $\mathcal{C}$ defines all instantaneously admissible states, we seek a subset of $\mathcal{C}$ from which a policy can guarantee evolution of Equation (2) remains within $\mathcal{C}$ indefinitely with a certain level of probability.

Definition 2 *The* probabilistically-safe *set $\mathcal{K}_\alpha \subseteq \mathcal{S}$ with confidence $\alpha \in [0, 1]$ comprises states for which there exists a $\pi \in \Pi_{\mathrm{SD}}$ that guarantees that the process remains in the constraint set $\mathcal{C}$ with probability of α:*

$$\mathcal{K}_\alpha := \left\{s \in \mathcal{S} \mid \exists \pi \in \Pi_{\mathrm{SD}}, \mathbb{P}_s^\pi \left[\tilde{x}_k \in \mathcal{C}, \forall k \in \mathbb{N}\right] \geq \alpha\right\}. \tag{3}$$

Let $\mathcal{K} := \mathcal{K}_1$ be the safe set.

It is easy to see that

$$1 \geq \alpha_1 \geq \alpha_2 \geq 0 \quad \Longrightarrow \quad \mathcal{K}_{\alpha_1} \subseteq \mathcal{K}_{\alpha_2}.$$

The notation $\mathbb{P}_s^\pi \left[\tilde{x}_k \in \mathcal{C}, \forall k \in \mathbb{N}\right]$ specifies that the entire future trajectory evolving from $\tilde{x}_k$ under policy π must remain within $\mathcal{C}$, ensuring safety over the entire infinite time horizon. The objective of the probabilistic safety problem, therefore, is to compute the probabilistic safe set $\mathcal{K}_\alpha$ and then find a policy $\pi \in \Pi_{SD}$ that can gurantee the system evolution according to Equation (2) remains in $\mathcal{C}$ with α level confidence.

3. Average Reward MDPs for Probabilistic Safety

This section addresses the probabilistic safety problem within the multichain average reward MDP. First, we define the components and properties of the specialized MDP model. Second, we provide lemmas and proofs to formally connect the average reward framework to the probabilistic safety problem. Finally, we establish the connection between average reward MDPs and linear programs.

Consider the MDP $(\mathcal{S}, \mathcal{A}, p, r, \mu)$ where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action sets, respectively. The function $p\colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the *transition-probability function*, where $p(s, a, s')$ indicates the probability of transitioning to the next state s' from the current state s under action a. We define the transitions as

$$p(s,a,s') := \begin{cases} \mathbb{P}\left[s' = f(s,a,\tilde{d}(s))\right] & \text{if } s \in \mathcal{C}, \\ \mathbb{1}\{s' = s\} & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathbb{1}$ denotes the indicator function. If the MDPs state s is admissible ($s \in \mathcal{C}$), the next state s' probabilistically follows the outcome of the system dynamics, $f(s,a,\tilde{d}(s))$, as described in Equation (1). Conversely, any occurrence of a violation ($s_k \notin \mathcal{C}$) results in all future states remaining inadmissible (i.e., they become absorbing states). The motivation for the construction in Equation (4) is to identify safety violations in the entire horizon by system evolution in Equation (2), as there is a possibility that the system evolution enters the inadmissible states and returns to the admissible state in the infinite horizon.

The *reward function* $r\colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ represents the immediate reward. In particular, $r(s,a,s')$ is the reward received when taking an action a in state s and transitioning to s'. We define the immediate reward function to reflect the safety objective as follows:

$$r(s,a,s') \ := \mathbb{1}\{s \in \mathcal{C}\}. \tag{5}$$

Definition in Equation (5) means the reward is 1 when the system is in a safe state and 0 otherwise, and the definition of reward function in Equation (5) will encourage the average reward MDP system to remain within the admissible state region. An initial distribution over $\mathcal{S}$ is represented by $\mu \in \Delta_{\mathcal{S}}$, and it is assumed that $\mu(s) > 0$ for all $s \in \mathcal{S}$.

The objective for the *average reward criterion* is:

$$\sup_{\pi \in \Pi_{\mathrm{SD}}} \limsup_{N \to \infty} \frac{1}{N} \mathbb{E}^{\pi}_{\mu} \left[\sum_{k=0}^{N-1} \mathbb{1}\{\tilde{s}_k \in \mathcal{C}\} \right], \tag{6}$$

where the superscript π denotes a policy from the stationary policy set Π_{SD}. Random variable $\tilde{s}_k$ represents the state of the MDP system at time k [1]. For the average reward criterion, the *gain* function (known as the average-reward value function) $g : \mathcal{S} \to \mathbb{R}$ and the *bias* function $h : \mathcal{S} \to \mathbb{R}$ are defined as:

$$g^{\pi}(s) = \limsup_{N \to \infty} \frac{1}{N} \mathbb{E}^{\pi}_{\mu} \left[\sum_{k=0}^{N-1} \mathbb{1}\{\tilde{s}_k \in \mathcal{C}\} \right], \quad h^{\pi}(s) = \limsup_{N \to \infty} \mathbb{E}^{\pi}_{\mu} \left[\sum_{k=0}^{N-1} (\mathbb{1}\{\tilde{s}_k \in \mathcal{C}\} - g^{\pi}(\tilde{s}_k)) \right]. \tag{7}$$

The optimal gain and the optimal policy are:

$$g^{\star} = \max_{\pi \in \Pi_{\mathrm{SD}}} g^{\pi}(s), \quad \pi^{\star} \in \arg\max_{\pi \in \Pi_{\mathrm{SD}}} g^{\pi}(s). \tag{8}$$

The gain function in Equation (7) is equivalent to the probabilistic safety value function, a key identity that tightly couples the system's long-run behavior to this value; as $N \to \infty$, this gain correctly becomes zero upon a state constraint violation. Within the multichain average-reward framework, the bias term is also necessary to satisfy the optimality equations, capturing the system's transient behavior while the gain captures its long-term behavior. The fundamental equations formally define

[1] To be clear in this paper, $\tilde{x}_k$ refers to the state trajectory evolution according to Equation (3) while $\tilde{s}_k$ is the trajectory of the MDP model when Equation (4) defined to formulate the probabilistic safety problem.

gain and bias, connecting them to the system's underlying transitions (Puterman, 1990, Theorem 8.2.6). Moreover, the multichain formulation is necessary because the optimal stationary policy may not be unique in Equation (8), leading to multiple recurrent classes (Puterman, 1990, Section 9.1.3).

The following theorem characterizes the probabilistic safe set in terms of the optimal gain function, establishing a fundamental link between probabilistic safety and long-run average rewards under optimal policies.

Theorem 3 *For every state $s \in S$ and confidence level $\alpha \in [0, 1]$:*

$$s \in \mathcal{K}_\alpha \iff \alpha \leq g^\star(s).$$

Proof Suppose that $s \in \mathcal{K}_\alpha$. There exists a policy $\hat{\pi} \in \Pi_{\text{SD}}$ such that for each $N \in \mathbb{N}$

$$\alpha \leq \mathbb{P}_s^{\hat{\pi}} \left[\tilde{x}_k \in \mathcal{C}, \ \forall k \in \mathbb{N}\right] = \mathbb{P}_s^{\hat{\pi}} \left[\tilde{s}_k \in \mathcal{C}, \ \forall k \in \mathbb{N}\right] \leq \mathbb{P}_s^{\hat{\pi}} \left[\tilde{s}_k \in \mathcal{C}, \ \forall k = 0, \dots, N-1\right]$$

Therefore, by the definition of gain and the fact that all rewards are non-negative, we get that

$$\begin{aligned} g^\star(s) \geq g^{\hat{\pi}}(s) &= \limsup_{N\to\infty} \frac{1}{N} \mathbb{E}_\mu^{\hat{\pi}} \left[\sum_{k=0}^{N-1} \mathbb{1}\{\tilde{s}_k \in \mathcal{C}\}\right] \\ &\geq \limsup_{N\to\infty} \frac{1}{N} N \cdot \mathbb{P}_s^{\hat{\pi}} \left[\tilde{s}_k \in \mathcal{C}, \ \forall k = 0, \dots, N-1\right] \geq \limsup_{N\to\infty} \frac{1}{N} N \cdot \alpha = \alpha. \end{aligned}$$

Now consider the optimal gain and summation of the Bellman equation over the entire $\mathcal{S}$:

$$g^\star(s) = \sum_{s'\in\mathcal{K}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right] g^\star(s') + \sum_{s'\in\mathcal{C}\setminus\mathcal{K}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right] g^\star(s') + \sum_{s'\in\mathcal{S}\setminus\mathcal{C}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right] g^\star(s').$$

According to Lemma 5 and Lemma 6 in Appendix A, we know that $g^\star(s') = 0$ if $s' \in \mathcal{S} \setminus \mathcal{C}$, and $g^\star(s') = 1$ if $s' \in \mathcal{K}$. Therefore,

$$g^\star(s) = \sum_{s'\in\mathcal{K}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right] \cdot 1 + \sum_{s'\in\mathcal{C}\setminus\mathcal{K}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right] g^\star(s').$$

Since $\mathcal{C} \setminus \mathcal{K}$ is a transient set based on Lemma 4 (see Appendix A), we have $\lim_{k\to\infty} \mathbb{P}_s^{\pi^\star} [\tilde{s}_k = s'] = 0, \quad \forall s' \in \mathcal{C} \setminus \mathcal{K}$. Taking the limit:

$$g^\star(s) = \lim_{k\to\infty} \left(\sum_{s'\in\mathcal{K}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right] \cdot 1 + \sum_{s'\in\mathcal{C}\setminus\mathcal{K}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right] g^\star(s') \right) = \lim_{k\to\infty} \sum_{s'\in\mathcal{K}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right].$$

Therefore, the limiting probability of being in $\mathcal{C}$ is entirely concentrated on $\mathcal{K}$:

$$\lim_{k\to\infty} \sum_{s'\in\mathcal{C}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right] = \sum_{s'\in\mathcal{K}} \lim_{k\to\infty} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right].$$

By the definition of the probabilistically–safe set

$$\mathbb{P}^{\hat{\pi}} \left[\tilde{x}_k(s) \in \mathcal{C}, \ \forall k \in \mathbb{N}\right] = \lim_{k\to\infty} \sum_{s'\in\mathcal{K}} \mathbb{P}_s^{\pi^\star} \left[\tilde{s}_k = s'\right] \ \geq \ \alpha.$$

This limiting sum defines the probability of eventually reaching a state in $\mathcal{C}$, which, by definition, equals $g^{\star}(s)$. Since $g^{\star}(s) \geq \alpha$, the condition is satisfied, and this completes the proof. ■

Intuitively, Theorem 3 states that the probability of staying safe at confidence level α is connected by the α probability of returning from a current transient state to the safe state in $\mathcal{K}$.

With the theoretical connection between multichain average reward MDPs and probabilistic safety established, we can now leverage the properties of infinite average reward MDPs to formulate this safety problem as a corresponding linear program. The infinite-horizon average reward problem proposed in Equation (8) can be solved using the following primal linear program (Puterman, 2005, Section 9.3):

$$\begin{aligned} \underset{g\in\mathbb{R}_+^{\mathcal{S}},\, h\in\mathbb{R}_+^{\mathcal{S}}}{\text{minimize}} \quad & \sum_{s\in\mathcal{S}} \alpha_s\, g(s) \\ \text{s. t.} \quad & g(s) \geq \sum_{s'\in\mathcal{S}} g(s')\, p(s,a,s'), \quad \forall\, s\in\mathcal{S}, a\in\mathcal{A} \\ & g(s)+h(s) \geq \mathbb{1}\{s_k\in\mathcal{C}\}+\sum_{s'\in\mathcal{S}} h(s')\, p(s,a,s'), \qquad \forall\, s\in\mathcal{S}, a\in\mathcal{A} \end{aligned} \tag{9}$$

where the value $\alpha_{s_0} \in \mathbb{R}_+^{\mathcal{S}}$ may be arbitrary as long as it satisfies that $\alpha_{s_0} > 0$, $\forall s \in \mathcal{S}$, $\sum_{s\in\mathcal{S}} \alpha_s = 1$. To determine the corresponding actions, solving the dual linear program is also necessary. Equation (6) can be solved using the following dual linear program.

$$\begin{aligned} \underset{z\in\mathbb{R}_+^{\mathcal{S}\times\mathcal{A}},\, y\in\mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}}{\text{maximize}} \quad & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} z(s,a)\mathbb{1}\{s_k\in\mathcal{C}\}, \\ \text{s. t.} \quad & \sum_{a\in\mathcal{A}} z(s',a) - \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} z(s,a)\, p(s,a,s') = 0, \forall\, s'\in\mathcal{S} \\ & \sum_{a\in\mathcal{A}} y(s',a) - \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} y(s,a)\, p(s,a,s') = \alpha_{s_0} - \sum_{a\in\mathcal{A}} z(s,a), \ \forall\, s'\in\mathcal{S}. \end{aligned} \tag{10}$$

We construct a policy from any feasible solution to the linear program in (10) for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$ as:

$$\pi(s) = \begin{cases} \frac{z(s,a)}{\sum_{a'\in\mathcal{A}} z(s,a)} & \text{if } \sum_{a'\in\mathcal{A}} z(s,a') > 0, \\ \frac{y(s,a)}{\sum_{a'\in\mathcal{A}} y(s,a')} & \text{otherwise.} \end{cases} \tag{11}$$

It should be mentioned that for each $s \in \mathcal{S}$, $z(s,a) > 0$ for exactly one $a \in \mathcal{A}$, and for each $s \notin \mathcal{S}$, $y(s,a) > 0$ for exactly one $a \in \mathcal{A}$, there exists a deterministic policy (Puterman, 2005).

4. Comparison with Existing Methods

In this section, we compare the proposed Average Reward (AVR) approach with three related studies: a robust safety framework based on HJI reachability (Fisac et al., 2018), the Minimum Discounted Reward (MDR) formulation for reachable sets (Akametalu et al., 2023), and the reachability of Markov chains using long-run average rewards (Ávila and Junca, 2021).

HJI-based approaches, formulated in the continuous domain, compute a single, robustly safe set for a deterministic system under worst-case disturbances. These methods numerically solve the HJI PDE using discrete-time Bellman equations, which typically require a specific initialization to

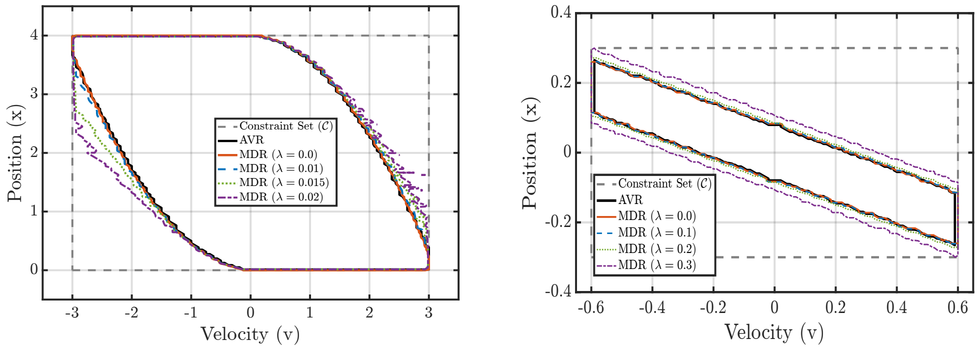

Figure 1: Safe sets computed by AVR and MDR for the Double Integrator (right) and Inverted Pendulum (left). The dashed gray line outlines the constraint set $\mathcal{C}$, the solid black line shows AVR's safe set, and the colored lines show MDR's safe set ($Z(\boldsymbol{x}) = 0$) for different λ.

converge. The MDR formulation (Akametalu et al., 2023), also continuous, addresses this by adding a discount factor, transforming the problem into a contraction mapping that ensures convergence from any initialization. In the MDR method, this discount factor directly influences the size of the final safe set, often represented by the zero-level set of a signed distance function.

In contrast, the AVR method is formulated directly as a discrete MDP with known stochastic transition probabilities. The AVR method uses the average reward criterion to solve the infinite-horizon problem and is formulated without a discount factor. For computation, the AVR method solves the problem using a standard linear program (Equations (9) and (10)). Finally, the AVR computation yields the optimal gain function, which provides a probabilistic safety function (Theorem 3) where different level sets correspond to different safety confidence levels.

Although our proposed and reachability for Markov chains methods both use the average reward criterion, the AVR method is formulated to solve the indefinite probabilistic safety problem. This objective differs from the standard reach-avoid task. While the reachability framework is applied to find the probability of reaching a set before avoiding, we aim to find the probability of remaining in the safe constraint set indefinitely (that is, the probability of never reaching the unsafe set). This difference in objective is reflected in the construction of the transition function. To solve the reach-avoid problem, the Avila and Junca method constructs a modified transition kernel where both the target set and the avoid set are made absorbing. In contrast, our method constructs a transition function (Equation (4)) where only the unsafe set is made absorbing. This specific construction of the transition model enables the optimal gain function to represent the probability of indefinite safety, allowing us to formally prove in Theorem 3 that the optimal gain function is directly equal to this maximum probability.

5. Numerical Validation

We evaluate our approach numerically on two benchmark control systems: a double integrator and an inverted pendulum. The simulations were implemented in Julia, using the JuMP modeling language and the MOSEK optimization solver. Our AVR method is solved as a linear program,

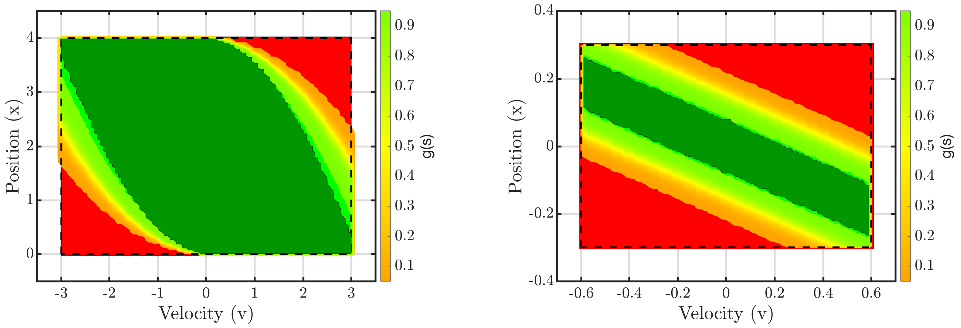

Figure 2: Safe sets computed using AVR as level sets of $g(s)$ for (a) Double Integrator and (b) Inverted Pendulum systems. The set $\mathcal{K}$ is dark green.

whereas the MDR method is implemented for comparison and solved using value iteration (VI). All computations were executed on a computer equipped with a 64-core AMD Ryzen Threadripper 3970X CPU and 256 GB of RAM.

The continuous state and action spaces of the two benchmark control systems are discretized into uniform grids. The stochastic transition probabilities are constructed using a Monte Carlo approach: for each discrete state-action pair, 100 simulations are run with disturbances from the known distribution and bounds. Each resulting next state in the continuous state space is mapped to the closest discretized state on the grid, identified using a k-nearest-neighbors search (KDTree and knn)[2]. The transition probabilities are thus computed from the frequency of these assignments.

The discrete-time Double Integrator dynamics is considered as

$$\mathrm{x}_{k+1} = \mathrm{x}_k + \mathrm{v}_k\,\Delta t, \quad \mathrm{v}_{k+1} = \mathrm{v}_k + \left(u_k + \tilde{d}_k\right)\Delta t.$$

x is position and v is velocity. The state space is bounded by $\mathcal{S} = \{(\mathrm{x},\mathrm{v}) | -1 \leq \mathrm{x} \leq 5, -5 \leq \mathrm{v} \leq 5\}$. The disturbance $\tilde{d}_k$ is drawn from a Normal distribution $\mathcal{N}(0,1)$, clamped to $[-1,1]$. The control input u_k is in $[-2,2]$, and the constraint set is $\mathcal{C} = \{(\mathrm{x},\mathrm{v}) \in \mathbb{R}^2 | 0 \leq \mathrm{x} \leq 4, -3 \leq \mathrm{v} \leq 3\}$. The time step is $\Delta t = 0.1$. For this specific system, we use a 161 × 161 state grid and 81 actions.

The discrete-time dynamics of a stochastic, nonlinear Inverted Pendulum are:

$$\theta_{k+1} = \theta_k + \omega_k \Delta t, \qquad \omega_{k+1} = \omega_k + (\frac{g_p}{l_p} sin(\theta_k) + \frac{1}{m_p l_p^2} u_k + \tilde{d}_k)\Delta t$$

Here, the state consists of the angle θ_k and the angular velocity ω_k, bounded by $\mathcal{S} = \{(\theta,\omega)| -0.5 \leq \theta \leq 0.5, -1.0 \leq \omega \leq 1.0\}$. The disturbance $\tilde{d}_k$ is drawn from a Normal distribution $\mathcal{N}(0,1)$, clamped to $[-0.75, 0.75]$. The control input u_k is in $[-3,3]$, and the constraint set is $\mathcal{C} = \{(\theta,\omega) \in \mathbb{R}^2 | -0.3 \leq \theta \leq 0.3, -0.6 \leq \omega \leq 0.6\}$. The time step is $\Delta t = 0.1$. For this system, we use a 201 × 201 state grid and 81 discrete actions.

Figure 1 illustrates the computed safe sets for the Double Integrator (left) and Inverted Pendulum (right). The figure shows the safe set boundary from our AVR method (solid black line) alongside

[2] Here, KNN serves as a non-parametric function approximator, where the value of a state is estimated from the values of its nearest neighbors on the discretized grid.

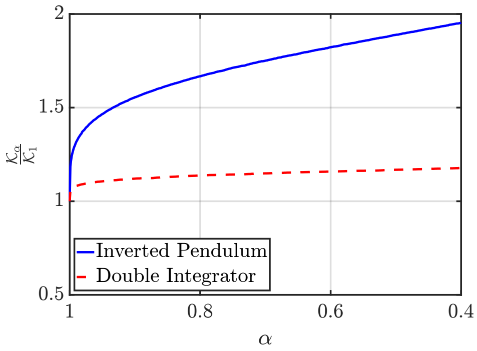

Figure 3: The relative size of the probabilistically-safe set $\mathcal{K}_\alpha$ as a function of the safety confidence level α. The y-axis plots the ratio of the size of $\mathcal{K}_\alpha$ to the 100% safe set, $\mathcal{K}$. Results are shown for the Double Integrator (dashed red) and Inverted Pendulum (solid blue).

the boundaries from the MDR method using various discount rates (λ). The boundaries from the MDR method are observed to change as λ is varied. The AVR result is aligned with the undiscounted ($\lambda = 0$) MDR case, which is considered the analytical solution (Akametalu et al., 2023).

Figure 2 presents the numerical results for the proposed AVR method, visualizing the optimal gain function, $g(s)$, as a safety distribution for Double Integrator (right) and Inverted Pendulum (left). In each subplot, the dashed black lines are the constraint set, $\mathcal{C}$, within which the color map indicates the level of probabilistic safety. The dark green area represents the safe set, $\mathcal{K}$, where the optimal gain $g^\star(s) = 1$. Conversely, the red region corresponds to the zero-level set where $g^\star(s) = 0$, representing unsafe states. The gradient from orange to light green illustrates the transient states between zero-level set and one-level set of $g^\star(s)$, showing states with varying probabilities of remaining safe as a function of the gain value.

Figure 3 plots the relative size of the probabilistically-safe set $\mathcal{K}_\alpha$ as a function of the safety confidence level α. The y-axis shows the ratio of the size of the α-safe set to the size of $\mathcal{K}$. Both systems start at a ratio of 1 when $\alpha = 1.0$. For the Double Integrator (dashed red line), the curve remains relatively flat, increasing only slightly as α decreases. In contrast, the Inverted Pendulum (solid blue line) shows a significant and steep increase, with the safe set's relative size approaching 2.0 as α moves toward 0.4.

Figure 4 compares the computational time for the AVR (LP) method and the MDR (VI) method on the Inverted Pendulum (left) and Double Integrator (right) systems. The plots show that the computation time for the AVR method, which is solved as a single linear program, scales much more efficiently with the number of states. In contrast, the time for the value iteration grows substantially faster. The plots also indicate that the value iteration computation time is largely insensitive to the specific discount rate chosen, as the lines for all λ values are close to each other.

6. Conclusion and Future Work

This study introduces a framework for probabilistic safety analysis in stochastic control systems by applying the average-reward MDP properties. A fundamental link was established between the probabilistic safe set $\mathcal{K}_\alpha$ and the optimal gain function $g^\star(s)$, which serves as the probabilistic

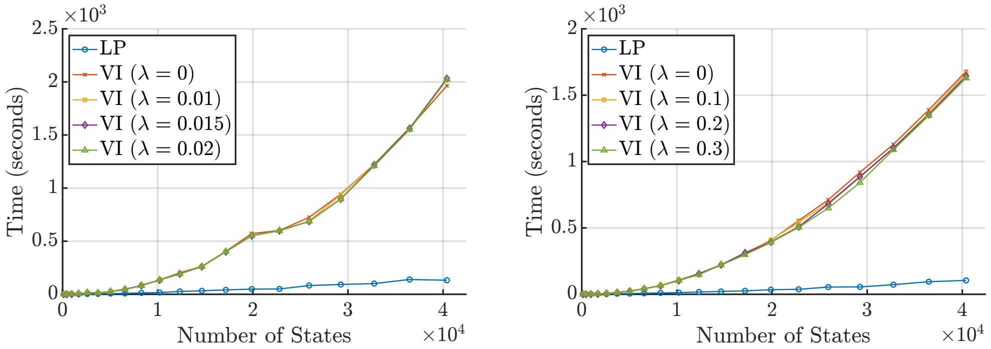

Figure 4: Runtime of the AVR (LP) and the MDR (VI) method for the Inverted Pendulum (left) and the Double Integrator (right) as a function of total number of discrete states.

safety value function. This approach eliminates dependence on the arbitrary discount factor, which is critical in MDR and other discounted methods, a factor known to affect both the convergence rate and the safe set size. By formulating the safety objective as an infinite-horizon average reward problem, the probabilistic safety value function can be computed efficiently by solving a standard linear program, which is significantly faster than value iteration (VI) methods typically used for discounted formulations. Numerical results on the Double Integrator and Inverted Pendulum systems confirm that the AVR method yields accurate and safe set boundaries.

Future research focuses on extending this framework to move beyond the computation of safe policies to selecting and utilizing only those safe policies that also satisfy predefined optimality conditions. Specifically, the objective will be to choose a subset of probabilistically safe policies that maximize performance metrics, thereby coupling safety guarantees with a desired level of system optimality.

References

Murad Abu-Khalaf, Frank L. Lewis, and Jie Huang. Policy iterations on the Hamilton-Jacobi-Isaacs equation for H_∞ state feedback control with input saturation. *IEEE Transactions on Automatic Control*, 51(12):1989–1995, 2006.

Anayo K Akametalu, Shromona Ghosh, Jaime F Fisac, Vicenc Rubies-Royo, and Claire J Tomlin. A minimum discounted reward Hamilton–Jacobi formulation for computing reachable sets. *IEEE Transactions on Automatic Control*, 69(2):1097–1103, 2023.

Anil Alan, Andrew J Taylor, Chaozhe R He, Aaron D Ames, and Gábor Orosz. Control barrier functions and input-to-state safety with application to automated vehicles. *IEEE Transactions on Control Systems Technology*, 31(6):2744–2759, 2023.

Daniel Arnström and André MH Teixeira. Data-driven and stealthy deactivation of safety filters. *arXiv preprint arXiv:2412.01346*, 2024.

Daniel Ávila and Mauricio Junca. On reachability of Markov chains: A long-run average approach. *IEEE Transactions on Automatic Control*, 67(4):1996–2003, 2021.

Azra Begzadić, Nikhil Uday Shinde, Sander Tonkens, Dylan Hirsch, Kaleb Ugalde, Michael C Yip, Jorge Cortés, and Sylvia Herbert. Back to base: Towards hands-off learning via safe resets with reach-avoid safety filters. *arXiv preprint arXiv:2501.02620*, 2025.

Robert S Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

William D Compton, Max H Cohen, and Aaron D Ames. Learning for layered safety-critical control with predictive control barrier functions. *arXiv preprint arXiv:2412.04658*, 2024.

James Dallas, John Talbot, Makoto Suminaka, Michael Thompson, Thomas Lew, Gabor Orosz, and John Subosits. Control barrier functions for shared control and vehicle safety. *arXiv preprint arXiv:2503.19994*, 2025.

Murad Dawood, Ahmed Shokry, and Maren Bennewitz. A dynamic safety shield for safe and efficient reinforcement learning of navigation tasks. *arXiv preprint arXiv:2412.04153*, 2024.

Lawrence C Evans and Panagiotis E Souganidis. Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations. *Indiana University Mathematics Journal*, 33 (5):773–797, 1984.

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

Yulong Gao, Alessandro Abate, Lihua Xie, and Karl Henrik Johansson. Distributional reachability for Markov decision processes: Theory and applications. *IEEE Transactions on Automatic Control*, 69(7):4598–4613, 2023.

Andrew H Jazwinski. *Stochastic processes and filtering theory*. Courier Corporation, 2007.

Kostas Margellos and John Lygeros. Hamilton-Jacobi formulation for reach–avoid differential games. *IEEE Transactions on automatic control*, 56(8):1849–1861, 2011.

Jun Moon. State and control path-dependent stochastic zero-sum differential games: Viscosity solutions of path-dependent Hamilton–Jacobi–Isaacs equations. *Mathematics*, 10(10):1766, 2022.

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.

Pedram Rabiee and Amirsaeid Safari. Safe exploration in reinforcement learning: training backup control barrier functions with zero training time safety violations. *arXiv preprint arXiv:2312.07828*, 2023.

Ehsan Soleimani, Irfan Ahmad Ganie, and S Jagannathan. Safe Optimal Control of Quadrotor Formations Using Multilayer Neural Networks and Continual Learning. *International Journal of Adaptive Control and Signal Processing*, 2025.

H Mete Soner and Nizar Touzi. Stochastic target problems, dynamic programming, and viscosity solutions. *SIAM Journal on Control and Optimization*, 41(2):404–424, 2002.

Kim Peter Wabersich and Melanie N Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*, 129:109597, 2021.

Yichao Zhong, Chong Zhang, Tairan He, and Guanya Shi. Bridging adaptivity and safety: Learning agile collision-free locomotion across varied physics. *arXiv preprint arXiv:2501.04276*, 2025.

Appendix A. Lemmas for Theorem 3

Lemma 4 *Each $s \in \mathcal{C} \setminus \mathcal{K}$ is transient:*

$$\lim_{k\to\infty} \mathbb{P}_s^{\pi}[\tilde{s}_k = s] = 0, \quad \forall \pi \in \Pi_{\mathrm{SD}}.$$

Proof Suppose, for the sake of deriving a contradiction, that $s \in \mathcal{C} \setminus \mathcal{K}$ is considered as *recurrent* if and only if the following summation of time step transition probability is:

$$\sum_{k=0}^{\infty} p^k(s|s) = \infty \quad \forall s \in \mathcal{C} \setminus \mathcal{K}$$

where

$$\sum_{k=0}^{\infty} p^k(s|s) = 1 + p(s|s) + p^2(s|s) + p^3(s|s) + \cdots + p^{\infty}(s|s).$$

Then, if and only if the limiting probability equals $p(s|s)^{\infty} = \lim_{k\to\infty} \mathbb{P}_s^{\pi}[\tilde{s}_k = s] = 1, \forall \pi \in \Pi_{\mathrm{SD}}$. Whereas this contradicts the fact that the state is transient if and only if

$$\sum_{k=0}^{\infty} p^k(s|s) < \infty \quad \forall s \in \mathcal{C} \setminus \mathcal{K}.$$

Hence the sum must be finite and $p(s|s)^{\infty} = \lim_{k\to\infty} \mathbb{P}_s^{\hat{\pi}}[\tilde{s}_k = s] = 0 \quad \forall s \in \mathcal{C} \setminus \mathcal{K}$. Therefore, $s \in \mathcal{C} \setminus \mathcal{K}$ is *transient*. ■

Lemma 5 *For all $s \in \mathcal{S} \setminus \mathcal{C}$,*

$$g^{\star}(s) = 0.$$

Proof This result follows directly from the structure of the transition probabilities defined in (4). States outside the constraint set $\mathcal{C}$ transition only to themselves, which are classified as unsafe and do not yield any reward. ■

Lemma 6 *For all $s \in \mathcal{S}$, we have*

$$s \in \mathcal{K} \iff g^{\star}(s) = 1.$$

Proof First, we show $s \in \mathcal{K} \implies g^\star(s) = 1$, there exists a deterministic stationary policy π such that

$$\mathbb{P}_s^\pi \left[\tilde{x}_k \in \mathcal{C}\right] = 1 \quad \text{a.s. for every } k \in \mathbb{N}.$$

Under this policy $\mathbb{1}\{\tilde{x}_k \in \mathcal{C}\} = 1$ for all k. Hence:

$$g^\pi(s) = \limsup_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} 1 = 1.$$

Since $r(s, a, s') \leq 1, \forall s, s' \in \mathcal{S}, a \in \mathcal{A}, g^\pi \leq g^\star(s) \leq 1$.

Second, we start $s \notin \mathcal{K} \implies g^\star(s) < 1$ (since $g^\star(s) \leq 1$). Since $s \notin \mathcal{K}$, we have that $\forall \pi \in \Pi_{\text{SD}}, \exists \zeta > 0$, such that

$$\mathbb{P}_s^\pi \left[\tilde{x}_\zeta \notin \mathcal{C}\right] > 0, \qquad \mathbb{P}_s^\pi \left[\tilde{x}_k \notin \mathcal{C}\right] = 0, \quad \forall k < \zeta.$$

By the construction of transition probabilities in (4):

$$\mathbb{P}_s^\pi \left[\tilde{s}_\zeta \notin \mathcal{C}\right] = \mathbb{P}_s^\pi \left[\tilde{x}_\zeta \notin \mathcal{C}\right] > 0, \qquad \mathbb{P}_s^\pi \left[\tilde{s}_k \notin \mathcal{C}\right] = 0, \quad \forall k < \zeta.$$

According to multichain optimality equation (Puterman, 2005):

$$\max_{a \in \mathcal{A}} \left\{ \sum_{s' \in \mathcal{S}} p(s, a, s') g^\pi(s') - g^\pi(s) \right\} = 0,$$

for a fixed policy and specific probability set up, the gain is:

$$g^\pi(s) = \sum_{s' \in \mathcal{S}} \mathbb{P}_s^\pi \left[\tilde{s}_\zeta = s'\right] \cdot g^\pi(s'), \quad \forall \pi \in \Pi_{\text{SD}}.$$

We know from Theorem 5 that for each $s' \in \mathcal{S} \setminus \mathcal{C}$ $g^\pi(s') = 0$, $\forall \pi \in \Pi_{\text{SD}}$. Then:

$$g^\pi(s) = \sum_{s' \in \mathcal{S}} \mathbb{P}_s^\pi \left[\tilde{s}_\zeta = s'\right] \cdot g^\pi(s') \leq (1 - \mathbb{P}_s^\pi \left[\tilde{s}_\zeta \in \mathcal{S} \setminus \mathcal{C}\right]) < 1, \quad \forall \pi \in \Pi_{\text{SD}}.$$

Therefore, from the existence of an optimal policy $g^\star(s) \leq 1$. Since both directions hold, we conclude $s \in \mathcal{K} \iff g^\star(s) = 1$. ■

With reference to Theorem 5, the following theorem shows that any state with a gain equal to 1 is considered a safe state, and the union of these states forms the safe set. Intuitively, the theorem implies that safety is ensured if there exists a policy such that the expected average reward, defined by the indicator function of the constraint set, remains 1 at all time steps.