Probabilistic Safety Guarantees for Markov Decision Processes

Rafal Wisniewski, *Senior Member, IEEE*, and Manuela L. Bujorianu

***Abstract*—This article aims to incorporate safety specifications into Markov decision processes. Explicitly, we address the minimization problem up to a stopping time with safety constraints. We establish a formalism leaning upon the evolution equation to achieve our goal. We show how to compute the safety function with dynamic programming. In the last part of this article, we develop several algorithms for safe stochastic optimization using linear and dynamic programming.**

***Index Terms*—Dynamic programming (DP), linear programming (LP), Markov decision processes (MDPs), safety.**

I. Introduction

The point of departure is a Markov decision process (MDP) with a finite number of states and actions. The overall objective of this article is twofold: 1) to formulate stochastic safety as dynamic programming (DP) and 2) to incorporate probabilistic safety guarantees into the stochastic optimization of MDPs. To the best of our knowledge, a method of directly unifying MDPs and safety is missing in the literature. Undoubtedly, several complementary approaches tackle safety in optimization. They will be described in the related work paragraph.

a) Motivation: Recently, the subject of DP has enjoyed a resurgence [1], [2]. The explanation for this increase in popularity is reinforcement learning (RL)—a powerful and prevalent method for learning from data and subsequently generating optimal decisions [3]. In a nutshell, DP provides the mathematical structure for RL. Applications of DP can be found in robotics [4], autonomous vehicles [5], drones [6], and water networks [7], to name a few examples. On the other hand, [8] showed that for optimization problems, where the constraints are formulated as cost functions, the principle of optimality does not hold (for multichain MDPs), and the value function depends on the initial distribution. Consequently, the solution of such optimization problems cannot be solved by DP. On the contrary, linear programming (LP) provides the means of solving constrained MDP problems [9] and [10]. Specifically, in this work, we strive to combine the results on constrained MDPs with safety [11]. Safety assigns the probability of reaching the undesired states—the forbidden set. The intended result is an optimization algorithm that keeps the system on the desired safety level. Specifically, the probability that the process realizations hit the forbidden states before reaching the target set remains below a certain value p. This is the concept of p-safety introduced in [12]. This definition of safety is related to the reach-avoidance problem thoroughly studied in formal verification, where the problem of safety can be formulated as a temporal logic specification [13].

b) Novelty: The approach of this article fruitfully combines ideas from constrained MDP with the concept of p-safety. The p-safety represents a rigorous mathematical formalism that encapsulates most of the probabilistic safety formulations in the literature: standard probabilistic reachability problem, reach-avoidance problem, and bound probabilistic reachability. The analytical reasoning in the current work leans upon elements of probabilistic potential theory. Already in [14] and [15], it has been shown that the analytical approach based on potential theory provides straightforward proofs for the barrier certificate's properties.

The list of original contributions of this work includes the following.

1) p-safety is reformulated as a DP problem.
2) The evolution equation, relating the initial, the hitting and the occupation measures, is introduced into DP.
3) The safe MDP is formulated as optimization with constraints. The resulting formulation involves two occupation measures for safety and optimality, which are subsequently combined in LP.

c) Related work: The subject of minimizing an expected cost without safety constraints is not new, and it is well-known that its solution is obtained by solving Bellman's equation [2] or LP [16]. The safety verification problem of stochastic systems has also been addressed in the literature [17]. This article [11] has extended the approach based on barrier certificates to discrete settings of Markov chains (MCs). Pragmatic methods for safe DP and RL have been addressed in [18]. In the abovementioned reference, safety is ensured with a barrier function, which serves as a soft constraint to the system. On the other hand, the work [19] proposes a supervisor that prevents the applied control action from driving the system into unsafe regions.

Several approaches have been grafted on different model predictive control (MPC) techniques in the context of safety learning. The shielding approach [20], [21] welds a backup policy that is proven safe and subsequently uses the backup policy to revoke the learned policy to guarantee safety. Another approach is verifying safety on the fly using MPC safety certification [22]. A similar research line is adaptive RL [23], [24], where safety is computed for the next k steps and unsafe actions are blocked. There is an intrinsic tradeoff when choosing the number k of steps. If it is too small, an MDP might end up in a state where all actions are unsafe even though a safe policy exists. If k is too long, the complexity of the shielding algorithm for blocking unsafe actions is too large.

d) Approach in this work: We take the starting point of an MDP with a (stationary) policy that, for each state, provides the probability of choosing a particular control action. Nonetheless, we face a challenge. To compute the optimal path to the target states, we need a random time when the process reaches the target set before hitting the forbidden set. Our solution to this challenge is to use the evolution equation [25], which relates the occupation measure with the hitting probability. The occupation measure corresponds to the expected number of the states' visits. When examining the hitting probability, we consider two sets: the set of target states and the set of forbidden states. Consequently, a safety function S_π is derived from the evolution

Manuscript received 2 October 2022; revised 19 May 2023; accepted 14 June 2023. Date of publication 4 July 2023; date of current version 5 December 2023. The first author's work has been supported by the Poul Due Jensens Foundation under project SWIft. Recommended by Associate Editor U. Topcu. *(Corresponding author: Rafal Wisniewski.)*

Rafal Wisniewski is with the Department of Electronic Systems, Section of Automation and Control, Aalborg University, 9220 Aalborg East, Denmark (e-mail: raf@es.aau.dk).

Manuela L. Bujorianu is with the Department of Computer Science, University College London, WC1E 6EA London, U.K. (e-mail: L.Bujorianu@ucl.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2023.3291952.

Digital Object Identifier 10.1109/TAC.2023.3291952

equation. The safety function provides the probability of hitting the forbidden set. It is shown that the safety function is the solution of Bellman's equation and can be computed by an iterative procedure analogous to the one used for computing the value function in DP. In the second part of this article, we combine stochastic optimization with safety guarantees. Consequently, we formulate the optimization with a constraint: minimizing the value function V_π subject to keeping system p-safe, $S_\pi \leq p$. Here, the cost and the safety are the expected values of accumulated rewards up to stopping times. To this end, we reformulate LP in [10] to address the time horizon specified by the two stopping times.

In the final section, we relax the concept of safety. Subsequently, we develop a local optimization algorithm, meaning that the control action at each state i is computed only using the information available from its neighbors.[1] We introduce a local concept of safety—relative safety. Equipped with this new concept, we define an optimization problem.

e) Organization of this article: We shed light on the preliminary objects of this work: MCs and MDPs in Sections II and III. Specifically, the focus in these sections is on formulating the evolution equation for the occupation measure and hitting measures. The stochastic optimization with stopping time is the matter of Section IV. It is shown in Section V that the safety function can be computed as the accumulated cost of the probability of getting to the unsafe set. Hence, the safety function is the solution to Bellman's equation. The main result—an algorithm for safe MDP is developed in Sections VI.

A. Notation

For a countable set U and a set R, we write $R^U := R^{|U|}$ for the Cartesian product of $|U|$ copies of R. We use I_U to denote the identity matrix on U, and $\mathbb{1}_U$ to denote the vector of ones on U. For two vectors $v, w \in \mathbb{R}^m$, we will use the Hadamard product $v \circ w$ defined by $(v \circ w)(i) = v(i)w(i)$. The notation $v \geq 0$ denotes $v(i) \geq 0$ for all $i \in \{1, \ldots, m\}$. The Kronecker delta denoted by δ_j is $\delta_j(i) = 1$ for $i = j$, otherwise it is 0.

II. MCs and Evolution Equation

Let $\mathcal{X}$ be a countable set of states denoted by letters $i, j, \ldots$ A probability distribution ν on $\mathcal{X}$, $(\nu(j))_{j\in\mathcal{X}}$, is thought of as a row vector $\nu \in \mathbb{R}^{\mathcal{X}}_{\geq 0}$ (with $\sum_{j\in\mathcal{X}} \nu_j = 1$). A function $f : \mathcal{X} \to \mathbb{R}$ is defined as a column vector $f = (f(j))^\top_{j\in\mathcal{X}}$.

Suppose that $(X_t) := (X_t)_{t\in\mathbb{N}}$ is a discrete-time (homogeneous) MC with transition probabilities

$$p_{ij} := \mathbb{P}[X_{t+1} = j | X_t = i] = \mathbb{P}[X_1 = j | X_0 = i]. \tag{1}$$

The transition matrix P of (X_t) is $P := (p_{ij})_{i,j\in\mathcal{X}}$. The k-step transition probabilities are $\mathbb{P}[X_k = j | X_0 = i] = (P^k)_{ij}$, where $P^k = PP\ldots P$ is the k-fold matrix product.

Let H be an arbitrary subset of $\mathcal{X}$, which will be kept fixed. We will call H the taboo set. Later in this article, the taboo set will be the complement of the union of the sets of all the goal states and the forbidden states. We restrict the transition probabilities of the MC (X_t) to the set H. These are the taboo transition probabilities [26]. We collect the taboo transition probabilities into the transition matrix $Q = (p_{ij})_{i,j\in H}$. In this case, the transition matrix Q is substochastic, i.e., the sum of row entries $\sum_{j\in H} p_{ij} \leq 1$.

We introduce the occupation (green) operator of H

$$G := \sum_{k=0}^{\infty} Q^k. \tag{2}$$

G is well defined if the states in H are transient, i.e., for all $i \in H$, we have $P[X_k = i \text{ for infinitely many } k | X_0 = i] = 0$.

From (2), it follows that:

$$G = I_H + QG = I_H + GQ \tag{3}$$

i.e., $G = (I_H - Q)^{-1}$. Recall I_H is the identity matrix on H.

A. Evolution of the MC

To study the reach-avoidance problem (reach the target set while avoiding the forbidden set), we examine the process up to the first hitting time of a target set or a forbidden set. We associate a reward to each state and ask two questions: What is the cost of getting to the target set, and what is the probability that the process reaches the forbidden set before the target set? To this end, we will use the evolution equation relating the occupation measure and the hitting probability, which we characterize first.

Suppose that τ is a stopping time, for instance, $\tau = \tau_E$ is the first hitting time of some set E, i.e., $\tau_E := \min\{t \geq 0 | \ X_t \in E\}$. The remaining part of this article assumes that the stopping time τ is finite almost surely (a.s.). Specifically, if the states in $\mathcal{X} \setminus E$ are transient, $\tau_E < \infty$ a.s.

Suppose that D is a subset of $\mathcal{X}$. Let $\rho_{<\tau}(D)$ be a random variable that describes the amount of time the MC spends in D before time τ has passed

$$\rho_{<\tau}(D) := \sum_{t=0}^{\tau-1} I_{\{X_t\in D\}}. \tag{4}$$

The (state) occupation measure $\gamma_{<\tau}$ for (X_n) is defined as the expectation of $\rho_{<\tau}(\cdot)$ in (4), i.e., $\gamma_{<\tau}(D) := \mathbb{E}\rho_{<\tau}(D)$

$$\begin{aligned}\gamma_{<\tau}(D) &= \mathbb{E}\sum_{t=0}^{\tau-1} I_{\{X_t\in D\}} = \mathbb{E}\sum_{t=0}^{\infty} I_{\{X_t\in D\}} I_{\{t<\tau\}} \\ &= \mathbb{E}\sum_{t=0}^{\infty} I_{\{X_t\in D, t<\tau\}} = \sum_{t=0}^{\infty} \mathbb{P}[t < \tau, X_t \in D]. \end{aligned} \tag{5}$$

As its name suggests, $\gamma_{<\tau}$ is a measure.

We define the integral w.r.t. $\gamma_{<\tau}$ of a vector function f as

$$\langle \gamma_{<\tau}, f \rangle := \mathbb{E}\sum_{t=0}^{\tau-1} f(X_t). \tag{6}$$

The abovementioned equation will be instrumental for computing the accumulated cost of the process until stopping time τ.

The (state) hitting measure $\lambda_\tau(D)$ is the expected time that the process lies in a set $D \subset \mathcal{X}$ at the time τ

$$\lambda_\tau(D) := \mathbb{P}[X_\tau \in D] = \sum_{t=0}^{\infty} \mathbb{P}[\tau = t, X_t \in D]. \tag{7}$$

We define the hitting operator corresponding to the stopping time τ as the integral of a function f with respect to λ_τ as

$$\langle \lambda_\tau, f \rangle = \mathbb{E}(f(X_\tau)). \tag{8}$$

Specifically, let E and U be disjoint subsets of $\mathcal{X}$. We think about E as a target set and U as a forbidden set. Suppose that $\tau = \tau_{U\cup E}$, the first hitting time of the union of U and E. Then $\langle \lambda_{\tau_{U\cup E}}, I_U \rangle$ is the

[1] By a neighbors of i, we understand a state with nonzero transition probability from the state i.

probability that the process hits the forbidden set before the target set. This relation will be instrumental for the computation of safety.

When the initial state is i, we employ the probability $\mathbb{P}^i$, and use the notations $\gamma^i_{<\tau}$ and λ^i_τ. Similarly, for the initial probability μ, we use $\mathbb{P}^\mu$, and the notations $\gamma^\mu_{<\tau}$ and λ^μ_τ.

The occupation measure and the hitting probability are connected by the adjoint or evolution equation [25]

$$\lambda^\mu_\tau = \mu + \gamma^\mu_{<\tau}\mathcal{L} \tag{9}$$

where $\mathcal{L}$ is the generator, $\mathcal{L} = P - I_{\mathcal{X}}$. The measure triplet $(\mu, \gamma^\mu_{<\tau}, \lambda^\mu_\tau)$ with $\gamma^\mu_{<\tau}(i) = 0$ for $i \in \mathcal{X} \setminus H$ and $\lambda^\mu_\tau(j) = 0$ for $j \in H$ uniquely characterizes the given Markov process [27].

III. MDPs and Evolution Equation

We suppose that (U_t) is a process with values in a countable set $\mathcal{U}$ of actions, and study the conditional probabilities $\mathbb{P}[X_{t+1} = j|\ X_t = i,\ U_t = u]$ for $i, j \in \mathcal{X}$, and $u \in \mathcal{U}$. We remark that Markov property holds for the MDPs, and introduce transition probabilities

$$p_{iuj} = \mathbb{P}[X_{t+1} = j | X_t = i,\ U_t = u]$$

where $(i, u, j) \in \mathcal{X} \times \mathcal{U} \times \mathcal{X}$.

By a Markov policy, we understand the family of stochastic kernels $(\pi_{iu}(t))_{(i,u)\in\mathcal{X}\times\mathcal{U}}$

$$\pi_{iu}(t) = \mathbb{P}[U_t = u | X_t = i].$$

We think about the policy π as the to-be-designed stochastic control. In this work, we entirely restrict our attention to stationary policies; the Markov policy is stationary if π_{iu} does not depend on the time, i.e.,

$$\pi_{iu} = \mathbb{P}[U_t = u | X_t = i] = \mathbb{P}[U_0 = u | X_0 = i].$$

To conclude, the stationary policy π is seen as the (possibly infinite-dimensional) matrix

$$\pi := (\pi_{iu})_{(i,u)\in\mathcal{X}\times\mathcal{U}} \tag{10}$$

with entries between 0 and 1, corresponding to the probability that at the state i, the control action has the value u.

Let $\mathcal{D}$ be the standard simplex in $\mathbb{R}^{\mathcal{U}}$

$$\mathcal{D} := \left\{ \alpha = (\alpha_u)_{u\in\mathcal{U}} | \alpha_u \geq 0,\ \sum_{u\in\mathcal{U}} \alpha_u = 1 \right\}. \tag{11}$$

So, for each fixed i, the probabilities $(\pi_{iu})_{u\in\mathcal{U}}$ belong to $\mathcal{D}$. We will write $\pi \in \mathcal{D}^{\mathcal{X}}$ (the Cartesian product of $|\mathcal{X}|$ copies of the set $\mathcal{D}$) even though we mean its matrix representation $(\pi_{iu})_{(i,u)\in\mathcal{X}\times\mathcal{U}}$ with $\sum_{u\in\mathcal{U}} \pi_{iu} = 1$, for each $i \in \mathcal{X}$.

For a stationary policy π, using the law of total probabilities the transition probability of the induced chain are

$$p_{ij}(\pi) = \sum_{u\in\mathcal{U}} \pi_{iu} p_{iuj}. \tag{12}$$

For the policy π, we define the transition probability matrix

$$P(\pi) = (p_{ij}(\pi))_{(i,j)\in\mathcal{X}\times\mathcal{X}}.$$

A straightforward calculation shows that the operator $P(\pi)$ on vector functions f has the following expression:

$$P(\pi)f = \sum_{u\in\mathcal{U}} \pi^u \circ (P(u)f)$$

where $\circ$ is the Hadamard product, $P(u) := (p_{iuj})_{i,j\in\mathcal{X}}$ and $\pi^u := (\pi_{iu})_{i\in\mathcal{X}}$ is thought of as a row measure associated to each action $u \in \mathcal{U}$. This "factorization" will be a key tool for obtaining most expressions in the following sections.

A. State-Action Occupation Measure

For a target set E. Let the taboo set $H = \mathcal{X} - E$. At the outset, we recall the notion of a reward

$$\rho : H \times \mathcal{U} \to \mathbb{R}.$$

The function ρ induces a process (ρ_t) by

$$\rho_t = \rho(X_t, U_t). \tag{13}$$

Let $R_\pi := (R_\pi(i))_{i\in H}$ with the components $R_\pi(i) = \sum_{u\in\mathcal{U}} \pi_{iu}\rho(i,u)$.

We suppose that the process (U_t) is generated by a policy π, which will be characterized in the following. Let τ be a stopping time, for example the first hitting time of a set. The cost for the policy π up to time τ is

$$V_\pi(i) := \mathbb{E}_\pi\left[\sum_{t=0}^{\tau-1} \rho_t | X_0 = i\right] \tag{14}$$

and the vector $V_\pi := (V_\pi(i))_{i\in H}$.

The aim of DP is to evaluate the cost function V_π, and subsequently to find a minimizing stationary policy. To meet this aim, this article will use the evolution equation defined in the last section. All the objects used below will depend on the policy π; herein, the expectation, the transition matrix, occupation measure, and hitting probability. Therefore, to enhance readability, we occasionally suppress π from the notation. At the outset, notice that in (14), since τ is a random variable, the expectation operator cannot be moved under the summation symbol, as it is customarily done in standard DP and RL (see [2] and [3]). In the realm of altering policies, we enhance the evolution equation to capture the frequencies of visiting the states and actions. We examine the process (X_t, U_t) with initial distribution $\overline{\mu}$ of (X_0, U_0). Moreover, because of policy stationarity, the initial distribution of U_0 has no effect on X_t nor U_t for $t > 0$. Let $\mu(\cdot) = \sum_{u\in\mathcal{U}} \overline{\mu}(\cdot, u)$. To this end, we define a state-action occupation measure by

$$\overline{\gamma}^\mu_{<\tau}(i,u) := \sum_{t=0}^{\infty} \mathbb{P}^\mu[X_t = i, U_t = u, t < \tau]$$

and a state-action hitting measure by

$$\overline{\lambda}^\mu_\tau(i,u) := \mathbb{P}^\mu[X_\tau = i, U_\tau = u]$$

for $(i,u) \in \mathcal{X} \times \mathcal{U}$. Then, the expectation of a function $f : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is

$$\mathbb{E}^\mu\left[\sum_{t=0}^{\tau-1} f(X_t, U_t)\right] = \sum_{u\in\mathcal{U}} \langle f(\cdot, u), \overline{\gamma}^\mu_{<\tau}(\cdot, u)\rangle. \tag{15}$$

For a given $\pi = (\pi_{iu})_{(i,u)\in\mathcal{X}\times\mathcal{U}}$, the state-action and state occupation measures are linked as follows:

$$\gamma^\mu_{<\tau}(i) = \overline{\gamma}^\mu(i, \mathcal{U}) = \sum_{u\in\mathcal{U}} \overline{\gamma}^\mu_{<\tau}(i,u), \text{ and} \tag{16}$$

$$\pi_{i,u}\gamma^\mu_{<\tau}(i) = \overline{\gamma}^\mu_{<\tau}(i,u). \tag{17}$$

The (17) follows from:

$$\begin{aligned}\pi_{i,u}\gamma^\mu_{<\tau}(i) &= \mathbb{P}[U_0 = u | X_0 = i] \sum_{t=0}^{\infty} \mathbb{P}[X_t = i, t < \tau] \\ &= \sum_{t=0}^{\infty} \mathbb{P}[U_t = u | X_t = i, t < \tau]\mathbb{P}[X_t = i, t < \tau] \\ &= \sum_{t=0}^{\infty} \mathbb{P}[X_t = i, U_t = u, t < \tau] = \overline{\gamma}^\mu_{<\tau}(i,u)\end{aligned}$$

where we have used the information that the policy is stationary, $\mathbb{P}[U_t = u|X_t = i] = \mathbb{P}[U_0 = u|X_0 = i]$.

Similarly, the state-action hitting measure and the state hitting measure are related by (16) and (17) with $\gamma_{<\tau}$ and $\overline{\gamma}_{<\tau}$ substituted by λ_τ and $\overline{\lambda}_\tau$.

We return to the evolution (9).

Lemma 1: The evolution equation for the state-action measures is

$$\sum_{u\in\mathcal{U}} \overline{\lambda}^\mu_\tau(\cdot, u) = \mu(\cdot) + \sum_{u\in\mathcal{U}} \overline{\gamma}^\mu_{<\tau}(\cdot, u)\mathcal{L}(u). \tag{18}$$

Proof: The state occupation measure satisfies (9). Left-hand side of (18), follows from (16); whereas, the right-hand side is the consequence of the following computation:

$$\begin{aligned}
[\gamma^\mu_{<\tau}|_H(Q(\pi) - I)](j) &= \sum_{i\in H} \gamma^\mu_{<\tau} p_{ij}(\pi) - \sum_{i\in H} \gamma^\mu_{<\tau}\delta_j(i) \\
&= \sum_{i\in H} \gamma^\mu_{<\tau}(i) \sum_{u\in\mathcal{U}} \pi_{iu} p_{iuj} - \gamma^\mu_{<\tau}(j) \\
&= \sum_{u\in\mathcal{U}} \sum_{i\in H} \gamma^\mu_{<\tau}(i)\pi_{iu}p_{iuj} - \sum_{u\in\mathcal{U}} \overline{\gamma}^\mu_{<\tau}(j, u) \\
&= \sum_{u\in\mathcal{U}} \sum_{i\in H} \overline{\gamma}^\mu_{<\tau}(j, u)(p_{iuj} - \delta_j(i)) \\
&= \left[\sum_{u\in\mathcal{U}} \overline{\gamma}^\mu_{<\tau}|_H(\cdot, u)(Q(u) - I)\right](j).
\end{aligned}$$

■

The evolution (18) can be reformulated as

$$\sum_{u\in\mathcal{U}} \left[\overline{\lambda}^\mu_\tau(i, u) - \overline{\mu}(i, u) - \overline{\gamma}^\mu_{<\tau}(i, u)\mathcal{L}(u)\right] = 0.$$

Hence, the evolution equation for the state-action measure is an average of the evolution equations over constant actions.

IV. Stochastic Optimization With Stopping Time

We shall call a policy π *transient on* H if the MDP X_t with the policy π is transient, i.e.,

$$\mathbb{P}_\pi\left[X_t \in H \text{ for infinitely many } t \mid X_0 \in H\right] = 0.$$

A. DP With Stopping Time

The following result shows that the cost V_π restricted to H can be computed as a potential of "charge" R_π.

Proposition 1: For an MDP (X_t) with the state-action space $(\mathcal{X}, \mathcal{U})$ and a target set $E \subset \mathcal{X}$, let π be a transient policy on $H := \mathcal{X} \setminus E$. Let $\tau := \tau_E$ be the first hitting time corresponding to E, and let ρ be the reward.

Then, the cost function V_π in (14) restricted to H is

$$V_\pi = G(\pi)R_\pi$$

where $G(\pi)$ is the kernel associated to H and π is given by (2).

We remark that the first assumption of π being transient ensures that the optimization problem is feasible for the policy π. If the probability of staying in the taboo set H was not 0 then the system would never reach the target set E.

Proof: Write $\mathbb{P} := \mathbb{P}_\pi$, and $\mathbb{E} := \mathbb{E}_\pi$. Notice that

$$\begin{aligned}
R_\pi(j) &= \sum_{u\in\mathcal{U}} \pi_{ju}\rho(j, u) = \sum_{u\in\mathcal{U}} \rho(X_t, u)\mathbb{P}[U_t = u|X_t = j] \\
&= \mathbb{E}[\rho(X_t, U_t)|X_t = j] = \mathbb{E}[\rho_t|X_t = j], \forall j \in H.
\end{aligned}$$

We claim that

$$V_\pi(i) = \mathbb{E}\left[\sum_{t=0}^{\tau-1} \rho_t|X_0 = i\right] = \mathbb{E}\left[\sum_{t=0}^{\tau-1} R_\pi(X_t)|X_0 = i\right]. \tag{19}$$

The claim follows from:

$$\begin{aligned}
V_\pi(i) &= \sum_{k=0}^{\infty} \mathbb{E}\left[\sum_{t=0}^{k} \rho_t|X_0 = i\right] \mathbb{P}[\tau = k + 1|X_0 = i] \\
&= \sum_{k=0}^{\infty} \sum_{t=0}^{k} \mathbb{E}\left[\rho_t|X_0 = i\right] \mathbb{P}[\tau = k + 1|X_0 = i]. \qquad (20)
\end{aligned}$$

We observe that

$$\begin{aligned}
\mathbb{E}\left[\rho_t|X_0 = i\right] &= \sum_{j\in\mathcal{X}} \mathbb{E}[\rho_t|X_t = j]\mathbb{P}[X_t = j|X_0 = i] \\
&= \sum_{j\in\mathcal{X}} R_\pi(j)\mathbb{P}[X_t = j|X_0 = i] = \mathbb{E}\left[R_\pi(X_t)|X_0 = i\right].
\end{aligned}$$

Inserting this equation in (20) shows the claim, i.e., (19).

From (6), we conclude that

$$V_\pi(i) = \mathbb{E}\left[\sum_{t=0}^{\tau-1} R_\pi(X_t)|X_0 = i\right] = \langle\gamma_{<\tau}, R_\pi\rangle.$$

In the second part of the proof, we will use the evolution (9), to evaluate $\langle\gamma^\mu_{<\tau}, R(\pi)\rangle$. We claim that

$$0 = \langle\mu|_H, G(\pi)f|_H\rangle - \langle\gamma^\mu_{<\tau}|_H, f|_H\rangle$$

for any f such that $f(e) = 0$ for all $e \in E$, and for any initial measure μ. The claim leads us to the conclusion

$$0 = G(\pi)R_\pi - V_\pi.$$

To prove the claim, without loss of the generality, suppose that the states are numbered such that the first states belong to H and the remaining to E. Then the (possibly infinite-dimensional) transition matrix $P := P(\pi)$ is decomposed as

$$P = \begin{bmatrix} Q & P^H_E \\ P^E_H & P^E_E \end{bmatrix}, \text{ and } \mathcal{L} = \begin{bmatrix} Q - I_H & P^H_E \\ P^E_H & P^E_E - I_E \end{bmatrix}$$

where $Q := P(\pi)|_H$. We define a matrix

$$\tilde{G} = \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix}$$

where G is the green operator defined in (2). By the relation (3), we have

$$\mathcal{L}\tilde{G} = -\begin{bmatrix} I_H & 0 \\ 0 & 0 \end{bmatrix}$$

and $\lambda^\mu_\tau|_H G = \mu|_H G - \gamma^\mu_{<\tau}|_H$. On the other hand, τ is the first hitting time of E, therefore $\lambda^\mu_\tau|_H = 0$. In conclusion

$$0 = \mu|_H G - \gamma^\mu_{<\tau}|_H. \tag{21}$$

Suppose that $f = \begin{bmatrix} f|_H & f|_E \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} f|_H & 0 \end{bmatrix}^{\mathrm{T}}$. From (9), it follows that:

$$0 = \langle\mu|_H, Gf|_H\rangle - \langle\gamma^\mu_{<T}|_H, f|_H\rangle.$$

This proves the claim as $f|_H$ is arbitrary. We conclude that $V_\pi = G(\pi)R_\pi$. ■

We strive to solve the following optimization problem:

$$V_*(i) = \min_{\pi\in\mathcal{D}^H} V_\pi(i) \tag{22}$$

where H is the taboo set, and $\mathcal{D}^H$ is the Cartesian product of $|H|$ copies of the simplex $\mathcal{D}$. V^* is called the *value function*.

From (3), we obtain that $G(\pi) = I_H + G(\pi)Q(\pi)$, hence

$$V_\pi = R_\pi + Q(\pi)G(\pi)R_\pi$$

and the result is the celebrated well-known formula in DP

$$V_\pi = R_\pi + Q(\pi)V_\pi \tag{23}$$

with the boundary condition $V_\pi(e) = 0$ for $e \in E$.

Let $\Delta(\pi) := I - Q(\pi)$ be the *discrete Laplacian operator* on the taboo set H, and then we write (23) as

$$\Delta(\pi)V_\pi = R_\pi.$$

Let us recall some standard results of DP [2], which will be instrumental in the following sections. Suppose that there is at least one transient policy π. The optimal cost V_* restricted to the set H satisfies Bellman's equation

$$V = \min_{\pi\in\mathcal{D}^H} \left[R_\pi + Q(\pi)V\right]. \tag{24}$$

Furthermore, the function $V_*|_H$ is the (coordinatewise) limit of the sequence (V^n) defined by

$$V^{n+1} = \min_{\pi\in\mathcal{D}^H} \left[R_\pi + Q(\pi)V^n\right] \tag{25}$$

with an arbitrary initial condition $V^0 \geq 0$.

B. Linear Programming

We follow the idea of [9] showing that the value function (22) is the largest among the functions $V : E \to \mathbb{R}$ satisfying the inequality

$$\Delta(\pi)V \leq R_\pi \text{ for all } \pi \in \mathcal{D}^H.$$

Such functions are called subharmonic vectors in the MDP context.

Lemma 2: The value function satisfies

$$V_* = \sup \mathcal{V} \tag{26}$$

where $\mathcal{V} := \{V \in \mathbb{R}^H |\ \Delta(\pi)V \leq R_\pi \text{ for all } \pi \in \mathcal{D}^H\}$, and sup is to be understood coordinatewise.

We use the action-state evolution (18) to formulate LP. At the outset, we define

$$V_\pi^\mu := \mathbb{E}^\mu\left[\sum_{t=0}^{\tau-1} r_\pi(X_t)\right]$$

where as before, τ is the first hitting time of E, $\tau = \tau_E$. Furthermore, we let $V_*^\mu = \min_{\pi\in\mathcal{D}^H} V_\pi^\mu$.

Lemma 3: Suppose that X_0 has the initial distribution μ with the support in H, then

$$V_*^\mu = \min \sum_{u\in\mathcal{U}} \langle \rho(\cdot, u), \overline{\alpha}^\mu(\cdot, u)\rangle$$

over the measure $\overline{\alpha}$ on $H \times \mathcal{U}$ that satisfies

$$0 = \mu(\cdot) + \sum_{u\in\mathcal{U}} \overline{\alpha}^\mu(\cdot, u)(Q(u) - I). \tag{27}$$

Furthermore, the policy is given by

$$\pi_{i,u} = \frac{\overline{\alpha}^\mu(i, u)}{\alpha^\mu(i)} \tag{28}$$

where $\alpha^\mu(i) = \sum_{u\in\mathcal{U}} \overline{\alpha}^\mu(i, u)$.

Proof: From (15), we have

$$V_\pi^\mu = \sum_{u\in\mathcal{U}} \langle \rho(\cdot, u), \overline{\gamma}_{<\tau}^\mu(\cdot, u)\rangle.$$

The state action evolution equation

$$\sum_{u\in\mathcal{U}} \overline{\lambda}_\tau^\mu(\cdot, u) = \mu(\cdot) + \sum_{u\in\mathcal{U}} \overline{\gamma}_{<\tau}^\mu(\cdot, u)\mathcal{L}(u)$$

uniquely characterizes the process (X_t, U_t). Subsequently, noticing that the support of $\overline{\lambda}_\tau^\mu(\cdot, u)$ is in the complement of H, from (21) in the proof of Proposition 1

$$0 = \mu(i) + \sum_{u\in\mathcal{U}}\sum_{j\in H} \overline{\gamma}_{<\tau}^\mu(j, u)(p_{jui} - \delta_i(j)) \text{ for } i \in H.$$

By substituting $\overline{\gamma}_{<\tau}^\mu$ by $\overline{\alpha}^\mu$, the equality (27) follows. Finally, the policy (28) follows from (16) and (17). ■

V. Safety

We formulate safety as a DP problem. In the previous section, we have considered the terminal set E and its complement, the taboo set H. We extend this situation by adding an extra set U, the set of forbidden states. We suppose that U is disjoint from E. Now, the taboo set is $H = \mathcal{X} \setminus (U \cup E)$.

The definition of safety is taken from [11]. For each state in $\mathcal{X}$, the safety function gives the probability that the realizations hit the forbidden set U before reaching the target set E.

We consider the problem of finding a policy π such that the safety function satisfies the following condition:

$$S_\pi(i) := \mathbb{P}^i[\tau_U < \tau_E] = \mathbb{P}[\tau_U < \tau_E|\ X_0 = i] \leq p$$

where τ_A is the first hitting time of a set A. We have again suppressed the policy π in the notation, $\mathbb{P} = \mathbb{P}(\pi)$.

To compute the safety function S_π, we apply the evolution (9) with the initial distribution μ concentrated at i, and $\tau = \tau_{E\cup U}$ equal to the first hitting time of $E \cup U$

$$\langle \lambda_\tau^i, f\rangle = f(i) + \langle \gamma_{<\tau}^i, \mathcal{L}(\pi)f\rangle \text{ for all } f : \mathcal{X} \to \mathrm{R}.$$

We observe that the safety function $S_\pi(i) = \lambda_\tau^i(U)$. We unfold the evolution equation

$$\sum_{k\in U\cup E} \lambda_\tau^i(k)f(k) = f(i) + \sum_{j\in H} \gamma_{<\tau}^i(j)(\mathcal{L}(\pi)f)(j). \tag{29}$$

Since the function f is arbitrary, for the specific choice of f such that $f(j) = 0$ for $j \in E$, $f(j) = 1$ for $j \in U$, and $(\mathcal{L}(\pi)f)(j) = 0$ for $j \in H$, we have $\sum_{k\in U} \lambda_\tau^i(k) = f(i)$.

In conclusion, the safety function S_π is the solution s of the following problem:

$$(\mathcal{L}(\pi)s)(j) = 0, \forall j \in H \tag{30a}$$

$$s(j) = 1, \forall j \in U \tag{30b}$$

$$s(j) = 0, \forall j \in E. \tag{30c}$$

The problem (30) is known as the Dirichlet problem. Its solution is unique. Since (30) is linear in s, we formulate it in terms of matrices. To this end, we suppose the state are numbered in the following order: the states in H are first, then in U, and finally in E. We decompose $P := P(\pi)$ as follows:

$$P = \begin{bmatrix} Q & P_H^U & P_H^E \\ P_U^H & P_U^U & P_U^E \\ P_E^H & P_E^U & P_E^E \end{bmatrix}. \tag{31}$$

Lemma 4: Suppose that the MDP (X_t) with a policy π is being transient on H. Let

$$K_\pi := P_H^U(\pi)\mathbb{1}_U \tag{32}$$

where $\mathbb{1}_U$ is the column vector of 1 s of length $|U|$. Then, the safety function is given by

$$S_\pi|_H = G(\pi)K_\pi \tag{33}$$

and it is the solution of the following Poisson equation:

$$S_\pi|_H = Q(\pi)S_\pi|_H + K_\pi. \tag{34}$$

Furthermore, the sequence (S_π^n) defined by

$$S_\pi^{n+1} = Q(\pi)S_\pi^n + K_\pi \tag{35}$$

for an arbitrary S_π^0 converges pointwise to $S_\pi|_H$.

Proof: Applying the transition matrix (31) to the Dirichlet problem (30) we get (33). Then (33) and (3) imply (34).

We regard (35) as a discrete-time dynamical systems, and observe that the eigenvalues of $Q(\pi)$ are in the open unit disk. Consequently, the sequence (S_π^n) converges to $G(\pi)K_\pi$. ■

The value of K_π at a state $i \in H$ is the probability of reaching one of the forbidden states in U in a single time-step. The characterization of the safety function in (33) corresponds to the probability of reaching U after staying entirely in the set H.

Safety function can be computed as the expectation of a cumulative reward. Suppose μ is the initial distribution of the process (X_t), $X_0 \sim \mu$. We consider the safety function

$$S_\pi^\mu := \mathbb{P}^\mu[\tau_U < \tau_E | \, X_0] = \langle \mu, S_\pi \rangle.$$

Corollary 1: Suppose μ is the initial distribution of the process (X_t) and the support of μ is in H. The safety function $S_\pi(\mu)$ is given by

$$S_\pi^\mu = \mathbb{E}^\mu \sum_{t=0}^{\tau-1} \kappa(X_t, U_t) \tag{36}$$

where $\tau = \tau_{U\cup E}$, and $\kappa(i,u) = \sum_{j\in U} p_{iuj}$, for all $i \in H$.

Proof: From (32)

$$K_\pi(i) = \sum_{j\in U} p_{ij}(\pi) = \sum_{j\in U}\sum_{u\in\mathcal{U}} \pi_{iu}p_{iuj} = \sum_{u\in\mathcal{U}} \pi_{iu} \sum_{j\in U} p_{iuj}$$
$$= \sum_{u\in\mathcal{U}} \pi_{iu}\kappa(i,u).$$

On the other hand, form Lemma 4, $S_\pi^\mu = G(\pi)K_\pi$. From Proposition 1, $S_\pi^\mu(i) = \mathbb{E}_\pi[\sum_t^{\tau-1} \kappa_t | X_0 = i]$ with $\kappa_t = \kappa(X_t, U_t)$. ■

Corollary 1 allows to formulate the safe optimization as a constrained MDP. Specifically, from Lemma 3, the safety function S_π^μ is computed from

$$S_\pi^\mu = \sum_{i\in H} \gamma_{<\tau}^\mu(i)\kappa(i,\pi(i)) \tag{37}$$

where $\gamma_{<\tau}^\mu$ is the occupation measure, the solution of the evolution equation

$$0 = \mu(j) + \sum_{i\in H} \gamma_{<\tau}^\mu(i)(p_{ij}(\pi) - \delta_j(i)), \; j \in H. \tag{38}$$

The measure $\gamma_{<\tau}^\mu$ (restricted to H) corresponds to the occupation measure of H, when there is no escape. Subsequently, the function κ is the reward to go into the set U.

VI. Safety Guarantee in Stochastic Optimization

We combine safety and stochastic optimization. The objective is to reach the goal states in the subset $E \subset \mathcal{X}$ with minimal expected cumulative reward subject to the safety constraints of reaching E before the set U of unsafe states with a probability below a prior level p. In other words, for $0 \le p \le 1$, and an initial distribution μ of X_0 with μ supported on H, we strive to find the minimum V^* of the cost

$$V_\pi^\mu := \mathbb{E}_\pi^\mu\left[\sum_{t=0}^{\tau_E-1} \rho(X_t, U_t)\right] \tag{39}$$

on H subject to $S_\pi^\mu \le p$, and over the stationary (mixed) policies $\pi \in \mathcal{D}^{\mathcal{X}\setminus E}$. First, we reformulate the safety function in the constraint as the cost

$$S_\pi^\mu = \mathbb{E}_\pi^\mu \sum_{t=0}^{\tau-1} \kappa(X_t, U_t) \tag{40}$$

with $\tau := \tau_{E\cup U}$, and $\kappa(i,u) := \sum_{j\in U} p_{iuj}$.

We notice the stopping times in the cost function (39) and in the constraint (40) are not the same; furthermore, $\tau = \tau_{U\cup E} \le \tau_E$ almost surely. As before, we define the (state-action) occupation measures and hitting measures for both stopping times. We use simplified notation, where we suppress the names of the stopping times and the initial measure

$$\overline{\gamma}^A(i,u) = \sum_{t=0}^{\infty} \mathbb{P}^\mu[X_t = i, U_t = u, t < \tau]$$

$$\overline{\gamma}^B(i,u) = \sum_{t=0}^{\infty} \mathbb{P}^\mu[X_t = i, U_t = u, \tau \le t < \tau_E]$$

and

$$\overline{\lambda}^A(i,u) = \mathbb{P}^\mu[X_\tau = i, U_\tau = u]$$

$$\overline{\lambda}^B(i,u) = \mathbb{P}^\mu[X_{\tau_E} = i, U_{\tau_E} = u]$$

for $(i,u) \in \mathcal{X} \times \mathcal{U}$. Consequently, there are two evolution equations: the first one governs the measures with the index A, and the second with the index B

$$\sum_{u\in\mathcal{U}} \overline{\lambda}^A(j,u) = \mu(j) + \sum_{u\in\mathcal{U}}\sum_{i\in\mathcal{X}} \overline{\gamma}^A(i,u)(p_{iuj} - \delta_j(i))$$
$$\sum_{u\in\mathcal{U}} \overline{\lambda}^B(j,u) = \sum_{u\in\mathcal{U}} \overline{\lambda}^A(j,u)$$
$$+ \sum_{u\in\mathcal{U}}\sum_{i\in\mathcal{X}} \overline{\gamma}^B(i,u)(p_{iuj} - \delta_j(i)). \tag{41}$$

Both $\overline{\lambda}^A$ and $\overline{\lambda}^B$ are zero on H, $\overline{\lambda}^B$ is additionally zero on U. By summing the two evolution equations in (41), we observe that $\overline{\gamma} := \overline{\gamma}^A + \overline{\gamma}^B$ satisfies the following evolution equation:

$$\sum_{u\in\mathcal{U}} \overline{\lambda}^B(\cdot,u) = \mu(\cdot) + \sum_{u\in\mathcal{U}} \overline{\gamma}(\cdot,u)\mathcal{L}(u).$$

In conclusion, the cost and the constraint are expressed in terms of the occupation measures that are

$$V_\pi^\mu = \sum_{(i,u)\in(H\cup U)\times\mathcal{U}} \overline{\gamma}(i,u)\rho(i,u)$$

subject to the inequality constraint

$$S_\pi^\mu = \sum_{(i,u)\in H\times\mathcal{U}} \overline{\gamma}^A(i,u)\kappa(i,u) \le p.$$

To conclude, safe stochastic optimization is formulated as the following linear program.

Proposition 2: Suppose that there is a transient policy $\pi \in \mathcal{D}^{\mathcal{X}\setminus E}$ then the minimum of V_π^μ over stationary policies $\pi \in \mathcal{D}^{\mathcal{X}\setminus E}$ is the solution of the following linear program:

$$\min_{\pi \in \mathcal{D}^{\mathcal{X}\setminus E}} V_\pi^\mu = \min \sum_{i\in H\cup U} \sum_{u\in\mathcal{U}} (\overline{\gamma}^A(i,u) + \overline{\gamma}^B(i,u))\rho(i,u)$$

subject to

$$0 \le \overline{\gamma}^A \text{ and } 0 \le \overline{\gamma}^B$$

$$0 = \overline{\gamma}^A(j,u), \text{ for } j \in U \cup E \text{ and } u \in \mathcal{U}$$

$$0 = \overline{\gamma}^B(j,u), \text{ for } j \in U \text{ and } u \in \mathcal{U}$$

$$0 = \mu(j) + \sum_{u\in\mathcal{U}} \sum_{i\in H} \overline{\gamma}^A(i,u)(p_{iuj} - \delta_j(i)), \text{ for } j \in H$$

$$0 = \sum_{u\in\mathcal{U}} \sum_{i\in H\cup U} \overline{\gamma}^B(i,u)(p_{iuj} - \delta_j(i)), \text{ for } j \in H$$

$$0 = \sum_{u\in\mathcal{U}} (\sum_{i\in H} \left(\overline{\gamma}^A(i,u) + \overline{\gamma}^B(i,u)\right)(p_{iuj} - \delta_j(i))$$

$$+ \sum_{i\in U} \overline{\gamma}^B(i,u)(p_{iuj} - \delta_j(i))), \text{ for } j \in U$$

$$p \ge \sum_{(i,u)\in H\times\mathcal{U}} \overline{\gamma}^A(i,u)\kappa(i,u).$$

The optimal policy is given by

$$\pi_{i,u} = \frac{\overline{\gamma}^A(i,u) + \overline{\gamma}^B(i,u)}{\gamma(i)} \tag{42}$$

where $\gamma(i) = \sum_{u\in\mathcal{U}} (\overline{\gamma}^A(i,u) + \overline{\gamma}^B(i,u))$.

Proof: The constraints follow from (41) noticing that the supports, $\text{supp}(\overline{\lambda}^A) \subseteq U \cup E$, $\text{supp}(\overline{\lambda}^B) \subseteq E$, $\text{supp}(\overline{\gamma}^A) \subseteq H$, and $\text{supp}(\overline{\gamma}^B) \subseteq H \cup U$.

Furthermore, the last equality constraint is the consequence of the following two qualities, for $j \in U$:

$$0 = \sum_{u\in\mathcal{U}} \overline{\lambda}^A(j,u) + \sum_{u\in\mathcal{U}} \sum_{i\in H\cup U} \overline{\gamma}^B(i,u)(p_{iuj} - \delta_j(i))$$

and

$$\sum_{u\in\mathcal{U}} \overline{\lambda}^A(j,u) = \sum_{u\in\mathcal{U}} \sum_{i\in H} \overline{\gamma}^A(i,u)(p_{iuj} - \delta_j(i)).$$

The policy (42) follows from the observation that:

$$\overline{\gamma}^\mu_{<\tau}(i,u) = \overline{\gamma}^A(j,u) + \overline{\gamma}^B(j,u)$$

and from (16) and (17). ■

A. Illustration

We provide a simple example illustrating how to use Proposition 2. Consider MDP in Fig. 1, where the state-space $\mathcal{X}$ consists of seven states. The initial state is 1. The unsafe set $U = \{4, 5\}$, the target set $E = \{6, 7\}$, and the taboo set $H = \mathcal{X} \setminus (E \cup U) = \{1, 2, 3\}$. There are two actions 1 and 2, i.e., $\mathcal{U} = \{1, 2\}$. We suppose that the reward $\rho(i,u) = 1$ for all states and actions. The decision variables in Proposition 2 are the occupation measures $\overline{\gamma}^A(i,u) \ge 0$ and $\overline{\gamma}^B(i,u) \ge 0$. The reward $\kappa(i,u)$ becomes 1 for $(i,u) \in \{(2,2),(3,2)\}$, q for $(i,u) = (3,1)$, and 0 for other state-action pairs. We minimize the sum

$$\overline{\gamma}^A(1,1) + \cdots + \overline{\gamma}^A(3,2) + \overline{\gamma}^B(1,1) + \cdots + \overline{\gamma}^B(5,2)$$

subject to

$$0 = 1 - \overline{\gamma}^A(1,1) - \overline{\gamma}^A(1,2)$$

$$0 = \overline{\gamma}^B(1,1) + \overline{\gamma}^B(1,2)$$

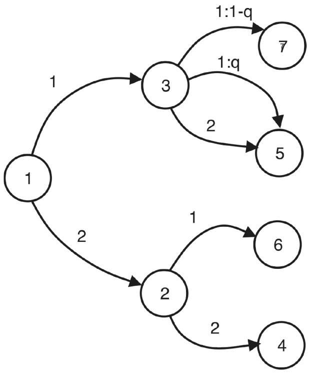

Fig. 1. Unsafe states are 4 and 5; whereas, the goal states are 6 and 7. At each state there two action 1 and 2, i.e., $\mathcal{U} = \{1, 2\}$.

$$0 = \overline{\gamma}^c(1,2) - \overline{\gamma}^c(2,1) - \overline{\gamma}^c(2,2) \text{ with } c \in \{A, B\}$$

$$0 = \overline{\gamma}^c(1,1) - \overline{\gamma}^c(3,1) - \overline{\gamma}^c(3,2) \text{ with } c \in \{A, B\}$$

$$0 = \overline{\gamma}^A(2,2) + \overline{\gamma}^B(2,2) - \overline{\gamma}^B(4,1) - \overline{\gamma}^B(4,2)$$

$$0 = (\overline{\gamma}^A(3,1) + \overline{\gamma}^B(3,1))(q) + (\overline{\gamma}^A(3,2) + \overline{\gamma}^B(3,2))$$

$$- \overline{\gamma}^B(5,1) - \overline{\gamma}^B(5,2)$$

$$p \ge \overline{\gamma}^A(2,2) + q\overline{\gamma}^A(3,1) + \overline{\gamma}^A(3,2).$$

VII. Conclusion

In this work, we have formulated the problem of stochastic optimization for MDPs with safety guarantees. First, we have expressed safety as the accumulated cost of the probability of getting into the unsafe set. Subsequently, we have used the evolution equation to devise a linear program for computing the optimal stationary policy that adheres to safety specifications.

References

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1. Nashua, NH, USA: Athena Scientific, 2017.

[2] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 2. Nashua, NH, USA: Athena Scientific, 2018.

[3] L. Busoniu, R. Babuska, B. D. Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*, 1st ed.Boca Raton, FL, USA: CRC Press, Inc., 2010.

[4] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: A survey,” *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1238–1274, Sep. 2013.

[5] S. Shammah and A. Shashua, “Safe, multi-agent, reinforcement learning for autonomous driving,” 2016, *arXiv:1610.03295*.

[6] C.-C. Chang, J. Tsai, P.-C. Lu, and C.-A. Lai, “Accuracy improvement of autonomous straight take-off, flying forward, and landing of a drone with deep reinforcement learning,” *Int. J. Comput. Intell. Syst.*, vol. 13, no. 1, pp. 914–919, 2020.

[7] J. Val Ledesma, R. Wisniewski, and C. Kallesøe, “Optimal control for water networks with unknown dynamics,” in *21st IFAC World Cong.*, 2020, pp. 6577–6582.

[8] M. Haviv, “On constrained Markov decision processes,” *Operations Res. Lett.*, vol. 19, no. 1, pp. 25–28, 1996.

[9] E. Altman, *Constrained Markov Decision Processes*, 1st ed. Routledge, New York, 1999.

[10] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Ser. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., 1994a.

[11] M. L. Bujorianu, R. Wisniewski, and E. Boulougouris, “Stochastic safety for Markov chains,” *IEEE Contr. Syst. Lett.*, vol. 5, no. 2, pp. 427–432, Apr. 2021.

[12] R. Wisniewski, M. L. Bujorianu, and C. Sloth, "p-safe analysis of stochastic hybrid processes," *IEEE Trans. Autom. Control*, vol. 65, no. 12, pp. 5220–5235, Dec. 2020.

[13] C. Baier and J.-P. Katoen, *Principles of Model Checking*. Cambridge, MA, USA: MIT Press, 2008.

[14] R. Wisniewski and M. L. Bujorianu, "Stochastic safety analysis of stochastic hybrid systems," in *Proc. IEEE 56th Annu. Conf. Decis. Control*, 2017, pp. 2390–2395.

[15] M. L. Bujorianu and R. Wisniewski, "New insights on p-safety of stochastic systems," in *Proc. IEEE 58th Conf. Decis. Control*, 2019, pp. 4433–4438.

[16] A. Hordijk and L. C. M. Kallenberg, "Linear programming and Markov decision chains," *Manage. Sci.*, vol. 25, no. 4, pp. 352–362, 1979.

[17] S. Prajna, A. Jadbabaie, and G. J. Pappas, "A framework for worst-case and stochastic safety verification using barrier certificates," *IEEE Trans. Autom. Control*, vol. 52, no. 8, pp. 1415–1428, Aug. 2007.

[18] Z. Marvi and B. Kiumarsi, "Safe reinforcement learning: A control barrier function optimization approach," *Int. J. Robust Nonlinear Control*, vol. 31, no. 6, pp. 1923–1940, 2021.

[19] Z. Li, U. Kalabić, and T. Chu, "Safe reinforcement learning: Learning with supervision using a constraint-admissible set," in *Proc. Annu. Amer. Control Conf.*, 2018, pp. 6390–6395.

[20] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," in *Proc. AAAI Conf. Art. Intell.*, 2017, pp. 2669–2678.

[21] O. Bastani, "Safe reinforcement learning with nonlinear dynamics via model predictive shielding," in *Amer. Control Conf.*2021, pp. 3488–3494.

[22] K. P. Wabersich, L. Hewing, A. Carron, and M. N. Zeilinger, "Probabilistic model predictive safety certification for learning-based control," *IEEE Trans. Automat. Control*, vol. 67, no. 1, pp. 176–188, Jan. 2022.

[23] S. Pranger, B. Konighofer, M. Tappler, M. Deixelberger, N. Jansen, and R. Bloem, "Adaptive shielding under uncertainty," in *Amer. Control Conf.*, 2021, pp. 3467–3474.

[24] N. Jansen, B. Könighofer, J. S. L Junges, A. C. Serban, R. Bloem, and Konnov I, "Safe reinforcement learning using probabilistic shields," in *Proc. 31st Int. Conf. Concurrency Theory*, 2020, pp. 3:1–3:16.

[25] K. Helmes, S. Röhl, and R. H. Stockbridge, "Computing moments of the exit time distribution for Markov processes by linear programming," *Oper. Res.*, vol. 49, no. 4, pp. 516–530, 2001.

[26] A. T. Bharucha-Reid, Ed., *Probabilistic Methods in Applied Mathematics*. vol. 3. New York, NY, USA: Academic Press [Harcourt Brace Jovanovich, Publishers], 1973.

[27] A. G. Bhatt and R. L. Karandikar, "Invariant measures and evolution equations for Markov processes characterized via martingale problems," *Ann. Probab.*, vol. 21, no. 4, pp. 2246–2268, 1993.