Online Learning of Safety function for Markov Decision Processes

Abhijit Mazumdar, Rafal Wisniewski and Manuela L. Bujorianu

Abstract— **In this paper, we aim to study safety specifications for a Markov decision process with stochastic stopping time in an *almost* model-free setting. Our approach involves characterizing a *proxy set* of the states that are *near* in a probabilistic sense to the set of unsafe states - *forbidden set*. We also provide results that relate safety function with reinforcement learning. Consequently, we develop an online algorithm based on the temporal difference method to compute the safety function. Finally, we provide simulation results that demonstrate our work in a simple example.**

Index Terms— **Markov decision processes, safety, online learning, temporal difference, proxy set.**

I. Introduction

Design of control policies for safety-critical systems, such as the electricity grid, power plant, and autonomous vehicles, must be such that the closed-loop system ensures optimal operation while being safe. Consequently, safety assessment is crucial for safety-critical systems. Finding the mathematical model of a system can be too tedious, or knowledge about the operating environment may not be available. In that case, one has to develop a method based on the observed data to make a safety assessment. To this end, we propose to use reinforcement learning (RL). Specifically, RL deals with learning value function and policy evaluation in a model-free setting.

Related literature: The concept of safety that we consider in this work is p-safety. It has been studied extensively in [1]–[4]. A system is called to be p-safe if the system states do not visit the dangerous states before reaching the control goal with a probability more than p.

An overview paper [5] discusses safety definitions and reinforcement learning methods with safe exploration. Three concepts of safety are examined: safety through cost, safety through labeling, and safety through ergodicity.

Data-driven safety verification has received considerable attention recently [6]–[10]. In [6], [7], a data-driven method based on barrier certificate is proposed to formally verify the safety of discrete-time continuous systems whose dynamics are not known. For networks of discrete-time subsystems, [8] presents a data-driven approach with formal guarantees. It involves finding a sub-barrier function for each subsystem. In [7], [9], [10], data-driven safety verification is considered. Based on the notion of barrier certificates, the safety verification problem is formulated as a robust convex problem. For a robust convex problem, certain constraints need to be satisfied over the whole state space and this makes it a semi-infinite linear program. To solve the robust convex problem efficiently, it is converted to a scenario, where finitely many data are collected and the robust convex problem is approximated with the finite data points.

Motivations and Contributions: Existing works on data-driven safety verification are based on the assumption that a rich offline data-set is available. However, many a times, safety verification needs to be performed in an online fashion so that an online safe learning algorithm can be designed.

In this paper, we present an online data-driven method to assess the safety of a Markov decision process (MDP) for a given policy with stochastic stopping time. For defining safety, we consider two subsets of the state-space: a set of target states, and a set of forbidden states that need to be avoided. Specifically, we define safety as the hitting probability of the forbidden set before the process hits the target set. Since our method is meant to be online, we do not allow the process to hit the set of forbidden states. Because, in practice, once the process hits the forbidden states, the process or the system might get severely destroyed. Thus, we define a set of states, and call it *proxy set*, visiting which, we find lower and upper bounds for the safety function. The following are our main contributions:

i) We define a subset of the state-space called *proxy set* in order to learn the safety function without visiting the forbidden states. In short, proxy states are in the vicinity, in a probabilistic sense, of the forbidden states.
ii) We present a result that gives a bound on the safety function if the upper and lower bounds of the transition probabilities from the proxy set to the forbidden set are known instead of the true transition probability.
iii) We present an online algorithm based on one-step temporal difference, i.e., TD(0), and illustrate the method on a simple numerical example.

The organization of the rest of the paper is as follows. In Section II, we recall the definition of an MDP and present various related notations. In section III, we formally define the safety function for an MDP. Further, we introduce the definition of the proxy set, and provide lower and upper bounds on the safety function. Section IV deals with the learning of the safety function without knowing the transition

The work of the first and the second authors has been supported by the Poul Due Jensens Foundation under project SWIft.The work of the third author has been funded by the EPSRC project EP/R006865/1: Interface reasoning for interacting systems (IRIS).

A. Mazumdar and R. Wisniewski are with the Section of Automation & Control, Aalborg University, 9220 Aalborg East, Denmark (e-mail: {abma, raf}@es.aau.dk).

Manuela L. Bujorianu is with the Department of Computer Science, University College London, UK (e-mail: l.bujorianu@ucl.ac.uk).

probabilities. An online learning algorithm based on TD(0) is presented in this section. In section V, by considering an example, we demonstrate that our proposed algorithm learns a better estimate of the safety function if the number of episodes increases. Furthermore, we present computation results corresponding to the upper and lower bounds of the safety function for the case when exact transition probabilities from the proxy set to the forbidden set are not known. Finally, we present the concluding remark and our future work plan in section VI.

II. Preliminaries

Suppose $\mathcal{X}$ is a finite set of states and $\mathcal{A}$ is a finite set of actions. We construct the sample space Ω of all sequences of the form $\omega = (x_1, a_1, x_2, a_2, \ldots) \in (\mathcal{X} \times \mathcal{A})^\infty$ with $x_i \in \mathcal{X}$ and $a_i \in \mathcal{A}$. The sample space Ω is equipped with the σ-algebra $\mathcal{F}$ generated by coordinate mappings: $X_t(\omega) = x_t$ and $A_t(\omega) = a_t$. With a slight abuse of notation, we shall use X_t and A_t to denote random variables, whereas x_t and a_t are deterministic values, their realizations, at time-step t. We suppose that μ is the distribution of the initial states X_0. In this work, we consider stationary policies, i.e., maps $\pi : \mathcal{X} \to \Delta(\mathcal{A})$, with $\Delta(\mathcal{A}) = \{(p_1, \ldots, p_{|\mathcal{A}|}) \in [0,1]^{\mathcal{A}} \mid p_1 + \ldots + p_{|\mathcal{A}|} = 1\}$. A sub-policy π' for a subset of $W \subseteq \mathcal{X}$ is defined as $\pi' : W \to \Delta(\mathcal{A})$. For a fixed initial distribution μ and a policy π, we define recursively the probability $\mathbb{P}^\mu_\pi$ on $\mathcal{F}$ by

$$\begin{aligned} \mathbb{P}^\mu_\pi[X_1 = x] &= \mu(x) \\ \mathbb{P}^\mu_\pi[A_t = a \mid X_t = x] &= \pi(x)(a) \\ \mathbb{P}^\mu_\pi[X_{t+1} = y \mid X_t = x, A_t = a] &= p_{x,a,y}. \end{aligned}$$

Specifically, the process (X_t) with stationary policy is characterized by the transition probability $p_{x,y} = \sum_{a \in \mathcal{A}} p_{x,a,y} \pi(x)(a)$. Hence, the process (X_t) is homogeneous.

We write $\mathbb{P}^x_\pi := \mathbb{P}^{\delta_x}_\pi$ for the delta distribution concentrated at x. The expectation with respect to $\mathbb{P}^x_\pi$ is denoted $\mathbb{E}^x_\pi$. For a set A, τ_A represents the first hitting time of the set.

III. Safety specification for Markov decision processes with proxy states:

Consider an MDP with the state-action space $(\mathcal{X}, \mathcal{A})$. We partition the state-space into a target set $E \subset \mathcal{X}$, a set of forbidden states U, and $H := \mathcal{X} \setminus (E \cup U)$ be the set of taboo state. The goal is to reach the target set E before reaching the forbidden set V.

Assumption 1: Following assumptions are followed throughout the paper.

1) Taboo set H is transient and hence, the hitting time of the target set E (τ_E) and the forbidden set U (τ_U) is finite, almost surely.
2) $\tau_{U \cup E} \le \tau_E$, almost surely.

For each state in $x \in \mathcal{X} \setminus U$, the safety function corresponds to the probability that the realizations hit the forbidden set U before the target set E is reached. Following [11], for a given policy π, we define the safety function for each state as follows:

$$S_\pi(x) := \mathbb{P}^x_\pi[\tau_U < \tau_E],$$

We call a state x to be p-safe, for a policy π, if $S_\pi(x) \le p$. An MDP, with a policy π, is called p-safe if $\max_{x \in \mathcal{X}} S_\pi(x) \le p$.

The reason for insisting on the probability of hitting the forbidden state before hitting the target set is that we assume the following: when the process (X_t) reaches the target set E, it is terminated as the decision objective is obtained. Otherwise, τ_E should be substituted by ∞.
From [11], if $\tau = \tau_{U \cup E}$ is almost surely finite, then the safety function is given by

$$S_\pi(x) = \mathbb{E}^x_\pi \sum_{t=0}^{\tau-1} \kappa(X_t, A_t), \tag{1}$$

where $\kappa(x, u) = \sum_{y \in U} p_{x,u,y}$ for all $x \in \mathcal{X} \backslash U$. In this work, the finiteness of τ will also be assumed. We strive to learn safety for a policy π. Nonetheless, we may not be allowed to visit the forbidden states. Hence, we estimate safety by only visiting certain *proxy states*.

A proxy set is characterized by the following definition.

Definition 1 (Proxy Set): Let π be a policy. Suppose that the sets $U, U' \subset \mathcal{X}$ are such that $U \subset U'$. Let q and w are in $[0,1]^{U' \backslash U}$. The subset U' is a (q, w)-*proxy set* or simply *proxy set* if it has the following properties:

P.1 $\tau_{U' \backslash U} < \tau_U$ almost surely.
P.2 $w(x) \le \mathbb{P}^x_\pi[\tau_U < \tau_E] \le q(x)$ for all $x \in U' \setminus U$. ■

Notice that $q(x)$ nor $w(x)$ are not necessarily the probabilities. They serve as a prior information the user have about $\mathbb{P}^x_\pi[\tau_U < \tau_E]$. In other words, the user is asked what she believes are the probability bounds $w(x)$ and $q(x)$ of hitting the forbidden set from the proxy states.

Although, we do not deal with the design of any policy in this work, we will shortly discuss the concept of a repelling policy. Suppose for a control goal, we wish to learn an optimal policy such that the forbidden states are not visited. In the process of learning this policy, we should visit the states in the set $\mathcal{X} \setminus U$ infinitely many times with probability 1, but repel the process from $U' \setminus U$ such that it does not hit U. In other words, we use $U' \setminus U$ as a safety buffer. We introduce the concept of repelling sub-policy as follows:

Definition 2 (Repelling Sub-Policy): For a policy π, a forbidden set U and a proxy set U', a repelling sub-policy π^R satisfies:

$$\mathbb{P}_{\pi^R}[X_{t+1} \in \mathcal{X} \setminus U | x_t \in U'] = 1 \tag{2}$$

■

The concept of the repelling policy can be interpreted in the following way. Repelling sub-policy π^R is able to keep the realizations of the process (X_t) away from the forbidden set U almost surely.

The design of a policy consists of learning a sub-policy π^L for the given MDP in $\mathcal{X} \setminus U'$ and applying repelling sub-policy π^R in $U' \setminus U$. As a result, we have policy π defined by

$$\pi(x) := \begin{cases} \pi^L(x), & \text{for } x \in \mathcal{X} \setminus U' \\ \pi^R(x), & \text{for } x \in U' \setminus U \end{cases} \tag{3}$$

We think about $U' \setminus U$ as a neighborhood of the forbidden set in the sense that the probability of hitting $U' \setminus U$ before hitting the forbidden set is 1.

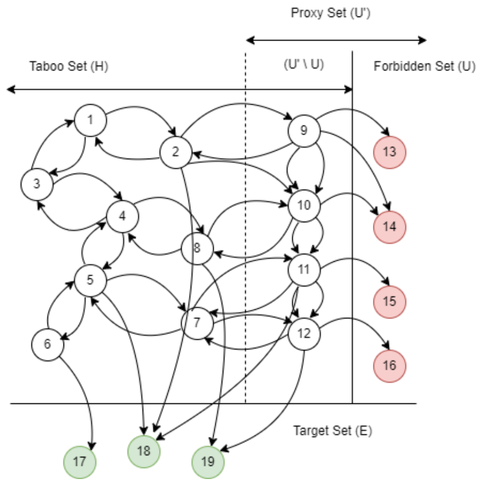

Fig. 1. Pictorial description of the taboo set, proxy set, forbidden set, and target set.

Proposition 1: Suppose U' is a (q, w)-proxy set. The functions q and w are extended to the whole state space by assuming that $q(x) = w(x) = 0$ for $x \in \mathcal{X} \setminus U'$. Let $\tau' = \tau_{(U' \setminus U) \cup E}$. Then

$$\mathbb{E}^x_\pi \sum_{t=0}^{\tau'} w(X_t) \leq S_\pi(x) \leq \mathbb{E}^x_\pi \sum_{t=0}^{\tau'} q(X_t). \tag{4}$$

The interpretation of (4) is that safety can be evaluated by the process of reinforcement learning of two cost functions $W(x) = \mathbb{E}^x_\pi \sum_{t=0}^{\tau'} w(X_t)$ and $Q(x) = \mathbb{E}^x_\pi \sum_{t=0}^{\tau'} q(X_t)$.

Proof: From the definition of the safety function, for each $x \in X \setminus U'$, we compute

$$\begin{aligned} S_\pi(x) &= \mathbb{P}^x_\pi[\tau_U < \tau_E] \\ &= \sum_{y \in \mathcal{X}} \mathbb{P}^y_\pi[\tau_U < \tau_E] \mathbb{P}^x_\pi[X_{\tau'} = y] \\ &= \sum_{y \in U' \setminus U} \mathbb{P}^y_\pi[\tau_U < \tau_E] \mathbb{P}^x_\pi[X_{\tau'} = y]. \end{aligned}$$

The second equality follows from the observation $\mathbb{P}^x_\pi[\tau_U < \tau_E | X_{\tau'} = y] = \mathbb{P}^y_\pi[\tau_U - \tau' < \tau_E - \tau'] = \mathbb{P}^y_\pi[\tau_U < \tau_E]$. On the other hand, the safety function is given by

$$\begin{aligned} S_\pi(x) &= \mathbb{E}^x_\pi \sum_{t=0}^{\tau-1} \kappa(X,A_t) \\ &= \mathbb{E}^x_\pi \left[\sum_{t=0}^{\tau-1} \kappa(X_t, A_t) \mid \tau' < \tau \right] \mathbb{P}^x_\pi[\tau' < \tau] \qquad (5) \\ &+ \mathbb{E}^x_\pi \left[\sum_{t=0}^{\tau-1} \kappa(X_t, A_t) \mid \tau' = \tau \right] \mathbb{P}^i_\pi[\tau' = \tau] \\ &= \mathbb{E}^x_\pi \left[\sum_{t=0}^{\tau-1} \kappa(X_t, A_t) \mid \tau' < \tau \right] \mathbb{P}^x_\pi[\tau' < \tau] \\ &= \mathbb{E}^x_\pi \left[\sum_{t=0}^{\tau'} \kappa(X_t, A_t) + \sum_{t=\tau'}^{\tau-1} \kappa(X_t, A_t) \mid \tau' < \tau \right] \\ &\cdot \ \mathbb{P}^x_\pi[\tau' < \tau] \qquad (6) \\ &= \mathbb{E}^x_\pi \left[\sum_{t=\tau'}^{\tau-1} \kappa(X_t, A_t) \right] \mathbb{P}^x_\pi[\tau' < \tau]. \qquad (7) \end{aligned}$$

with $\tau = \tau_{U \cup E}$, and $\kappa(x, a) = \sum_{y \in U} p_{x,a,y}$ for all $x \in U' \setminus U$, and $\kappa(x, a) = 0$ for $x \in \mathcal{X} \setminus U'$ and $\forall a \in \mathcal{A}$. In (7), we have used Property P.2 in the definition of the proxy set, i.e., either $\tau'(\omega) = \tau(\omega)$ when the realization of the process hits E before U' else $\tau' < \tau$. We re-write

$$\begin{aligned} S_\pi(x) &= \mathbb{E}^x_\pi \sum_{t=\tau'}^{\tau-1} \kappa(X_t, A_t) \ \mathbb{P}^x_\pi[\tau' < \tau] \\ &= \sum_{y \in U' \setminus U} \mathbb{E}_\pi \left[\sum_{t=\tau'}^{\tau-1} \kappa(X_t, A_t) \mid X_{\tau'} = y \right] \mathbb{P}^x_\pi[X_{\tau'} = y] \\ &\cdot \ \mathbb{P}^x_\pi[\tau' < \tau]. \end{aligned}$$

On the other hand,

$$\mathbb{E}_\pi \left[\sum_{t=\tau'}^{\tau-1} \kappa(X_t, A_t) \mid X_{\tau'} = y \right] = \mathbb{E}^y_\pi \sum_{t=0}^{\tau-\tau'-1} \kappa(X_t, A_t).$$

The above equation relates the prior information $\mathbb{P}^y[\tau_U < \tau_E]$ with the information we strive to characterize - $\kappa(X_t, A_t)$. We observe that

$$\mathbb{E}^y_\pi \sum_{t=0}^{\tau-\tau'-1} \kappa(X_t, A_t) = \mathbb{P}^y_\pi[\tau_U < \tau_E] \leq q(y).$$

Hence

$$\begin{aligned} S_\pi(x) &= \sum_{y \in U' \setminus U} \mathbb{P}^y_\pi[\tau_U < \tau_E] \mathbb{P}^x_\pi[X_{\tau'} = y] \mathbb{P}^x_\pi[\tau' < \tau] \\ &\leq \sum_{y \in U' \setminus U} q(y) \mathbb{P}^x_\pi[X_{\tau'} = y] \mathbb{P}^x_\pi[\tau' < \tau] \\ &= \sum_{y \in \mathcal{X}} q(y) \mathbb{P}^x_\pi[X_{\tau'} = y] \mathbb{P}^x_\pi[\tau' < \tau] \\ &= \mathbb{E}^x_\pi q(X_{\tau'}) \mathbb{P}^x_\pi[\tau' < \tau] \\ &= \mathbb{E}^x_\pi q(X_{\tau'}) \\ &= \mathbb{E}^x_\pi \sum_{t=0}^{\tau'} q(X_t). \end{aligned}$$

The last line follows from the fact that that X_t up to the stopping time $\tau' - 1$ belong to $\mathcal{X} \setminus U'$ and $q(X_t)$ is zero.

Similarly, for the lower bound

$$S_\pi(x) \geq \sum_{y \in U' \setminus U} w(y) \mathbb{P}^x_\pi[X_{\tau'} = y] = \mathbb{E}^x_\pi \sum_{t=0}^{\tau'} w(X_t).$$

■

Remark 1: Since we examine safety learning with unknown transition probabilities, once the process is in the set $U' \setminus U$ the probability of hitting U is unknown and can not be learned without going into U. Therefore, our approach is to assume a priori knowledge about this probability. ■

We do not consider the necessity of knowing the functions w and q to be too demanding as without any prior information $w(x)$ and $q(x)$ can be assumed to be 0 and 1.

IV. Learning safety function

Learning of safety involves two ingredients: the reward $\kappa(x,a) = \sum_{y \in U} p_{x,a,y}$, and learning with a stopping time. Since the transition probabilities are unknown, we estimate upper and lower bounds for the safety function $S_\pi(x)$ by using the prior knowledge q and w in the definition of the proxy state as the reward. Learning with a stopping time is discussed next. Suppose that $\mathcal{K}$ is the set of episodes, and each episode $k \in \mathcal{K}$ is a sequence $(x_1, a_1, r_1, x_2, a_2, r_2, x_3, \ldots, x_{k_{f-1}}, a_{k_{f-1}}, r_{k_{f-1}}, x_{k_f})$ of states, action and rewards, where k_f is the terminal iteration for episode k. Here, the reward function is defined as, either $r_t := \kappa(a_t, a_t)$ if one knows the transition probability from the set $U' \setminus U$ to the forbidden set U, or $r_t := q(x_t)$ ($r_t := w(x_t)$) if only upper bound (lower bound) of the transition probabilities to the forbidden set is known. We concatenate the episodes and generate a sequence of the form $(x'_t, a'_t, r'_t, x'_{t+1})$, where $x'_t = x_{t'}$, $x'_{t+1} = x_{t'+1}$, $r'_t = 0$, for all t', if $x_{t'} \notin U' \setminus U$ and $x_{t'+1} \in U' \setminus U$. If $x_{t'} \in U' \setminus U$, then $x'_t = x_{t'}$ and $x'_{t+1} = x_1$ due to the repelling action $a'_t = a_{t'}$ at iteration t', where x_1 will be the initial state of the next episode. Further, if $x \in U' \setminus U$, $r_t = q(x)$ for computing the upper bound and $r_t = w(x)$ for computing the upper bound of the safety function. Then, with a given policy π, the learning with TD(0) follows the computation [12]

$$\begin{aligned} V_{t+1}\left(x'_t\right) = &\left(1 - \alpha_t\left(x'_t\right)\right) V_t\left(x'_t\right) \\ &+ \alpha_t\left(x'_t\right)\left[r_t + V_t\left(x'_{t+1}\right)\right], \end{aligned}$$

By Lemma 1 and Theorem 1 in [13], for each $x \in \mathcal{X} \setminus U'$, $V_t(x)$ converges to $\mathbb{E}^x_\pi \sum_{t=0}^{\tau'} q\left(X_t\right)$ ($\tau' = \tau\left(U' \setminus U\right) \cup E$), if the reward is assumed to be $r_t = q(X_t)$, with probability 1 if the following two conditions are satisfied:

i) Each state is visited infinitely often

ii) The learning rate $\alpha_t(x)$, for each $x \in \mathcal{X} \setminus U'$, satisfies $0 \leq \alpha_t(x) \leq 1, \sum_t \alpha_t(x) = \infty$ and $\sum_t \alpha_t^2(x) < \infty$.

Similarly, under the above sufficient conditions on the learning rate, the learned safety function $V_t(x)$ converges to $\mathbb{E}^x_\pi \sum_{t=0}^{\tau'} w\left(X_t\right)$.

We now present an online algorithm based on one-step temporal difference (TD-(0)) to learn the safety function where one only needs to know the transition probability to the forbidden set U. It is not necessary to know about the transition probabilities in the taboo set H or to the target set E. Then, we use the algorithm to find bounds for the safety function without the need of knowing about the transition probabilities. To this end, we use the proxy set U' and the prior belief about the transition probabilities from the set $U' \setminus U$ to the forbidden set U.

Details of the Algorithm: Since stopping time is finite, the algorithm works in an episodic manner. Each episode starts with a random initial state and ends whenever the process hits any state $x \in U' \setminus U$. Thus the states $x \in U' \setminus U$ are the terminal states. If one knows the exact transition probabilities from the proxy set to the forbidden set then the true safety function can be estimated by using $r_t = \kappa(X_t, A_t)$ where $X_t \in U' \setminus U$ and $r_t = 0$ where $X_t \in \mathcal{X} \setminus U'$. For any state $x \in \mathcal{X} \setminus U'$, we denote the estimated safety function at any episode k by $\hat{S}_k(x)$ and will demonstrate that, as $k \to \infty$, $\hat{S}_k(x)$ converges to $S_\pi(x)$, almost surely.

Algorithm 1 : TD(0) algorithm for estimating the safety function:

```
Input: The policy $\pi$ for which safety to be evaluated, algorithm parameter $\alpha_t(x)$ for each $x \in H$.
Initialize: $V_1(x)$ for each $x \in H$ arbitrarily, $V_1(x) = 0$ for each $x \in U \cup E$, initial state $x_1$.
for Episodes ($k = 1, 2, ..., \mathcal{K}$) do
    if $k \geq 2$ then
        Initial state $x_1$ due to the repelling action
        Initialize $V_1(x)$ to $V_{t'+1}(x)$ for each $x \in H$.
        Initialize $\alpha_t(x)$ to $\alpha_{t'}(x)$ for each $x \in H$.
    end if
    for Iterations ($t = 1, 2, ..., \mathcal{T}$) do
        Choose action $A_t$ according to the policy $\pi$
        Observe reward $r_t$ and $X_{t+1}$
        Compute the following
            $V_{t+1}(x_t) \leftarrow V_t(x_t) + \alpha_t(x_t)[r_t + V_t(X_{t+1}) - V_t(x_t)]$   (8)
        if $x_t$ is a terminal state, i.e., $x_t \in U' \setminus U$ then
            Terminate the loop.
        end if
    end for
    Set $t'$ as the terminal iteration.
    Update safety function estimation $\hat{S}_k(x)$ as $\hat{S}_k(x) = V_{t'+1}(x)$ for each $x \in H$.
    Apply the repelling action for $x_{t'}$
end for
```

V. Illustrating Example

We consider a Markov decision process as shown in Fig. 2. The MDP has 11 states and 2 actions. The taboo set $H =$

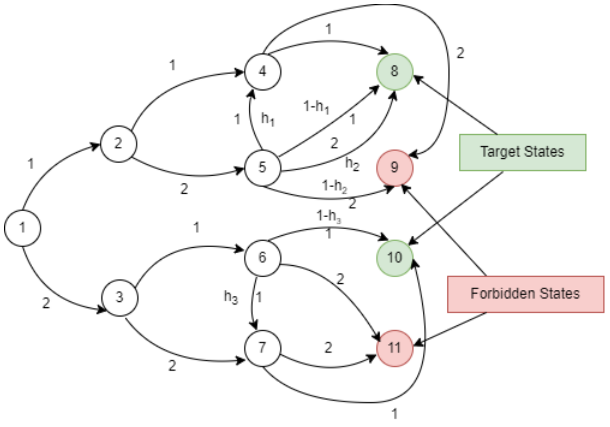

Fig. 2. An MDP for estimating the safety function

$\{1, 2, 3, 4, 5, 6, 7\}$, target set is $E = \{8, 10\}$, the forbidden set is $U = \{9, 11\}$ and the proxy set $U' = \{4, 5, 6, 7, 9, 11\}$ and $U' \setminus U = \{4, 5, 6, 7\}$. The MDP has the following parameters: $p_{1,1,2} = 1$, $p_{1,2,3} = 1$, $p_{2,1,4} = 1$, $p_{2,2,5} = 1$, $p_{3,1,6} = 1$, $p_{3,2,7} = 1$, $p_{4,1,8} = 1$, $p_{4,2,9} = 1$, $p_{5,1,4} = h_1$, $p_{5,1,8} = 1 - h_1$, $p_{5,2,8} = h_2$, $p_{5,2,9} = 1 - h_2$, $p_{6,1,7} = h_3$, $p_{6,1,10} = 1 - h_3$, $p_{6,2,11} = 1$, $p_{7,1,10} = 1$, $p_{7,2,11} = 1$.

For evaluating the safety function, we consider a uniform random policy, i.e., $\pi(x)(a) = 0.5, \forall x \in \mathcal{X}$ and $\forall a \in \{1, 2\}$. Thus the transition probabilities are given by $p_{1,2} = p_{1,3} = 0.5$, $p_{2,4} = p_{2,5} = 0.5$, $p_{3,6} = p_{3,7} = 0.5$, $p_{4,8} = 0.5$, $p_{4,9} = 0.5$, $p_{5,4} = 0.5h_1$, $p_{5,8} = 0.5(1-h_1)+0.5h_2$, $p_{5,9} = 0.5(1-h_2)$, $p_{6,7} = 0.5h_3$, $p_{6,10} = 0.5(1-h_3)$, $p_{6,11} = 0.5$, $p_{7,10} = 0.5$, $p_{7,11} = 0.5$. Other transition probabilities are 0.

We first calculate the safety function just knowing the transition probabilities from the set U' to the forbidden set U and assume that we do not have access to the transition probabilities of the set $\mathcal{X} \setminus U'$. We assume that we know the transition probabilities from the set $U' \setminus U$ to the forbidden set U in order to verify our learning algorithm as these transition probabilities are the rewards. Thus, our problem resembles that of standard reinforcement learning-based value function evaluation methods. However, in practice, one might only have bounds on these transition probabilities instead of the true value. The number of iterations within each episode is $\mathcal{T} = 3$ ($\mathcal{T}$ has to be ≥ 3 for the example given in Fig. 2). For all $x \in \mathcal{X}$, we use a learning rate $\alpha_t(x) = 0.001$ for $t < 10000$ and $\alpha_t(x) = \frac{1}{t}$ for ≥ 10000. We set $h_1 = 0.4$, $h_2 = 0.6$ and $h_3 = 0.5$. Thus $S_\pi(4) = 0.5$, $S_\pi(6) = 0.3$, $S_\pi(7) = 0.625$ and $S_\pi(8) = 0.5$. We learn $S_\pi(1)$, $S_\pi(2)$ and $S_\pi(3)$ using TD(0). Moreover, we calculate the safety function using the model-based approach given in [11] and compare that with the ones learned using Algorithm 1. Table I below presents the results. From the table, it is evident that as the number of episodes increased, the estimated value of the safety function for each state moved closer to the true value of the safety function.

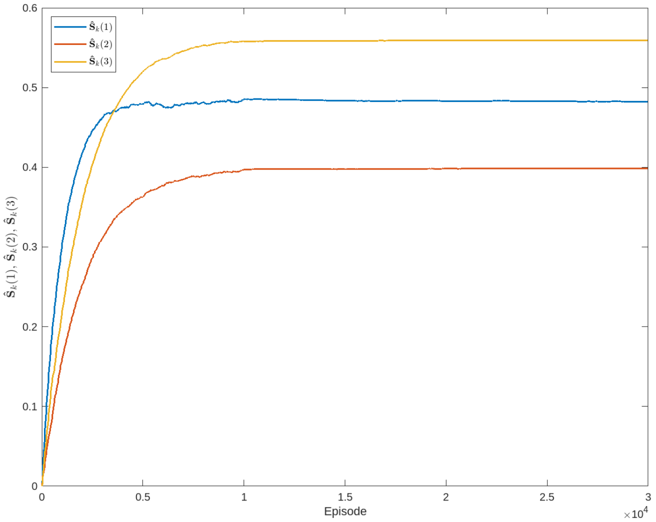

Fig. 3. Convergence of the safety functions with $\mathcal{T} = 30000$ number of episodes

State (x)	$S_\pi(x)$ using [11]	$\hat{S}_{\mathcal{K}}(x)$ using Algorithm 1 ($\mathcal{K} = 15000$)	$\hat{S}_{\mathcal{K}}(x)$ using Algorithm 1 ($\mathcal{K} = 30000$)	$\hat{S}_{\mathcal{K}}(x)$ using Algorithm 1 ($\mathcal{K} = 50000$)
1	0.4813	0.4809	0.4816	0.4816
2	0.4	0.3992	0.4003	0.4000
3	0.5625	0.5585	0.5593	0.5609

TABLE I

SAFETY FUNCTION WITH DIFFERENT NUMBER OF EPISODES.

Then we consider the case when we do not have any exact knowledge about the transition probabilities instead we have a belief about the bounds on the transition probabilities from the proxy set to the forbidden set. For that, we assume the reward as $r_t = w(X_t)$ for the lower bound and $r_t = q(X_t)$ for the upper bound of the safety function. In this case, unlike the previous case, we assume that we do not have any knowledge about the true transition probabilities even from the proxy set to the forbidden set. Instead, we have bounds for $\mathbb{P}^x_\pi[\tau_U < \tau_E]$ for all $x \in U' \setminus U$. With $h_1 = 0.4$, $h_2 = 0.6$, $h_3 = 0.5$, true $\mathbb{P}^x_\pi[\tau_U < \tau_E]$s are: $\mathbb{P}^4_\pi[\tau_U < \tau_E] = 0.5$, $\mathbb{P}^5_\pi[\tau_U < \tau_E] = 0.3$, $\mathbb{P}^6_\pi[\tau_U < \tau_E] = 0.625$ and $\mathbb{P}^7_\pi[\tau_U < \tau_E] = 0.5$. Upper bound for $\mathbb{P}^x_\pi[\tau_U < \tau_E]$ is considered to be 0.7, 0.5, 0.8 and 0.7 for state 4, 5, 6 and 7, respectively. Similarly the lower bound for $\mathbb{P}^x_\pi[\tau_U < \tau_E]$ is assumed to be 0.3, 0.15, 0.5 and 0.35 for state 4, 5, 6 and 7, respectively.

VI. CONCLUSION AND FUTURE WORK

In this paper, we considered the problem of estimating the safety function for an MDP only knowing the bounds on the transition probabilities to the forbidden states. We characterized a *proxy set* visiting which we learn bounds on safety function. Our algorithm is based on the one-step temporal difference method TD(0).

In this work, we focused on estimating the safety function of an MDP with a given policy. In our future work, we shall

State (x)	$S_\pi(x)$ using [11]	$\hat{S}_{\mathcal{K}}(x)$ using Algorithm 1 ($\mathcal{K} = 50000$)	Estimated Upper bound of $S_\pi(x)$ with $\mathcal{K} = 50000$	Estimated lower bound of $S_\pi(x)$ with $\mathcal{K} = 50000$
1	0.4813	0.4816	0.6740	0.3253
2	0.4	0.4	0.5967	0.2242
3	0.5625	0.5609	0.7475	0.4248

TABLE II

Estimated bounds of the Safety function.

consider calculating the safety function and designing safe optimal policies side by side.

References

[1] R. Wisniewski and L.-M. Bujorianu, "Safety of stochastic systems: An analytic and computational approach," *Automatica*, vol. 133, p. 109839, 2021.

[2] M. L. Bujorianu, R. Wisniewski, and E. Boulougouris, "Stochastic safety for random dynamical systems," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 1340–1345.

[3] ——, "p-safety and stability," *IFAC-PapersOnLine*, vol. 54, no. 9, pp. 665–670, 2021.

[4] ——, "Stochastic safety for markov chains," *IEEE Control Systems Letters*, vol. 5, no. 2, pp. 427–432, 2020.

[5] M. Pecka and T. Svoboda, "Safe exploration techniques for reinforcement learning – an overview," in *Modelling and Simulation for Autonomous Systems*, J. Hodicky, Ed. Cham: Springer International Publishing, 2014, pp. 357–375.

[6] A. Lavaei, A. Nejati, P. Jagtap, and M. Zamani, "Formal safety verification of unknown continuous-time systems: a data-driven approach," in *Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control*, 2021, pp. 1–2.

[7] A. Salamati, A. Lavaei, S. Soudjani, and M. Zamani, "Data-driven safety verification of stochastic systems via barrier certificates," *IFAC-PapersOnLine*, vol. 54, no. 5, pp. 7–12, 2021.

[8] N. Noroozi, A. Salamati, and M. Zamani, "Data-driven safety verification of discrete-time networks: A compositional approach," *IEEE Control Systems Letters*, vol. 6, pp. 2210–2215, 2021.

[9] A. Salamati and M. Zamani, "Data-driven safety verification of stochastic systems via barrier certificates: A wait-and-judge approach," in *Learning for Dynamics and Control Conference*. PMLR, 2022, pp. 441–452.

[10] ——, "Safety verification of stochastic systems: A repetitive scenario approach," *IEEE Control Systems Letters*, vol. 7, pp. 448–453, 2022.

[11] R. Wisniewski and M. L. Bujorianu, "Probabilistic Safety Guarantees for Markov Decision Processes," *Submitted to IEEE Transactions on Automatic Control*, 2022.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[13] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvari, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine Learning*, vol. 38, pp. 287–308, 2000.