

Policy Gradients for Probabilistic Constrained Reinforcement Learning

Wei Qin Chen
Department of Electrical,
Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY, USA
chenw18@rpi.edu

Dharmashankar Subramanian
IBM T. J. Watson Research Center
Yorktown Heights, NY, USA
dharmash@us.ibm.com

Santiago Paternain
Department of Electrical,
Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY, USA
paters@rpi.edu

Abstract—This paper considers the problem of learning safe policies in the context of reinforcement learning (RL). In particular, we consider the notion of probabilistic safety. This is, we aim to design policies that maintain the state of the system in a safe set with high probability. This notion differs from cumulative constraints often considered in the literature. The challenge of working with probabilistic safety is the lack of expressions for their gradients. Indeed, policy optimization algorithms rely on gradients of the objective function and the constraints. To the best of our knowledge, this work is the first one providing such explicit gradient expressions for probabilistic constraints. It is worth noting that the gradient of this family of constraints can be applied to various policy-based algorithms. We demonstrate empirically that it is possible to handle probabilistic constraints in a continuous navigation problem.

Index Terms—reinforcement learning, probabilistic constraint, safe policy, policy gradient

I. INTRODUCTION

Reinforcement learning (RL) has gained traction as a solution to the problem of computing policies to perform challenging and high-dimensional tasks, e.g., playing video games [1], mastering Go [2], robotic manipulation [3] and locomotion [4], etc. However, in general, RL algorithms are only concerned with maximizing a cumulative reward [5], [6], which may lead to risky behaviors [7] in realistic domains such as robot navigation [8].

Taking into account the safety requirements motivates the development of policy optimization under safety guarantees [9]–[11]. Some approaches consider risk-aware objectives or regularized solutions where the reward is modified to take into account the safety requirements [12]–[14]. Other formulations include Constrained Markov Decision Processes (CMDPs) [15] where additional cumulative (or average) rewards need to be kept above a desired threshold. This approach has been commonly used to induce safe behaviors [16]–[24]. To solve these constrained problems, primal-dual algorithms [16]–[20] combined with classical and state-of-the-art policy-based algorithms, e.g., REINFORCE [25], DDPG [26], TRPO [27], PPO [28] are generally used.

In this cumulative constraint setting, safety violations are acceptable as long as the amount of violations does not exceed

the desired thresholds. This makes them often not suitable for safety-critical applications. For instance, in the context of autonomous driving, even one collision is unacceptable.

A more suitable notion of safety in this context is to guarantee that the whole trajectory of the system remains within a set that is deemed to be safe (see e.g., [29]). Ideally, one would like to achieve this goal for every possible trajectory. This being an ambitious goal, in this work we settle for solutions that guarantee every time safety with high probability. We describe this setting in detail in Section II.

The main challenge in solving problems under probabilistic safety constraints is that explicit policy gradient-like expressions for such constraints are not readily available. Indeed in [20], [30] replace this constraint with a suitable lower bound. In [31] an approximation of the gradient is also provided. These limitations, prevent us from running the aforementioned policy-based algorithms. In Section III, we present the main contribution of this work: an expression for the gradient that enables stochastic approximations. Other than concluding remarks, the paper finishes (Section IV) with numerical examples that demonstrate the use of the probabilistic safe gradient to train safe policies in a navigation task.

II. PROBLEM FORMULATION

In this work, we consider the problem of finding optimal policies for Markov Decision Processes (MDPs) under probabilistic safety guarantees. In particular, we are interested in situations where the state transition distributions are unknown and thus the policies need to be computed from data. An MDP [32] is defined by a tuple $(\mathcal{S}, \mathcal{A}, r, \mathbb{P}, \mu, T)$, where \mathcal{S} is the state space, \mathcal{A} is the action space, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function describing the quality of the decision, $\mathbb{P}_{s_t \rightarrow s_{t+1}}^{a_t}(\hat{\mathcal{S}}) := \mathbb{P}(s_{t+1} \in \hat{\mathcal{S}} \mid s_t, a_t)$ where $\hat{\mathcal{S}} \subset \mathcal{S}$, $s_t \in \mathcal{S}$, $a_t \in \mathcal{A}$, $t \in \mathbb{N}$ is the transition probability describing the dynamics of the system, $\mu(\hat{\mathcal{S}}) := \mathbb{P}(s \in \hat{\mathcal{S}})$ is the initial state distribution, and T is the time horizon. The state and action at time t are random variables denoted respectively by S_t and A_t . A *policy* is a conditional distribution $\pi_\theta(a|s)$ parameterized by $\theta \in \mathbb{R}^d$ (for instance the weights and biases of neural networks), from which the agent draws action $a \in \mathcal{A}$ when in the corresponding state $s \in \mathcal{S}$. In the context of MDPs the

objective is to find a policy that maximizes the value function. The latter is defined as

$$V(\theta) = \mathbb{E}_{\mathbf{a} \sim \pi_\theta(\mathbf{a}|\mathbf{s}), S_0 \sim \mu} \left[\sum_{t=0}^T r(S_t, A_t) \right], \quad (1)$$

where \mathbf{a} and \mathbf{s} denote the sequences of actions and states for the whole episode, this is, from time $t = 0$ to $t = T$. The subscripts of this expectation are omitted in the remaining of the paper for simplicity.

In this paper, we are concerned with imposing safety requirements. In particular, we focus on the notion of probabilistic safety which we formally define next.

Definition 1. A policy π_θ is $(1 - \delta)$ -safe for the set $\mathcal{S}_{\text{safe}} \subset \mathcal{S}$ if and only if $\mathbb{P} \left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta \right) \geq 1 - \delta$.

With this definition, we formulate the following probabilistic safe RL problem as a constrained optimization problem

$$\begin{aligned} P^* &= \max_{\theta \in \mathbb{R}^d} V(\theta) \\ \text{s.t. } &\mathbb{P} \left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta \right) \geq 1 - \delta. \end{aligned} \quad (2)$$

Note that the previous problem differs from most of the safe RL literature that work with the cumulative constraint setting (see e.g., [19]–[21], [24]). To solve problem (2), it is conceivable to employ gradient-based methods e.g., regularization [33] or primal-dual [34] to achieve local optimal solutions. For instance, consider the regularization method with a fixed penalty. This is, for $\lambda > 0$ we formulate the following *unconstrained* problem as an approximation to the *constrained* problem (2)

$$\mathbb{E} \left[\sum_{t=0}^T r(S_t, A_t) \right] + \lambda \left(\mathbb{P} \left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta \right) - (1 - \delta) \right). \quad (3)$$

Note that λ trades-off safety and task performance. Indeed, for large values of λ solutions to (3) will prioritize safe behaviors, whereas for small values of λ the solutions will focus on maximizing the expected cumulative rewards.

Then, to solve problem (3) locally, gradient ascent [35] or its stochastic versions are generally employed. Note that the gradient of the first term in (3) can be computed using the Policy Gradient Theorem [6]. Nevertheless, the lack of an expression for the gradient of the probabilistic safety constraint, i.e., $\nabla_\theta \mathbb{P} \left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta \right)$ prevents us from applying the gradient ascent family of methods to solve (3). In the next section, we provide such an expression for the gradient that allows us to overcome this limitation.

III. THE GRADIENT OF PROBABILISTIC CONSTRAINTS

We start this section by defining an important quantity in what follows next. For any t such that $0 \leq t \leq T$ define

$$G_t = \prod_{u=t}^T \mathbb{1}(S_u \in \mathcal{S}_{\text{safe}}). \quad (4)$$

Having defined this quantity we are now in conditions of providing an expression for the gradient of the probabilistic constraint. This is the subject of the following Theorem.

Theorem 1. Let $S_0 \in \mathcal{S}_{\text{safe}}$, the gradient of the probability of being safe for a given policy π_θ yields

$$\begin{aligned} &\nabla_\theta \mathbb{P} \left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta, S_0 \right) \\ &= \mathbb{E} \left[\sum_{t=0}^{T-1} G_1 \nabla_\theta \log \pi_\theta(A_t | S_t) | \pi_\theta, S_0 \right]. \end{aligned} \quad (5)$$

Proof. We proceed by presenting and proving the following two technical lemmas (Lemma 1 and Lemma 2).

Lemma 1. Given $S_{t-1} \in \mathcal{S}_{\text{safe}}$ and $G_t, t = 1, 2, \dots, T-1$ defined in (4), it holds that

$$\begin{aligned} \nabla_\theta \mathbb{E}[G_t | S_{t-1}] &= \mathbb{E}[\nabla_\theta \mathbb{E}[G_{t+1} | S_t] \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}) | S_{t-1}] \\ &\quad + \mathbb{E}[G_t \nabla_\theta \log \pi_\theta(A_{t-1} | S_{t-1}) | S_{t-1}]. \end{aligned} \quad (6)$$

Proof. We start the proof by rewriting the expectation of G_1 using the towering property of the expectation

$$\begin{aligned} \mathbb{E}[G_1 | S_0] &= \mathbb{E}[\mathbb{E}[G_1 | S_1] | S_0] \\ &= \mathbb{E}[\mathbb{E}[G_2 \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_1] | S_0], \end{aligned} \quad (7)$$

where the second equality follows from (4). Since S_1 is measurable with respect to the σ -algebra \mathcal{F}_1 it follows that

$$\mathbb{E}[G_1 | S_0] = \mathbb{E}[\mathbb{E}[G_2 | S_1] \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_0]. \quad (8)$$

Rewriting the outer expectation in terms of the probability distribution of S_1 , the previous expression reduces to

$$\begin{aligned} \mathbb{E}[G_1 | S_0] &= \int_{\mathcal{S}} \mathbb{E}[G_2 | s_1] \mathbb{1}(s_1 \in \mathcal{S}_{\text{safe}}) p(s_1 | S_0) ds_1 \\ &= \int_{\mathcal{S} \times \mathcal{A}} \mathbb{E}[G_2 | s_1] \mathbb{1}(s_1 \in \mathcal{S}_{\text{safe}}) \\ &\quad p(s_1 | S_0, a_0) \pi_\theta(a_0 | S_0) ds_1 da_0, \end{aligned} \quad (9)$$

where the last equality follows from $p(s_1 | S_0) = \int_{\mathcal{A}} p(s_1 | S_0, a_0) \pi_\theta(a_0 | S_0) da_0$. Taking the gradient of the previous expression with respect to the policy parameters θ results in

$$\begin{aligned} \nabla_\theta \mathbb{E}[G_1 | S_0] &= \int_{\mathcal{S} \times \mathcal{A}} \nabla_\theta (\mathbb{E}[G_2 | s_1]) \mathbb{1}(s_1 \in \mathcal{S}_{\text{safe}}) \\ &\quad p(s_1 | S_0, a_0) \pi_\theta(a_0 | S_0) ds_1 da_0 \\ &\quad + \int_{\mathcal{S} \times \mathcal{A}} \mathbb{E}[G_2 | s_1] \mathbb{1}(s_1 \in \mathcal{S}_{\text{safe}}) \\ &\quad p(s_1 | S_0, a_0) \nabla_\theta \pi_\theta(a_0 | S_0) ds_1 da_0. \end{aligned} \quad (10)$$

Notice that the first in the right-hand side of the previous expression can be presented by

$$\begin{aligned} &\int_{\mathcal{S} \times \mathcal{A}} \nabla_\theta (\mathbb{E}[G_2 | s_1]) \mathbb{1}(s_1 \in \mathcal{S}_{\text{safe}}) \\ &\quad p(s_1 | S_0, a_0) \pi_\theta(a_0 | S_0) ds_1 da_0 \\ &= \mathbb{E}[\nabla_\theta \mathbb{E}[G_2 | S_1] \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_0]. \end{aligned} \quad (11)$$

The second term using the “log-trick” yields

$$\begin{aligned} & \int_{\mathcal{S} \times \mathcal{A}} \mathbb{E}[G_2 | s_1] \mathbb{1}(s_1 \in \mathcal{S}_{\text{safe}}) \\ & \quad p(s_1 | S_0, a_0) \nabla_{\theta} \pi_{\theta}(a_0 | S_0) ds_1 da_0 \\ &= \int_{\mathcal{S} \times \mathcal{A}} \mathbb{E}[G_2 | s_1] \mathbb{1}(s_1 \in \mathcal{S}_{\text{safe}}) p(s_1 | S_0, a_0) \\ & \quad \pi_{\theta}(a_0 | S_0) \nabla_{\theta} \log \pi_{\theta}(a_0 | S_0) ds_1 da_0. \end{aligned} \quad (12)$$

Likewise, since s_1 is measurable with respect to the σ -algebra \mathcal{F}_1 , (12) can be simplified as $\int_{\mathcal{S} \times \mathcal{A}} \mathbb{E}[G_1 | s_1] p(s_1 | S_0, a_0) \pi_{\theta}(a_0 | S_0) \nabla_{\theta} \log \pi_{\theta}(a_0 | S_0) ds_1 da_0$, which is also equivalent to $\mathbb{E}[\mathbb{E}[G_1 | S_1] \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0]$. Since $\log \pi_{\theta}(A_0 | S_0)$ is measurable given S_1 , we have

$$\begin{aligned} & \mathbb{E}[\mathbb{E}[G_1 | S_1] \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0] \\ &= \mathbb{E}[\mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_1] | S_0]. \end{aligned} \quad (13)$$

Using the towering property of the expectation the previous expressions yield

$$\begin{aligned} & \int_{\mathcal{S} \times \mathcal{A}} \mathbb{E}[G_2 | s_1] \mathbb{1}(s_1 \in \mathcal{S}_{\text{safe}}) p(s_1 | S_0, a_0) \\ & \quad \nabla_{\theta} \pi_{\theta}(a_0 | S_0) ds_1 da_0 \\ &= \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0]. \end{aligned} \quad (14)$$

Then, combining (11) with (14) yields

$$\begin{aligned} \nabla_{\theta} \mathbb{E}[G_1 | S_0] &= \mathbb{E}[\nabla_{\theta} \mathbb{E}[G_2 | S_1] \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_0] \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0]. \end{aligned} \quad (15)$$

Repeating the process above i times for $1 \leq i \leq T-1$, we obtain the following recursive definition of the gradient of the probability in (2) with respect to θ

$$\begin{aligned} \nabla_{\theta} \mathbb{E}[G_i | S_{i-1}] &= \mathbb{E}[\nabla_{\theta} \mathbb{E}[G_{i+1} | S_i] \mathbb{1}(S_i \in \mathcal{S}_{\text{safe}}) | S_{i-1}] \\ & \quad + \mathbb{E}[G_i \nabla_{\theta} \log \pi_{\theta}(A_{i-1} | S_{i-1}) | S_{i-1}]. \end{aligned} \quad (16)$$

This completes the proof of Lemma 1. \square

Lemma 2. Given $S_{t-1} \in \mathcal{S}_{\text{safe}}$ and $G_t, t = 1, 2, \dots, T-1$ defined in (4), it holds that

$$\begin{aligned} \nabla_{\theta} \mathbb{E}[G_1 | S_0] &= \sum_{t=0}^{T-2} \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_t | S_t) | S_0] \\ & \quad + \mathbb{E}\left[\nabla_{\theta} \mathbb{E}[G_T | S_{T-1}] \prod_{t=1}^{T-1} \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}) | S_0\right]. \end{aligned} \quad (17)$$

Proof. We proceed by employing Lemma 1 to derive the gradient of the expectation of G_1 and G_2 , respectively

$$\begin{aligned} \nabla_{\theta} \mathbb{E}[G_1 | S_0] &= \mathbb{E}[\nabla_{\theta} \mathbb{E}[G_2 | S_1] \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_0] \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0]. \end{aligned} \quad (18)$$

$$\begin{aligned} \nabla_{\theta} \mathbb{E}[G_2 | S_1] &= \mathbb{E}[\nabla_{\theta} \mathbb{E}[G_3 | S_2] \mathbb{1}(S_2 \in \mathcal{S}_{\text{safe}}) | S_1] \\ & \quad + \mathbb{E}[G_2 \nabla_{\theta} \log \pi_{\theta}(A_1 | S_1) | S_1]. \end{aligned} \quad (19)$$

Then, substituting (19) into (18) yields

$$\begin{aligned} & \nabla_{\theta} \mathbb{E}[G_1 | S_0] \\ &= \mathbb{E}[\mathbb{E}[\nabla_{\theta} \mathbb{E}[G_3 | S_2] \mathbb{1}(S_2 \in \mathcal{S}_{\text{safe}}) | S_1] \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_0] \\ & \quad + \mathbb{E}[G_2 \nabla_{\theta} \log \pi_{\theta}(A_1 | S_1) | S_1] \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_0] \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0]. \end{aligned} \quad (20)$$

As $\mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}})$ is measurable given S_1 , we have

$$\begin{aligned} & \nabla_{\theta} \mathbb{E}[G_1 | S_0] \\ &= \mathbb{E}[\mathbb{E}[\nabla_{\theta} \mathbb{E}[G_3 | S_2] \mathbb{1}(S_2 \in \mathcal{S}_{\text{safe}}) \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_1] \\ & \quad + \mathbb{E}[G_2 \nabla_{\theta} \log \pi_{\theta}(A_1 | S_1) \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_1] | S_0] \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0]. \end{aligned} \quad (21)$$

By definition of G_1 we can simplify the second term of the right-hand side of the previous equation,

$$\begin{aligned} & \nabla_{\theta} \mathbb{E}[G_1 | S_0] \\ &= \mathbb{E}[\mathbb{E}[\nabla_{\theta} \mathbb{E}[G_3 | S_2] \mathbb{1}(S_2 \in \mathcal{S}_{\text{safe}}) \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_1] \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_1 | S_1) | S_1] | S_0] \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0]. \end{aligned} \quad (22)$$

Using the towering property of expectation, (22) reduces to

$$\begin{aligned} & \nabla_{\theta} \mathbb{E}[G_1 | S_0] \\ &= \mathbb{E}[\nabla_{\theta} \mathbb{E}[G_3 | S_2] \mathbb{1}(S_2 \in \mathcal{S}_{\text{safe}}) \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_0] \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_1 | S_1) | S_0] \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0]. \end{aligned} \quad (23)$$

Then repeatedly unwrapping $\nabla_{\theta} \mathbb{E}[G_1 | S_0]$ in terms of G_3, \dots, G_T by Lemma 1 yields

$$\begin{aligned} & \nabla_{\theta} \mathbb{E}[G_1 | S_0] \\ &= \mathbb{E}[\nabla_{\theta} \mathbb{E}[G_T | S_{T-1}] \mathbb{1}(S_{T-1} \in \mathcal{S}_{\text{safe}}) \\ & \quad \dots \mathbb{1}(S_2 \in \mathcal{S}_{\text{safe}}) \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) | S_0] \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_{T-2} | S_{T-2}) | S_0] + \dots \\ & \quad + \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_0 | S_0) | S_0] \\ &= \sum_{t=0}^{T-2} \mathbb{E}[G_1 \nabla_{\theta} \log \pi_{\theta}(A_t | S_t) | S_0] \\ & \quad + \mathbb{E}\left[\nabla_{\theta} \mathbb{E}[G_T | S_{T-1}] \prod_{t=1}^{T-1} \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}) | S_0\right]. \end{aligned} \quad (24)$$

This completes the proof of Lemma 2. \square

We are now in conditions to prove Theorem 1. We start by rewriting the probability of remaining safe in terms of G_0 defined in (4). By definition of probability we have

$$\begin{aligned} & \mathbb{P}\left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_{\theta}, S_0\right) \\ &= \mathbb{E}\left[\mathbb{1}\left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\}\right) | \pi_{\theta}, S_0\right]. \end{aligned} \quad (25)$$

Note that the indicator function in the previous expression takes the value one, if and only if each $S_t \in \mathcal{S}_{\text{safe}}$. Hence, it

is possible to rewrite the previous expression in terms of the product of indicator functions of states satisfying the safety condition at each time

$$\mathbb{P}\left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta, S_0\right) = \mathbb{E}\left[\prod_{t=0}^T \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}) | \pi_\theta, S_0\right] = \mathbb{E}[G_0 | S_0], \quad (26)$$

where π_θ is omitted in the last equation for simplicity. By virtue of $S_0 \in \mathcal{S}_{\text{safe}}$, we obtain $\mathbb{E}[G_0 | S_0] = \mathbb{E}[G_1 \cdot \mathbb{1}(S_0 \in \mathcal{S}_{\text{safe}}) | S_0] = \mathbb{E}[G_1 | S_0]$. Then, using (26), the gradient of the probability of remaining safe reduces to

$$\nabla_\theta \mathbb{P}\left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta, S_0\right) = \nabla_\theta \mathbb{E}[G_1 | S_0]. \quad (27)$$

In Lemma 1 we derive a recursive relationship for the gradient of $\mathbb{E}[G_t | S_{t-1}]$, $t = 1, 2, \dots, T-1$. By virtue of Lemma 2, to complete the proof of the result it suffices to establish that

$$\mathbb{E}\left[\nabla_\theta \mathbb{E}[G_T | S_{T-1}] \prod_{t=1}^{T-1} \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}) | S_0\right] = \mathbb{E}[G_1 \nabla_\theta \log \pi_\theta(A_{T-1} | S_{T-1}) | S_0]. \quad (28)$$

We establish this result next. Let us start by working with the gradient of the inner expectation on the left-hand side of the previous expression.

Using the fact that $G_T = \mathbb{1}(S_T \in \mathcal{S}_{\text{safe}})$ and the definition of expectation one can write $\nabla_\theta \mathbb{E}[G_T | S_{T-1}]$ in the left-hand side of the previous expression as

$$\nabla_\theta \mathbb{E}[G_T | S_{T-1}] = \nabla_\theta \int_{\mathcal{S}} \mathbb{1}(s_T \in \mathcal{S}_{\text{safe}}) p(s_T | S_{T-1}) ds_T, \quad (29)$$

where $p(s_T | S_{T-1})$ denotes the conditional probability of S_T given S_{T-1} . Marginalizing the probability distribution it follows that

$$p(s_T | S_{T-1}) = \int_{\mathcal{A}} p(s_T | S_{T-1}, a_{T-1}) \pi_\theta(a_{T-1} | S_{T-1}) da_{T-1}. \quad (30)$$

Consequently, (29) can be converted to

$$\begin{aligned} \nabla_\theta \mathbb{E}[G_T | S_{T-1}] &= \nabla_\theta \int_{\mathcal{S} \times \mathcal{A}} \mathbb{1}(s_T \in \mathcal{S}_{\text{safe}}) p(s_T | S_{T-1}, a_{T-1}) \\ &\quad \pi_\theta(a_{T-1} | S_{T-1}) ds_T da_{T-1}. \end{aligned} \quad (31)$$

Note that in the previous expression, the only term dependent on θ is the policy, hence we have that

$$\nabla_\theta \mathbb{E}[G_T | S_{T-1}] = \int_{\mathcal{S} \times \mathcal{A}} \mathbb{1}(s_T \in \mathcal{S}_{\text{safe}}) p(s_T | S_{T-1}, a_{T-1}) \nabla_\theta \pi_\theta(a_{T-1} | S_{T-1}) ds_T da_{T-1}. \quad (32)$$

Applying the “log-trick” to the right-hand side of (32) yields

$$\begin{aligned} \nabla_\theta \mathbb{E}[G_T | S_{T-1}] &= \int_{\mathcal{S} \times \mathcal{A}} \mathbb{1}(s_T \in \mathcal{S}_{\text{safe}}) p(s_T | S_{T-1}, a_{T-1}) \\ &\quad \pi_\theta(a_{T-1} | S_{T-1}) \nabla_\theta \log \pi_\theta(a_{T-1} | S_{T-1}) ds_T da_{T-1}. \end{aligned} \quad (33)$$

Since $p(s_T | S_{T-1}, a_{T-1}) \pi_\theta(a_{T-1} | S_{T-1}) = p(s_T, a_{T-1} | S_{T-1})$ is the joint probability distribution of S_T and A_{T-1} given S_{T-1} the previous expression reduces to

$$\nabla_\theta \mathbb{E}[G_T | S_{T-1}] = \mathbb{E}[G_T \nabla_\theta \log \pi_\theta(A_{T-1} | S_{T-1}) | S_{T-1}]. \quad (34)$$

Since S_1, \dots, S_{T-1} are measurable with respect to S_{T-1} it follows that

$$\begin{aligned} \nabla_\theta \mathbb{E}[G_T | S_{T-1}] &\prod_{t=1}^{T-1} \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}) \\ &= \mathbb{E}[G_1 \nabla_\theta \log \pi_\theta(A_{T-1} | S_{T-1}) | S_{T-1}], \end{aligned} \quad (35)$$

where we have used that $G_1 = G_T \prod_{t=1}^{T-1} \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}})$. Substituting the previous expression in the left-hand side of (28) it follows that

$$\begin{aligned} \mathbb{E}\left[\nabla_\theta \mathbb{E}[G_T | S_{T-1}] \prod_{t=1}^{T-1} \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}) | S_0\right] &= \mathbb{E}[\mathbb{E}[G_1 \nabla_\theta \log \pi_\theta(A_{T-1} | S_{T-1}) | S_{T-1}] | S_0]. \end{aligned} \quad (36)$$

The law of total expectation completes the result claimed in (28) and therefore completes the proof of Theorem 1. \square

The expression for the gradient in Theorem 1, allows us to directly tackle safe RL problems that take the form of (2) using policy optimization techniques. It is worth pointing out that the proof of Theorem 1 is similar to policy gradient theorems in the literature [6], [25]. Although promising, stochastic approximations of the gradient introduce challenges. Unlike the classic policy gradient (where policy parameters update each iteration), under this framework, the parameters in (5) only update when every step of the trajectory is safe, i.e., when $G_1 = \prod_{t=1}^T \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}) = 1$. Addressing this issue is beyond the scope of this work and opens up several interesting future research avenues. Despite this limitation, we demonstrate in the next section that the estimate proposed can solve problems of the form (2).

IV. NUMERICAL EXPERIMENTS

To illustrate the ability of using (5) to train safe policies, we consider a continuous navigation task in an environment populated with hazardous obstacles (see Figure 1). The coordinates of the obstacles' centers are (7, 7), (3, 7), (1.5, 4), (4.5, 3), (8, 3) with the corresponding radii $\{2, 1, 0.5, 1.5, 0.75\}$. The state in this example is the position of the agent on the x - and y -axis, namely, $s = (x, y)$. We set the continuous state space as $\mathcal{S} = [0, 10] \times [0, 10]$. The goal of the agent is to reach a goal position $s_{\text{goal}} = (9, 1.5)$ within the time horizon

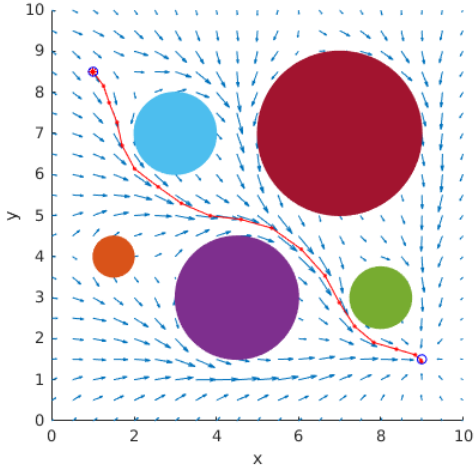


Fig. 1. Navigation policy learned after 40,000 iterations for probabilistic constraint formulation selecting $\lambda = 6$, $\eta = 0.002$. The agent is trained to navigate starting from (1, 8.5) to a goal (9, 1.5).

$T = 20$, while avoiding 5 obstacles. Accordingly, the safe set is defined as the whole map/state space excluding regions of 5 obstacles.

The agent's action a is a two-dimensional velocity. For a given state and action at time t , the state evolves according to $s_{t+1} = s_t + a_t T_s$ with $T_s = 0.05$. The policy of the agent is a multivariate Gaussian distribution

$$\pi_\theta(a|s) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(a - \mu_\theta(s))^\top \Sigma^{-1}(a - \mu_\theta(s))\right), \quad (37)$$

where $\mu_\theta(s)$ and Σ denote the mean and covariance matrix of the Gaussian policy. We parameterize $\mu_\theta(s)$ as a linear combination of Radial Basis Functions (RBFs)

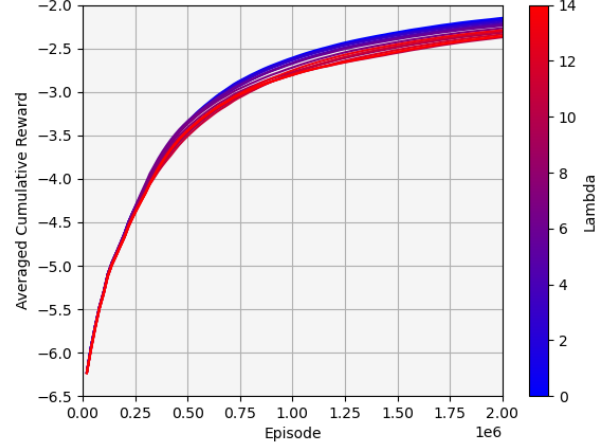
$$\mu_\theta(s) = \sum_{k=1}^d \theta_k \exp\left(-\frac{\|s - \bar{s}_k\|^2}{2\sigma^2}\right), \quad (38)$$

where $\theta = [\theta_1, \theta_2, \dots, \theta_d]^\top$ are parameters that need to be learned, \bar{s}_k are centers of each RBFs kernel and σ their bandwidth. In this experiment we set $\Sigma = \text{diag}(0.5, 0.5)$, $\sigma = 0.5$, $d = 1681$ and $\bar{s}_k = (x_k, y_k)$, $k = 1, 2, \dots, 1681$ where \bar{s}_k forms a uniform lattice with separation 0.25 in each direction. The reward is the negative squared distance to the goal position s_{goal} , i.e., $r(s_t, a_t) = -\|s_t - s_{goal}\|^2$.

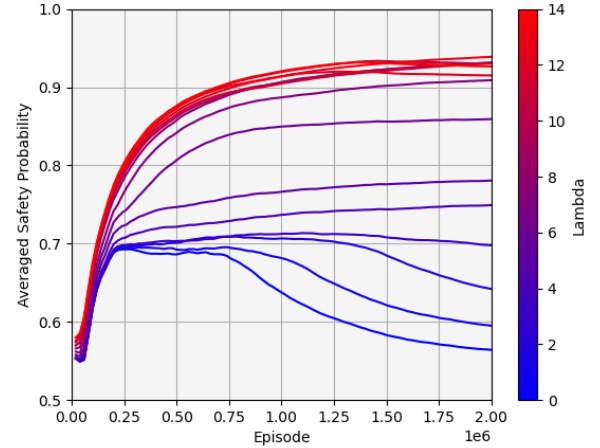
To solve problem (3) in this set-up, we consider a stochastic approximation of the gradient ascent, which yields the following update rule for the parameters θ of the policy

$$\theta_{k+1} = \theta_k + \eta \left(\hat{\nabla}_\theta V(\theta_k) + \lambda \hat{\nabla}_\theta \mathbb{P} \left(\bigcap_{t=0}^T \{S_t \in \mathcal{S}_{\text{safe}}\} \right) \right), \quad (39)$$

where the first term in bracket on the right-hand side is computed by a stochastic approximation of the Policy Gradient



(a) Evolution of the averaged cumulative reward



(b) Evolution of the averaged safety probability

Fig. 2. Evolution of the averaged cumulative reward and averaged safety probability as the algorithm iterates. The step-size η and time horizon T are fixed to 0.0006 and 2000000. The color bar represents that λ is selected from $[0.5, 14]$ in which red and blue denote large and small values of λ respectively.

Theorem [6] and the second term is a stochastic approximation of (5).

Figure 1 demonstrates that the agent with probabilistic safety constraints is trained to safely navigate to the goal position (9, 1.5) from the initial state (1, 8.5) after 40,000 episodes of training, during which λ is fixed to be 6 and with $\eta = 0.002$.

Note that to attain different levels of safety, in general, different values of λ are required. Subsequently, we run (39) with different λ to find solutions to problem (3). The worst-case scenario requires $\eta = 0.0006$ and 2,000,000 episodes. As depicted in Figure 2, the color bar represents that λ is chosen from $[0.5, 14]$ in which red and blue denote large and small values of λ respectively. To quantitatively analyze the effect of λ on safety and task performance, the safety probability is defined as the fraction of safe episodes over 1000

independent episodes under different random seeds and then averaged across the evaluation episodes. Similarly, we estimate the value function by averaging the cumulative reward across the evaluation episodes. It is not surprising that larger λ yields larger safety probability and lower cumulative rewards.

We thus demonstrated that stochastic approximations of the gradient of the probabilistic constraints (Theorem 1) can successfully be employed for solving the task at hand. Notice that the algorithms used in this numerical section are Monte Carlo methods [32, Chapter 5]. In the RL literature there exist algorithms that exploit temporal differences [5] and/or trust regions [27] which result in faster convergence rates. As mentioned in Section III the estimation of the gradient of the probabilistic constraint is zero for every unsafe episode, thus hindering the rate of convergence. Analyzing alternatives to overcome this issue is out of the scope of this work.

V. CONCLUSIONS

In this work, we considered the problem of learning probabilistic safe policies. Unlike cumulative constraints often considered in the literature, we aim to guarantee that the state of the agent remains in the safe set with high probability.

We have provided the first expression for the gradient of a probabilistic constraint safety requirement, thus enabling the application of Policy Optimization methods in these settings as well. We have also demonstrated that updates based on this gradient can be used to solve continuous navigation problems in cluttered environments. The stochastic approximation presented is not without issues. In particular, the gradient estimate is zero unless the agent remains on the safe set during the episode. Future work includes improving this estimate and characterizing the convergence and data-efficiency of algorithms that use this gradient.

REFERENCES

- [1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.
- [2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.
- [3] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.
- [4] Y. Duan, X. Chen, R. Houthoofd, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," in *International conference on machine learning*, pp. 1329–1338, PMLR, 2016.
- [5] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.
- [6] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.
- [7] J. Garcia and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.
- [8] G. Kahn, A. Villafior, B. Ding, P. Abbeel, and S. Levine, "Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5129–5136, IEEE, 2018.
- [9] P. Geibel, "Reinforcement learning for mdps with constraints," in *European Conference on Machine Learning*, pp. 646–653, Springer, 2006.
- [10] Y. Kadota, M. Kurano, and M. Yasuda, "Discounted markov decision processes with utility constraints," *Computers & Mathematics with Applications*, vol. 51, no. 2, pp. 279–284, 2006.
- [11] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6070–6120, 2017.
- [12] R. A. Howard and J. E. Matheson, "Risk-sensitive markov decision processes," *Management science*, vol. 18, no. 7, pp. 356–369, 1972.
- [13] M. Sato, H. Kimura, and S. Kobayashi, "Td algorithm for the variance of return and mean-variance reinforcement learning," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 16, no. 3, pp. 353–362, 2001.
- [14] P. Geibel and F. Wysotzki, "Risk-sensitive reinforcement learning applied to control under constraints," *Journal of Artificial Intelligence Research*, vol. 24, pp. 81–108, 2005.
- [15] E. Altman, *Constrained Markov decision processes*, vol. 7. CRC Press, 1999.
- [16] V. S. Borkar, "An actor-critic algorithm for constrained markov decision processes," *Systems & control letters*, vol. 54, no. 3, pp. 207–213, 2005.
- [17] S. Bhatnagar and K. Lakshmanan, "An online actor-critic algorithm with function approximation for constrained markov decision processes," *Journal of Optimization Theory and Applications*, vol. 153, no. 3, pp. 688–708, 2012.
- [18] Q. Liang, F. Que, and E. Modiano, "Accelerated primal-dual policy optimization for safe reinforcement learning," *arXiv preprint arXiv:1802.06480*, 2018.
- [19] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*, pp. 22–31, PMLR, 2017.
- [20] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *arXiv preprint arXiv:1805.11074*, 2018.
- [21] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge, "Projection-based constrained policy optimization," *arXiv preprint arXiv:2010.03152*, 2020.
- [22] Y. Zhang, Q. Vuong, and K. Ross, "First order constrained optimization in policy space," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15338–15349, 2020.
- [23] Y. Liu, J. Ding, and X. Liu, "Ipo: Interior-point policy optimization under constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 4940–4947, 2020.
- [24] L. Zhang, L. Shen, L. Yang, S. Chen, B. Yuan, X. Wang, and D. Tao, "Penalized proximal policy optimization for safe reinforcement learning," *arXiv preprint arXiv:2205.11814*, 2022.
- [25] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.
- [26] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
- [27] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, pp. 1889–1897, PMLR, 2015.
- [28] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [29] A. Castellano, H. Min, E. Mallada, and J. A. Bazerque, "Reinforcement learning with almost sure constraints," in *Learning for Dynamics and Control Conference*, pp. 559–570, PMLR, 2022.
- [30] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, "Safe policies for reinforcement learning via primal-dual methods," *IEEE Transactions on Automatic Control*, 2022.
- [31] B. Peng, Y. Mu, J. Duan, Y. Guan, S. E. Li, and J. Chen, "Separated proportional-integral lagrangian for chance constrained reinforcement learning," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, pp. 193–199, IEEE, 2021.
- [32] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [33] Y. Censor, "Pareto optimality in multiobjective problems," *Applied Mathematics and Optimization*, vol. 4, no. 1, pp. 41–59, 1977.
- [34] K. J. Arrow, H. Azawa, L. Hurwicz, and H. Uzawa, *Studies in linear and non-linear programming*, vol. 2. Stanford University Press, 1958.
- [35] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.