Policy Gradients for Probabilistic Constrained Reinforcement Learning

Weiqin Chen, Dharmashankar Subramanian and Santiago Paternain

Abstract— **This paper considers the problem of learning safe policies in the context of reinforcement learning (RL). In particular, a safe policy or controller is one that, with high probability, maintains the trajectory of the agent in a given safe set. We relate this notion of safety to the notion of average safety often considered in the literature by providing theoretical bounds in terms of their safety and performance. The challenge of working with the probabilistic notion of safety considered in this work 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 such probabilistic gradients are naturally algorithm independent, which provides possibilities for them to be applied to various policy-based algorithms. In addition, we consider a continuous navigation problem to empirically illustrate the advantages (in terms of safety and performance) of working with probabilistic constraints as compared to average constraints.**

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]. A limitation of these approaches is that reward shaping (the process of combining reward with safety requirements) is, in general, a time-consuming process of hyper-parameter tuning that requires human intervention and is problem dependent [15], [16].

To mitigate this issue, a common approach is to employ the framework of Constrained Markov Decision Processes (CMDPs) [17] where additional cumulative (or average) rewards need to be kept above a desired threshold. This approach has been commonly used to induce safe behaviors [18]–[26]. To solve these constrained problems, primal-dual algorithms are generally used [18]–[22]. In this 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 in practice.

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. Ideally, one would like to achieve this goal for every possible trajectory. This being an ambitious goal, in this work we settle for problems that guarantee safety with high probability. We describe this setting in detail in Section II. Problems of this form have been considered in [9], [27]. The main challenge to solve problems under probabilistic safety constraints is that policy gradient-like expressions for such constraints are not readily available. This, in turn, prevents from running classical and state-of-the-art policy-based algorithms, e.g., REINFORCE [28], DDPG [29], TRPO [30], PPO [31] in the context of probabilistic constraints. Perhaps, for this reason, cumulative constraints as relaxations of the probabilistic constraints are considered in the literature [32], [33]. In Section IV we provide an expression for the gradient of the probabilistic constraint. Working directly with these constraints provides advantages as compared to the cumulative setting, in terms of an improved safety-optimality trade-off. In Section III we provide such theoretical bounds. In Section V we consider a navigation task that illustrates (i) the ability to use the gradient of the probabilistic constraint to train safe policies and (ii) the theoretical bounds established in Section III relating the problems with cumulative and probabilistic constraints. In particular, the latter traces a better safety-performance trade-off.

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 [34] 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, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function describing the quality of the decision, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability describing the dynamics of the system, $\mu : \mathcal{S} \to [0, 1]$ is the initial state distribution, and T is the time horizon. The state and action at time $t \in \{0, 1, \dots, T\}$ 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

Weiqin Chen and Santiago Paternain are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute. Email: {chenw18, paters}@rpi.edu

Dharmashankar Subramanian is with IBM T.J. Watson Research Center. Email: dharmash@us.ibm.com

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

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

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

As mentioned in Section I, the actions taken by the agent might be unsafe or result in risky behaviors. Thus, we impose safety requirements to overcome this issue. 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 $\mathcal{S}_{safe} \subset \mathcal{S}$ if and only if $\mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{safe}\}|\pi_\theta\right) \geq 1-\delta$.*

From the previous definition, a safe policy is such that the state remains within the safe set $\mathcal{S}_{\text{safe}}$ for the whole episode with probability at least $1-\delta$. With this definition, to find an optimal and safe policy, we can formulate a probabilistic safe RL problem as a constrained optimization problem

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

To solve problem (2), it is conceivable to employ gradient methods e.g., regularization [35] or primal-dual [36] 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)

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

It is important to point out that, in general, there is no guarantee that a fixed coefficient λ achieves the same solution as (2) (an exception is, for instance, in cases where (2) are convex [37]). However, λ trades-off safety and 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 rewards.

Then, to solve problem (3) locally gradient ascent [38] or its stochastic versions can be used. 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, 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 this family of methods to solve (3).

Notice that other state-of-the-art algorithms, e.g., CPO [21], RCPO [22], PCPO [23], FOCOPS [24] also rely on computing the gradients of objective functions and constraints. As such, it is not surprising that when considering safe formulations, instead of dealing with problems of the form (2) they consider cumulative constraints. In these problems, an auxiliary reward function $r_c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined, and the following problem is formulated

$$\begin{aligned} \tilde{P}^\star(\xi) = &\max_{\theta\in\mathbb{R}^d} V(\theta) \\ &\text{s.t.} \quad V_c(\theta) := \mathbb{E}\left[\sum_{t=0}^{T} r_c(S_t, A_t)\right] \geq \xi, \end{aligned} \tag{4}$$

where ξ is a hyper-parameter that induces different safety levels. Problems of the form (4) are known as CMDPs. When formulating safety as a CMDP, the function r_c is designed so that it induces safe behaviors when attaining large values. This results in a design that is problem dependent and in general, it necessitates a time-consuming process of hyper-parameter tuning. Moreover, probabilistic safety as in Definition 1 is not guaranteed by formulation (4). Both of these limitations are avoided under formulation (2).

Although different, problems (2) and (4) are not unrelated. In Section III we study the relationship between both problems. In particular, we exploit ideas from [32], [39] to provide explicit bounds on safety guarantees and the values of the solutions to problems (2) and (4). As discussed earlier, an immediate challenge in solving the probabilistic safety constrained problem in (2) is the computation of the gradient of the probabilistic safety. In Section IV we provide such expressions. Other than concluding remarks, this paper finishes with numerical experiments (Section V) that explore empirically the safety-performance trade-offs of both formulations.

III. The Relationship between Probabilistic and Cumulative Constraints

In the context of guaranteeing that the agent remains in a safe subset of the state space, a possibility [14] for choosing $r_c(S_t, A_t)$ in (4) is $r_c(S_t, A_t) = \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}})/(T+1)$. In this case, the cumulative notion of safety is related to the function

$$\mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{1}(S_t \in \mathcal{S}_{\text{safe}})|\pi_\theta\right]. \tag{5}$$

We turn our attention then, to understanding what is the relationship between the probabilistic safety constraint in Definition 1 and the cumulative notion given by (5). In the remaining of this section, we discuss such relationship in terms of agent's performance of task completion. To proceed we consider the following formulation that replaces the left-hand side of the constraint of problem (4) with (5)

$$\begin{aligned} \hat{P}^\star = &\max_{\theta\in\mathbb{R}^d} V(\theta) \\ &\text{s.t.} \quad \mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right)|\pi_\theta\right] \geq 1-\frac{\delta}{T+1}, \end{aligned} \tag{6}$$

where ξ is set to $1-\delta/(T+1)$, specifically. The formulation presented here guarantees that any θ that satisfies the constraint in (6) is guaranteed to be safe as in Definition 1 with probability $1-\delta$. We formally state and prove this claim in the following proposition.

Proposition 1. *Denote by $\tilde{\theta}$ a feasible solution to problem* (6)*. Then, $\tilde{\theta}$ is a feasible solution to problem* (2) *as well, i.e., the policy induced by $\tilde{\theta}$ guarantees safety in the sense of Definition 1.*

Proof. We start by writing the probability of being safe at time t as the expectation of the indicator function. Hence we have that

$$\begin{aligned}&\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{P}\left(\{S_t\in\mathcal{S}_{\text{safe}}\}|\pi_{\tilde{\theta}}\right)\\&=\mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{1}\left(S_t\in\mathcal{S}_{\text{safe}}\right)|\pi_{\tilde{\theta}}\right]\geq 1-\frac{\delta}{T+1},\end{aligned}\tag{7}$$

where the inequality follows from the fact that $\tilde{\theta}$ is feasible for problem (6). By the rule of the probability of complementary events, the previous expression can be rewritten as

$$\frac{1}{T+1}\sum_{t=0}^{T}\left(1-\mathbb{P}\left(\overline{\{S_t\in\mathcal{S}_{\text{safe}}\}|\pi_{\tilde{\theta}}}\right)\right)\geq 1-\frac{\delta}{T+1}.\tag{8}$$

The previous inequality is equivalent to

$$\sum_{t=0}^{T}\mathbb{P}\left(\overline{\{S_t\in\mathcal{S}_{\text{safe}}\}|\pi_{\tilde{\theta}}}\right)\leq\delta.\tag{9}$$

Then applying Boole-Fréchet-Bonferroni inequality [40, Chapter 1] to the previous inequality yields

$$\mathbb{P}\left(\bigcup_{t=0}^{T}\overline{\{S_t\in\mathcal{S}_{\text{safe}}\}|\pi_{\tilde{\theta}}}\right)\leq\sum_{t=0}^{T}\mathbb{P}(\overline{\{S_t\in\mathcal{S}_{\text{safe}}\}|\pi_{\tilde{\theta}}})\leq\delta.\tag{10}$$

Using the rule of the probability of complementary events and DeMorgan's law, (10) is equivalent to

$$\mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t\in\mathcal{S}_{\text{safe}}\}|\pi_{\tilde{\theta}}\right)\geq 1-\delta.\tag{11}$$

This completes the proof of Proposition 1. ■

Although Proposition 1 confirms that any feasible solution to problem (6) is also feasible to problem (2), the converse is not true. This indicates that $\hat{P}^\star$ in (6) is smaller than $P^\star$ in (2). In the next Theorem, we provide upper and lower bounds for the latter in terms of the former. Before doing so, we recall the definition of the Lagrange multiplier associated with (6). Let us define the dual function associated with (6)

$$d(\hat{\lambda})=\max_{\theta\in\mathbb{R}^d}V(\theta)+\hat{\lambda}(V_c(\theta)-(1-\delta/(T+1))),\tag{12}$$

where $\hat{\lambda}>0$. The dual function provides an upper bound on problem (6) [37, Chapter 5]. Thus in general, one is interested in finding the $\hat{\lambda}$ that provides the tightest of the upper bounds. The $\hat{\lambda}$ where this is achieved is termed the optimal Lagrange multiplier

$$\hat{\lambda}^\star=\underset{\hat{\lambda}\in\mathbb{R}}{\operatorname{argmin}}\ d(\hat{\lambda}).\tag{13}$$

Having defined $\hat{\lambda}^\star$, we are in conditions of stating the bounds between the optimal values of (2) and (6). This is the subject of the following theorem.

Theorem 1. *Let $P^\star$ and $\hat{P}^\star$ denote the optimal values of problem* (2) *and problem* (6)*. Denote by $\hat{\lambda}^\star$ the solution to the dual problem associated with problem* (6)*, as defined in* (13)*. Then it holds that*

$$\hat{P}^\star+\hat{\lambda}^\star\frac{\delta T}{T+1}\geq P^\star\geq\hat{P}^\star.\tag{14}$$

Proof. We start by proving the rightmost inequality. Denote by $\tilde{\theta}^\dagger$ the optimal solution to problem (6). By virtue of Proposition 1 the policy $\pi_{\tilde{\theta}^\dagger}$ is $(1-\delta)$-safe, thus, it is a feasible solution to problem (2). It follows, by definition of the optimal solution to problem (2), that $P^\star\geq V(\tilde{\theta}^\dagger)=\hat{P}^\star$.

Having established the rightmost inequality of the claim, we set our focus on proving the leftmost inequality. To do so, consider the following perturbation to problem (6)

$$\begin{aligned}\bar{P}^\star=&\max_{\theta\in\mathbb{R}^d}V(\theta)\\&\text{s.t.}\quad\mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{1}\left(S_t\in\mathcal{S}_{\text{safe}}\right)|\pi_\theta\right]\geq 1-\delta.\end{aligned}\tag{15}$$

Notice that problem (15) is the same problem as (6) with a looser constraint. Similar to the rightmost inequality in (14), employing Lemma 1 in Appendix A yields $\bar{P}^\star\geq P^\star$. In addition, applying Lemma 2 in Appendix A to (6) and (15) yields the following relationship between their optimal values

$$\bar{P}^\star\leq\hat{P}^\star+\hat{\lambda}^\star\frac{\delta T}{T+1}.\tag{16}$$

Combining with the fact that $P^\star\leq\bar{P}^\star$ completes the proof of Theorem 1. ■

Theorem 1 provides a lower bound and upper bound on problem (2) that depends on the the optimal value of problem (6) and its optimal Lagrange multiplier. Note that for large horizons the optimality gap between the two problems can be approximated by $\delta\hat{\lambda}^\star$. This suggests that an increasing safety requirement, i.e., δ approaching zero would make the two problems equivalent. In fact, when $\delta=0$ the two problems are equivalent. A caveat is that the bound also depends on the value of the Lagrange multiplier which in turn depends on δ. For a given level of safety δ, we can interpret how the bound changes for different systems. An interpretation of Lagrange multipliers is that they provide a measurement of how hard it is to satisfy the constraint in (6). Thus, for systems, where satisfying the constraint is easier (smaller Lagrange multiplier), the bound between (2) and (6) will be tighter.

We illustrate the intuition numerically by a navigation task in Section V. Before doing so, in the next section, we provide an expression that allows us to compute the gradient of the probabilistic safety constraints.

IV. The Gradients of Probabilistic Constraints

To solve probabilistic constrained reinforcement learning formulations, i.e., problems of the form (2), we are required to compute the gradient of the probabilistic constraints with respect to policy parameters θ. We proceed by defining an important quantity in what follows next. Let G_t be the product of indicator functions $\mathbb{1}\left(S_u \in \mathcal{S}_{\text{safe}}\right)$ from $u = t$ $(0 \leq t \leq T)$ to $u = T$

$$G_t = \prod_{u=t}^{T} \mathbb{1}\left(S_u \in \mathcal{S}_{\text{safe}}\right). \tag{17}$$

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

Theorem 2. *Let $S_0 \in \mathcal{S}_{safe}$, the gradient of the probability of being safe for a given policy π_θ yields*

$$\begin{aligned}&\nabla_\theta \mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{safe}\}|\pi_\theta, S_0\right)\\ &= \mathbb{E}\left[\sum_{t=0}^{T-1} G_1 \nabla_\theta \log \pi_\theta(A_t \mid S_t) \mid \pi_\theta, S_0\right].\end{aligned} \tag{18}$$

Proof. See Appendix B. ■

It is worth pointing out that the proof of this result is similar to policy gradient theorems in the literature [6], [28]. To draw parallelisms between them, let us define the cumulative rewards until horizon T starting from t, i.e., the so-called "reward to-go" [41] from the transition (S_t, A_t)

$$R_t = \sum_{u=t}^{T} r(S_u, A_u). \tag{19}$$

Then, the gradient of the value function (1) can be computed using Policy Gradient Theorem [6] as

$$\nabla_\theta V(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} R_t \nabla_\theta \log \pi_\theta(A_t \mid S_t) \mid \pi_\theta, S_0\right]. \tag{20}$$

Despite the similarities between (18) and (20) an important difference is evident. In (20) the gradient of the logarithm at time t is multiplied by the cumulative reward from time t until the end of the episode. This means that action A_t is only concerned with the rewards in the episode's future. While in (18), this is not the case, and action A_t is concerned with the whole episode. Intuitively, if the episode is unsafe for the first few steps the actions in the future are irrelevant since the episode is unsafe as a whole. This poses some additional challenges in estimating the gradient as we discuss next.

Let us start however, describing a challenge in the computation of $\nabla_\theta \mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{\text{safe}}\} \mid \pi_\theta, S_0\right)$ which is also shared by the computation of $\nabla_\theta V(\theta)$. This is, the need to compute expectations with respect to the trajectories of the system. In order to avoid sampling a multitude of trajectories, stochastic approximation methods [42] are often considered. Namely, one can use only one sample trajectory to approximate (20)

$$\hat{\nabla}_\theta V(\theta) = \sum_{t=0}^{T-1} R_t \nabla_\theta \log \pi_\theta(A_t \mid S_t), \tag{21}$$

where R_t is the cumulative return from time t until the end of the episode as defined in (19). Likewise, applying the stochastic approximation to (18) yields

$$\hat{\nabla}_\theta \mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta, S_0\right) = \sum_{t=0}^{T-1} G_1 \nabla_\theta \log \pi_\theta(A_t | S_t). \tag{22}$$

From (22) the additional challenges in estimating the gradient of the probabilistic safety constraint become explicit. Unlike the classic policy gradient (21) (where policy parameters update each iteration), under this framework the parameters in (22) only update when every step of the trajectory is safe, i.e., when $G_1 = \prod_{t=1}^{T} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right) = 1$. To address this issue, one can use a different expression for the gradient which we provide in the next corollary.

Corollary 1. *Let G_t^c be the product of indicator functions $\mathbb{1}\left(S_u \in \mathcal{S}_{safe}\right)$ from $u = 0$ to $u = t$, i.e., $G_t^c = \prod_{u=0}^{t} \mathbb{1}\left(S_u \in \mathcal{S}_{safe}\right)$. Under the same hypothesis of Theorem 2 it follows that*

$$\begin{aligned}&\nabla_\theta \mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{safe}\}|\pi_\theta, S_0\right)\\ &= \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{E}\left[G_{t+1} \,|\, S_t, A_t\right] G_t^c \nabla_\theta \log \pi_\theta(A_t \mid S_t) \,|\, \pi_\theta, S_0\right].\end{aligned} \tag{23}$$

Proof. The proof follows directly from Theorem 2. Conditioning the right hand side of (18) with respect to S_t and A_t yields

$$\begin{aligned}&\nabla_\theta \mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{\text{safe}}\}|\pi_\theta, S_0\right)\\ &= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=0}^{T-1} G_1 \nabla_\theta \log \pi_\theta(A_t \mid S_t) \mid S_t, A_t\right] \mid \pi_\theta, S_0\right].\end{aligned} \tag{24}$$

Using the fact that $G_1 = G_{t+1}\, G_t^c$ and that $S_0, \ldots, S_t$ are measurable with respect to S_t, A_t yields (23). ■

Note that $\mathbb{E}\left[G_{t+1} \mid S_t, A_t\right]$ is the probability of remaining safe from time $t+1$ until the end of the episode. This information about the future is akin to the Q function in RL problems [34, Chapter 6]. Hence, assuming that an estimate of the probability is available, one can run actor-critic [43] type algorithms e.g., DDPG [29], TRPO [30], PPO [31] in this setting as well. In particular, the estimator of the gradient takes the form

$$\hat{\nabla}_\theta \mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{\text{safe}}\} \mid \pi_\theta, S_0\right) \tag{25}$$
$$= \sum_{t=0}^{T-1} G_t^c\, \mathbb{E}\left[G_{t+1}\right] \nabla_\theta \log \pi_\theta(A_t \mid S_t).$$

An advantage of the previous expression is that if the episode is safe until some time $0 < t < T$, we have $G_u^c = 1$, with $u \leq t$. Hence the estimate of the gradient is not zero and the policy will be updated. This is in contrast with the expression in (22).

V. Numerical Experiments

To compare the optimal values of problem (2) and problem (6), 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 $s_{goal} = (9, 1.5)$ within the time horizon $T = 20$, 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

$$\pi_\theta(a|s) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}\left(a - \mu_\theta(s)\right)^\top \Sigma^{-1}(a - \mu_\theta(s))\right), \tag{26}$$

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)

$$\mu_\theta(s) = \sum_{k=1}^{d} \theta_k \exp\left(-\frac{||s - \bar{s}_k||^2}{2\sigma^2}\right), \tag{27}$$

where $\theta = [\theta_1, \theta_2, \cdots, \theta_d]^\top$ 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 $\bar{s}_k = (x_k, y_k), k = 1, 2, \cdots, 1681$ 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., $r(s_t, a_t) = -\|s_t - s_{goal}\|^2$. We attempt to find solutions to (2) and (6) through regularization. This is, instead of solving (2) we consider the regularized version in (3) and instead of (6) the following regularization of the constraints

$$V_\mu(\theta) = \mathbb{E}\left[\sum_{t=0}^{T}\Big(r(S_t, A_t) + \mu\mathbb{1}(S_t \in \mathcal{S}_{\text{safe}})\Big)\Big|\pi_\theta\right], \tag{28}$$

where μ is a parameter that trades-off the reward for safety. For large μ policies are induced to be excessively safe, whereas if μ is small the agent is more concerned with rewards.

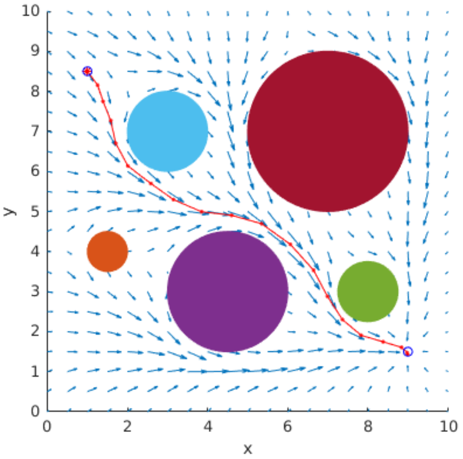

Fig. 1. Navigation policy learned after 40,000 iterations for probabilistic constraint formulation selecting $\lambda = 6, \eta = 0.002$. The agent is trained to navigate starting from (1, 8.5) to a goal (9, 1.5).

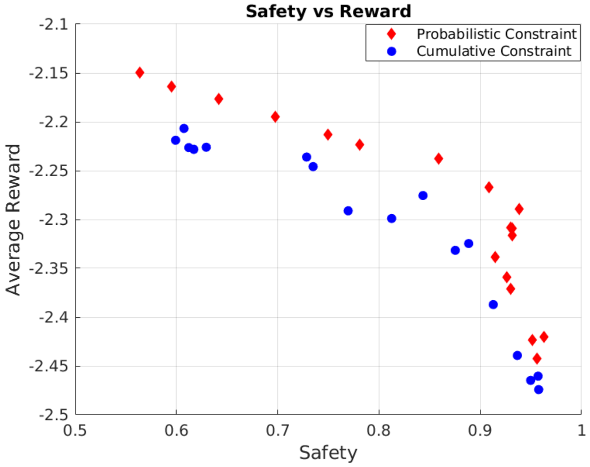

Fig. 2. Depicted in blue is the safety-reward for the cumulative constraint formulation (28) for different values of $\mu \in [0.45, 14]$ applied to the navigation problem depicted in Figure 1. Depicted in red is the safety-reward for the probabilistic constraint formulation (3) for different values of $\lambda \in [0.45, 14]$ applied to the navigation problem depicted in Figure 1. As it can be observed the probabilistic constraint traces a better trade-off of reward and safety as defined in Definition 1.

As claimed by Theorem 1 for the same level of safety the probabilistic formulation (2) yields a value that is always larger than the CMDP formulation (6). Therefore we expect both formulations to trace different Pareto fronts as we modify the hyper-parameters λ and μ in (3) and (28).

To solve (28) we define the reward

$$r_\mu(S_t, A_t) = r(S_t, A_t) + \mu\mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}), \tag{29}$$

and we consider off-the-shelf reinforcement learning algorithms. In particular, we consider a stochastic approximation of the Policy Gradient Theorem which yields the following update rule for the parameters θ of the policy

$$\begin{aligned}\theta_{k+1} &= \theta_k + \eta \hat{\nabla} V_\mu(\theta) \\ &= \theta_k + \eta \left(\sum_{t=0}^{T-1} R_{t,\mu} \nabla_\theta \log \pi_\theta(A_t \mid S_t) \right), \end{aligned} \tag{30}$$

where η denotes the stepsize and $R_{t,\mu} = \sum_{u=t}^{T} r_\mu(S_u, A_u)$. Likewise, for problem (3), the stochastic approximation of gradient ascent yields

$$\theta_{k+1} = \theta_k + \eta \left(\hat{\nabla}_\theta V(\theta_k) + \lambda \hat{\nabla}_\theta \mathbb{P}\left(\bigcap_{t=0}^{T} \{S_t \in \mathcal{S}_{\text{safe}}\} \right) \right), \tag{31}$$

where the first term in bracket on the right-hand side is computed using (21) and the second term takes the form in (22).

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 $\eta = 0.002$.

As mentioned in Section III, μ and λ trade-off performance in task completion and safety. It is important to point out that to attain the same level of safety, in general, different values of λ and μ are required. We run algorithms (30) and (31) with different weights $\mu, \lambda \in (0, 14]$ to find solutions to problems (28) and (3) respectively. For different values of λ and μ different optimal step-sizes can be used which results in faster or slower convergence of the algorithms. The worst case scenario requires $\eta = 0.0006$. In this case 2,000,000 episodes are needed.

The level of safety is the fraction of safe episodes (with all their states in the safe set as per Definition 1) that the trained policy yields over 1000 independent episodes under different random seeds. The initial state is uniformly drawn from the safe set for each episode. Similarly, we evaluate the value function by averaging the cumulative reward across the evaluation episodes.

The scatter plot in Figure 2 depicts the results of the safety and average reward for problems (3) and (28) for different values $\lambda, \mu \in [0.45, 14]$. Observe that in both formulations, the higher the desired safety, the lower the average reward is. Despite this common trend, for the same level of safety, the probabilistic constrained problem yields an overall larger average reward. Hence, demonstrating empirically, that the probabilistic constraint formulation yields a better safety-performance trade-off. This result is in accordance with the theoretical bounds provided in Theorem 1. It is worth pointing that the curve of reward-safety in Figure 2 is approximately concave, which is consistent with (16) and shows an empirical evidence of the concavity of $\tilde{P}^\star(\xi)$ in (4) with respect to ξ.

Notice that the algorithms used in this numerical section are Monte Carlo methods [34, Chapter 5]. In the RL literature there exist algorithms that exploit temporal differences [5] and/or trust regions [30] which result in faster convergence rates. The goal of this numerical section is not to show faster convergence but to demonstrate the theoretical underpinnings established in Section III as well as to show that the estimation of the gradient of the probabilistic constraints (Theorem 2) can successfully be employed for solving the task at hand. As mentioned in Section IV 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.

VI. Conclusions

In this work, we studied the problem of learning safe policies. Concretely, we consider probabilistic safety. This is, a safe policy as one that guarantees, with high probability, that the state of the agent remains in the desired safe throughout the whole trajectory. This concept is different than the cumulative safety constraints often consider in the literature. Albeit the differences, we have established conditions where cumulative constraints can guarantee feasibility in the probabilistic safety sense. To achieve this, in general, the set of feasible policies needs to be shrunk which results in a loss in performance. In particular, we provided upper and lower bounds to characterize this effect. The result is that the probabilistic safety setting shows a better safety and performance trade-off.

Regardless of the notion of safety considered, a natural way to solve RL problems is to find the corresponding gradients of the objective functions and constraints. Despite that classic policy gradient algorithms apply in cumulative constraint setting directly, the gradients for solving probabilistic formulations are not available. In this work, we have provided the first explicit gradient expression for the probabilistic constraint. We have also demonstrated that updates based on these gradients can be used to solve continuous navigation problems in cluttered environments. The expression for the gradient, however, has some limitations. In particular, the fact that the gradient estimate is zero unless the agent remains on the safe set for all steps in the episode. We have provided an alternative expression that can allow for actor-critic type of algorithms. Developing these algorithms and characterizing their convergences rates and data-efficiency are topics beyond the scope of this work. In addition, using the navigation task, we have confirmed empirically our theoretical results regarding the safety-performance trade-off that both notions of safety induce.

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. Houthooft, 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. Garcıa 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. Villaflor, 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] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg, “Ai safety gridworlds,” *arXiv preprint arXiv:1711.09883*, 2017.

[16] H. Mania, A. Guy, and B. Recht, “Simple random search provides a competitive approach to reinforcement learning,” *arXiv preprint arXiv:1803.07055*, 2018.

[17] E. Altman, *Constrained Markov decision processes*, vol. 7. CRC Press, 1999.

[18] V. S. Borkar, “An actor-critic algorithm for constrained markov decision processes,” *Systems & control letters*, vol. 54, no. 3, pp. 207–213, 2005.

[19] 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.

[20] Q. Liang, F. Que, and E. Modiano, “Accelerated primal-dual policy optimization for safe reinforcement learning,” *arXiv preprint arXiv:1802.06480*, 2018.

[21] J. Achiam, D. Held, A. Tamar, and P. Abbeel, “Constrained policy optimization,” in *International conference on machine learning*, pp. 22–31, PMLR, 2017.

[22] C. Tessler, D. J. Mankowitz, and S. Mannor, “Reward constrained policy optimization,” *arXiv preprint arXiv:1805.11074*, 2018.

[23] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge, “Projection-based constrained policy optimization,” *arXiv preprint arXiv:2010.03152*, 2020.

[24] 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.

[25] 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.

[26] L. Shen, L. Yang, S. Chen, B. Yuan, X. Wang, D. Tao, *et al.*, “Penalized proximal policy optimization for safe reinforcement learning,” *arXiv preprint arXiv:2205.11814*, 2022.

[27] E. Delage and S. Mannor, “Percentile optimization for markov decision processes with parameter uncertainty,” *Operations research*, vol. 58, no. 1, pp. 203–213, 2010.

[28] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.

[29] 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.

[30] 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.

[31] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.

[32] 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.

[33] M. Calvo-Fullana, L. F. Chamon, and S. Paternain, “Towards safe continuing task reinforcement learning,” in *2021 American Control Conference (ACC)*, pp. 902–908, IEEE, 2021.

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

[35] Y. Censor, “Pareto optimality in multiobjective problems,” *Applied Mathematics and Optimization*, vol. 4, no. 1, pp. 41–59, 1977.

[36] K. J. Arrow, H. Azawa, L. Hurwicz, and H. Uzawa, *Studies in linear and non-linear programming*, vol. 2. Stanford University Press, 1958.

[37] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[38] D. P. Bertsekas, “Nonlinear programming,” *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[39] N. Wagener, B. Boots, and C.-A. Cheng, “Safe reinforcement learning using advantage-based intervention,” *arXiv preprint arXiv:2106.09110*, 2021.

[40] R. Durrett, “Probability: Theory and examples cambridge university press,” 2010.

[41] J. Vitay, “Deep reinforcement learning,” 2020.

[42] H. Robbins and S. Monro, “A stochastic approximation method,” *The annals of mathematical statistics*, pp. 400–407, 1951.

[43] V. Konda and J. Tsitsiklis, “Actor-critic algorithms,” *Advances in neural information processing systems*, vol. 12, 1999.

Appendix

A. Technical Lemmas for the Proof of Theorem 1

Lemma 1. *Let $\hat{\mathcal{F}}$, $\mathcal{F}$ and $\bar{\mathcal{F}}$ represent the feasible sets to problem* (6)*, problem* (2) *and problem* (15)*, respectively. Then, it holds that $\hat{\mathcal{F}} \subseteq \mathcal{F} \subseteq \bar{\mathcal{F}}$.*

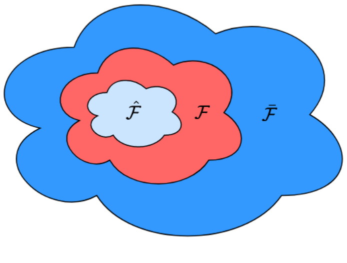

Fig. 3. The illustration for feasible sets of problem (6)–$\hat{\mathcal{F}}$ (light blue), problem (2)–$\mathcal{F}$ (red), and problem (15)–$\bar{\mathcal{F}}$ (navy blue).

Proof. We start by proving the leftmost inclusion. Denote by $\hat{\theta}$ the any feasible solution to problem (6), i.e., $\hat{\theta} \in \hat{\mathcal{F}}$. By virtue of Proposition 1 the policy $\pi_{\hat{\theta}}$ is $(1-\delta)$ safe as in Definition 1. In turn, this means that $\hat{\theta}$ is a feasible solution to problem (2) as well. Then it holds that $\hat{\theta} \in \mathcal{F}$ for $\forall \hat{\theta} \in \hat{\mathcal{F}}$, thus $\hat{\mathcal{F}} \subseteq \mathcal{F}$.

We now focus on establishing the second inclusion. Denote by $\bar{\theta}$ by any point in $\mathcal{F}$. $\bar{\theta}$ is thus a feasible solution to problem (2), i.e.,

$$\mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{\text{safe}}\} \mid \pi_{\bar{\theta}}\right) \geq 1-\delta. \tag{32}$$

Observe that the previous inequality is equivalent to

$$\mathbb{P}\left(\sum_{t=0}^{T} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right) = T+1 \mid \pi_{\bar{\theta}}\right) \geq 1-\delta. \tag{33}$$

Indeed, for the state to belong to $\mathcal{S}_{\text{safe}}$ for all times, all the indicator functions in (33) needs to take the value 1. Since $\sum_{t=0}^{T} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right)$ is a non-negative random variable, it follows that

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=0}^{T} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right) | \pi_{\bar{\theta}}\right] \qquad (34)\\ &\quad \geq \mathbb{P}\left(\sum_{t=0}^{T} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right) = T+1 | \pi_{\bar{\theta}}\right)(T+1).\end{aligned}$$

Combining (33) and (34), it follows that

$$\mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right) | \pi_{\bar{\theta}}\right] \geq 1-\delta. \tag{35}$$

Hence, $\bar{\theta}$ is a feasible point in $\bar{\mathcal{F}}$ for $\forall \bar{\theta} \in \mathcal{F}$, i.e., $\mathcal{F} \subseteq \bar{\mathcal{F}}$. This completes the proof of Lemma 1. ■

Lemma 2. *Consider the function $\tilde{P}^\star(\xi)$ defined in* (4). *Let $\xi_0, \xi_1 \in \mathbb{R}$. Let $\tilde{\lambda}^\star(\xi_0)$ be the dual optimal solution to* (4) *with $\xi = \xi_0$, defined tantamount to* (13). *It holds that*

$$\tilde{P}^\star(\xi_1) \leq \tilde{P}^\star(\xi_0) + \tilde{\lambda}^\star(\xi_0)(\xi_0 - \xi_1). \tag{36}$$

Proof. Recall the definition of the dual problem associated to (4)

$$\tilde{D}^\star(\xi) = \min_{\tilde{\lambda} \in \mathbb{R}} \max_{\theta \in \mathbb{R}^d} V(\theta) + \tilde{\lambda}(V_c(\theta) - \xi), \tag{37}$$

where $\tilde{\lambda} \in \mathbb{R}, \theta \in \mathbb{R}^d$. It follows from Theorem 3 in [32] that zero duality gap holds for problem (4)

$$\tilde{P}^\star(\xi_1) = \tilde{D}^\star(\xi_1) = V(\theta^\star(\xi_1)) + \tilde{\lambda}^\star(\xi_1)(V_c(\theta^\star(\xi_1)) - \xi_1), \tag{38}$$

where $(\theta^\star(\xi_1), \tilde{\lambda}^\star(\xi_1))$ denote the primal-dual optimal solution of (4) with $\xi = \xi_1$. Likewise, we can also write

$$\tilde{P}^\star(\xi_0) = V(\theta^\star(\xi_0)) + \tilde{\lambda}^\star(\xi_0)(V_c(\theta^\star(\xi_0)) - \xi_0), \tag{39}$$

where the primal-dual solution with respect to ξ_0 is denoted by $(\theta^\star(\xi_0), \tilde{\lambda}^\star(\xi_0))$. By definition of $\tilde{\lambda}^\star(\xi_1)$, i.e., the minimizer of (37) with $\xi = \xi_1$, it follows that for any $\lambda > 0$ we have that

$$\begin{aligned}\tilde{P}^\star(\xi_1) &= V(\theta^\star(\xi_1)) + \tilde{\lambda}^\star(\xi_1)(V_c(\theta^\star(\xi_1)) - \xi_1)\\ &\leq V(\theta^\star(\xi_1)) + \lambda(V_c(\theta^\star(\xi_1)) - \xi_1).\end{aligned} \tag{40}$$

In particular, this holds for $\lambda = \tilde{\lambda}^\star(\xi_0)$

$$\tilde{P}^\star(\xi_1) \leq V(\theta^\star(\xi_1)) + \tilde{\lambda}^\star(\xi_0)(V_c(\theta^\star(\xi_1)) - \xi_1). \tag{41}$$

By adding and subtracting $\tilde{\lambda}^\star(\xi_0)\,\xi_0$ to the previous expression yields

$$\begin{aligned}\tilde{P}^\star(\xi_1) \leq\, & V(\theta^\star(\xi_1)) + \tilde{\lambda}^\star(\xi_0)(V_c(\theta^\star(\xi_1)) - \xi_0)\\ &+ \tilde{\lambda}^\star(\xi_0)(\xi_0 - \xi_1).\end{aligned} \tag{42}$$

Likewise, $\theta^\star(\xi_0)$ is the primal maximizer of the Lagrangian with $\xi = \xi_0$

$$\theta^\star(\xi_0) = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmax}}\, V(\theta) + \tilde{\lambda}^\star(\xi_0)\left(V_c(\theta) - \xi_0\right). \tag{43}$$

Thus $\tilde{P}^\star(\xi_1)$ in (42) is upper bounded as

$$\begin{aligned}\tilde{P}^\star(\xi_1) \leq\, & V(\theta^\star(\xi_0)) + \tilde{\lambda}^\star(\xi_0)(V_c(\theta^\star(\xi_0)) - \xi_0)\\ &+ \tilde{\lambda}^\star(\xi_0)(\xi_0 - \xi_1).\end{aligned} \tag{44}$$

Substituting (39) into (44) reduces to

$$\tilde{P}^\star(\xi_1) \leq \tilde{P}^\star(\xi_0) + \tilde{\lambda}^\star(\xi_0)(\xi_0 - \xi_1). \tag{45}$$

This completes the proof of Lemma 2. ■

B. Proof of Theorem 2

We proceed by presenting and proving the following two technical lemmas (Lemma 3 and Lemma 4).

Lemma 3. *Given $S_{t-1} \in \mathcal{S}_{safe}$ and $G_t, t = 1, 2, \cdots, T-1$ defined in* (17), *it holds that*

$$\begin{aligned}\nabla_\theta \mathbb{E}\left[G_t \mid S_{t-1}\right] =\, & \mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_{t+1} | S_t\right] \mathbb{1}\left(S_t \in \mathcal{S}_{safe}\right) | S_{t-1}\right]\\ &+ \mathbb{E}\left[G_t \nabla_\theta \log \pi_\theta(A_{t-1} \mid S_{t-1}) \mid S_{t-1}\right].\end{aligned} \tag{46}$$

Proof. We start the proof by rewriting the expectation of G_1 with respect to S_0 by using the towering property of the expectation

$$\begin{aligned}\mathbb{E}\left[G_1 \mid S_0\right] &= \mathbb{E}\left[\mathbb{E}\left[G_1 \mid S_1\right] \mid S_0\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[G_2 \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_1\right] \mid S_0\right],\end{aligned} \tag{47}$$

where the second equality follows from (17). Since S_1 is measurable with respect to the σ-algebra $\mathcal{F}_1$ it follows that

$$\mathbb{E}\left[G_1 \mid S_0\right] = \mathbb{E}\left[\mathbb{E}\left[G_2 \mid S_1\right] \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_0\right]. \tag{48}$$

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

$$\mathbb{E}\left[G_1 \mid S_0\right] = \int_{\mathcal{S}} \mathbb{E}\left[G_2 \mid s_1\right] \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right) p\left(s_1 \mid S_0\right) ds_1. \tag{49}$$

Notice that the conditional probability of s_1 given S_0 can be rewritten as

$$p(s_1 \mid S_0) = \int_{\mathcal{A}} p(s_1 \mid S_0, a_0) \pi_\theta(a_0 \mid S_0)\, da_0. \tag{50}$$

Hence, substituting the previous expression into (49) yields

$$\begin{aligned}\mathbb{E}\left[G_1 \mid S_0\right] = \int_{\mathcal{S}\times\mathcal{A}} & \mathbb{E}\left[G_2 \mid s_1\right] \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right)\\ & p(s_1 \mid S_0, a_0) \pi_\theta(a_0 \mid S_0)\, ds_1 da_0.\end{aligned} \tag{51}$$

Taking the gradient of the previous expression with respect to the policy parameters θ results in the following expression

$$\begin{aligned}\nabla_\theta \mathbb{E}\left[G_1 \mid S_0\right] = &\int_{\mathcal{S}\times\mathcal{A}} \nabla_\theta \left(\mathbb{E}\left[G_2 \mid s_1\right]\right) \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right) \\ &\qquad p(s_1 \mid S_0,\, a_0)\pi_\theta(a_0 \mid S_0)\, ds_1 da_0 \\ &+ \int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[G_2 \mid s_1\right] \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right) \\ &\qquad p(s_1 \,|\, S_0,\, a_0)\nabla_\theta \pi_\theta(a_0 \,|\, S_0)\, ds_1 da_0.\end{aligned} \tag{52}$$

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

$$\begin{aligned}&\int_{\mathcal{S}\times\mathcal{A}} \nabla_\theta \left(\mathbb{E}\left[G_2 \mid s_1\right]\right) \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right) \\ &\qquad p(s_1 \mid S_0,\, a_0)\pi_\theta(a_0 \mid S_0)\, ds_1 da_0 \\ &= \mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_2 \mid S_1\right] \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_0\right].\end{aligned} \tag{53}$$

The second term using the "log-trick", i.e. the fact that $\nabla_\theta \pi_\theta(a_0 \mid S_0) = \pi_\theta(a_0 \mid S_0)\nabla_\theta \log \pi_\theta(a_0 \mid S_0)$ yields

$$\begin{aligned}&\int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[G_2 \mid s_1\right] \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right) \\ &\qquad p(s_1 \mid S_0,\, a_0)\nabla_\theta \pi_\theta(a_0 \mid S_0)\, ds_1 da_0 \\ &= \int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[G_2 \mid s_1\right] \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right) p(s_1 \mid S_0,\, a_0) \\ &\qquad \pi_\theta(a_0 \mid S_0)\nabla_\theta \log \pi_\theta(a_0 \mid S_0)\, ds_1 da_0.\end{aligned} \tag{54}$$

Likewise, since s_1 is measurable with respect to the σ-algebra $\mathcal{F}_1$, (54) can be simplified as follows

$$\begin{aligned}&\int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[G_2 \mid s_1\right] \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right) p(s_1 \mid S_0,\, a_0) \\ &\qquad \pi_\theta(a_0 \mid S_0)\nabla_\theta \log \pi_\theta(a_0 \mid S_0)\, ds_1 da_0 \\ &= \int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[G_2 \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right) \mid s_1\right] p(s_1 \mid S_0,\, a_0) \\ &\qquad \pi_\theta(a_0 \mid S_0)\nabla_\theta \log \pi_\theta(a_0 \mid S_0)\, ds_1 da_0 \\ &= \int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[G_1 \mid s_1\right] p(s_1 \mid S_0,\, a_0)\pi_\theta(a_0 \mid S_0) \\ &\qquad \nabla_\theta \log \pi_\theta(a_0 \mid S_0)\, ds_1 da_0.\end{aligned} \tag{55}$$

Notice that the previous expression is equivalent to

$$\begin{aligned}&\int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[G_1 \mid s_1\right] p(s_1 \mid S_0,\, a_0)\pi_\theta(a_0 \mid S_0) \\ &\qquad \nabla_\theta \log \pi_\theta(a_0 \mid S_0)\, ds_1 da_0 \\ &= \mathbb{E}\left[\mathbb{E}\left[G_1 \mid S_1\right] \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0\right].\end{aligned} \tag{56}$$

Since $\log \pi_\theta(A_0 \mid S_0)$ is measurable given S_1, the expression above can be rewritten as

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[G_1 \mid S_1\right] \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_1\right] \mid S_0\right].\end{aligned} \tag{57}$$

Using the towering property of the expectation the previous expressions yield

$$\begin{aligned}&\int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[G_2 \mid s_1\right] \mathbb{1}\left(s_1 \in \mathcal{S}_{\text{safe}}\right) p(s_1 \mid S_0,\, a_0) \\ &\qquad \nabla_\theta \pi_\theta(a_0 \mid S_0)\, ds_1 da_0 \\ &= \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0\right].\end{aligned} \tag{58}$$

Then, combining (53) with (58) yields

$$\begin{aligned}\nabla_\theta \mathbb{E}\left[G_1 \mid S_0\right] &= \mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_2 \mid S_1\right] \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_0\right] \\ &\quad + \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0\right].\end{aligned} \tag{59}$$

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

$$\begin{aligned}\nabla_\theta \mathbb{E}\left[G_i \mid S_{i-1}\right] &= \mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_{i+1} \,|\, S_i\right] \mathbb{1}\left(S_i \in \mathcal{S}_{\text{safe}}\right) \,|\, S_{i-1}\right] \\ &\quad + \mathbb{E}\left[G_i \nabla_\theta \log \pi_\theta(A_{i-1} \mid S_{i-1}) \mid S_{i-1}\right].\end{aligned} \tag{60}$$

This completes the proof of Lemma 3. ■

Lemma 4. *Given $S_{t-1} \in \mathcal{S}_{\text{safe}}$ and $G_t, t = 1, 2, \cdots, T-1$ defined in* (17)*, it holds that*

$$\begin{aligned}\nabla_\theta \mathbb{E}\left[G_1 \mid S_0\right] &= \sum_{t=0}^{T-2} \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_t \mid S_t) \mid S_0\right] \\ &\quad + \mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_T \,|\, S_{T-1}\right] \prod_{t=1}^{T-1} \mathbb{1}\left(S_t \in \mathcal{S}_{\textit{safe}}\right) \,|\, S_0\right].\end{aligned} \tag{61}$$

Proof. We proceed by employing Lemma 3 to derive the gradient of the expectation of G_1 and G_2, respectively

$$\begin{aligned}\nabla_\theta \mathbb{E}\left[G_1 \mid S_0\right] &= \mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_2 \mid S_1\right] \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_0\right] \\ &\quad + \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0\right].\end{aligned} \tag{62}$$

$$\begin{aligned}\nabla_\theta \mathbb{E}\left[G_2 \mid S_1\right] &= \mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_3 \mid S_2\right] \mathbb{1}\left(S_2 \in \mathcal{S}_{\text{safe}}\right) \mid S_1\right] \\ &\quad + \mathbb{E}\left[G_2 \nabla_\theta \log \pi_\theta(A_1 \mid S_1) \mid S_1\right].\end{aligned} \tag{63}$$

Then, substituting (63) into (62) yields

$$\begin{aligned}&\nabla_\theta \mathbb{E}[G_1 \mid S_0] \\ &= \mathbb{E}[\mathbb{E}[\nabla_\theta \mathbb{E}[G_3 \mid S_2] \mathbb{1}(S_2 \in \mathcal{S}_{\text{safe}}) \mid S_1] \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) \\ &+ \mathbb{E}[G_2 \nabla_\theta \log \pi_\theta(A_1 \mid S_1) \mid S_1] \mathbb{1}(S_1 \in \mathcal{S}_{\text{safe}}) \mid S_0] \\ &+ \mathbb{E}[G_1 \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0].\end{aligned} \tag{64}$$

As $\mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right)$ is measurable given S_1, the previous equation can be transformed to

$$\begin{aligned}&\nabla_\theta \mathbb{E}\left[G_1 \mid S_0\right] \\ &= \mathbb{E}[\mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_3 \mid S_2\right] \mathbb{1}\left(S_2 \in \mathcal{S}_{\text{safe}}\right) \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_1\right] \\ &+ \mathbb{E}\left[G_2 \nabla_\theta \log \pi_\theta(A_1 \mid S_1) \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_1\right] \mid S_0] \\ &+ \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0\right].\end{aligned} \tag{65}$$

By definition of G_1 we can simplify the second term of the right hand side of the previous equation. Then we have

$$\begin{aligned}&\nabla_\theta \mathbb{E}\left[G_1 \mid S_0\right]\\ &= \mathbb{E}[\mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_3 \mid S_2\right] \mathbb{1}\left(S_2 \in \mathcal{S}_{\text{safe}}\right) \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_1\right]\\ &+ \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_1 \mid S_1) \mid S_1\right] \mid S_0]\\ &+ \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0\right]. \qquad (66)\end{aligned}$$

Using the towering property of the expectation (66) reduces to

$$\begin{aligned}&\nabla_\theta \mathbb{E}\left[G_1 \mid S_0\right]\\ &= \mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_3 \mid S_2\right] \mathbb{1}\left(S_2 \in \mathcal{S}_{\text{safe}}\right) \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_0\right]\\ &+ \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_1 \mid S_1) \mid S_0\right]\\ &+ \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0\right]. \qquad (67)\end{aligned}$$

Then repeatedly unwrapping $\nabla_\theta \mathbb{E}\left[G_1 \mid S_0\right]$ in terms of $G_3, \dots, G_T$ by Lemma 3 yields

$$\begin{aligned}&\nabla_\theta \mathbb{E}\left[G_1 \mid S_0\right]\\ &= \mathbb{E}[\nabla_\theta \mathbb{E}\left[G_T \mid S_{T-1}\right] \mathbb{1}\left(S_{T-1} \in \mathcal{S}_{\text{safe}}\right)\\ &\quad \cdots \mathbb{1}\left(S_2 \in \mathcal{S}_{\text{safe}}\right) \mathbb{1}\left(S_1 \in \mathcal{S}_{\text{safe}}\right) \mid S_0]\\ &+ \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_{T-2} \mid S_{T-2}) \mid S_0\right] + \cdots\\ &+ \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_0 \mid S_0) \mid S_0\right]\\ &= \sum_{t=0}^{T-2} \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_t \mid S_t) \mid S_0\right]\\ &+ \mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_T \mid S_{T-1}\right] \prod_{t=1}^{T-1} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right) \mid S_0\right]. \qquad (68)\end{aligned}$$

This completes the proof of Lemma 4. ■

We are now in conditions to prove Theorem 2. We start by rewriting the probability of remaining safe in terms of G_0 defined in (17). By definition of probability we have

$$\begin{aligned}&\mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta, S_0\right)\\ &\quad = \mathbb{E}\left[\mathbb{1}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{\text{safe}}\}\right) | \pi_\theta, S_0\right]. \qquad (69)\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta, S_0\right) &= \mathbb{E}\left[\prod_{t=0}^{T} \mathbb{1}(S_t \in \mathcal{S}_{\text{safe}}) | \pi_\theta, S_0\right]\\ &= \mathbb{E}\left[G_0 | S_0\right], \qquad (70)\end{aligned}$$

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 (70), the gradient of the probability of remaining safe reduces to

$$\nabla_\theta \mathbb{P}\left(\bigcap_{t=0}^{T}\{S_t \in \mathcal{S}_{\text{safe}}\} | \pi_\theta, S_0\right) = \nabla_\theta \mathbb{E}\left[G_1 | S_0\right]. \qquad (71)$$

In Lemma 3 we derive a recursive relationship for the gradient of $\mathbb{E}\left[G_t \mid S_{t-1}\right], t = 1, 2, \cdots, T-1$. By virtue of Lemma 4, to complete the proof of the result it suffices to establish that

$$\begin{aligned}&\mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_T \mid S_{T-1}\right] \prod_{t=1}^{T-1} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right) \mid S_0\right]\\ &\quad = \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_{T-1} \mid S_{T-1}) \mid S_0\right]. \qquad (72)\end{aligned}$$

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

$$\nabla_\theta \mathbb{E}\left[G_T \mid S_{T-1}\right] = \nabla_\theta \int_{\mathcal{S}} \mathbb{1}(s_T \in \mathcal{S}_{\text{safe}}) p(s_T | S_{T-1})\, ds_T, \qquad (73)$$

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

$$p(s_T|S_{T-1}) = \int_{\mathcal{A}} p(s_T|S_{T-1}, a_{T-1}) \pi_\theta(a_{T-1}|S_{T-1}) da_{T-1}. \qquad (74)$$

Consequently, (73) can be converted to

$$\begin{aligned}&\nabla_\theta \mathbb{E}\left[G_T \mid S_{T-1}\right]\\ &\quad = \nabla_\theta \int_{\mathcal{S}\times\mathcal{A}} \mathbb{1}\left(s_T \in \mathcal{S}_{\text{safe}}\right) p(s_T \mid S_{T-1}, a_{T-1})\\ &\qquad\qquad \pi_\theta(a_{T-1} \mid S_{T-1})\, ds_T da_{T-1}. \qquad (75)\end{aligned}$$

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

$$\begin{aligned}\nabla_\theta \mathbb{E}\left[G_T \mid S_{T-1}\right] &= \int_{\mathcal{S}\times\mathcal{A}} \mathbb{1}\left(s_T \in \mathcal{S}_{\text{safe}}\right) p(s_T \,|\, S_{T-1}, a_{T-1})\\ &\qquad \nabla_\theta \pi_\theta(a_{T-1} \mid S_{T-1})\, ds_T da_{T-1}. \qquad (76)\end{aligned}$$

Applying the "log-trick" to the right hand side of the previous equation yields

$$\begin{aligned}&\nabla_\theta \mathbb{E}\left[G_T \mid S_{T-1}\right]\\ &= \int_{\mathcal{S}\times\mathcal{A}} \mathbb{1}\left(s_T \in \mathcal{S}_{\text{safe}}\right) p(s_T \mid S_{T-1}, a_{T-1})\\ &\quad \pi_\theta(a_{T-1} \mid S_{T-1}) \nabla_\theta \log \pi_\theta(a_{T-1} \mid S_{T-1})\, ds_T da_{T-1}. \qquad (77)\end{aligned}$$

Since $p\left(s_T \mid S_{T-1}, a_{T-1}\right) \pi_\theta\left(a_{T-1} \mid S_{T-1}\right) = p\left(s_T, a_{T-1} \mid S_{T-1}\right)$ is the joint probability distribution of S_T and A_{T-1} given S_{T-1} the previous expression can be rewritten as

$$\nabla_\theta \mathbb{E}\left[G_T \mid S_0\right] = \mathbb{E}\left[G_T \nabla_\theta \log \pi_\theta(A_{T-1} \mid S_{T-1}) \mid S_{T-1}\right]. \qquad (78)$$

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

$$\begin{aligned} &\nabla_\theta \mathbb{E}\left[G_T \mid S_0\right] \prod_{t=1}^{T-1} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right) \\ &\quad = \mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_{T-1} \mid S_{T-1}) \mid S_{T-1}\right], \end{aligned} \tag{79}$$

where we have used that $G_1 = G_T \prod_{t=1}^{T-1} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right)$. Substituting the previous expression in the left hand side of (72) it follows that

$$\begin{aligned} &\mathbb{E}\left[\nabla_\theta \mathbb{E}\left[G_T \mid S_{T-1}\right] \prod_{t=1}^{T-1} \mathbb{1}\left(S_t \in \mathcal{S}_{\text{safe}}\right) \mid S_0\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[G_1 \nabla_\theta \log \pi_\theta(A_{T-1} \mid S_{T-1}) \mid S_{T-1}\right] \mid S_0\right]. \end{aligned} \tag{80}$$

The law of total expectation completes the result claimed in (72) and therefore completes the proof of Theorem 2.