

Off-Policy Reinforcement Learning with Anytime Safety Guarantees via Robust Safe Gradient Flow

Pol Mestres Arnau Marzabal Jorge Cortés

Abstract—This paper considers the problem of solving constrained reinforcement learning (RL) problems with anytime guarantees, meaning that the algorithmic solution must yield a constraint-satisfying policy at every iteration of its evolution. Our design is based on a discretization of the Robust Safe Gradient Flow (RSGF), a continuous-time dynamics for anytime constrained optimization whose forward invariance and stability properties we formally characterize. The proposed strategy, termed RSGF-RL, is an off-policy algorithm which uses episodic data to estimate the value functions and their gradients and updates the policy parameters by solving a convex quadratically constrained quadratic program. Our technical analysis combines statistical analysis, the theory of stochastic approximation, and convex analysis to determine the number of episodes sufficient to ensure that safe policies are updated to safe policies and to recover from an unsafe policy, both with an arbitrary user-specified probability, and to establish the asymptotic convergence to the set of KKT points of the RL problem almost surely. Simulations on a navigation example and the cart-pole system illustrate the superior performance of RSGF-RL with respect to the state of the art.

I. INTRODUCTION

Reinforcement learning (RL) seeks to find an optimal decision policy by having an agent interact with its environment through trial and error. At any given state, an action taken by the agent makes them transition to a new state with some probability, after which they incur an associated reward. The optimal policy is that which maximizes a prespecified long-horizon cumulative reward. Today, RL-based methods are pervasive in a wide range of technological applications of machine learning and artificial intelligence in society. However, the use of RL in safety-critical applications (e.g., autonomous driving, healthcare, or energy management) requires additional precautions, because the process of trial-and-error can lead the agent towards unsafe regions, with potentially catastrophic consequences. This has sparked the development of safe RL techniques that seek to find optimal policies meeting desired safety specifications. In this paper, we design an algorithm to solve constrained RL problems in an anytime fashion, meaning that the algorithm satisfies the constraints at every iterate.

Literature Review: Safe RL has been actively pursued in recent years, see [2]–[5] for comprehensive surveys on the subject. Here, we discuss the works best aligned with the approach to safe RL taken here. Safety constraints in RL are often expressed as *cumulative constraints*, which require the expected value of a sum of costs over a given time horizon to

be kept below a certain threshold [6]–[9]. Markov Decision Processes with such type of constraints are referred to as Constrained Markov Decision Processes (CMDPs). A standard approach to solve CMDPs are primal-dual methods [10], [11], which take a gradient ascent step in the primal variable and a gradient descent step in the dual variable. For finite state and action MDPs with a special type of transition functions, [12] shows that such primal-dual scheme converges to the optimal policy. Similarly, for continuous state and action spaces, [9] also provides a primal-dual scheme that provably converges to the optimal policy. However, these guarantees require solving an unconstrained RL algorithm at every iteration, which makes the algorithm computationally hard to execute (although practical implementations are given in [10]). Furthermore, primal-dual schemes can lead to safety violations during the training process, which compromises their implementation in physical domains. Although there exist implementations of primal-dual methods that guarantee safety during training, these are either limited to particular policy parametrizations [13] or solve a relaxed version of the problem and hence introduce an optimality gap [14]. Beyond primal-dual methods, there exist other algorithms in the literature whose goal is to provide safety guarantees during training. For example, [7] proposes CPO, an algorithm that is solely based on primal updates and that enjoys safety guarantees at every iteration. However, performing the exact policy update is computationally intensive, and the proposed practical implementations employ a first-order approximation of the objective and constraints that might violate the safety constraints during training. On the other hand, [15] introduces IPO, another primal method that includes the safety constraints as penalty terms in the objective function, and also guarantees the satisfaction of the safety constraints during training. However, this algorithm presupposes the existence of a safe initial policy and its convergence properties are not studied. The method proposed in [16] leverages Lyapunov functions to guarantee the satisfaction of constraints during training. However, the method proposed to search for such Lyapunov functions might be computationally intensive, and it is only shown to converge for a limited class of problems. On the other hand, [17] introduces an algorithm for finite state and action CMDPs that guarantees that trajectories satisfy budget constraints at all times. It is also possible to optimize over a class of truncated policies so that unsafe actions have probability zero, as in [18], but such restrictions also introduce an optimality gap, which is not formally quantified.

The methods described above are all on-policy, i.e., they rely on trajectories from the current policy iterate to generate the estimates needed for the algorithm execution. Instead, here we pursue the design of off-policy methods, where trajectories

A preliminary version of this work appeared as [1] at the 2025 Learning for Dynamics and Control Conference. This work was partially supported by ONR Award N00014-23-1-2353.

The authors are with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, {pmestres,amarzabal,cortes}@ucsd.edu

from other policies can be used to generate the estimates. Such methods enable the use of datasets of trajectories obtained offline or in previous iterations, significantly enhancing the efficiency of the algorithm implementation.

Statement of Contributions: The paper contributions are:

- (i) we introduce a continuous-time algorithm for anytime constrained optimization termed Robust Safe Gradient Flow (RSGF). We identify conditions under which the RSGF is well defined and locally Lipschitz. We also establish the equivalence between its equilibria and the KKT points of the constrained optimization problem, and show forward invariance of the constraint set and convergence to the set of KKT points;
- (ii) we define estimates for the value functions defining the constrained RL problem as well as their gradients. These estimates are off-policy, i.e. the estimates of any given policy can be constructed using trajectories generated by other policies. We establish a range of statistical properties of these estimates, including their mean and bounds on their variance and tail probabilities;
- (iii) we combine (i) and (ii) to introduce the off-policy Robust Safe Gradient Flow-based Reinforcement Learning (RSGF-RL). This algorithm is based on a discretization of RSGF and employs the off-policy estimates of the value function and their gradients. By leveraging the statistical properties of the latter, we determine a sufficient number of episodes such that RSGF-RL updates safe (and unsafe but close to safe) policies to safe policies for any prescribed confidence. Combining the properties of RSGF with the theory of stochastic approximation [19], [20], we also show that the iterates of RSGF-RL asymptotically converge to a KKT point almost surely, and characterize its rate of convergence.
- (iv) we illustrate the performance of RSGF-RL on a navigation example and the cart-pole system, and compare it against the state of the art.

Preliminary results were presented in the conference article [1], whose focus was restricted to on-policy data and a single safety constraint. Furthermore, the convergence to the set of KKT points was only ensured in expectation. All of these are special cases of the present work. The generalization here from on-policy to off-policy data and the establishment of almost sure convergence, along with the novel technical treatment based on the dynamical properties of RSGF and the theory of stochastic approximation, are instrumental in expanding the applicability of the proposed framework.

II. PRELIMINARIES

We introduce here the notation and basic notions on stability of dynamical systems, Markov decision processes, and constraint qualification in nonlinear programming.

Notation: We denote by $\mathbb{Z}_{>0}$, \mathbb{R} , and $\mathbb{R}_{\geq 0}$ the set of positive integers, real, and nonnegative real numbers, respectively. Given $N \in \mathbb{Z}_{>0}$, we let $[N] = \{1, 2, \dots, N\}$. For $N_1, N_2 \in \mathbb{Z}$, we let $[N_1 : N_2] = \{N_1 + 1, N_1 + 2, \dots, N_2\}$. For $x \in \mathbb{R}^n$, $\|x\|$ denotes its Euclidean norm, and for $l \in [n]$, $x^{(l)}$ is its l -th component. Given a set $\mathcal{C} \subset \mathbb{R}^n$, $\mathbb{1}_{\mathcal{C}}$ is the indicator function of \mathcal{C} , which is such that $\mathbb{1}_{\mathcal{C}}(x) = 1$ if $x \in \mathcal{C}$

and $\mathbb{1}_{\mathcal{C}}(x) = 0$ otherwise. We let \mathbf{I}_n be the n -dimensional identity matrix. Given a function $V : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we let $\text{Im}(V) = \{V(\theta) \in \mathbb{R}^m : \theta \in \mathbb{R}^n\}$ denote its image. Given a random variable X taking scalar values, $X \sim \eta$ indicates X is distributed according to a probability distribution η , $\mathbb{E}[X]$ denotes its expectation, and $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$ its variance. Given a set S , $P(S)$ denotes its power set, i.e., the collection of all subsets of S . The collection $\Sigma \subset P(S)$ is a σ -algebra if and only if: (i) S is in Σ , (ii) if $A \in \Sigma$, the complement of A is also in Σ , (iii) if $\{A_i\}_{i \in \mathbb{Z}_{>0}}$ is a countable union of sets in Σ , then $\bigcup_{i \in \mathbb{Z}_{>0}} A_i \in \Sigma$.

Stability of Dynamical Systems: We recall here concepts on stability of dynamical systems following [21]. Let $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a locally Lipschitz vector field and consider the dynamical system $\dot{z} = F(z)$. Local Lipschitzness ensures that for every initial condition $x \in \mathbb{R}^n$ there exists $T > 0$ and a unique trajectory $z : [0, T] \rightarrow \mathbb{R}^n$ such that $z(0) = x$ and $\dot{z}(t) = F(z(t))$. If the solution is defined for all $t \geq 0$, then it is *forward complete*. If every solution is forward complete, for each $t \geq 0$, the *flow map* is defined by the function $\Phi_t : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\Phi_t(x) = z(t)$. A set $\mathcal{K} \subset \mathbb{R}^n$ is forward invariant if, for every initial condition $x \in \mathcal{K}$, the trajectory with initial condition at x is forward complete and $\Phi_t(x) \in \mathcal{K}$ for all $t \geq 0$.

Constrained Markov Decision Processes: Here we recall concepts on Constrained Markov Decision Processes (CMDP) following [6], [22]. A CMDP is given by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R_0, \{R_j\}_{j=1}^q)$, with $q \in \mathbb{Z}_{>0}$. Here, \mathcal{S} is a set of states, \mathcal{A} is a set of actions, and $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is a probability transition function, where $P(s, a, s')$ represents the probability that the agent transitions to state $s' \in \mathcal{S}$ given that it is at state $s \in \mathcal{S}$ and takes action $a \in \mathcal{A}$. Further, $R_0 : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ and $R_j : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ for $j \in [q]$ are functions: for every $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $s' \in \mathcal{S}$, $R_0(s, a, s')$ is the reward associated with completing a task when an agent is at state s , takes action a , and transitions to state s' . Instead, $R_j(s, a, s')$ is the cost associated with a safety constraint when an agent is at state s , takes action a , and transitions to state s' . A policy π for the CMDP is a function that maps every state $s \in \mathcal{S}$ to a distribution over \mathcal{A} , denoted as $\pi(\cdot|s)$: here, $\pi(a|s)$ is the probability of taking action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$.

Constraint Qualifications in Nonlinear Programming: We summarize here various constraint qualification conditions from nonlinear programming following [23]–[25]. Let $f, g_1, \dots, g_q : \mathbb{R}^d \rightarrow \mathbb{R}$ be differentiable functions, and consider a nonlinear optimization problem of the form

$$\begin{aligned} \min_{x \in \mathbb{R}^d} & f(x) \\ \text{s.t. } & g_j(x) \leq 0, \quad j \in [q], \end{aligned} \quad (1)$$

where f, g_1, \dots, g_q are continuously differentiable. Let the active and inactive constraint sets be defined by $I_0(x) = \{j \in [q] : g_j(x) = 0\}$ and $I_-(x) = \{j \in [q] : g_j(x) < 0\}$. Then,

- Slater's condition (SC) holds for (1) if there exists $x \in \mathbb{R}^d$ such that $g_j(x) < 0$ for all $j \in [q]$;
- the Mangasarian-Fromovitz Constraint Qualification (MFCQ) holds for (1) at $x \in \mathbb{R}^d$ if there exist $\xi \in \mathbb{R}^d$ such that $\nabla g_j(x)^\top \xi < 0$ for all $j \in I_0(x)$;

- the constant-rank condition (CRC) holds for (1) at $x \in \mathbb{R}^d$ if there exists a neighborhood \mathcal{N} of x such that for all $I \subset I_0(x)$, $\text{rank}(\{\nabla g_j(\bar{x})\}_{j \in I})$ is constant for all $\bar{x} \in \mathcal{N}$. If x^* is a local minimizer of (1), and MFCQ or CRC hold at x^* , then there exist $u^* \in \mathbb{R}^q$ (which we refer to as a Lagrange multiplier vector) such that the *Karush-Kuhn-Tucker* (KKT) conditions hold:

$$\nabla f(x^*) + \sum_{i=1}^q u_i^* \nabla g_i(x^*) = 0, \quad (2a)$$

$$g_j(x^*) \leq 0, \quad u_j^* \geq 0, \quad u_j^* g_j(x^*) = 0, \quad j \in [q]. \quad (2b)$$

If (1) is convex, x^* is a local minimizer, and SC holds, then the KKT conditions (2) also hold. Any $x^* \in \mathbb{R}^d$ for which there exists $u^* \in \mathbb{R}^q$ satisfying (2) is referred to as a KKT point of (1). We note that u_j^* is the Lagrange multiplier associated with the j -th constraint.

Given differentiable functions $\tilde{f}, \tilde{g}_1, \dots, \tilde{g}_q : \mathbb{R}^d \times \mathbb{R}^c \rightarrow \mathbb{R}$, consider the parametric nonlinear optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \tilde{f}(x, z) \\ \text{s.t.} \quad & \tilde{g}_j(x, z) \leq 0, \quad j \in [q]. \end{aligned} \quad (3)$$

Let $\tilde{I}_0(x, z) = \{j \in [q] : g_j(x, z) = 0\}$ be the set of active constraints. We say that the constant-rank condition (CRC) holds for (3) at $(x_0, z_0) \in \mathbb{R}^d \times \mathbb{R}^c$ if there exists a neighborhood \mathcal{N} of (x_0, z_0) such that for any $\tilde{I} \subset \tilde{I}_0(x_0, z_0)$ and $(x, z) \in \mathcal{N}$, $\{\nabla_x g_j(x, z)\}_{j \in \tilde{I}}$ has a constant rank.

III. PROBLEM STATEMENT

In this section we formalize the problem of solving constrained reinforcement learning (RL) problems in an anytime fashion. Given a CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R_0, \{R_j\}_{j=1}^q)$, the goal is to maximize the cumulative reward while keeping the cumulative costs below a threshold. We consider a parametric class of policies indexed by a vector $\theta \in \mathbb{R}^d$. We denote the policy associated with θ as π_θ . Given a distribution η of initial states, a discount factor $\gamma \in (0, 1)$, and a time horizon $T \in \mathbb{Z}_{>0}$, we consider the following problem:

$$\begin{aligned} \min_{\theta \in \mathbb{R}^d} \quad & V_0(\theta) = \mathbb{E}_{\substack{a \sim \pi_\theta(\cdot|s) \\ s_0 \sim \eta}} \left[\sum_{k=0}^{T-1} -\gamma^k R_0(s_k, a_k, s_{k+1}) \right] \\ \text{s.t.} \quad & V_j(\theta) = \mathbb{E}_{\substack{a \sim \pi_\theta(\cdot|s) \\ s_0 \sim \eta}} \left[\sum_{k=0}^{T-1} \gamma^k R_j(s_k, a_k, s_{k+1}) \right] \leq 0, \quad j \in [q]. \end{aligned} \quad (4)$$

Problem (4) seeks to find the policy π_θ that maximizes the expected cumulative reward given by R_0 (for convenience, we have changed the sign of R_0 to turn (4) into a minimization problem) over T time steps and also maintains the expected cumulative costs given by R_j for all $j \in [q]$ over T time steps below zero. Throughout the paper, we refer to V_0, \dots, V_q as *value functions*. The discount factor γ determines how much future rewards are valued compared to immediate rewards.

Remark III.1. (Ensuring safety of state trajectories): Throughout the paper, the notion of *safety* refers to the satisfaction of the constraints in (4), and therefore pertains the policy parameter θ . Interestingly, with an appropriate selection

of the cost function R_j , this safety guarantee implies the forward invariance of a desired set $\mathcal{C}_j \subset \mathcal{S}$ with a prescribed confidence. In fact, let

$$R_j(s_t) = 1 - \mathbb{1}_{\mathcal{C}_j}(s_t) + \frac{\gamma^T \delta_j}{\sum_{t=0}^{T-1} \gamma^t},$$

where $0 < \delta_j < 1$, for all $j \in [q]$, are prescribed confidence levels. According to [10, Theorems 1 and 2], the satisfaction of the cumulative constraints in (4) implies that

$$\mathbb{P}\left(\bigcap_{t=0}^{T-1} \{s_t \in \mathcal{C}_j\}\right) \geq 1 - \delta_j, \quad \forall j \in [q],$$

i.e., the probability that the states remain within \mathcal{C}_j in the next T timesteps is at least $1 - \delta_j$. •

The functions $\{V_i\}_{i=0}^q$ are in general non-convex, and this makes solving (4) NP-hard. Therefore, we aim to find local minimizers (or, more generally, KKT points) of (4). Additionally, because of their definition, the values of V_0, \dots, V_q and their gradients at arbitrary $\theta \in \mathbb{R}^d$ are not readily available, and instead need to be estimated through episodic data (i.e., trajectories generated by the policy π_θ) of the CMDP.

Formally, we seek to solve the following problem.

Problem 1. *Develop an RL algorithm that,*

- *converges to a KKT point of (4);*
- *is anytime, meaning that at every iteration, the constraints of (4) are satisfied.*

Due to the probabilistic nature of the CMDP dynamics, Problem 1 can only be solved in a probabilistic sense, i.e., given a finite number of available episodes, one can only expect to obtain convergence and constraint satisfaction results that hold in probability. As the number of available episodes grows, one can also expect that the convergence and constraint satisfaction guarantees hold with arbitrarily high probability.

IV. THE ROBUST SAFE GRADIENT FLOW

Here we describe the Robust Safe Gradient Flow (RSGF), a continuous-time anytime algorithm for constrained optimization that is a variation of the *Safe Gradient Flow* [26], [27]. We later rely on the RSGF to design our solution to Problem 1. Even though our proposed RL algorithm will eventually be defined in discrete time, the properties of the continuous-time flow established here are key, as we will leverage them using the theory of stochastic approximation, cf. [19], [20].

Let $V_0, \dots, V_q : \mathbb{R}^d \rightarrow \mathbb{R}$ be continuously differentiable functions and consider the constrained optimization problem

$$\begin{aligned} \min_{\theta \in \mathbb{R}^d} \quad & V_0(\theta) \\ \text{s.t.} \quad & V_j(\theta) \leq 0, \quad j \in [\tilde{q}]. \end{aligned} \quad (5)$$

We let $\mathcal{C} = \{\theta \in \mathbb{R}^d : V_j(\theta) \leq 0, \forall j \in [\tilde{q}]\}$ denote the feasible set. Given $\alpha > 0$ and a continuously differentiable function $\beta : \mathbb{R}^d \rightarrow \mathbb{R}_{>0}$, let $\mathcal{R}_{\alpha, \beta} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be defined by

$$\begin{aligned} \mathcal{R}_{\alpha, \beta}(\theta) = \arg \min_{\xi \in \mathbb{R}^d} \quad & \frac{1}{2} \|\xi + \nabla V_0(\theta)\|^2 \\ \text{s.t.} \quad & \alpha V_j(\theta) + \nabla V_j(\theta)^\top \xi + \frac{\beta(\theta)}{2} \|\xi\|^2 \leq 0, \quad j \in [\tilde{q}]. \end{aligned} \quad (6)$$

We note that if $\beta \equiv 0$, this definition recovers the Safe Gradient Flow [26]. We study the properties of the flow

$$\dot{\theta} = \mathcal{R}_{\alpha,\beta}(\theta), \quad (7)$$

which we refer to as the Robust Safe Gradient Flow (RSGF). In particular, we seek to determine conditions under which the dynamics is well-posed and characterize the transient and asymptotic behavior of its trajectories.

A. Well-Posedness and Regularity Properties

We start by introducing some regularity and constraint qualification assumptions regarding the optimization problem (5).

Assumption 1. (Regularity): *The functions $V_0, \dots, V_{\bar{q}} : \mathbb{R}^d \rightarrow \mathbb{R}$, and $\beta : \mathbb{R} \rightarrow \mathbb{R}$ are twice continuously differentiable.*

Assumption 2. (Constraint qualifications in the feasible set): *For all $\theta \in \mathcal{C}$, (5) satisfies MFCQ. Additionally, for each $\theta \in \mathcal{C}$, the parametric problem (6) satisfies CRC at $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$.*

Assumption 3. (Constraint qualifications outside the feasible set): *For all $\theta \in \mathbb{R}^d \setminus \mathcal{C}$, Slater's condition holds for (6) and the parametric problem (6) satisfies CRC at $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$.*

Assumption 1 is standard in the literature [28] and is satisfied by considering smooth policies π_θ . MFCQ and CRC in Assumptions 2, 3 are standard constraint qualification conditions for constrained optimization problems such as (5) and (6), and ensure that $\mathcal{R}_{\alpha,\beta}$ enjoys good regularity properties, as we establish in the sequel. Lemma A.2 provides conditions under which Slater's condition holds for (6) for each $\theta \in \mathbb{R}^d \setminus \mathcal{C}$, and Lemma A.3 provides conditions under which CRC holds for (6) at $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$ for some $\theta \in \mathcal{C}$.

The next result provides a closed-form expression for $\mathcal{R}_{\alpha,\beta}$ in terms of the Lagrange multipliers of (6).

Lemma IV.1. (Alternative expression for RSGF): *Let $u_j : \mathbb{R}^d \rightarrow \mathbb{R}$ map $\theta \in \mathbb{R}^d$ to the Lagrange multiplier associated with the j -th constraint of (6). If MFCQ holds for (5) at $\theta \in \mathcal{C}$,*

$$\mathcal{R}_{\alpha,\beta}(\theta) = - \frac{\nabla V_0(\theta) + \sum_{j=1}^{\bar{q}} u_j(\theta) \nabla V_j(\theta)}{1 + \beta(\theta) \sum_{j=1}^{\bar{q}} u_j(\theta)}. \quad (8)$$

Proof. Note that since $\theta \in \mathcal{C}$, $[\tilde{q}] = I_0(\theta) \cup I_-(\theta)$. Since MFCQ holds for (5) at θ , there exists $\xi \in \mathbb{R}^d$ such that $\nabla V_j(\theta)^\top \xi < 0$ for all $j \in I_0(\theta)$. Hence, by taking $\epsilon_j < \frac{2|\nabla V_j(\theta)^\top \xi|}{\beta(\theta)\|\xi\|^2}$ and $\hat{\xi} = \epsilon \xi$ with $\epsilon \in (0, \min_{j \in I_0(\theta)} \epsilon_j)$,

$$\alpha V_j(\theta) + \nabla V_j(\theta)^\top \hat{\xi} + \frac{\beta(\theta)}{2} \|\hat{\xi}\|^2 < 0, \quad \forall j \in I_0(\theta).$$

On the other hand, for every $j \in I_-(\theta)$, let ϵ_j be sufficiently small so that $\alpha V_j(\theta) + \epsilon_j \nabla V_j(\theta)^\top \hat{\xi} + \epsilon_j^2 \frac{\beta(\theta)}{2} \|\hat{\xi}\|^2 < 0$. Now, taking $\epsilon \in (0, \min_{j \in [\bar{q}]} \epsilon_j)$ and $\tilde{\xi} = \epsilon \hat{\xi}$, we conclude $\alpha V_j(\theta) + \nabla V_j(\theta)^\top \tilde{\xi} + \frac{\beta(\theta)}{2} \|\tilde{\xi}\|^2 < 0$, for all $j \in [\bar{q}]$, and hence Slater's condition holds for (6). Since (6) is convex, this means that $\mathcal{R}_{\alpha,\beta}$ satisfies the KKT equations associated to (6). Hence,

$$\mathcal{R}_{\alpha,\beta}(\theta) + \nabla V_0(\theta) + \sum_{j=1}^{\bar{q}} u_j(\theta) \left(\nabla V_j(\theta) + \beta(\theta) \mathcal{R}_{\alpha,\beta}(\theta) \right) = 0,$$

from where the expression (8) follows. \square

The next result provides conditions under which (6) is feasible and locally Lipschitz.

Lemma IV.2. (Feasibility and Lipschitzness): *Suppose Assumption 1 holds. Then,*

- (i) *under Assumption 2, $\mathcal{R}_{\alpha,\beta}$ is well-defined and locally Lipschitz on an open neighborhood containing \mathcal{C} ;*
- (ii) *under Assumptions 2 and 3, $\mathcal{R}_{\alpha,\beta}$ is well-defined and locally Lipschitz on \mathbb{R}^d .*

Proof. (i): By the argument employed in the proof of Lemma IV.1, Slater's condition holds for (6) at any $\theta \in \mathcal{C}$. This means that there exists $\xi \in \mathbb{R}^d$ such that $\alpha V_j(\theta) + \nabla V_j(\theta)^\top \xi + \frac{\beta(\theta)}{2} \|\xi\|^2 < 0$ for all $j \in [q]$. Since $V_j, \nabla V_j$ and β are continuous, there exists a neighborhood U_θ of θ such that $\alpha V_j(\bar{\theta}) + \nabla V_j(\bar{\theta})^\top \xi + \frac{\beta(\bar{\theta})}{2} \|\xi\|^2 < 0$ for all $\bar{\theta} \in U_\theta$. In particular, the constraints in the definition of $\mathcal{R}_{\alpha,\beta}$ are feasible at all points in U_θ and hence $\mathcal{R}_{\alpha,\beta}$ is well-defined at all points in U_θ . Hence $\mathcal{R}_{\alpha,\beta}$ is well-defined in the open set $\bigcup_{\theta \in \mathcal{C}} U_\theta$ containing \mathcal{C} . Since SC implies MFCQ for convex problems [29, Proposition 5.39], the functions $V_0, \dots, V_{\bar{q}}$, and β are twice continuously differentiable, and for each $\theta \in \mathcal{C}$, (6) satisfies CRC at $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$, $\mathcal{R}_{\alpha,\beta}$ is locally Lipschitz on an open neighborhood of \mathcal{C} , invoking [25, Theorem 3.6].

(ii): by assumption, for any $\theta \in \mathbb{R}^d$, Slater's condition holds for (6) and CRC holds at $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$ for (6). Hence, by [25, Theorem 3.6], $\mathcal{R}_{\alpha,\beta}$ is locally Lipschitz at θ . \square

The local Lipschitzness of $\mathcal{R}_{\alpha,\beta}$ on a neighborhood of \mathcal{C} , cf. Lemma IV.2(i) (resp., in all of \mathbb{R}^d , cf. Lemma IV.2(ii)) ensures (7) is well-defined and has a unique solution for any initial condition in a neighborhood of \mathcal{C} (resp., in all of \mathbb{R}^d). We refer to [30] for other conditions that guarantee local Lipschitzness of parametric optimization problems such as (6).

B. Equilibria, Forward Invariance, and Stability

Next we establish the equivalence between the equilibrium points of (7) and the KKT points of (5).

Proposition IV.3. (Equivalence between equilibria and KKT points): *Let (6) be feasible at $\theta^* \in \mathbb{R}^d$. If $\mathcal{R}_{\alpha,\beta}(\theta^*) = 0$, then $\theta^* \in \mathcal{C}$. If MFCQ holds for (5) at $\theta^* \in \mathbb{R}^d$, then $\mathcal{R}_{\alpha,\beta}(\theta^*) = 0$ if and only if θ^* is a KKT point of (5).*

Proof. If (6) is feasible at $\theta^* \in \mathbb{R}^d$ and $\mathcal{R}_{\alpha,\beta}(\theta^*) = 0$, then $\alpha V_j(\theta^*) + \nabla V_j(\theta^*)^\top \mathcal{R}_{\alpha,\beta}(\theta^*) + \frac{\beta(\theta^*)}{2} \|\mathcal{R}_{\alpha,\beta}(\theta^*)\|^2 = \alpha V_j(\theta^*) \leq 0$, for all $j \in [\bar{q}]$, and therefore $\theta^* \in \mathcal{C}$. Next, suppose MFCQ holds for (5) at $\theta^* \in \mathbb{R}^d$ and $\mathcal{R}_{\alpha,\beta}(\theta^*) = 0$. As shown in the proof of Lemma IV.1(i), Slater's condition holds for (6). Hence, since $\mathcal{R}_{\alpha,\beta}$ is the local minimizer, it satisfies the KKT equations for (6). Enforcing that the solution is $\xi = 0$, these read exactly as the KKT equations for (5). Since MFCQ holds for (5), it follows that θ^* is a KKT point of (5). Conversely, if θ^* is a KKT point of (5), then there exist a Lagrange multiplier vector $u \in \mathbb{R}^{\bar{q}}$ satisfying the KKT equations. Since the solution of (6) is unique because the problem is strongly convex, we conclude $\mathcal{R}_{\alpha,\beta}(\theta^*) = 0$. \square

The next result shows that \mathcal{C} is forward invariant under (7).

Proposition IV.4. (Safety of RSGF): *Suppose Assumptions 1 and 2 hold. Then, \mathcal{C} is forward invariant under (7).*

Proof. By Lemma IV.2(i), every solution of (7) with initial condition in \mathcal{C} is unique and well-defined as long as it stays in a neighborhood of \mathcal{C} . Due to the constraints in (6),

$$\nabla V_j(\theta)^\top \mathcal{R}_{\alpha,\beta}(\theta) \leq -\alpha V_j(\theta) - \frac{\beta(\theta)}{2} \|\mathcal{R}_{\alpha,\beta}(\theta)\|^2, \quad (9)$$

and hence $j \in [\tilde{q}]$, $\nabla V_j(\theta)^\top \mathcal{R}_{\alpha,\beta}(\theta) \leq 0$ whenever $V_j(\theta) = 0$. The result then follows from Nagumo's Theorem [31]. \square

The final result of this section characterizes the convergence properties of (7).

Proposition IV.5. (Convergence of RSGF): *Suppose Assumptions 1 and 2 hold. Then,*

- (i) *every bounded trajectory of (7) starting in \mathcal{C} converges to the set of KKT points of (5).*
- (ii) *if Assumption 3 holds, then every bounded trajectory of (7) converges to the set of KKT points of (5).*

In either case, if every KKT point is isolated, convergence is to a point.

Proof. (i): From the proof of Lemma IV.1(i), we have that $\mathcal{R}_{\alpha,\beta}(\theta)$ satisfies the KKT equations for (6),

$$\begin{aligned} u_j(\theta)(\alpha V_j(\theta) + \nabla V_j(\theta)^\top \mathcal{R}_{\alpha,\beta}(\theta) + \frac{\beta(\theta)}{2} \|\mathcal{R}_{\alpha,\beta}(\theta)\|^2) &= 0, \\ u_j(\theta) \geq 0, \quad \alpha V_j(\theta) + \nabla V_j(\theta)^\top \mathcal{R}_{\alpha,\beta}(\theta) + \frac{\beta(\theta)}{2} \|\mathcal{R}_{\alpha,\beta}(\theta)\|^2 &\leq 0. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{d}{dt} V_0(\theta) &= \nabla V_0(\theta)^\top \mathcal{R}_{\alpha,\beta}(\theta) = \\ &= -\mathcal{R}_{\alpha,\beta}(\theta)^\top \left((1 + \beta(\theta) \sum_{j=1}^q u_j(\theta)) \mathcal{R}_{\alpha,\beta}(\theta) + \sum_{j=1}^q u_j(\theta) \nabla V_j(\theta) \right) \\ &= -(1 + \frac{\beta(\theta)}{2} \sum_{j=1}^q u_j(\theta)) \|\mathcal{R}_{\alpha,\beta}(\theta)\|^2 + \sum_{j=1}^q \alpha u_j(\theta) V_j(\theta), \end{aligned} \quad (10)$$

where in the second equality we have used (8) and in the third we have used the KKT equations above. Now, since $u_j(\theta) \geq 0$ for all $j \in [q]$, and $V_j(\theta) \leq 0$ for all $j \in [q]$ if $\theta \in \mathcal{C}$, we deduce that $\frac{d}{dt} V_0(\theta) \leq 0$ for all $\theta \in \mathcal{C}$, with equality if and only if θ is a KKT point of (5) by Proposition IV.3. The fact that all bounded trajectories converge to the set of KKT points follows then from [32, Proposition 5.3] using V_0 as a LaSalle function. Convergence to a point when the KKT points are isolated follows from [32, Corollary 5.2].

(ii): Let $\epsilon > 0$ and define $V_{\epsilon_*} : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$V_{\epsilon_*}(\theta) = V_0(\theta) + \frac{1}{\epsilon_*} \sum_{j=1}^q [V_j(\theta)]_+.$$

From [33, Proposition 3], V_{ϵ_*} is directionally differentiable and its directional derivative in the direction $\xi \in \mathbb{R}^n$ is

$$\begin{aligned} V'_{\epsilon_*}(\theta; \xi) &= \nabla V_0(\theta)^\top \xi + \frac{1}{\epsilon_*} \sum_{j \in I_+(\theta)} \nabla V_j(\theta)^\top \xi \\ &\quad + \frac{1}{\epsilon_*} \sum_{j \in I_0(\theta)} [\nabla V_j(\theta)^\top \xi]_+, \end{aligned} \quad (11)$$

where $I_0(\theta)$ and $I_+(\theta)$ correspond to the optimization problem (5). From the KKT equations above, we have that $\nabla V_j(\theta)^\top \mathcal{R}_{\alpha,\beta}(\theta) \leq -\alpha V_j(\theta)$ for all $j \in [q]$. Using (10) in (11) for $\xi = \mathcal{R}_{\alpha,\beta}(\theta)$, we have

$$\begin{aligned} V'_{\epsilon_*}(\theta; \mathcal{R}_{\alpha,\beta}(\theta)) &\leq -(1 + \frac{\beta(\theta)}{2} \sum_{j=1}^q u_j(\theta)) \|\mathcal{R}_{\alpha,\beta}(\theta)\|^2 \\ &\quad + \sum_{j=1}^q \alpha u_j(\theta) V_j(\theta) - \frac{1}{\epsilon_*} \sum_{j \in I_+(\theta)} \alpha V_j(\theta). \end{aligned}$$

Now, by an argument analogous to [26, Lemma D.1], for any compact set Ω , there exists $B_\Omega > 0$ such that $u_j(\theta) \leq B_\Omega$ for all $j \in [\tilde{q}]$ and $\theta \in \Omega$. Then, for $\epsilon_* \in (0, \frac{1}{B_\Omega})$, and since $u_j(\theta) V_j(\theta) \leq 0$ for all $j \in I_0(\theta) \cup I_-(\theta)$, we have

$$V'_{\epsilon_*}(\theta; \mathcal{R}_{\alpha,\beta}(\theta)) \leq -(1 + \frac{\beta(\theta)}{2} \sum_{j=1}^q u_j(\theta)) \|\mathcal{R}_{\alpha,\beta}(\theta)\|^2 \leq 0,$$

for all $\theta \in \Omega$, where the last inequality is an equality if and only if θ is a KKT point of (5) (cf. Proposition IV.3). Now the fact that all bounded trajectories in Ω converge to the set of KKT points of (5) follows from [32, Proposition 5.3] using V_{ϵ_*} as a LaSalle function. Convergence to a point when the KKT points are isolated follows from [32, Corollary 5.2]. \square

Remark IV.6. (Boundedness of trajectories): Regarding Proposition IV.5(i), note that all trajectories of (7) starting in \mathcal{C} remain in it by Proposition IV.4, and hence are bounded if this set is compact. Regarding Proposition IV.5(ii), if there is $i_* \in [\tilde{q}]$ and c such that $\Gamma = \{\theta \in \mathbb{R}^d : V_{i_*}(\theta) \leq c\}$ is compact, note that (9) implies that this set is forward invariant under (7). Therefore, all trajectories of (7) starting in Γ are bounded. In particular, this holds if V_{i_*} is radially unbounded, since all its sublevel sets are compact. \bullet

Remark IV.7. (Robustness to error): The introduction of the strictly positive term β in the definition (6) strengthens the robustness against errors and disturbances of the robust safe gradient flow (as compared, for instance, with the safe gradient flow [26], which corresponds to $\beta \equiv 0$). An indication of this fact can be observed, for instance, in the contributions of the β term to the decrease of the LaSalle functions in the proof of Proposition IV.5. We quantify more precisely this robustness to model errors in Section VI and exploit it to handle imperfect knowledge of the functions $\{V_j\}_{j=0}^{\tilde{q}}$ and their gradients $\{\nabla V_j\}_{j=0}^{\tilde{q}}$ in the algorithm implementation. \bullet

Remark IV.8. (Discretization): We note that the forward-Euler discretization of (7) is equivalent to the discrete-time dynamics introduced in [1]. This follows by performing a change of variables ($\xi = \frac{y-\theta}{h}$ in the optimization problem (2) in [1], with the variables y and h as defined therein). This discrete-time dynamics is a special case of the Moving Balls Algorithm (MBA) [34]. Both [1] and [34] study the safety and convergence properties of the discrete-time dynamics directly, instead of their continuous-time counterpart (7), as we have done here. We leverage the latter in what follows using the theory of stochastic approximation [19], [20]. \bullet

V. ROBUST SAFE GRADIENT FLOW-BASED REINFORCEMENT LEARNING

In this section, we introduce our algorithmic solution to Problem 1. Consider the optimization problem (4) defining the optimal policy for the CMDP \mathcal{M} . Instead of dealing directly with (4), we consider (5) with $\tilde{q} = q + 1$, and include the additional function $V_{q+1}(\theta) = \|\theta\|^2 - C$, where $C > 0$ is a design parameter. As we justify later, this has the effect of keeping the iterates of the algorithm bounded.

Given $\alpha > 0$ and $\beta : \mathbb{R}^d \rightarrow \mathbb{R}_{>0}$, let $\mathcal{R}_{\alpha,\beta} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be defined by (6). To solve Problem 1, consider the forward-Euler discretization of the RSGF (7),

$$\theta_{i+1} = \theta_i + h_i \mathcal{R}_{\alpha,\beta}(\theta_i), \quad (12)$$

where $\{h_i\}_{i \in \mathbb{Z}_{>0}}$ is a sequence of stepsizes. Note that, since closed-form expressions for the value functions V_0, \dots, V_q are not readily available, one cannot directly implement this iteration. Instead, our strategy consists of relying on the robustness properties of (12), when viewed as a discrete-time dynamical system, and employing estimates of V_1, \dots, V_q , and $\nabla V_0, \dots, \nabla V_q$ constructed with episodic data, as detailed next (note that V_{q+1} and ∇V_{q+1} are known).

Episodic data available: Let Λ be a given set of policies for \mathcal{M} and \mathcal{I}_0 a batch of episodes obtained offline with policies from Λ . Formally,

$$\mathcal{I}_0 = \{[s_0^n, a_0^n, s_1^n, a_1^n, \dots, s_T^n, a_T^n, s_{T+1}^n]\}_{n \in [N_\zeta], \zeta \in \Lambda},$$

where N_ζ is the number of episodes obtained with policy ζ . Given $i \in \mathbb{Z}_{>0}$, let \mathcal{I}_i be the collection of episodes at iteration i obtained using policy π_{θ_i} (with $N_i = |\mathcal{I}_i|$ its number).

At iteration i , we construct the estimates of the value functions and their gradients using episodes from $\cup_{j=0}^i \mathcal{I}_j$ as follows. Although one could potentially use all such episodes, for flexibility we assume that we only use a subset $\mathcal{J}_i \subset \cup_{j=0}^i \mathcal{I}_j$. We enumerate the episodes in \mathcal{J}_i as

$$\mathcal{J}_i = \{[s_0^n, a_0^n, s_1^n, a_1^n, \dots, s_T^n, a_T^n, s_{T+1}^n]\}_{n=1}^{|\mathcal{J}_i|}.$$

For each $n \in [|\mathcal{J}_i|]$, we denote by ζ_n the policy utilized to obtain the corresponding episode.

Assumption 4. *There exists $\nu > 0$ such that, for any $a \in \mathcal{A}$, $s \in \mathcal{S}$, $\theta \in \mathbb{R}^d$ and $\zeta \in \Lambda$, we have $\pi_\theta(a|s) > \nu$, $\zeta(a|s) > \nu$.*

Assumption 4 is standard in the context of importance-sampling methods in RL [35], [36]. For any given state, it requires that any action has a positive probability lower bounded by ν for any policy in the parametric family $\{\pi_\theta\}$ as well as in Λ .

Estimates of value functions and their gradients: For each $j \in [q] \cup \{0\}$, we consider the following estimate of the value function at iteration i ,

$$\widehat{V}_j(\theta_i) = \frac{\sigma_j}{|\mathcal{J}_i|} \left(\sum_{n=1}^{|\mathcal{J}_i|} \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right), \quad (13)$$

where $\sigma_0 = -1$, and $\sigma_j = 1$ for $j \in [q]$. Under Assumption 4, $\widehat{V}_j(\theta_i)$ is well defined, because the denominator in the ratio $\frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)}$ is strictly positive.

For any $a \in \mathcal{A}$ and $s \in \mathcal{S}$, define $\chi_{a,s} : \mathbb{R}^d \rightarrow \mathbb{R}$ as $\chi_{a,s}(\theta) = \log \pi_\theta(a|s)$. Note that $\chi_{a,s}$ is well-defined for all $\theta \in \mathbb{R}^d$ under Assumption 4. Let $b : \mathcal{S} \rightarrow \mathbb{R}$ be a baseline function whose absolute value is bounded by $\hat{B} > 0$. For each $j \in [q] \cup \{0\}$, we consider the following estimates of the gradients of the value functions at iteration i ,

$$\widehat{\nabla V}_j(\theta_i) = \frac{\sigma_j}{|\mathcal{J}_i|} \left(\sum_{n=1}^{|\mathcal{J}_i|} \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \sum_{t=0}^T \gamma^t \nabla \chi_{a_t^n, s_t^n}(\theta_i) D_{j,t}^n \right). \quad (14a)$$

where

$$D_{j,t}^n = \sum_{t'=t}^T \gamma^{t'-t} R_j(s_{t'}^n, a_{t'}^n, s_{t'+1}^n) - b(s_t^n). \quad (14b)$$

Under Assumption 4, $\widehat{\nabla V}_j(\theta_i)$ is well defined. Given these estimates, we define an approximated version of (6) as follows:

$$\hat{\mathcal{R}}_{\alpha,\beta}(\theta) = \arg \min_{\xi \in \mathbb{R}^d} \frac{1}{2} \|\xi + \widehat{\nabla V}_0(\theta)\|^2 \quad (15a)$$

$$\text{s.t. } \alpha \widehat{V}_j(\theta) + \widehat{\nabla V}_j(\theta)^\top \xi + \frac{\beta(\theta)}{2} \|\xi\|^2 \leq 0, \quad j \in [q], \quad (15b)$$

$$\alpha V_{q+1}(\theta) + \nabla V_{q+1}(\theta)^\top \xi + \frac{\beta(\theta)}{2} \|\xi\|^2 \leq 0. \quad (15c)$$

Note that this can be computed with the episodic data available to the agent.

Algorithm 1 presents the pseudocode for our proposal to solve Problem 1. We refer to it as Robust Safe Gradient Flow-based Reinforcement Learning (RSGF-RL).

Algorithm 1 RSGF-RL

- 1: **Parameters:** $\alpha, \beta, k, m, \{h_i\}_{i=1}^k, T, \gamma, \mathcal{I}_0, \{N_i\}_{i=1}^k$
- 2: **Initial Policy Parameter:** θ_1
- 3: **for** $i \in [k]$ **do**
- 4: Generate N_i episodes of length $T + 1$ using π_{θ_i}
- 5: Select the set \mathcal{J}_i of episodes at iteration i
- 6: Compute estimates $\{\widehat{V}_j(\theta_i)\}_{j=0}^q$ using (13)
- 7: Compute estimates $\{\widehat{\nabla V}_j(\theta_i)\}_{j=0}^q$ using (14)
- 8: Update policy according to

$$\theta_{i+1} = \theta_i + h_i \hat{\mathcal{R}}_{\alpha,\beta}(\theta_i) \quad (16)$$

9: **end for**

10: **return** θ_{k+1}

In Algorithm 1, we do not detail a specific scheme to select the sets of episodes \mathcal{J}_i from the available ones in $\cup_{j=0}^i \mathcal{I}_j$. Instead, in what follows, we study the properties of RSGF-RL for arbitrary sets \mathcal{J}_i and provide conditions on these sets that guarantee a desired level of algorithmic performance.

VI. ANYTIME SAFETY AND CONVERGENCE GUARANTEES OF RSGF-RL

In this section we present our technical analysis of RSGF-RL. We start by establishing different statistical properties of the value function and gradient estimates, and then characterize the safety and convergence properties of RSGF-RL.

A. Statistical Properties of Estimates

Here, we establish the statistical properties of the estimates (13) and (14) of the value functions and their gradients, resp. In our analysis, we make the following assumptions.

Assumption 5. (Boundedness of reward functions): *For each $j \in [q] \cup \{0\}$, there exist $B_j > 0$ such that $|R_j(s, a, s')| < B_j$, for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $s' \in \mathcal{S}$.*

Assumption 6. (Differentiability and Lipschitzness of policy): *The function $\chi_{a,s}$ is continuously differentiable and there exist $L > 0$ and $\tilde{B} > 0$ such that*

$$\begin{aligned} \|\nabla \chi_{a,s}(\theta) - \nabla \chi_{a,s}(\bar{\theta})\| &\leq L \|\theta - \bar{\theta}\|, \\ \forall \theta, \bar{\theta} \in \mathbb{R}^d, a \in \mathcal{A}, s \in \mathcal{S}, \\ \|\nabla \chi_{a,s}(\theta)^{(l)}\| &\leq \tilde{B}, \forall \theta \in \mathbb{R}^d, l \in [d], a \in \mathcal{A}, s \in \mathcal{S}. \end{aligned}$$

Assumptions 5 and 6 are standard in the literature, cf. [28], [37]. By the Policy Gradient Theorem [22, Section 13.2], under Assumption 6, the functions $\{V_j\}_{j=0}^q$ in (4) are differentiable. Moreover, Lemma A.1 ensures that, for all $j \in \{0\} \cup [q]$, ∇V_j is globally Lipschitz on \mathbb{R}^d (we denote by L_j its Lipschitz constant). Additionally, we let $L_{q+1} = 2\sqrt{C}$ be the Lipschitz constant of V_{q+1} on $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|^2 \leq C\}$.

In what follows, all expectations, variances, and probabilities are taken with respect to $s_0 \sim \eta$, $a_t^n \sim \zeta_n(\cdot|s_t^n)$, for $t \in [T]$, and $n \in [|\mathcal{J}_i|]$. We next characterize the mean, variance, and tail probabilities of the value function estimates.

Proposition VI.1. (Value function estimates): *Suppose Assumptions 4 and 5 hold. Let $i \in \mathbb{Z}_{>0}$ and assume that \mathcal{J}_i contains \tilde{N}_i episodes generated with π_{θ_i} (without loss of generality, we label them as the first \tilde{N}_i episodes in \mathcal{J}_i). Let*

$$\tilde{N}_i = |\mathcal{J}_i| - \bar{N}_i, \phi_j = \frac{B_j(1 - \gamma^{T+1})}{1 - \gamma}, \bar{\phi}_j = \frac{B_j(1 - \gamma^{T+1})}{(1 - \gamma)\nu^{T+1}}.$$

Then, for $j \in \{0\} \cup [q]$,

- (i) $\mathbb{E}[\widehat{V}_j(\theta_i)] = V_j(\theta_i)$ (unbiased function estimates);
- (ii) $\text{Var}[\widehat{V}_j(\theta_i)] = \frac{\tilde{N}_i \phi_j^2 + \bar{N}_i \bar{\phi}_j^2}{|\mathcal{J}_i|^2}$ and $|\widehat{V}_j(\theta_i)| \leq \frac{\tilde{N}_i \phi_j + \bar{N}_i \bar{\phi}_j}{|\mathcal{J}_i|}$;
- (iii) $\mathbb{P}(|\widehat{V}_j(\theta_i) - V_j(\theta_i)| \leq \epsilon) \geq 1 - 2 \exp\left(-\frac{\epsilon^2 |\mathcal{J}_i|^2}{2\tilde{N}_i \phi_j^2 + 2\bar{N}_i \bar{\phi}_j^2}\right)$.

Further assume that $\chi_{a,s}$ is globally Lipschitz, uniformly in a, s , i.e., there exists $\tilde{L} > 0$ such that

$$|\chi_{a,s}(\theta) - \chi_{a,s}(\theta')| \leq \tilde{L} \|\theta - \theta'\|, \quad (17)$$

for all $\theta, \theta' \in \mathbb{R}^d$, $a \in \mathcal{A}$, and $s \in \mathcal{S}$, and that the policies in Λ belong to $\{\pi_\theta\}$. Let $\{\bar{\theta}_n\}_{n=1}^{|\mathcal{J}_i|}$ denote the parameters that describe all the policies in \mathcal{J}_i and define $\tilde{\phi}_{i,j,n} = \phi_j \exp((T+1)\tilde{L}\|\theta_i - \bar{\theta}_n\|)$. Then,

- (iv) $\text{Var}[\widehat{V}_j(\theta_i)] \leq \frac{\sum_{n=1}^{|\mathcal{J}_i|} \tilde{\phi}_{i,j,n}^2}{|\mathcal{J}_i|^2}$ and $|\widehat{V}_j(\theta_i)| \leq \frac{\sum_{n=1}^{|\mathcal{J}_i|} \tilde{\phi}_{i,j,n}}{|\mathcal{J}_i|}$;
- (v) $\mathbb{P}(|\widehat{V}_j(\theta_i) - V_j(\theta_i)| \leq \epsilon) \geq 1 - 2 \exp\left(-\frac{\epsilon^2 |\mathcal{J}_i|^2}{2 \sum_{n=1}^{|\mathcal{J}_i|} \tilde{\phi}_{i,j,n}^2}\right)$.

Proof. (i): Let $d\Omega = \prod_{n=1}^{|\mathcal{J}_i|} ds_{T+1}^n \prod_{t=0}^T ds_t^n da_t^n$, where ds_t^n and da_t^n are the differential elements associated with the variables s_t^n and a_t^n , respectively. Define, for $j \in \{0\} \cup [q]$,

$$E_{j,n} = \sum_{t'=0}^T \gamma^{t'} R_j(s_{t'}^n, a_{t'}^n, s_{t'+1}^n), \quad \Gamma = \mathcal{S}^{|\mathcal{J}_i|(T+2)} \times \mathcal{A}^{|\mathcal{J}_i|(T+1)},$$

Using (13), we have

$$\begin{aligned} \mathbb{E}[\widehat{V}_j(\theta_i)] &= \frac{\sigma_j}{|\mathcal{J}_i|} \int_{\Gamma} \left(\eta(s_0) \sum_{n=1}^{\bar{N}_i} \prod_{t=0}^T \pi_{\theta_i}(a_t^n | s_t^n) E_{j,n} \right. \\ &\quad \left. + \eta(s_0) \sum_{n=\bar{N}_i+1}^{|\mathcal{J}_i|} \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} E_{j,n} \zeta_n(a_t^n | s_t^n) \right) d\Omega \\ &= \frac{1}{\bar{N}_i + \tilde{N}_i} (\bar{N}_i V_j(\theta_i) + \tilde{N}_i V_j(\theta_i)) = V_j(\theta_i). \end{aligned}$$

(ii): By Assumption 4, $\zeta_n(a_t^n | s_t^n) > \nu$ for all $n \in [|\mathcal{J}_i|]$ and $t \in [T]$. This implies that, for each $n \in [\bar{N}_i : |\mathcal{J}_i|]$,

$$\left| \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \left(\sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right) \right| \leq B_j \frac{1 - \gamma^{T+1}}{1 - \gamma} \frac{1}{\nu^{T+1}} = \bar{\phi}_j \quad (18)$$

By Popovicius' inequality [38, Corollary 1], we have

$$\text{Var} \left[\prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \left(\sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right) \right] \leq \bar{\phi}_j^2.$$

Since the random variables

$$\left\{ \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \left(\sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right) \right\}_{n \in [\bar{N}_i : |\mathcal{J}_i|]}$$

are independent, it follows that

$$\text{Var} \left[\sum_{n=\bar{N}_i+1}^{|\mathcal{J}_i|} \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \left(\sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right) \right] \leq \tilde{N}_i \bar{\phi}_j^2.$$

On the other hand, note that

$$\left| \sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right| \leq B_j \frac{1 - \gamma^{T+1}}{1 - \gamma} = \phi_j. \quad (19)$$

By Popovicius' inequality [38, Corollary 1],

$$\text{Var} \left[\sum_{\bar{n}=1}^{\tilde{N}_i} \sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right] \leq \tilde{N}_i \phi_j^2,$$

from where the bound on the variance follows. Note also that (18) and (19) imply that $\widehat{V}_j(\theta_i)$ is uniformly upper bounded by $\frac{\tilde{N}_i \phi_j + \bar{N}_i \bar{\phi}_j}{|\mathcal{J}_i|}$.

(iii): This follows from Hoeffding's inequality [39] using (18) and (19).

(iv): Under (17), we have

$$\frac{\pi_{\theta}(a|s)}{\pi_{\theta'}(a|s)} \leq \exp(\tilde{L}\|\theta - \theta'\|),$$

for any $\theta, \theta' \in \mathbb{R}^d$. Therefore,

$$\begin{aligned} \left| \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\pi_{\bar{\theta}_n}(a_t^n | s_t^n)} \left(\sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right) \right| \\ \leq B_j \frac{1 - \gamma^{T+1}}{1 - \gamma} \exp((T+1)\tilde{L}\|\theta_i - \bar{\theta}_n\|) = \tilde{\phi}_{i,j,n}. \end{aligned} \quad (20)$$

By Popovicius' inequality [38, Corollary 1], this implies

$$\text{Var}\left[\prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\pi_{\bar{\theta}_n}(a_t^n | s_t^n)} \left(\sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right)\right] \leq \tilde{\phi}_{i,j,n}^2.$$

The result now follows by noting that the random variables

$$\left\{ \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\pi_{\bar{\theta}_n}(a_t^n | s_t^n)} \left(\sum_{t=0}^T \gamma^t R_j(s_t^n, a_t^n, s_{t+1}^n) \right) \right\}_{n \in [\mathcal{J}_i]}$$

are independent. Note also that (20) implies that $\widehat{V}_j(\theta_i)$ is uniformly upper bounded by $\frac{\sum_{n=1}^{|\mathcal{J}_i|} \tilde{\phi}_{i,j,n}}{|\mathcal{J}_i|}$.

(v): This follows from Hoeffding's inequality [39] using (20). \square

In Proposition VI.1, the bounds in (iv) and (v) derived under the additional assumption (17) are tighter than the ones in (ii) and (iii). This is because $\tilde{\phi}_{i,j,n}$, which appears in (iv) and (v), depends on the difference between the policy parameters $\bar{\theta}_n$ and θ_i , so it takes advantage of their proximity. Instead, $\tilde{\phi}_j$, which appears in (ii) and (iii), is insensitive to this proximity. We next study the statistical properties of the gradients.

Proposition VI.2. (Gradient of value function estimates): Suppose Assumptions 4, 5, and 6 hold. Let $i \in \mathbb{Z}_{>0}$, \bar{N}_i and \tilde{N}_i as in Proposition VI.1, and

$$\begin{aligned} \psi_j &= \tilde{B} \sum_{t=0}^T \gamma^t \sum_{t'=t}^T (\gamma^{t'-t} B_j + \hat{B}), \\ \bar{\psi}_j &= \frac{\tilde{B}}{\nu^{T+1}} \sum_{t=0}^T \gamma^t \sum_{t'=t}^T (\gamma^{t'-t} B_j + \hat{B}). \end{aligned}$$

Then, for $j \in \{0\} \cup [q]$

- (i) $\mathbb{E}[\widehat{\nabla V}_j(\theta_i)] = \nabla V_j(\theta_i)$ (unbiased gradient estimates);
- (ii) $\text{Var}[\widehat{\nabla V}_j(\theta_i)^{(l)}] = \frac{\bar{N}_i \psi_j^2 + \tilde{N}_i \bar{\psi}_j^2}{|\mathcal{J}_i|^2}$ and $|\widehat{\nabla V}_j(\theta_i)^{(l)}| \leq \frac{\bar{N}_i \psi_j + \tilde{N}_i \bar{\psi}_j}{|\mathcal{J}_i|}$, for all $l \in [d]$;
- (iii) $\mathbb{P}(\|\widehat{\nabla V}_j(\theta_i) - \nabla V_j(\theta_i)\| \leq \epsilon) \geq 1 - 2d \exp\left(-\frac{\epsilon^2 |\mathcal{J}_i|^2}{2d(\bar{N}_i \psi_j^2 + \tilde{N}_i \bar{\psi}_j^2)}\right)$.

Further assume that $\chi_{a,s}$ is globally Lipschitz, uniformly in a, s , i.e., (17) holds, and that the policies in Λ belong to $\{\pi_\theta\}$. Let $\{\bar{\theta}_n\}_{n=1}^{|\mathcal{J}_i|}$ denote the parameters that describe all the policies in \mathcal{J}_i and define $\tilde{\psi}_{i,j,n} = \exp((T+1)\tilde{L}\|\theta_i - \bar{\theta}_n\|)\psi_j$. Then,

- (iv) $\text{Var}[\widehat{\nabla V}_j(\theta_i)^{(l)}] \leq \frac{\sum_{n=1}^{|\mathcal{J}_i|} \tilde{\psi}_{i,j,n}^2}{|\mathcal{J}_i|^2}$ and $|\widehat{\nabla V}_j(\theta_i)^{(l)}| \leq \frac{\sum_{n=1}^{|\mathcal{J}_i|} \tilde{\psi}_{i,j,n}}{|\mathcal{J}_i|}$, for all $l \in [d]$;
- (v) $\mathbb{P}(\|\widehat{\nabla V}_j(\theta_i) - \nabla V_j(\theta_i)\| \leq \epsilon) \geq 1 - 2d \exp\left(-\frac{\epsilon^2 |\mathcal{J}_i|^2}{2 \sum_{n=1}^{|\mathcal{J}_i|} \tilde{\psi}_{i,j,n}^2}\right)$.

Proof. (i): Let $j \in [q] \cup \{0\}$. With the notation of (14), by the Policy Gradient Theorem with baseline (cf. [22, Section 13.4]), for each $n \in [\bar{N}_i]$, we have

$$\mathbb{E}\left[\sigma_j \sum_{t=0}^T \gamma^t \nabla \chi_{a_t^n, s_t^n}(\theta_i) D_{j,t}^n\right] = \nabla V_j(\theta_i).$$

On the other hand, for $n \in [\bar{N}_i : |\mathcal{J}_i|]$,

$$\begin{aligned} \mathbb{E}\left[\sigma_j \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \sum_{t=0}^T \gamma^t \nabla \chi_{a_t^n, s_t^n}(\theta_i) D_{j,t}^n\right] &= \\ \int_{\Gamma} \sigma_j \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \sum_{t=0}^T \gamma^t \nabla \chi_{a_t^n, s_t^n}(\theta_i) D_{j,t}^n \zeta_n(a_t^n | s_t^n) d\Omega &= \\ \int_{\Gamma} \sigma_j \eta(s_0) \prod_{t=0}^T \pi_{\theta_i}(a_t^n | s_t^n) \sum_{t=0}^T \gamma^t \nabla \chi_{a_t^n, s_t^n}(\theta_i) D_{j,t}^n d\Omega &= \\ = \nabla V_j(\theta_i), \end{aligned}$$

where $d\Omega$ and Γ are defined as in the proof of Proposition VI.1, and for the last equality we have also used the Policy Gradient Theorem with baseline (cf. [22, Section 13.4]). Therefore,

$$\mathbb{E}[\widehat{\nabla V}_j(\theta_i)] = \frac{\bar{N}_i}{\bar{N}_i + \tilde{N}_i} \nabla V_j(\theta_i) + \frac{\tilde{N}_i}{\bar{N}_i + \tilde{N}_i} \nabla V_j(\theta_i) = \nabla V_j(\theta_i).$$

(ii): Note that for each $l \in [d]$,

$$\begin{aligned} \left| \sigma_j \sum_{t=0}^T \gamma^t \nabla \chi_{a_t^n, s_t^n}(\theta_i)^{(l)} \sum_{t'=t}^T (\gamma^{t'-t} R_j(s_{t'}^n, a_{t'}^n, s_{t'+1}^n) - b(s_t^n)) \right| &= \\ \leq \tilde{B} \sum_{t=0}^T \gamma^t \sum_{t'=t}^T (\gamma^{t'-t} B_j + \hat{B}) = \psi_j. \end{aligned} \quad (21)$$

By Popovicius' inequality [38, Corollary 1], this implies

$$\text{Var}\left[\sigma_j \sum_{t=0}^T \gamma^t \nabla \chi_{a_t^n, s_t^n}(\theta_i)^{(l)} D_{j,t}^n\right] \leq \psi_j^2,$$

for $n \in [\bar{N}_i]$. On the other hand, for $n \in [\bar{N}_i : |\mathcal{J}_i|]$,

$$\begin{aligned} \left| \sigma_j \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \sum_{t=0}^T \gamma^t \nabla \chi_{a_t^n, s_t^n}(\theta_i)^{(l)} D_{j,t}^n \right| &\leq \\ \frac{\tilde{B}}{\nu^{T+1}} \sum_{t=0}^T \gamma^t \sum_{t'=t}^T (\gamma^{t'-t} B_j + \hat{B}) = \bar{\psi}_j. \end{aligned} \quad (22)$$

Again, by Popoviciu's inequality [38, Corollary 1],

$$\text{Var}\left[\sigma_j \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \sum_{t=0}^T \gamma^t \frac{\partial}{\partial \theta(l)} \chi_{a_t^n, s_t^n}(\theta_i) D_{j,t}^n\right] \leq \bar{\psi}_j^2,$$

for $n \in [\bar{N}_i : |\mathcal{J}_i|]$. Since the random variables

$$\left\{ \sigma_j \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\zeta_n(a_t^n | s_t^n)} \sum_{t=0}^T \gamma^t \nabla \chi_{a_t^n, s_t^n}(\theta_i)^{(l)} D_{j,t}^n \right\}_{n \in [\mathcal{J}_i]}$$

are independent, it follows that

$$\text{Var}[\widehat{\nabla V}_j(\theta_i)^{(l)}] \leq \frac{\bar{N}_i \psi_j^2 + \tilde{N}_i \bar{\psi}_j^2}{|\mathcal{J}_i|^2}, \quad \forall l \in [d].$$

Note also that (21) and (22) imply that $\widehat{\nabla V}_j(\theta_i)^{(l)}$ is uniformly upper bounded by $\frac{\bar{N}_i \psi_j + \tilde{N}_i \bar{\psi}_j}{|\mathcal{J}_i|}$.

(iii): From Hoeffding's inequality, using (21), (22), for any $\epsilon > 0$ and $l \in [d]$,

$$\mathbb{P}\left(\left|\widehat{\nabla V}_j(\theta_i)^{(l)} - \nabla V_j(\theta_i)^{(l)}\right| \leq \frac{\epsilon}{\sqrt{d}}\right) \geq$$

$$1 - 2 \exp \left\{ - \frac{\epsilon^2 |\mathcal{J}_i|^2}{2d(\bar{N}_i \psi_j^2 + 2\bar{N}_i \bar{\psi}_j^2)} \right\}.$$

Now, note that if $|\widehat{\nabla V_j}(\theta_i)^{(l)} - \nabla V_j(\theta_i)^{(l)}| \leq \frac{\epsilon}{\sqrt{d}}$ for all $l \in [d]$, then $\|\widehat{\nabla V_j}(\theta_i) - \nabla V_j(\theta_i)\| \leq \epsilon$, which means that

$$\begin{aligned} & \mathbb{P} \left(\|\widehat{\nabla V_q}(\theta_i) - \nabla V_q(\theta_i)\| \leq \epsilon \right) \\ & \geq \mathbb{P} \left(\bigcap_{l=1}^d \left\{ |\widehat{\nabla V_q}(\theta_i)^{(l)} - \nabla V_q(\theta_i)^{(l)}| \leq \frac{\epsilon}{\sqrt{d}} \right\} \right). \end{aligned}$$

Using Fréchet's Inequality [40],

$$\begin{aligned} & \mathbb{P} \left(\bigcap_{l=1}^d \left\{ |\widehat{\nabla V_q}(\theta_i)^{(l)} - \nabla V_q(\theta_i)^{(l)}| \leq \frac{\epsilon}{\sqrt{d}} \right\} \right) \geq \\ & 1 - 2d \exp \left\{ - \frac{\epsilon^2 |\mathcal{J}_i|^2}{2d(\bar{N}_i \psi_j^2 + \bar{N}_i \bar{\psi}_j^2)} \right\}, \end{aligned}$$

and the result follows.

(iv): Under (17), for each $n \in [|\mathcal{J}_i|]$,

$$\left| \sigma_j \prod_{t=0}^T \frac{\pi_{\theta_i}(a_t^n | s_t^n)}{\pi_{\bar{\theta}_n}(a_t^n | s_t^n)} \sum_{t=0}^T \gamma^t \frac{\partial}{\partial \theta^{(l)}} \chi_{a_t^n, s_t^n}(\theta_i) D_{j,t}^n \right| \leq \quad (23)$$

$$\bar{B} \exp \left\{ (T+1) \tilde{L} \|\theta_i - \bar{\theta}_n\| \right\} \sum_{t=0}^T \gamma^t \sum_{t'=t}^T (\gamma^{t'-t} B_j + \hat{B}) = \tilde{\psi}_{i,j,n}.$$

The argument is analogous to the one used in (ii). Note (23) implies $\widehat{\nabla V_j}(\theta_i)^{(l)}$ is uniformly upper bounded by $\frac{\sum_{n=1}^{|\mathcal{J}_i|} \tilde{\psi}_{i,j,n}}{|\mathcal{J}_i|}$.

(v): This follows analogously to item (iii) by using (23). \square

Propositions VI.1 and VI.2 characterize the statistical properties of the estimates of the value functions and their gradients, generalizing to the on/off-policy case our previous result in [1, Lemma 2], which was limited to the on-policy case. These results show that, by increasing the number of episodes (either on-policy or off-policy) used, the distribution of the estimates of the value functions and their gradients concentrates around their true values, with the rate of concentration depending on the constants defined in Assumptions 4, 5, 6.

Remark VI.3. (Assumption on global Lipschitzness): Assumption (17) is standard in the literature (cf. [28, Assumption 3.1]). We note that, if the parameterized policy π_θ is globally Lipschitz uniformly in a and s , then (17) is satisfied. Indeed, using the Mean Value Theorem [41, Theorem 5.10], and under Assumption 4, we deduce

$$|\log \pi_\theta(a|s) - \log \pi_{\theta'}(a|s)| \leq \frac{1}{p^*} |\pi_\theta(a|s) - \pi_{\theta'}(a|s)|,$$

for some $p^* \in [\pi_\theta(a|s), \pi_{\theta'}(a|s)]$. Note that such p^* is strictly positive because of Assumption 4. Hence, if π_θ is globally Lipschitz uniformly in a and s , it follows that (17) is satisfied. This is the case for truncated Gaussian policies with compact state and action spaces (cf. [42, Section 6], [1, Section 5]). •

B. Safety Guarantees

In this section we study the safety guarantees of RSGF-RL.

Theorem VI.4. (Safety guarantees): *Suppose Assumptions 4, 5, and 6 hold. Let $i \in \mathbb{Z}_{>0}$, \bar{N}_i , \bar{N}_i , ϕ_j , and ϕ_j as in Proposition VI.1, and ψ_j , $\bar{\psi}_j$ as in Proposition VI.2. Suppose (15) is feasible at $\theta_i \in \mathbb{R}^d$ and the stepsize satisfies*

$$h_i < \min \left\{ \frac{1}{\alpha}, \frac{\beta(\theta_i)}{L_1}, \dots, \frac{\beta(\theta_i)}{L_q}, \frac{\beta(\theta_i)}{L_{q+1}} \right\}. \quad (24)$$

For $j \in [q]$, define

$$\hat{M}_{i,j} = \frac{-(1 - \alpha h_i) \widehat{V}_j(\theta_i) + \frac{h_i}{2} (\beta(\theta_i) - L_j h_i) \|\hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)\|^2}{1 + h_i \|\hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)\|}.$$

Then, for any $\delta \in (0, 1)$, under (16)

(i) if $\widehat{V}_j(\theta_i) \leq 0$ and

$$\frac{|\mathcal{J}_i|^2}{\bar{N}_i \phi_j^2 + \bar{N}_i \bar{\phi}_j^2} \geq -\frac{2}{\hat{M}_{i,j}^2} \log \frac{\delta}{2}, \quad (25a)$$

$$\frac{|\mathcal{J}_i|^2}{\bar{N}_i \psi_j^2 + \bar{N}_i \bar{\psi}_j^2} \geq -\frac{2d}{\hat{M}_{i,j}^2} \log \frac{\delta}{2d}, \quad (25b)$$

then $\mathbb{P}(V_j(\theta_{i+1}) \leq 0) \geq 1 - 2\delta$;

(ii) if $\widehat{V}_j(\theta_i) > 0$ is such that $\hat{M}_{i,j} > 0$, and (25) holds, then, $\mathbb{P}(V_j(\theta_{i+1}) \leq 0) \geq 1 - 2\delta$;

(iii) if for each $j \in [q]$ such that $\widehat{V}_j(\theta_i) > 0$, it holds that $\hat{M}_{i,j} > 0$, and (25) holds for all $j \in [q]$, then $\mathbb{P}(V_j(\theta_{i+1}) \leq 0, \forall j \in [q]) \geq 1 - 2q\delta$;

(iv) if $V_{q+1}(\theta_i) \leq 0$, then $V_{q+1}(\theta_{i+1}) \leq 0$.

Proof. (i): Since ∇V_j is Lipschitz with Lipschitz constant L_j , cf. Lemma A.1, we invoke [43, Lemma 1.2.3] to deduce

$$V_j(\theta_{i+1}) \leq V_j(\theta_i) + \nabla V_j(\theta_i)^\top (\theta_{i+1} - \theta_i) + \frac{L_j}{2} \|\theta_{i+1} - \theta_i\|^2. \quad (26)$$

This implies, using the Cauchy-Schwartz inequality, that

$$\begin{aligned} V_j(\theta_{i+1}) & \leq V_j(\theta_i) - \widehat{V}_j(\theta_i) + \widehat{V}_j(\theta_i) + \\ & \quad \|\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i)\| \|\theta_{i+1} - \theta_i\| + \\ & \quad \widehat{\nabla V}_j(\theta_i)^\top (\theta_{i+1} - \theta_i) + \frac{L_j}{2} \|\theta_{i+1} - \theta_i\|^2. \end{aligned} \quad (27)$$

Since $\theta_{i+1} = \theta_i + h_i \hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)$, and by using the constraints in (15), inequality (27) implies

$$\begin{aligned} V_j(\theta_{i+1}) & \leq V_j(\theta_i) - \widehat{V}_j(\theta_i) + (1 - \alpha h_i) \widehat{V}_j(\theta_i) + \\ & \quad \|\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i)\| h_i \mathcal{R}_{\alpha,\beta}(\theta_i) \\ & \quad - \frac{h_i}{2} (\beta(\theta_i) - L_j h_i) \|\mathcal{R}_{\alpha,\beta}(\theta_i)\|^2. \end{aligned} \quad (28)$$

Note that (24), together with the fact that $\widehat{V}_j(\theta_i) \leq 0$, implies that $\hat{M}_{i,j} > 0$. Now, by Proposition VI.1(iii), if $\frac{|\mathcal{J}_i|^2}{\bar{N}_i \phi_j^2 + \bar{N}_i \bar{\phi}_j^2} \geq -\frac{2}{\hat{M}_{i,j}^2} \log \frac{\delta}{2}$, then $\mathbb{P}(|\widehat{V}_j(\theta_i) - V_j(\theta_i)| \leq \hat{M}_{i,j}) \geq 1 - \delta$.

On the other hand, by Proposition VI.2(iii), if $\frac{|\mathcal{J}_i|^2}{\bar{N}_i \psi_j^2 + \bar{N}_i \bar{\psi}_j^2} \geq -\frac{2d}{\hat{M}_{i,j}^2} \log \frac{\delta}{2d}$, then $\mathbb{P}(\|\widehat{\nabla V}_j(\theta_i) - \nabla V_j(\theta_i)\| \leq \hat{M}_{i,j}) \geq 1 - \delta$. Using (28) and the definition of $\hat{M}_{i,j}$, we deduce that, if

$|\widehat{V}_j(\theta_i) - V_j(\theta_i)| \leq \hat{M}_{i,j}$ and $\|\widehat{\nabla V}_j(\theta_i) - \nabla V_j(\theta_i)\| \leq \hat{M}_{i,j}$, then $V_j(\theta_{i+1}) \leq 0$. Now, the result follows by Fréchet's inequality [40].

(ii): if $\hat{M}_{i,j} > 0$, $|\widehat{V}_j(\theta_i) - V_j(\theta_i)| \leq \hat{M}_{i,j}$, and $\|\widehat{\nabla V}_j(\theta_i) - \nabla V_j(\theta_i)\| \leq \hat{M}_{i,j}$, then $V_j(\theta_{i+1}) \leq 0$, even if $\widehat{V}_j(\theta_i) \geq 0$. The result follows by using a similar argument to (i).

(iii): this follows from (i), (ii), and Fréchet's inequality [40].

(iv): this follows from employing (15c) in (26), combined with the hypothesis that $V_{q+1}(\theta_i) \leq 0$. \square

Theorem VI.4(i) shows that if the number of episodes utilized to estimate $V_j(\theta_i)$ is sufficiently large and $\widehat{V}_j(\theta_i) \leq 0$ (i.e., we estimate that the j -th safety constraint is satisfied at iteration i), then the next iterate of RSGF-RL satisfies the j -th safety constraint with arbitrarily high probability. Similarly, Theorem VI.4(ii) provides such guarantees when $\widehat{V}_j(\theta_i) \geq 0$ (i.e., we estimate that the j -th safety constraint is not satisfied at iteration i). We note that $\hat{M}_{i,j} > 0$ holds when $\widehat{V}_j(\theta_i) \leq 0$ and $\|\hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)\|$ is nonzero (which is a reasonable assumption if θ_i is away from a KKT point). By continuity, this suggests that $\hat{M}_{i,j} > 0$ is also satisfied in a neighborhood of $\{\theta \in \mathbb{R}^d : \widehat{V}_j(\theta) \leq 0\}$ (again provided that $\|\hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)\|$ is nonzero), and it becomes increasingly more difficult to satisfy with large $\widehat{V}_j(\theta_i)$. Intuitively, this means that safety can be ensured in the next iteration as long as safety violations in the current iteration are not too extreme.

Remark VI.5. (Feasibility): Since the estimates of the value functions and their gradients converge to their true values as the number of episodes increases, cf. Propositions VI.1 and VI.2, the requirement in Theorem VI.4 that (15) is feasible at θ_i is satisfied for large enough number of episodes with high probability under Assumptions 2 and 3, cf. Lemma IV.2. \bullet

We state next a result that provides safety guarantees over a finite time horizon. Its proof follows from Theorem VI.4 and Fréchet's inequality [40]. We omit it for space reasons.

Corollary VI.6. (Safety guarantees over a finite time horizon): Suppose Assumptions 4, 5 and 6 hold. Let $H \in \mathbb{Z}_{>0}$. If, for each $i \in [H]$, the assumptions in Theorem VI.4(iii) hold, then under (16), $\mathbb{P}\left(\bigcap_{i=1}^{H+1} \{V_j(\theta_i) \leq 0, \forall j \in [q]\}\right) \geq 1 - 2qH\delta$.

Corollary VI.6 provides conditions under which consecutive iterates of RSGF-RL probabilistically satisfy the constraints. Since δ is a design parameter, this guarantee can be ensured with arbitrarily high probability. Smaller values of δ , however, require a larger number of episodes, as reflected in (25).

C. Convergence Guarantees

Here we provide convergence guarantees for RSGF-RL.

Theorem VI.7. (Almost sure convergence): Suppose Assumptions 1, 2, 4, 5, 6 hold. Further suppose that:

- (i) $V_{q+1}(\theta_0) \leq 0$;
- (ii) for all $\theta \in \Theta \setminus \mathcal{C}$, Slater's condition holds for (6) and CRC holds for (6) at $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$;
- (iii) (15) is feasible for all $i \in \mathbb{Z}_{>0}$;
- (iv) $\lim_{i \rightarrow \infty} \|\mathcal{R}_{\alpha,\beta}(\theta_i) - \hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)\| = 0$ with probability one;

- (v) $\lim_{i \rightarrow \infty} h_i = 0, \sum_{i=1}^{\infty} h_i = \infty$;

Then, under (16), the sequence $\{\theta_i\}_{i \in \mathbb{Z}_{>0}}$ converges to the set of KKT points of (4) in Θ almost surely.

Proof. Our proof proceeds by verifying that the hypotheses required by [19, Theorem 2.3.1] hold and then invoking this result. First, note that $\{\theta_i\}_{i \in \mathbb{Z}_{>0}}$ is bounded with probability one. Indeed, since $V_{q+1}(\theta_0) \leq 0$ by (i), it follows from (iii) and Theorem VI.4(iv) that $V_{q+1}(\theta_i) \leq 0$ for all $i \in \mathbb{Z}_{>0}$. This guarantees that $\|\theta_i\| \leq \sqrt{C}$ for all $i \in \mathbb{Z}_{>0}$. Second, $\mathcal{R}_{\alpha,\beta}$ is continuous on Θ . Indeed, $\mathcal{R}_{\alpha,\beta}$ is locally Lipschitz on \mathcal{C} by Lemma IV.2 and, since for all $\theta \in \Theta \setminus \mathcal{C}$, Slater's condition holds for (6) and CRC holds for (6) at $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$, cf. (ii), a similar argument to the one in the proof of Lemma IV.2 guarantees that $\mathcal{R}_{\alpha,\beta}$ is locally Lipschitz on $\Theta \setminus \mathcal{C}$. Third, the set of KKT points of (4) in Θ is globally asymptotically stable in Θ by an argument analogous to that of Proposition IV.5(ii). Furthermore, we write the dynamics as

$$\theta_{i+1} = \theta_i + h_i \mathcal{R}_{\alpha,\beta}(\theta_i) + h_i (\hat{\mathcal{R}}_{\alpha,\beta}(\theta_i) - \mathcal{R}_{\alpha,\beta}(\theta_i)).$$

Note that the noise sequence $\hat{\mathcal{R}}_{\alpha,\beta}(\theta_i) - \mathcal{R}_{\alpha,\beta}(\theta_i)$ is asymptotically vanishing with probability one, cf. (iv) and the stepsize sequence satisfies $\lim_{i \rightarrow \infty} h_i = 0, \sum_{i=1}^{\infty} h_i = \infty$, cf. (v). Finally, taking the sequence $\{\xi_n\}$ in the notation of [19, Theorem 2.3.1] equal to zero, we conclude that the sequence $\{\theta_i\}_{i \in \mathbb{Z}_{>0}}$ converges to the set of KKT points in Θ almost surely. \square

Remark VI.8. (Assumptions in Theorem VI.7): Requirement (i) on the initial policy estimate and (ii) on constraint qualification conditions are reasonable, given our discussion above. The feasibility requirement in (iii) follows in the setting considered in Remark VI.5. Regarding requirement (iv), we note that, by the same argument as in Lemma IV.2, the function $\hat{\xi} : \mathbb{R}^{d(2\bar{q}+1)+1} \rightarrow \mathbb{R}^d$ defined as

$$\hat{\xi}(\{A_j\}_{j=1}^{\bar{q}}, \{B_j\}_{j=0}^{\bar{q}}, C) = \arg \min_{\xi \in \mathbb{R}^d} \|\xi + B_0\|^2 \quad (29)$$

$$\text{s.t. } A_j + B_j^\top \xi + \frac{C}{2} \|\xi\|^2 \leq 0, \quad j \in [\bar{q}],$$

is locally Lipschitz. This means that small perturbations in $\{\nabla V_j\}_{j=1}^{\bar{q}}$ and $\{V_j\}_{j=1}^{\bar{q}}$ (like the ones obtained from using estimates of such quantities) result in small perturbations in $\mathcal{R}_{\alpha,\beta}$. In particular, this implies that, for $\bar{\epsilon} > 0$, there exists $\bar{\delta}$ such that, if $\|\widehat{\nabla V}_j(\theta) - \nabla V_j(\theta)\| < \bar{\delta}$ for all $j \in [q] \cup \{0\}$ and $\|\widehat{V}_j(\theta) - V_j(\theta)\| < \bar{\delta}$ for all $j \in [q]$, then $\|\hat{\mathcal{R}}_{\alpha,\beta}(\theta) - \mathcal{R}_{\alpha,\beta}(\theta)\| < \bar{\epsilon}$. Since the estimates of the value functions and their gradients become arbitrarily close to their true values if a sufficiently large number of episodes is used (cf. Propositions VI.1 and VI.2), this means that the condition $\lim_{i \rightarrow \infty} \|\mathcal{R}_{\alpha,\beta}(\theta_i) - \hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)\| = 0$ is satisfied if the number of episodes used to create the estimates of the value functions and their gradients increases as the number of iterations increases. Finally, an example of a stepsize sequence verifying (v) is $h_i = \frac{1}{i}$. \bullet

The following result complements the almost sure convergence established in Theorem VI.7 by providing a bound on the number of iterations required to converge to a neighborhood of a KKT point. This finite iteration convergence result is based on ideas from [28, Theorem 4.3].

Theorem VI.9. (Finite iteration convergence): *Suppose Assumptions 4, 5, 6 hold and that (15) is feasible at every $\{\theta_i\}_{i \in \mathbb{Z}_{\geq 0}}$. Let $h_0 = \frac{1}{\alpha}$ and $h_i = \frac{1}{\alpha\sqrt{i}}$ for $i \in \mathbb{Z}_{>0}$, and assume that for each $\theta \in \Theta$, Slater's condition holds for (15). Let $\hat{\ell} > 0$ be such that $\|\hat{\mathcal{R}}_{\alpha,\beta}(\theta)\| \leq \hat{\ell}$ for all $\theta \in \Theta$ and $B_L > 0$ such that $\hat{u}_j(\theta) \leq B_L$ for all $j \in [q]$ and $\theta \in \Theta$. Let $\epsilon_* < \frac{1}{B_L}$. For $\epsilon > 0$, define*

$$\text{It}_\epsilon = \min\{i \in \mathbb{Z}_{>0} : \inf_{0 \leq j \leq i} \mathbb{E}[\|\hat{\mathcal{R}}_{\alpha,\beta}(\theta_j)\|^2] \leq \epsilon\}.$$

Let $\epsilon > 0$ and $\bar{\sigma} > 0$ such that $\text{Var}(\widehat{\nabla V}_j(\theta_i)^{(l)}) \leq \bar{\sigma}$ for all $i \in [\text{It}_\epsilon]$, $j \in \{0\} \cup [q]$, and $l \in [d]$, $\|\hat{\mathcal{R}}_{\alpha,\beta}(\theta_0)\|^2 > \epsilon$, and $\epsilon > \frac{3}{2}\hat{\ell}\bar{\sigma}(\frac{q}{\epsilon_} + 1)$. Then, there exists $\kappa > 0$ such that*

$$\text{It}_\epsilon \leq \left(\frac{\kappa}{\epsilon - \frac{3}{2}\hat{\ell}\bar{\sigma}(\frac{q}{\epsilon_*} + 1)} \right)^2.$$

Proof. First, since Slater's condition holds for (15), $\hat{\mathcal{R}}_{\alpha,\beta}$ is continuous on Θ [44, Theorem 5.3], and since Θ is compact, $\hat{\ell}$ as in the statement exists. Using (26), we deduce

$$V_j(\theta_{i+1}) \leq V_j(\theta_i) + (\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i))^\top (\theta_{i+1} - \theta_i) + \widehat{\nabla V}_j(\theta_i)^\top (\theta_{i+1} - \theta_i) + \frac{L_j \hat{\ell}^2 h_i^2}{2}, \quad (30)$$

for all $j \in \{0\} \cup [q]$. Define $J_+^i = \{j \in [q] : \widehat{V}_j(\theta_i) \geq 0\}$. Let $\epsilon_* > 0$ as in the statement and define $V_{\epsilon_*}^i = \widehat{V}_0(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^i} \widehat{V}_j(\theta_i)$. Equivalently, we write

$$V_{\epsilon_*}^{i+1} = \widehat{V}_0(\theta_{i+1}) - V_0(\theta_{i+1}) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\widehat{V}_j(\theta_{i+1}) - V_j(\theta_{i+1})) + V_0(\theta_{i+1}) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} V_j(\theta_{i+1})$$

Using (30), we have

$$\begin{aligned} V_{\epsilon_*}^{i+1} &\leq \widehat{V}_0(\theta_{i+1}) - V_0(\theta_{i+1}) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\widehat{V}_j(\theta_{i+1}) - V_j(\theta_{i+1})) \\ &\quad + V_0(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} V_j(\theta_i) \\ &\quad + (\nabla V_0(\theta_i) - \widehat{\nabla V}_0(\theta_i))^\top (\theta_{i+1} - \theta_i) \\ &\quad + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i))^\top (\theta_{i+1} - \theta_i) \\ &\quad + \widehat{\nabla V}_0(\theta_i)^\top (\theta_{i+1} - \theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} \widehat{\nabla V}_j(\theta_i)^\top (\theta_{i+1} - \theta_i) \\ &\quad + \left(\frac{L_0}{2} + \frac{\sum_{j=1}^q L_j}{2\epsilon_*} \right) \hat{\ell} h_i^2. \end{aligned} \quad (31)$$

Equivalently,

$$\begin{aligned} V_{\epsilon_*}^{i+1} &\leq \widehat{V}_0(\theta_{i+1}) - V_0(\theta_{i+1}) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\widehat{V}_j(\theta_{i+1}) - V_j(\theta_{i+1})) \\ &\quad + V_0(\theta_i) - \widehat{V}_0(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (V_j(\theta_i) - \widehat{V}_j(\theta_i)) \end{aligned}$$

$$\begin{aligned} &+ \widehat{V}_0(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1} \cap J_+^i} \widehat{V}_j(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1} \setminus J_+^i} \widehat{V}_j(\theta_i) \\ &+ (\nabla V_0(\theta_i) - \widehat{\nabla V}_0(\theta_i))^\top (\theta_{i+1} - \theta_i) \\ &+ \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i))^\top (\theta_{i+1} - \theta_i) \\ &+ \widehat{\nabla V}_0(\theta_i)^\top (\theta_{i+1} - \theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} \widehat{\nabla V}_j(\theta_i)^\top (\theta_{i+1} - \theta_i) \\ &+ \left(\frac{L_0}{2} + \frac{\sum_{j=1}^q L_j}{2\epsilon_*} \right) \hat{\ell} h_i^2. \end{aligned} \quad (32)$$

Using an argument analogous to the one in the proof of Proposition IV.5(i) to obtain equation (10), but now with the estimates and the definition (15) of the approximated RSGF, one can derive, for all $i \in \mathbb{Z}_{>0}$,

$$\begin{aligned} \widehat{\nabla V}_0(\theta_i)^\top \hat{\mathcal{R}}_{\alpha,\beta}(\theta_i) &= - \left(1 + \frac{\beta(\theta_i)}{2} \sum_{j=1}^q \hat{u}_j(\theta_i) \right) \|\hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)\|^2 \\ &+ \sum_{j=1}^q \alpha \hat{u}_j(\theta_i) \widehat{V}_j(\theta_i) \leq - \|\hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)\|^2 + \sum_{j=1}^q \alpha \hat{u}_j(\theta_i) \widehat{V}_j(\theta_i), \end{aligned} \quad (33)$$

where $\hat{u}_j(\theta)$ denotes the Lagrange multiplier associated to constraint j in (15). Furthermore, from the constraints in (15),

$$\widehat{\nabla V}_j(\theta_i)^\top \hat{\mathcal{R}}_{\alpha,\beta}(\theta_i) \leq -\alpha \widehat{V}_j(\theta_i), \quad (34)$$

for all $j \in [q]$. Substituting (33) and (34) into (32), we get

$$\begin{aligned} V_{\epsilon_*}^{i+1} &\leq \widehat{V}_0(\theta_{i+1}) - V_0(\theta_{i+1}) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\widehat{V}_j(\theta_{i+1}) - V_j(\theta_{i+1})) \\ &\quad + V_0(\theta_i) - \widehat{V}_0(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (V_j(\theta_i) - \widehat{V}_j(\theta_i)) \\ &\quad + \widehat{V}_0(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1} \cap J_+^i} \widehat{V}_j(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1} \setminus J_+^i} \widehat{V}_j(\theta_i) \\ &\quad + (\nabla V_0(\theta_i) - \widehat{\nabla V}_0(\theta_i))^\top (\theta_{i+1} - \theta_i) \\ &\quad + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i))^\top (\theta_{i+1} - \theta_i) \\ &\quad - h_i \|\hat{\mathcal{R}}_{\alpha,\beta}(\theta_i)\|^2 + \sum_{j=1}^q \alpha h_i \hat{u}_j(\theta_i) \widehat{V}_j(\theta_i) \\ &\quad - \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} \alpha h_i \widehat{V}_j(\theta_i) + \left(\frac{L_0}{2} + \frac{\sum_{j=1}^q L_j}{2\epsilon_*} \right) \hat{\ell} h_i^2. \end{aligned} \quad (35)$$

Now note that by an argument analogous to that of [26, Lemma D.1], there exists $B_L > 0$ as in the statement. Since the iterates $\{\theta_i\}_{i \in \mathbb{Z}_{>0}}$ remain bounded in Θ , we have $\hat{u}_j(\theta_i) \leq B_L$ for all $j \in [q]$ and $i \in \mathbb{Z}_{>0}$. Since $\widehat{V}_j(\theta_i) \leq 0$ for $j \notin J_+^i$, $\sum_{j=1}^q \alpha h_i \hat{u}_j(\theta_i) \widehat{V}_j(\theta_i) \leq \sum_{j \in J_+^i} \alpha h_i \hat{u}_j(\theta_i) \widehat{V}_j(\theta_i)$ and using $\epsilon_* < \frac{1}{B_L}$ it follows that

$$\sum_{j=1}^q \alpha h_i \hat{u}_j(\theta_i) \widehat{V}_j(\theta_i) - \sum_{j \in J_+^{i+1}} \frac{\alpha h_i}{\epsilon_*} \widehat{V}_j(\theta_i) \leq$$

$$\begin{aligned}
& \sum_{j \in J_+^i} \alpha h_i \hat{u}_j(\theta_i) \widehat{V}_j(\theta_i) - \sum_{j \in J_+^{i+1}} \frac{\alpha h_i}{\epsilon_*} \widehat{V}_j(\theta_i) \leq \\
& \sum_{j \in J_+^i \setminus J_+^{i+1}} \alpha h_i \hat{u}_j(\theta_i) \widehat{V}_j(\theta_i) - \sum_{j \in J_+^{i+1} \setminus J_+^i} \frac{\alpha h_i}{\epsilon_*} \widehat{V}_j(\theta_i) \leq \\
& \sum_{j \in J_+^i \setminus J_+^{i+1}} \frac{\alpha h_i}{\epsilon_*} \widehat{V}_j(\theta_i) - \sum_{j \in J_+^{i+1} \setminus J_+^i} \frac{\alpha h_i}{\epsilon_*} \widehat{V}_j(\theta_i). \quad (36)
\end{aligned}$$

Using the fact that $\sum_{j \in J_+^i \setminus J_+^{i+1}} \frac{\widehat{V}_j(\theta_i)}{\epsilon_*} + \sum_{j \in J_+^i \cap J_+^{i+1}} \frac{\widehat{V}_j(\theta_i)}{\epsilon_*} = \sum_{j \in J_+^i} \frac{\widehat{V}_j(\theta_i)}{\epsilon_*}$ along with (36) and $\alpha h_i < 1$, we get

$$\begin{aligned}
V_{\epsilon_*}^{i+1} & \leq \widehat{V}_0(\theta_{i+1}) - V_0(\theta_{i+1}) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\widehat{V}_j(\theta_{i+1}) - V_j(\theta_{i+1})) \\
& + V_0(\theta_i) - \widehat{V}_0(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (V_j(\theta_i) - \widehat{V}_j(\theta_i)) \\
& + \widehat{V}_0(\theta_i) + \frac{1}{\epsilon_*} \sum_{j \in J_+^i} \widehat{V}_j(\theta_i) \\
& + (\nabla V_0(\theta_i) - \widehat{\nabla V}_0(\theta_i))^\top (\theta_{i+1} - \theta_i) + \\
& \frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i))^\top (\theta_{i+1} - \theta_i) \\
& - h_i \|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_i)\|^2 + \left(\frac{L_0}{2} + \frac{\sum_{j=1}^q L_j}{2\epsilon_*} \right) \hat{\ell} h_i^2. \quad (37)
\end{aligned}$$

Taking expectations on both sides of (37) with respect to the σ -algebra generated by $(\{\theta_j\}_{j \in [\text{It}_\epsilon]}, \mathcal{I}_0, \{\mathcal{J}_j\}_{j \in [\text{It}_\epsilon - 1]})$, we get

$$\begin{aligned}
\mathbb{E}[V_{\epsilon_*}^{i+1}] & \leq \mathbb{E}[V_{\epsilon_*}^i] - h_i \mathbb{E}[\|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_i)\|^2] \\
& + \mathbb{E}[(\nabla V_0(\theta_i) - \widehat{\nabla V}_0(\theta_i))^\top (\theta_{i+1} - \theta_i)] \\
& + \mathbb{E}\left[\frac{1}{\epsilon_*} \sum_{j \in J_+^{i+1}} (\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i))^\top (\theta_{i+1} - \theta_i)\right] \\
& + \left(\frac{L_0}{2} + \frac{\sum_{j=1}^q L_j}{2\epsilon_*} \right) \hat{\ell} h_i^2. \quad (38)
\end{aligned}$$

Let V_* be such that $V_{\epsilon_*}^i \geq V_*$ for all $i \in \mathbb{Z}_{>0}$ (note that such value exists because the value function estimates are uniformly bounded as shown in Proposition VI.1). Define $U_i = V_{\epsilon_*}^i - V_*$ and $L_* = \frac{L_0}{2} + \frac{\sum_{j=1}^q L_j}{2\epsilon_*}$. Summing (38) for $i \in \{0\} \cup [\text{It}_\epsilon]$,

$$\begin{aligned}
\sum_{i=0}^{\text{It}_\epsilon} \mathbb{E}[\|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_i)\|^2] & \leq \sum_{i=0}^{\text{It}_\epsilon} \left(\frac{\mathbb{E}[U_i]}{h_i} - \frac{\mathbb{E}[U_{i+1}]}{h_i} \right) \\
& + \sum_{i=0}^{\text{It}_\epsilon} L_* \hat{\ell} h_i + \sum_{i=0}^{\text{It}_\epsilon} \frac{\mathbb{E}[(\theta_{i+1} - \theta_i)^\top (\nabla V_0(\theta_i) - \widehat{\nabla V}_0(\theta_i))]}{h_i} \\
& + \sum_{i=0}^{\text{It}_\epsilon} \sum_{j \in J_+^{i+1}} \frac{\mathbb{E}[(\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i))^\top (\theta_{i+1} - \theta_i)]}{\epsilon_* h_i}. \quad (39)
\end{aligned}$$

Note that

$$\sum_{i=0}^{\text{It}_\epsilon} \left(\frac{\mathbb{E}[U_i]}{h_i} - \frac{\mathbb{E}[U_{i+1}]}{h_i} \right) =$$

$$\frac{\mathbb{E}[U_0]}{h_0} + \sum_{i=1}^{\text{It}_\epsilon} \left(\frac{1}{h_i} - \frac{1}{h_{i-1}} \right) \mathbb{E}[U_i] - \frac{\mathbb{E}[U_{\text{It}_\epsilon+1}]}{h_{\text{It}_\epsilon}}.$$

Since $\{U_i\}_{i \in \mathbb{Z}_{>0}}$ is positive and uniformly upper bounded (cf. Proposition VI.1), by letting B_u be such that $|U_i| \leq B_u$ for all $i \in \mathbb{Z}_{>0}$, and noting that $\frac{1}{h_i} \geq \frac{1}{h_{i-1}}$ for all $i \in [\text{It}_\epsilon]$,

$$\sum_{i=0}^{\text{It}_\epsilon} \left(\frac{\mathbb{E}[U_i]}{h_i} - \frac{\mathbb{E}[U_{i+1}]}{h_i} \right) \leq \frac{B_u}{h_0} + \sum_{i=1}^{\text{It}_\epsilon} \left(\frac{1}{h_i} - \frac{1}{h_{i-1}} \right) B_u = \frac{B_u}{h_{\text{It}_\epsilon}}.$$

On the other hand, by the Cauchy-Schwartz inequality,

$$\begin{aligned}
\mathbb{E}[(\theta_{i+1} - \theta_i)^\top (\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i))] & \leq \\
& \sqrt{\mathbb{E}[\|\theta_{i+1} - \theta_i\|^2]} \sqrt{\mathbb{E}[\|\nabla V_j(\theta_i) - \widehat{\nabla V}_j(\theta_i)\|^2]},
\end{aligned}$$

for all $j \in \{0\} \cup [q]$. Moreover, since $\|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_i)\| \leq \hat{\ell}$ for all $i \in \mathbb{Z}_{>0}$, and since $\max_{i \in [K_\epsilon]} \text{Var}(\widehat{\nabla V}_j(\theta_i)) \leq \bar{\sigma}$ for all $j \in \{0\} \cup [q]$, we have from (39) that

$$\begin{aligned}
\frac{1}{\text{It}_\epsilon} \sum_{i=0}^{\text{It}_\epsilon} \mathbb{E}[\|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_i)\|^2] & \leq \frac{B_u}{\text{It}_\epsilon h_{\text{It}_\epsilon}} + \frac{\sum_{i=1}^{\text{It}_\epsilon} L_* \hat{\ell} h_i}{\text{It}_\epsilon} \\
& + \hat{\ell} \bar{\sigma} \left(\frac{q}{\epsilon_*} + 1 \right) \frac{\text{It}_\epsilon + 1}{\text{It}_\epsilon}.
\end{aligned}$$

Using the fact that $\sum_{i=1}^{\text{It}_\epsilon} i^{-a} \leq \text{It}_\epsilon^{1-a} - 1$ (cf. [42, page 31]) for any $a \in (0, 1)$, substituting $h_i = \frac{1}{\alpha \sqrt{i}}$, and using $\frac{\text{It}_\epsilon + 1}{\text{It}_\epsilon} \leq \frac{3}{2}$ (because $\|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_0)\| > \epsilon$), we obtain

$$\begin{aligned}
\frac{1}{\text{It}_\epsilon} \sum_{i=0}^{\text{It}_\epsilon} \mathbb{E}[\|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_i)\|^2] & \leq \frac{B_u \alpha}{\sqrt{\text{It}_\epsilon}} + \frac{L_* \hat{\ell}}{\alpha} \left(\frac{1}{\sqrt{\text{It}_\epsilon}} - \frac{1}{\text{It}_\epsilon} \right) \\
& + \frac{3}{2} \hat{\ell} \bar{\sigma} \left(\frac{q}{\epsilon_*} + 1 \right). \quad (40)
\end{aligned}$$

By definition of It_ϵ , $\mathbb{E}[\|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_i)\|^2] > \epsilon$, for all $i \in \{0\} \cup [\text{It}_\epsilon - 1]$, and therefore from (40) by taking $\kappa = B_u \alpha + \frac{L_* \hat{\ell}}{\alpha}$,

$$\epsilon \leq \frac{1}{\text{It}_\epsilon} \sum_{i=0}^{\text{It}_\epsilon - 1} \mathbb{E}[\|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_i)\|^2] \leq \frac{\kappa}{\sqrt{\text{It}_\epsilon}} + \frac{3}{2} \hat{\ell} \bar{\sigma} \left(\frac{q}{\epsilon_*} + 1 \right),$$

from where the result follows. \square

We note the result in Theorem VI.9 ensures the existence of $j \in [\text{It}_\epsilon]$ such that $\mathbb{E}[\|\hat{\mathcal{R}}_{\alpha, \beta}(\theta_j)\|^2] \leq \epsilon$, but does not imply that the convergence in expectation of the norm of $\hat{\mathcal{R}}_{\alpha, \beta}$ is monotonic. This is akin to the convergence results obtained for policy gradient methods (cf. [28, Theorem 4.3]). We also point out that the iteration number It_ϵ is defined in terms of $\hat{\mathcal{R}}_{\alpha, \beta}$, instead of $\mathcal{R}_{\alpha, \beta}$. As justified in Remark VI.8, by using a sufficiently large number of episodes when estimating the value functions and their gradients, $\hat{\mathcal{R}}_{\alpha, \beta}$ and $\mathcal{R}_{\alpha, \beta}$ can be made arbitrarily close at any point with high probability. This means that if the estimates of all policies obtained for $i \in [\text{It}_\epsilon]$ are computed with a sufficiently large number of episodes, Theorem VI.9 provides a bound for the number of iterations needed to reach a KKT point with high probability.

Remark VI.10. (Assumptions in Theorem VI.9): The argument in the proof of Theorem VI.9 is valid for any sequence $h_i = \frac{i^{-a}}{\alpha}$ for $a \in (0, 1)$, but by following an argument similar

to that of [28, Theorem 4.3], the optimal rate is $a = 1/2$, which is the one adopted in the statement. Moreover, Proposition VI.2 provides a way to compute the number of episodes necessary to ensure that the condition $\text{Var}(\nabla \bar{V}_j(\theta_i)^{(l)}) \leq \bar{\sigma}$ is satisfied for all $j \in \{0\} \cup [q]$, $i \in [\text{It}_\epsilon]$ and $l \in [d]$. Finally, since Θ is compact, if $\mathcal{R}_{\alpha,\beta}$ is locally Lipschitz on Θ (e.g., if for all $\theta \in \Theta$, Slater's condition holds for (6) and CRC holds for (6) at $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$, cf. Lemma IV.2), then $\mathcal{R}_{\alpha,\beta}$ is bounded in Θ . Hence, $\hat{\ell}$ exists provided that the value function and gradient estimates are taken so that $\|\mathcal{R}_{\alpha,\beta} - \hat{\mathcal{R}}_{\alpha,\beta}\|$ is bounded. This holds, for example, under the asymptotically vanishing noise assumption discussed in Remark VI.8. •

VII. SIMULATIONS

Here we test RSGF-RL in two scenarios: a robot solving a navigation task in a 2D environment and a cart-pole system seeking to keep the pole upright by moving the cart. We compare its performance against other approaches¹ that also seek to solve constrained Markov decision processes in an anytime fashion: constrained policy optimization (CPO) algorithm [45] and on-policy RSGF-RL [1].

Navigation 2D: We test the algorithm in a 2D environment, where a robot with single-integrator dynamics navigates to a target point while avoiding unknown obstacles (cf. Figure 1). The obstacles' location approximates a simplified real-world floorplan. The state space is given by $s = (x, y) \in [0, 10]^2$, representing the position of the agent, and the action space is continuous, with $a \in [-5, 5]^2$ representing the velocity in the x and y directions. The target point is $s^* = (8.5, 8)$. The reward is $R_0(s, a) = -\|s - s^*\|$ and the constraint reward is

$$R_1(s, a) = \begin{cases} \varepsilon(e^{d(s)} - 1), & \text{if } s \in \mathcal{C} \\ 1 - \varepsilon, & \text{otherwise} \end{cases} \quad (41)$$

where $\varepsilon = 0.01$, $d(s)$ is the distance between s and the closest obstacle border, and the safe set \mathcal{C} is the obstacle-free region inside $[0, 10]^2$. We use the family of Gaussian policies $\pi_\theta(a|s) \sim \mathcal{N}(\mu_\theta(s), \Sigma)$, where $\Sigma = 0.5\mathbf{I}_2$ and the mean function is defined by radial basis functions (RBF) kernels,

$$\mu_\theta(s) = \sum_{i=1}^{N_c} \tanh(\theta_i) \exp\left(-\frac{\|s - c_i\|^2}{2\sigma^2}\right) \quad (42)$$

Here, \tanh is applied element-wise, $\{c_i\}_{i=1}^{N_c}$ are the RBF centers, and $\{\theta_i\}_{i=1}^{N_c} \subset \mathbb{R}^2$ are the training parameters. We choose the centers to be evenly spaced points over the state space. To make a fair comparison, all algorithms collect the same amount of episodes per iteration and perform the same number of iterations. Table I summarizes the training setup and the hyperparameters. For RSGF-RL, the estimators are constructed using data from both the current and the immediately preceding policies. To mitigate the high variance of the estimators during the training process, we clip the values of the importance sampling weights between 0.8 and 1.2.

Figure 1 shows the policy evolution under the RSGF-RL algorithm starting from an initial safe policy. The anytime

nature of the algorithm is reflected in the fact that intermediate iterations remain safe. Figure 2 shows the evolution of the performance (V_0) and safety (V_1) metrics for the different strategies. One can see that both on- and off-policy RSGF-RL outperform CPO, while remaining safe during the whole training procedure or recovering from an initial unsafe state. Interestingly, off-policy RSGF-RL without clipping the importance sampling weights performs similarly to on-policy RSGF-RL, while RSGF-RL with clipping significantly outperforms both in terms of sampling efficiency, converging with fewer iterations. This suggests off-policy data can improve the training process but introduces a high variance on the estimators that needs to be compensated by variance reduction techniques.

TABLE I: Hyperparameters for the simulation environments

Parameter	Navigation 2D	Cart-pole
Environment		
Time horizon (T)	50	200
Discount factor (γ)	0.98	0.995
Reward (R_0)	$-\ s - s^*\ $	1
Constraint reward (R_1)	Eq. (41)	Eq. (41)
ε	0.01	0.1
Policy		
Type	Gaussian	Gaussian
Centers evenly spaced over	$[0, 10]^2$	$[-3, 3] \times [-\frac{\pi}{4}, \frac{\pi}{4}] \times [-1, 1] \times [-1.5, 1.5]$
Number of centers (N_c)	400	1000
RBF centers variance	0.5	0.5
Policy variance (Σ)	$0.5\mathbf{I}_2$	$0.5\mathbf{I}_4$
Common elements		
Number of iterations (k)	1500	300
Episodes per iteration (N_i)	100	30
V_0 baseline ($b_0(s)$)	0	neural network
V_1 baseline ($b_1(s)$)	0	neural network
CPO		
δ	0.15	4×10^{-4}
On-policy RSGF-RL		
Step size (h)	0.1	10^{-3}
α	9	0.1
Off-policy RSGF-RL		
Step size (h)	0.1	$\min\{10^{-3}, \frac{0.02}{\ \mathcal{R}_{\alpha,\beta}\ }\}$
Episodes available ($ \mathcal{I}_i $)	200	15
Updates per iteration	1	2
α	9	0.1
β (constant)	1	1

Cart-pole: We also evaluate RSGF-RL on the Gymnasium *Inverted Pendulum-v4* environment [46], where the objective is to learn a policy that keeps a pole upright by applying forces to a cart while avoiding hitting a wall. The state is $s = (x, \theta, \dot{x}, \dot{\theta}) \in \mathbb{R}^4$, where x is the cart position, θ is the pole angle (relative to vertical), \dot{x} is the cart velocity, and $\dot{\theta}$ is the pole angular velocity. The action space is continuous, with $a \in [-3, 3]$ representing the force applied to the cart. The reward is $R_0(s, a) = 1$, encouraging the pole to remain upright as long as possible. The wall is at $x = 0.5$ and hence the safe set is $\mathcal{C} = \{s = [x, \theta, \dot{x}, \dot{\theta}] \in \mathbb{R}^4 : x < 0.5\}$.

We define the constraint reward as in (41), with $\varepsilon = 0.1$ and $d(s) = [1, 0, 0, 0]^\top s - 0.5$ and use the same Gaussian policy with centers uniformly distributed over the state space. Table I gathers the training details. To showcase the flexibility of the off-policy approach, we update the policy twice per iteration using two minibatches, instead of relying on previous trajectories (i.e., we run steps 5-8 twice in Algorithm 1).

¹The interested reader can find in [1] a comparative analysis of on-policy RSGF-RL with primal-dual approaches [10], [11].

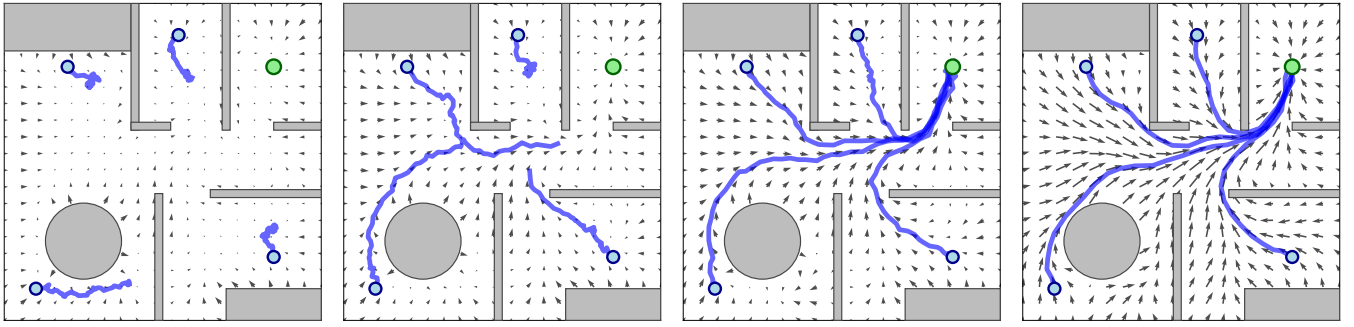


Fig. 1: Policy evolution under RSGF-RL in the Navigation 2D example. Obstacles are depicted in gray. Target point in green and different robot initial conditions in light blue. Initial policy on the left, final policy on the right, with intermediate policies obtained during the algorithm evolution in the middle.

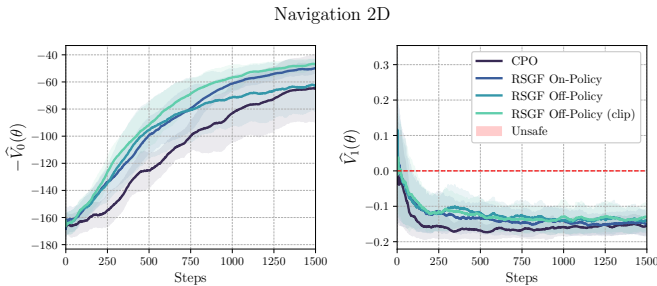


Fig. 2: Comparison between CPO and different RSGF-RL training strategies in the Navigation 2D environment. Left plot shows the average $V_0(\theta)$ as a performance metric, while the right plot shows the average $V_1(\theta)$ as a safety metric. Averages are computed over 5 seeds and the shaded area represents the standard deviation.

Figure 3 shows the evolution of the performance (V_0) and safety (V_1) metrics for the different algorithms. All approaches maintain the safety constraint below 0, remaining safe during training. Both on-policy and off-policy RSGF-RL outperform CPO, and RSGF-RL with clipping of the importance sampling weights significantly outperforms all the other approaches.

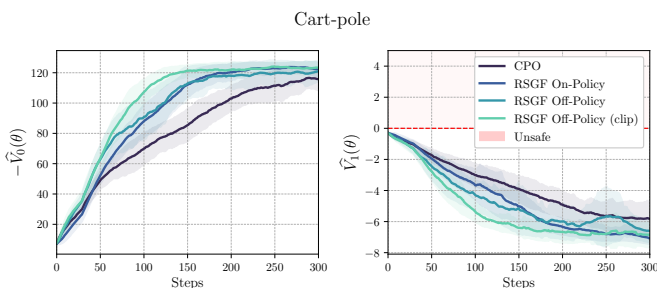


Fig. 3: Comparison between CPO and different RSGF-RL training strategies in the Cart-pole environment. Left plot shows the average $V_0(\theta)$ as a performance metric, while the right plot shows the average $V_1(\theta)$ as a safety metric. The max reward is 126.61. Averages are computed over 5 seeds and the shaded area represents the standard deviation.

VIII. CONCLUSIONS

We have introduced the Robust Safe Gradient Flow-based Reinforcement Learning (RSGF-RL) algorithm for constrained reinforcement learning with anytime safety guarantees. RSGF-RL's design is based on the Robust Safe Gradient Flow, a continuous-time algorithm for anytime constrained optimization whose forward invariance and asymptotic stability

properties we have also characterized. At every iteration, RSGF-RL uses off-policy episodic data to construct estimates of the value functions defining the constrained RL problem, as well as of their gradients. We have rigorously characterized the statistical properties of such estimates. Building on this, we have determined the number of episodes needed to ensure, with a user-specified reliability, that safe policies remain safe at the next iteration or, alternatively, that the algorithm returns to safety from an unsafe policy. Leveraging the theory of stochastic approximation, we have also shown that RSGF-RL converges to a KKT point almost surely, and we have provided a bound on the number of iterations required for convergence. Simulations have compared the performance of RSGF-RL with the state of the art. Future work will focus on extensions to other safety constraints commonly used in safe RL, such as probabilistic or conditional value-at-risk. We also plan to explore schemes that adaptively tune algorithm parameters for improved convergence, safety, and memory allocation requirements. Finally, we will extend the framework to actor-critic methods and perform tests in physical hardware.

REFERENCES

- [1] P. Mestres, A. Marzabal, and J. Cortés, “Anytime safe reinforcement learning,” in *Learning for Dynamics and Control Conference*, N. Ozay, L. Balzano, D. Panagou, and A. Abate, Eds. Proceedings of Machine Learning Research, 2025, vol. 283, pp. 221–232.
- [2] L. Brunkel, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, “Safe learning in robotics: from learning-based control to safe reinforcement learning,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.
- [3] Y. Liu, H. Avishai, and X. Liu, “Policy learning with constraints in model-free reinforcement learning: a survey,” in *International Joint Conference on Artificial Intelligence*, Montreal, Canada, 2021, pp. 4508–4515.
- [4] J. Garcia and F. Fernandez, “A comprehensive survey on safe reinforcement learning,” *Journal of Machine Learning Research*, vol. 16, pp. 1437–1480, 2015.
- [5] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, “A review of safe reinforcement learning: methods, theory and applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 11 216–11 235, 2024.
- [6] E. Altman, *Constrained Markov Decision Processes*. CRC Press, 1999, vol. 7.
- [7] J. Achiam, D. Held, A. Tamar, and P. Abbeel, “Constrained policy optimization,” in *International Conference on Machine Learning*. Proceedings of Machine Learning Research, 2017, vol. 70, pp. 22–31.
- [8] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, “Risk-constrained reinforcement learning with percentile risk criteria,” *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6070–6120, 2017.

- [9] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, “Constrained reinforcement learning has zero duality gap,” in *Proceedings of the 32nd Conference in Neural Information Processing Systems*, vol. 32, Vancouver, Canada, 2019, pp. 7555–7565.
- [10] S. Paternain, M. Calvo-Fullana, L. F. O. Chamon, and A. Ribeiro, “Safe policies for reinforcement learning via primal-dual methods,” *IEEE Transactions on Automatic Control*, vol. 68, no. 3, pp. 1321–1336, 2023.
- [11] D. Ding, K. Zhang, T. Basar, and M. Jovanovic, “Natural policy gradient primal-dual method for constrained Markov decision processes,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Curran Associates, Inc., 2020, vol. 33, pp. 8378–8390.
- [12] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic, “Provably efficient safe exploration via primal-dual policy optimization,” in *International Conference on Artificial Intelligence and Statistics*, A. Banerjee and K. Fukumizu, Eds. Proceedings of Machine Learning Research, 2021, vol. 130, pp. 3304–3312.
- [13] S. Zeng, T. T. Doan, and J. Romberg, “Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained Markov decision processes,” in *IEEE Conf. on Decision and Control*, Cancun, Mexico, 2022, pp. 4028–4033.
- [14] Q. Bai, A. S. Bedi, M. Agarwal, A. Koppel, and V. Aggarwal, “Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach,” in *AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, Vancouver, Canada, 2022, pp. 3682–3689.
- [15] Y. Liu, J. Ding, and X. Liu, “IPO: interior-point policy optimization under constraints,” in *AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, New York, NY, 2020, pp. 4940–4947.
- [16] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, “A Lyapunov-based approach to safe reinforcement learning,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, vol. 31, pp. 8103–8112.
- [17] J. McMahan and X. Zhu, “Anytime-constrained reinforcement learning,” in *International Conference on Artificial Intelligence and Statistics*, S. Dasgupta, S. Mandt, and Y. Li, Eds., vol. 238. Valencia, Spain: Proceedings of Machine Learning Research, 2024, pp. 4321–4329.
- [18] W. Suttle, V. K. Sharma, K. C. Kosaraju, S. Seetharaman, J. Liu, V. Gupta, and B. M. Sadler, “Sampling-based safe reinforcement learning for nonlinear dynamical systems,” in *International Conference on Artificial Intelligence and Statistics*, S. Dasgupta, S. Mandt, and Y. Li, Eds. Valencia, Spain: Proceedings of Machine Learning Research, 2024, vol. 238, pp. 4420–4428.
- [19] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York: Springer, 1978.
- [20] V. S. Borkar, *Stochastic Approximation A Dynamical Systems Viewpoint*. New Delhi, India: Hindustan Book Agency, 2008.
- [21] H. Khalil, *Nonlinear Systems*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 2002.
- [22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.
- [23] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.
- [24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2009.
- [25] J. Liu, “Sensitivity analysis in nonlinear programs and variational inequalities via continuous selections,” *SIAM Journal on Control and Optimization*, vol. 33, no. 4, pp. 1040–1060, 1995.
- [26] A. Allibhoy and J. Cortés, “Control barrier function-based design of gradient flows for constrained nonlinear programming,” *IEEE Transactions on Automatic Control*, vol. 69, no. 6, pp. 3499–3514, 2024.
- [27] A. Allibhoy and J. Cortés, “Sequential convex programming via discretization of continuous-time flows,” 2024, unpublished work.
- [28] K. Zhang, A. Koppel, H. Zhu, and T. Başar, “Global convergence of policy gradient methods to (almost) locally optimal policies,” *SIAM Journal on Control and Optimization*, vol. 58, no. 6, pp. 3586–3612, 2020.
- [29] N. Andréasson, A. Evgrafov, and M. Patriksson, *An Introduction to Continuous Optimization: Foundations and Fundamental Algorithms*. Courier Dover Publications, 2020.
- [30] P. Mestres, A. Allibhoy, and J. Cortés, “Regularity properties of optimization-based controllers,” *European Journal of Control*, vol. 81, p. 101098, 2025.
- [31] M. Nagumo, “Über die Lage der Integralkurven gewöhnlicher Differentialgleichungen,” *Proceedings of the Physico-Mathematical Society of Japan*, vol. 24, pp. 551–559, 1942.
- [32] S. P. Bhat and D. S. Bernstein, “Nontangency-based Lyapunov tests for convergence and stability in systems having a continuum of equilibria,” *SIAM Journal on Control and Optimization*, vol. 42, no. 5, pp. 1745–1775, 2003.
- [33] G. Di Pillo and L. Grippo, “Exact penalty functions in constrained optimization,” *SIAM Journal on Control and Optimization*, vol. 27, no. 6, pp. 1333–1360, 1989.
- [34] A. Auslender, R. Shefi, and M. Teboulle, “A moving balls approximation method for a class of smooth constrained minimization problems,” *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 3232–3259, 2010.
- [35] T. Xu, Z. Yang, Z. Wang, and Y. Liang, “Doubly robust off-policy actor-critic: convergence and optimality,” in *International Conference on Machine Learning*. Proceedings of Machine Learning Research, 2021, vol. 139, pp. 11 581–11 591.
- [36] J. Huang and N. Jiang, “On the convergence rate of off-policy policy optimization methods with density-ratio correction,” in *International Conference on Artificial Intelligence and Statistics*, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Eds. Proceedings of Machine Learning Research, 2022, vol. 151, pp. 2658–2705.
- [37] Q. Bai, W. U. Mondal, and V. Aggarwal, “Regret analysis of policy gradient algorithm for infinite horizon average reward Markov decision processes,” in *AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, Vancouver, Canada, 2024, pp. 10980–10988.
- [38] R. Bhatia and C. Davis, “A better bound on the variance,” *The American Mathematical Monthly*, vol. 107, no. 4, pp. 353–357, 2000.
- [39] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.
- [40] M. Fréchet, “Généralisations du théorème des probabilités totales,” *Fundamenta Mathematicae*, vol. 1, no. 25, pp. 379–387, 1935.
- [41] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1953.
- [42] K. Zhang, A. Koppel, H. Zhu, and T. Başar, “Global convergence of policy gradient methods to (almost) locally optimal policies,” *arXiv preprint arXiv:1506.08472*, 2020.
- [43] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed., ser. Springer Optimization and Its Applications. Springer International Publishing, 2018, vol. 137.
- [44] A. V. Fiacco and J. Kyparisis, “Sensitivity analysis in nonlinear programming under second order assumptions,” *Systems and Optimization*, vol. 66, pp. 74–97, 1985.
- [45] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [46] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. D. Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG, R. Perez-Vicente, A. Pierré, S. Schulhoff, J. J. Tai, H. Tan, and O. G. Younis, “Gymnasium: A standard interface for reinforcement learning environments,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.17032>
- [47] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

APPENDIX

Lemma A.1. (Lipschitzness of gradient of value functions): *Suppose Assumptions 5 and 6 hold. Let $j \in [q] \cup \{0\}$. Then, ∇V_j is Lipschitz with constant*

$$B_j L \left(\frac{1 - \gamma^T}{1 - \gamma} \right)^2 + 2B_j \tilde{B}^2 \gamma \frac{1 - (T+1)\gamma^T + T\gamma^{T+1}}{(1 - \gamma)^2} + B_q \tilde{B}^2 \left(\frac{1 - \gamma^T}{1 - \gamma} \right)^2.$$

Proof. By the Policy Gradient Theorem [22, Section 13.2], for any $\theta \in \mathbb{R}^d$,

$$\nabla V_j(\theta) = \sum_{t=0}^T \sum_{\tau=0}^T \int \gamma^{t+\tau} R_j(s_{t+\tau}, a_{t+\tau}, s_{t+\tau+1}) \nabla \chi_{a_t, s_t} p_\theta d\sigma,$$

where $\mathcal{I} = \mathcal{S}^{T+1} \times \mathcal{A}^{T+1}$, $d\sigma = ds_0 ds_1 \dots ds_T da_0 da_1 \dots da_T$, and

$$p_\theta = \left(\prod_{k=0}^{t+\tau} P(s_{k+1}, s_k, a_k) \right) \left(\prod_{k=0}^{t+\tau} \pi_\theta(a_k | s_k) \right) \eta(s_0).$$

Now, by following the same steps as in the proof of [28, Lemma 3.2],

$$\begin{aligned} \|\nabla V_j(\theta_1) - \nabla V_j(\theta_2)\| &\leq \sum_{t=0}^T \sum_{\tau=0}^T \gamma^{t+\tau} B_j L \|\theta_1 - \theta_2\| + \\ &\quad \sum_{t=0}^T \sum_{\tau=0}^T \gamma^{t+\tau} B_j \tilde{B}^2(t+\tau+1) \|\theta_1 - \theta_2\|. \end{aligned}$$

Now, by using the formulas

$$\sum_{t=0}^T \gamma^t = \frac{1-\gamma^{T+1}}{1-\gamma}, \quad \sum_{t=0}^T t\gamma^t = \gamma \frac{1-(T+1)\gamma^T + T\gamma^{T+1}}{(1-\gamma)^2},$$

we get

$$\begin{aligned} \|\nabla V_j(\theta_1) - \nabla V_j(\theta_2)\| &\leq B_j L \|\theta_1 - \theta_2\| \left(\frac{1-\gamma^{T+1}}{1-\gamma} \right)^2 \\ &\quad + 2B_j \tilde{B}^2 \|\theta_1 - \theta_2\| \gamma \frac{1-(T+1)\gamma^T + T\gamma^{T+1}}{(1-\gamma)^2} \\ &\quad + B_j \tilde{B}^2 \|\theta_1 - \theta_2\| \left(\frac{1-\gamma^{T+1}}{1-\gamma} \right)^2, \end{aligned}$$

from where the result follows. \square

The following result provides a sufficient condition under which Slater's condition holds for (6) for each $\theta \in \mathbb{R}^d \setminus \mathcal{C}$.

Lemma A.2. (Slater's condition): *Let $\delta : \mathbb{R}^d \rightarrow \mathbb{R}_+$ be a continuous function, and suppose that for each $\theta \in \mathbb{R}^d \setminus \mathcal{C}$, there exists $\xi \in \mathbb{R}^d$ that satisfies $\alpha V_j(\theta) + \nabla V_j(\theta)^\top \xi < -\delta(\theta)$. Consider $\xi^* : \mathbb{R}^d \rightarrow \mathbb{R}^d$ defined as*

$$\begin{aligned} \xi^*(\theta) &= \arg \min_{\xi \in \mathbb{R}^d} \|\xi\|^2 \\ \text{s.t. } \alpha V_j(\theta) + \nabla V_j(\theta)^\top \xi &\leq 0, \quad j \in [q]. \end{aligned} \quad (43)$$

Select a differentiable function β such that

$$\frac{\beta(\theta)}{2} \|\xi^*(\{V_j(\theta), \nabla V_j(\theta)\}_{j=1}^q)\| < \delta(\theta).$$

Then, Slater's condition holds for (6) for every $\theta \in \mathbb{R}^d \setminus \mathcal{C}$.

Proof. Since (43) satisfies Slater's condition for all $\theta \in \mathbb{R}^d \setminus \mathcal{C}$, ξ^* is continuous at every $\theta \in \mathbb{R}^d \setminus \mathcal{C}$ [44, Theorem 5.3]. Therefore, a differentiable β as required in the statement exists. Now, it follows that $\xi^*(\{V_j(\theta), \nabla V_j(\theta)\}_{j=1}^q)$ is strictly feasible for (6) for each $\theta \in \mathbb{R}^d \setminus \mathcal{C}$, and the result follows. \square

In particular, if δ is uniformly lower bounded by a positive constant and ξ^* is uniformly upper bounded, there exists a constant β function that makes Slater's condition hold for $\theta \in \mathbb{R}^d \setminus \mathcal{C}$. We also note that the feasibility of the linear inequalities $\alpha V_j(\theta) + \nabla V_j(\theta)^\top \xi < -\delta(\theta)$ can be verified using Farkas' Lemma [47, Theorem 22.1]. Therefore, one can verify the feasibility of such linear inequalities and select an appropriate β to satisfy Slater's condition in $\mathbb{R}^d \setminus \mathcal{C}$.

Next, we provide a condition for CRC to hold for (6).

Lemma A.3. (Constant rank condition): *Let $\tilde{q} = 1$, $\theta \in \mathcal{C}$ and suppose (5) satisfies MFCQ. Then, (6) satisfies CRC at $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$.*

Proof. If the single constraint of (6) is not active, then CRC trivially holds. Suppose that it is active. The gradient with

respect to ξ of the single constraint of (6) evaluated at $\xi = \mathcal{R}_{\alpha,\beta}(\theta)$ is $g_\theta = \nabla V_1(\theta) + \beta(\theta) \mathcal{R}_{\alpha,\beta}(\theta)$. Note that if $g_\theta \neq \mathbf{0}_d$, then there exists a neighborhood \mathcal{N} of $(\theta, \mathcal{R}_{\alpha,\beta}(\theta))$ such that if $(\bar{\theta}, \bar{\xi}) \in \mathcal{N}$, then $\nabla V_1(\bar{\theta}) + \beta(\bar{\theta}) \bar{\xi}$ has the same rank as g_θ . Alternatively, if $g_\theta = 0$, then $0 = \alpha V_1(\theta) + \nabla V_1(\theta)^\top \mathcal{R}_{\alpha,\beta}(\theta) + \frac{\beta(\theta)}{2} \|\mathcal{R}_{\alpha,\beta}(\theta)\|^2 = \alpha V_1(\theta) - \frac{\beta(\theta)}{2} \|\mathcal{R}_{\alpha,\beta}(\theta)\|^2$. This implies that $V_1(\theta) = 0$ and $\mathcal{R}_{\alpha,\beta}(\theta) = \mathbf{0}_d$. Since MFCQ holds for (5) and $V_1(\theta) = 0$, then $\nabla V_1(\theta) \neq \mathbf{0}_d$ necessarily. However, since $\mathcal{R}_{\alpha,\beta}(\theta) = \mathbf{0}_d$ this contradicts the fact that $g_\theta = \mathbf{0}_d$. \square

Finally, we state a few inequalities from probability theory used along the paper.

Lemma A.4. (Popoviciu's inequality [38, Corollary 1]): *Let X be a real-valued random variable. Let $m, M \in \mathbb{R}$ be such that $m \leq X \leq M$ almost surely. Then, $\text{Var}(X) \leq \frac{(M-m)^2}{4}$.*

Lemma A.5. (Hoeffding's inequality [39]): *Let X_1, \dots, X_n be independent random variables. Suppose there exist $a_i, b_i \in \mathbb{R}$ for $i \in [n]$ such that $a_i \leq X_i \leq b_i$ almost surely. Let $S_n = X_1 + \dots + X_n$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq \epsilon) \leq 2 \exp \left(- \frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2} \right).$$

Lemma A.6. (Fréchet's Inequality [40]): *Let $\{A_i\}_{i=1}^n$ be $n \in \mathbb{Z}_{>0}$ events. Then,*

$$\mathbb{P} \left(\bigcap_{i=1}^n A_i \right) \geq \max \left\{ 0, \sum_{i=1}^n \mathbb{P}(A_i) - (n-1) \right\}.$$



Pol Mestres received the Bachelor's degree in mathematics and the Bachelor's degree in engineering physics from the Universitat Politècnica de Catalunya, Barcelona, Spain, in 2020, and the Master's degree in mechanical engineering in 2021 from the University of California, San Diego, La Jolla, CA, USA, where he is currently a Ph.D. candidate. His research interests include safety-critical control, optimization-based controllers, distributed optimization and motion planning.



Arnau Marzabal received the Bachelor's degrees in Engineering Physics and in Industrial Engineering from the Universitat Politècnica de Catalunya, Barcelona, Spain, in 2025. He conducted his bachelor thesis at the University of California, San Diego, La Jolla, CA, USA, under the supervision of Prof. Jorge Cortés. His research interests include robotics, machine learning and data-driven control.



Jorge Cortés (M'02, SM'06, F'14) received the Licenciatura degree in mathematics from Universidad de Zaragoza, Spain, in 1997, and the Ph.D. degree in engineering mathematics from Universidad Carlos III de Madrid, Spain, in 2001. He held postdoctoral positions with the University of Twente, Twente, The Netherlands, and the University of Illinois at Urbana-Champaign, Illinois, USA. He is a Professor and Cymer Corporation Endowed Chair in High Performance Dynamic Systems Modeling and Control at the Department of Mechanical and Aerospace Engineering, UC San Diego, California, USA. He is a Fellow of IEEE, SIAM, and IFAC. His research interests include distributed control and optimization, network science, nonsmooth analysis, reasoning and decision making under uncertainty, network neuroscience, and multi-agent coordination in robotic, power, and transportation networks.