Proceedings of Machine Learning Research vol vvv:1–30, 2027

Towards Probabilistically Safe Reinforcement Learning: Theory and Deployment on Humanoid and Aerial Robots

Pol Mestres MESTRES@CALTECH.EDU
Varun Gabbita VGABBITA@CALTECH.EDU
Jason H.K. Siu JASONSIU@CALTECH.EDU
Sergio A. Esteban SESTEBAN@CALTECH.EDU
Hampus Serneke HSERNEKE@CALTECH.EDU
Joel W. Burdick JBURDICK@CALTECH.EDU
Aaron D. Ames AMES@CALTECH.EDU
Department of Mechanical and Civil Engineering, Caltech, Pasadena, CA, USA

Abstract

Reinforcement learning (RL) enables robots to acquire complex skills, but standard RL formulations do not guarantee that learned policies satisfy safety constraints, which limits their use in safety-critical settings. Existing safe RL methods with provable guarantees typically support only restricted classes of constraints and require unbiased estimators of the objective and constraint functions, which prevents their direct integration with state-of-the-art algorithms such as proximal policy optimization (PPO). In this paper, we extend the recently introduced robust safe gradient flow-based reinforcement learning (RSGF-RL) framework to overcome these limitations. Our formulation accommodates a general class of trajectory-level safety constraints, including chance constraints such as collision avoidance with a prescribed probability, and our analysis allows for biased value and gradient estimators. We show that, with high probability, the constraint violation throughout training is bounded by the estimator bias plus a term that vanishes as the number of rollouts grows, and decays to this level if the initial policy is unsafe. We also establish the convergence properties of the iterates to approximate KKT points of the original safe RL problem. Finally, we show that RSGF-RL implemented with PPO-style estimates enables zero-shot sim-to-real deployment of whole-body control policies on a Unitree G1 humanoid and collision-free waypoint navigation on a Crazyflie drone.

Keywords: Safe reinforcement learning, model-free reinforcement learning, humanoid robots.

1. Introduction

Reinforcement learning (RL) has emerged as a powerful tool for enabling robots to acquire a wide range of complex skills (Hwangbo et al., 2019; OpenAI et al., 2020; Kaufmann et al., 2023). However, standard RL algorithms primarily optimize task performance, without guaranteeing that the learned policies satisfy safety constraints, either during training or after deployment. This lack of formal safety guarantees remains a major obstacle to the adoption of RL in safety-critical applications.

Motivated by this challenge, a growing body of work on *safe RL* seeks to optimize policy performance subject to prescribed safety specifications (Brunkel et al., 2022; Garcia and Fernandez, 2015; Gu et al., 2024). A common approach formulates safe RL as a

© 2027 P. Mestres, V. Gabbita, J.H. Siu, S.A. Esteban, H. Serneke, J.W. Burdick & A.D. Ames.

constrained policy optimization problem, where the constraint requires the expected cumulative value of a safety-related cost to remain below a prescribed threshold (Achiam et al., 2017; Chow et al., 2017; Paternain et al., 2019). Several algorithms have been proposed to solve such problems, including primal-dual methods (Paternain et al., 2023; Ding et al., 2020), constrained policy optimization (CPO) (Achiam et al., 2017), projection-based constrained policy optimization (PCPO) (Yang et al., 2020), and state augmentation (Sootla et al., 2022). Despite their empirical success, these methods generally do not guarantee constraint satisfaction throughout the learning process, and their convergence guarantees rely on restrictive assumptions.

An alternative approach enforces safety at the level of individual control actions, typically by combining an unconstrained RL policy with a safety filter based on control barrier functions (CBFs) (Bejarano et al., 2025; Alshiekh et al., 2018; Cheng et al., 2019; Yang et al., 2026; Hsu et al., 2024). While such methods can provide guarantees on the safety of the resulting closed-loop system, they do not directly address the problem of optimizing a policy subject to safety constraints.

In this paper, we adopt the constrained policy optimization perspective, building on the recently introduced *robust safe gradient flow-based reinforcement learning* (RSGF-RL) framework (Mestres et al., 2025b,a). RSGF-RL provides convergence guarantees to critical points of the constrained RL problem while maintaining constraint satisfaction throughout the learning process with high probability. However, several limitations restrict its integration with state-of-the-art RL algorithms and modern robotic systems:

1. **Limited class of safety constraints.** The existing framework considers constraints that require that the expected value of *cumulative safety costs* remains below a given threshold. However, many safety specifications arising in robotics are naturally expressed as *probabilistic constraints*, such as requiring that the probability of colliding with an obstacle during an episode remains below a prescribed threshold, which are not equivalent to the expected cumulative cost formulations.

2. **Restrictive assumptions.** The theoretical guarantees of RSGF-RL as provided in (Mestres et al., 2025b,a) require *unbiased estimators* of the objective and constraint functions and their gradients. In contrast, modern RL algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), Soft Actor-Critic (SAC) (Haarnoja et al., 2024), and DreamerV3 (Hafner et al., 2025) employ estimation procedures that can introduce bias. This limits the integration of RSGF-RL with state-of-the-art policy optimization methods. Additionally, the conditions under which (Mestres et al., 2025a) establishes convergence of RSGF-RL (such as the requirement that the estimated RSGF-RL update law converges to the true one with probability one) are quite restrictive.

3. **Lack of hardware validation.** Previous evaluations of RSGF-RL have been limited to relatively simple simulation environments, leaving open whether its theoretical guarantees can be combined with the computational demands and practical challenges of RL policies for real robotic systems.

In this paper, we develop an extension of RSGF-RL that addresses these shortcomings, and show that the resulting algorithm can be used to train RL policies for a variety of different tasks, with zero-shot sim-to-real transfer. Our contributions are as follows.

1. First, we extend RSGF-RL to a general class of trajectory-level constraints, which includes cumulative cost and probabilistic constraints. Similarly to the Policy Gradient Theorem (Sutton and Barto, 2018, Chapter 13), we show that gradient estimates for this class of constraints follow from the classical likelihood-ratio method in (Glynn, 1990);

2. Second, we show that, with high probability, the constraint violation of the RSGF-RL iterates is bounded by the estimator bias plus a sampling-error term that vanishes as the number of rollouts N used for value and gradient estimation grows. If the initial policy is safe, this bound holds throughout training; if it is unsafe, the violation decays to this level. We also show that, almost surely, the RSGF-RL iterates have a limit point that is an approximate KKT point of the constrained RL problem, with the approximation level again given by the estimator bias plus a term that vanishes as $N \to \infty$.

3. Third, we show that RSGF-RL implemented with (biased) PPO gradient estimates enables zero-shot sim-to-real deployment of whole-body motor skills on a Unitree G1 humanoid robot and collision-free waypoint navigation on a Crazyflie drone. Some assumptions of our theoretical results are not met in the practical implementation of RSGF-RL used in these experiments. We discuss this theory-to-practice gap in Section 4.3.

2. Background

2.1. Safe Reinforcement Learning

Consider a Markov Decision Process $(\mathcal{S}, \mathcal{A}, P)$ with unknown transition function P. Let $T \in \mathbb{N}$ be a finite time horizon, and let $g_0, g_1 : \mathcal{S}^{T+2} \times \mathcal{A}^{T+1} \to \mathbb{R}$ be real-valued functions. We also consider a distribution η over initial states and a parametric class of policies indexed by $\theta \in \mathbb{R}^d$, with π_θ denoting the policy associated with parameter θ. Let $s = [s_0, \ldots, s_{T+1}]$ and $a = [a_0, \ldots, a_T]$. We consider the constrained RL problem

$$\min_{\theta \in \mathbb{R}^d} \quad V_0(\theta) = \mathbb{E}[g_0(s, a)] \qquad \text{s.t.} \quad V_1(\theta) = \mathbb{E}[g_1(s, a)] \leq 0. \tag{1}$$

The expectations in (1) are taken with respect to the probability measure over trajectories induced by the initial-state distribution η, the policy π_θ, and the transition function P. Unless stated otherwise, all expectations are taken with respect to this measure.

The function g_0 encodes the objective to be minimized over the horizon T, whereas g_1 encodes a safety constraint over the same horizon. For simplicity, we consider (1) with a single safety constraint and same time horizon for both the objective and constraint, but the results developed in the paper extend naturally to multiple constraints and different time horizons.

We next present different instances of (1) that illustrate the generality of this formulation and connect it with safe RL problems previously studied in the literature.

Example 1 (Cumulative constraints): *Let $\gamma \in (0,1]$ be a discount factor, $R_0 : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ a reward function, and $R_1 : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ a cost function. Consider $g_j(s,a) = (-1)^{j+1}\sum_{t=0}^{T}\gamma^t R_j(s_t,a_t,s_{t+1})$. Then* (1) *becomes the standard constrained RL formulation considered, e.g., in (Altman, 1999; Achiam et al., 2017; Chow et al., 2017; Paternain et al., 2019; Mestres et al., 2025b). The objective is to maximize the expected discounted cumulative reward while constraining the expected discounted cumulative cost.* •

Example 2 (Probabilistic constraints): *Let $\bar{\mathcal{C}} \subset \mathcal{S}$ be a safe set and let $\delta \in (0,1)$. Taking g_0 as in Example 1 and defining $g_1(s,a) = -\mathbb{I}\Big(\bigcap_{t=0}^{T}\{s_t \in \bar{\mathcal{C}}\}\Big) + 1 - \delta$, where $\mathbb{I}(A)$ denotes the indicator function of an event A, problem* (1) *becomes*

$$\max_{\theta\in\mathbb{R}^d}\ \mathbb{E}\Big[\sum_{t=0}^{T}\gamma^t R_0(s_t,a_t,s_{t+1})\Big], \qquad s.t.\quad \mathbb{P}\Big(\bigcap_{t=0}^{T}\{s_t \in \bar{\mathcal{C}}\}\Big) \geq 1-\delta. \tag{2}$$

This formulation requires the state trajectory to remain in the safe set $\bar{\mathcal{C}}$ over the entire horizon with probability at least $1-\delta$, and is considered, for example, in (Chen et al., 2024; Paternain et al., 2023). As shown in Appendix A, an appropriate choice of R_1 in Example 1 can also imply the probabilistic constraint in (2)*, but the two formulations are not equivalent.* •

Beyond examples 1-2, problem (1) also covers the percentile risk constraints considered in (Chow et al., 2017; Ying et al., 2021) and the conditional chance constraints in (Shen et al., 2024). Hence, (1) captures several common notions of safety in RL.

2.2. Robust Safe Gradient Flow

We next introduce the Robust Safe Gradient Flow (RSGF) (Mestres et al., 2025a), a constrained optimization method that will serve as the basis for solving (1). Let $V_0, V_1 : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable functions and consider the constrained optimization problem

$$\min_{\theta\in\mathbb{R}^d} V_0(\theta) \qquad \text{s.t.}\quad V_1(\theta) \leq 0. \tag{3}$$

Let $\mathcal{C} = \{\theta \in \mathbb{R}^d : V_1(\theta) \leq 0\}$ denote its feasible set. Given $\alpha, \beta > 0$, define the RSGF vector field $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}^d$ by

$$\mathcal{R}(\theta) = \arg\min_{\xi\in\mathbb{R}^d} \frac{1}{2}\left\|\xi + \nabla V_0(\theta)\right\|^2 \quad \text{s.t.}\quad \alpha V_1(\theta) + \nabla V_1(\theta)^\top \xi + \frac{\beta}{2}\left\|\xi\right\|^2 \leq 0. \tag{4}$$

Thus, $\mathcal{R}(\theta)$ selects a direction that remains as close as possible to the negative gradient $-\nabla V_0(\theta)$ while satisfying a constraint designed to preserve feasibility. Given a sequence of positive stepsizes $\{h_t\}_{t\in\mathbb{N}}$, we consider the discrete-time RSGF iteration

$$\theta_{t+1} = \theta_t + h_t \mathcal{R}(\theta_t). \tag{5}$$

As shown in (Mestres et al., 2025b, Lemma 1) and (Mestres et al., 2025a, Lemma IV.2, Prop. IV.3) (consolidated in Proposition 5 in Appendix B for completeness), under a constraint qualification and suitable Lipschitz and stepsize conditions, the limit points of (5) are KKT points of (3). Besides ensuring convergence to KKT points of (3), this result highlights the *anytime nature of RSGF*: provided that the initial iterate is feasible and the stepsizes satisfy the stated conditions, every subsequent iterate remains feasible.

3. Safe Reinforcement Learning with the Robust Safe Gradient Flow

Although the RSGF iterates (5) have KKT points of the constrained optimization problem (1) as limit points, they cannot be applied directly in the model-free RL setting. In particular, because the transition kernel P of the MDP is unknown, the value functions V_0, V_1 and their gradients $\nabla V_0, \nabla V_1$ cannot in general be evaluated exactly. In this section, we therefore construct estimates of these quantities from policy rollouts.

The following result provides an expression for the gradient of the general functions V_0 and V_1 introduced in (1), and will serve as the basis for estimating such gradients from sampled trajectories. This result is a direct consequence of the likelihood-ratio gradient identity (Glynn, 1990), and its proof is deferred to Appendix C, where we also discuss the special cases it recovers (including the standard Policy Gradient Theorem (Sutton and Barto, 2018, Chapter 13)).

Proposition 1 (Policy gradient for general objective and constraints): *Suppose that, for every $a \in \mathcal{A}$ and $s \in \mathcal{S}$, the mapping $\theta \mapsto \log \pi_\theta(a|s)$ is continuously differentiable, the functions g_0, g_1 are uniformly bounded, and for every $\theta \in \mathbb{R}^d$, there exists a neighborhood U_θ and a constant $M_\theta > 0$ where*

$$\sup_{\theta' \in U_\theta} \sup_{s \in \mathcal{S}, a \in \mathcal{A}} \|\nabla_{\theta'} \log \pi_{\theta'}(a|s)\| \leq M_\theta. \tag{6}$$

Let $H_t = (s_0, a_0, s_1, a_1, \ldots, a_{t-1}, s_t) \in (\mathcal{S} \times \mathcal{A})^t \times \mathcal{S}$ denote the trajectory history available when action a_t is selected. Further,[1]*, for each $t \in 0:T$, let $b_{j,t} : (\mathcal{S} \times \mathcal{A})^t \times \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ be an arbitrary integrable function. Then, for $j \in \{0, 1\}$,*

$$\nabla V_j(\theta) = \mathbb{E}\Big[\sum_{t=0}^{T} \big(g_j(s,a) - b_{j,t}(H_t, \theta)\big) \nabla_\theta \log \pi_\theta(a_t|s_t)\Big]. \tag{7}$$

Fix $j \in \{0, 1\}$. Given $N \in \mathbb{N}$ independent episodes generated under π_θ with initial distribution η, empirical averages of g_j and (7) yield unbiased estimators of V_j and ∇V_j, respectively. Practical RL algorithms, however, often modify these estimators to reduce variance or improve numerical performance, generally at the cost of bias: PPO-clip (Schulman et al., 2017, Section 3) optimizes a clipped surrogate objective, and actor–critic methods (Konda and Tsitsiklis, 2003) build lower-variance gradient estimates from learned value functions. We therefore consider general estimators

$$\widehat{V}_j(\theta) = \frac{1}{N} \sum_{n=1}^{N} G_{j,n}^\theta, \qquad \widehat{\nabla V}_j(\theta) = \frac{1}{N} \sum_{n=1}^{N} \tilde{G}_{j,n}^\theta, \tag{8}$$

where $G_{j,n}^\theta \in \mathbb{R}$ and $\tilde{G}_{j,n}^\theta \in \mathbb{R}^d$ satisfy, for all $n \in 1:N$, $i \in 1:d$, and some constants $\zeta_j, \tilde{\zeta}_j \geq 0$ and $B_j, \tilde{B}_j$ (independent of θ),

$$\big|\mathbb{E}[G_{j,n}^\theta] - V_j(\theta)\big| \leq \zeta_j, \qquad \big|\mathbb{E}[(\tilde{G}_{j,n}^\theta)_i] - (\nabla V_j(\theta))_i\big| \leq \tilde{\zeta}_j, \tag{9}$$

$$|G_{j,n}^\theta| \leq B_j, \qquad |(\tilde{G}_{j,n}^\theta)_i| \leq \tilde{B}_j \quad \text{a.s.} \tag{10}$$

1. For non-negative integers n_1, n_2, we use the notation $n_1 : n_2 = \{n_1, n_1 + 1, \ldots, n_2\}$.

Assumption (10) is standard in RL (Zhang et al., 2020; Bai et al., 2024; Mestres et al., 2025b,a); for the estimators induced by (7), it holds whenever $\nabla_\theta \log \pi_\theta$, $\{g_j\}$, and $\{b_{j,t}\}$ are uniformly bounded over (s, a). Bounds of the form (9) are available for PPO (Döring et al., 2026, Theorem 4.2) and actor–critic algorithms (Mu and Klabjan, 2024, Lemma 4.10). Using (8), we define the empirical RSGF direction

$$\hat{\mathcal{R}}(\theta) = \underset{\xi \in \mathbb{R}^d}{\arg\min} \frac{1}{2} \left\| \xi + \widehat{\nabla V_0}(\theta) \right\|^2 \quad \text{s.t.} \quad \alpha \widehat{V}_1(\theta) + \widehat{\nabla V}_1(\theta)^\top \xi + \frac{\beta}{2} \|\xi\|^2 \leq 0, \tag{11}$$

and the stochastic policy update

$$\theta_{t+1} = \theta_t + h_t \hat{\mathcal{R}}(\theta_t), \tag{12}$$

where for every θ_t, $\hat{\mathcal{R}}(\theta_t)$ is computed through (11) (with $\theta = \theta_t$) and estimates obtained through N independent rollouts as in (8), and $\{h_t\}$ is a deterministic sequence of stepsizes. We note that $\hat{\mathcal{R}}(\theta)$ (and $\mathcal{R}(\theta)$) can be computed in closed-form (cf. Mestres et al., 2025b, Lemma 3). We refer to the algorithm that updates policy parameters according to (12) as RSGF-RL. If problem (11) is infeasible at $\theta = \theta_t$, we set $\hat{\mathcal{R}}(\theta_t) := -\widehat{\nabla V}_1(\theta_t)/\beta$. We next investigate the extent to which RSGF-RL retains the safety and convergence properties of the nominal RSGF iterates (5). In the following results, we let $\mathcal{F}_t$ be the filtration obtained from all stochasticity up to iteration t (but not including the rollouts from π_{θ_t}). We begin by establishing a bound on the number of samples required to achieve a probabilistic safety guarantee.

Proposition 2 (Probabilistic safety): *Consider a sequence $\{\theta_t\}_{t\in\mathbb{N}}$ generated by (12), and let $K \in \mathbb{N}$. Suppose that ∇V_1 is L_1-Lipschitz, that the stepsizes satisfy $0 < h_t \leq h_{\max} < \min\{1/\alpha, \beta/L_1\}$, and that, almost surely, (11) is feasible at $\theta = \theta_t$ for every $t \in 0 : K-1$. Fix $\epsilon > 0$ and $\bar{\delta} \in (0, 1)$, and let*

$$N \geq \max\left\{ \tfrac{2B_1^2 \log(4/\bar{\delta})}{\epsilon^2},\ \tfrac{2d\tilde{B}_1^2 \log(4d/\bar{\delta})}{\epsilon^2} \right\}. \tag{13}$$

Define $\Pi_t := \prod_{k=0}^{t-1}(1 - \alpha h_k)$ and

$$\bar{c} := \zeta_1 + \epsilon + \frac{(\sqrt{d}\,\tilde{\zeta}_1 + \epsilon)^2}{2\alpha(\beta - L_1 h_{\max})}. \tag{14}$$

Then, with probability at least $1 - K\bar{\delta}$,

$$V_1(\theta_t) \leq \Pi_t\, V_1(\theta_0) + (1 - \Pi_t)\, \bar{c} \qquad \text{for all } t \in 0 : K. \tag{15}$$

In particular, with probability at least $1 - K\bar{\delta}$: (i) if $V_1(\theta_0) \leq 0$, then $V_1(\theta_t) \leq \bar{c}$ for all $t \in 0 : K$; (ii) if $V_1(\theta_0) > 0$, then $V_1(\theta_t) \leq \Pi_t V_1(\theta_0) + \bar{c}$ for all $t \in 0 : K$.

The proof of Proposition 2 follows an argument similar to that of (Mestres et al., 2025b, Proposition 5) and is included for completeness in Appendix E. Proposition 2 quantifies how estimation error and bias affect the safety of policy updates. It shows that with high probability (which degrades linearly with the safety horizon length K), (i) RSGF-RL

iterates with safe initial policy stay in a neighborhood of the safe set (with the size of the neighborhood determined by $\bar{c}$, which shrinks as the estimator bias decreases and the number of rollouts N increases); (ii) RSGF-RL iterates with unsafe initial policy exponentially converge (e.g. by choosing constant stepsizes) to this neighborhood of the safe set. Note that if we want to ensure exact constraint satisfaction with high probability, we can simply tighten the constraint in (3) by a margin $\bar{c}$.

We next study the convergence properties of (12). For $\theta \in \mathbb{R}^d$ define

$$D(\theta) := \frac{\|\nabla V_1(\theta)\|^2}{\beta^2} - \frac{2\alpha V_1(\theta)}{\beta}, \qquad \hat{D}(\theta) := \frac{\|\widehat{\nabla V}_1(\theta)\|^2}{\beta^2} - \frac{2\alpha \widehat{V}_1(\theta)}{\beta}. \tag{16}$$

As shown in Lemma 8, problem (4) (resp. (11)) is feasible if and only if $D(\theta) \geq 0$ (resp. $\hat{D}(\theta) \geq 0$).

Proposition 3 (Convergence): *Suppose that:*

(i) $\Theta \subset \mathbb{R}^d$ *is compact and convex, and* $\theta_t \in \Theta$ *for all* $t \in \mathbb{N}$ *almost surely;*

(ii) $\nabla V_0, \nabla V_1$ *are* L_0*-,* L_1*-Lipschitz on* Θ*;*

(iii) problem (4) *is feasible for all* $\theta \in \Theta$ *and* $\rho := \min_{\theta\in\Theta} \sqrt{D(\theta)} > 0$*;*

Define the constants $G_j := \max_{\theta\in\Theta} \|\nabla V_j(\theta)\|$ *for* $j \in \{0, 1\}$*,*

$$\bar{u} := \frac{(G_0 + G_1/\beta)^2}{\beta\rho^2}, \qquad c_1 := \frac{1}{\beta} + \frac{\sqrt{d}\,\tilde{B}_1 + G_1}{\beta^2\rho}, \qquad c_V := \frac{2\alpha}{\beta\rho}, \tag{17}$$

$$\varepsilon_N^2 := 3\Big[d\big(\tilde{\zeta}_0^2 + \tfrac{\tilde{B}_0^2}{N}\big) + c_1^2\, d\big(\tilde{\zeta}_1^2 + \tfrac{\tilde{B}_1^2}{N}\big) + c_V^2\big(\zeta_1^2 + \tfrac{B_1^2}{N}\big)\Big]. \tag{18}$$

Fix any $\kappa \geq \bar{u}$*, and let* $\Phi_\kappa(\theta) := V_0(\theta) + \kappa[V_1(\theta)]_+$*,* $G_\Phi := G_0 + \kappa G_1$*,* $L_\Phi := L_0 + \kappa L_1$*,* $\Delta_0 := \Phi_\kappa(\theta_0) - \min_\Theta V_0$*, and* $H_K := \sum_{t=0}^{K-1} h_t$ *for* $K \in \mathbb{N}$*. Let* $\{\theta_t\}$ *be generated by* (12) *with* $0 < h_t \leq \bar{h} := \min\{1/\alpha,\, 1/(2L_\Phi)\}$*. Then:*

(a) (Non-asymptotic bound) *For every* $K \in \mathbb{N}$*,*

$$\frac{1}{H_K}\sum_{t=0}^{K-1} h_t\, \mathbb{E}\|\mathcal{R}(\theta_t)\|^2 \;\leq\; \frac{2\Delta_0}{H_K} + 2G_\Phi\,\varepsilon_N + \frac{2L_\Phi\varepsilon_N^2}{H_K}\sum_{t=0}^{K-1} h_t^2. \tag{19}$$

In particular, for a constant stepsize $h_t \equiv h \leq \bar{h}$*,*

$$\min_{0\leq t<K} \mathbb{E}\|\mathcal{R}(\theta_t)\|^2 \;\leq\; \frac{1}{K}\sum_{t=0}^{K-1} \mathbb{E}\|\mathcal{R}(\theta_t)\|^2 \;\leq\; \frac{2\Delta_0}{hK} + 2G_\Phi\,\varepsilon_N + \varepsilon_N^2. \tag{20}$$

(b) (Almost-sure asymptotic bound) *If* $\sum_t h_t = \infty$*, then almost surely*

$$\limsup_{K\to\infty} \frac{1}{H_K}\sum_{t=0}^{K-1} h_t \|\mathcal{R}(\theta_t)\|^2 \;\leq\; 2G_\Phi\,\varepsilon_N + \varepsilon_N^2, \tag{21}$$

so in particular $\liminf_{t\to\infty} \|\mathcal{R}(\theta_t)\|^2 \leq 2G_\Phi\varepsilon_N + \varepsilon_N^2$ *a.s.*

The proof of Proposition 3 is included in Appendix F. Let us comment on the assumptions of Proposition 3. First, (i) can easily be guaranteed by adding an additional deterministic constraint in (11) that guarantees that Θ is forward invariant, as done in Mestres et al. (2025a). On the other hand, (ii) is standard in the literature, and the Lipschitz constants can even be obtained explicitly as in (Zhang et al., 2020, Lemma 3.2). Finally, since feasibility of problem (4) at parameter θ is equivalent to $D(\theta) \geq 0$, condition (iii) requires that problem (4) has a strict feasibility margin in the set Θ.

Since Θ is compact and $\mathcal{R}$ is continuous, Proposition 3(b) implies that, almost surely, a subsequence of the RSGF-RL iterates (12) converges to a point which satisfies the KKT conditions of (3) up to a tolerance that vanishes with ε_N. Hence, as $\varepsilon_N \to 0$ (equivalently, as the estimator biases approach 0 and the number of rollouts N increases), such points form an approximate KKT sequence as introduced in (Andreani et al., 2011), whose limit points are KKT points of (3) (because assumption (iii) implies MFCQ). In particular, if estimates are constructed so that they become asymptotically unbiased, the approximation error only depends on the number of rollouts N.

4. Experiments

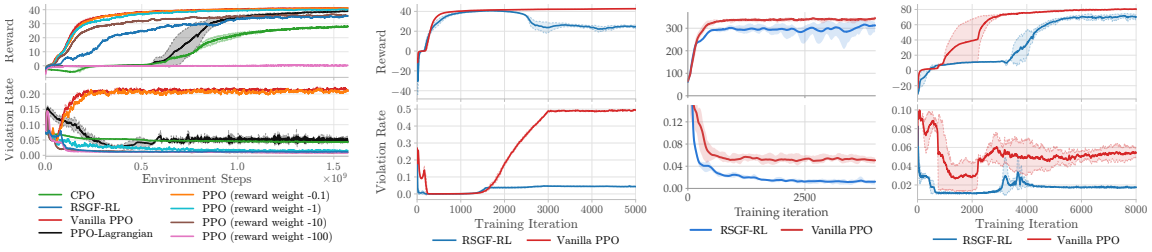
We implement RSGF-RL (12) for humanoid whole-body control and drone obstacle avoidance tasks[2]. For all tasks we use Gaussian policies with mean parametrized by a multi-layer perceptron (MLP). Furthermore, the policy acts on observations rather than on the full state; this fits our framework by augmenting the state with such observations[3]. We obtain estimates $\widehat{\nabla V}_0$ through PPO-clip, and estimates $\widehat{\nabla V}_1, \widehat{V}_1$ through an actor-critic scheme. Both provide biased estimates (cf. Döring et al., 2026, Theorem 4.2 and Kumar et al., 2023, Theorem 2 for bias bounds). Since the choice of g_1 in Example 2 tends to stall learning, possibly due to its sparse nature, we use g_0, g_1 as in Example 1, with R_1 as in (23). Rewards, hyperparameters, and details on how the estimates are computed are reported in Appendices G and H.

4.1. Humanoid whole-body control

We evaluate RSGF-RL on whole-body control tasks for a Unitree G1 humanoid robot, whose high dimensionality, contact-rich dynamics, and underactuation make safety-critical controller design challenging. In all tasks, we train the RL policy in mjlab (Zakka et al., 2026). Simulated dynamics run at 200 Hz and the policy at 50 Hz, with 10 second episodes. The learned policies output joint position set-point commands that are tracked by a low-level PD controller. For sim-to-real transfer, we randomize base disturbances, contact friction, and center-of-mass and joint-encoder estimation errors.

First, we learn a fixed forward-gait walking policy, where the safe set requires a base height of at least 0.67 m, roll in $[-0.35, 0.35]$ rad, pitch in $[-0.5, 0.5]$ rad, and the center of

2. Videos of the different experiments are available at the following link: `https://drive.google.com/file/d/19_EAcHGyLCidhLtGAe3-6qqPSirQZppJ/view?usp=sharing`.
3. In this setting, conditions (6) and (10) do not hold for Gaussian policies with unbounded support, but they hold for Gaussian policies truncated to a compact action set $\mathcal{A}$ with standard deviation bounded away from zero on the compact parameter space Θ and observations bounded on a compact set. Truncating actions and observations on a sufficiently large set retains the same performance in practice.



(*a*) Squatting (*b*) Crawling (*c*) Obstacle Avoid. (*d*) Walking

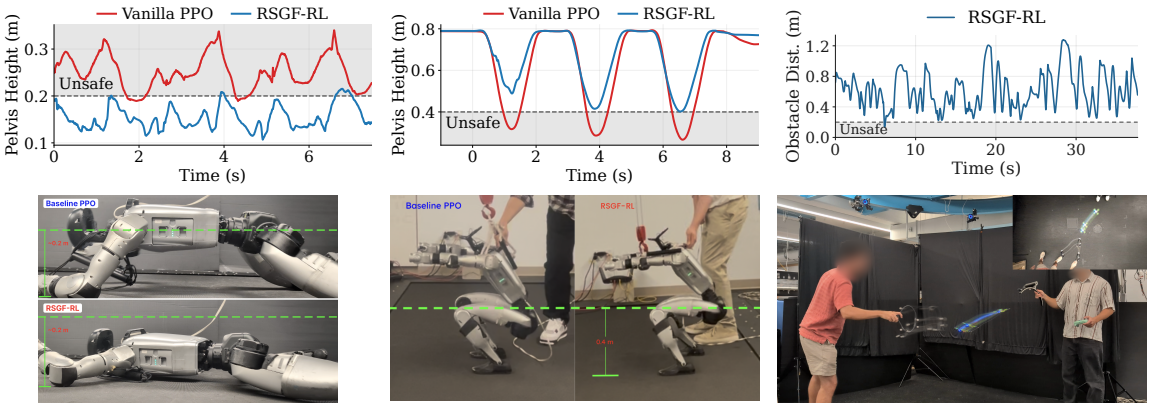
Figure 1: Training curves for vanilla PPO and RSGF-RL. In each panel, the top plot shows the mean reward and the bottom plot the safety violation rate per episode (i.e., the fraction of episodes that contain at least one safety violation). Shaded areas represent ± 1 standard deviation over 5 seeds. Panel (a) also compares against CPO (Achiam et al., 2017), PPO-Lagrangian (Ray et al., 2019) and vanilla PPO with a safety-aware reward penalty for several penalty weights, and uses *environment steps* on the horizontal axis since CPO uses a different number of steps per environment.

pressure (CoP) inside the support polygon of foot contact points. Figure 1(d) shows that, when trained with the same g_0, RSGF-RL learns a policy with substantially fewer safety constraint violations than vanilla PPO, while asymptotically achieving comparable rewards.

Next, we test RSGF-RL on crawling and squatting motion imitation tasks (Figures 2(a)-2(b)), training a separate policy for each with motion imitation rewards (Liao et al., 2026) and reference motions from a human motion dataset (BONES Studio, 2025). We constrain the base height to be at least 0.4 m for squatting, and the base and head heights to be at most 0.2 m for crawling (with a curriculum from 0.3 m to 0.2 m). Figures 1(a)-1(b) show that, with the same g_0, RSGF-RL learns the squatting and crawling motions while respecting the height constraint, whereas vanilla PPO does not. Figure 1(a) also shows that for the squatting motion task, RSGF-RL improves upon CPO (Achiam et al., 2017) both in terms of safety and performance, and upon PPO-Lagrangian (Ray et al., 2019) in safety, with very similar performance. Additionally, vanilla PPO with an added safety-aware reward has safety and performance similar to RSGF-RL, as long as the penalty weight of the safety-aware reward is chosen appropriately. However, the correct penalty weight is very task-dependent, and it is unclear how to select it a priori.

4.2. Drone collision avoidance

We learn a policy that steers a Crazyflie drone to a hover target that changes every 5 seconds, in an arena with a randomized number of obstacles, some of which move randomly and others toward the drone. For sim-to-real transfer, we use domain randomization with random perturbations on the velocity, thrust coefficient, drag, command delays, and low-level controller feedback delay. The learned policy outputs velocity commands that are tracked by the drone's onboard low-level controller. We train the RL policy in IsaacLab (Mittal et al., 2025). The simulated dynamics run at 400 Hz and the policy at 100 Hz, with 10 second episodes. In this case, we add safety penalties to vanilla PPO for entering, or being predicted to enter within 0.5 seconds, the safety margin of obstacles or the arena boundary, with a terminal penalty for collisions or leaving the arena. Figure 1(c) again shows that RSGF-RL achieves a task return comparable to vanilla PPO while maintaining a low safety



(a) Crawling (b) Squatting (c) Obstacle Avoidance

Figure 2: (a)–(b) Humanoid crawling and squatting: vanilla PPO violates the safety constraints whereas RSGF-RL does not. (c) The RSGF-RL drone policy avoids two approaching obstacles.

violation rate throughout training, whereas vanilla PPO sacrifices safety for higher return. Increasing these penalties made vanilla PPO training unstable. Although more careful tuning might improve its safety, RSGF-RL has no penalty parameter, and its own parameters $(\alpha, \beta, \{h_t\})$ required little tuning in this experiment. On hardware, the RSGF-RL policy successfully avoids obstacles while staying within the arena boundary (Fig. 2(c)).

4.3. Discussion on theory to practice gap

Some assumptions of our results, such as the bound (13) on the number of rollouts and the stepsize condition $0 < h_t \leq \bar{h}$, are not met in the experiments, and the probabilistic safety guarantees in Propositions 2 and 4 are loose compared to the safety achieved in practice.

Moreover, our safety and convergence results depend on the estimator biases $\zeta_1, \tilde{\zeta}_0, \tilde{\zeta}_1$. The estimates used in Section 4, detailed in Appendix H, rely on standard practical RL techniques, such as learned value critics, multiple Adam steps on the PPO loss, and omitting the discount weighting of the policy gradient, each of which introduces bias. Although techniques such as those in (Döring et al., 2026), (Kumar et al., 2023) can be used to bound these biases, we do not have tight characterizations for them. Finally, the curriculum in the crawling task makes the constraint iteration dependent, and is not directly covered by our analysis. However, Proposition 2 still applies for policy updates with the same constraint.

5. Conclusions

We have extended RSGF-RL to a general class of constraints, including probabilistic and cumulative ones, and to biased gradient and value function estimators, enabling its integration with state-of-the-art RL algorithms such as PPO. We established safety and convergence guarantees for the resulting algorithm, and demonstrated it on whole-body motor skills for a G1 humanoid robot and collision-free waypoint navigation for a Crazyflie drone. Future work will refine our assumptions and bias bounds and design estimators with good practical performance and provable guarantees, in order to narrow the gap between theory and practice discussed in Section 4.3.

Acknowledgments

The authors used various AI assistants to write code, check arguments, and edit the paper. The authors verified all results in the paper and take full responsibility for the content.

References

J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 22–31, Sydney, Australia, 2017.

M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, and S. Niekum. Safe reinforcement learning via shielding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 2669–2678, 2018.

E. Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

R. Andreani, G. Haeser, and J. M. Martínez. On sequential optimality conditions for smooth constrained optimization. *Optimization*, 60(5):627–641, 2011.

T. M. Apostol. *Mathematical analysis*. Addison-Wesley, 1974.

Q. Bai, W. U. Mondal, and V. Aggarwal. Regret analysis of policy gradient algorithm for infinite horizon average reward Markov decision processes. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 10980–10988, Vancouver, Canada, 2024.

H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Nature, 1991.

F. P. Bejarano, L. Brunke, and A. P. Schoellig. Safety filtering while training: improving the performance and sample efficiency of reinforcement learning agents. *IEEE Robotics and Automation Letters*, 10(1):788–795, 2025.

BONES Studio. BONES-SEED. `https://huggingface.co/datasets/bones-studio/seed`, 2025.

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*, 5:411–444, 2022.

W. Chen, D. Subramanian, and S. Paternain. Probabilistic Constraint for Safety-Critical Reinforcement Learning. *IEEE Transactions on Automatic Control*, 69(10):6789–6804, 2024.

R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 3387–3395, 2019.

Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1): 6070–6120, 2017.

D. Ding, K. Zhang, T. Basar, and M. Jovanovic. Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pages 8378–8390, Online Conference, 2020.

R. Durrett. *Probability: Theory and Examples*. Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5th edition, 2010. ISBN 9781108473682.

L. Döring, D. Schmidt, M. Melcher, S. Kassing, B. Wille, T. Aach, and S. Weissmann. An approximate ascent approach to prove convergence of PPO. *arXiv preprint arXiv:2602.03386*, 2026.

M. Fréchet. Généralizations du théorème des probabilités totales. *Fundamenta Mathematicae*, 1(25):379–387, 1935.

J. Garcia and F. Fernandez. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16:1437–1480, 2015.

P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll. A review of safe reinforcement learning: methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11216 – 11235, 2024.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 36455–36485, 2024.

D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

K. Hsu, H. Hu, and J. F. Fisac. The safety filter: a unified view of safety-critical control in autonomous systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 7: 47–72, 2024.

J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26): eaau5872, 2019.

E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620:982–987, 2023.

V. R. Konda and J. N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.

H. Kumar, A. Koppel, and A. Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning*, 112:2433–2467, 2023.

K. Li, Z. Olkin, Y. Yue, and A. D. Ames. CLF-RL: Control Lyapunov function guided reinforcement learning. *IEEE Robotics and Automation Letters*, 2026.

Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and K. C. Liu. BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.

Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *Science Robotics*, 11(117):eadx8924, 2026.

P. Mestres, A. Marzabal, and J. Cortés. Off-Policy Reinforcement Learning with Anytime Safety Guarantees via Robust Safe Gradient Flow. 2025a. Available at https://arxiv.org/pdf/2510.01492.

P. Mestres, A. Marzabal, and J. Cortés. Anytime safe reinforcement learning. In *Learning for Dynamics and Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pages 221–232. 2025b.

Mayank Mittal et al. Isaac lab: A GPU-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. URL `https://arxiv.org/abs/2511.04831`.

S. Mu and D. Klabjan. On the second-order convergence of biased policy gradient algorithms. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 36455–36485, 2024.

Siddharth Mysore, Bassel Mabsout, Renato Mancuso, and Kate Saenko. Regularizing action policies for smooth control with reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1810–1816. IEEE, 2021.

Y. Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer International Publishing, 2nd edition, 2018. ISBN 3319915770.

OpenAI. Safety starter agents. `https://github.com/openai/safety-starter-agents`, 2019. Accessed November 2026.

OpenAI, M. Andrychowitz, and B. Baker et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro. Constrained Reinforcement Learning Has Zero Duality Gap. In *Proceedings of the 33rd Conference in Neural Information Processing Systems*, volume 32, pages 7555–7565, Vancouver, Canada, 2019.

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*, 68(3):1321–1336, 2023.

A. Ray, J. Achiam, and D. Amodei. Benchmarking safe exploration in deep reinforcement learning. Technical report, OpenAI, 2019. URL `https://cdn.openai.com/safexp-short.pdf`.

N. Rudin, D. Hoeller, P. Reist, and M. Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning (CoRL)*, pages 91–100, 2022.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. 2017. Available at https://arxiv.org/pdf/1707.06347.

X. Shen, S. Jiang, A. Wachi, K. Hashimoto, and S. Gros. Flipping-based policy for chance-constrained Markov decision processes. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 64377–64411, 2024.

J. Siekmann, Y. Godse, A. Fern, and J. Hurst. Sim-to-real learning of all common bipedal gaits via periodic reward composition. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 7309–7315, 2021.

A. Sootla, A. I. Cowen-Rivers, T. Jafferjee, Z. Wang, D. H. Mguni, J. Wang, and H. Ammar. Saute RL: almost surely safe reinforcement learning using state augmentation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20423–20442, 2022.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

L. Yang, B. Werner, M. de Sa, and A. D. Ames. CBF-RL: Safety filtering reinforcement learning in training with control barrier functions. In *IEEE International Conference on Robotics and Automation*, pages 17825–17823, 2026.

T. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge. Projection-based constrained policy optimization. In *Proceedings of the 8th International Conference on Learning Representations*, Online Conference, 2020.

C. Ying, X. Zhou, D. Yan, and J. Zhu. Towards safe reinforcement learning via constraining conditional value at risk. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

K. Zakka, Q. Liao, B. Yi, L. L. Lay, K. Sreenath, and P. Abbeel. mjlab: A lightweight framework for GPU-accelerated robot learning. *arXiv preprint arXiv:2601.22074*, 2026.

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*, 58(6): 3586–3612, 2020.

Appendix A. Connection between cumulative and probabilistic constraint

Here we show that, with an appropriate choice of the cost function R_1, the cumulative constraint $V_1(\theta) \leq 0$ induced by the choice of g_1 in Example 1 implies (up to a term that vanishes as $\gamma \to 1$) the probabilistic constraint in (2).

Consider

$$R_1(s_t, a_t, s_{t+1}) = -\mathbb{I}(s_t \in \bar{\mathcal{C}}) + 1 - \frac{\delta}{T+1}, \tag{22}$$

which assigns a cost $1 - \frac{\delta}{T+1}$ to unsafe states and $-\frac{\delta}{T+1}$ to safe ones. In practice, we use a smoothened version of (22) given by

$$R_1(s_t, a_t, s_{t+1}) = \begin{cases} \frac{\delta}{T+1}\big(e^{-\tilde{h}(s_t)} - 1\big), & \text{if } s_t \in \bar{\mathcal{C}}, \\ 1 - \frac{\delta}{T+1}, & \text{else}, \end{cases} \tag{23}$$

where $\tilde{h} : \mathcal{S} \to \mathbb{R}$ is a function whose 0-superlevel set includes $\bar{\mathcal{C}}$. Since $\tilde{h}(s_t) \geq 0$ for $s_t \in \bar{\mathcal{C}}$, (23) upper bounds (22) pointwise, while additionally encouraging the state to stay away from the boundary of $\bar{\mathcal{C}}$. For $\gamma \in (0, 1]$, define

$$\kappa(\gamma) = \frac{1}{T+1} \sum_{k=0}^{T} \gamma^{-k} = \frac{1 - \gamma^{T+1}}{(T+1)(1-\gamma)\gamma^T} \quad (\gamma < 1), \tag{24}$$

$\kappa(1) = 1$, and $\kappa(\gamma) \geq 1$ for all $\gamma \in (0, 1]$.

Proposition 4 *Let $\gamma \in (0, 1]$, let g_1 be as in Example 1, and let R_1 be either* (22) *or* (23)*. If $V_1(\theta) \leq 0$, then*

$$\mathbb{P}\Big(\bigcap_{t=0}^{T} \{s_t \in \bar{\mathcal{C}}\}\Big) \geq 1 - \delta - \varepsilon(\gamma), \qquad \varepsilon(\gamma) = \delta\big(\kappa(\gamma) - 1\big) \leq \frac{\delta\, T(1-\gamma)}{2\gamma^T}. \tag{25}$$

In particular, for $\gamma = 1$ the probabilistic constraint in (2) *holds exactly, and for fixed T its violation vanishes linearly in $1 - \gamma$ as $\gamma \to 1$.*

Proof Let $U_t = \mathbb{I}(s_t \notin \bar{\mathcal{C}})$, so that (22) reads $R_1 = U_t - \frac{\delta}{T+1}$. Since (23) upper bounds (22), in both cases

$$0 \geq V_1(\theta) \geq \mathbb{E}\Big[\sum_{t=0}^{T} \gamma^t U_t\Big] - \frac{\delta}{T+1} \sum_{t=0}^{T} \gamma^t. \tag{26}$$

Using the union bound on probabilities (Durrett, 2010, Theorem 1.1.1) and $\gamma^t \geq \gamma^T$ for $t \leq T$,

$$\mathbb{P}\Big(\bigcup_{t=0}^{T} \{s_t \notin \bar{\mathcal{C}}\}\Big) \leq \mathbb{E}\Big[\sum_{t=0}^{T} U_t\Big] \leq \gamma^{-T}\, \mathbb{E}\Big[\sum_{t=0}^{T} \gamma^t U_t\Big] \leq \frac{\delta}{T+1} \sum_{t=0}^{T} \gamma^{t-T} = \delta\, \kappa(\gamma),$$

where in the last inequality we used (26), and gives the first part of (25). For the bound on $\varepsilon(\gamma)$, note that $1-\gamma^k = (1-\gamma)\sum_{i=0}^{k-1}\gamma^i \le k(1-\gamma)$ and $\gamma^k \ge \gamma^T$, so $\gamma^{-k}-1 = \frac{1-\gamma^k}{\gamma^k} \le \frac{k(1-\gamma)}{\gamma^T}$. Hence

$$\kappa(\gamma) - 1 = \frac{1}{T+1}\sum_{k=0}^{T}\left(\gamma^{-k}-1\right) \le \frac{1-\gamma}{(T+1)\gamma^T}\sum_{k=0}^{T} k = \frac{T(1-\gamma)}{2\gamma^T}.$$

■

Remarks. The case $\gamma = 1$ with (22) was established in (Paternain et al., 2023, Theorem 1). However, in our experiments we found that taking $\gamma < 1$, together with the choice of R_1 in (23), leads to better practical performance. Proposition 4 shows that the price paid for this choice is a violation of the probabilistic constraint of at most $\varepsilon(\gamma) = \mathcal{O}\big(\delta T(1-\gamma)\big)$. If exact satisfaction is required for $\gamma < 1$, it suffices to replace the constant $\frac{\delta}{T+1}$ in (22) and (23) by $\frac{\delta}{T+1}\kappa(\gamma)^{-1} = \frac{\delta\gamma^T(1-\gamma)}{1-\gamma^{T+1}}$, at the cost of a more conservative constraint. Finally, since Proposition 4 only provides a sufficient condition for the satisfaction of the probabilistic constraint (2), (assuming $\gamma = 1$ or correct constraint tightening), the two formulations are not equivalent.

Appendix B. Properties of the RSGF

We next summarize the properties of (5), adapted from (Mestres et al., 2025b, Lemma 1) and (Mestres et al., 2025a, Lemma IV.2, Prop. IV.3).

Proposition 5 (RSGF Properties): *Suppose that* (3) *satisfies the Mangasarian–Fromovitz constraint qualification (MFCQ)*[4] *at every* $\theta \in \mathcal{C}$. *Then* (4) *is well-defined on an open neighborhood of* $\mathcal{C}$. *Moreover,* $\theta \in \mathcal{C}$ *is a KKT point of* (3) *if and only if* $\mathcal{R}(\theta) = 0$. *Furthermore, suppose that* ∇V_1 *is* L_1*-Lipschitz and that* $h_t < \min\{1/\alpha, \beta/L_1\}$ *for all* $t \in \mathbb{N}$. *Then* $\mathcal{C}$ *is forward invariant under* (5). *If, in addition,* ∇V_0 *is* L_0*-Lipschitz and* $h_t < 1/L_0$, *then every limit point of a sequence* $\{\theta_t\}_{t\in\mathbb{N}}$ *initialized at* $\theta_0 \in \mathcal{C}$ *is a KKT point of* (3).

Proposition 5 is a direct consequence of (Mestres et al., 2025b, Lemma 1) and (Mestres et al., 2025a, Lemma IV.2). However, we note that (Mestres et al., 2025a, Lemma IV.2) requires that (4) satisfies the Constant Rank Condition (CRC) at $(\theta, \mathcal{R}(\theta))$ for every $\theta \in \mathcal{C}$, but that hypothesis is not stated in the statement of Proposition 5.

The following result explains the reason behind this mismatch: given that (3) only has a single constraint, the satisfaction of MFCQ for (3) at every $\theta \in \mathcal{C}$ implies the satisfaction of CRC for (4) at $(\theta, \mathcal{R}(\theta))$ for every $\theta \in \mathcal{C}$.

Lemma 6 (MFCQ implies CRC for single constraint problems): *Suppose that* (3) *satisfies MFCQ for each* $\theta \in \mathcal{C}$. *Then,* (4) *satisfies CRC at* $(\theta, \mathcal{R}(\theta))$ *for every* $\theta \in \mathcal{C}$.

4. For the single-constraint problem considered here, MFCQ requires that, for every $\theta \in \mathcal{C}$ satisfying $V_1(\theta) = 0$, there exists $q \in \mathbb{R}^d$ such that $\nabla V_1(\theta)^\top q < 0$. Equivalently, $\nabla V_1(\theta) \neq 0$ whenever $V_1(\theta) = 0$.

Proof Suppose that CRC does not hold for (4) at some $\theta \in \mathcal{C}$. Then we necessarily have that

$$\alpha V_1(\theta) + \nabla V_1(\theta)^\top \mathcal{R}(\theta) + \frac{\beta}{2} \|\mathcal{R}(\theta)\|^2 = 0, \tag{27}$$

and

$$\frac{\partial}{\partial \xi}\Big(\alpha V_1(\theta) + \nabla V_1(\theta)^\top \xi + \frac{\beta}{2} \|\xi\|^2\Big)\Big|_{\xi=\mathcal{R}(\theta)} = \nabla V_1(\theta) + \beta \mathcal{R}(\theta) = 0. \tag{28}$$

Plugging (28) into (27) we get

$$\alpha V_1(\theta) - \frac{\beta}{2} \|\mathcal{R}(\theta)\|^2 = 0. \tag{29}$$

Since $\theta \in \mathcal{C}$, we have $V_1(\theta) \leq 0$. Hence, (29) implies that $V_1(\theta) = 0$ and $\mathcal{R}(\theta) = 0$, which from (28) implies that $\nabla V_1(\theta) = 0$, violating MFCQ for (3) at $\theta \in \mathcal{C}$. Hence, CRC necessarily holds for (4) at $(\theta, \mathcal{R}(\theta))$ for every $\theta \in \mathcal{C}$. ■

Appendix C. Proof of Proposition 1

When g_j is a discounted cumulative return, as in Example 1, and $b_{j,t} \equiv 0$, Proposition 1 recovers the standard Policy Gradient Theorem (Sutton and Barto, 2018, Section 13.2). If the baseline is time-independent and depends only on s_t, it recovers the corresponding baseline formulation (Sutton and Barto, 2018, Section 13.4). Likewise, for the probabilistic constraint in Example 2 with $b_{j,t} \equiv 0$, it reduces to (Chen et al., 2024, Theorem 2). As usual, the baselines $\{b_{j,t}\}$ leave the expected gradient unchanged (as shown next in the proof) while potentially reducing its variance.

Proof The proof is standard and follows arguments analogous to (Glynn, 1990) and the Policy Gradient Theorem (Sutton and Barto, 2018, Chapter 13). We include it here for completeness. Let us first show that

$$\nabla V_j(\theta) = \mathbb{E}\Big[\sum_{t=0}^{T} g_j(s,a) \nabla_\theta \log \pi_\theta(a_t|s_t)\Big]. \tag{30}$$

Let $\sigma : \mathcal{S}^{T+2} \times \mathcal{A}^{T+1} \times \mathbb{R}^d \to \mathbb{R}$ be defined as

$$\sigma(s,a,\theta) = \eta(s_0) \prod_{t=0}^{T} \pi_\theta(a_t|s_t) P(s_{t+1}|a_t, s_t).$$

Since the hypothesis for differentiation under the integral sign hold (Apostol, 1974, Theorem 10.39) (because $\theta \to \log \pi_\theta(a|s)$ is continuously differentiable and for each $\theta \in \mathbb{R}^d$, (6) holds), we have

$$\nabla V_j(\theta) = \int_\Gamma g_j(s,a) \sigma(s,a,\theta) \nabla_\theta \log \sigma(s,a,\theta) ds da,$$

where $\Gamma = \mathcal{S}^{T+2} \times \mathcal{A}^{T+1}$. Now, note that since

$$\log \sigma(s, a, \theta) = \log \eta(s_0) + \sum_{t=0}^{T} \Big(\log P(s_{t+1}|a_t, s_t) + \log \pi_\theta(a_t|s_t) \Big),$$

we have

$$\nabla V_j(\theta) = \int_\Gamma g_j(s, a) \sigma(s, a, \theta) \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) ds da,$$

proving (30). Next, let us show that for each $t \in 0 : T$, we have

$$\mathbb{E}[b_{j,t}(H_t, \theta) \nabla_\theta \log \pi_\theta(a_t|s_t)] = 0,$$

which will complete the proof. Using the tower property of conditional expectation (Williams, 1991, Section 9.7),

$$\mathbb{E}[b(H_t, \theta) \nabla_\theta \log \pi_\theta(a_t|s_t)] = \mathbb{E}_{H_t}[\mathbb{E}_{a_t}[b(H_t, \theta) \nabla_\theta \log \pi_\theta(a_t|s_t)|H_t]],$$

where on the right hand side expression, the outer expectation is with respect to the history H_t and the inner expectation is with respect to action a_t only. Now, note that for any fixed H_t, since H_t does not depend on a_t and $a_t \sim \pi_\theta(\cdot|s_t)$, we have

$$\begin{aligned} \mathbb{E}_{a_t}[b(H_t, \theta) \nabla_\theta \log \pi_\theta(a_t|s_t)|H_t] &= b(H_t, \theta) \int_{\mathcal{A}} \nabla_\theta \log \pi_\theta(a_t|s_t) \pi_\theta(a_t|s_t) da_t \\ &= b(H_t, \theta) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a_t|s_t) da_t = 0, \end{aligned}$$

where in the last equality we have used the fact that since $1 = \int_{\mathcal{A}} \pi_\theta(a_t|s_t) da_t$, by differentiating under the integral sign we have $0 = \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a_t|s_t) da_t$. ■

Appendix D. Statistical properties of estimators

The following result establishes different statistical properties of the estimators (8), which are needed for the proof of Proposition 2.

Proposition 7 (Statistical properties of estimators): *Let $j \in \{0, 1\}$, $\theta \in \mathbb{R}^d$, and $\widehat{V}_j(\theta)$, $\widehat{\nabla V}_j(\theta)$ be defined as in* (8)*, and suppose that conditions* (10), (9) *hold. Then,*

$$\begin{aligned} &|\mathbb{E}[\widehat{V}_j(\theta)] - V_j(\theta)| \leq \zeta_j, \ \Big\| \mathbb{E}[\widehat{\nabla V}_j(\theta)] - \nabla V_j(\theta) \Big\| \leq \sqrt{d} \tilde{\zeta}_j, \\ &\mathbb{P}\Big(|\widehat{V}_j(\theta) - V_j(\theta)| \leq \zeta_j + \epsilon \Big) \geq 1 - 2 \exp\Big\{ - \frac{N \epsilon^2}{2 B_j^2} \Big\}, \\ &\mathbb{P}\Big(\Big\| \widehat{\nabla V}_j(\theta) - \nabla V_j(\theta) \Big\| \leq \sqrt{d} \tilde{\zeta}_j + \epsilon \Big) \geq 1 - 2d \exp\Big\{ - \frac{N \epsilon^2}{2 d \tilde{B}_j^2} \Big\}, \end{aligned}$$

for all $\epsilon > 0$.

Proof The proof of Proposition 7 follows an argument analogous to that of (Mestres et al., 2025b, Lemma 2), but we include it here for completeness. Fix $j \in \{0,1\}$ and $\theta \in \mathbb{R}^d$. We first establish the bounds on the bias. By definition of $\widehat{V}_j(\theta)$ and condition (9),

$$\left|\mathbb{E}[\widehat{V}_j(\theta)] - V_j(\theta)\right| = \left|\frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}[G^\theta_{j,n}] - V_j(\theta)\right)\right| \leq \frac{1}{N}\sum_{n=1}^{N}\left|\mathbb{E}[G^\theta_{j,n}] - V_j(\theta)\right| \leq \zeta_j.$$

Similarly, for every $i \in 1:d$,

$$\left|\left(\mathbb{E}[\widehat{\nabla V_j}(\theta)] - \nabla V_j(\theta)\right)_i\right| = \left|\frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}[(\tilde{G}^\theta_{j,n})_i] - (\nabla V_j(\theta))_i\right)\right| \leq \tilde{\zeta}_j.$$

Consequently,

$$\left\|\mathbb{E}[\widehat{\nabla V_j}(\theta)] - \nabla V_j(\theta)\right\| \leq \sqrt{\sum_{i=1}^{d}\tilde{\zeta}_j^2} = \sqrt{d}\,\tilde{\zeta}_j.$$

We next establish the concentration bounds. Since the N episodes are independent, the random variables $\{G^\theta_{j,n}\}_{n=1}^N$ are independent. Moreover, by (10), $G^\theta_{j,n} \in [-B_j, B_j]$ almost surely. Hence, Hoeffding's inequality (Hoeffding, 1963) yields, for every $\epsilon > 0$,

$$\mathbb{P}\left(\left|\widehat{V}_j(\theta) - \mathbb{E}[\widehat{V}_j(\theta)]\right| > \epsilon\right) \leq 2\exp\left\{-\frac{N\epsilon^2}{2B_j^2}\right\}.$$

Using

$$\left|\widehat{V}_j(\theta) - V_j(\theta)\right| \leq \left|\widehat{V}_j(\theta) - \mathbb{E}[\widehat{V}_j(\theta)]\right| + \left|\mathbb{E}[\widehat{V}_j(\theta)] - V_j(\theta)\right| \leq \left|\widehat{V}_j(\theta) - \mathbb{E}[\widehat{V}_j(\theta)]\right| + \zeta_j,$$

we obtain

$$\mathbb{P}\left(\left|\widehat{V}_j(\theta) - V_j(\theta)\right| \leq \zeta_j + \epsilon\right) \geq 1 - 2\exp\left\{-\frac{N\epsilon^2}{2B_j^2}\right\}.$$

We finally consider the gradient estimator. For every $i \in 1:d$, condition (10) implies $(\widetilde{G}^\theta_{j,n})_i \in [-\tilde{B}_j, \tilde{B}_j]$ almost surely. Thus, Hoeffding's inequality (Hoeffding, 1963) gives

$$\mathbb{P}\left(\left|(\widehat{\nabla V_j}(\theta))_i - (\mathbb{E}[\widehat{\nabla V_j}(\theta)])_i\right| > \frac{\epsilon}{\sqrt{d}}\right) \leq 2\exp\left\{-\frac{N\epsilon^2}{2d\tilde{B}_j^2}\right\}.$$

Therefore, by Fréchet's inequality (Fréchet, 1935),

$$\mathbb{P}\left(\bigcap_{i=1}^{d}\left\{\left|(\widehat{\nabla V_j}(\theta))_i - (\mathbb{E}[\widehat{\nabla V_j}(\theta)])_i\right| \leq \frac{\epsilon}{\sqrt{d}}\right\}\right) \geq 1 - 2d\exp\left\{-\frac{N\epsilon^2}{2d\tilde{B}_j^2}\right\}. \tag{31}$$

Now, if the event inside the probability in (31) holds, we have that

$$\left\|\widehat{\nabla V_j}(\theta) - \mathbb{E}[\widehat{\nabla V_j}(\theta)]\right\| \leq \sqrt{\sum_{i=1}^{d}\frac{\epsilon^2}{d}} = \epsilon.$$

Combining this inequality with the bias bound established above gives

$$\left\|\widehat{\nabla V_j}(\theta) - \nabla V_j(\theta)\right\| \le \left\|\widehat{\nabla V_j}(\theta) - \mathbb{E}[\widehat{\nabla V_j}(\theta)]\right\| + \left\|\mathbb{E}[\widehat{\nabla V_j}(\theta)] - \nabla V_j(\theta)\right\| \le \epsilon + \sqrt{d}\,\tilde{\zeta}_j.$$

Hence,

$$\mathbb{P}\left(\left\|\widehat{\nabla V_j}(\theta) - \nabla V_j(\theta)\right\| \le \sqrt{d}\,\tilde{\zeta}_j + \epsilon\right) \ge 1 - 2d\exp\left\{-\frac{N\epsilon^2}{2d\tilde{B}_j^2}\right\},$$

which completes the proof. ■

Appendix E. Proof of Proposition 2

Proof Throughout, fix $t \in 0 : K-1$ and let $\Delta_t := \theta_{t+1} - \theta_t = h_t\hat{\mathcal{R}}(\theta_t)$.

Step 1 (one-step safety evolution). Since ∇V_1 is L_1-Lipschitz, by (Nesterov, 2018, Lemma 1.2.3) we have

$$V_1(\theta_{t+1}) \le V_1(\theta_t) + \nabla V_1(\theta_t)^\top \Delta_t + \frac{L_1}{2}\|\Delta_t\|^2.$$

Adding and subtracting $\widehat{\nabla V}_1(\theta_t)^\top \Delta_t$ and using the Cauchy–Schwarz inequality,

$$V_1(\theta_{t+1}) \le V_1(\theta_t) + \widehat{\nabla V}_1(\theta_t)^\top \Delta_t + \left\|\nabla V_1(\theta_t) - \widehat{\nabla V}_1(\theta_t)\right\| \|\Delta_t\| + \frac{L_1}{2}\|\Delta_t\|^2.$$

Since (11) is feasible at θ_t, $\hat{\mathcal{R}}(\theta_t)$ satisfies its constraint; multiplying it by $h_t > 0$ and using $\Delta_t = h_t\hat{\mathcal{R}}(\theta_t)$ gives $\widehat{\nabla V}_1(\theta_t)^\top \Delta_t \le -\alpha h_t \widehat{V}_1(\theta_t) - \frac{\beta}{2h_t}\|\Delta_t\|^2$. Using also the identity $V_1(\theta_t) - \alpha h_t \widehat{V}_1(\theta_t) = (1-\alpha h_t)V_1(\theta_t) - \alpha h_t\big(\widehat{V}_1(\theta_t) - V_1(\theta_t)\big)$, we obtain

$$\begin{aligned} V_1(\theta_{t+1}) \le{}& (1-\alpha h_t)V_1(\theta_t) + \alpha h_t\left|\widehat{V}_1(\theta_t) - V_1(\theta_t)\right| \\ &+ \left\|\nabla V_1(\theta_t) - \widehat{\nabla V}_1(\theta_t)\right\| \|\Delta_t\| - \frac{1}{2}\Big(\frac{\beta}{h_t} - L_1\Big)\|\Delta_t\|^2. \end{aligned} \tag{32}$$

Inequality (32) holds deterministically, i.e., for every realization of the N rollouts collected at iteration t.

Step 2 (eliminating the step length). Since $h_t < \beta/L_1$, the coefficient $c_t := \frac{1}{2}(\beta/h_t - L_1)$ is positive. For every $e \ge 0$, we have $\max_{x\ge 0}\{ex - c_t x^2\} = e^2/(4c_t)$. Hence the last two terms of (32) are at most $h_t\|\nabla V_1(\theta_t) - \widehat{\nabla V}_1(\theta_t)\|^2/(2(\beta - L_1 h_t))$, and

$$V_1(\theta_{t+1}) \le (1-\alpha h_t)V_1(\theta_t) + \alpha h_t\left|\widehat{V}_1(\theta_t) - V_1(\theta_t)\right| + \frac{h_t\left\|\nabla V_1(\theta_t) - \widehat{\nabla V}_1(\theta_t)\right\|^2}{2(\beta - L_1 h_{\max})}, \tag{33}$$

where we also used $h_t \le h_{\max}$.

Step 3 (bounding estimates deviation from true values). Define the events

$$\begin{aligned} \mathcal{E}_{V,t} &:= \Big\{\left|\widehat{V}_1(\theta_t) - V_1(\theta_t)\right| \le \zeta_1 + \epsilon\Big\}, \\ \mathcal{E}_{\nabla,t} &:= \Big\{\left\|\widehat{\nabla V}_1(\theta_t) - \nabla V_1(\theta_t)\right\| \le \sqrt{d}\,\tilde{\zeta}_1 + \epsilon\Big\}, \end{aligned}$$

and $\mathcal{E}_t := \mathcal{E}_{V,t} \cap \mathcal{E}_{\nabla,t}$. Note that $\epsilon > 0$ is fixed before sampling. Moreover, θ_t is $\mathcal{F}_t$-measurable and, conditionally on $\mathcal{F}_t$, the N episodes collected at iteration t are independent and generated under π_{θ_t}. Hence Proposition 7 applies conditionally on $\mathcal{F}_t$ with $\theta = \theta_t$, and yields

$$\mathbb{P}\big(\mathcal{E}_{V,t} \,\big|\, \mathcal{F}_t\big) \geq 1 - 2\exp\Big\{-\frac{N\epsilon^2}{2B_1^2}\Big\} \geq 1 - \frac{\bar{\delta}}{2},$$
$$\mathbb{P}\big(\mathcal{E}_{\nabla,t} \,\big|\, \mathcal{F}_t\big) \geq 1 - 2d\exp\Big\{-\frac{N\epsilon^2}{2d\tilde{B}_1^2}\Big\} \geq 1 - \frac{\bar{\delta}}{2},$$

where the second inequality in each line follows from (13). By Fréchet's inequality (Fréchet, 1935), $\mathbb{P}(\mathcal{E}_t \mid \mathcal{F}_t) \geq 1 - \bar{\delta}$. By the union bound (cf. Durrett, 2010, Theorem 1.1.1) and the tower property (cf. Williams, 1991, Section 9.7),

$$\mathbb{P}\Big(\bigcup_{t=0}^{K-1} \mathcal{E}_t^c\Big) \leq \sum_{t=0}^{K-1} \mathbb{P}(\mathcal{E}_t^c) = \sum_{t=0}^{K-1} \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_t^c}\big] = \sum_{t=0}^{K-1} \mathbb{E}\Big[\mathbb{E}\big[\mathbf{1}_{\mathcal{E}_t^c} \,\big|\, \mathcal{F}_t\big]\Big] = \sum_{t=0}^{K-1} \mathbb{E}\big[\mathbb{P}(\mathcal{E}_t^c \mid \mathcal{F}_t)\big] \leq K\bar{\delta}, \tag{34}$$

Step 4 (conclusion). On $\mathcal{E}_t$, inequality (33) and the definition (14) of $\bar{c}$ give

$$V_1(\theta_{t+1}) \leq (1 - \alpha h_t)V_1(\theta_t) + \alpha h_t(\zeta_1 + \epsilon) + \frac{h_t(\sqrt{d}\,\tilde{\zeta}_1 + \epsilon)^2}{2(\beta - L_1 h_{\max})} = (1 - \alpha h_t)V_1(\theta_t) + \alpha h_t\,\bar{c},$$

or, equivalently, $V_1(\theta_{t+1}) - \bar{c} \leq (1 - \alpha h_t)\big(V_1(\theta_t) - \bar{c}\big)$. On $\bigcap_{t=0}^{K-1} \mathcal{E}_t$, which by (34) has probability at least $1 - K\bar{\delta}$, this holds for all $t \in 0 : K-1$. Since $0 < 1 - \alpha h_t < 1$ (because $0 < h_t < 1/\alpha$), iterating it yields $V_1(\theta_t) - \bar{c} \leq \Pi_t\big(V_1(\theta_0) - \bar{c}\big)$ for all $t \in 0 : K$, which is (15). Item (i) follows because $\Pi_t V_1(\theta_0) \leq 0$ and $1 - \Pi_t \leq 1$ when $V_1(\theta_0) \leq 0$, and item (ii) follows from $1 - \Pi_t \leq 1$. ■

Appendix F. Proof of Proposition 3

Throughout this section, we let $[x]_+ = \max\{x, 0\}$, $\mathcal{R}_t := \mathcal{R}(\theta_t)$, $\hat{\mathcal{R}}_t := \hat{\mathcal{R}}(\theta_t)$, $E_t := \|\hat{\mathcal{R}}_t - \mathcal{R}_t\|$, and $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$ in order to simplify notation. We also drop the dependency in θ wherever it is clear. Given $p \in \mathbb{R}^d$ and a compact set $A \subset \mathbb{R}^d$, $\mathrm{Proj}_A(p)$ denotes the projection of p onto A.

We introduce a few technical lemmas which are crucial for the proof of Proposition 3.

Lemma 8 (Projection interpretation of RSGF): *For $v \in \mathbb{R}$, $p, q \in \mathbb{R}^d$, the set $\{\xi \in \mathbb{R}^d : \alpha v + q^\top \xi + \frac{\beta}{2}\|\xi\|^2 \leq 0\}$ is nonempty iff $\|q\|^2/\beta^2 - 2\alpha v/\beta \geq 0$, in which case it equals the closed ball $\mathbb{B}(-q/\beta, r)$ with $r = \sqrt{\|q\|^2/\beta^2 - 2\alpha v/\beta}$. Hence, whenever (4) is feasible we have*

$$\mathcal{R}(\theta) = P_{\mathbb{B}(-\nabla V_1(\theta)/\beta,\, r(\theta))}(-\nabla V_0(\theta)).$$

Proof Completing the square, $\alpha v + q^\top \xi + \frac{\beta}{2}\|\xi\|^2 = \frac{\beta}{2}\big(\|\xi + q/\beta\|^2 - r^2\big)$, and the objective function $\frac{1}{2}\|\xi + \nabla V_0(\theta)\|^2$ is minimized over the closed convex set $\mathbb{B}(-q/\beta, r)$ by the projection of $-\nabla V_0(\theta)$ onto this set $\mathbb{B}(-q/\beta, r)$. ■

Lemma 9 (Multiplier bound): *Suppose that $\Theta \subset \mathbb{R}^d$ is compact and that assumption (iii) of Proposition 3 holds, and let $\bar{u}$ be as defined in* (17)*. Let $\theta \in \Theta$ and let $u(\theta)$ be the (unique) multiplier of* (4)*. Then $u(\theta) \le \bar{u}$.*

Proof The center $\bar{\xi} = -\nabla V_1/\beta$ satisfies

$$\alpha V_1(\theta) + \nabla V_1(\theta)^\top \bar{\xi} + \frac{\beta}{2}\left\|\bar{\xi}\right\|^2 = -\tfrac{\beta}{2} D(\theta) \le -\tfrac{\beta}{2}\rho^2 < 0.$$

Hence, Slater's condition holds and $\mathcal{R}(\theta)$ admits a KKT multiplier $u \ge 0$. Since $u \ge 0$, the Lagrangian

$$L(\xi, u) = \tfrac{1}{2}\|\xi + \nabla V_0\|^2 + u\Big(\alpha V_1 + \nabla V_1^\top \xi + \frac{\beta}{2}\|\xi\|^2\Big)$$

is convex in ξ, and stationarity gives $\nabla_\xi L(\mathcal{R}, u) = 0$, so $\mathcal{R}$ minimizes $L(\cdot, u)$. Together with complementary slackness of (4) this yields $\frac{1}{2}\|\mathcal{R} + \nabla V_0\|^2 = L(\mathcal{R}, u) \le L(\bar{\xi}, u) \le \frac{1}{2}\|\bar{\xi} + \nabla V_0\|^2 - u\frac{\beta}{2}\rho^2$, and therefore $u \le \|\nabla V_0 - \nabla V_1/\beta\|^2/(\beta\rho^2) \le \bar{u}$.

■

Lemma 10 (RSGF perturbation bound): *Suppose that $\Theta \subset \mathbb{R}^d$ is compact and that assumption (iii) of Proposition 3 holds, and let c_1, c_V be as defined in* (17) *and ε_N as defined in* (18)*. Let $\theta \in \Theta$, $e_V := \widehat{V}_1(\theta) - V_1(\theta)$, $e_j := \widehat{\nabla V}_j(\theta) - \nabla V_j(\theta)$. Then, almost surely,*

$$\|\hat{\mathcal{R}}(\theta) - \mathcal{R}(\theta)\| \le \|e_0\| + c_1\|e_1\| + c_V|e_V|. \tag{35}$$

If, in addition, $\theta_t \in \Theta$ almost surely for some $t \in \mathbb{N}$, then $\mathbb{E}_t[E_t^2] \le \varepsilon_N^2$, $\mathbb{E}_t[E_t] \le \varepsilon_N$ and $\|\mathcal{R}(\theta_t) - \mathbb{E}_t[\hat{\mathcal{R}}(\theta_t)]\| \le \varepsilon_N$.

Proof Throughout, we let $P(y; c, s) := \mathrm{Proj}_{\mathbb{B}(c,s)}(y)$. We also write $\hat{s}(\theta) := \sqrt{[\hat{D}(\theta)]_+}$. By Lemma 8 and since $\hat{\mathcal{R}}(\theta) = -\widehat{\nabla V}_1(\theta)/\beta$ by convention whenever (11) is infeasible, we have

$$\mathcal{R}(\theta) = P\big(-\nabla V_0(\theta);\, -\nabla V_1(\theta)/\beta,\, \sqrt{D(\theta)}\big), \qquad \hat{\mathcal{R}}(\theta) = P\big(-\widehat{\nabla V}_0(\theta);\, -\widehat{\nabla V}_1(\theta)/\beta,\, \hat{s}(\theta)\big), \tag{36}$$

where $\mathcal{R}(\theta)$ is well defined since $r(\theta) \ge \rho > 0$ for all $\theta \in \Theta$ by assumption. We split the proof in three steps.

Step 1 (Lipschitz dependence of ball projections). Let us show that for all $y, y', c, c' \in \mathbb{R}^d$ and $s, s' \ge 0$,

$$\|P(y; c, s) - P(y'; c', s')\| \le \|y - y'\| + \|c - c'\| + |s - s'|. \tag{37}$$

To prove it, we use three standard facts. First, for any nonempty, closed, convex $C \subseteq \mathbb{R}^d$, both Proj_C and $I - \mathrm{Proj}_C$ are (firmly) nonexpansive (Bauschke and Combettes, 1991,

Prop. 4.16 and 4.4). Second, we have $P(y;c,s) = c + P(y-c;0,s)$ (Bauschke and Combettes, 1991, Prop. 29.1). Third, $P(0;0,s) = 0$ and $P(z;0,s) = \min\{s, \|z\|\}\, z/\|z\|$ for $z \neq 0$. By the triangle inequality,

$$\begin{aligned}\left\|P(y;c,s) - P(y';c',s')\right\| \leq \|P(y;c,s) - P(y';c,s)\| + \|P(y';c,s) - P(y';c',s)\| \\ + \|P(y';c',s) - P(y';c',s')\|\end{aligned}$$

Now, by the non-expansiveness, we have

$$\left\|P(y;c,s) - P(y';c,s)\right\| \leq \left\|y - y'\right\|.$$

With $Q := I - \mathrm{Proj}_{\mathbb{B}(0,s)}$, the translation property gives $P(y';c,s) = y' - Q(y'-c)$, so the (firm) nonexpansiveness of Q yields

$$\|P(y';c,s) - P(y';c',s)\| = \|Q(y'-c') - Q(y'-c)\| \leq \left\|c - c'\right\|.$$

Finally,

$$\begin{aligned}\|P(y';c',s) - P(y';c',s')\| &= \|P(y'-c';0,s) - P(y'-c';0,s')\| \\ &= |\min\{s, \left\|y'-c'\right\|\}\frac{y'-c'}{\|y'-c'\|} - \min\{s', \left\|y'-c'\right\|\}\frac{y'-c'}{\|y'-c'\|}| \\ &= |\min\{s, \left\|y'-c'\right\|\} - \min\{s', \left\|y'-c'\right\|\}| \leq |s-s'|,\end{aligned}$$

from where (37) follows. Now, applying (37) to (36) with $y = -\widehat{\nabla V}_0$, $c = -\widehat{\nabla V}_1/\beta$, $s = \hat{s}$ and $y' = -\nabla V_0$, $c' = -\nabla V_1/\beta$, $s' = \sqrt{D}$ (all evaluated at θ),

$$\|\hat{\mathcal{R}} - \mathcal{R}\| \leq \|e_0\| + \frac{\|e_1\|}{\beta} + |\hat{s} - \sqrt{D}|. \tag{38}$$

Step 2 (bounding the radius error). We first note that for $a \in \mathbb{R}$ and $b > 0$,

$$\left|\sqrt{[a]_+} - \sqrt{b}\right| \leq \frac{|a-b|}{\sqrt{b}}. \tag{39}$$

If $a \geq 0$, this follows from $|\sqrt{a} - \sqrt{b}| = |a-b|/(\sqrt{a}+\sqrt{b})$ and $\sqrt{a}+\sqrt{b} \geq \sqrt{b}$. If $a < 0$, then $|\sqrt{[a]_+} - \sqrt{b}| = \sqrt{b} = b/\sqrt{b} \leq (b-a)/\sqrt{b}$. Applying (39) with $a = \hat{D}(\theta)$, $b = D(\theta) \geq \rho^2 > 0$ gives $|\hat{s} - \sqrt{D}| \leq |\hat{D} - D|/\rho$. By (16),

$$|\hat{D} - D| \leq \frac{\left|\|\widehat{\nabla V}_1\|^2 - \|\nabla V_1\|^2\right|}{\beta^2} + \frac{2\alpha|e_V|}{\beta}.$$

Moreover, $\left|\|\widehat{\nabla V}_1\|^2 - \|\nabla V_1\|^2\right| = |(\widehat{\nabla V}_1 - \nabla V_1)^\top(\widehat{\nabla V}_1 + \nabla V_1)| \leq \|e_1\|\,(\|\widehat{\nabla V}_1\| + \|\nabla V_1\|) \leq \|e_1\|(\sqrt{d}\,\tilde{B}_1 + G_1)$, where we used the Cauchy–Schwarz inequality, $\|\nabla V_1\| \leq G_1$, and $\|\widehat{\nabla V}_1\| \leq \sqrt{d}\,\tilde{B}_1$ a.s. Therefore

$$|\hat{s} - \sqrt{D}| \leq \frac{\sqrt{d}\,\tilde{B}_1 + G_1}{\beta^2\rho}\|e_1\| + \frac{2\alpha}{\beta\rho}|e_V|,$$

and substituting into (38) yields (35) with c_1, c_V as in (17).

Step 3 (moment bounds). Let $\theta = \theta_t$, which is $\mathcal{F}_t$-measurable. Squaring (35) and using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ for all $a, b, c > 0$,

$$\mathbb{E}_t[E_t^2] \le 3\big(\mathbb{E}_t\|e_0\|^2 + c_1^2\,\mathbb{E}_t\|e_1\|^2 + c_V^2\,\mathbb{E}_t[e_V^2]\big). \tag{40}$$

Fix $j \in \{0,1\}$ and write $\widehat{\nabla V}_j = \frac{1}{N}\sum_{n=1}^N \tilde{G}_{j,n}^{\theta_t}$. Now, since $\mathbb{E}_t[\widehat{\nabla V}_j - \mathbb{E}_t[\widehat{\nabla V}_j]] = 0$, we have

$$\mathbb{E}_t\|e_j\|^2 = \|\mathbb{E}_t[\widehat{\nabla V}_j] - \nabla V_j(\theta_t)\|^2 + \mathbb{E}_t\|\widehat{\nabla V}_j - \mathbb{E}_t[\widehat{\nabla V}_j]\|^2.$$

For the first term, by (9) each coordinate of $\mathbb{E}_t[\widehat{\nabla V}_j] - \nabla V_j(\theta_t)$ is at most $\tilde{\zeta}_j$ in absolute value, so the term is at most $d\tilde{\zeta}_j^2$. For the second term, $\widehat{\nabla V}_j - \mathbb{E}_t[\widehat{\nabla V}_j] = \frac{1}{N}\sum_n(\tilde{G}_{j,n}^{\theta_t} - \mathbb{E}_t\tilde{G}_{j,n}^{\theta_t})$. By assumption, the $\tilde{G}_{j,n}^{\theta_t}$ are conditionally independent given $\mathcal{F}_t$, so $\mathbb{E}_t\langle \tilde{G}_{j,n}^{\theta_t} - \mathbb{E}_t\tilde{G}_{j,n}^{\theta_t}, \tilde{G}_{j,m}^{\theta_t} - \mathbb{E}_t\tilde{G}_{j,m}^{\theta_t}\rangle = 0$ for $n \ne m$, and

$$\mathbb{E}_t\|\widehat{\nabla V}_j - \mathbb{E}_t[\widehat{\nabla V}_j]\|^2 = \frac{1}{N^2}\sum_{n=1}^N \mathbb{E}_t\|\tilde{G}_{j,n}^{\theta_t} - \mathbb{E}_t\tilde{G}_{j,n}^{\theta_t}\|^2 = \frac{1}{N^2}\sum_{n=1}^N \left(\mathbb{E}_t\left\|\tilde{G}_{j,n}^{\theta_t}\right\|^2 - \left\|\mathbb{E}_t\tilde{G}_{j,n}^{\theta_t}\right\|^2\right)$$
$$\le \frac{1}{N^2}\sum_{n=1}^N \mathbb{E}_t\|\tilde{G}_{j,n}^{\theta_t}\|^2 \le \frac{d\tilde{B}_j^2}{N},$$

where we used $\|\tilde{G}_{j,n}^{\theta_t}\|^2 \le d\tilde{B}_j^2$ a.s. by (10). Hence $\mathbb{E}_t\|e_j\|^2 \le d(\tilde{\zeta}_j^2 + \tilde{B}_j^2/N)$. The same argument applied to the scalar samples $G_{1,n}^{\theta_t}$, using $|\mathbb{E}_t[\widehat{V}_1] - V_1(\theta_t)| \le \zeta_1$ and $|G_{1,n}^{\theta_t}| \le B_1$, gives $\mathbb{E}_t[e_V^2] \le \zeta_1^2 + B_1^2/N$. Substituting into (40) gives $\mathbb{E}_t[E_t^2] \le \varepsilon_N^2$. By Jensen's inequality, $\mathbb{E}_t[E_t] \le (\mathbb{E}_t[E_t^2])^{1/2} \le \varepsilon_N$. Finally, by Jensen's inequality we have $\|\mathcal{R}(\theta_t) - \mathbb{E}_t[\hat{\mathcal{R}}(\theta_t)]\| = \|\mathbb{E}_t[\hat{\mathcal{R}}(\theta_t) - \mathcal{R}(\theta_t)]\| \le \mathbb{E}_t[E_t] \le \varepsilon_N$.

■

Lemma 11 (Φ_κ function descent): *Suppose that assumptions (i)–(iii) of Proposition 3 hold, fix $\kappa \ge \bar{u}$, and let Φ_κ, G_Φ and L_Φ be as defined in Proposition 3. If $h_t \le 1/\alpha$ then,*

$$\Phi_\kappa(\theta_{t+1}) \le \Phi_\kappa(\theta_t) - h_t\|\mathcal{R}_t\|^2 + h_t G_\Phi E_t + \frac{L_\Phi h_t^2}{2}\|\hat{\mathcal{R}}_t\|^2 \quad a.s. \tag{41}$$

If moreover $h_t \le 1/(2L_\Phi)$, then $\Phi_\kappa(\theta_{t+1}) \le \Phi_\kappa(\theta_t) - \frac{h_t}{2}\|\mathcal{R}_t\|^2 + h_t G_\Phi E_t + L_\Phi h_t^2 E_t^2$.

Proof Throughout the proof, all gradients are evaluated at θ_t unless stated otherwise. We let $\Delta_t := \theta_{t+1} - \theta_t = h_t\hat{\mathcal{R}}_t$, and $u_t := u(\theta_t)$ be the multiplier of (4) at θ_t, which by Lemma 9 satisfies $0 \le u_t \le \bar{u}$. No expectation is taken in this proof: all inequalities hold for every realization with $\theta_t, \theta_{t+1} \in \Theta$ (by assumption, this means that all inequalities hold almost surely).

Since $\theta_t, \theta_{t+1} \in \Theta$, Θ is convex and ∇V_j is L_j-Lipschitz on Θ, (Nesterov, 2018, Lemma 1.2.3) gives, for $j \in \{0,1\}$,

$$V_j(\theta_{t+1}) \le V_j(\theta_t) + \nabla V_j(\theta_t)^\top\Delta_t + \tfrac{L_j}{2}\|\Delta_t\|^2$$

$$= V_j(\theta_t) + h_t \nabla V_j(\theta_t)^\top \mathcal{R}_t + h_t \nabla V_j(\theta_t)^\top (\hat{\mathcal{R}}_t - \mathcal{R}_t) + \tfrac{L_j}{2}\|\Delta_t\|^2.$$

By the Cauchy–Schwarz inequality and $\|\nabla V_j(\theta_t)\| \le G_j$, we have $h_t \nabla V_j(\theta_t)^\top (\hat{\mathcal{R}}_t - \mathcal{R}_t) \le h_t G_j E_t$, so

$$V_j(\theta_{t+1}) \le V_j(\theta_t) + h_t \nabla V_j(\theta_t)^\top \mathcal{R}_t + h_t G_j E_t + \tfrac{L_j}{2}\|\Delta_t\|^2, \qquad j \in \{0,1\}. \tag{42}$$

The map $[\cdot]_+$ is nondecreasing, subadditive ($[a+b]_+ \le [a]_+ + [b]_+$), and satisfies $[c]_+ = c$ for $c \ge 0$. Applying these properties to (42) with $j = 1$ (the last two terms there are nonnegative) gives

$$[V_1(\theta_{t+1})]_+ \le [V_1(\theta_t) + h_t \nabla V_1(\theta_t)^\top \mathcal{R}_t]_+ + h_t G_1 E_t + \tfrac{L_1}{2}\|\Delta_t\|^2.$$

Multiplying this inequality by $\kappa \ge 0$ and adding (42) with $j = 0$ yields

$$\Phi_\kappa(\theta_{t+1}) \le \Psi_t + h_t G_\Phi E_t + \tfrac{L_\Phi h_t^2}{2}\|\hat{\mathcal{R}}_t\|^2, \tag{43}$$

where

$$\Psi_t := V_0(\theta_t) + h_t \nabla V_0^\top \mathcal{R}_t + \kappa [V_1(\theta_t) + h_t \nabla V_1^\top \mathcal{R}_t]_+,$$

and we used $G_\Phi = G_0 + \kappa G_1$, $L_\Phi = L_0 + \kappa L_1$ and $\|\Delta_t\|^2 = h_t^2 \|\hat{\mathcal{R}}_t\|^2$. The result follows if we show that

$$\Psi_t \le \Phi_\kappa(\theta_t) - h_t \|\mathcal{R}_t\|^2. \tag{44}$$

To see this, note that by (4):

$$V_1(\theta_t) + h_t \nabla V_1^\top \mathcal{R}_t \le (1 - \alpha h_t) V_1(\theta_t) - \tfrac{\beta h_t}{2}\|\mathcal{R}_t\|^2 \le (1 - \alpha h_t) V_1(\theta_t). \tag{45}$$

On the other hand, as shown in (Mestres et al., 2025a, Proposition IV.5 (i)) we have

$$h_t \nabla V_0^\top \mathcal{R}_t \le -h_t \|\mathcal{R}_t\|^2 + u_t \, \alpha h_t \, V_1(\theta_t). \tag{46}$$

Now we distinguish two cases:

Case $V_1(\theta_t) \le 0$. Combining (45), (46), $V_1(\theta_t) \le 0$ and the definition of $[\cdot]$ we get

$$\Psi_t \le V_0(\theta_t) - h_t \left\|\mathcal{R}_t\right\|^2 \le \Phi_\kappa(\theta_t) - h_t \left\|\mathcal{R}_t\right\|^2,$$

where in the last inequality we used that $[V_1(\theta_t)]_+ \ge 0$. Hence, (44) holds.

Case $V_1(\theta_t) > 0$. Combining (45) and (46) we get

$$\begin{aligned}\Psi_t &\le V_0(\theta_t) - h_t\|\mathcal{R}_t\|^2 + \bar{u}\alpha h_t V_1(\theta_t) + \kappa V_1(\theta_t) - \kappa\alpha h_t V_1(\theta_t) \\ &= \Phi_\kappa(\theta_t) - h_t\|\mathcal{R}_t\|^2 - (\kappa - \bar{u})\,\alpha h_t\, V_1(\theta_t) \le \Phi_\kappa(\theta_t) - h_t\|\mathcal{R}_t\|^2,\end{aligned}$$

where the last inequality uses $\kappa \ge \bar{u}$. ■

Proof [Proof of Proposition 3] (a) Summing the second inequality of Lemma 11 over $t = 0, \dots, K-1$ and using $\Phi_\kappa(\theta_K) \ge \min_\Theta V_0$,

$$\frac{1}{2}\sum_{t=0}^{K-1} h_t \|\mathcal{R}_t\|^2 \le \Delta_0 + G_\Phi \sum_{t=0}^{K-1} h_t E_t + L_\Phi \sum_{t=0}^{K-1} h_t^2 E_t^2, \tag{47}$$

for every realization. Taking expectations and using $\mathbb{E}[E_t] \le \varepsilon_N$, $\mathbb{E}[E_t^2] \le \varepsilon_N^2$ (Lemma 10 and the tower property) and dividing by $H_K/2$ gives (19). For constant h, $2L_\Phi h \le 1$ gives (20).

(b) By Lemma 8 (and the convention for infeasible (11)), $\hat{\mathcal{R}}_t$ lies in the closed ball with center $-\widehat{\nabla V}_1(\theta_t)/\beta$ and radius $\sqrt{[\hat{D}(\theta_t)]_+}$. Since $\|\widehat{\nabla V}_1(\theta_t)\| \le \sqrt{d}\,\tilde{B}_1$ and $|\hat{V}_1(\theta_t)| \le B_1$ a.s. by (10), so $\sqrt{[\hat{D}(\theta_t)]_+} \le \sqrt{d\tilde{B}_1^2/\beta^2 + 2\alpha B_1/\beta} \le \sqrt{d}\tilde{B}_1/\beta + \sqrt{2\alpha B_1/\beta}$, which means that $\|\hat{\mathcal{R}}_t\| \le 2\sqrt{d}\,\tilde{B}_1/\beta + \sqrt{2\alpha B_1/\beta}$ a.s. Moreover, $\mathcal{R}$ is continuous on the compact set Θ (by Lemma 10) and $\theta_t \in \Theta$ a.s., so $\|\mathcal{R}_t\| \le \max_\Theta \|\mathcal{R}\| < \infty$. Hence $E_t \le \|\hat{\mathcal{R}}_t\| + \|\mathcal{R}_t\| \le \bar{E}$ a.s. for a deterministic constant $\bar{E}$. Hence, $D_t := E_t - \mathbb{E}_t[E_t]$ and $D_t' := E_t^2 - \mathbb{E}_t[E_t^2]$ are bounded martingale differences adapted to $\mathcal{F}_{t+1}$. With $H_{t+1} = \sum_{k\le t} h_k$, we have $(h_t/H_{t+1})^2 \le \bar{h}\,(1/H_t - 1/H_{t+1})$ for $t \ge 1$, so $\sum_t (h_t/H_{t+1})^2 < \infty$ and therefore the martingales $\sum_{k\le t}(h_k/H_{k+1})D_k$, $\sum_{k\le t}(h_k/H_{k+1})D_k'$ are bounded in L^2, hence converge a.s by (Durrett, 2010, Theorem 4.4.6). Since $H_K \to \infty$, Kronecker's lemma (Durrett, 2010, Theorem 2.5.9) yields

$$\frac{1}{H_K}\sum_{t<K} h_t D_t \to 0, \quad \frac{1}{H_K}\sum_{t<K} h_t D_t' \to 0, \quad a.s.,$$

so $\limsup_K \frac{1}{H_K}\sum_{t<K} h_t E_t \le \varepsilon_N$ and $\limsup_K \frac{1}{H_K}\sum_{t<K} h_t E_t^2 \le \varepsilon_N^2$ a.s. Dividing (47) by $H_K/2$, using $L_\Phi h_t^2 E_t^2 \le \frac{1}{2} h_t E_t^2$ and letting $K \to \infty$ gives (21).

For the lim inf statement, fix a realization for which (21) holds, and let $a_t := \|\mathcal{R}(\theta_t)\|^2 \ge 0$ and $c := 2G_\Phi\varepsilon_N + \varepsilon_N^2$. Suppose, by contradiction, that $\liminf_{t\to\infty} a_t > c$. Then there exist $m > c$ and $T \in \mathbb{N}$ such that $a_t \ge m$ for all $t \ge T$. Since $a_t \ge 0$ and $h_t > 0$, for every $K > T$,

$$\frac{1}{H_K}\sum_{t=0}^{K-1} h_t a_t \;\ge\; \frac{1}{H_K}\sum_{t=T}^{K-1} h_t a_t \;\ge\; m\,\frac{H_K - H_T}{H_K}.$$

Since H_T is fixed and $H_K \to \infty$, the right-hand side converges to m as $K \to \infty$, so $\limsup_{K\to\infty} \frac{1}{H_K}\sum_{t=0}^{K-1} h_t a_t \ge m > c$, which contradicts (21). Hence $\liminf_{t\to\infty} \|\mathcal{R}(\theta_t)\|^2 \le c$ for every realization for which (21) holds, i.e., almost surely. ■

Appendix G. Experiments reward definitions

Table 1 lists the reward terms for all experiments. Each term contributes $w \cdot r$ to the total reward, where r is the expression shown and w its weight. All expressions are non-negative; penalties carry negative weights.

For the humanoid squatting and crawling, terms follow BeyondMimic (Liao et al., 2025) as implemented in mjlab (Zakka et al., 2026): p_r, q_r root position/orientation; p_i, q_i, v_i, ω_i position, orientation, linear and angular velocity of tracked body i; d_q geodesic quaternion distance; c_{joint} soft joint-limit violation; $F_{i,b}$ contact force on body h. $\Delta x = x^{\text{ref}} - x$; $\overline{(\cdot)}$ is the mean over the N_b tracked bodies.

For the humanoid walking, generic terms follow (Rudin et al., 2022); the gait term follows (Siekmann et al., 2021). $v_x^\star$ commanded forward speed; e_y, e_ψ lateral and heading drift;

$s_f, s_f^\star$ stride length and target; s_ℓ lateral stride offset; d_L, d_R left/right leg deviation; $h_i, h^\star$ swing-foot height and target; c_i contact of foot i; td, i touchdown of foot i; $g_{b,xy}$ projected gravity in base frame; z_b base height; $\mathcal{G}_\sigma(e) = e^{-e^2/(2\sigma^2)}$; $a \to b$ denotes a curriculum.

For the Crazyflie experiment, p, v denote drone position and velocity, p^{ref} the hover target, $d = \|p - p^{\mathrm{ref}}\|$, v_b, ω_b body-frame linear and angular velocities, and $a_t, \ddot{p}_t$ action and acceleration at step t. V denotes a quadratic CLF on position and attitude errors and their rates (Li et al., 2026); $\bar{\lambda} = \lambda_{\max}(P)$, and $\Delta_{\max}$ clips the decrease condition $\dot{V} \le -0.02\, V$. CAPS (Mysore et al., 2021) uses $\mu_t = \mu(s_t)$ policy mean action, ϵ small random state perturbation, $\mathrm{sg}[\cdot]$ stop-gradient. We also denote by p_i, v_i, r_i the center, velocity, and radius of obstacle i; drone radius $r = 0.05\,\mathrm{m}$ safety margin $m = 0.1\,\mathrm{m}$, $c_{\mathrm{obs}} = \min_i \|p_i - p\| - (\rho_i + r + m)$, and c_{arena} is the clearance to the arena boundary minus m. The safety reward is $R_1 = 1 - \frac{\delta}{T+1}$ outside the safe set $\bar{\mathcal{C}} = \{\min(c_{\mathrm{obs}}, c_{\mathrm{arena}}) \ge 0\}$ (Table 2); episodes terminate only on contact. $\hat{c}_{\mathrm{obs}} = \max\{0, \min_i \min_{t\in[0,\tau]} \|(p_i - p) + t(v_i - v)\| - (\rho_i + r + m)\}$, with $\tau = 0.5\,\mathrm{s}$, is the clearance predicted at constant relative velocity, so $\tilde{h} \ge 0$ on $\bar{\mathcal{C}}$.

Appendix H. Training details

Algorithm 1 proposes a practical implementation of RSGF-RL as used in Section 4. It summarizes how the estimates $\widehat{V}_1$, $\widehat{\nabla V}_1$ and $\widehat{\nabla V}_0$ entering (11) are computed at each iteration. At iteration t, the policy π_{θ_t} is rolled out in N parallel environments for T_{r} steps each (cf. Table 2); environments that terminate are reset and keep running, so each environment produces a segment containing one or more, possibly partial, episodes. The constraint estimates are built from the discounted cost-to-go of R_1 along these segments, using a learned cost critic W_ψ as the baseline $b_{1,t}$ in (7). For the objective, instead of forming a gradient explicitly, we run a standard PPO-clip update starting from θ_t and define $\widehat{\nabla V}_0(\theta_t)$ as the resulting parameter displacement divided by h_t, so that $-h_t \widehat{\nabla V}_0(\theta_t)$ is exactly the PPO step. Although this quantity is not an unbiased estimate of $\nabla V_0(\theta_t)$ (as it accumulates J Adam-preconditioned minibatch steps evaluated along the PPO iterates), it is uniformly bounded, and hence satisfies (9)-(10) for some $\tilde{\zeta}_0$, and we have found works well in practice. In this way, problem (11) acts as a filter on the PPO step: if $\xi = -\widehat{\nabla V}_0(\theta_t)$ is feasible for (11), then $\hat{\mathcal{R}}(\theta_t) = -\widehat{\nabla V}_0(\theta_t)$ and the update coincides with the PPO iterate; otherwise, the PPO step is replaced by $h_t \hat{\mathcal{R}}(\theta_t)$.

In Algorithm 1, L^{clip} is as defined in (Schulman et al., 2017, Equation 9) and $\overline{\mathrm{KL}}$ is the KL divergence averaged over the rollout states. Adam uses gradient-norm clipping, and its moment estimates are carried over across iterations. The cost critic W_ψ has the same architecture as the value critic and receives the same observations.

Regarding the baseline comparison in Figure 1(a), we point out that PPO-Lagrangian follows the implementation in OpenAI (2019) (Lagrange multiplier $\lambda = \mathrm{softplus}(\vartheta)$ initialized to 1 and updated with Adam, learning rate 0.05), with all PPO settings matching the PPO baselines in Table 2 and the cost limit applied to the mean per-step cost. On the other hand, CPO (Achiam et al., 2017) uses the parameters in Table 3.

(a) Humanoid squatting and crawling (motion tracking)

Term	Expression	Weight
Task reward R_0		
Root position	$e^{-\|\Delta p_r\|^2/0.3^2}$	0.5
Root orientation	$e^{-d_q(q_r^{\mathrm{ref}},q_r)^2/0.4^2}$	0.5
Body position	$e^{-\overline{\|\Delta p_i\|^2}/0.3^2}$	1.0
Body orientation	$e^{-\overline{d_q(q_i^{\mathrm{ref}},q_i)^2}/0.4^2}$	1.0
Body lin. velocity	$e^{-\overline{\|\Delta v_i\|^2}/1.0^2}$	1.0
Body ang. velocity	$e^{-\overline{\|\Delta \omega_i\|^2}/3.14^2}$	1.0
Action rate	$\|a_t - a_{t-1}\|^2$	-0.2
Joint limits	c_{joint}	-10
Self-collision	$\sum_{b=1}^{B} \mathbb{I}\big[\max_i \|F_{i,b}\| > 10\big]$	-10
Safety reward R_1		
Base / head height	R_1 in (23), $\delta = 0.1$	–

(b) Humanoid walking

Term	Expression	Weight	Term	Expression	Weight
Task			*Posture / stability*		
Forward velocity	$\mathcal{G}_{\sigma_v}(v_x - v_x^\star)$	$1 \to 3$	Arm posture	$\mathrm{mean}\left((q-q^\star)^2/\sigma^2\right)$	-1
Straight line	$\mathcal{G}_{\sigma_y}(e_y)\,\mathcal{G}_{\sigma_\psi}(e_\psi)$	$1.5 \to 3$	Leg posture	$\mathrm{mean}\left((q-q^\star)^2/\sigma^2\right)$	-0.7
Gait			Lateral motion	$v_{b,y}^2 + v_{b,z}^2$	-0.5
Foot gait	phase-dependent contacts	1	Orientation rate	$\omega_x^2 + \omega_y^2$	-1
Stride length	$\mathcal{G}_{\sigma_s}(s_f - s_f^\star)\,\mathcal{G}_{\sigma_\ell}(s_\ell)$	1	Upright	$\mathcal{G}_{0.15}(\|g_{b,xy}\|)$	1.8
Leg symmetry	$e^{-\mathrm{mean}(d_L+d_R)^2/\sigma_{\mathrm{sym}}^2}$	3	Base height	$\mathcal{G}_{0.05}(z_b - 0.793)$	1
Foot clearance	$\sum_i \lvert h_i - h^\star\rvert \, \|\dot{p}_{i,xy}\|$	-2	*Regularization*		
Foot slip	$\sum_i (\|\dot{p}_{i,xy}\|^2 + 0.25\,\omega_{i,z}^2)\,1_{c_i}$	-1	Alive	1	0.05
Soft landing	$\sum_i \|F_i\|\,1_{\mathrm{td},i}$	-10^{-3}	Action	$\|a_t\|^2$	-0.1
			Action rate	$\|a_t - a_{t-1}\|^2$	-0.15
			Joint acceleration	$\|\ddot{q}\|^2$	-10^{-6}
			Joint limits	soft joint-limit violation	-0.5
Safety reward R_1					
Height / attitude / CoP	R_1 in (23), $\delta = 0.1$	–			

(c) Crazyflie obstacle avoidance

Term	Expression	Weight
Task reward R_0		
Settle	$\big(1 - \tanh\frac{\|p-p^{\mathrm{ref}}\|}{0.2}\big)\big(1 - \tanh\frac{\|v\|}{0.3}\big)$	20
Target distance	$10 - d$	0.5
CLF	$e^{-V/(0.7^2\,\bar{\lambda})}$	10
CLF decrease	$\mathrm{clip}\left([\dot{V} + 0.02\,V]_+/\Delta_{\max},\, 0,\, 1\right)$	-5
Lin. / ang. velocity	$\|v_b\|^2$ / $\|\omega_b\|^2$	-0.01 / -0.02
Action rate	$\|a_t - a_{t-1}\|^2$	-0.01
Jerk	$\|\ddot{p}_t - \ddot{p}_{t-1}\|^2$	-10^{-4}
Policy-loss regularization (CAPS)		
Temporal / spatial	$\|\mu_t - \mu_{t-1}\|^2$ / $\|\mu(s_t+\epsilon) - \mathrm{sg}[\mu(s_t)]\|^2$	1 / 1
Safety reward R_1		
Hazard proximity	R_1 in (23) with $\delta = 0.05$ and $\tilde{h} = \min\{6\,\hat{c}_{\mathrm{obs}},\, 10\,c_{\mathrm{arena}}\}$	–

Table 1: Reward terms for all experiments.

	Squat./crawl.	**Hum. walking**	**Drone flight**
Training parameters			
Actor/Critic MLP size	(512, 256, 128)	(256, 128, 64)	(256, 256, 256)
Activation function	ELU	ELU	ELU
Entropy coef.	0.005	0.003	0.001
Epochs / minibatches	5 / 4	5 / 4	5 / 4
J	20	20	20
E_c	4	4	4
$KL_{\max}$	0.08	0.08	0.08
PPO learning rate*	10^{-3}	3×10^{-4}	3×10^{-4}
PPO clip	0.2	0.2	0.2
Discount factor γ	0.99	0.99	0.99
GAE parameter λ	0.95	0.95	0.95
Rollout (steps/env)	24	32	24
Parallel envs	4096	2048	8192
Total environment steps	$\approx (5.9/7.9) \times 10^8$	$\approx 5.2 \times 10^8$	$\approx 7.9 \times 10^8$
Safety constraint	Base height ≥ 0.4 m / Base and head ≤ 0.2 m	Base height ≥ 0.67 m, $\lvert\text{roll}\rvert \leq 0.35$ rad, $\lvert\text{pitch}\rvert \leq 0.50$ rad, CoP in support polygon	Distance to obstacles and arena edge ≥ 0.1 m
RSGF-RL parameters			
$\{h_t\}$	constant, 0.5	constant, 0.5	constant, 0.15
α	2×10^{-5}	2×10^{-5}	0.1
β	1	1	1

Table 2: Training and RSGF-RL parameters for each experiment. All experiments use observation normalization. *Adaptive, except for drone flight. The reported values refer to the initial PPO learning rate.

Parameter	**Value**
Returns and advantages	
Reward discount γ	0.99
Cost discount γ_c	0.99
Reward / cost GAE parameter λ	0.95 / 0.95
Cost limit d	0
Critics	
Critic learning rate	10^{-3}
Critic epochs / minibatches	5 / 4

Parameter	**Value**
Trust-region policy step	
KL trust-region limit δ	0.01
Conjugate-gradient iterations	15
Conjugate-gradient damping	0.1
Fisher-vector-product subsampling	5
Line-search attempts	15
Line-search decay	0.8

Table 3: CPO parameters for the squatting comparison in Figure 1(a).

Algorithm 1 One iteration of RSGF-RL as implemented in Section 4

Require: Policy parameter θ_t; cost critic W_ψ; PPO value critic and Adam state; stepsize h_t; parameters α, β, γ

Rollout: for $n \in 1 : N$, $k \in 0 : T_{\mathrm{r}} - 1$, sample $a_k^n \sim \pi_{\theta_t}(\cdot \,|\, s_k^n)$ and record $(s_k^n, a_k^n, r_k^n, c_k^n, d_k^n)$, where $r_k^n = R_0(s_k^n, a_k^n, s_{k+1}^n)$, $c_k^n = R_1(s_k^n, a_k^n, s_{k+1}^n)$, and $d_k^n \in \{0, 1\}$ indicates termination of an episode at step k

Cost-to-go: set $G_{T_{\mathrm{r}}}^n = 0$ and $G_k^n = c_k^n + \gamma(1 - d_k^n)\, G_{k+1}^n$ for $k = T_{\mathrm{r}} - 1, \ldots, 0$

Constraint value: $\widehat{V}_1(\theta_t) = \frac{1}{NT_{\mathrm{r}}} \sum_{n=1}^{N} \sum_{k=0}^{T_{\mathrm{r}}-1} G_k^n$

Cost critic: fit W_ψ to the targets $\{G_k^n\}$ by minimizing the squared error with E_{c} epochs of minibatch Adam

Constraint gradient:
$\widehat{\nabla V}_1(\theta_t) = \frac{1}{NT_{\mathrm{r}}} \sum_{n,k} \big(G_k^n - W_\psi(s_k^n)\big)\, \nabla_\theta \log \pi_{\theta_t}(a_k^n \,|\, s_k^n)$

PPO advantages: compute GAE advantages $\hat{A}_k^n$ from $\{r_k^n\}$ and the value critic; set $\theta^{(0)} = \theta_t$

for $j = 0, \ldots, J - 1$ (J = epochs $\times$ minibatches) **do**

  Obtain $\theta^{(j+1)}$ through Adam update on $\theta^{(j)}$, using gradient of L^{clip} computed with minibatch $\mathcal{B}_j$.

end for

Objective gradient: $\widehat{\nabla V}_0(\theta_t) = -\big(\theta^{(J)} - \theta_t\big)/h_t$

RSGF direction: compute $\hat{\mathcal{R}}(\theta_t)$ from (11) in closed form (Mestres et al., 2025b, Lemma 3), or set $\hat{\mathcal{R}}(\theta_t) = -\widehat{\nabla V}_1(\theta_t)/\beta$ if (11) is infeasible

if $\xi = -\widehat{\nabla V}_0(\theta_t)$ is feasible for (11) **then**

  $\theta_{t+1} = \theta^{(J)}$ (constraint inactive: $\hat{\mathcal{R}}(\theta_t) = -\widehat{\nabla V}_0(\theta_t)$)

else

  $\theta_{t+1} = \theta_t + \sigma h_t \hat{\mathcal{R}}(\theta_t)$, where $\sigma = 2^{-m}$ and m is the smallest integer in $0 : 4$ such that $\overline{\mathrm{KL}}(\pi_{\theta_t} \,\|\, \pi_{\theta_{t+1}}) \le \mathrm{KL}_{\max}$ ($m = 4$ if none)

end if