A UNIFIED BELLMAN OPERATOR FOR SAFETY-CRITICAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning in safety-critical domains requires maximizing task performance while strictly adhering to safety constraints. Existing safe reinforcement learning paradigms *typically* force a trade-off: they either require *a priori* knowledge to provide strict safety guarantees (e.g., safety filters), or they enable joint learning but only satisfy safety constraints on average. In this work, we propose a novel Bellman operator that unifies performance and safety objectives into a *joint* value function. We show that temporal difference learning with the joint Bellman operator converges under a two-timescale stochastic approximation framework. On the fast timescale, the *safety value* of the learning joint policy is estimated, while the *joint value* is estimated on the slow timescale. Convergence is ensured by formulating the limiting dynamics as an *occupation-averaged* differential inclusion, and showing that it asymptotically converges to a set of limiting optimal safety-constrained task value functions. *Theoretically*, once converged, the resulting optimal policy maximizes task return while maintaining safety at all times. Empirical evaluations on continuous control tasks with *neural approximations* demonstrate stable convergence with near-zero safety violations at test time.

1 INTRODUCTION

Biological agents instinctively prioritize safety while mastering diverse tasks to ensure efficient, harm-free exploration. Safe Reinforcement Learning (Safe RL) captures this paradigm, aiming to learn near-optimal policies with strict safety assurances. Existing methods typically formalize this balance either by treating *safety as a soft constraint* during optimization or by theoretically enforcing *hard safety guarantees*.

Some of the early works formalized safe RL as a *Constrained Markov Decision Process* (CMDP), solving safety-constrained optimization problems using Lagrangian relaxation with dual multiplier methods (Altman, 1999; García & Fernández, 2015; Ray et al., 2019), with subsequent extensions via trust-region projections (Achiam et al., 2017; Yang et al., 2020; Xu et al., 2021; Milosevic et al., 2025), state-augmentation (Sootla et al., 2022), and offline distribution shaping (Yao et al., 2024). While these model-free techniques jointly optimize task performance and safety, they inherently bound *expected cumulative costs*, yielding only probabilistic safety guarantees. Conversely, model-based methods achieve strict safety via *predictive rollouts*, *epigraph formulation* of the value function (for deterministic control problems), or by strictly avoiding unsafe future sets (Huang et al., 2024; Jayant & Bhatnagar, 2022; Pfrommer et al., 2022; Altarovici et al., 2013). However, they suffer from compounding model biases and computational overhead to infer safe actions in real-time, particularly in environments with complex or unmodeled dynamics. Motivated by these tradeoffs, we seek a model-free approach that strictly adheres to safety specifications while optimizing task performance.

Some of the control-theoretic approaches that rigorously hard-guarantee safety are based on *Hamilton-Jacobi (HJ) reachability analysis*, casting safety and reach-avoid problems as *zero-sum differential games* (Isaacs, 1965; Mitchell et al., 2005). While deep RL techniques have enabled tractable neural solutions for the time-discounted HJI avoid/reach-avoid Bellman operators for safety analysis (Fisac et al., 2019; Hsu et al., 2021; Bajcsy et al., 2019), it is not immediately obvious how these techniques can be extended to generic tasks specified by dense, continuous reward functions. Consequently, a

*Equal contribution

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dominant paradigm employs a *decoupled approach*, where an *a priori learned* task-oblivious *safety filter* preemptively intercepts and projects nominal unsafe task actions onto a plausible safe set (Bui et al., 2021; Hsu et al., 2024; Nguyen et al., 2024; Cai et al., 2025; Banerjee et al., 2024). However, these switch-type interventions only occur near the boundary, causing erratic, jittery interventions when a safety-oblivious task policy persistently tries to exit the boundary. This is seldom desirable, especially in human-in-the-loop scenarios (Oh et al., 2025).

Recent offline *feasibility-guided models* instead utilize diffusion, flow-matching, or world-models to approximate the largest feasible region using HJ-reachability and steer task-driven actions to strictly lie within the safe envelope (Zheng et al., 2024; Tayal et al., 2026; Cao et al., 2025). However, they either introduce iterative online inference latency or rely on post-hoc distillation that weakens strict safety guarantees. Alternatively, using *precomputed* safety value functions as *Control Barrier Functions* (Q-CBF) (Hsu et al., 2024; Lu et al., 2025; Oh et al., 2025; 2026) avoids planning latency, but necessitates computing action projections online. Additionally, relying on external projections introduces *action aliasing*, which can be detrimental (Markgraf et al., 2026). If the safe action set is known a priori, neural-based methods such as (Suttle et al., 2024) can hard-guarantee safety; we instead assume *no access to an oracle safe action set* and learn from data. Thus, a growing subset of literature advocates for *unified architectures* that embed safety directly into the policy (Gu et al., 2024a;b; Cen et al., 2024). Yet, jointly optimizing rewards and costs frequently destabilizes learning due to gradient conflicts, necessitating algorithmic workarounds like gradient manipulation and explicit latent decoupling (Gu et al., 2024a;b; Kim et al., 2025; Cen et al., 2024; Yang et al., 2025).

In this paper, we combine *safety and task objectives directly at the value-function level*, rather than at the policy level, and jointly optimize *safety-constrained task performance* using a single *joint policy* via a novel, *joint Bellman operator* under adversarial disturbances. Our contributions are:

Unified Formulation: We propose a robust Bellman operator and derive the corresponding objective function from robust optimal stopping theory. We propose a two-timescale stochastic approximation framework that enables learning of its fixed point, the *joint value function*.

Theoretical Safety and Optimality Guarantees: We prove forward invariance of the safe set under the joint policy, and show convergence to the optimal robust task performance within the safe set.

Empirical Validation: We conduct experiments using neural approximations on a set of continuous control tasks from the `Gymnasium` and `SafetyGymnasium` suites under adversarial disturbances, demonstrating stable training convergence and near-zero safety violations during evaluation.

2 Preliminaries

This section introduces key definitions, followed by preliminaries on Constrained Markov Decision Processes and Hamilton-Jacobi safety analysis. Refer to Appendix A for more details.

2.1 Definitions and Notations

We consider a discounted MDP with finite state, action and disturbance spaces: $\mathcal{S} \subset \mathbb{R}^n$, $\mathcal{A} \subset \mathbb{R}^m$, and $\mathcal{D} \subset \mathbb{R}^l$ respectively, with the transition kernel $P(\cdot|s,a,d) \in \Delta(\mathcal{S})$. A deterministic joint policy $\pi = (\mu, \nu)$ pairs the mappings $\mu : \mathcal{S} \to \mathcal{A}$ and $\nu : \mathcal{S} \times \mathcal{A} \to \mathcal{D}$, yielding $\pi(s) = (\mu(s), \nu(s, \mu(s)))$. The reward and the safety margin are given by $r : \mathcal{S} \times \mathcal{A} \times \mathcal{D} \to [0, R_{\text{max}}], R_{\text{max}} > 0$ w.l.o.g, and $g : \mathcal{S} \to [-G_{\text{min}}, G_{\text{max}}]$, with $G_{\text{min}}, G_{\text{max}} > 0$ respectively. We subsume the extended class-$\mathcal{K}_e$ scaling function $\beta(\cdot)$ into $g(s)$ to align with task return scales (unless stated otherwise). The *constraint set* is $\mathcal{F}^c := \{s \in \mathcal{S} : g(s) \geq 0\}$, complementing the failure set $\mathcal{F}$. The *viability kernel* $\Omega \subseteq \mathcal{F}^c$ is the subset of states from which a control policy exists that can keep the agent away from the failure set $\mathcal{F}$ for all times, regardless of the adversarial disturbance. Finally, the value and action-value function spaces are defined as $\mathcal{V} := \mathbb{R}^{|\mathcal{S}|}$ and $Q := \mathbb{R}^{|\mathcal{S}||\mathcal{A}||\mathcal{D}|}$, equipped with the supremum norm $\|\cdot\|_\infty$.

2.2 Constrained Markov Decision Processes

A Constrained Markov Decision Process (CMDP) (Altman, 1999) extends an MDP by introducing auxiliary costs to encode safety constraints. It is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, c, \gamma, \mu_0)$, where r and c are the task reward and constraint cost functions, respectively, $\gamma \in [0, 1)$ is the discount factor, and $\mu_0 : \mathcal{S} \to [0, 1]$ denotes the initial state distribution. For a stochastic policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, let $J_r(\pi)$ and $J_c(\pi)$ denote the expected discounted cumulative return and cost.

The CMDP objective seeks an optimal policy that maximizes returns subject to a predefined safety threshold $d \in \mathbb{R}$, typically solved via Lagrangian relaxation, yielding an equivalent unconstrained min-max saddle-point problem: $\min_{\lambda \geq 0} \max_{\pi \in \Pi} (J_r(\pi) - \lambda(J_c(\pi) - d))$, where $\lambda \geq 0$ is the Lagrange multiplier and Π is the set of stationary stochastic policies. More details can be found in (Altman, 1999; Achiam et al., 2017; Tessler et al., 2019; Stooke et al., 2020).

2.3 Hamilton-Jacobi Safety Analysis

Hamilton-Jacobi (HJ) reachability analysis seeks to find the *safe set* $\Omega \subseteq \mathcal{F}^c$ (Mitchell et al., 2005). Formulated as a *zero-sum differential game*, the robust safety value function is $V^\phi(s) = \inf_{\mathcal{N}[\boldsymbol{d}]} \sup_{\boldsymbol{a}} \inf_{t \geq 0} g(\xi_{s_0}^{\boldsymbol{a}, \mathcal{N}[\boldsymbol{d}]}(t))$, where $\xi_{s_0}^{\boldsymbol{a}, \mathcal{N}[\boldsymbol{d}]}(\cdot)$ denotes the trajectory starting at s_0 and evolving according to the control and disturbance trajectories $(\boldsymbol{a}, \mathcal{N}[\boldsymbol{d}])$, with $\mathcal{N}[\boldsymbol{d}]$ the set of *non-anticipative disturbance strategies* (Evans & Souganidis, 1983; Fisac et al., 2018). By Dynamic Programming, $V(s)$ is the unique viscosity solution to the Hamilton-Jacobi-Isaacs (HJI) variational inequality. As the corresponding infinite-horizon Bellman operator does not induce a contraction mapping, (Fisac et al., 2019) introduces a discount factor $\gamma \in (0, 1)$ to enable TD-learning, resulting in the time-discounted robust Bellman equation under stochastic dynamics $V(s) = (1-\gamma)g(s) + \gamma \min \left\{ g(s), \max_{a \in \mathcal{A}} \min_{d \in \mathcal{D}} \mathbb{E}_{s' \sim p(\cdot|s,a,d)}[V(s')] \right\}$ More details can be found in (Mitchell et al., 2005; Fisac et al., 2019; 2018; Evans & Souganidis, 1983).

3 Learning Safety-Constrained Task Action-Value Function

This section presents our unified formulation to learn a joint safety-constrained task value function. We show that the corresponding *joint Bellman operator* induces a contraction mapping when the *auxiliary* safety value function is held fixed. We further see that the coupled system possesses a discontinuity due to changing joint greedy policies, and discuss its resolution.

3.1 The Joint Formulation

For clarity, we first introduce the joint Bellman equation to intuit its formulation. In the presence of adversarial disturbances, the *time-discounted task action-value function* Q^ψ purely corresponding to task return maximization can be extended to the following robust stochastic equation:

$$Q^{\text{task}}(s, a, d) = r(s, a, d) + \gamma \mathop{\mathbb{E}}_{s' \sim p(\cdot|s,a,d)} \left[\max_{a' \in \mathcal{A}} \min_{d' \in \mathcal{D}} Q^{\text{task}}(s', a', d') \right] \tag{1}$$

Next, consider the *time-discounted* safety value function, as introduced in (Fisac et al., 2019). We'd like to use this value function to instead assess the safety values of some given policy $\pi := (\mu, \nu)$, yielding the following *safe action-value function* evaluated π:

$$Q_\pi^\phi(s, a, d) = (1-\gamma)g(s) + \gamma \min \left\{ g(s), \mathop{\mathbb{E}}_{s' \sim p(\cdot|s,a,d)} \left[\max_{a' \in \mu(s')} \min_{d' \in \nu(s', a')} Q_\pi^\phi(s', a', d') \right] \right\} \tag{2}$$

We can use Q_π^ϕ to cap the further accrual of task rewards, i.e., *truncate task accrual* for the policy π when the corresponding task action-value goes above Q_π^ϕ, enabling the agent to instead optimize *task performance subject to safety*. Since γ serves as the *probability of episode continuation*, by the notion of *principled time-discounting* Fisac et al. (2019), we can combine Eq. 2 with Eq. 1, yielding the stochastic time-discounted *joint action-value function* as follows:

$$\begin{aligned} Q(s, a, d) =& (1-\gamma) \min\{r(s, a, d), g(s)\} \\ &+ \gamma \min \left\{ r(s, a, d) + \mathop{\mathbb{E}}_{s' \sim p(\cdot|s,a,d)} \left[\max_{a' \in \mathcal{A}} \min_{d' \in \mathcal{D}} Q(s', a', d') \right], Q_\pi^\phi(s, a, d) \right\} \end{aligned} \tag{3}$$

where, with probability $1-\gamma$, the agent receives the reward $\min\{r(s, a, d), g(s)\}$, or continues with probability γ. The corresponding *maximin joint policy* $\pi := (\mu, \nu)$ is defined through a *measurable selection* as:

$$\mu(s) := \arg\max_{a \in \mathcal{A}} \min_{d \in \mathcal{D}} Q(s, a, d), \qquad \nu(s, \mu(s)) := \arg\min_{d \in \mathcal{D}} Q(s, \mu(s), d) \tag{4}$$

Here, the disturbance gets the advantage of observing the controller's action, while the controller must *optimize* for the adversarial disturbance. Note that we do not assume Q_π^ϕ to be given; instead,

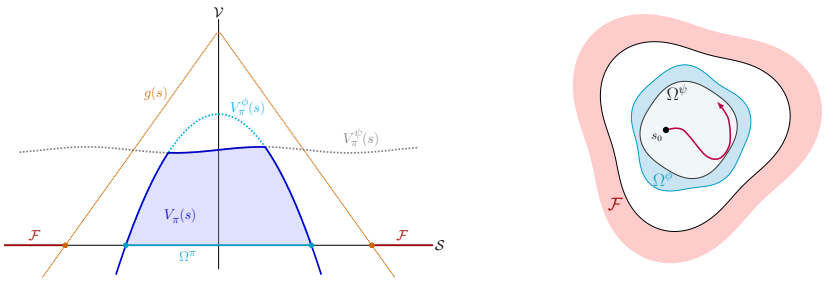

Figure 1: (Left) The Joint Value function is *conceptualized* as a 2D figure, with the state-space indicated on the x-axis, and joint values on the y-axis. (Right) The corresponding regions within the safe set Ω^π are depicted. The locus of points where $V_\pi^\phi = V_\pi^\psi$ indicates the boundary between Ω^ϕ and Ω^ψ as defined in Definition 1. Trajectories must stay within Ω^ψ to maximize episodic return.

Q_π^ϕ is learned alongside Q, estimating the safety values of the joint policy π. Note the *coupling*: the joint policy π is derived from the joint value Q, which is evaluated by Q_π^ϕ, whose values in turn update Q. Next, we show that the safety and joint Bellman operators are contractions given a fixed policy and fixed safety value, respectively. The joint value function is *conceptualized* in Fig. 1.

3.2 Safety Action-Value of the Joint Policy

Given a joint policy π, the safety Bellman operator $\mathcal{T}_\pi^\phi : \mathcal{Q} \to \mathcal{Q}$ corresponding to Eq. 2 is defined as:

$$\mathcal{T}_\pi^\phi \left[Q^\phi\right](s,a,d) = (1-\gamma)g(s) + \gamma \min\left\{g(s), \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mu(s')} \min_{d'\in\nu(s,a')} Q^\phi(s',a',d')\right]\right\} \tag{5}$$

Note that $\nu(\cdot, a')$ refers to the specific sample a' drawn from $\mu(\cdot)$ by the $\max_{a'\in\mu}$ operation.

Proposition 1 (Contraction Mapping of $\mathcal{T}_\pi^\phi$)**.** *For any arbitrary fixed joint policy π, the time-discounted Bellman operator defined in Eq. 5 induces a γ-contraction mapping under the supremum norm. That is, for a given discount factor $\gamma \in [0,1)$, and any $Q_\pi^\phi, \widetilde{Q}_\pi^\phi \in \mathcal{Q}$, we have $\left\|\mathcal{T}_\pi^\phi\left[Q_\pi^\phi\right] - \mathcal{T}_\pi^\phi\left[\widetilde{Q}_\pi^\phi\right]\right\|_\infty \leq \gamma \left\|Q_\pi^\phi - \widetilde{Q}_\pi^\phi\right\|_\infty$. (Proof in Appendix B.1)*

Hence, there exists a unique fixed point, Q_π^ϕ, which we call the safety action-value of the fixed joint policy π. The corresponding safety value function is obtained as $V_\pi^\phi(s) = \max_{a\in\mu}\min_{d\in\nu} Q_\pi^\phi(s,a,d)$. The set $\Omega^\pi = \{s \in \mathcal{S} : V_\pi^\phi(s) \geq 0\}$ is then the *safe set of policy* π (under $\gamma \to 1$ as discussed in Fisac et al. (2019) i.e., the set of all states from which the joint policy π keeps the agent safe. Note that, in general, $\Omega^\pi \subseteq \Omega$, where the maximal robust safe set, Ω, (viability kernel) is given by $\Omega = \{s \in \mathcal{S} : \exists\ \mu \text{ such that } \forall \nu,\ V_{(\mu,\nu)}^\phi(s) \geq 0\}$.

3.3 Joint Action-Value Function

Next, for any fixed safety action-value function, $Y \in \mathcal{Q}$, the joint Bellman operator $\mathcal{T}_Y : \mathcal{Q} \to \mathcal{Q}$ corresponding to Eq. 3 is given by:

$$\mathcal{T}_Y[Q](s,a,d) = (1-\gamma)m(s,a,d) + \gamma \min\left\{r(s,a,d) + \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[V_Q(s')\right], Y(s,a,d)\right\} \tag{6}$$

where $m(s,a,d) = \min\{r(s,a,d), g(s)\}$ and $V_Q(s) = \max_{a\in\mathcal{A}}\min_{d\in\mathcal{D}} Q(s,a,d)$.

Proposition 2 (Contraction Mapping of $\mathcal{T}_Y$)**.** *For any arbitrary, fixed $Y \in \mathcal{Q}$, the Bellman operator defined in Eq. 6 induces a γ-contraction mapping under the supremum norm. That is, for a given discount factor $\gamma \in [0,1)$, and any $Q, \widetilde{Q} \in \mathcal{Q}$, we have $\left\|\mathcal{T}_Y[Q] - \mathcal{T}_Y[\widetilde{Q}]\right\|_\infty \leq \gamma \left\|Q - \widetilde{Q}\right\|_\infty$. (Proof in Appendix B.2)*

Thus, for a fixed safety action-value function Y, there exists a unique fixed point Q_Y of the joint Bellman operator.

3.4 Objective Function of Joint Action-Value Function

Given the value functions (Q, Y), we derive the *joint objective function* from robust optimal stopping theory to provide a principled interpretation of the joint value function. To this end, we first define the task-optimizing and safety-maximizing regions, as illustrated in Fig. 1. For a given joint action-value function q, safety critic y and the joint policy $\pi = (\mu, \nu)$, let $V_q^\psi(s) := r(s, \mu(s), \nu(s, \mu(s))) + \mathbb{E}[V_q(s')]$, where $V_q(s) = \max_a \min_d q(s, a, d)$. Define the safety value function $V_\pi^\phi(s) := Q_\pi^\phi(s, \mu(s), \nu(s, \mu(s))$.

Definition 1 (Task and safety optimizing regions)**.** *The region where the task value function is strictly lower than the safe value function is defined as the task-optimizing region:* $\Omega^\psi = \left\{ s \in \mathcal{S} \;\middle|\; V_q^\psi(s) < V_\pi^\phi(s) \right\}$*. The region where the task value function is greater than the safe value function is defined as the safety-optimizing region:* $\Omega^\phi = \left\{ s \in \mathcal{S} \;\middle|\; V_q^\psi(s) \geq V_\pi^\phi(s) \right\}$.

Theorem 1 (Objective Function)**.** *Given a Borel-measurable safety maximizing region,* $\Omega^\phi \in \mathcal{B}(\mathcal{S})$*, define the first switching time* $\tau : \mathcal{W} \to \{0, \cdots, T\}$ *as* $\tau = \inf\{k \in \{0, \cdots, T\} : s_k \in \Omega^\phi\}$*, where* T *is the time horizon and* $\inf \emptyset = T$ *(*$T \to \infty$ *for infinite horizon setting). Let* Λ_k *be the set of stopping times* τ *bounded by time horizon* k*, i.e.,* $0 \leq \tau \leq k$*. Let* ω *be trajectories sampled using the joint policy as* $a_t, d_t \sim \pi(s_t)$*, with the dynamics* $s_{t+1} \sim p(\cdot|s_t, a_t, d_t)$*. Then, the corresponding joint objective function that the joint policy maximizes is given by: (Proof in Appendix B.3)*

$$J(\pi) := \inf_{\tau \in \Lambda_T} \mathbb{E}_{\omega \sim \pi} \left[\sum_{t=0}^{\tau-1} \gamma^t r(s_t, a_t, d_t) + \gamma^\tau \left(\underbrace{(1-\gamma) \min\{r(s_\tau, a_\tau, d_\tau), g(s_\tau)\} + \gamma V_Y^\phi(s_\tau)}_{:= J_\tau} \right) \middle| s_0 = s \right] \quad (7)$$

Intuitively, the stopping time $\tau \in \Lambda_T$ serves as a *safety inspector*: once the trajectory leaves Ω^ψ, the further accrual of expected return is replaced by the *terminal safety value* J_τ (dominated by V_Y^ϕ as $\gamma \to 1$). Hence, task performance can be optimized as long as the trajectory remains inside Ω^ψ. We next analyze the coupled evolution of the safety and joint value functions.

3.5 Coupled Fixed Point

Let $\pi_Q = (\mu_Q, \nu_Q)$ be the maximin policy derived from Q, as in Eq. 4. Since the map $Q \mapsto \pi_Q$ is generally discontinuous, small perturbations of Q can change the argmax/argmin completely. The optimal solution we seek is a pair (Q^*, Y^*) satisfying $Y^* = Q^\phi_{\pi_{Q^*}}, \quad Q^* = \mathcal{T}_{Y^*} Q^*$. This is a coupled system: the safety critic Y^* equals the safety value of the greedy policy of Q^*, and Q^* is the fixed point of its own Bellman operator, "*truncated*" by Y^*. Thus, the iteration $Q_{k+1} = \mathcal{T}_{Q^\phi_{\pi_{Q_k}}}(Q_k)$, the composition $Q \mapsto \mathcal{T}_{Q^\phi_{\pi_Q}}(Q)$ is not guaranteed to induce a contraction by our current analysis, as $Q^\phi_{\pi_Q}$ can change discontinuously due the discontinuous mapping $Q \mapsto \pi_Q$. This motivates a two-timescale stochastic approximation learning framework in Section 4, where the joint and safety value functions are learned on separate timescales.

4 Two-Timescale Stochastic Approximation

In this section, we analyze the coupled learning dynamics in finite/tabular settings. For clarity, we present a synchronous update analysis here, under which every state-action-disturbance coordinate is updated at each iteration; extension to asynchronous settings is discussed in Appendix B.8. Let $\hat{\mathcal{T}}^\phi_{n+1, \pi_{Q_n}}$ and $\hat{\mathcal{T}}_{n+1, Y}$ denote the stochastic estimators of the Bellman targets, with $\pi_{Q_n} \in \Pi(Q_n)$, where for $q \in \mathcal{Q}$, $\Pi(q)$ denotes the finite correspondence of the maximin policies:

$$\Pi(q) := \left\{ (\mu, \nu) : \mu(s) \in \arg\max_a \min_d q(s, a, d), \;\; \nu(s, \mu(s)) \in \arg\min_d q(s, \mu(s), d) \right\} \quad (8)$$

Assumption 1 (Conditional unbiasedness and Martingale-difference noise (Tsitsiklis & Van Roy, 1997))**.** *Let* $\{\mathcal{H}_n\}_{n \geq 0}$ *denote the filtration generated by the learning history up to iteration* n*. Then,*

$$\mathbb{E}\left[\hat{\mathcal{T}}^\phi_{n+1, \pi_{Q_n}}[Y_n] \middle| \mathcal{H}_n \right] = \mathcal{T}^\phi_{\pi_{Q_n}}[Y_n], \quad \mathbb{E}\left[\hat{\mathcal{T}}_{n+1, Y_n}[Q_n] \middle| \mathcal{H}_n \right] = \mathcal{T}_{Y_n}[Q_n] \quad (9)$$

Note that Assumption 1 concerns *unbiased estimation* of the Bellman target itself.

Assumption 2 (Step-size conditions (Robbins & Monro, 1951)). *The step-sizes, $\alpha_n, \beta_n \in (0, 1]$, satisfy Robbins-Monro conditions: $\sum_{n=0}^{\infty} \alpha_n = \sum_{n=0}^{\infty} \beta_n = \infty$, $\sum_{n=0}^{\infty}(\alpha_n^2 + \beta_n^2) < \infty$, $\lim_{n\to\infty} \alpha_n \to 0$, $\lim_{n\to\infty} \beta_n \to 0$, and additionally, $\lim_{n\to\infty} \alpha_n/\beta_n = 0$.*

Additionally, the total martingale noise is assumed to have uniformly bounded conditional variance under bounded iterates. Next, under Temporal difference learning, the bootstrapped updates with step-size sequences $\{\alpha_n\}, \{\beta_n\}$ are given by:

$$Y_{n+1} = Y_n + \beta_n \left[\hat{\mathcal{T}}^{\phi}_{n+1,\pi_{Q_n}}[Y_n] - Y_n + \xi^{\phi}_{n+1}\right], \quad Q_{n+1} = Q_n + \alpha_n \left[\hat{\mathcal{T}}_{n+1,Y_n}[Q_n] - Q_n + \xi_{n+1}\right] \tag{10}$$

where ξ^{ϕ}_{n+1} and ξ_{n+1} are martingale difference perturbations (Borkar, 2009), with $\mathbb{E}[\xi^{\phi}_{n+1}|\mathcal{H}_n] = \mathbb{E}[\xi_{n+1}|\mathcal{H}_n] = 0$. As the map $Q \mapsto Y_{\pi_Q}$ need not be continuous or Lipschitz, the standard two-timescale stochastic approximation need not apply. Thus, we analyze the fast timescale recursion as a stochastic recursive inclusion in Section 4.1 using the *set-valued two-timescale framework* (Yaji & Bhatnagar, 2020; Borkar, 2023).

4.1 Fast Timescale: Safety Critic Tracking

Under Assumption 2, even with a *quasi-static* Q_n, at the *switching time*, two or more policies can have ties in Eq. 8, but can have different safety-critic evaluations. This motivates a *set-valued* fast tracking analysis. To characterize the asymptotic behavior of the discrete TD recursion, let $h_\pi(y) := \mathcal{T}^{\phi}_{\pi}[y] - y$ denote the limiting dynamics induced by Assumption 2 for each $\pi \in \Pi(q)$. Thus, for a given $q \in \mathcal{Q}$, the limiting differential inclusion is given by:

$$\dot{y}(t) \in \Psi(q, y(t)), \quad \text{with } \Psi(q, y) := \overline{\text{conv}}\{h_\pi(y) : \pi \in \Pi(q)\} \tag{11}$$

where $\overline{\text{conv}}\{\cdot\}$ denotes a closed convex hull of a set. Next, let $\Lambda(q) := \{y(0) : y : \mathbb{R} \to \mathcal{Y}$ is a bounded complete solution of Eq. 11$\}$

Lemma 1 (Set-valued fast timescale tracking). *Under Assumptions 1-2, suppose the iterates are almost surely bounded. Then,*

1. *for every quasi-static q, Eq. 11 admits a nonempty compact global attracting set $\Lambda(q)$;*
2. *$q \mapsto \Lambda(q)$ is upper semicontinuous; and*
3. *the coupled iterates satisfy $\operatorname{dist}((Y_n, Q_n), \operatorname{Gr}\Lambda) \to 0$ almost surely, where $\operatorname{Gr}\Lambda := \{(y, q) : y \in \Lambda(q)\}$ is the graph of Λ. (Proof in Appendix B.4)*

Note that when there are no ties, $\Pi(q) = \{\pi_q\}$, and Eq. 11 reduces to an ODE, and $\Lambda(q) = \{Y_{\pi_q}\}$, recovering the usual single-policy fast timescale behavior locally.

4.2 Slow Timescale: Averaged Differential Inclusion

Section 4.1 shows that the coupled iterates asymptotically approach the graph of the compact global attractor $\Lambda(q)$ for some quasi-static q. In this section, we show that the slow recursion is governed by a differential inclusion obtained by averaging over admissible occupation measures associated with the equilibrated fast time-scale dynamics. For $y \in \mathcal{Y}$ and $q \in \mathcal{Q}$, define the *slow-timescale Bellman residual map* to be $G(y, q) := \mathcal{T}_y[q] - q$. Next, following (Yaji & Bhatnagar, 2020), let $\mathfrak{D}(q)$ denote the set of admissible limiting occupation measures describing the *equilibrated* fast time-scale dynamics at some q; in particular, every $\mathbb{P} \in \mathfrak{D}(q)$ is supported on $\Lambda(q)$. The corresponding *averaged* slow-timescale residual is given by:

$$\overline{G}(q) := \left\{ \int_{\Lambda(q)} G(y, q)\mathbb{P}(dy) \ : \ \mathbb{P} \in \mathfrak{D}(q) \right\} \tag{12}$$

Note that when $\Lambda(q) = \{y_{\pi_q}\}$, $\mathfrak{D}(q) = \{\delta_{y_{\pi_q}}\}$, and thus, $\overline{G}(q) = \{G(y_{\pi_q}, q)\}$. The integral in Eq. 12 captures the *long-run average effect* of the equilibrated fast timescale behavior in y on the slow-timescale variable q.

Theorem 2 (Slow-timescale differential-inclusion limit). *Let $\overline{Q}(\cdot)$ denote the continuous affine interpolation of $\{Q_n\}$ on the slow-timescale. Under Assumptions 1-2, Lemma 1, and almost surely bounded iterates $\{(Y_n, Q_n)\}$, $\overline{Q}(\cdot)$ is almost surely an asymptotic pseudotrajectory of the averaged differential inclusion $\dot{q}(t) \in \overline{G}(q(t))$. Consequently, its ω-limit set $\mathcal{U} := \omega(\overline{Q})$ is a nonempty compact invariant internally chain-transitive set of the limiting differential-inclusion dynamics. (Proof in Appendix B.5)*

5 SAFE CONFIGURATION SET AND FORWARD INVARIANCE

Having established the limiting dynamics of the discrete TD learning iterates, we now show that these dynamics preserve the safety certificate: the set of joint action-value functions whose induced policy is everywhere safe is forward invariant, so a configuration that certifies safety cannot evolve into one that does not.

Definition 2 (Kernel-preserving actions). *An action is kernel-preserving at state s, if it keeps the next state inside the viability kernel against every disturbance $\mathcal{A}^{\Omega}(s) := \{a \in \mathcal{A} : P(s' \in \Omega | s, a, d) = 1 \quad \forall\, d \in \mathcal{D}\}$, where $\kappa, \epsilon_0 > 0$ without loss of generality.*

Definition 3 (Index Sets and Relaxed Constraint Sets). *For each state $s \in \Omega$ and each action $a \notin \mathcal{A}^{\Omega}(s)$, fix a disturbance $\hat{d}(s,a) \in \mathcal{D}$ with $P(s' \notin \Omega | s, a, \hat{d}(s,a)) > 0$. Then the index sets are defined as follows: $\mathcal{I} := \{(s,a,d) : s \in \Omega, a \in \mathcal{A}^{\Omega}(s), d \in \mathcal{D}\}$ and $\mathcal{J} := \{(s,a,\hat{d}(s,a)) : s \in \Omega, a \notin \mathcal{A}^{\Omega}(s)\}$. Let $\mathcal{C}_\epsilon := \{q \in \mathcal{Q} : q(x) \geq -\epsilon \;\forall\, x \in \mathcal{I}, q(x) \leq -\kappa + \epsilon \;\forall\, x \in \mathcal{J}\}$ denote the relaxed constraint set, for $\epsilon \geq 0$ i.e., $\mathcal{C}_\epsilon$ relaxes each defining inequality of $\mathcal{C}_0$ by ϵ.*

Definition 4 (Safe Configuration Set). *The safe configuration set is defined as the set of joint Q-functions whose induced policy is kernel-preserving:*

$$\mathcal{Q}_{cfg} := \mathcal{C}_0 = \{q \in \mathcal{Q} : q(x) \geq 0 \;\forall\, x \in \mathcal{I}, q(x) \leq -\kappa \;\forall\, x \in \mathcal{J}\} \tag{13}$$

Next, let $\rho(q) := \mathrm{dist}(q, \mathcal{Q}_{\mathrm{cfg}})$ denote the distance to the set $\mathcal{Q}_{\mathrm{cfg}}$. Details on the geometry of the set is provided in Appendix B.6.1. Throughout, we take $\epsilon_0 \in (0, \kappa/2)$, which ensures that the two families of constraints remain strictly separated on all of $\mathcal{C}_{\epsilon_0}$. As the coupled iterates approach the graph of $\Lambda(\cdot)$, safety certification (both positive and negative) must hold throughout $\Lambda(q)$.

Assumption 3 (Negative Certification of the fast attractor). $\exists\, \sigma_\Lambda > 0$, *such that, at unit margin scale,* $y(x) \;\leq\; -\sigma_\Lambda \;\forall\, x \in \mathcal{J}, y \in \Lambda_1(q), q \in \mathcal{C}_{\epsilon_0}$.

The corresponding condition on $\mathcal{I}$ is preserved in the fast inclusion, and hence, the certificate holds throughout $\Lambda(q)$. Note that $\Pi(q)$ is greedy with respect to q and is hence independent of margin scale, and the safety Bellman operator of Eq. 5 is positively homogeneous of degree one in the margin, i.e., $\Lambda_K(q) = K\Lambda_1(q)$. Let the Assumption 3 hold and let $\Theta := ((1-\gamma)R_{\max} + \kappa)/\gamma$. Then, for every $K \geq \underline{K} := \frac{(1-\gamma)R_{\max}+\kappa}{\gamma\sigma_\Lambda}$, one has $y(x) \leq -\Theta$ for all $x \in \mathcal{J}$, all $y \in \Lambda(q)$, and all $q \in \mathcal{C}_{\epsilon_0}$. If $\mathcal{J} = \emptyset$, the bound on K and the conclusion hold trivially. (Proof in Appendix B.6)

Theorem 3 (Forward invariance and exponential attraction). *Let Assumption 3 hold and let $K \geq \underline{K}$. If $\rho(q(0)) < \epsilon_0$, then every solution of $\dot{q}(t) \in \overline{G}(q(t))$ satisfies $\rho(q(t)) \;\leq\; \rho(q(0))\, e^{-(1-\gamma)t} \quad \forall\, t \geq 0$. In particular, if $q(0) \in \mathcal{Q}_{cfg}$ then $q(t) \in \mathcal{Q}_{cfg}\; \forall\, t \geq 0$, and $\mathcal{Q}_{cfg}$ is exponentially attracting with basin containing the open neighbourhood $\{q : \rho(q) < \epsilon_0\}$. (Proof in Appendix B.6)*

We next check that membership in $\mathcal{Q}_{\mathrm{cfg}}$ certifies safety of the induced policy.

Corollary 1 (Safety of the induced policy). *Let $q(0) \in \mathcal{Q}_{cfg}$. Then at every $t \geq 0$ the induced policy $\pi_{q(t)}$ satisfies $\inf_{s \in \Omega} V^{\phi}_{\pi_{q(t)}}(s) \geq 0$ from any $s_0 \in \Omega$, the state trajectory generated by $\mu_{q(t)}$ remains in Ω almost surely, for all time and against every disturbance sequence. (Proof in Appendix B.6)*

6 CONVERGENCE TO OPTIMAL SAFETY-CONSTRAINED TASK PERFORMANCE

Forward invariance ensures that the agent eventually learns a policy that is guaranteed to be safe. In this section, we show that the system not only remains safe but also maximizes task performance in the safe set.

Definition 5 (Optimal safety-constrained task value function). *Let $\mathcal{Q}_\Omega := \mathbb{R}^{\mathcal{I}} = \{\mathfrak{q} : \mathcal{I} \to \mathbb{R}\}$ denote the restricted action-value space on the index set $\mathcal{I}$. Define the robust safety-constrained task Bellman operator $\mathcal{T}_\Omega : \mathcal{Q}_\Omega \to \mathcal{Q}_\Omega$:*

$$\mathcal{T}_\Omega[\mathfrak{q}](s,a,d) = r(s,a,d) + \gamma \mathbb{E}\left[\max_{a' \in A^{\Omega}(s')} \min_{d' \in \mathcal{D}} \mathfrak{q}(s',a',d')\right], \quad (s,a,d) \in \mathcal{I} \tag{14}$$

Note that since $A^{\Omega}(s) \neq \emptyset$ for $s \in \Omega$, by contraction, $\mathcal{T}_\Omega[\cdot]$ has a unique fixed point $Q^*_\Omega = \mathcal{T}_\Omega Q^*_\Omega$, and thus, the optimal safety-constrained task value function is given by $V^*_\Omega(s) =$

$\max_{a\in A^\Omega(s)} \min_{d\in\mathcal{D}} Q^*_\Omega(s,a,d)$*, which is the maximum expected return that a safe policy can accrue by keeping the agent in* Ω*. Moreover, let* $\mathcal{Q}^* := \{q \in \mathcal{Q}_{\text{cfg}} : q|_{\mathcal{I}} = Q^*_\Omega\}$*. Thus, every* $q \in \mathcal{Q}^* \subseteq \mathcal{Q}$ *corresponds to a policy that is safe and attains the maximal possible return among all safe policies.*

Let $Q^{\phi,1}_\pi$ denote the safety critic obtained by scaling the margin $g(\cdot)$ by $K = 1$. By homogeneity, $Q^\phi_\pi := Q^{\phi,K}_\pi = KQ^{\phi,1}_\pi$. Next, let $\delta_\Omega := \inf_{q\in\mathcal{Q}_{\text{cfg}}} \min_{\pi\in\Pi(q)} \min_{s\in\Omega} V^{\phi,1}_\pi > 0$ represent the margin of the safe policy that comes closest to $\partial\Omega$.

Theorem 4 (Convergence to optimal safety-constrained task performance)**.** *Let the sufficient margin scale* K^* *be defined as* $K^* := \frac{R_{\max}+\overline{V}}{\delta_\Omega} = \frac{(2-\gamma)R_{\max}}{(1-\gamma)\delta_\Omega}$*, with* $\overline{V} := \frac{R_{\max}}{1-\gamma}$*. Let* $\mathcal{U} \subset \{q : \rho(q) < \epsilon_0\}$*. Then, under Theorems 2-3, almost surely bounded iterates* (Y_n, Q_n) *and* $K \geq K^* \geq \underline{K}$*,* $\delta_\Omega > 0$*, every internally chain-transitive set of the averaged slow-timescale differential inclusion contained in* $\mathcal{Q}_{\text{cfg}}$ *is a subset of* $\mathcal{Q}^*$*:* $\mathcal{U} \subseteq \mathcal{Q}^* := \{q \in \mathcal{Q}_{\text{cfg}} : q|_{\mathcal{I}} = Q^*_\Omega\}$*, and thus,* $\operatorname{dist}(Q_n, \mathcal{Q}^*) \to 0$ *almost surely. Moreover, on* $\mathcal{U}$ *the averaged slow-timescale residual on kernel-preserving coordinates* $\mathcal{I}$ *is given by* $\overline{G}(q)|_{\mathcal{I}} = \{\mathcal{T}_\Omega[q|_{\mathcal{I}}] - q|_{\mathcal{I}}\}$*. Thus, every corresponding limiting controller attains the optimal task value* V^*_Ω *on* Ω *while selecting only kernel-preserving actions. (Proof in Appendix B.7).*

The sufficient margin scale K^* ensures that the safety truncation is never active when $\delta_\Omega > 0$. Consequently, if $\delta_\Omega \to 0$, then $K^* \to \infty$; if $\delta_\Omega = 0$, Theorem 4 no longer guarantees convergence to the optimal safety-constrained task value for any finite K, motivating a finite-K characterization. Let Γ_Ω be the set of kernel-preserving controller policies, and for $\zeta \geq 0$, define $V^*_{\Omega,\zeta}(s) := \sup_{\mu\in\Gamma_{\Omega,\zeta}} \inf_\nu V^\psi_{\mu,\nu}(s)$ and $\Gamma_{\Omega,\zeta} := \left\{\mu \in \Gamma_\Omega : \inf_\nu \inf_{s\in\Omega} V^{\phi,1}_{(\mu,\nu)}(s) \geq \zeta\right\}$. Let $\Delta_{\text{marg}}(\zeta) := \|V^*_\Omega - V^*_{\Omega,\zeta}\|_\infty$, and let V^*_K denote the optimal joint value of Eq. 3 over kernel-preserving controllers, for some finite $K \geq \underline{K}$.

Theorem 5 (Finite-K characterization)**.** *Define* $\zeta_K := \frac{R_{\max}+\overline{V}}{K} = \frac{(2-\gamma)R_{\max}}{(1-\gamma)K}$*. Assume there exists* $\overline{\zeta}$ *such that* $\Gamma_{\Omega,\zeta_K} \neq \emptyset$ *for all* $0 < \zeta_K < \overline{\zeta}$ *(sufficiently large* K*). Then,* $0 \leq V^*_\Omega(s) - V^*_K(s) \leq \Delta_{\text{marg}}(\zeta_K)$ *for* $s \in \Omega$*. Hence, if* $\Delta_{\text{marg}}(\zeta) \to 0$ *as* $\zeta \downarrow 0$*, then* $V^*_K \to V^*_\Omega$ *as* $K \to \infty$*. (Proof in Appendix B.7).*

Thus, when $\delta_\Omega > 0$, with sufficiently large, finite-K, the tabular two-timescale recursion converges to the optimal safety-constrained task value; when $\delta_\Omega = 0$, the finite-K optimality gap vanishes as $K \to \infty$.

7 Experiments

In this section, we perform empirical evaluations on our proposed framework, where we choose to answer the following questions: (1) *Does the two-timescale learning achieve stable convergence on continuous-control tasks via neural approximations?* (2) *How does our joint policy compare against CMDP baselines in maintaining near-zero violations during evaluation without sacrificing task performance?* (3) *How does the learned safe set dynamically expand during training?*

Appendix C.1 discusses the exact tabular DP solution of the joint Bellman equation on a discrete double integrator, with comparisons to projection and safety-filtering baselines. Here, we instead discuss the results of scaling the optimal joint policy to high-dimensional continuous spaces where exact dynamic programming is intractable. To this end, we introduce `JointSAC`: a deep RL approximation that closely follows the ISAACS framework (Hsu et al., 2023) (built on SAC (Haarnoja et al., 2018)). Controller and disturbance policies are trained using *gradient descent-ascent finite timescale separation* (τ-GDA) (Fiez & Ratliff, 2021).

7.1 Experimental Setup

We benchmark our framework on high-dimensional continuous control tasks in `MuJoCo` (Todorov et al., 2012), utilizing three `Gymnasium` locomotion environments (Towers et al., 2026) and eight `SafeVelocity` environments from `SafetyGymnasium` (Ji et al., 2023). We augment the environment with an adversarial disturbance that actively attacks the agent to either violate safety constraints or to prevent liveness.

Baselines: We compare our method against least-restrictive safety filters (LRSF), and three well-established CMDP baselines. We specifically select these CMDP baselines, as they represent the two

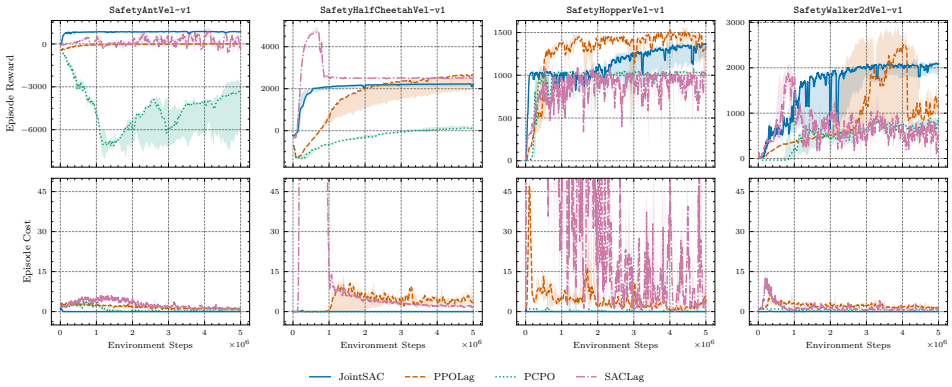

Figure 2: Training performance across 5 seeds (mean ± standard deviation). The top row indicates episodic returns, while the bottom row indicates episodic costs as training progresses. `JointSAC` demonstrates stable convergence with near-zero episode costs.

dominant paradigms of model-free Safe RL: *Lagrangian dual-descent* (`PPOLag`, `SACLag`) and *trust-region projections* `PCPO` from omnisafe (Ji et al., 2024). Additional experiments and details can be found in Appendix C (including the safe set expansion); an ablation of episode return/cost as K varies can be found in Appendix C.4.

7.2 Main Results

Fig. 2 illustrates training curves across four `SafeVelocity` (`v1`) environments (5 seeds). `JointSAC` demonstrates stable convergence to high task rewards with near-zero costs, whereas baselines exhibit higher variance, with only a subset that meets the permissible cost threshold ($d = 0$). Table 1 summarizes final rollout evaluations (300 evaluations per training seed, totaling $1,500$ distinct eval seeds). Across all 11 environments, `JointSAC` maintains near-zero violations (with maximum cost of 0.03 ± 0.08 on `Hopper-v4`), while baselines exhibit inconsistent, sometimes, major cost violations (`PCPO` walks backwards on `SafetyAntVel`). Overall, `JointSAC` yields stable, high task performance while enforcing near-zero safety violations.

Table 1: Evaluation Metrics on Adversarial Environments with values indicating mean ± standard deviation across 5 training seeds and 300 rollouts each (totalling $1,500$ evals). Results indicate that `JointSAC` (Ours) has near-zero costs (maximum of 0.03 ± 0.08 on only `Hopper-v4`).

	Reward					Cost				
Environment	**JointSAC (Ours)**	**Safety Filter** (LRSF)	**PCPO**	**PPO-Lag**	**SAC-Lag**	**JointSAC (Ours)**	**Safety Filter** (LRSF)	**PCPO**	**PPO-Lag**	**SAC-Lag**
Hopper-v4	3266.89 ± 103.80	1169.92 ± 326.17	1373.82 ± 303.63	1074.26 ± 562.70	754.89 ± 495.76	0.03 ± 0.08	0.00 ± 0.00	0.00 ± 0.00	0.97 ± 0.07	0.40 ± 0.54
InvertedPendulum-v4	999.43 ± 1.28	1000.00 ± 0.00	34.77 ± 4.93	506.30 ± 464.77	1000.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.60 ± 0.55	0.00 ± 0.00
Walker2d-v4	2869.59 ± 319.83	973.26 ± 3.10	849.93 ± 94.91	3962.76 ± 1138.50	700.85 ± 357.47	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.25 ± 0.42	0.27 ± 0.39
SafetyAntVel-v0	941.08 ± 8.26	674.19 ± 127.31	-1509.56 ± 1724.06	112.66 ± 230.86	1251.74 ± 562.78	0.00 ± 0.00	0.05 ± 0.04	0.04 ± 0.07	0.89 ± 0.38	0.12 ± 0.12
SafetyAntVel-v1	940.70 ± 9.06	632.65 ± 62.56	-2915.70 ± 2404.71	62.72 ± 103.62	582.13 ± 963.46	0.00 ± 0.00	0.00 ± 0.00	0.38 ± 0.47	1.03 ± 0.15	0.97 ± 1.18
SafetyHalfCheetahVel-v0	1783.42 ± 292.07	986.44 ± 76.15	453.22 ± 247.44	2214.42 ± 656.13	2255.89 ± 100.72	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.02	152.87 ± 341.14	0.15 ± 0.34
SafetyHalfCheetahVel-v1	2238.10 ± 107.59	1152.01 ± 110.50	422.09 ± 204.15	2295.79 ± 725.03	1686.74 ± 1505.66	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.36 ± 0.59	0.47 ± 0.68
SafetyHopperVel-v0	1056.06 ± 0.50	974.38 ± 24.15	936.84 ± 269.07	1016.16 ± 375.81	1045.57 ± 61.95	0.00 ± 0.00	0.00 ± 0.00	0.90 ± 1.99	1.38 ± 1.95	17.06 ± 32.27
SafetyHopperVel-v1	1329.36 ± 89.75	1001.40 ± 13.29	1039.24 ± 8.99	1274.01 ± 557.27	1027.49 ± 18.75	0.02 ± 0.03	0.00 ± 0.00	0.01 ± 0.02	4.11 ± 8.84	0.00 ± 0.00
SafetyWalker2dVel-v0	1842.44 ± 309.17	547.07 ± 138.62	640.29 ± 352.84	1583.43 ± 1248.16	462.38 ± 353.09	0.01 ± 0.01	0.35 ± 0.33	0.22 ± 0.36	1.56 ± 1.48	0.76 ± 0.61
SafetyWalker2dVel-v1	2056.53 ± 332.37	575.99 ± 131.44	828.76 ± 141.58	1178.47 ± 1061.40	433.04 ± 366.89	0.02 ± 0.04	0.21 ± 0.19	0.11 ± 0.14	1.52 ± 1.52	0.60 ± 0.46

8 Conclusion and Limitations

In this paper, we present a novel robust Bellman operator that jointly enables safety-constrained task optimization. We propose a two-timescale stochastic approximation algorithm to compute its fixed point and prove theoretical convergence to a set of optimal safety-constrained task value functions. The optimal joint policy theoretically guarantees safety while maximizing task returns. To scale the computation of the optimal fixed point to continuous-control tasks with a high-dimensional state space, we propose `JointSAC`. Empirical evaluations demonstrate that the neural approximation yields stable, low-variance training, while achieving near-zero safety violations.

Limitations. Firstly, while our theoretical analysis ensures strict safety, scaling to high-dimensional continuous spaces necessitates neural approximations, violating strict forward invariance. Secondly, theoretical convergence relies on a heavily scaled safety margin ($K \geq K^*$). This can introduce numerical instabilities; in practice, K needs to be heuristically tuned as high as practically possible. Finally, our approach assumes well-defined, continuous safety margins under full state observability, which are often difficult to perfectly specify in unstructured real-world environments.

AI USE STATEMENT

In this work, we used generative AI tools for grammar and spelling checks, word choice, rephrasing, and assistance with editing code and launching experiments.

We did not use generative AI tools to develop theoretical results, write full pages, identify citations, or generate research ideas. Any AI-generated code was reviewed by at least one author. The authors take full responsibility for the final content of this work, including the text, claims, code, and experimental results.

ETHICS STATEMENT

Our work develops a unified, theoretical framework for RL algorithms to optimize for task performance, while *adhering to safety at all times*. Thus, the work does not have immediate ethical concerns.

REPRODUCIBILITY STATEMENT

We will provide the source code for all of our experiments upon publication. The hyperparameters used for the main experiments and additional details are given in Appendix C, and the proofs for all theoretical results are provided in Appendix B.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 22–31. JMLR.org, 2017.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Erick Trevino Aguilar. T-systems and the lower snell envelope, 2009. URL https://arxiv.org/abs/0902.4245.

Albert Altarovici, Olivier Bokanowski, and Hasnaa Zidani. A general hamilton-jacobi framework for non-linear state-constrained control problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(2):337–357, 2013.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Jean-Pierre Aubin and Arrigo Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2012. ISBN 9783642695124. URL https://books.google.com/books?id=rVnsCAAAQBAJ.

Andrea Bajcsy, Somil Bansal, Eli Bronstein, Varun Tolani, and Claire J. Tomlin. An efficient reachability-based framework for provably safe autonomous navigation in unknown environments. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 1758–1765, 2019. doi: 10.1109/CDC40024.2019.9030133.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta mathematicae*, 3(1):133–181, 1922.

Arko Banerjee, Kia Rahmani, Joydeep Biswas, and Isil Dillig. Dynamic model predictive shielding for provably safe reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=x2zY4hZcmg.

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

Michel Benaïm and Morris W Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005. doi: 10.1137/S0363012904439301.

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Texts and Readings in Mathematics. Hindustan Book Agency, 2009. ISBN 9789386279385. URL https://books.google.com/books?id=t_JdDwAAQBAJ.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48 of *Texts and Readings in Mathematics*. Springer Singapore, 2nd edition, 2023. doi: 10.1007/978-981-99-8277-6.

Minh Bui, Michael Lu, Reza Hojabr, Mo Chen, and Arrvindh Shriraman. Real-time hamilton-jacobi reachability analysis of autonomous system with an fpga. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1666–1673. IEEE, 2021.

Kaitong Cai, Jusheng zhang, Jing Yang, and Keze Wang. Guardian: Decoupling exploration from safety in reinforcement learning, 2025. URL https://openreview.net/forum?id=e9T6ZZFRZl.

Chenyang Cao, Yucheng Xin, Silang Wu, Longxiang He, Zichen Yan, Junbo Tan, and Xueqian Wang. FOSP: Fine-tuning offline safe policy through world models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dbuFJg7eaw.

Zhepeng Cen, Yihang Yao, Zuxin Liu, and Ding Zhao. Feasibility consistent representation learning for safe reinforcement learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6002–6019. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cen24b.html.

Lawrence C. Evans and Panagiotis E. Souganidis. Differential games and representation formulas for solutions of hamilton-jacobi-isaacs equations. *Indiana University Mathematics Journal*, 33:773–797, 1983. URL https://api.semanticscholar.org/CorpusID:118892068.

Tanner Fiez and Lillian J Ratliff. Local convergence analysis of gradient descent ascent with finite timescale separation. In *International Conference on Learning Representations*, 2021.

Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

Jaime F Fisac, Neil F Lugovoy, Vicenç Rubies-Royo, Shromona Ghosh, and Claire J Tomlin. Bridging hamilton-jacobi safety analysis and reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8550–8556. IEEE, 2019.

Fernando García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

Shangding Gu, Bilgehan Sel, Yuhao Ding, Lu Wang, Qingwei Lin, Ming Jin, and Alois Knoll. Balance reward and safety optimization for safe reinforcement learning: A perspective of gradient manipulation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024a. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i19.30102. URL https://doi.org/10.1609/aaai.v38i19.30102.

Shangding Gu, Laixi Shi, Yuhao Ding, Alois Knoll, Costas Spanos, Adam Wierman, and Ming Jin. Enhancing efficiency of safe reinforcement learning via sample manipulation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=oPFjhl6DpR.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018.

Kai-Chieh Hsu, Vicenç Rubies-Royo, Claire J. Tomlin, and Jaime F. Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. In *Proceedings of Robotics: Science and Systems*, Held Virtually, July 2021. doi: 10.15607/RSS.2021.XVII.077.

Kai-Chieh Hsu, Duy Phuong Nguyen, and Jaime Fernàndez Fisac. Isaacs: Iterative soft adversarial actor-critic for safety. In Nikolai Matni, Manfred Morari, and George J. Pappas (eds.), *Proceedings of the 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*. PMLR, 15–16 Jun 2023. URL https://proceedings.mlr.press/v211/hsu23a.html.

Kai-Chieh Hsu, Haimin Hu, and Jaime 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.

Weidong Huang, Jiaming Ji, Borong Zhang, Chunhe Xia, and Yaodong Yang. Safedreamer: Safe reinforcement learning with world models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tsE5HLYtYg.

Rufus Isaacs. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. John Wiley & Sons, 1965.

Ashish Kumar Jayant and Shalabh Bhatnagar. Model-based safe deep reinforcement learning via a constrained proximal policy optimization algorithm. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=WZmlxIuIGR.

Jiaming Ji, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, Ruiyang Sun, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang. Omnisafe: An infrastructure for accelerating safe reinforcement learning research. *Journal of Machine Learning Research*, 25(285):1–6, 2024.

Dohyeong Kim, Mineui Hong, Jeongho Park, and Songhwai Oh. Conflict-averse gradient aggregation for constrained multi-objective reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogXkmugNZw.

Michael Lu, Jashanraj Gosain, Luna Sang, and Mo Chen. Safe learning in the real world via adaptive shielding with hamilton-jacobi reachability. In Necmiye Ozay, Laura Balzano, Dimitra Panagou, and Alessandro Abate (eds.), *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pp. 1257–1270. PMLR, 04–06 Jun 2025. URL https://proceedings.mlr.press/v283/lu25a.html.

Hannah Markgraf, Shambhuraj Sawant, Hanna Krasowski, Lukas Schäfer, Sebastien Gros, and Matthias Althoff. Safe reinforcement learning using action projection: Safeguard the policy or the environment? *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=DDrGSEYxGU. Expert Certification.

Nikola Milosevic, Johannes Müller, and Nico Scherf. Embedding safety into RL: A new take on trust region methods. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4zRb89SbzG.

Ian M. Mitchell, Alexandre M. Bayen, and Claire J. Tomlin. A time-dependent Hamilton–Jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control*, 50(7):947–957, 2005.

Duy P. Nguyen, Kai-Chieh Hsu, Wenhao Yu, Jie Tan, and Jaime Fernández Fisac. Gameplay filters: Robust zero-shot safety through adversarial imagination. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=Ke5xrnBFAR.

Donggeon David Oh, Justin Lidard, Haimin Hu, Himani Sinhmar, Elle Lazarski, Deepak Gopinath, Emily S Sumner, Jonathan A DeCastro, Guy Rosman, Naomi Ehrich Leonard, and Jaime Fernández Fisac. Safety with agency: Human-centered safety filter with application to ai-assisted motorsports. In *Robotics: Science and Systems*, 2025.

Donggeon David Oh, Duy P Nguyen, Haimin Hu, and Jaime Fernández Fisac. Synthesis and deployment of maximal robust control barrier functions through adversarial reinforcement learning. *arXiv preprint arXiv:2604.13192*, 2026.

Steven Perkins and David S Leslie. Asynchronous stochastic approximation with differential inclusions. *Stochastic Systems*, 2(2):409–446, 2013.

Samuel Pfrommer, Tanmay Gautam, Alec Zhou, and Somayeh Sojoudi. Safe reinforcement learning with chance-constrained model predictive control. In *Learning for Dynamics and Control Conference*, pp. 291–303. PMLR, 2022.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International conference on machine learning*, pp. 2817–2826. PMLR, 2017.

Arunselvan Ramaswamy and Shalabh Bhatnagar. Stochastic recursive inclusion in two timescales with an application to the lagrangian dual problem. *Stochastics*, 88(8):1173–1187, 2016. doi: 10.1080/17442508.2016.1215450. URL https://doi.org/10.1080/17442508.2016.1215450.

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

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, and Haitham Ammar. Sauté rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*, pp. 20423–20443. PMLR, 2022.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Wesley Suttle, Vipul Kumar Sharma, Krishna Chaitanya Kosaraju, Sivaranjani Seetharaman, Ji Liu, Vijay Gupta, and Brian M Sadler. Sampling-based safe reinforcement learning for nonlinear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pp. 4420–4428. PMLR, 2024.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

Mumuksh Tayal, Manan Tayal, and Ravi Prakash. Safe flow q-learning: Offline safe reinforcement learning with reachability-based flow policies, 2026. URL https://arxiv.org/abs/2603.15136.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkfrvsA9FX.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Mark Towers, Ariel Kwiatkowski, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun KG, Rodrigo De Lazcano Perez-Vicente, J K Terry, Andrea Pierré, Sander V Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=qPMLvJxtPK.

J.N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pp. 11480–11491. PMLR, 2021.

Vinayaka G. Yaji and Shalabh Bhatnagar. Stochastic recursive inclusions in two timescales with nonadditive iterate-dependent markov noise. *Math. Oper. Res.*, 45(4):1405–1444, November 2020. ISSN 0364-765X. doi: 10.1287/moor.2019.1037. URL https://doi.org/10.1287/moor.2019.1037.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J. Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rke3TJrtPS.

Zhihe Yang, Yunjian Xu, and Yang Zhang. Q-supervised contrastive representation: A state decoupling framework for safe offline reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hsoyRfvMGu.

Yihang Yao, Zhepeng Cen, Wenhao Ding, Haohong Lin, Shiqi Liu, Tingnan Zhang, Wenhao Yu, and Ding Zhao. OASIS: Conditional distribution shaping for offline safe reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3uDEmsf3Jf.

Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=j5JvZCaDM0.

A PRELIMINARIES

In this section, we extend Sec. 2 to provide definitions of some of the notations and concepts used in the paper, followed by brief preliminaries on Reinforcement Learning, Differential Inclusions, Asymptotic Pseudotrajectories, Internally Chain-Transitive Sets, Constrained Markov Decision Processes, and Hamilton-Jacobi Reachability analysis.

A.1 DEFINITIONS AND NOTATIONS

Definition 6. *(State, Action, and Disturbance Spaces) We consider a discounted MDP with finite state, action and disturbance spaces: $\mathcal{S} \subset \mathbb{R}^n$, $\mathcal{A} \subset \mathbb{R}^m$, and $\mathcal{D} \subset \mathbb{R}^l$ respectively, with the transition kernel $P(\cdot|s,a,d) \in \Delta(\mathcal{S})$.*

Definition 7. *(Reward, Safety Margin, and the Viability Kernel) The reward and the safety margin are given by $r : \mathcal{S} \times \mathcal{A} \times \mathcal{D} \to [0, R_{max}]$, $R_{max} > 0$, and $g : \mathcal{S} \to [-G_{min}, G_{max}]$, with $G_{min}, G_{max} > 0$ respectively. We subsume the extended class-$\mathcal{K}_e$ scaling function $\beta(\cdot)$ into $g(s)$ to align with task return scales (unless stated otherwise). The constraint set is $\mathcal{F}^c := \{s \in \mathcal{S} : g(s) \geq 0\}$, complementing the failure set $\mathcal{F}$. The viability kernel $\Omega \subseteq \mathcal{F}^c$ is the subset of states from which a control policy exists that can keep the agent away from the failure set $\mathcal{F}$ for all times, regardless of the adversarial disturbance.*

Definition 8. *(Action and Disturbance Policies) A deterministic joint policy $\pi = (\mu, \nu)$ pairs the mappings $\mu : \mathcal{S} \to \mathcal{A}$ and $\nu : \mathcal{S} \times \mathcal{A} \to \mathcal{D}$, yielding $\pi(s) = (\mu(s), \nu(s, \mu(s)))$.*

Definition 9. *(Value and Action-Value Functions) the value and action-value function spaces are defined as* $\mathcal{V} := \mathbb{R}^{|\mathcal{S}|}$ *and* $Q := \mathbb{R}^{|\mathcal{S}||\mathcal{A}||\mathcal{D}|}$*, equipped with the supremum norm* $\|\cdot\|_\infty$.

A.2 Reinforcement Learning

Reinforcement Learning (RL) (Sutton & Barto, 1998; Agarwal et al., 2019) consists of data-driven methods enabling agents to compute approximations of the optimal value function and/or optimal policy to an optimal control problem. In RL, the objective is to maximize the expected cumulative sum of rewards over a finite horizon T, or the expected discounted cumulative sum in the infinite horizon setting. Accordingly, the general form of the objective function without disturbances is written as

$$J(\pi) := \mathop{\mathbb{E}}_{\omega\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\middle| s_0\sim\mu_0\right] \tag{15}$$

where ω denotes a sampled trajectory $\omega = (s_0, a_0, s_1, a_1, \dots)$ generated by the policy π and $\mu_0 : \mathcal{S} \to [0,1]$ denotes the initial state distribution. Considering stochastic dynamics, the dynamic programming principle associated with the action-value function $Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ representing the maximum expected cumulative discounted reward takes the form of a discrete-time Bellman equation (Bellman, 1957)

$$Q(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} Q(s',a')\right] \tag{16}$$

For any $Q \in \mathcal{Q}$, the Bellman operator $\mathcal{T}_\mathcal{M} : \mathcal{Q} \to \mathcal{Q}$ associated with Eq. 16 is defined as:

$$\mathcal{T}[Q](s,a) := r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} Q(s',a')\right] \tag{17}$$

The Bellman operator defined in Eq. 17 is a contraction mapping on $\mathcal{Q}$ (under the L_∞ norm). By the Banach Fixed-Point Theorem (Banach, 1922), this implies that iteratively applying $\mathcal{T}_\mathcal{M}$ to any initial action-value function Q converges to a unique fixed point Q^*, which is the solution to Eq. 16. This theoretical guarantee underpins modern Deep RL, justifying the use of neural networks to approximate this fixed point in high-dimensional state spaces.

A.3 Differential Inclusions, Asymptotic Pseudotrajectories, and Internally Chain-Transitive Sets

Let $E = \mathbb{R}^d$ be equipped with the supremum norm $\|\cdot\|_\infty$. Consider a set-valued map $\Psi : E \rightrightarrows E$. Ψ is *Marchaud* if it is upper semicontinuous, has nonempty compact convex values, and has at most linear growth. Under these conditions, the differential inclusion

$$\dot{x}(t) \in \Psi(x(t)) \tag{18}$$

admits absolutely continuous solutions from every initial condition (Aubin & Cellina, 2012; Benaïm et al., 2005). Since solutions need not be unique, Eq. 18 induces a set-valued dynamical system $\Phi_t(x) := \{z(t) : z(0) = x, \dot{z}(\tau) \in \Psi(z(\tau)) \text{ a.e.}\}$

Definition 10. *(Asymptotic Pseudotrajectory (Benaïm & Hirsch, 1996)) Let* Φ_t *denote the set-valued flow of Eq. 18. A continuous function* $y : [0,\infty) \to E$ *is an asymptotic pseudotrajectory (APT) of* Φ*, if for every T > 0,*

$$\lim_{t\to\infty}\sup_{0\le h\le T}\operatorname{dist}\left(y(t+h), \Phi_h\left(x(t)\right)\right) = 0 \tag{19}$$

Definition 11. *(Internally Chain-Transitive Set (Benaïm et al., 2005)) A compact invariant set* $A \subseteq E$ *is internally chain-transitive (ICT), if for any* $a, b \in A$ *and any* $\epsilon > 0, T > 0$*, there exist points* $a = x_0, x_1, \dots, x_k = b$ *in A and times* $t_1, \dots, t_k \ge T$ *such that*

$$\operatorname{dist}\left(\Phi_{t_i}\left(x_{i-1}\right), x_i\right) < \epsilon \qquad i = 1, \dots, k \tag{20}$$

Theorem 6. *(Limit set of a bounded APT (Benaïm et al., 2005)) Let Ψ satisfy the preceding conditions, and let $y : [0, \infty) \to E$ be a bounded APT of $\dot{x}(t) \in \Psi(x(t))$. Then, its ω-limit set:*

$$\omega(y) = \bigcap_{t \geq 0} \overline{\{y(s) : s \geq t\}} \tag{21}$$

is a non-empty, compact, invariant and internally chain-transitive set-valued dynamics Φ.

In the two-timescale setting considered in this work, the required APT property is obtained from the *stochastic recursive inclusion* framework of (Yaji & Bhatnagar, 2020): after the "*equilibration*" of the fast timescale, the slow-timescale Bellman residual is averaged w.r.t admissible limiting occupation measures of the quasi-static fast timescale dynamics, yielding the averaged differential inclusion in Theorem 2.

A.4 Constrained Markov Decision Processes

A Constrained Markov Decision Process (CMDP) (Altman, 1999) extends the standard Markov Decision Process (MDP) framework by introducing auxiliary cost functions that encode safety constraints and perhaps, resource limitations. Formally, a CMDP is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, c, \gamma, \mu_0)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defines the transition dynamics, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the task reward function, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the constraint cost function, $\gamma \in [0, 1)$ denotes the discount factor and $\mu_0 : \mathcal{S} \to [0, 1]$ denotes the initial state distribution.

Let $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ denote a stochastic policy, with the corresponding *expected discounted cumulative return* $J_r(\pi)$ and *expected discounted cumulative cost* $J_c(\pi)$ being defined as:

$$J_r(\pi) := \mathbb{E}_{\omega \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 \sim \mu_0\right], \quad J_c(\pi) := \mathbb{E}_{\omega \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \middle| s_0 \sim \mu_0\right] \tag{22}$$

where $\omega = (s_0, a_0, s_1, a_1, \dots)$ denotes a trajectory generated by the policy π, with $s_0 \in \mu_0$, $a_t \sim \pi(\cdot|s_t)$, and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. Then, the objective in a CMDP is to find an optimal policy π^* that maximizes the expected return, while ensuring that the cost remains below a predefined safety threshold $d \in \mathbb{R}$:

$$\max_{\pi \in \Pi} J_r(\pi) \quad \text{s.t.} \quad J_c(\pi) \leq d \tag{23}$$

where Π is the set of all stationary stochastic policies. To solve this constrained optimization problem, a standard approach is to employ the *Lagrangian relaxation* by introducing a dual variable $\lambda \geq 0$ (Lagrange multiplier). One can then obtain the unconstrained Lagrangian function, and the resulting *min-max saddle-point problem*:

$$\min_{\lambda \geq 0} \max_{\pi \in \Pi} \mathcal{L}(\pi, \lambda), \quad \text{where} \quad \mathcal{L}(\pi, \lambda) := J_r(\pi) - \lambda(J_c(\pi) - d) \tag{24}$$

Under Slater's condition (assuming that at least one strictly safe policy exists, where $J_c(\pi) \leq d$), strong duality holds, and optimizing this unconstrained minimax objective is equivalent to solving the primal CMDP. More details can be found in seminal works on CMDPs(Altman, 1999; Achiam et al., 2017; Tessler et al., 2019; Stooke et al., 2020).

A.5 Hamilton-Jacobi Safety Analysis

Consider a continuous-time dynamical system $\dot{s}(t) = f(s(t), a(t), d(t))$, where $s \in \mathcal{S}, a \in \mathcal{A}$ and $d \in \mathcal{D}$ as in Definition 2.1. We assume that the specification of a failure set is given by a margin function $l : \mathcal{S} \to \mathbb{R}$ such that $\mathcal{F}^c := \{s : g(s) \geq 0\}$. The goal of safety analysis is then to determine the *safe set* $\Omega \subseteq \mathcal{F}^c \subseteq \mathcal{S}$ which is the set of all initial states from which there exists a control policy that can keep the system away from the failure set $\mathcal{F}$, at all times, regardless of the adversarial disturbance (Mitchell et al., 2005). This is formalized as a *zero-sum differential game*, yielding the robust safety value function $V(s)$:

$$V(s) = \inf_{\mathcal{N}[\boldsymbol{d}](\cdot)} \sup_{\boldsymbol{a}(.)} \inf_{t \geq 0} \; g\left(\xi_{s_0}^{\boldsymbol{a}, \mathcal{N}[\boldsymbol{d}]}(t)\right) \tag{25}$$

where $\xi_{s_0}^{\boldsymbol{a},\mathcal{N}[\boldsymbol{d}]}(\cdot)$ denotes the trajectory starting at the state s_0, and evolving according to the control and disturbance trajectories $(\boldsymbol{a}, \mathcal{N}[\boldsymbol{d}])$, with $\mathcal{N}[\boldsymbol{d}](\cdot)$ is defined to be the set of *non-anticipative strategies* for the disturbance, as defined in Evans & Souganidis (1983) and Fisac et al. (2018). By the principle of Dynamic Programming, $V(s)$ is the unique viscosity solution to the Hamilton-Jacobi-Isaacs (HJI) variational inequality:

$$0 = \min\left\{g(s) - V(s,t), \frac{\partial V}{\partial t}(s,t) + \max_{a\in\mathcal{A}} \min_{d\in\mathcal{D}} \nabla_s V^\top(s,t) f(s,a,d)\right\} \tag{26}$$

where $V(s,T) = g(s), \forall s \in \mathcal{S}, t \in [0,T]$. In the infinite-horizon $T \to \infty$ case, $V(s,t)$ becomes independent of t as long as there exists a nonempty safe set. Under stochastic dynamics, one can obtain the infinite-horizon Bellman equation as:

$$V(s) = \min\left\{g(s), \max_{a\in\mathcal{A}} \min_{d\in\mathcal{D}} \mathbb{E}_{s'\sim p(\cdot|s,a,d)}\left[V(s')\right]\right\} \tag{26}$$

where $s' \sim p(\cdot|s,a,d) \iff s' = s + f(s,a,d)\Delta t + \varepsilon_t$. However, the Bellman operator that governs Eq. 26 doesn't induce a contraction mapping in the space of safe value functions. To leverage techniques from TD-learning and deep RL, Fisac et al. (2019) introduces a discount factor $\gamma \in (0,1)$ that ensures contraction on the space of value functions. This results in the following *time-discounted Bellman equation*:

$$V(s) = (1-\gamma)g(s) + \gamma \min\left\{g(s), \max_{a\in\mathcal{A}} \min_{d\in\mathcal{D}} \mathbb{E}_{s'\sim p(\cdot|s,a,d)}\left[V(s')\right]\right\} \tag{27}$$

For more details, please refer to prior work(Evans & Souganidis, 1983; Mitchell et al., 2005; Fisac et al., 2018; 2019).

B Proofs of Theoretical Results

In this section, we provide detailed proofs and remarks on the theoretical results in Sec. 3, Sec. 3.4, Sec. 4, Sec. 5, and Sec. 6.

B.1 Safety Action-Value of the Joint Policy (Sec. 3.2)

Given a joint policy π, the safety Bellman equation is defined as:

$$Q_\pi^\phi(s,a,d) = (1-\gamma)g(s) + \gamma \min\left\{g(s), \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mu(s')} \min_{d'\in\nu(s',a')} Q_\pi^\phi(s',a',d')\right]\right\} \tag{2}$$

Note that $\nu(\cdot,a')$ refers to the specific sample a' drawn from $\mu(\cdot)$ by the $\max_{a'\in\mu}$ operation. The corresponding safety Bellman operator $\mathcal{T}_\pi^\phi : \mathcal{Q} \to \mathcal{Q}$ is defined as:

$$\mathcal{T}_\pi^\phi[Q_\pi^\phi](s,a,d) = (1-\gamma)g(s) + \gamma \min\left\{g(s), \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mu(s')} \min_{d'\in\nu(s',a')} Q_\pi^\phi(s',a',d')\right]\right\} \tag{5}$$

Proposition 1 (Contraction Mapping of $\mathcal{T}_\pi^\phi$)**.** *For any arbitrary fixed joint policy π, the time-discounted Bellman operator defined in Eq. 5 induces a γ-contraction mapping under the supremum norm. That is, for a given discount factor $\gamma \in [0,1)$, and any $Q_\pi^\phi, \widetilde{Q}_\pi^\phi \in \mathcal{Q}$, we have $\left\|\mathcal{T}_\pi^\phi\left[Q_\pi^\phi\right] - \mathcal{T}_\pi^\phi\left[\widetilde{Q}_\pi^\phi\right]\right\|_\infty \leq \gamma \left\|Q_\pi^\phi - \widetilde{Q}_\pi^\phi\right\|_\infty$. (Proof in Appendix B.1)*

Proof. It will suffice to show for all tuples $(s,a,d) \in \mathcal{S}\times\mathcal{A}\times\mathcal{D}$,

$\left|\mathcal{T}_\pi^\phi\left[Q_\pi^\phi\right](s,a,d) - \mathcal{T}_\pi^\phi[\widetilde{Q}_\pi^\phi](s,a,d)\right| \leq \gamma \left\|Q_\pi^\phi - \widetilde{Q}_\pi^\phi\right\|_\infty$. We have:

$$\begin{aligned}
\left|\mathcal{T}_\pi^\phi\left[Q_\pi^\phi\right](s,a,d) - \mathcal{T}_\pi^\phi[\widetilde{Q}_\pi^\phi](s,a,d)\right| &= \Big|(1-\gamma)g(s) \\
&\quad + \gamma \min\left\{g(s), \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mu(s')}\min_{d'\in\nu(s',a')} Q_\pi^\phi(s',a',d')\right]\right\} - (1-\gamma)g(s) \\
&\quad - \gamma \min\left\{g(s), \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mu(s')}\min_{d'\in\nu(s',a')} \widetilde{Q}_\pi^\phi(s',a',d')\right]\right\}\Big| \\
&= \Big|\gamma \min\left\{g(s), \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mu(s')}\min_{d'\in\nu(s',a')} Q_\pi^\phi(s',a',d')\right]\right\} \\
&\quad - \gamma \min\left\{g(s), \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mu(s')}\min_{d'\in\nu(s',a')} \widetilde{Q}_\pi^\phi(s',a',d')\right]\right\}\Big|
\end{aligned}$$

Observing that$|\min\{b,a\} - \min\{c,a\}| \leq |b-c|\ \forall a,b,c\in\mathbb{R}$,

$$\begin{aligned}
&\left|\mathcal{T}_\pi^\phi\left[Q_\pi^\phi\right](s,a,d) - \mathcal{T}_\pi^\phi[\widetilde{Q}_\pi^\phi](s,a,d)\right| \\
&\quad \leq \gamma \left|\mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mu(s')}\min_{d'\in\nu(s',a')} Q_\pi^\phi(s',a',d') - \max_{a'\in\mu(s')}\min_{d'\in\nu(s',a')} \widetilde{Q}_\pi^\phi(s',a',d')\right]\right| \\
&\quad \text{(Applying Jensen's inequality)} \\
&\quad \leq \gamma \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\left|\max_{a'\in\mu(s')}\min_{d'\in\nu(s',a')} Q_\pi^\phi(s',a',d') - \max_{a'\in\mu(s')}\min_{d'\in\nu(s',a')} \widetilde{Q}_\pi^\phi(s',a',d')\right|\right]
\end{aligned}$$

Without loss of generality, let the first min-max term on the right-hand side be the larger one, and denote by a^* a maximizer of this term. Then, we have:

$$\left|\mathcal{T}_\pi^\phi\left[Q_\pi^\phi\right](s,a,d) - \mathcal{T}_\pi^\phi[\widetilde{Q}_\pi^\phi](s,a,d)\right| \leq \gamma \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\left|\min_{d'\in\nu(s',a')} Q_\pi^\phi(s',a^*,d') - \min_{d'\in\nu(s',a')} \widetilde{Q}_\pi^\phi(s',a^*,d')\right|\right]$$

Similarly, let d^* be a minimizer of the second, smaller term.

$$\begin{aligned}
\left|\mathcal{T}_\pi^\phi\left[Q_\pi^\phi\right](s,a,d) - \mathcal{T}_\pi^\phi[\widetilde{Q}_\pi^\phi](s,a,d)\right| &\leq \gamma \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\left|Q_\pi^\phi(s',a^*,d^*) - \widetilde{Q}_\pi^\phi(s',a^*,d^*)\right|\right] \\
&\leq \gamma \sup_{s'}\max_{a'\in\mathcal{A}}\max_{d'\in\mathcal{D}}\left|Q_\pi^\phi(s',a',d') - \widetilde{Q}_\pi^\phi(s',a',d')\right| \\
&= \gamma \left\|Q_\pi^\phi - \widetilde{Q}_\pi^\phi\right\|_\infty
\end{aligned}$$

□

B.2 Joint Action-Value Function (Sec. 3.3)

For any fixed safety action-value function, $Y \in \mathcal{Q}$, the joint Bellman operator $\mathcal{T}_Y : \mathcal{Q} \to \mathcal{Q}$ is given by:

$$\mathcal{T}_Y[Q](s,a,d) = (1-\gamma)m(s,a,d) + \gamma \min\left\{r(s,a,d) + \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[V_Q(s')\right], Y(s,a,d)\right\} \quad (6)$$

where $m(s,a,d) = \min\{r(s,a,d), g(s)\}$ and $V_Q(s) = \max_{a\in\mathcal{A}}\min_{d\in\mathcal{D}} Q(s,a,d)$.

Proposition 2 (Contraction Mapping of $\mathcal{T}_Y$)**.** *For any arbitrary, fixed $Y \in \mathcal{Q}$, the Bellman operator defined in Eq. 6 induces a γ-contraction mapping under the supremum norm. That is, for a given discount factor $\gamma \in [0,1)$, and any $Q, \widetilde{Q} \in \mathcal{Q}$, we have $\left\|\mathcal{T}_Y[Q] - \mathcal{T}_Y[\widetilde{Q}]\right\|_\infty \leq \gamma \left\|Q - \widetilde{Q}\right\|_\infty$. (Proof in Appendix B.2)*

Proof. It will suffice to show for all tuples $(s,a,d) \in \mathcal{S}\times\mathcal{A}\times\mathcal{D}$, $\left|\mathcal{T}_Y\left[Q\right](s,a,d) - \mathcal{T}_Y[\widetilde{Q}](s,a,d)\right| \leq \gamma \left\|Q-\widetilde{Q}\right\|_\infty$. We have:

$$\begin{aligned}
\left|\mathcal{T}_Y\left[Q\right](s,a,d) - \mathcal{T}_Y[\widetilde{Q}](s,a,d)\right| &= \Big|(1-\gamma)m(s,a,d) \\
&+\gamma\min\Big\{r(s,a,d)+\mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} Q(s',a',d')\right], Y(s,a,d)\Big\} - (1-\gamma)m(s,a,d) \\
&-\gamma\min\Big\{r(s,a,d)+\mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} \widetilde{Q}(s',a',d')\right], Y(s,a,d)\Big\}\Big| \\
&= \Big|\gamma\min\Big\{r(s,a,d)+\mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} Q(s',a',d')\right], Y(s,a,d)\Big\} \\
&-\gamma\min\Big\{r(s,a,d)+\mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} \widetilde{Q}(s',a',d')\right], Y(s,a,d)\Big\}\Big|
\end{aligned}$$

Observing that $|\min\{b,a\} - \min\{c,a\}| \leq |b-c|\ \forall a,b,c\in\mathbb{R}$,

$$\begin{aligned}
\left|\mathcal{T}_Y\left[Q\right](s,a,d) - \mathcal{T}_Y[\widetilde{Q}](s,a,d)\right| &\leq \gamma\Big|r(s,a,d)+\mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} Q(s',a',d')\right] \\
&\quad - r(s,a,d) - \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} \widetilde{Q}(s',a',d')\right]\Big| \\
&= \gamma\left|\mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} Q(s',a',d')\right] - \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} \widetilde{Q}(s',a',d')\right]\right| \\
&= \gamma\left|\mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} Q(s',a',d') - \max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} \widetilde{Q}(s',a',d')\right]\right| \\
&\text{(Applying Jensen's inequality)} \\
&\leq \gamma \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\left|\max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} Q(s',a',d') - \max_{a'\in\mathcal{A}}\min_{d'\in\mathcal{D}} \widetilde{Q}(s',a',d')\right|\right]
\end{aligned}$$

Without loss of generality, let the first min-max term on the right-hand side be the larger one, and denote by a^* a maximizer of this term. Then, we have:

$$\left|\mathcal{T}_Y\left[Q\right](s,a,d) - \mathcal{T}_Y[\widetilde{Q}](s,a,d)\right| \leq \gamma \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\left|\min_{d'\in\mathcal{D}} Q(s',a^*,d') - \min_{d'\in\mathcal{D}} \widetilde{Q}(s',a^*,d')\right|\right]$$

Similarly, let d^* be a minimizer of the second, smaller term.

$$\begin{aligned}
\left|\mathcal{T}_Y\left[Q\right](s,a,d) - \mathcal{T}_Y[\widetilde{Q}](s,a,d)\right| &\leq \gamma \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\left|Q(s',a^*,d^*) - \widetilde{Q}(s',a^*,d^*)\right|\right] \\
&\leq \gamma \sup_{s'}\max_{a'\in\mathcal{A}}\max_{d'\in\mathcal{D}}\left|Q(s',a',d') - \widetilde{Q}(s',a',d')\right| = \gamma\left\|Q-\widetilde{Q}\right\|_\infty
\end{aligned}$$

□

B.3 Objective Function of Joint Action-Value Function (Sec. 3.4)

In this section, we derive the objective function from robust optimal stopping theory to provide a principled theoretical interpretation of the joint value function. To this end, we first prove supporting Lemmas.

Lemma 2. *(Global upper bound and non-negativity of Q_π^ϕ under Joint Trajectory Evaluation) For all $(s,a,d)\in\mathcal{S}\times\mathcal{A}\times\mathcal{D}$ and all greedy joint policies π, $Q_\pi^\phi(s,a,d) \leq g(s)$ independent of the policy evaluated. Consequently, for all states within the failure set, $s\in\mathcal{F}$, $Q^{safe}_{\pi^a,\pi^d}(s,a,d) < 0$, where the failure set is defined as the set $\mathcal{F} = \{s\in\mathcal{S} : g(s) < 0\}$.*

Proof. By definition,

$$Q_\pi^\phi(s,a,d) = (1-\gamma)g(s) + \gamma \min\left\{g(s), \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a,d)}\left[\max_{a'\in\mu(s')}\min_{d'\in\nu(s',a')} Q_\pi^\phi(s',a',d')\right]\right\}$$

Noting that $\min\{x,y\} \leq x$, $Q_\pi^\phi(s,a,d) \leq (1-\gamma)g(s) + \gamma g(s) = g(s)$. For all states $s \in \mathcal{F}$, by definition, $g(s) < 0$. Furthermore, applying our previous result gives us $Q_\pi^\phi(s,a,d) < 0$. Finally, noting that $g(s)$ is independent of π, completes the proof. □

Remark 1. *As seen in Lemma 2, the upper bound of Q_π^ϕ relies solely on the minimum operator in the safety Bellman equation. Therefore, it holds universally for any arbitrary policy and does not require that the policy be greedy with respect to safety Q-function.*

Remark 2. *A positive immediate margin ($g(s) > 0$) does not guarantee viability. In inevitable failure states (unsafe states) where no safe policy exists, the long-term safety value correctly captures this future inevitable failure by evaluating to $Q_\pi^\phi(s,a,d) < 0$. Note that Lemma 2 only provides a global upper bound and does not impose a lower bound, allowing $Q_\pi^\phi(s,a,d)$ to reflect the unsafe states.*

Definition 1 (Task and safety optimizing regions)**.** *The region where the task value function is strictly lower than the safe value function is defined as the task-optimizing region: $\Omega^\psi = \left\{s \in \mathcal{S} \;\middle|\; V_q^\psi(s) < V_\pi^\phi(s)\right\}$. The region where the task value function is greater than the safe value function is defined as the safety-optimizing region: $\Omega^\phi = \left\{s \in \mathcal{S} \;\middle|\; V_q^\psi(s) \geq V_\pi^\phi(s)\right\}$.*

Lemma 3. *(First Switching time as valid stopping time) Let $(\mathcal{W}, \mathcal{G}, \mathbb{P})$ be a probability space and let $(s_k)_{k=0}^T$ be a stochastic process taking values in a measurable state space $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$. Let $\mathbb{F} = (\mathcal{G}_k)_{k=0}^T$ be the natural filtration where $\mathcal{G}_k = \sigma(s_0, \ldots, s_k)$. Given a Borel-measurable safety maximizing region, $\Omega^\phi \in \mathcal{B}(\mathcal{S})$, define the first switching time $\tau : \mathcal{W} \to \{0, \cdots, T\}$ as:*

$$\tau = \inf\{k \in \{0, \cdots, T\} : s_k \in \Omega^\phi\}$$

where T is the time horizon ($T \to \infty$ for infinite horizon setting) and $\inf \emptyset = T$. Then the first switching time, τ, is a stopping time with respect to $\mathbb{F}$.

Proof. τ is a stopping time if and only if for every $k \in \{0, \cdots, T\}$, the set of $\{\tau \leq k\} \in \mathcal{G}_k$. We start by observing that the event $\{\tau \leq k\}$ is equivalent to union of all events where the trajectory switched the first time from task optimizing region to safety maximizing region, $\{\tau \leq k\} = \bigcup_{j=0}^k \{s_j \in \Omega^\phi\}$. Noting that the state s_j is $\mathcal{G}_j$-measurable by definition of natural filtration, $\{s_j \in \Omega^\phi\} \in \mathcal{G}_j$ because Ω^ϕ is a Borel set in $\mathcal{S}$. Moreover, we have $\{s_j \in \Omega^\phi\} \in \mathcal{G}_k$ due to monotonicity of filtration (i.e., $\mathcal{G}_j \subseteq \mathcal{G}_k \; \forall j \leq k$). Finally, because σ-algebras are closed under finite union, we get $\{\tau \leq k\} \in \mathcal{G}_k$. This concludes the proof that τ is a valid stopping time. □

Remark 3. *(Reduction of the Joint Action-Value Function to Only Task Action-Value Function Prior to Switching)*
Prior to the first switching time τ, the trajectory remains entirely within the task optimizing region Ω^ψ. In this region, by definition, $V_q^\psi(s) = r_\pi(s) + \mathbb{E}[V_q(s')] < V_\pi^\phi(s)$. Furthermore, recalling from Lemma 2 that the safety value is globally bounded by the immediate margin, and the fact that, since the successor state remains in the domain $V_q(s') \geq 0$, we have the strict ordering $r_\pi(s) \leq V_\pi^\psi(s) < V_\pi^\phi(s) \leq g(s)$. This implies, $r(s,a) < g(s)$. Consequently, $V_q(s) = (1-\gamma)r_\pi(s) + \gamma V_q^\psi(s) = r_\pi(s) + \gamma\mathbb{E}[V_q(s')]$.This ensures that before entering the safety-maximizing region, the joint formulation evaluates only the task objective while remaining implicitly upper-bounded by $g(s)$.

Let the cumulative return from time $t = 0$, evaluated at a stopping time τ be given by $\mathcal{R}_\tau := \sum_{t=0}^{\tau-1} \gamma^t r(s_t) + \gamma^\tau \Big((1-\gamma)\min\{r(s_\tau), g(s_\tau)\} + \gamma V_Y^\phi(s_\tau)\Big)$, where $V_Y^\phi(s_\tau) = Y(s_\tau, \mu(s), \nu(s, \mu(s)))$. For a process observed from the next state s_1 (which we will *alias* as s'_0), let τ' denote a stopping time. Then the cumulative return from $t = 1$ is defined as : $\mathcal{R}'_\tau := \sum_{t=0}^{\tau'-1} \gamma^t r(s'_t) + \gamma^{\tau'} \Big((1-\gamma)\min\{r(s'_{\tau'}), g(s'_{\tau'})\} + \gamma V_Y^\phi(s'_{\tau'})\Big)$. Then, consider the following lemma:

Lemma 4. *Let Λ_k be the set of stopping times τ bounded by time horizon k, i.e., $0 \leq \tau \leq k$. Let $\mathcal{R}_\tau = (\mathcal{R}_t)_{t \leq \tau}$ be the stochastic process representing the cumulative discounted reward. Let $V_{Q_k}(s_1)$ be the value function evaluated at the next state s_1, defined as the optimal stopping problem $V_{Q_k}(s_1) := \inf_{\tau' \in \Lambda_k} \mathbb{E}_\omega [\mathcal{R}'_{\tau'} | s'_0 = s_1]$. Then the following equality holds, evaluated from $s_0 = s$:*

$$r(s) + \gamma \mathop{\mathbb{E}}_{\omega(s_1)} [V_{Q_k}(s_1) | s_0 = s] = \inf_{\tau \in \Lambda_{k+1},\, \tau \geq 1} \mathop{\mathbb{E}}_{\omega} [\mathcal{R}_\tau | s_0 = s]$$

where $\omega(s_1)$ denotes the s_1 element of the trajectory ω.

Proof. We will prove this by showing inequality in both directions. We first prove:

$$r(s) + \gamma \mathop{\mathbb{E}}_{\omega(s_1)} [V_{Q_k}(s_1) | s_0 = s] \geq \inf_{\tau \in \Lambda_{k+1},\, \tau \geq 1} \mathop{\mathbb{E}}_{\omega} [\mathcal{R}_\tau | s_0 = s]$$

To prove the above inequality, we will construct a global stopping time by *pasting together* ϵ-optimal local stopping times.

1. Fix an arbitrarily small $\epsilon > 0$. For every realization z of the state s_1, there exists a valid stopping time $\tau'_z \in \Lambda_k$ such that its expected value is ϵ-close to the optimal value $V_{Q_k}(z)$, i.e., $\mathbb{E}_\omega[\mathcal{R}'_{\tau'_z} | s'_0 = z] \leq V_{Q_k}(z) + \epsilon$.

2. Define a global candidate stopping time $\tau^* \in \Lambda_{k+1}$ by waiting one step and then *executing* τ'_z, i.e., $\tau^* := 1 + \tau'_z$. By the recursive definition of cumulative return, $\mathcal{R}_{\tau^*} = r(s) + \gamma \mathcal{R}'_{\tau'_z}$.

3. We now invoke Lemma 2.8 in Aguilar (2009) (Stability under Pasting, from optimal robust stopping time theory), which states that the set of stopping times is closed under such concatenation. Since τ'_z depends on the history after $t = 1$, the constructed variable τ^* is non-anticipative and is a valid stopping time in Λ_{k+1} with $\tau^* \geq 1$.

4. The global infimum over Λ_{k+1} must be less than or equal to the expectation of the specific candidate:

$$\inf_{\tau \in \Lambda_{k+1},\, \tau \geq 1} \mathop{\mathbb{E}}_{\omega} [\mathcal{R}_\tau | s_0 = s] \leq \mathop{\mathbb{E}}_{\omega} [\mathcal{R}_{\tau^*} | s_0 = s]$$

5. Finally, applying the Law of Iterated Expectations and our ϵ-bound yields:

$$\begin{aligned}
\inf_{\tau \in \Lambda_{k+1},\, \tau \geq 1} \mathop{\mathbb{E}}_{\omega} [\mathcal{R}_\tau | s_0 = s] &\leq \mathop{\mathbb{E}}_{\omega(s_1)} \left[r(s) + \gamma \mathbb{E}_\omega [\mathcal{R}'_{\tau'_{s_1}} | s'_0 = s_1] \right] \\
&\leq \mathop{\mathbb{E}}_{\omega(s_1)} [r(s) + \gamma (V_{Q_k}(s_1) + \epsilon)] \\
&= r(s) + \gamma \mathop{\mathbb{E}}_{\omega(s_1)} [V_{Q_k}(s_1)] + \gamma \epsilon
\end{aligned}$$

Since this inequality holds for any arbitrary small $\epsilon > 0$, letting $\epsilon \to 0$ yields the first direction.

Next, we prove the opposite direction:

$$r(s) + \gamma \mathop{\mathbb{E}}_{\omega(s_1)} [V_{Q_k}(s_1) | s_0 = s] \leq \inf_{\tau \in \Lambda_{k+1},\, \tau \geq 1} \mathop{\mathbb{E}}_{\omega} [\mathcal{R}_\tau | s_0 = s]$$

1. Let $\tau \in \Lambda_{k+1}$ be any arbitrary global stopping time with $\tau \geq 1$.

2. Conditioning on the first transition $s_1 = z$, the remainder of the trajectory induces a local stopping time $\tau'_z = \tau - 1$. The induced stopping time satisfies $0 \leq \tau'_z \leq k$, and thus, $\tau'_z \in \Lambda_k$.

3. By the definition of value function as an infimum over Λ_k, we have $\mathbb{E}_\omega[\mathcal{R}'_{\tau'_z}|s'_0 = z] \geq V_{Q_k}(z)$.

4. . Using Law of Iterated Expectations,

$$\begin{aligned}\mathop{\mathbb{E}}_{\omega}[\mathcal{R}_\tau|s_0 = s] &= r(s) + \gamma \mathop{\mathbb{E}}_{\omega(s_1)}\left[\mathbb{E}_\omega\left[\mathcal{R}'_{\tau'_{s_1}}\middle|s'_0 = s_1\right]\right] \\ &\geq r(s) + \gamma \mathop{\mathbb{E}}_{\omega(s_1)}[V_{Q_k}(s_1)]\end{aligned}$$

5. Finally, since this inequality holds for any arbitrary $\tau \in \Lambda_{k+1}$, $(\tau \geq 1)$, it must hold for the infimum over all such τ, completing the proof.

□

Remark 4. *(On Non-Anticipativity) The invocation of Lemma 2.8 in Aguilar (2009) guarantees that our constructed variable τ^* uses only current information, since the specific local stopping time τ'_z is solely determined by the realization of state s_1 observed at time $t = 1$. Lemma 4 justifies the interchange of the infimum and expectation operators. Intuitively, it proves that the optimal value of the global stopping time is equivalent to the expectation of local optimal values of the local stopping times.*

Theorem 7. *(Objective Function for a fixed trajectory) Let $\omega = (s_0, s_1, \cdots, s_T)$ be a trajectory generated by action policy, $\mu(s)$, and disturbance policy $\nu(s, \mu(s))$, with associated rewards $r(s_t)$. Here, T denotes the time horizon. Then the objective function that the joint action-value function optimizes is given by:*

$$J(s) := \inf_{\tau \in \Lambda_T} \mathop{\mathbb{E}}_{\omega}\left[\sum_{t=0}^{\tau-1} \gamma^t r(s_t) + \gamma^\tau \left((1-\gamma)\min\{r(s_\tau), g(s_\tau)\} + \gamma V_Y^\phi(s_\tau)\right)\middle| s_0 = s\right]$$

Proof. We begin by observing that the first switching time τ from target optimizing region to safety maximizing region is a valid stopping time (Refer to Lemma 3). We prove the theorem by induction on the time horizon T.

1. **Base Case ($T = 1$):** Consider a time horizon $T = 1$. The set of valid stopping times is given by $\Lambda_1 = \{0, 1\}$. When $\tau = 0$, the initial state $s_0 \in \Omega^\phi$ and the trajectory is terminated immediately, yielding the return $\mathcal{R}_0 = (1-\gamma)\min\{r(s_0), g(s_0)\} + \gamma V_Y^\phi(s_0)$. When $\tau = 1$, the agent takes one step and is stopped and thus, the next state $s_1 \in \Omega^\phi$, with $s_0 \in \Omega^\psi$. This yields the expected return $\mathbb{E}[\mathcal{R}_1] = r(s_0) + \gamma \mathbb{E}_{\omega(s_1)}[(1-\gamma)\min\{r(s_1), g(s_1)\} + \gamma V_Y^\phi(s_1)|s_0 = s]$. As the value function is the infimum over Λ_1, it evaluates to the minimum of these two choices:

$$V_{Q_1}(s) = \min\Bigg\{(1-\gamma)\min\{r(s_0), g(s_0)\} + \gamma V_Y^\phi(s_0), \\ r(s_0) + \gamma \mathbb{E}_{\omega(s_1)}[(1-\gamma)\min\{r(s_1), g(s_1)\} + \gamma V_Y^\phi(s_1)|s_0 = s]\Bigg\}$$

   Since the minimum is equivalent to the infimum over the set of valid stopping times $\tau \in \Lambda_1$, we have $V_{Q_1}(s) = \inf_{\tau \in \Lambda_1} \mathbb{E}_\omega[\mathcal{R}_\tau|s_0 = s]$.

2. **Induction Hypothesis ($T = k$):** Assume for a time horizon $T = k$, the value function is given by:
$$V_{Q_k}(s) = \inf_{\tau \in \Lambda_k} \mathop{\mathbb{E}}_{\omega}[\mathcal{R}_\tau|s_0 = s]$$

3. **Inductive Step ($T = k + 1$):** Consider the time horizon $T = k + 1$. According to the Bellman equation, the value at the current state is the minimum of stopping immediately

($\tau = 0$) or continuing for one step and behaving optimally thereafter (which is horizon k starting at $t = 1$):

$$V_{Q_{k+1}}(s) = \min\Bigg\{(1-\gamma)\min\{r(s_0), g(s_0)\} + \gamma V_Y^{\phi}(s_0), \\ r(s_0) + \gamma \mathbb{E}_{\omega(s_1)}[V_{Q_k}(s_1)|s_0 = s]\Bigg\}$$

Applying Lemma 4 to the above yields:

$$V_{Q_{k+1}}(s) = \min\left\{(1-\gamma)\min\{r(s_0), g(s_0)\} + \gamma V_Y^{\phi}(s_0), \inf_{\tau\in\Lambda_{k+1},\, \tau\geq 1} \mathop{\mathbb{E}}_{\omega}[\mathcal{R}_\tau|s_0 = s]\right\}$$

Finally, observing that the first term is exactly the return $\mathcal{R}_0$ corresponding to the case where $\tau = 0$, the outer minimum operator simplifies into the infimum over the strict subset $\tau \geq 1$, yielding an infimum over the entire domain $0 \leq \tau \leq k+1$:

$$\begin{aligned} V_{Q_{k+1}}(s) &= \inf_{\tau\in\Lambda_{k+1}} \mathbb{E}_{\omega}\left[\mathcal{R}_\tau|s_0 = s\right] \\ &= \inf_{\tau\in\Lambda_T} \mathop{\mathbb{E}}_{\omega}\left[\sum_{t=0}^{\tau-1}\gamma^t r(s_t) + \gamma^\tau\left((1-\gamma)\min\{r(s_\tau), g(s_\tau)\} + \gamma V_Y^{\phi}(s_\tau)\right)\middle| s_0 = s\right] \end{aligned}$$

□

Theorem 1 (Objective Function)**.** *Given a Borel-measurable safety maximizing region, $\Omega^\phi \in \mathcal{B}(\mathcal{S})$, define the first switching time $\tau : \mathcal{W} \to \{0, \cdots, T\}$ as $\tau = \inf\{k \in \{0, \cdots, T\} : s_k \in \Omega^\phi\}$, where T is the time horizon and $\inf\emptyset = T$ ($T \to \infty$ for infinite horizon setting). Let Λ_k be the set of stopping times τ bounded by time horizon k, i.e., $0 \leq \tau \leq k$. Let ω be trajectories sampled using the joint policy as $a_t, d_t \sim \pi(s_t)$, with the dynamics $s_{t+1} \sim p(\cdot|s_t, a_t, d_t)$. Then, the corresponding joint objective function that the joint policy maximizes is given by: (Proof in Appendix B.3)*

$$J(\pi) := \inf_{\tau\in\Lambda_T} \mathbb{E}_{\omega\sim\pi}\left[\sum_{t=0}^{\tau-1}\gamma^t r(s_t, a_t, d_t) + \gamma^\tau\left(\underbrace{(1-\gamma)\min\{r(s_\tau, a_\tau, d_\tau), g(s_\tau)\} + \gamma V_Y^{\phi}(s_\tau)}_{:=J_\tau}\right)\middle| s_0 = s\right] \quad (7)$$

Proof. This follows directly from Theorem 7. Without loss of generality, the Theorem used state-based rewards $r(s_t)$ in the main proof for notational simplicity. The extension to state-action-based rewards $r(s_t, a_t, d_t)$ holds because the expectation is taken over the distribution of trajectories. □

Remark 5 (Boundary Condition Without Joint Actions)**.** *When the stopping time τ occurs such that no subsequent action is executed at the terminal state s_τ, the state-action-disturbance reward $r(s_\tau, a_\tau, d_\tau)$ is not accrued. Consequently, the terminal condition collapses entirely to the safety value function, evaluating to $(1-\gamma)g(s_\tau) + \gamma V_Y^{\phi}(s_\tau)$, with no contribution from the task reward.*

B.4 Fast Timescale: Safety Critic Tracking (Sec. 4.1)

Assumption 1 (Conditional unbiasedness and Martingale-difference noise (Tsitsiklis & Van Roy, 1997))**.** *Let $\{\mathcal{H}_n\}_{n\geq 0}$ denote the filtration generated by the learning history up to iteration n. Then,*

$$\mathbb{E}\left[\hat{\mathcal{T}}^{\phi}_{n+1,\pi_{Q_n}}[Y_n]\middle|\mathcal{H}_n\right] = \mathcal{T}^{\phi}_{\pi_{Q_n}}[Y_n], \quad \mathbb{E}\left[\hat{\mathcal{T}}_{n+1,Y_n}[Q_n]\middle|\mathcal{H}_n\right] = \mathcal{T}_{Y_n}[Q_n] \quad (9)$$

Assumption 2 (Step-size conditions (Robbins & Monro, 1951))**.** *The step-sizes, $\alpha_n, \beta_n \in (0, 1]$, satisfy Robbins-Monro conditions: $\sum_{n=0}^{\infty}\alpha_n = \sum_{n=0}^{\infty}\beta_n = \infty$, $\sum_{n=0}^{\infty}(\alpha_n^2 + \beta_n^2) < \infty$, $\lim_{n\to\infty}\alpha_n \to 0$, $\lim_{n\to\infty}\beta_n \to 0$, and additionally, $\lim_{n\to\infty}\alpha_n/\beta_n = 0$.*

The updates with step-size sequences $\{\alpha_n\}, \{\beta_n\}$ are given by:

$$Y_{n+1} = Y_n + \beta_n \left[\hat{\mathcal{T}}^{\phi}_{n+1,\pi_{Q_n}}[Y_n] - Y_n + \xi^{\phi}_{n+1} \right] \tag{28}$$

$$Q_{n+1} = Q_n + \alpha_n \left[\hat{\mathcal{T}}_{n+1,Y_n}[Q_n] - Q_n + \xi_{n+1} \right] \tag{29}$$

where ξ^{ϕ}_{n+1} and ξ_{n+1} are martingale difference perturbations (Borkar, 2009), with $\mathbb{E}[\xi^{\phi}_{n+1}|\mathcal{H}_n] = \mathbb{E}[\xi_{n+1}|\mathcal{H}_n] = 0$. As the map $Q \mapsto Y_{\pi_Q}$ need not be continuous or Lipschitz, the standard two-timescale stochastic approximation need not apply. Thus, we analyze the fast timescale recursion as a stochastic recursive inclusion in Section 4.1 using the *set-valued two-timescale framework* (Yaji & Bhatnagar, 2020; Borkar, 2023).

Lemma 1 (Set-valued fast timescale tracking)**.** *Under Assumptions 1-2, suppose the iterates are almost surely bounded. Then,*

1. *for every quasi-static q, Eq. 11 admits a nonempty compact global attracting set $\Lambda(q)$;*
2. *$q \mapsto \Lambda(q)$ is upper semicontinuous; and*
3. *the coupled iterates satisfy $\operatorname{dist}\left((Y_n, Q_n), \operatorname{Gr}\Lambda\right) \to 0$ almost surely, where $\operatorname{Gr}\Lambda := \{(y,q) : y \in \Lambda(q)\}$ is the graph of Λ. (Proof in Appendix B.4)*

Proof. We first establish some properties of the fast timescale differential inclusion.

- Let $y \in \mathcal{Y} \subseteq \mathbb{R}^{|\mathcal{S}||\mathcal{A}||\mathcal{D}|}$ denote the fast timescale safety critic with the limiting dynamics $h_\pi(y) := \mathcal{T}^{\phi}_{\pi}[y] - y$. Here, π is quasi-static, as induced by the quasi-static joint critic $q \in \mathcal{Q}$. Thus, note that for any $y \in \mathcal{Y}$, the operator $\mathcal{T}^{\phi}_{\pi}[y]$ induces a γ-contraction, as discussed in Section 3.2, and thus possesses a unique fixed point.

- Next, note that, since $\mathcal{S}, \mathcal{A}, \mathcal{D}$ are finite, $\Pi(q)$ is finite and nonempty for every q. For a given q, any action that is not maximin-greedy at q has a strictly positive gap from the maximum, and any disturbance that is not minimizing on some greedy action has a strictly positive gap from the minimum. Consequently, there exists $\vartheta(q) > 0$ such that $\|q' - q\|_\infty < \vartheta(q)$, which implies that $\Pi(q') \subseteq \Pi(q)$. Hence, $\Pi(\cdot)$ is upper semicontinuous. Moreover, by continuity of $h_\pi(\cdot)$, the map $\Psi(\cdot,\cdot)$ (as defined in Eq. 11) is Marchaud, since it has nonempty compact convex values, and is upper semicontinuous with at most linear growth.

- Let $\overline{g}_K := K \max_{s\in\mathcal{S}} |g(s)|$ with $B := [-\overline{g}_K, \overline{g}_K]^{|\mathcal{S}||\mathcal{A}||\mathcal{D}|}$. Note that if $y \in B$, then $-\overline{g}_K \leq \mathbb{E}\left[y(s, \pi(s)\right] \leq \overline{g}_K$. Hence, $\mathcal{T}^{\phi}_{\pi}[B] \subseteq B$, $\forall \pi$. By convexity, B is forward invariant for the differential inclusion $\dot{y} \in \Psi(q, y)$. Moreover, any solution to the differential inclusion can be represented almost everywhere as:

$$\dot{y}(t) = \sum_{\pi \in \Pi(q)} \lambda_\pi(t) \left(\mathcal{T}^{\phi}_{\pi}[y(t)] - y(t) \right), \quad \text{with } \lambda_\pi(t) \geq 0, \sum_{\pi} \lambda_\pi(t) = 1 \tag{30}$$

  Next, let $y(\cdot), z(\cdot)$ be two trajectories with the same weight $\lambda_\pi(\cdot)$, and let $e(t) := y(t) - z(t)$. Let $I(t) := \{i : |e_i(t)| = \|e(t)\|_\infty\}$ be the coordinate set on which $|e(t)|$ achieves the supremum norm. For the supremum norm, the upper Dini derivative satisfies:

$$\mathbb{D}^+ \|e(t)\|_\infty \leq \max_{i \in I(t)} \operatorname{sgn}(e_i(t)) \dot{e}(t) \tag{31a}$$

$$\implies \mathbb{D}^+ \|y(t) - z(t)\|_\infty \leq -(1-\gamma)\|y(t) - z(t)\|_\infty \text{ (by } \gamma\text{-contraction of } \mathcal{T}^{\phi}_{\pi}) \tag{31b}$$

$$\therefore \|y(t) - z(t)\|_\infty \leq e^{-(1-\gamma)(t-s)} \|y(s) - z(s)\|_\infty \text{ (by Grönwall)} \tag{31c}$$

  Thus, choosing $z(\cdot) \in B$ and the fact that $\mathcal{T}^{\phi}_{\pi}[B] \subseteq B$ shows that every fast timescale trajectory $y(\cdot)$ (driven by the same weights λ_π) for a quasi-static q is attracted to the bounded set B.

1. Since $\Psi(q, \cdot)$ is a Marchaud map, the differential inclusion for a quasi-static q admits solutions from every initial condition. As the compact set B is forward invariant, and globally attracting, the limiting dynamics $y(\cdot)$ admit a nonempty compact global attracting set $\Lambda(q) \subseteq B$ (Benaïm et al., 2005).

2. Let $q_m \to q$ and let $y_m \in \Lambda(q_m)$ with $y_m \to y$. Note that for sufficiently large m, $\Pi(q_m) \subseteq \Pi(q)$, and thus, $\Psi(q_m, \cdot) \subseteq \Psi(q, \cdot)$. Next, for any given y, $\{h_\pi(y) : \pi \in \Pi(q_m)\} \subseteq \{h_\pi(y) : \pi \in \Pi(q)\}$. Since the convex hull preserves set inclusions, $\Psi(q_m, y) \subseteq \Psi(q, y)$ for every y, and thus, every complete solution of q_m-differential inclusion is also a complete solution of the q-differential inclusion. Consequently, $\Lambda(q_m) \subseteq \Lambda(q)$ for sufficiently large m. Moreover, since $\Lambda(q)$ is closed, $y \in \Lambda(q)$, and thus $\mathrm{Gr}\Lambda$ is closed. This implies that the map $q \mapsto \Lambda(q)$ is upper semicontinuous.

3. Finally, under Assumptions 1-2 and by the two-timescale stochastic recursive inclusion framework (Yaji & Bhatnagar, 2020; Ramaswamy & Bhatnagar, 2016), the limit set of the coupled iterates satisfies $\omega(Y, Q) \subseteq \mathrm{Gr}\Lambda$ a.s. Since the iterates are bounded and $\mathrm{Gr}\Lambda$ is closed,

$$\mathrm{dist}\left((Y_n, Q_n), \mathrm{Gr}\Lambda\right) \to 0 \quad \text{a.s.}, \tag{32}$$

□

Remark 6. *(The Quasi-Static Joint Critic) Intuitively, Lemma 1 guarantees that the safety critic learns much faster than the joint critic. This implies that the joint critic is provided with an accurate estimate of its corresponding policy's safety evaluation before it updates its estimate and changes the policy.*

B.5 Slow Timescale: Averaged Differential Inclusion (Sec. 4.2)

Section 4.1 shows that the coupled iterates asymptotically approach the graph of the compact, globally attracting set $\Lambda(q)$ for some quasi-static q. By Assumption 2, the slow-timescale iterate Q_n remains quasi-static while the fast timescale dynamics evolve. Thus, the increments of Q_n depend on the *long-run average* of $G(Y_n, Q_n)$. These averages are encoded by the occupation measures in $\mathfrak{D}(q)$, leading to the set-valued $\overline{G}(q)$ in Eq. 12.

Theorem 2 (Slow-timescale differential-inclusion limit)**.** *Let $\overline{Q}(\cdot)$ denote the continuous affine interpolation of $\{Q_n\}$ on the slow-timescale. Under Assumptions 1-2, Lemma 1, and almost surely bounded iterates $\{(Y_n, Q_n)\}$, $\overline{Q}(\cdot)$ is almost surely an asymptotic pseudotrajectory of the averaged differential inclusion $\dot{q}(t) \in \overline{G}(q(t))$. Consequently, its ω-limit set $\mathcal{U} := \omega(\overline{Q})$ is a nonempty compact invariant internally chain-transitive set of the limiting differential-inclusion dynamics. (Proof in Appendix B.5)*

Proof. Let $M_{n+1} := \hat{\mathcal{T}}_{n+1,Y_n}[Q_n] - \mathcal{T}_{Y_n}[Q_n] + \xi_{n+1}$ be the total martingale-difference error (error due to the stochastic Bellman target and additional perturbation noise). Then, Eq. 29 can be written as:

$$Q_{n+1} := Q_n + \alpha_n \left[G(Y_n, Q_n) + M_{n+1}\right], \quad \text{where } G(y, q) := \mathcal{T}_y[q] - q \tag{33}$$

1. In the finite tabular setting, G is continuous in both arguments; $V_q(s) = \max_a \min_d q(s, a, d)$ is a 1-Lipschitz map in q, and $x \mapsto \min \cdot, x$ is 1-Lipschitz map. Consequently,

$$\|G(y, q) - G(y', q')\|_\infty \leq \gamma \|y - y'\|_\infty + (1 + \gamma)\|q - q'\|_\infty \tag{34}$$

   Thus, $G(\cdot, \cdot)$ is continuous and has at most linear growth.

2. Next, let $\mathfrak{D}(q)$ denote the admissible limiting occupation measure set, appearing in the two-timescale stochastic recursive inclusion theorem of (Yaji & Bhatnagar, 2020); these measures account for the long-run fast timescale behavior at a given q; the slow-timescale differential inclusion is then formed through probability-measure averaging (Yaji & Bhatnagar, 2020):

$$\overline{G}(q) = \left\{ \int_{\Lambda(q)} G(y, q) \mathbb{P}(\mathrm{d}y) : \mathbb{P} \in \mathfrak{D}(q) \right\} \tag{35}$$

3. Next, by Assumptions 1-2, the *accumulated noise* on the slow-timescale vanishes on every finite time-window:

$$\lim_{n\to\infty} \sup_{m\geq n:\sum_{k=n}^{m}\alpha_k\leq T} \left\| \sum_{k=n}^{m} \alpha_k M_{k+1} \right\|_{\infty} = 0 \quad \text{a.s.} \tag{36}$$

for every $T > 0$. Hence, by the slow-timescale APT result of (Yaji & Bhatnagar, 2020), the continuous affine interpolation $\overline{Q}(\cdot)$ is almost surely an asymptotic pseudotrajectory of $\dot{q}(t) \in \overline{G}(q(t))$. Finally, by assumption, the iterates are almost surely bounded, and thus, by the limit-set theorem for bounded asymptotic pseudotrajectories (Benaïm et al., 2005), $\mathcal{U} := \omega(\overline{\mathcal{Q}})$ is nonempty, compact, invariant and internally chain transitive for the differential inclusion $\dot{q} \in \overline{G}(q)$.

□

B.6 Safe Configuration Set and Forward Invariance (Sec. 5)

The state, action and disturbance spaces are finite, so the action-value space $\mathcal{Q} = \mathcal{B}(\mathcal{S} \times \mathcal{A} \times \mathcal{D})$ is identified with $\mathbb{R}^N, N = |\mathcal{S} \times \mathcal{A} \times \mathcal{D}|$, and the limiting dynamics are an inclusion in finite dimension. The extended class-$\mathcal{K}_e$ function is written explicitly as $Kg(s)$, since the margin scale affects the bounds. We write $\|\cdot\|_\infty$ for the supremum-norm. Recalling also that the reward function is $r : \mathcal{S} \times \mathcal{A} \times \mathcal{D} \to [0, R_{\max}]$ and $\Omega \subseteq \mathcal{F}^c = \{s : g(s) > 0\}$.

For a fixed critic $y \in \mathcal{Q}$, let $\mathcal{T}_y[q]$ denote the joint Bellman operator of Eq. 6, and for policy π let $\mathcal{T}_\pi^\phi[y]$ denote the safety Bellman operator of Eq. 5.

Definition 2 (Kernel-preserving actions)**.** *An action is kernel-preserving at state s, if it keeps the next state inside the viability kernel against every disturbance $\mathcal{A}^\Omega(s) := \{a \in \mathcal{A} : P(s' \in \Omega | s, a, d) = 1 \quad \forall\, d \in \mathcal{D}\}$, where $\kappa, \epsilon_0 > 0$ without loss of generality.*

Since Ω is the set of states from which a control policy exists keeping the agent safe at all times regardless of the adversarial disturbances, taking $a = \mu(s)$ for such a policy shows $\mathcal{A}^\Omega(s) \neq \emptyset\ \forall s \in \Omega$. This is a property of the viability kernel. Since the transition kernel is stochastic, the kernel preservation is *almost sure*, and the safety conclusions are almost surely with respect to the transition randomness.

Definition 3 (Index Sets and Relaxed Constraint Sets)**.** *For each state $s \in \Omega$ and each action $a \notin \mathcal{A}^\Omega(s)$, fix a disturbance $\hat{d}(s, a) \in \mathcal{D}$ with $P(s' \notin \Omega | s, a, \hat{d}(s, a)) > 0$. Then the index sets are defined as follows: $\mathcal{I} := \{(s, a, d) : s \in \Omega, a \in \mathcal{A}^\Omega(s), d \in \mathcal{D}\}$ and $\mathcal{J} := \{(s, a, \hat{d}(s, a)) : s \in \Omega, a \notin \mathcal{A}^\Omega(s)\}$. Let $\mathcal{C}_\epsilon := \{q \in \mathcal{Q} : q(x) \geq -\epsilon\ \forall\ x \in \mathcal{I}, q(x) \leq -\kappa + \epsilon\ \forall\ x \in \mathcal{J}\}$ denote the relaxed constraint set, for $\epsilon \geq 0$ i.e., $\mathcal{C}_\epsilon$ relaxes each defining inequality of $\mathcal{C}_0$ by ϵ.*

The index sets are determined by the transition kernel and by Ω; they depend on neither q nor any policy. Throughout we take $\kappa > 0$ and $\epsilon_0 \in (0, \kappa/2)$. The upper bound on ϵ_0 ensures the two families of constraints remain strictly separated on all of $\mathcal{C}_{\epsilon_0}$.

We fix a deterministic Borel-measurable selection in Eq. 4: we choose a single action $\mu_q(s)$ from the maximizers of $\min_d q(s, \cdot, d)$, and a single disturbance $\nu_q(s, \mu_q(s))$ from the minimizers of $q(s, \mu_q(s), \cdot)$. Existence of such selections is guaranteed by the Kuratowski-Ryll-Nardzewski theorem. Fixing the selection makes $q \mapsto \pi_q$ single-valued but not continuous; this is why fast timescale is analyzed through $\Pi(q)$ and $\Lambda(q)$ rather than through a single equilibrium map, and why the slow limit is the inclusion Eq. 37.

$$\dot{q}(t) \in \overline{G}(q(t)) \tag{37}$$

One might expect a condition of the form $\min_d q(s, a, d) \leq -\kappa$ for kernel-exiting a; but the disturbance attaining the inner minimum of q need not be the one under which the action leaves the kernel (the truncation bound in Lemma 10 is available only for the latter). Fixing $\hat{d}$ removes the mismatch while still yielding $\min_d q(s, a, d) \leq q(s, a, \hat{d}(s, a)) \leq -\kappa$.

Definition 4 (Safe Configuration Set)**.** *The safe configuration set is defined as the set of joint Q-functions whose induced policy is kernel-preserving:*

$$\mathcal{Q}_{cfg} := \mathcal{C}_0 = \{q \in \mathcal{Q} : q(x) \geq 0\ \forall\ x \in \mathcal{I}, q(x) \leq -\kappa\ \forall\ x \in \mathcal{J}\} \tag{13}$$

The intuition is as follows: the lower bound states that the joint action value has not *underestimated* kernel-preserving actions, and the upper bound states that it has not *overestimated* actions that leave the kernel. We now define the distance to the safe configuration set as $\rho(q) := \text{dist}(q, \mathcal{Q}_{\text{cfg}})$, the supremum-norm distance.

B.6.1 Geometry of the Safe Configuration Set

Lemma 5 (Structure of $\mathcal{C}_\epsilon$)**.** *(i) $\mathcal{I} \cap \mathcal{J} = \emptyset$ and $\mathcal{Q}_{cfg} \neq \emptyset$ (ii) Every $\mathcal{C}_\epsilon$ is a closed convex polyhedron. (iii) For every q,*

$$\rho(q) = \max\left\{ \max_{x\in\mathcal{I}}(-q(x))_+ \ , \ \max_{x\in\mathcal{J}}(q(x)+\kappa)_+ \right\}$$

and consequently $\mathcal{C}_\epsilon = \{q : \rho(q) \leq \epsilon\}$.

Proof. (i) A triple of $\mathcal{I}$ has $a \in \mathcal{A}^\Omega(s)$ and a triple of $\mathcal{J}$ has $a \notin \mathcal{A}^\Omega(s)$, so the index sets are disjoint. Next, set $q(x) = 0$ on $x \in \mathcal{I}$, $q(x) = -\kappa$ on $x \in \mathcal{J}$ and arbitrary values elsewhere, therefore this point lies in $\mathcal{Q}_{\text{cfg}}$.
(ii) Each condition is a closed half-space of the form $\{q : \pm q(x) \leq \beta\}$, and a finite intersection of half-spaces is a closed convex polyhedron.
(iii) Let $p \in \mathcal{Q}_{\text{cfg}}$. We first check if the distance $\|q-p\|_\infty$ is at least the asserted quantity. For $x \in \mathcal{I}$ with $q(x) < 0$ we have $p(x) \geq 0$, hence $\|q(x)-p(x)\|_\infty \geq -q(x)$; for $x \in \mathcal{J}$ with $q(x) > -\kappa$ we have $p(x) \leq -\kappa$, hence $\|q(x)-p(x)\|_\infty \geq q(x)+\kappa$. Thus $\|q-p\|_\infty$ is at least the asserted quantity. Conversely, the coordinatewise clip defined by $p(x) = \max(q(x), 0)$ on $x \in \mathcal{I}$, $p(x) = \min(q(x), -\kappa)$ on $x \in \mathcal{J}$ and $p(x) = q(x)$ on the remaining coordinates belongs to $\mathcal{Q}_{\text{cfg}}$ and attains it. Comparison with the defining inequalities of $\mathcal{C}_\epsilon$ gives the last claim. □

The intuition is that $\mathcal{Q}_{\text{cfg}}$ *is a box*: its constraints are coordinate-separated, so its metric projection is a clip and its distance function has a closed form. By Lemma 5(iii), the family $\mathcal{C}_\epsilon$ is precisely the family of enlargements of $\mathcal{Q}_{\text{cfg}}$.

Lemma 6 (Regularity of the distance)**.** *Under the supremum norm, for any convex $C \subseteq \mathbb{R}^N$, the function dist$(\cdot, C)$ is convex and 1-Lipschitz.*

Proof. For convexity: Given $q, q' \in \mathcal{Q}$, $\lambda \in [0,1]$ and $\eta > 0$, choose $p, p' \in C$ with $\|q-p\|_\infty \leq \text{dist}(q, C) + \eta$ and similarly for p'. By convexity of C, $\lambda p + (1-\lambda)p' \in C$. Then, applying to distance function, $\text{dist}(\lambda q + (1-\lambda)q', C) \leq \|(\lambda q + (1-\lambda)q') - (\lambda p + (1-\lambda)p')\|_\infty = \|\lambda(q-p) + (1-\lambda)(q'-p')\|_\infty \leq \lambda\text{dist}(q, C) + (1-\lambda)\text{dist}(q', C) + \eta$. Then taking $\eta \to 0$, gives us the claim of convexity.
For 1-Lipschitz: As $\text{dist}(q, C) \leq \|q-p\|_\infty \leq \|q-q'\|_\infty + \|q'-p\|_\infty \ \forall p \in C$, taking the infimum over p and exchanging q and q' yields $|\text{dist}(q, C) - \text{dist}(q', C)| \leq \|q-q'\|_\infty$. □

Lemma 7 (Convex distance decay)**.** *Let $C \subseteq \mathbb{R}^N$ be closed, convex and nonempty, $\rho_C := dist(\cdot, C)$ be the supremum-norm, $\theta \in [0,1)$, $R \in (0, \infty]$, and $\Phi : \mathbb{R}^N \to \mathbb{R}^N$ be a set-valued map. Suppose that for every x with $\rho_C(x) < R$, every $v \in \Phi(x)$ has the form $v = z - x$ with $\rho_C(z) \leq \theta\rho_C(x)$. Then every absolutely continuous $x(\cdot)$ with $x(t) \in \Phi(x(t))$ almost everywhere and $\rho_C(x(0)) < R$ satisfies*

$$\rho_C(x(t)) \leq \rho_C(x(0))e^{-(1-\theta)t} \quad \forall\, t \geq 0$$

Proof. Let $u(t) := \rho_C(x(t))$. Since ρ_C is 1-Lipschitz (Lemma 6), u is absolutely continuous and hence differentiable almost everywhere. Let $t^* := \sup\{\tau \geq 0 : u(t) < R \ \forall\, t \leq \tau\}$. Notice that as $u(0) < R$ and u is continuous, $t^* > 0$. Consider $t < t^*$ at which u is differentiable and fix $v := \dot{x}(t)$ such that $v = z - x(t)$ with $\rho_C(z) \leq \theta u(t)$. For $h \in (0,1)$, absolute continuity gives $x(t+h) = x(t) + hv + o(h)$. By the Lipschitz property, we get,

$$u(t+h) \leq \rho_C(x(t)+hv) + o(h) = \rho_C((1-h)x(t) + hz) + o(h)$$

Convexity of ρ_C gives us,

$$u(t+h) \leq (1-h)u(t) + h\rho_C(z) + o(h) \leq (1-h)u(t) + h\theta u(t) + o(h)$$

Dividing by h and letting $h \to 0$ yields us $\dot{u}(t) \leq -(1-\theta)u(t)$ almost everywhere on $[0, t^*)$, so Grönwall's inequality gives us $u(t) \leq u(0)e^{-(1-\theta)t}$. In particular u is non-increasing, so

$u(t) \leq u(0) < R$. If $t^* < \infty$, continuity gives us $u(t^*) \leq u(0) < R$ and hence $u < R$ on neighbourhood of t^*, contradicting the definition of t^* as a supremum. Therefore $t^* = \infty$. □

Lemma 8 (Selection gap). *Let $0 \leq \epsilon \leq \epsilon_0$ and $q \in \mathcal{C}_\epsilon$. Then,*

$$\max_{a \in \mathcal{A}^\Omega(s)} \min_d q(s,a,d) - \max_{a \notin \mathcal{A}^\Omega(s)} \min_d q(s,a,d) \geq \kappa - 2\epsilon > 0 \; \forall \; s \in \Omega$$

and consequently every maximin greedy policy $\pi = (\mu, \nu) \in \Pi_q$ satisfies $\mu(s) \in \mathcal{A}^\Omega(s)$. For $q \in \mathcal{Q}_{cfg}$, the gap is atleast κ.

Proof. Noting that $\mathcal{A}^\Omega(s) \neq \emptyset \; \forall \; s \in \Omega$, so the first maximin is not over an empty set. For $a \in \mathcal{A}^\Omega(s)$, we have $(s,a,d) \in \mathcal{I} \; \forall \; d \in \mathcal{D}$, by definitions of kernel-preserving action set and index set. Therefore, $\min_d q(s,a,d) \geq -\epsilon$. We now consider two cases for the second maximin

- If $\mathcal{A}^\Omega(s) = \mathcal{A}$, then the second maximin is over an empty set and so $\max_{a \notin \mathcal{A}^\Omega(s)} \min_d q(s,a,d) = -\infty$ by convention. Hence, the claims of the lemma immediately follow.
- If $\mathcal{A}^\Omega(s) \neq \mathcal{A}$, for $a \notin \mathcal{A}^\Omega(s)$, the triple $(s, a, \hat{d}(s,a)) \in \mathcal{J}$ by definition of $\mathcal{J}$. Hence, $\min_d q(s,a,d) \leq q(s,a,\hat{d}(s,a)) \leq -\kappa + \epsilon$ and the gap is therefore at least $\kappa - 2\epsilon$. Noting that this selection gap of $\kappa - 2\epsilon$ is strictly positive since $\epsilon < \epsilon_0 < \kappa/2$ completes the proof.

□

Lemma 8 depends only on the coordinate inequalities definition $\mathcal{C}_\epsilon$, and does not depend on any property of the critic. This is what allows us to use this in the proof of forward invariance without creating circular dependency. Note that this is stated for the whole family $\mathcal{C}_\epsilon$ and $\forall \; \pi \in \Pi_q$ because the lemma applies to arbitrary greedy branch and at configurations in $\mathcal{C}_{\epsilon_0}$ rather than only in $\mathcal{C}_0$. A further consequence is that on $\mathcal{Q}_{\text{cfg}}$ no tie between kernel-preserving and kernel-exiting action can arise, so choice of tie-breaking rule in the selection convention does not affect the following safety conclusions.

B.6.2 Safety Certification of the fast attractor

As the coupled iterated approach the graph of $\Lambda(\cdot)$, $\operatorname{Gr}\Lambda$, rather than a single critic, safety certification (both positive and negative) must hold throughout $\Lambda(q)$, uniformly over $q \in \mathcal{C}_{\epsilon_0}$. For $\sigma > 0$, we define the certificate as,

$$\mathcal{Y}_\sigma = \{y \in \mathcal{Q} : y(x) \geq 0 \; \forall \; x \in \mathcal{I} \; , \; y(x) \leq -\sigma \; \forall \; x \in \mathcal{J}\},$$

a closed convex polyhedron by Lemma 5(ii).

Only the negative half of the certificate that constrains the safety critic on kernel-exiting triples is assumed as follows.

Assumption 3 (Negative Certification of the fast attractor). *$\exists \; \sigma_\Lambda > 0$, such that, at unit margin scale, $y(x) \leq -\sigma_\Lambda \; \forall \; x \in \mathcal{J}, y \in \Lambda_1(q), q \in \mathcal{C}_{\epsilon_0}$.*

The corresponding condition on $\mathcal{I}$ is not assumed because it is preserved by every branch of the fast inclusion (proved in the following lemma and corollary) and therefore the certificate holds throughout $\Lambda(q)$.

Lemma 9 (Branchwise absorption of the positive constraints). *Let $q \in \mathcal{C}_{\epsilon_0}$ and $\pi = (\mu, \nu) \in \Pi(q)$. If $y(x) \geq 0 \; \forall \; x \in \mathcal{I}$, then $\mathcal{T}_\pi^\phi[y](x) \geq 0 \; \forall \; x \in \mathcal{I}$.*

Proof. Let $x = (s,a,d) \in \mathcal{I}$. Since $s \in \Omega \subseteq \mathcal{F}^c$, we have $Kg(s) > 0$ by definition of the scaled safety margin function. As $a \in \mathcal{A}^\Omega(s)$, every successor state $s' \in \Omega \; \forall \; d \in \mathcal{D}$. By Lemma 8, $\mu(s') \in \mathcal{A}^\Omega(s') \implies (s', \mu(s'), \nu(s', \mu(s'))) \in \mathcal{I} \implies y(s', \mu(s'), \nu(s', \mu(s'))) \geq 0$. Hence, $\mathbb{E}_{s'}[y(\cdot)] \geq 0$. Now, substituting these inequalities in the safety Bellman equation, Eq. 5, we get

$$\mathcal{T}_\pi^\phi[y](x) = (1-\gamma)Kg(s) + \gamma \min\{Kg(s), \mathbb{E}_{s'}[y(\cdot)]\} \geq 0$$

□

Corollary 2 (Positive certification)**.** *For every $q \in \mathcal{C}_{\epsilon_0}$ and every $y \in \Lambda_1(q)$, $y(x) \geq 0 \; \forall \; x \in \mathcal{I}$. Together with Assumption 3 this gives $\Lambda_1(q) \subseteq \mathcal{Y}_{\sigma_\Lambda}$.*

Proof. Let $\mathcal{P} := \{y : y(x) \geq 0 \; \forall \; x \in \mathcal{I}\}$, a closed convex polyhedron, and let $\rho_{\mathcal{P}}$ denote the supremum-norm distance to $\mathcal{P}$.

1. As in Lemma 5, $\rho_{\mathcal{P}}(y) = \max_{x \in \mathcal{I}}(-y(x))_+$. If y violates the constraints of $\mathcal{P}$ by at most $\epsilon \geq 0$, then $\mathcal{T}_\pi^\phi[y](x) \geq \gamma \min\{Kg(s), -\epsilon\} \geq -\gamma\epsilon \; \forall \; x \in \mathcal{I}, \; \pi \in \Pi(q)$ as in Lemma 9. This implies $\rho_{\mathcal{P}}(\mathcal{T}_\pi^\phi[y]) \leq \gamma \rho_{\mathcal{P}}(y)$.
2. Next, every element of $\Psi(q, y)$ has the form $w - y$ with $\rho_{\mathcal{P}}(w) \leq \gamma \rho_{\mathcal{P}}(y)$, because of convexity of the relaxation of $\mathcal{P}$. Lemma 7 therefore applies, with $C = \mathcal{P}$, $\theta = \gamma$, and $R = \infty$, at every y.
3. For $y \in \Lambda_1(q)$, by Eq. 11 there is a bounded complete solution $y(\cdot)$ of $\dot{y} \in \Psi(q, y)$ with $y(0) = y$, and by shift invariance $y(s) \in \Lambda_1(q) \; \forall \; s \in \mathbb{R}$.
4. Noting that $\Lambda_1(q)$ is compact and $\rho_{\mathcal{P}}$ is continuous, we get $M := \max_{w \in \Lambda_1(q)} \rho_{\mathcal{P}}(w)$ is finite and $\rho_{\mathcal{P}}(y(-t)) \leq M \; \forall \; t > 0$.
5. Applying Lemma 7 on interval $[-t, 0]$ with initial point of $y(-t)$, we get,
$$\rho_{\mathcal{P}}(y(0)) \leq \rho_{\mathcal{P}}(y(-t))\, e^{-(1-\gamma)t} \; \forall \; t > 0$$
6. Taking the limit of $t \to \infty$, we get, $\rho_{\mathcal{P}}(y(0)) = 0$ and $\mathcal{P}$ is closed. Hence, $y \in \mathcal{P}$.

□

Remark 7 (On the selection of $\hat{d}$ and index sets)**.** *The selection $\hat{d}$ is unconstrained, and can be made to satisfy Assumption 3; this assumption is a condition on the existence of a suitable selection; however, it is not a condition on every kernel-exiting disturbance. Enlarging $\mathcal{J}$ to contain every kernel-exiting triple would shrink $\mathcal{Q}_{cfg}$ and strengthen Assumption 3 without improving any conclusion, since Lemma 8 requires only $\min_d q(s, a, d) \leq -\kappa$, which a single coordinate already delivers.*

Remark 8 (Only the condition on $\mathcal{J}$ is assumed)**.** *We now want to reason whether the Corollary 2 extends to $\mathcal{J}$. Towards this end, let us first consider $x = (s, a, \hat{d}(s, a)) \in \mathcal{J}$, where $s \in \Omega$ by definition of index set. This means $Kg(s) > 0$ and hence substituting in Eq. 5, we get $(1-\gamma)Kg(s) > 0$. If $\mathcal{T}_\pi^\phi[y](x) \leq -\sigma$, then $\mathbb{E}_{s'}[y(\cdot)] \leq Kg(s)$ and $\mathbb{E}_{s'}[y(\cdot)] \leq \frac{-\sigma + (1-\gamma)Kg(s)}{\gamma}$. This second term of minimum in Eq. 5 is controlled by the values of $y(x')$, where $x' = (s', \cdot, \cdot)$ and $s' \notin \Omega$, at successor states outside of Ω. Note that the index sets, $\mathcal{I}$ and $\mathcal{J}$ do not provide any information on the value of y at such successor states outside of Ω. Hence, we get a structural asymmetry: positive certification propagates since kernel-preserving action keeps every successor in Ω where we have the constraints, whereas negative certification must propagate into a region where nothing has been assumed. Certifying the condition on J therefore requires the Assumption 3 and Corollary 2 cannot extend to $\mathcal{J}$.*

Proposition 3 (Margin Separation)**.** *Let Assumption 3 hold and set $\Theta := ((1-\gamma)R_{max} + \kappa)/\gamma$. Then, for every*

$$K \geq \underline{K} := \frac{(1-\gamma)R_{max} + \kappa}{\gamma \sigma_\Lambda} \tag{38}$$

one has $y(x) \leq -\Theta$ for all $x \in \mathcal{J}$, all $y \in \Lambda(q)$, and all $q \in \mathcal{C}_{\epsilon_0}$. If $\mathcal{J} = \emptyset$ the bound on K and the conclusion holds trivially.

Proof.

1. The safety Bellman operator is positively homogeneous of degree one in the margin-replacing g by $Kg(s)$ and Q^ϕ by KQ^ϕ leaves the Eq. 5 invariant for all $K > 0$.
2. $\Pi(q)$ is greedy with respect to q and is therefore independent of the positive margin scale.
3. Noting that positive scaling commutes with the closed convex hull and graph completion, $\Psi_K(q, Ky) = K\Psi_1(q, y)$.

4. $y(\cdot)$ is a bounded complete solution at unit scale if and only if $Ky(\cdot)$ is a complete solution at positive scale K.
5. Taking time zero values in Eq. 11, gives us $\Lambda_K(q) = K\Lambda_1(q)$.
6. By Assumption 3, every $y_1 \in \Lambda_1(q)$ satisfies $y_1(x) \leq -\sigma_\Lambda \forall x \in \mathcal{J}$.
7. Similarly, for every $y = Ky_1 \in \Lambda_K(q)$ satisfies $y(x) \leq -K\sigma \leq -\Theta$ whenever $K\sigma \geq \Theta$, which is the stated bound.

□

Note only the Assumption 3 interacts with K. The condition on $\mathcal{I}$ obtained in Corollary 2 is positive scale-invariant, that is, multiplying by $K > 0$ does not change the sign.

B.6.3 Uniform violation contraction

Lemma 10 (Target Bounds)**.** *Let $0 \leq \epsilon \leq \epsilon_0$, $q \in \mathcal{C}_\epsilon$, and $y \in \Lambda(q)$. Then (i) $T_y[q](x) \leq -\kappa$ for every $x \in \mathcal{J}$, and (ii) $T_y[q](x) \geq -\gamma\epsilon$ for every $x \in \mathcal{I}$.*

Proof. (i) Let $x = (s, a, d) \in \mathcal{J}$. Bounding each $\min$ in Eq. 6 by one of its arguments and using $m(s, a, d) = \min\{r, Kg(s)\} \leq r \leq R_{\max}$,

$$T_y[q](x) = (1-\gamma)m(s,a,d) + \gamma \min\{r + \mathbb{E}[V_q(s')], y(x)\} \leq (1-\gamma)R_{\max} + \gamma y(x).$$

Since $q \in \mathcal{C}_{\epsilon_0}$ and $y \in \Lambda(q)$, Proposition 3 gives $y(x) \leq -\Theta$, so $T_y[q](x) \leq (1-\gamma)R_{\max} - \gamma\Theta = -\kappa$ (again using definition of Θ as defined in Proposition 3). The bound depends only on $R_{\max}$, γ and Θ, and is therefore uniform in q and in $y \in \Lambda(q)$.
(ii) Let $x = (s, a, d) \in \mathcal{I}$. Since $s \in \Omega \subseteq F^c$ we have $g(s) > 0$, and $r \geq 0$, so $m(s, a, d) \geq 0$ and the first term of Eq. 6 is non-negative. For the continuation term, $a \in \mathcal{A}^{\overline{\Omega}}(s)$, so every successor s' lies in Ω - against every disturbance, not only the one appearing in x. As $s' \in \Omega$, the set $\mathcal{A}^{\Omega}(s') \neq \emptyset$; fix $a^+ \in \mathcal{A}^{\Omega}(s')$. Since $\mathcal{I}$ contains every disturbance coordinate, $(s', a^+, d') \in \mathcal{I}$ for all $d' \in D$, and $q \in \mathcal{C}_\epsilon$ gives $q(s', a^+, d') \geq -\epsilon$ for each such d'. Hence

$$V_q(s') = \max_{a'} \min_{d'} q(s', a', d') \geq \min_{d'} q(s', a^+, d') \geq -\epsilon$$

the first inequality because a maximum over a' is at least the value at the particular action a^+. Therefore $r + E[V_q(s')] \geq -\epsilon$, while $y(x) \geq 0$ by Corollary 2 on $\mathcal{I}$. Since $\min\{X, Y\} \geq \min\{x, y\}$ whenever $X \geq x$ and $Y \geq y$,

$$T_y[q](x) \geq (1-\gamma) \cdot 0 + \gamma \cdot \min\{-\epsilon, 0\} = -\gamma\epsilon$$

□

Part (ii) of the above proof never identifies which action the greedy policy of Eq. 4 selects at s', nor which disturbance is applied there: it requires only that some kernel-preserving action exist at s' and carry value at least $-\epsilon$. This is why the lower constraints propagate without any hypothesis on the disturbance policy, whereas the upper constraints require Assumption 3 on $\mathcal{J}$.

Corollary 3 (The joint operator contracts the violation)**.** *For $0 \leq \epsilon \leq \epsilon_0$ and every $y \in \Lambda(q)$, $T_y[\mathcal{C}_\epsilon] \subseteq \mathcal{C}_{\gamma\epsilon}$.*

Proof. Let $q \in \mathcal{C}_\epsilon$ and $y \in \Lambda(q)$. Lemma 10(ii) gives $T_y[q](x) \geq -\gamma\epsilon$ for $x \in \mathcal{I}$, and Lemma 10(i) gives $T_y[q](x) \leq -\kappa \leq -\kappa + \gamma\epsilon$ for $x \in \mathcal{J}$. These are precisely the defining inequalities of $\mathcal{C}_{\gamma\epsilon}$. □

This is the property on which the invariance argument depends on: the joint Bellman operator maps each relaxed constraint set into a smaller one, uniformly over the critics in the fast attractor. It is *not* a consequence of Proposition 2, which holds the safety action-value fixed. Here, the critic ranges over $\Lambda(q)$ while q itself moves and the composed map $q \mapsto T_{\mathbf{Y}_q}[q]$ need not induce a contraction. Corollary 3 asserts the weaker but sufficient property that the operator contracts the *distance* to $\mathcal{Q}_{\text{cfg}}$.

We now proceed to transfer the property from the individual operators to their *occupation-measure average* in the next section of the proof.

B.6.4 AVERAGING OVER THE FAST ATTRACTOR PRESERVES THE VIOLATION CONTRACTION

Lemma 11 (Averaging preserves the violation contraction)**.** *Let $\rho(q) = \epsilon < \epsilon_0$. Then every $v \in \overline{G}(q)$ has the form $v = z - q$ with $z \in \mathcal{C}_{\gamma\epsilon}$.*

Proof. 1. By Eq. 12, there is a probability measure, $\mathbb{P} \in \mathfrak{D}(q)$ supported on $\Lambda(q)$ with

$$v = \int_{\Lambda(q)} (T_y[q] - q)\mathbb{P}(dy) = \int_{\Lambda(q)} T_y[q]\mathbb{P}(dy) - \int_{\Lambda(q)} q\mathbb{P}(dy) = \left(\int_{\Lambda(q)} T_y[q]\mathbb{P}(dy)\right) - q$$

2. Comparing this to $v = z - q$, gives us that $z = \int_{\Lambda(q)} T_y[q]\mathbb{P}(dy)$. In words, z is the averaged target.

3. From Corollary 3, we know $T_y[q] \in \mathcal{C}_{\gamma\epsilon}$ for every $y \in \Lambda(q)$.

4. From Lemma 5(ii), we know that $\mathcal{C}_{\gamma\epsilon}$ is closed and convex. By extension, the averaged target $z \in \mathcal{C}_{\gamma\epsilon}$.

□

Averaging over occupation measures still preserves the violation contraction because of the convexity of $\mathcal{C}_\epsilon$, since $\text{conv}(\mathcal{C}_\epsilon) = \mathcal{C}_\epsilon$.

B.6.5 FORWARD INVARIANCE

Theorem 3 (Forward invariance and exponential attraction)**.** *Let Assumption 3 hold and let $K \geq \underline{K}$. If $\rho(q(0)) < \epsilon_0$, then every solution of $\dot{q}(t) \in \overline{G}(q(t))$ satisfies $\rho(q(t)) \;\leq\; \rho(q(0))\, e^{-(1-\gamma)t} \quad \forall\, t \geq 0$. In particular, if $q(0) \in \mathcal{Q}_{cfg}$ then $q(t) \in \mathcal{Q}_{cfg}\ \forall\, t \geq 0$, and $\mathcal{Q}_{cfg}$ is exponentially attracting with basin containing the open neighbourhood $\{q : \rho(q) < \epsilon_0\}$. (Proof in Appendix B.6)*

Proof. 1. Solutions of Eq. 37 are by definition absolutely continuous functions of time satisfying $\dot{q}(t) \in \overline{G}(q(t))$ for almost every t.

2. Such solutions exist because $\overline{G}$ is upper semicontinuous with nonempty compact convex values, Λ is nonempty, compact and upper semicontinuous by Lemma 1 and the barycentre map preserves these properties.

3. By Lemma 11, for every q with $\rho(q) < \epsilon_0$ each $v \in \overline{G}(q)$ has the form $v = z - q$ with $z \in \mathcal{C}_{\gamma\rho(q)}$.

4. By Lemma 5(iii), $\rho(z) \leq \gamma\rho(q)$.

5. The hypothesis of Lemma 7 therefore holds with $C = \mathcal{Q}_{\text{cfg}}$, $\theta = \gamma$ and $R = \epsilon_0$. Since $\rho(q(0)) < \epsilon_0$, applying Lemma 7, yielding:

$$\rho(q(t)) \;\leq\; \rho(q(0))\, e^{-(1-\gamma)t} \quad \forall\, t \geq 0. \tag{39}$$

6. If $q(0) \in \mathcal{Q}_{\text{cfg}}$ then $\rho(q(0)) = 0$, so $\rho \equiv 0$ and $q(t) \in \mathcal{Q}_{\text{cfg}}\ \forall\ t$ by Lemma 5(iii). The exponential attractivity statement is the bound in Eq. 39.

□

B.6.6 THE SAFE CONFIGURATION SET CERTIFIES SAFETY

Lemma 12 (Robust invariance of Ω)**.** *Let $q \in \mathcal{Q}_{cfg}$. Then (i) $\inf_{s\in\Omega} V^{\phi}_{\pi_q}(s) \geq 0$; and (ii) for every $s_0 \in \Omega$ and every disturbance sequence in $\mathcal{D}$, the state trajectory generated by μ_q satisfies $\mathbb{P}(s_t \in \Omega\ \forall\, t \geq 0) = 1$.*

Proof. We will first prove Part(i) below.

1. The fixed point $Q^{\phi}_{\pi_q}$ of the safety Bellman operator $\mathcal{T}^{\phi}_{\pi_q}$ satisfies $\mathcal{T}^{\phi}_{\pi_q}[Q^{\phi}_{\pi_q}] - Q^{\phi}_{\pi_q} = 0 \in \Psi(q, Q^{\phi}_{\pi_q})$ by Eq. 11.
2. So $Q^{\phi}_{\pi_q}$ is an equilibrium of the frozen inclusion. Therefore, by the second consequence of Eq. 11, $Q^{\phi}_{\pi_q} \in \Lambda(q)$.
3. Now, applying Corollary 2, gives us $Q^{\phi}_{\pi_q}(x) \geq 0 \; \forall \; x \in \mathcal{I}$.
4. By Lemma 8, $\mu_q(s) \in \mathcal{A}^{\Omega}(s)$ and therefore $(s, \mu_q(s), \nu_q(s, \mu_q(s))) \in \mathcal{I} \; \forall \; \nu_q(s, \mu_q(s)) \in \mathcal{D}$.
5. Noting that μ_q and ν_q are deterministic, the maximum and minimum in the definition of the safety value function is over singletons, giving us
$$V^{\phi}_{\pi_q} = \max_{a \in \mu_q(s)} \min_{d \in \nu_q(s, \mu_q(s))} Q^{\phi}_{\pi_q}(s, a, d) = Q^{\phi}_{\pi_q}(s, \mu_q(s), \nu_q(s, \mu_q(s))) \geq 0$$
6. As the above reasoning holds $\forall \; s \in \Omega$, taking infimum over Ω, yields, $\inf_{s \in \Omega} V^{\phi}_{\pi_q} \geq 0$.

We will next prove Part(ii). Let $E_t := \{s_t \in \Omega\}$. We show $\mathbb{P}(E_t) = 1$ for every t by induction.

1. $\mathbb{P}(E_0) = 1$ by hypothesis.
2. Suppose $\mathbb{P}(E_t) = 1$. On E_t, Lemma 8 gives $\mu_q(s_t) \in \mathcal{A}^{\Omega}(s_t)$, so by the definition of $\mathcal{A}^{\Omega}$, $\mathbb{P}(s_{t+1} \in \Omega \mid s_t, \mu_q(s_t), d) = 1 \; \forall \; d \in \mathcal{D}$; taking expectations over E_t yields $\mathbb{P}(E_{t+1}) = 1$, whichever disturbance is applied.
3. Since the time index is discrete, $\{s_t \in \Omega \; \forall \; t \geq 0\} = \bigcap_{t \geq 0} E_t$ is a countable intersection of probability-one events and therefore has probability one.

□

Corollary 1 (Safety of the induced policy)**.** *Let $q(0) \in \mathcal{Q}_{cfg}$. Then at every $t \geq 0$ the induced policy $\pi_{q(t)}$ satisfies $\inf_{s \in \Omega} V^{\phi}_{\pi_{q(t)}}(s) \geq 0$ from any $s_0 \in \Omega$, the state trajectory generated by $\mu_{q(t)}$ remains in Ω almost surely, for all time and against every disturbance sequence. (Proof in Appendix B.6)*

Proof. From Theorem 3, we get $q(t) \in \mathcal{Q}_{\text{cfg}} \; \forall \; t \geq 0$ because $q(0) \in \mathcal{Q}_{\text{cfg}}$. Finally, applying Lemma 12 at $q = q(t)$, both assertions follow. □

Remark 9 (Scope)**.** *Corollary 1 concerns the trajectory generated by the policy induced at a given configuration. Safety throughout training would additionally require safe initialization together with a safe exploration mechanism, neither of which is assumed here.*

Remark 10 (Intuition for the Geometric Argument)**.** *The drift points from the current configuration toward an average of its Bellman images under the critics in $\Lambda(q)$, each of which lies closer to $\mathcal{Q}_{cfg}$ by the factor γ. A step along the drift is therefore a chord from q into $\mathcal{C}_{\gamma\rho(q)}$, and convexity of the distance function converts this into decay of ρ. Forward invariance is thus a statement about convex geometry.*

B.7 Convergence to Optimal Safety-Constrained Task Performance (Sec. 6)

Definition 5 (Optimal safety-constrained task value function)**.** *Let $\mathcal{Q}_{\Omega} := \mathbb{R}^{\mathcal{I}} = \{\mathfrak{q} : \mathcal{I} \to \mathbb{R}\}$ denote the restricted action-value space on the index set $\mathcal{I}$. Define the robust safety-constrained task Bellman operator $\mathcal{T}_{\Omega} : \mathcal{Q}_{\Omega} \to \mathcal{Q}_{\Omega}$:*

$$\mathcal{T}_{\Omega}[\mathfrak{q}](s, a, d) = r(s, a, d) + \gamma \mathbb{E}\left[\max_{a' \in A^{\Omega}(s')} \min_{d' \in \mathcal{D}} \mathfrak{q}(s', a', d')\right], \quad (s, a, d) \in \mathcal{I} \tag{14}$$

Note that since $A^{\Omega}(s) \neq \emptyset$ for $s \in \Omega$, by contraction, $\mathcal{T}_{\Omega}[\cdot]$ has a unique fixed point $Q^{*}_{\Omega} = \mathcal{T}_{\Omega} Q^{*}_{\Omega}$, and thus, the optimal safety-constrained task value function is given by $V^{*}_{\Omega}(s) =$

$\max_{a\in A^\Omega(s)} \min_{d\in\mathcal{D}} Q^*_\Omega(s,a,d)$*, which is the maximum expected return that a safe policy can accrue by keeping the agent in* Ω*. Moreover, let* $\mathcal{Q}^* := \{q \in \mathcal{Q}_{\text{cfg}} : q|_{\mathcal{I}} = Q^*_\Omega\}$*. Thus, every* $q \in \mathcal{Q}^* \subseteq \mathcal{Q}$ *corresponds to a policy that is safe and attains the maximal possible return among all safe policies.*

Let $Q^{\phi,1}_\pi$ denote the safety critic obtained by scaling the margin $g(\cdot)$ by $K = 1$. By homogeneity, $Q^\phi_\pi := Q^{\phi,K}_\pi = KQ^{\phi,1}_\pi$. Next, let $\delta_\Omega := \inf_{q\in\mathcal{Q}_{\text{cfg}}} \min_{\pi\in\Pi(q)} \min_{s\in\Omega} V^{\phi,1}_\pi > 0$ represent the lowest safety value (represents the *closest* a safe policy ever comes to $\partial\Omega$).

Theorem 4 (Convergence to optimal safety-constrained task performance)**.** *Let the sufficient margin scale* K^* *be defined as* $K^* := \frac{R_{\max}+\overline{V}}{\delta_\Omega} = \frac{(2-\gamma)R_{\max}}{(1-\gamma)\delta_\Omega}$*, with* $\overline{V} := \frac{R_{\max}}{1-\gamma}$*. Let* $\mathcal{U} \subset \{q : \rho(q) < \epsilon_0\}$*. Then, under Theorems 2-3, almost surely bounded iterates* (Y_n, Q_n) *and* $K \geq K^* \geq \underline{K}$*,* $\delta_\Omega > 0$*, every internally chain-transitive set of the averaged slow-timescale differential inclusion contained in* $\mathcal{Q}_{\text{cfg}}$ *is a subset of* $\mathcal{Q}^*$*:* $\mathcal{U} \subseteq \mathcal{Q}^* := \{q \in \mathcal{Q}_{\text{cfg}} : q|_{\mathcal{I}} = Q^*_\Omega\}$*, and thus,* $\operatorname{dist}(Q_n, \mathcal{Q}^*) \to 0$ *almost surely. Moreover, on* $\mathcal{U}$ *the averaged slow-timescale residual on kernel-preserving coordinates* $\mathcal{I}$ *is given by* $\overline{G}(q)|_{\mathcal{I}} = \{\mathcal{T}_\Omega[q|_{\mathcal{I}}] - q|_{\mathcal{I}}\}$*. Thus, every corresponding limiting controller attains the optimal task value* V^*_Ω *on* Ω *while selecting only kernel-preserving actions. (Proof in Appendix B.7).*

Proof. For any $q_0 \in \mathcal{U}$, since $\mathcal{U}$ is compact and $\mathcal{U} \subseteq \{q : \rho(q) < \epsilon_0\}$, applying Theorem 3 yields $\mathcal{U} \subseteq \mathcal{Q}_{\text{cfg}}$. Then, $\mathcal{U} \subseteq \mathcal{Q}_{\text{cfg}}$ is a compact invariant set of the averaged slow-timescale differential inclusion $\dot{q}(t) \in \overline{G}(q(t))$. We will first prove that $\mathcal{U} \subseteq \mathcal{Q}^*$.

1. By definition of δ_Ω, for every $q \in \mathcal{Q}_{\text{cfg}}$, every $\pi \in \Pi(q)$, and every $s \in \Omega$, we have $V^{\phi,1}_\pi(s) \geq \delta_\Omega$, and thus, $g(s) \geq \delta_\Omega$ for $s \in \Omega$. We will now show that every $y \in \Lambda(q)$ satisfies $y(x_i) \geq K\delta_\Omega$ for $x_i \in \mathcal{I}$.

   Define $\zeta(y) := \max_{x_i\in\mathcal{I}} (K\delta_\Omega - y(x_i))^+$ which, by definition gives $y(x_i) \geq K\delta_\Omega - \zeta(y)$. Next, note that, since $x_i = (s,a,d) \in \mathcal{I}$, by definition, $s \in \Omega$, and $a \in A^\Omega(s)$, which implies that $P(s' \in \Omega \mid s,a,d) = 1$. Thus, almost surely, $\mathbb{E}[y(s',\pi(s'))] \geq K\delta_\Omega - \zeta(y)$. Applying the above facts to the safety Bellman operator yields:

   $$\mathcal{T}^\phi_\pi[y](x_i) = (1-\gamma)Kg(s) + \gamma \min\{Kg(s), \mathbb{E}[y(s',\pi(s'))]\} \tag{40a}$$
   $$\geq (1-\gamma)Kg(s) + \gamma\left(K\delta_\Omega - \zeta(y)\right) \tag{40b}$$
   $$= K\delta_\Omega - \gamma\zeta(y) \tag{40c}$$

   Since this holds for all policies $\pi \in \Pi(q)$, every convex combination has the same lower bound. Next, under Assumption 2, the induced limiting fast timescale dynamics is given by $\dot{y}(x_i) = \mathcal{T}^\phi_\pi[y](x_i) - y(x_i)$. Plugging the above result yields $\dot{y}(x_i) \geq (1-\gamma)\zeta(y)$. Next, let $E(t) := \{x_i \in \mathcal{I} : x_i = \arg\max_{x_i\in\mathcal{I}} K\delta_\Omega - y(x_i)(t)\}$. On these coordinates, the upper Dini derivative satisfies:

   $$\mathbb{D}^+\zeta(y(t)) \leq \max_{i\in E(t)} \mathbb{D}^+ \left(K\delta_\Omega - y(x_i)(t)\right)^+ \tag{41a}$$
   $$\leq -(1-\gamma)\zeta(y(t)) \tag{41b}$$
   $$\therefore\ \zeta(y(t)) \leq e^{-(1-\gamma)(t-s)}\zeta(y(s)),\ t \geq s \quad \text{(By Grönwall)} \tag{41c}$$

   For contradiction, assume $\zeta(y(0)) > 0$, which yields $\zeta(y(s)) \geq e^{(1-\gamma)(0-s)}\zeta(y(0))$. As $s \to -\infty$ (since $s < t$), $\zeta(y(s)) \to \infty$ which is a contradiction, since $y(\cdot)$ is bounded, and thus, $\zeta(y(\cdot))$ must also be bounded. Thus, $\zeta(y(0)) = 0$, and since $y(0)$ is an arbitrary point of $\Lambda(q)$, we must have $y(x_i) \geq K\delta_\Omega$, for all $x_i \in \mathcal{I}$ and for all $y \in \Lambda(q)$.

2. Next, let $\overline{V} := \frac{R_{\max}}{1-\gamma}$ be the maximum task return possible. For a quasi-static $q \in \mathcal{U}, y \in \Lambda(q)$ and $x_i = (s,a,d) \in \mathcal{I}$, since $K^* \geq K$, we have:

   $$Kg(s) \geq K\delta_\Omega \geq R_{\max} + \overline{V} \geq R_{\max} \geq r(x_i) \tag{42}$$

   which implies that the safety truncation is inactive. Moreover, we showed above that $y(x_i) \geq K\delta_\Omega$, and thus, the joint Bellman operator of Eq. 6 reduces to:

   $$\mathcal{T}_y[q](x_i) = r(x_i) + \gamma\mathbb{E}\left[V_q(s')\right] \tag{43}$$

Since $a \in A^{\Omega}(s)$, $s' \in \Omega$ almost surely; moreover, since $q \in \mathcal{Q}_{\text{cfg}}$ implies that every maximin greedy controller action at $s' \in \Omega$ is kernel preserving, we have:

$$V_q(s') = \max_{a \in A^{\Omega}(s)} \min_{d' \in \mathcal{D}} q(s', a', d') \tag{44}$$

thus, yielding Eq. 14. Consequently, with $\mathfrak{q} = q|_{\mathcal{I}}$, we have:

$$G(y, q)|_{\mathcal{I}} = \mathcal{T}_{\Omega}[\mathfrak{q}] - \mathfrak{q} \quad \forall y \in \Lambda(q) \tag{45}$$

Averaging over any $\mathbb{P} \in \mathfrak{D}(q)$yields: (46)

$$\overline{G}(q)|_{\mathcal{I}} = \{\mathcal{T}_{\Omega}[q|_{\mathcal{I}}] - q|_{\mathcal{I}}\}, \quad q \in \mathcal{U} \tag{47}$$

3. Next, note that the operator $\mathcal{T}_{\Omega}$ is a γ-contraction in the supremum norm:

$$\|\mathcal{T}_{\Omega}[\mathfrak{q}] - \mathcal{T}_{\Omega}[\mathfrak{q}']\|_{\infty} \leq \gamma \|\mathfrak{q} - \mathfrak{q}'\|_{\infty} \tag{48}$$

and thus, possesses a unique fixed point Q^*_{Ω}. Moreover, let $q(\cdot)$ be any complete solution contained in $\mathcal{U}$. Let $\mathfrak{q}(t) := q(t)|_{\mathcal{I}}$. Then, by Eq. 47, $\dot{\mathfrak{q}}(t) = \mathcal{T}_{\Omega}[\mathfrak{q}(t)] - \mathfrak{q}(t)$. Now, since $\mathcal{T}_{\Omega}[Q^*_{\Omega}] = Q^*_{\Omega}$, its contraction yields:

$$\mathbb{D}^+ \|\mathfrak{q}(t) - Q^*_{\Omega}\|_{\infty} \leq -(1-\gamma) \|\mathfrak{q}(t) - Q^*_{\Omega}\|_{\infty} \tag{49}$$

If $\mathfrak{q}(0) \neq Q^*_{\Omega}$, then by Grönwall:

$$\|\mathfrak{q}(0) - Q^*_{\Omega}\|_{\infty} \leq e^{-(1-\gamma)(0-s)} \|\mathfrak{q}(s) - Q^*_{\Omega}\|_{\infty}, \quad s < t = 0 \tag{50}$$

which contradicts the compactness of $\mathcal{U}$. Thus, $q|_{\mathcal{I}} = Q^*_{\Omega}$ for all $q \in \mathcal{U}$. Moreover, since $\mathcal{U} \subseteq \mathcal{Q}_{\text{cfg}}$, we have

$$\mathcal{U} \subseteq \{q \in \mathcal{Q}_{\text{cfg}} : q|_{\mathcal{I}} = Q^*_{\Omega}\} = \mathcal{Q}^* \tag{51}$$

4. Finally, by Theorem 2, the slow-timescale continuous affine interpolation $\overline{Q}(\cdot)$ has a nonempty compact invariant internally chain-transitive ω-limit set. Moreover, by Theorem 3, this limit set lies in $\mathcal{Q}_{\text{cfg}}$. Hence, by the previous argument, we have almost surely $\omega(\overline{Q}) \subseteq \mathcal{Q}^*$. Since the iterates $\{Q_n\}$ are almost surely bounded, and $\mathcal{Q}^*$ is closed, every subsequential limit of $\{Q_n\}$ belongs to $\mathcal{Q}^*$. Hence,

$$\text{dist}\,(Q_n, \mathcal{Q}^*) \to 0 \qquad \text{a.s.} \tag{52}$$

Thus, every corresponding limiting controller attains the optimal task value V^*_{Ω} on Ω while selecting only kernel-preserving actions. □

Intuitively, the sufficient margin scale K^* scales the safety critic *just enough* (i.e., based on δ_{Ω}) to ensure that the safety truncation is never active (as illustrated in Fig. 1). Consequently, if $\delta_{\Omega} \to 0$, i.e., if the safe policy approaches $\partial\Omega$ arbitrarily closely, then $K^* \to \infty$. If $\delta_{\Omega} = 0$, Theorem 4 no longer guarantees convergence to the optimal safety-constrained task value for any finite K; we can characterize the resulting finite-K conservatism as follows. Let Γ_{Ω} be the set of kernel-preserving controller policies, and for $\zeta \geq 0$, define:

$$\Gamma_{\Omega,\zeta} := \left\{\mu \in \Gamma_{\Omega} : \inf_{\nu} \inf_{s \in \Omega} V^{\phi,1}_{(\mu,\nu)}(s) \geq \zeta\right\} \qquad V^*_{\Omega,\zeta}(s) := \sup_{\mu \in \Gamma_{\Omega,\zeta}} \inf_{\nu} V^{\psi}_{\mu,\nu}(s) \tag{53}$$

Let $\Delta_{\text{marg}}(\zeta) := \|V^*_{\Omega} - V^*_{\Omega,\zeta}\|_{\infty}$, and let V^*_K denote the optimal joint value of Eq. 3 over kernel-preserving controllers, for some finite $K \geq \underline{K}$.

Theorem 5 (Finite-K characterization)**.** *Define $\zeta_K := \frac{R_{\max}+\overline{V}}{K} = \frac{(2-\gamma)R_{\max}}{(1-\gamma)K}$. Assume there exists $\overline{\zeta}$ such that $\Gamma_{\Omega,\zeta_K} \neq \emptyset$ for all $0 < \zeta_K < \overline{\zeta}$ (sufficiently large K). Then, $0 \leq V^*_{\Omega}(s) - V^*_K(s) \leq \Delta_{\text{marg}}(\zeta_K)$ for $s \in \Omega$. Hence, if $\Delta_{\text{marg}}(\zeta) \to 0$ as $\zeta \downarrow 0$, then $V^*_K \to V^*_{\Omega}$ as $K \to \infty$. (Proof in Appendix B.7).*

Proof. For a given policy $\pi = (\mu, \nu)$, let $V^K_{\mu,\nu}$ denote the unique fixed point of the finite-K joint Bellman equation of Eq. 6, written as:

$$\mathcal{T}_{V^{\phi,1}_{\mu,\nu}}[v](s) = (1-\gamma)\min\{r_{\mu,\nu}(s), Kg(s)\} + \gamma \min\left\{r_{\mu,\nu}(s) + \mathbb{E}[v(s')], KV^{\phi,1}_{\mu,\nu}(s)\right\} \tag{54}$$

Define $V^*_K(s) := \sup_{\mu \in \Gamma_{\Omega}} \inf_{\nu} V^K_{\mu,\nu}(s)$. Then, we have the following:

1. Let $\mathcal{T}^{\psi}_{\mu,\nu}[v](s) := r_{\mu,\nu}(s) + \gamma\mathbb{E}[v(s')]$ be the task Bellman operator of Eq. 1. Note that by definition of $\min\{\cdot\}$, we have $\mathcal{T}_{V^{\phi,1}_{\mu,\nu}}[v] \leq \mathcal{T}^{\psi}_{\mu,\nu}[v]$; taking $\inf_\nu$ and $\sup_{\mu\in\Gamma_\Omega}$ on both sides yield:

$$V^*_K(s) \leq \sup_{\mu\in\Gamma_\Omega}\inf_{\nu} V^{\psi}_{\mu,\nu}(s) = V^*_\Omega(s). \tag{55}$$

Thus, finite-K truncation can only *decrease the task return* obtained by the agent.

2. Next, define:

$$K^* := \frac{R_{\max}+\overline{V}}{\delta_\Omega} = \frac{(2-\gamma)R_{\max}}{(1-\gamma)\delta_\Omega}, \qquad \overline{V} := \frac{R_{\max}}{1-\gamma} \tag{56}$$

Next, note that, for $\mu_{\zeta_K} \in \Gamma_{\Omega,\zeta_K}$, $V^K_{\mu_{\zeta_K},\nu} = V^{\psi}_{\mu_{\zeta_K},\nu}$ i.e., if the controller μ is uniformly "ζ_K units" away from $\partial\Omega$, scaling by K prevents safety truncation to be active. This argument follows analogously from Eq. 56 and Eq. 42. Then, the following holds:

$$\begin{aligned}
V^*_K(s) &= \sup_{\mu\in\Gamma_\Omega}\inf_{\nu} V^K_{\mu,\nu}(s) && \text{(57a)}\\
&\geq \sup_{\mu\in\Gamma_{\Omega,\zeta_K}}\inf_{\nu} V^K_{\mu,\nu}(s) \quad \text{(finite-}K\text{ suboptimality)} && \text{(57b)}\\
&= \sup_{\mu\in\Gamma_{\Omega,\zeta_K}}\inf_{\nu} V^{\psi}_{\mu_{\zeta_K},\nu}(s) && \text{(57c)}\\
&= V^*_{\Omega,\zeta_K}(s) && \text{(57d)}
\end{aligned}$$

Combining the above inequalities, we have:

$$\begin{aligned}
V^*_{\Omega,\zeta_K}(s) \leq& V^*_K(s) \leq V^*_\Omega(s) && \text{(58a)}\\
\implies 0 \leq V^*_\Omega(s) - V^*_K(s) &\leq V^*_\Omega - V^*_{\Omega,\zeta_K}(s) && \text{(58b)}\\
&\leq \Delta_{\text{marg}}(\zeta_K) && \text{(58c)}
\end{aligned}$$

3. Finally, as $K\to\infty$, $\zeta_K = \frac{R_{\max}+\overline{V}}{K} \to 0$, and thus, if $\Delta_{\text{marg}}(\zeta)\to 0$ as $\zeta\downarrow 0$, then:

$$\|V^*_\Omega - V^*_K\|_\infty \leq \Delta_{\text{marg}}(\zeta_K) \to 0 \implies V^*_K \to V^*_\Omega \qquad \text{uniformly on } \Omega \tag{59}$$

4. To characterize $\overline{G}(\cdot)$ in this regime, consider the following analysis. For some $q\in\mathcal{Q}_{\text{cfg}}$, let $\mathscr{q} = q|_{\mathcal{I}}$ with $x_i = (s,a,d)\in\mathcal{I}$. Since $q\in\mathcal{Q}_{\text{cfg}}$, we have $\mathcal{T}_\Omega[\mathscr{q}](x_i) = r(x_i)+\gamma\mathbb{E}[V_q(s')]$. Then, using the property $\min\{a,b\} = a - [a-b]^+$, we have:

$$\begin{aligned}
\mathcal{T}_y[q](x_i) &= (1-\gamma)\min\{r(x_i), Kg(s)\} + \gamma\min\{r(x_i)+\mathbb{E}[V_q(s')], y(x_i)\} && \text{(60a)}\\
&= \mathcal{T}_\Omega[\mathscr{q}](x_i) - (1-\gamma)\left[r(x_i) - Kg(s)\right]^+ - \gamma\left[r(x_i)+\mathbb{E}[V_q(s')] - y(x_i)\right]^+ && \text{(60b)}
\end{aligned}$$

Hence, the slow-timescale Bellman residual is given by:

$$G_K(y,q)|_{\mathcal{I}} = \mathcal{T}_\Omega[\mathscr{q}] - \mathscr{q} - e_K(y,q) \tag{61a}$$

$$\text{where,}\quad e_K(y,q) := (1-\gamma)\left[r(x_i) - Kg(s)\right]^+ + \gamma\left[r(x_i)+\mathbb{E}[V_q(s')] - y(x_i)\right]^+ \tag{61b}$$

Averaging over an admissible occupation measure yields:

$$\overline{G}_K(q)|_{\mathcal{I}} = \left\{\mathcal{T}_\Omega[\mathscr{q}] - \mathscr{q} - e^{\mathbb{P}}_K(q) : \mathbb{P}\in\mathfrak{D}_K(q)\right\} \tag{62a}$$

$$\text{with}\quad e^{\mathbb{P}}_K(q) := (1-\gamma)\left[r(x_i) - Kg(s)\right]^+ + \gamma\int_{\Lambda_K(q)}\left[r(x_i)+\mathbb{E}[V_q(s')] - y(x_i)\right]^+\mathbb{P}(dy) \tag{62b}$$

Note that when $\delta_\Omega > 0$ i.e., when $K\geq K^*$, we have $y(x_i)\geq r(x_i)+\mathbb{E}[V_q(s')]$ with $Kg(s)\geq r(x_i)$, yielding $e_K(y,q) = e^{\mathbb{P}}_K(q) = 0$, resulting in $\overline{G}(q)\big|_{\mathcal{I}} = \{\mathcal{T}_\Omega[q|_{\mathcal{I}}] - q|_{\mathcal{I}}\}$.

□

B.8 Extension to Asynchronous settings

In this section, we discuss how our proposed framework extends to asynchronous coordinate updates using the *asynchronous stochastic-approximation* construction of (Perkins & Leslie, 2013). Let $\mathcal{E}_Q$ and $\mathcal{E}_Y$ denote the finite coordinate sets of joint and safety critics; at iteration n, only subsets $E_n^Q \subseteq \mathcal{E}_Q$ and $E_n^Y \subseteq \mathcal{E}_Y$ need to be updated. For each coordinate $x \in \mathcal{S} \times \mathcal{A} \times \mathcal{D}$, let:

$$N_n^Q(x) := \sum_{k=1}^{n} \mathbb{1}_{\{x \in E_k^Q\}}, \quad N_n^Y(x) := \sum_{k=1}^{n} \mathbb{1}_{\{x \in E_k^Y\}} \tag{63}$$

be its local update counters, with corresponding local step sizes $\alpha_{N_n^Q(x)}$ and $\beta_{N_n^Y(x)}$ respectively. We assume the standard asynchronous regularity conditions of (Perkins & Leslie, 2013):

- update process is recurrent so that every coordinate gets a uniformly asymptotic fraction of updates,
- the local step-size sequences satisfy the required regularity and noise conditions,
- the effective two-timescale separation is preserved.

Then, there exists $\epsilon_Q, \epsilon_Y > 0$ such that the *effective relative update rates* belong to:

$$\mathfrak{M}_{\epsilon_Q}^m := \{M = \mathrm{diag}(\lambda_1, \ldots, \lambda_m) : \lambda_i \in [\epsilon_Q, 1]\} \quad \text{with } \epsilon = \epsilon_Q \text{ or } \epsilon_Y \tag{64}$$

Consequently, the limiting mean fields get premultiplied by positive diagonal rescaling matrices. In particular, for a quasi-static q, the fast-timescale differential inclusion becomes:

$$\dot{y}(t) \in \mathfrak{M}_{\epsilon_Y}^{|\mathcal{E}_Y|} \Psi(q, y(t)) := \left\{ M h_\pi : M \in \mathfrak{M}_{\epsilon_Y}^{|\mathcal{E}_Y|}, h_\pi(y(t)) \in \Psi(q, y(t)) \right\} \tag{65}$$

and its globally attracting set replaces $\Lambda(q)$ in the two-timescale analysis. The corresponding slow-timescale limiting field is obtained analogously by averaging the diagonally rescaled Bellman residual over the admissible occupation measures.

Note that this rescaling (due to asynchronity) does not alter any qualitative conclusions made in the synchronous setting. As every diagonal entry is positive, the coordinatewise inequalities retain their sign, with only their contraction rates being reduced at most by the corresponding ϵ_Q, ϵ_Y factors. For instance, Eq. 41b and Eq. 49 become:

$$\mathbb{D}^+ \zeta(y(t)) \leq -\epsilon_Y \, (1-\gamma) \, \zeta(y(t)), \qquad \mathbb{D}^+ \|\mathfrak{q}(t) - Q_\Omega^*\|_\infty \leq -\epsilon_Q (1-\gamma) \|\mathfrak{q}(t) - Q_\Omega^*\|_\infty \tag{66}$$

Hence, the asynchronous extension preserves the limiting safety and optimality results; only the effective convergence rates alter.

C Additional Experimental Details and Results

In this section, we provide more details of Sec. 7, where we choose to answer the following questions: (1) *Does the proposed two-timescale learning framework achieve stable convergence on continuous-control tasks with TD bootstrapping via deep neural approximations?* (2) *How does our joint policy perform against well-established CMDP-based Safe RL benchmarks, particularly concerning its ability to achieve near-zero violations during evaluation without sacrificing task performance?* (3) *How does the set of safe states under the joint policy dynamically expand during the learning process?*

While our theoretical analysis establishes that the joint policy corresponding to the exact fixed point of $Y^* = Q^\phi_{\pi_{Q^*}}, \quad Q^* = \mathcal{T}_{Y^*} Q^*$ *hard-guarantees* safety and *optimal* safety-constrained task performance within the safe set, computing it using exact Dynamic Programming solutions is seldom feasible for high-dimensional continuous state spaces; hence, to scale our framework, we adopt techniques from deep RL to obtain *neural approximations* to our joint value function and policy. In particular, we closely follow the ISAACS framework (Hsu et al., 2023) with *gradient descent-ascent finite timescale separation* (τ-GDA) Fiez & Ratliff (2021) (between the adversarial disturbance and controller), built on the Soft Actor-Critic algorithm (Haarnoja et al., 2018), and call it `JointSAC`.

C.1 Low-dimensional Tabular Experiment

In this subsection, we consider a low-dimensional dynamical system, namely the *double integrator*, and provide *exact* computation of the Joint Value function using tabular Dynamic Programming. **Double-integrator System:** We consider a deterministic double integrator with state $z = (x, v)$ and control a:

$$v^+ = v + \Delta t\, a \qquad x^+ = x + \Delta t\, v^+ \qquad |a| \le 1. \tag{67}$$

The safety margin is defined as $g(x, v) = 1 - |x|$, so the system must remain in $|x| < 1$. The state domain is given by the box $x \in [-2.5, 2.5]$, $v \in [-3, 3]$. The system is rewarded for moving to $x_g = 1.2$, with the reward signal:

$$r(x, v, a) = \exp\left(-\frac{|x - x_g|^2}{\sigma_x^2} - \frac{v^2}{\sigma_v^2} - \lambda_a |a|^2\right) \in (0, 1], \tag{68}$$

where $\sigma_x = 0.8$, $\sigma_v = 1.0$ indicate the *bump width* of the reward signal over $x_g = 1.2$, $v_g = 0$ respectively, and $\lambda_a = 1.0$ denotes the cost on control effort. Thus, the task encourages exiting the right boundary, while safety requires braking early enough to remain inside $|x| < 1$.

Discretization: We use $\Delta x = 0.01$, $\Delta v = 0.05$, $\Delta a = 0.25$, and $\Delta t = 0.2$, yielding 60,379 states and 9 actions. All value functions and policies are computed by full tabular dynamic programming to numerical convergence, with no neural network approximations.

Safe set Ω and admissible actions: $A_\Omega(z)$: For reference, the continuous-time safe set of the disturbance-free double integrator is

$$\Omega_{\mathrm{ct}} = \left\{(x, v) : |x| \le 1, \quad \left| x + \frac{v|v|}{2} \right| \le 1 \right\}. \tag{69}$$

The experiment itself uses the exact kernel Ω_h of the discretized dynamics. It is computed by finite backward reachability:

$$S_0 = z : g(z) > 0, \qquad S_{k+1} = S_k \cap \{z : \exists a \in A,\ f_h(z, a) \in S_k\}. \tag{70}$$

Because the state space is finite, this sequence terminates at its greatest fixed point Ω_h. The admissible action set is then

$$A_{\Omega_h}(z) = \{a \in A : f_h(z, a) \in \Omega_h\}, \tag{71}$$

which is nonempty for every $z \in \Omega_h$ by construction.

Decoupled baselines: We independently compute an unconstrained task critic Q_{task} and a task-oblivious optimal safety critic Q^*_{safe}. We compare against three runtime compositions:

1. A "*hard mask*":

$$\mu_{\mathrm{mask}}(z) \in \operatorname*{arg\,max}_{a : Q^*_{\mathrm{safe}}(z,a) \ge 0} Q_{\mathrm{task}}(z, a). \tag{72}$$

   We verify exhaustively that

$$\{a : Q^*_{\mathrm{safe}}(z, a) \ge 0\} = A_{\Omega_h}(z) \tag{73}$$

   for every viable state-action pair. Hence, any performance difference is caused by the ranking of admissible actions, not by an inaccurate safety critic or a mismatched feasible set.

2. A least-restrictive filter, which executes the unconstrained greedy action when it is admissible and otherwise switches to $\arg\max_a Q^*_{\mathrm{safe}}(z, a)$.
3. A nearest-safe-action, Q-CBF-style projection, which selects the admissible action closest to the unconstrained greedy action.

The oracle value Q^* is obtained by directly running value iteration, restricted to $A_{\Omega_h}(z)$ (unlike the hard mask, where Q_{task} is first obtained).

Computation of the joint value: On the fast timescale (inner loop), we compute Eq. (5) in the paper for a given joint value function (and its joint policy π_{Q_k}) until convergence, and on the slow timescale (outer loop), we compute Eq. (6) in the paper, given the previously converged safety critic Y_k. This implements the two-timescale separation corresponding to $\alpha_n/\beta_n \to 0$.

The following table provides optimality gap characterization (mean, median, max, and violations) for each method, where the optimality gap is defined as

$$\Delta = V^*_{\Omega_h} - V^{\text{method}}(z) : \tag{74}$$

Policy	Mean	Median	Max	Violations
Oracle	0	0	0	0
Hard mask	0.301	0.304	0.507	0
Q-CBF	0.301	0.304	0.507	0
Least-restrictive	0.452	0.442	0.820	0
Joint ($K = 2 \times 10^4$)	0	0	0	0
Joint ($K = 1 \times 10^4$)	0.541	0.5366	3.0394	0
Joint ($K = 5 \times 10^3$)	1.1076	1.1	3.5627	0
Joint ($K = 1 \times 10^3$)	5.6149	5.6476	11.2038	0

We see that an exact runtime composition of independently computed reward and safety critics is already near-optimal. Our joint framework recovers the viability-restricted optimum at a large margin scale (K). The improvement over the hard mask and the least-restrictive filter is likely due to propagating the policy-conditioned safety truncation backward through the Bellman updates; earlier actions are ranked according to their safety-constrained evaluation (conditional on overcoming finite-K conservatism).

C.2 Safe Set Expansion

Since the safety value function V^ϕ_π evaluates the reward-maximizing joint policy π, in theory, the safe set $\left\{s : V^\phi_\pi(s) > 0\right\}$ expands in the direction of maximum task return, before asymptotically reaching the maximal safe set. We demonstrate this on `InvertedPendulum-v4`, rewarding positive cart velocity within ± 0.8m limits. Thus, the agent slides towards the positive x-limit, after which it stays there, balancing the pole. The safe set expansion towards 0.8m during training is illustrated in Fig. 3.

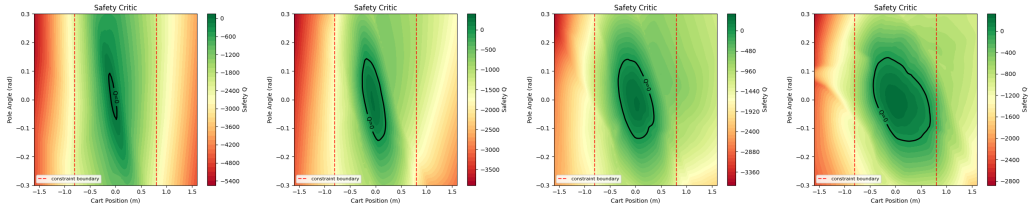

Figure 3: Safe set expands initially in the direction of high rewards during training (`InvertedPendulum-v4` rewarded for positive cart velocity), before converging to the maximal safe set. Constraint boundary (red dotted line) and 0-level set of safety critic (black line) are visualized.

C.3 Continuous Control Experimental Setup

We evaluate our proposed framework across a suite of high-dimensional continuous control tasks implemented in the `MuJoCo` simulator (Todorov et al., 2012), utilizing environments from both `Gymnasium` (Towers et al., 2026) and `SafetyGymnasium` (Ji et al., 2023). The primary objective in all tasks is to maximize returns while strictly satisfying safety constraints at all times.

Simulation Tasks We consider three locomotion environments from `Gymnasium MuJoCo` suite (Towers et al., 2026) and eight single agent environments from the `SafeVelocity` suite from `Safety-Gymnasium`(Ji et al., 2023).

1. `Gymnasium` (Liveness constraints): Here, safety is implicitly defined by the environment's *liveness bounds*, which mandate postural stability to avoid premature episode termination

while moving forward. To rigorously test the safety guarantees of our formulation, we augment the environment with a *physical adversary*. The adversary applies continuous bounded disturbances directly to the agent's torso. Thus, while the controller attempts to safely maximize forward locomotion, the adversary actively seeks to destabilize the agent and prematurely terminate the episode.

2. `SafeVelocity` (Conflicting Objectives): To evaluate our framework under explicitly conflicting reward-cost signals, we utilize the `SafeVelocity` suite of `SafetyGymnasium`. Here, agents receive monotonically increasing rewards for forward velocity, but incur safety violations if they exceed a *predefined safety limit*. As with standard Gymnasium tasks, we subject these agents to adversarial torso disturbances, causing the agent to either prematurely terminate (violating liveness bounds) or push the agent over the speed limits.

Baselines We compare our method against a safety filter baseline (least-restrictive safety filter) and three well-established CMDP baselines. We specifically select these CMDP baselines, as they represent the two dominant paradigms of model-free Safe RL: *Lagrangian dual-descent* and *trust-region projections*:

1. **Safety Filters (`LRSF`)**: Safety filters (Hsu et al., 2024; Fisac et al., 2019) serve as the standard Hamilton-Jacobi-reachability-based baseline, where a *task-oblivious* safety value function (often termed a "safety critic") is learned for an entire *operational design domain* (ODD), where the constraint set (and hence its complement, the failure set) is often encoded by a margin function. Given any *safety-oblivious* task policy, one can query for *the safety* of its actions (by querying the safety-action value function). If the action is deemed unsafe, only then does the *filter intervene* by choosing a greedy action w.r.t. the safety-action value function. Hence, this style of intervention is called *least-restrictive* (LR-safety filters).
2. **Projection-based Constrained Policy Optimization (`PCPO`)**: `PCPO` (Yang et al., 2020) serves as the standard projection-based baseline, employing a two-step trust region framework. It first optimizes the reward advantage function subject to a Kullback-Leibler (KL) divergence constraint, ignoring safety entirely. In the second step, it explicitly projects this intermediate policy onto a feasible safe set. To preserve task performance, this projection minimizes either an L^2 norm or a KL-divergence distance penalty with a second-order Taylor expansion.
3. **PPO-Lagrangian (`PPOLag`)**: Ray et al. (2019); Ji et al. (2024) combine the Lagrangian multiplier approach with Proximal Policy Optimization (PPO) to obtain `PPOLag`. It enforces safety bounds using dynamic penalty coefficients, solving the equivalent min-max optimization problem via alternating gradient ascent on the reward objective and dual gradient descent on the constraint violation costs.
4. **SAC-Lagrangian (`SACLag`)** : This baseline (`SACLag`) (Ray et al., 2019; Ji et al., 2024) is conceptually identical to `PPOLag`, but replaces the underlying architecture with Soft-Actor Critic (Haarnoja et al., 2018).

Robust Adversarial Integration: To incorporate adversarial disturbances, we modify all baselines into adversarial *zero-sum* formulations using the principles of Robust Adversarial RL (Pinto et al., 2017). We introduce a secondary adversarial disturbance $\nu(s', a')$ that can observe the controller's actions $a' \in \mu(s')$ and injects *adversarial* bounded perturbations into the environment dynamics. This yields robust versions of all three baselines, where the controller seeks to maximize the dynamic Lagrangian objective (maximize rewards, minimize violation costs), while the adversarial disturbance actively attacks and minimizes it.

Metrics We use cumulative *Episodic Return* and *Episodic Cost* tracked across training to evaluate stable convergence. We also use *cumulative reward* and *episode cost* as the evaluation metrics. For Lagrangian-based methods, a permissive cost violation threshold of $d = 0$ is allowed. We take safety as the primary criterion for evaluations, and pursue higher rewards on the basis of meeting safety requirements.

Compute We train our algorithms on a `slurm` cluster with a single Nvidia L40, 48GB of VRAM. Each run takes 4 CPU cores. We implement the code in `PyTorch`, and all experiments take up to 4 days to fully train for 5M timesteps.

Hyperparameters

- We choose seeds 5 training seeds: $[0, 5, 10, 15, 20]$ as the algorithms in `SafetyGymnasium` have been reported to be stable on these seeds.
- We tune the scaling of the margin function $\beta(g(s)) = K \cdot g(s)$ based on the maximum task return possible for each environment. Additionally, we pick the maximum value of K such that there are no safety violations / algorithmic instabilities. Specifically, we choose the following:

  1. `Hopper-v4`: 800.0
  2. `Walker2d-v4`: 1200.0
  3. `InvertedPendulum-v4`: 200.0
  4. `SafetyWalker2dVelocity`: 1200.0
  5. `SafetyHopperVelocity`: 800.0
  6. `SafetyHalfCheetahVelocity`: 1200.0
  7. `SafetyAntVelocity`: 1200.0

  We pick the disturbance magnitude (applied to the torso) that is meaningfully challenging to the agent. Any higher value resulted in high safety violations for all algorithms:

  1. `InvertedPendulum-v4`: 5.0
  2. `Walker2d-v4`: 30.0
  3. `Hopper2d-v4`: 7.5
  4. `SafetyWalker2dVelocity`: 25.0
  5. `SafetyHopperVelocity`: 5.0
  6. `SafetyHalfCheetahVelocity`: 20.0
  7. `SafetyAntVelocity`: 2.5

- For For `JointSAC`, we use a safety-critic learning rate of 1.0×10^{-3}, a joint-critic learning rate of 5×10^{-4}, a controller actor learning rate of 1×10^{-4}, a disturbance learning rate of 5×10^{-4}, and autotune alpha (entropy term) with a learning rate of 5×10^{-4}. We implement timescale separation through the critic learning rates: the joint critic is updated at half the learning rate of the safety critic, with both critics updated at every step (this led to stable training while completing in reasonable amount of time, requiring around 5M training steps; *polyak averaging* ($\tau = 5 \times 10^{-3}$) is used only for the target networks of both critics. Finally, we use different discount factors γ for the joint Bellman equation and $\widetilde{\gamma}$ for the safety Bellman equation; we anneal $\widetilde{\gamma}$ from 0.85 to $\gamma = 0.99$ while training progresses, as discussed in (Fisac et al., 2019).

C.4 Margin Scale K Variation

We perform an ablation experiment on the variation of episode rewards and cost as K increases progressively (in steps of 200, from $K = 200$ to $K = 2000$). Results are illustrated in Fig. 4 for 4 environments - `Hopper, Walker2d` (from `Gymnasium` suite), and `SafetyHalfCheetah, SafetyWalker2d` (from `SafetyGymnasium` suite). Training was conducted on 5 seeds, and each training seed was evaluated for 300 unique episodes, totaling $1,500$ unique evaluation episodes.

The general trend agrees with our previous discussion in Sections 4-6: scaling K *generally improves* episode reward (as the region Ω^{ψ} increases), while safety is *generally preserved*, with small variations. However, one can also observe that increasing K too much may lead to *training instabilities* (as seen in `Walker2d` environment).

C.5 Additional Results

The training curves for 5 seeds on:

- three locomotion tasks from `Gymnasium` suite incorporating adversarial disturbance is presented in Fig. 5
- eight single agent environments from `SafeVelocity (v0)` suite from `Safety-Gymnasium` incorporating adversarial disturbance is presented in Fig. 6

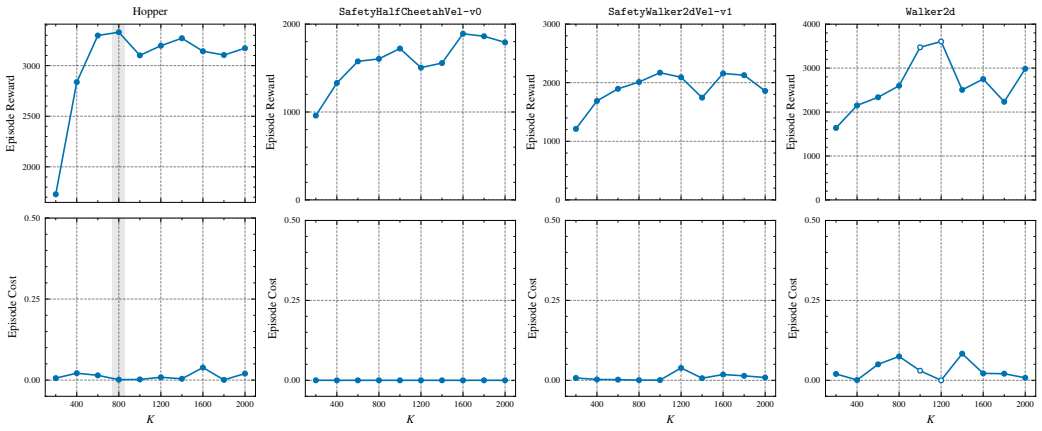

Figure 4: Figure shows variation of episode reward and episode cost as K increases from 200 to 2000. While episode reward generally tends to increase with an increase in K, scaling beyond a point can introduce *training instabilities* with neural approximations, as seen with `Walker2d` environment.

These results align with the results presented in Sec. 7.2. Throughout training, `JointSAC` demonstrates stable convergence to high task rewards, while consistently incurring near-zero episode costs. A single cost violation incurs $+1$ penalty each timestep. The baseline methods exhibit higher variance, with only a subset of them reliably converging below the permissible cost threshold ($d = 0$).

For evaluations, we perform 300 rollout evaluations for each of the 5 training seeds (totaling $1,500$ distinct episodes per environment), with results summarized in Table 1. Results indicate that `JointSAC` maintains near-zero costs across all 11 environments (maximum of 0.03 ± 0.08 on the `Hopper-v4` adversarial environment), while the baseline exhibits varied cost violations, occasionally achieving high episode costs for some environments and staying below the permissible cost violation for others. Overall, it is seen that `JointSAC` yields stable performance while maintaining near-zero violations.

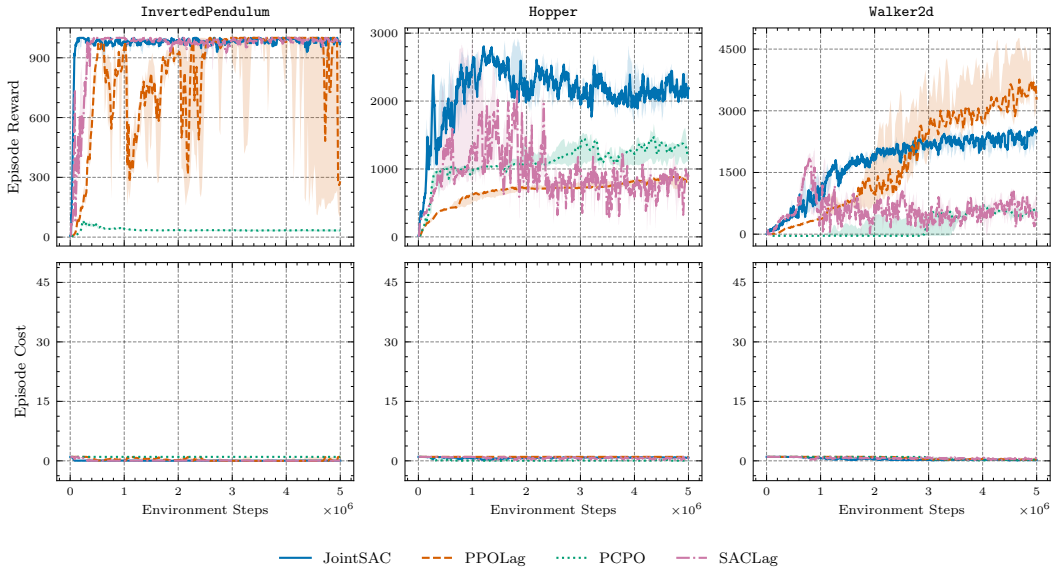

Figure 5: Training performance across 5 seeds (mean $\pm$ standard deviation) on three locomotion tasks from `Gymnasium` suite with adversarial disturbances. The top row indicates episodic returns, while the bottom row indicates episodic costs as training progresses. `JointSAC` demonstrates stable convergence with near-zero episode costs on two out of three tasks. A single cost violation incurs +1 penalty each timestep.

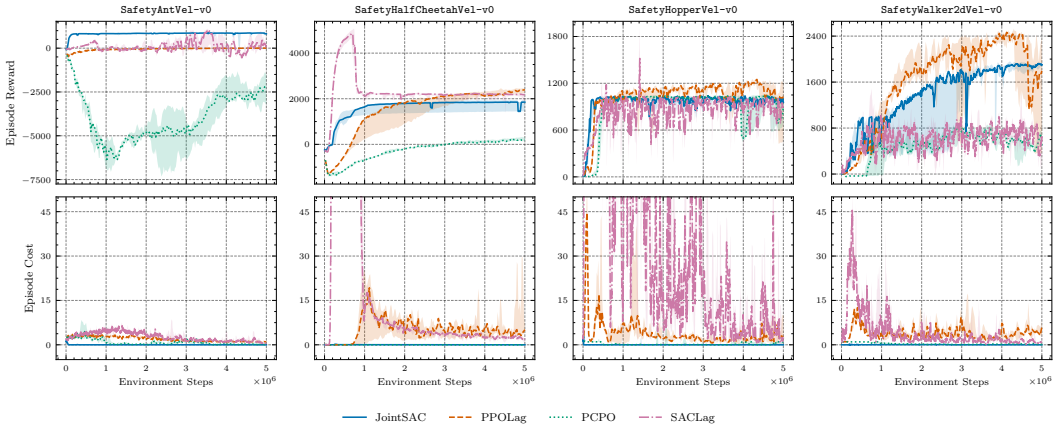

Figure 6: Training performance across 5 seeds (mean ± standard deviation) on eight single agent environments from `SafeVelocity` (`v0`) suite from `SafetyGymnasium` with adversarial disturbances. The top two rows indicate episodic returns, while the bottom two rows indicate episodic costs as training progresses. `JointSAC` demonstrates stable convergence with near-zero episode costs.