

Safe Reinforcement Learning via Hierarchical Adaptive Chance-Constraint Safeguards

Zhaorun Chen¹, Zhuokai Zhao¹, Tairan He², Binhao Chen³, Xuhao Zhao³, Liang Gong³ and Chengliang Liu³

Abstract—Ensuring safety in Reinforcement Learning (RL), typically framed as a Constrained Markov Decision Process (CMDP), is crucial for real-world exploration applications. Current approaches in handling CMDP struggle to balance optimality and feasibility, as direct optimization methods cannot ensure state-wise in-training safety, and projection-based methods correct actions inefficiently through lengthy iterations. To address these challenges, we propose Adaptive Chance-constrained Safeguards (ACS), an adaptive, model-free safe RL algorithm using the safety recovery rate as a surrogate chance constraint to iteratively ensure safety during exploration and after achieving convergence. Theoretical analysis indicates that the relaxed probabilistic constraint sufficiently guarantees forward invariance to the safe set. And extensive experiments conducted on both simulated and real-world safety-critical tasks demonstrate its effectiveness in enforcing safety (nearly zero-violation) while preserving optimality (+23.8%), robustness, and fast response in stochastic real-world settings.

I. INTRODUCTION

Reinforcement learning (RL) has demonstrated remarkable success in handling nonlinear stochastic control problems with large uncertainties [5], [21]. Although solving unconstrained optimization problems in simulations incurs no safety concerns, ensuring safety during training is crucial in real-world applications [18]. However, the inclusion of safety constraints in RL is non-trivial. First, safety and goal objectives are often competitive [4]. Second, the constrained state-space is usually non-convex [42]. And third, ensuring safety via iterative action corrections is time-consuming and impractical for safety-critical tasks requiring fast responses.

Numerous studies strive to address these challenges, enhancing safety assurances in RL. Early works utilized trust-region [1] and fixed penalty methods [7] to enforce cumulative constraints satisfaction in expectation [24], [36]. However, these methods are sensitive to hyperparameters and often result in policies being either too aggressive or too conservative [26]. Furthermore, these methods only ensure safe behaviors asymptotically upon the completion of training, resulting in a gap in safety assurance during the training exploration process [6].

Recognizing the limitations of above methods in achieving immediate safety during the training phase, many works seek to employ hierarchical agents to project task-oriented

action into a prior [38], or learned safe region [43] to satisfy state-wise constraints. However, these methods usually make additional assumptions like white-box system dynamics [9] or default safe controller [22], which is not always available and strict feasibility is not guaranteed [42]. Besides, applying step-wise projection to satisfy instantaneous hard constraints is often time-consuming, which hampers the optimality of the task objectives [46].

Alternatively, integrating safety into RL via chance or probabilistic constraints, which involves unfolding predictions to estimate safety probability for future states, presents a compelling advantage. Chance constraints have received significant attention within both safe control [20], [40] and RL communities [6], [28], [48]. However, while these methods mark a step forward, they often fall short in efficiently enforcing such constraints [6], [14] or necessitate the use of an independent safety probability estimator [17].

Motivated by these challenges, we propose Adaptive Chance-constrained Safeguards (ACS), an efficient model-free safe RL algorithm that models safety recovery rate as a surrogate chance constraint to adaptively guarantee safety in exploration and after convergence. Unlike existing work [28] that approximates the safety critic through lengthy Monte Carlo sampling, or [6] that learns a conservative policy with relaxed upper-bounds in conservative Q-learning [23], ACS directly constrains the safety advantage critics, which can be interpreted as *safety recovery rate*. We show theoretically in §IV that this is a sufficient condition to certify in-training safety convergence in expectation. The introduction of recovery rate mitigates the objective trade-off commonly encountered in safe RL [18] by encouraging agents to explore risky states with more confidence, while enforcing strict recovery to the desired safety threshold. We also validate empirically in §V that ACS can find a near-optimal policy in tasks with stochastic moving obstacles where almost all other state-of-the-art (SOTA) algorithms fail.

To summarize, the contributions of this paper include: (1) proposing adaptive chance-constrained safeguards (ACS), an advantage-based algorithm mitigating exploration-safety trade-offs with surrogate probabilistic constraints that theoretically certifies safety recovery; (2) extensive experiments on various simulated safety-critical tasks demonstrating that ACS not only achieves superior safety performance (nearly zero in-training violation), but also surpasses SOTA methods in cumulative reward and time efficiency with a significant increase ($23.8\% \pm 10\%$); and (3) two real-world manipulation experiments showing that ACS boosts the success rate by 30% and reduces safety violations by 65%, while

¹Zhaorun Chen and Zhuokai Zhao are with the Department of Computer Science, University of Chicago, USA. {zhaorun, zhuokai}@uchicago.edu

²Tairan He is with the Robotics Institute, Carnegie Mellon University, USA. tairanh@andrew.cmu.edu.

³Binhao Chen, Xuhao Zhao, Liang Gong and Chengliang Liu are with the Department of Mechanical Engineering, Shanghai Jiao Tong University {cbh_mage, rachmaninov, gongliang_mi, chlilu}@sjtu.edu.cn

requiring fewer iterations compared to existing methods.

II. RELATED WORK

A. Safe RL

Safe RL focuses on algorithms that can learn optimal behaviors while ensuring safety constraints are met during both the training and deployment phases [18], [35]. Existing safe RL methods can be generally divided into three categories:

1) End-to-end

End-to-end agent augments task objective with safety cost and solves unconstrained optimization (or its dual problem) directly thereafter [24], [36]. For example, [12] augments safety constraint as L_2 regularization; [27] adopts an additional network to approximate the Lagrange multiplier; and [19] searches for intrinsic cost to achieve zero-violation performance. However, the resulting policies are only asymptotically safe and lack in-training safety assurance, while the final convergence on safety constraints is not guaranteed [8].

2) Direct policy optimization (DPO)

Instead of augmenting safety cost into the reward function, DPO methods such as [1] leverage trust-regions to update task policy inside the feasible region. More specifically, [42] refines the sampling distribution by solving a constrained cross-entropy problem. And [44] confines the safe region via a convex approximation to the surrogate constraints with first-order Taylor expansion. However, these methods are usually inefficient and are prone to being overly conservative.

3) Projection-based methods

To ensure strict certification in safety, recent work leverages a hierarchical safeguard/shield [2] to project unsafe actions into the safe set. Projection can be conducted by various approaches, including iterative-sampling [6], gradient-descent [43], quadratic-programming (QP) [10], and control-barrier-function [9]. More specifically, [6] proposes an upper-bounded safety critic via CQL [23] and iteratively collects samples until a conservative action that satisfies the safety constraint is found. However, iterative sampling is not time-efficient for safety-critical tasks that require immediate responses. Similarly, [16], [33] conduct black-box reachability analysis to iteratively search for entrance to the feasible set. On the other hand, safety layer method [11] parameterizes the system dynamics and solves the QP problem with the learned dynamics. However, we show in §V that these methods fail in tasks with complex cost functions. Some other methods seek to achieve zero-violation with a hand-crafted energy function, such as ISSA [47], RL-CBF [41], and ShieldNN [15]. However, these methods require prior knowledge of the task, which is intractable for general model-free RL case. Compared with existing SOTA methods, our proposed method ACS, effectively addresses and surpasses them by tackling two major challenges in this field: (1) balancing the trade-off between task optimality and safety feasibility; and (2) efficiently conducting projection.

B. Chance-Constrained Safe Control

A tentative approach for better trading-off between objectives is through unrolling future predictions, and deriving a surrogate chance-constraint to prevent future safety violations [40]. However, the inclusion of chance-constraints in

control optimization is non-trivial. [17] proposes to linearize the chance-constraint into a myopic controller to guarantee long-term safety. However, their method requires a refined system model and an extra differentiable safe probability estimator. In model-free RL, state-wise safety probability depends on the policy and can thus be approximated with a critic [34]. Early works [28], [42] approximate this critic through Monte Carlo sampling, which are lengthy and slow.

To further address the issue, [6] approximates the critic via CQL [23] and learns a conservative policy via iterative sampling. Similarly, [14] proposes to learn an ensemble of critics and train both a forward task-oriented policy and a reset goal-conditioned policy that kicks in when the agent is in an unsafe state. Given that resetting is not always necessary and efficient, [37] proposes to learn a dedicated policy that recovers unsafe states. However, these approaches require an additional recovery policy and are prone to being overly conservative. In this paper, we show that using only one advantage-based safeguarded policy can achieve comparable recovery capability, while maintaining high efficiency.

III. PRELIMINARIES AND PROBLEM FORMULATION

A. Markov Decision Process with Safety Constraint

Let $x_k \in \mathcal{X} \subset \mathbb{R}^{n_x}$, $u_k \in \mathcal{U} \subset \mathbb{R}^{n_u}$ be the discrete sample of system state and control input of continuous time t (i.e., $t = k\Delta t$), where n_x and n_u are the dimension of the state space \mathcal{X} and control space \mathcal{U} , the partially observable system with stochastic disturbances can be essentially represented by a probability distribution, that is:

$$x_{k+1} = \mathbf{F}(x_k, u_k, \hat{\epsilon}) + w_k \Rightarrow x_{k+1} \sim P(x_{k+1}|x_k, u_k) \quad (1)$$

where \mathbf{F} denotes the system dynamics under parametric uncertainty, and $\hat{\epsilon}$, w_k denote the parametric uncertainties and the additive disturbance of the system respectively. RL policy seeks to maximize rewards in an infinite-horizon Constrained Markov Decision Process (CMDP) [3], which can be specified by a tuple $(\mathcal{X}, \mathcal{U}, \gamma, R, P)$, where $R : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$ is the reward function, $0 \leq \gamma \leq 1$ is the discount factor and $P : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \rightarrow [0, 1]$ is the system state transition probability function defined in Eq. (1). Therefore, the safe RL problem can be formulated as

$$\arg \max_{\pi_\theta} J(\pi) = \mathbb{E}_{x \sim P, u \sim \pi_\theta} \left[\sum_{k=0}^{\infty} \gamma^k R(x_k, u_k) \right] \quad (2a)$$

$$\text{s.t. } \pi_\theta \in \Pi_C \quad (2b)$$

$$\Pi_C = \{\pi \in \Pi \mid \forall u_k \sim \pi_\theta, x_k \in \mathcal{S}_C\} \quad (2c)$$

$$\mathcal{S}_C = \{x \mid \forall i, J_{C_i}(x) \leq d_i\} \quad (2d)$$

where Π denotes the set of all stationary policies, $\Pi_C \subset \Pi$ represents the set of *feasible* policies that satisfy all safety constraints $C = \{C_0, C_1, \dots, C_n\}$. Accordingly, J_{C_i} denotes the cost measure with respect to one specific safety constraint $C_i \in C$, and is evaluated as $J_{C_i} = \mathbb{E}_{\tau \sim \pi} [\sum_{k=0}^{\infty} \gamma^k C_i(x_k, u_k, x_{k+1})]$, where τ is a trajectory (i.e., $\tau = \{x_0, u_0, x_1, u_1, \dots\}$) resulted from the policy π . Finally, \mathcal{S}_C represents the set of *safe* states, where each state satisfies the i^{th} safety constraints $C_i \in C$ through not exceeding its corresponding permitted threshold d_i .

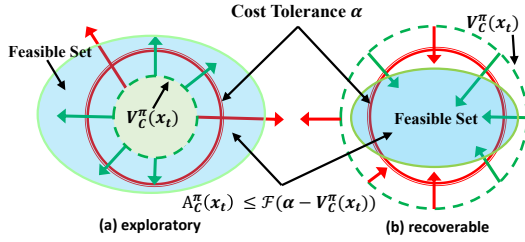


Fig. 1. The proposed adaptive chance constraint. The green-dashed and red circle denotes the current safety cost and unified cost tolerance level respectively. The blue oval denotes the adaptive chance-constrained feasible set. Green/red arrows denote feasible/infeasible actions. When current cost $V_C^\pi(x_k)$ is within tolerance, the agent is encouraged to explore more risky states. Otherwise, the next action is constrained in a more conservative set which satisfies Eq. (6a), so that long-term safety recovery is certified.

B. Chance-constrained Safety Probability

However, enforcing all the states in a trajectory to stay within \mathcal{S}_C as defined in Eq. (2d) is impractical in real-world settings. Because, first, the environment is stochastic with large uncertainties, which is impossible to satisfy Eq. (2b) all the time. Second, the penalty feedback induced by the final disastrous behavior is often sparse and delayed. Thus solely abiding by these constraints will result in myopic, unrecoverable policies [40]. Therefore, we propose to instead consider a chance constraint [17] by unrolling future predictions and ensuring $x_k \in \mathcal{S}_C$ during an outlook time window $\mathcal{T}(k) = \{k, k+1, \dots\}$ with probability $1 - \alpha$. Precisely, we define the safety probability Ψ of a single state x_k as

$$\Psi(x_k) := P\left(\bigcap_{j \in \mathcal{T}(k)} x_j \in \mathcal{S}_C\right) \geq 1 - \alpha \quad (3)$$

where α represents the tolerance level of the unsafe event.

To evaluate the constraint in Eq. (3), we draw from the Bellman equation [35] and approximate the expected long-term chance-constrained safety probability Ψ with a value network $V_C^\pi(x_k)$ that is learned through trials and errors. The associated theorem (Theorem VI.1) and its proof are detailed in Appendix VI-A.

While other works may evaluate the state-action value $Q_C^\pi(x_k, u_k)$ [6], [43], in this paper, we follow [39] and approximate the advantage $A_C^\pi(x_k, u_k)$ of control u_k with a critic network, since it could better adjust with the change of safety probability, as will be shown later in §IV-A.

IV. ADAPTIVE CHANCE-CONSTRAINED SAFEGUARDS

In this section, we illustrate the proposed advantage-based chance-constraint safeguard (ACS), which derives a relaxed constraint on RL exploration to achieve better task-oriented performance while theoretically guaranteeing recovery to the safe region. The schematic of ACS is shown in Fig. 1.

A. Learning to Recover

As mentioned earlier in §III-B, strictly guaranteeing Eq. (2d) is neither practical nor necessary for real-world settings. To help mitigate the difficulty in specifying a rational safety boundary that balances between task and safety objectives, we relax the safety chance constraint following [40] and propose a plug-and-play sufficient condition of the safe recovery in ACS for generic RL controllers. First, we define a discrete-time generator G for any $x_k \in \mathcal{X} \subset \mathbb{R}^{n_x}$, which

can be considered as a stochastic process [13] taking form:

$$G\Psi(x_k) = \mathbb{E}[\Psi(x_{k+1}) | x_k, \pi_\theta(x_k)] - \Psi(x_k) \quad (4)$$

where $\pi_\theta(x_k)$ is also conditioned on state x_k . Essentially, Eq. (4) captures the expected improvement or degeneration of safety probability Ψ as the stochastic process proceeds. However, rather than imposing constraints directly on $\Psi(x_k)$ like in [6], [43], we propose to apply the chance constraint on the generator output (recovery rate of $\Psi(x_k)$), that is:

$$G\Psi(x_k) \geq -\mathcal{F}(\Psi(x_k) - (1 - \alpha)) \quad (5)$$

where $\mathcal{F}(p)$ can be any concave function that is upper-bounded by p . Consequently, Eq. (5) defines a lower bound for its recovery rate, such that when $\Psi(x_k) \leq (1 - \alpha)$ (i.e. the safety assurance is compromised), the controller is enforced to recover to safety at rate $G\Psi(x_k)$. Otherwise, the controller is free to explore to achieve better task-oriented performance, which balances the trade-off between safety and optimality.

More specifically, the following chance constraint is proposed in ACS framework to certify in-training safety, which is specified in Theorem IV.1. The detailed proof of the theorem is illustrated in Appendix VI-B.

Theorem IV.1. Let $A_C^{\pi_\theta}(x_k, u_k)$ denote the advantage function of control u_k at x_k , the sufficient condition that can ensure asymptotic safety satisfaction both in training and after convergence is

$$A_{C_i}^{\pi_\theta}(x_k, u_k) \leq \mathcal{F}_i(\alpha_i - V_{C_i}^{\pi_\theta}(x_k)) \quad (6a)$$

$$\text{s.t. } H(\mathcal{F}_i(q)) \preceq 0 \text{ and } \mathcal{F}_i(q) \leq q \quad (6b)$$

where $H(\mathcal{F}_i(q))$ is the Hessian of $\mathcal{F}_i(q)$, and C_i, α_i denotes the cost function and the tolerance level of the i^{th} safety constraint, respectively.

To evaluate the constraint Eq. (6a) in implementation, we construct a fully-connected multi-layer network for both the value and advantage function separately based on [29]¹. These two networks are updated iteratively together with the task policy in RL exploration through trials-and-errors.

B. Hierarchical Safeguarded Controller

In practice, the actual implementation of ACS is illustrated in Fig. 2. To strictly enforce the safety chance constraint, we evaluate the proposed control action at every time step and project it into the safe action set [46]. Similar to [12], [43], we adopt a hierarchical architecture in which the upper policy first solves for a sub-optimal action and iteratively corrects it to satisfy the chance constraint in Eq. (6a). Specifically, we employ the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method [25] to enforce the feasibility of actions provided by the safety-aware policy approximator while preserving time efficiency. Note that L-BFGS is theoretically promised to converge much faster to the safety region starting from a sub-optimal solution embedded in a locally-convex space [25], as compared to other gradient-descent methods [43]. Hence the implementation framework of ACS is consisted of the following two sub-modules.

¹<https://github.com/DLR-RM/stable-baselines3>

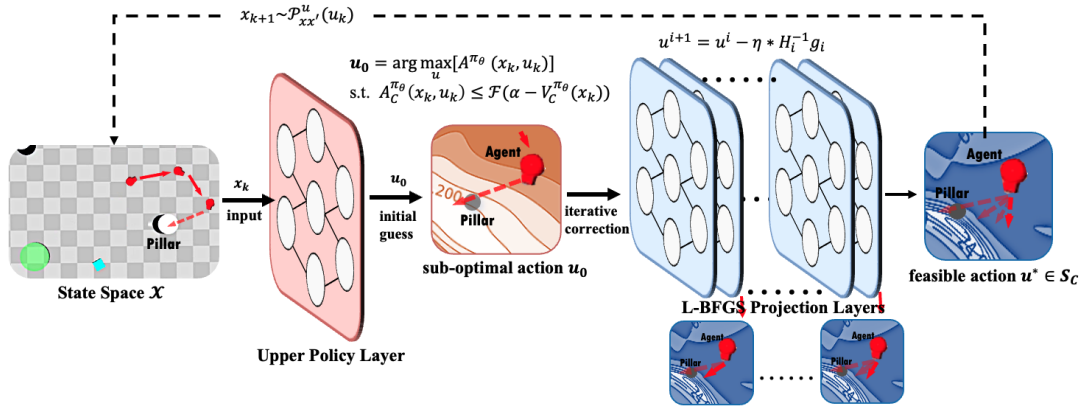


Fig. 2. The hierarchical framework of the proposed ACS. A Lagrangian-based upper policy layer first generates a near-optimal initial action u_0 by solving Eq. (7), then the quasi-newton-based projection layers iteratively correct it into the safe set that satisfies Eq. (6a) via efficient back-propagation Eq. (8), enabling ACS to balance task objective and certified safety by constraining actions in an adaptive feasible set while ensuring immediate response.

Sub-optimal policy layer. For the upper policy layer, we follow [36] and train a policy optimizer to solve task objective Eq. (2a) with augmented penalty of the chance constraint Eq. (6a) weighted by a Lagrangian multiplier λ :

$$\max_{\theta} \min_{\lambda} \mathcal{L}(\pi_{\theta}, \gamma) = \mathbb{E}_{\tau \sim \pi_{\theta}} [A_C^{\pi_{\theta}} - \lambda(A_C^{\pi_{\theta}} - \mathcal{F}(\alpha - V_C^{\pi_{\theta}}))] \quad (7)$$

Note that the goal here is to find a safety-aware policy that can produce a sub-optimal initial guess u_0 with respect to both task objective and safety constraint. While extensive existing works solve various forms of Eq. (7) [12], [27], [43], we focus on the sub-optimality of the initial guess rather than strictly solving the cumulative-constrained MDP. Thus, any model-free safety-aware RL algorithm, such as [1], [6], can serve as the policy solver of our hierarchical framework. We show empirically in §V that simply optimizing Eq. (7) is efficient enough to find a near-optimal solution via ACS.

Fast ACS projection. The initial guess produced by the upper policy layer does not strictly satisfy Eq. (6a), given that Eq. (7) optimizes cumulative costs and λ is hard to tune. Therefore, to strictly satisfy constraint in Eq. (6a), we propose to employ L-BFGS [25], an efficient Quasi-Newton method to iteratively correct u_0 . L-BFGS is a memory-efficient method that updates control action by:

$$u^{k+1} = u^k - \eta * H_k^{-1} g_k \quad (8)$$

where η is the learning rate, $g_k = \frac{\delta}{\delta u^k} [A_C^{\pi_{\theta}} - \mathcal{F}(\alpha - V_C^{\pi_{\theta}})]$ is the gradient vector, and H_k^{-1} is the approximation of the inversed Hessian matrix. BFGS finds a better projection axis and step-length through approximating an additional second-order Hessian inverse. We show in §V that ACS can recover safety in few steps, even when against an adversarial policy.

We claim L-BFGS to be a better projection strategy in ACS for two reasons: 1) safety-critical tasks usually require immediate response, and BFGS methods exhibit faster convergence rate by trading space for time; 2) the optimal solution is embedded in the locally-convex sub-optimal region found by Eq. (7) [12], making it possible to correct the action in one step. In addition, it is insensitive to η [25], which differentiates from existing work [43] requiring tedious hyperparameters (e.g., learning rate) tuning.

V. EXPERIMENT

To thoroughly assess the effectiveness of ACS, we conduct experiments on both simulated and real-world safety-critical tasks. During simulations, we test on one speed-planning problem, *Ant-Run*, in addition to three manipulation tasks, including *Kuka-Reach*, *Kuka-Pick* and *InMoov-Stretch*. In real-world scenarios, we implement the real-world versions of *Kuka-Pick* and *InMoov-Stretch*, which share the same tasks and safety specifications as their simulated counterparts, to further assess ACS’s robustness in real-world settings.

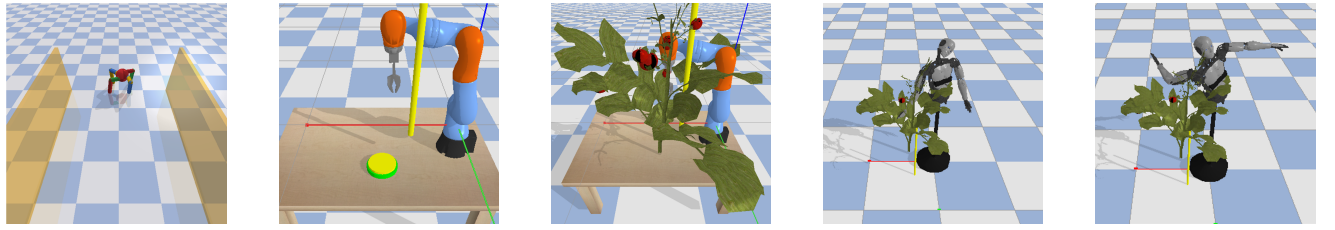
A. Experimental Setup

Simulated tasks. As shown in Fig. 3, four tasks are designed to evaluate ACS along with other safe RL methods. We have:

- *Ant-Run*: utilizes a simple quadrupedal robot to assess both effectiveness and time efficiency by constraining the robot to run within a certain velocity limit.
- *Kuka-Reach*: employs a 7-DOF Kuka robot arm to assess the effectiveness of ACS during human-robot interaction. As shown in Fig. 3(b), the robot arm is trying to reach for a table button while avoiding collision with a static cylinder (yellow in the figure) which represents the human.
- *Kuka-Pick*: extends *Kuka-Reach* to include dynamic obstacles where, as shown in Fig. 3(c), the robot arm seeks to pick a tomato on the tree while avoiding collisions with a moving cylinder (yellow) which represents moving human and other tomatoes (static obstacles) on the tree.
- *InMoov-Stretch*: evaluates high-dimensional control generalizability and robustness. It utilizes an *InMoov* humanoid which has 53 actively controllable joints [45]. As shown in Fig. 3(d), the robot is trying to reach for the fruit on the tree with a human-like arm stretch.

Real-world tasks. To assess the robustness of ACS on real-world control tasks, we convert *Kuka-Pick* and *InMoov-Stretch* to their real-world counterparts through point-cloud reconstruction and key-point matching [45]. The task and safety specifications remain identical as in the simulations. Examples from successful ACS episode runs of initial and end joint states for both tasks are shown in Fig. 4.

Baselines. Six SOTA safe RL approaches, including Safety Layer [11], Proximal Policy Optimization (PPO) [32], Re-



(a) Ant-Run: a quadrupedal robot seeks to run within a velocity limit. (b) Kuka-Reach: a 7-DOF manipulator Kuka navigates to the button collision-free. (c) Kuka-Pick: Kuka picks up a fruit while not bumping into a moving cylinder. (d) InMoov-Stretch: a humanoid InMoov reaches a fruit with a natural posture. (e) An ill-formed morphology to reach the fruit while violating safety constraint.

Fig. 3. (a)-(d): Four simulated safe-critical tasks where we assess five safe RL algorithms; (e): An illustration of safety constraint violation.



(a) Initial and end Kuka arm positions reaching the target while avoiding the cylinder obstacle in-between.

(b) Initial and end InMoov arm postures searching for a natural arm stretch trajectory to reach the target.

Fig. 4. The initial and end pose of the robots in real-world *Kuka-Pick* and *InMoov-Stretch*.

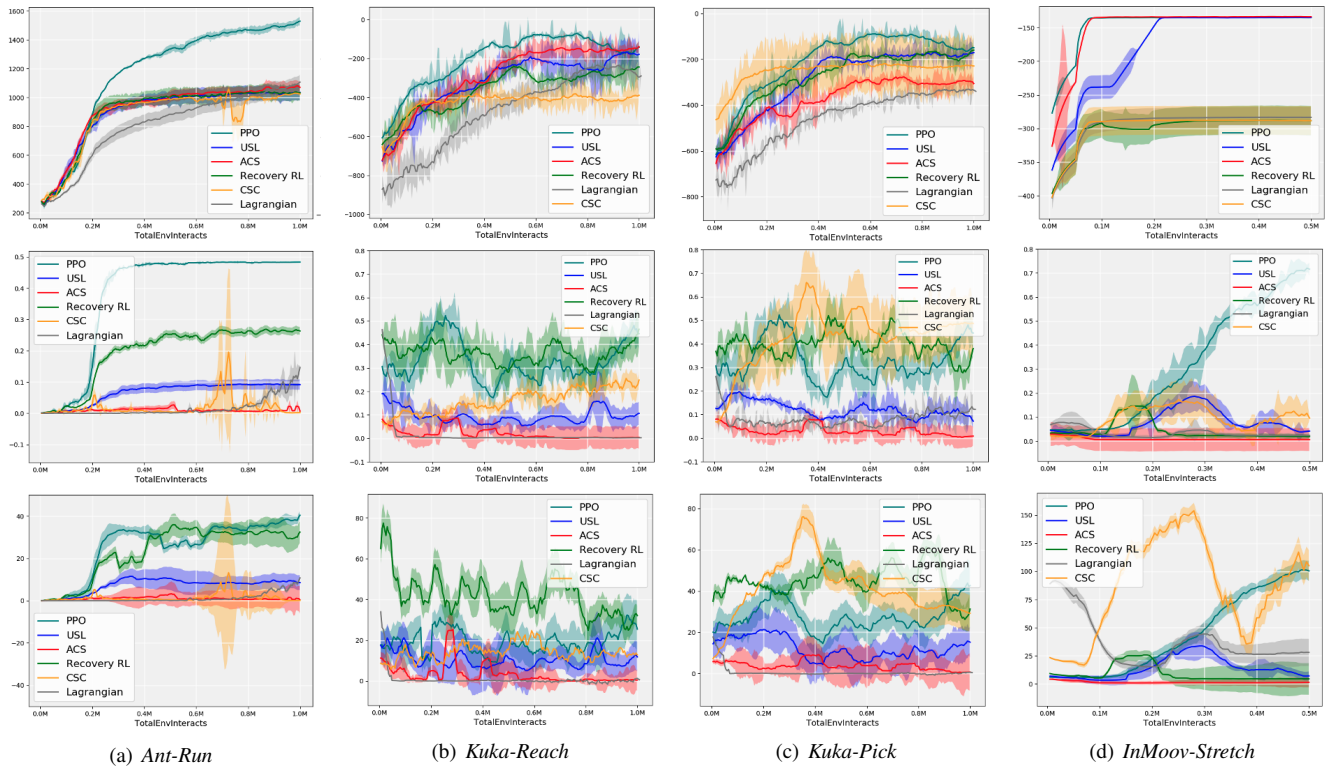


Fig. 5. In-training curves of episodic return J_r (top row), total cost rate J_C (middle row), and temporal safety cost rate J_{TC} (bottom row) w.r.t. the number of interactions of different algorithms on four safety-critical simulation tasks.

covery RL [37], Conservative Safe Critics (CSC) [6] Unrolling Safety Layer (USL) [43], and PPO-Lagrangian [30], are included in our experiments. Official implementations and recommended parameters are used for each method.

Metrics. We first consider common metrics [30] such as episodic return J_r , total cost rate J_C , success rate, average number of collisions and average number of iterations for

iterative methods such as USL, CSC, and ACS. Besides, we argue that episodic inference time is also a critical dimension to consider coupled with safety performance. Therefore, we propose a novel metric, *temporal cost rate*, which measures the safety performance along with forward inference time. Specifically, we have $J_{TC} = \frac{\text{accumulated cost}}{\text{length of the episode}} * \bar{t}_{\text{forward}}$, where lower values indicate better and faster safety performance.

TABLE I

MEAN PERFORMANCE OF 20 EPISODES AT CONVERGENCE ON FOUR SAFETY-CRITICAL SIMULATION TASKS.

Task	Safety Layer	PPO	Recovery RL	CSC	USL	PPO-Lagrangian	ACS (Ours)	
<i>Ant-run</i>	$J_r \uparrow$	933.2 \pm 12.98	1512.7 \pm 32.10	1028.1 \pm 87.05	1024.7 \pm 11.68	981.7 \pm 26.96	1103.7 \pm 50.10	
	$J_C \downarrow$	0.087 \pm 0.01	0.48 \pm 0.001	0.26 \pm 0.00	0.002 \pm 0.00	0.09 \pm 0.01	0.016 \pm 0.00	
	$J_{TC} \downarrow$	1.3e-3	9e-4	1e-3	9.6e-3	9e-3	8e-4	6e-3
<i>Kuka-Reach</i>	$J_r \uparrow$	-256.9 \pm 116.8	-141.8 \pm 80.7	-243.1 \pm 145.10	-388.3 \pm 114.5	-178.7 \pm 142.4	-289.1 \pm 106.2	-138.4 \pm 60.89
	$J_C \downarrow$	0.24 \pm 0.03	0.46 \pm 0.06	0.44 \pm 0.15	0.22 \pm 0.05	0.05 \pm 0.02	0.08 \pm 0.001	0.0032 \pm 0.04
	$J_{TC} \downarrow$	5.5e-3	1e-3	6e-3	5.4e-3	1.4e-2	1e-3	1e-2
<i>Kuka-Pick</i>	$J_r \uparrow$	-183.5 \pm 51.6	-149.7 \pm 81.2	-159.6 \pm 45.4	-229.0 \pm 247.3	-170.6 \pm 66.8	-339.7 \pm 64.0	-306.1 \pm 57.9
	$J_C \downarrow$	0.29 \pm 0.03	0.47 \pm 0.06	0.34 \pm 0.05	0.49 \pm 0.12	0.08 \pm 0.03	0.13 \pm 0.08	0.007 \pm 0.08
	$J_{TC} \downarrow$	8.9e-3	1e-3	9.3e-3	6e-3	1.2e-3	9e-4	9e-3
<i>InMoov-Stretch</i>	$J_r \uparrow$	-288.2 \pm 44.7	-134.9 \pm 0.2	-288.3 \pm 43.5	288.2 \pm 42.8	-134.6 \pm 0.23	-283.3 \pm 30.3	-133.7 \pm 1.1
	$J_C \downarrow$	0.08 \pm 0.04	0.04 \pm 0.01	0.02 \pm 0.009	0.11 \pm 0.05	0.04 \pm 0.01	0.03 \pm 0.009	0.007 \pm 0.05
	$J_{TC} \downarrow$	8e-3	4.4e-3	1.8e-2	3.2e-2	2.8e-2	4.5e-3	2e-2

TABLE II

QUANTITATIVE RESULTS OVER 50 EPISODES ON REAL-WORLD TASKS.

Method	PPO		CSC		USL		ACS (ours)	
	<i>Kuka-Pick</i>	<i>InMoov-Stretch</i>	<i>Kuka-Pick</i>	<i>InMoov-Stretch</i>	<i>Kuka-Pick</i>	<i>InMoov-Stretch</i>	<i>Kuka-Pick</i>	<i>InMoov-Stretch</i>
Success Rate \uparrow	86%	64%	42%	44%	56%	52%	62%	50%
Avg # of Collisions \downarrow	3.6	1.0	1.74	0.62	0.88	0.48	0.24	0.28
Avg # of Iterations \downarrow	0	0	4.0	4.6	2.6	9.0	1.9	6.8

B. Results

In all experimental results, we set α to 0.2. More investigations on the impact of varying α are elaborated in §VI-C. **Simulation tasks results.** The results of in-training performance of all methods across four simulation tasks are shown in Fig. 5. The corresponding numerical result is shown in Table I. We see that ACS achieves the best task performance on *Ant-Run* (+11.2%), *Kuka-Reach* (+48.9%) and *InMoov-Stretch* (+47.9%) while preserving nearly zero safety violation, indicating that ACS can quickly learn from failures and find a better trade-off boundary to balance between task optimality and safety. We also notice from temporal cost rate J_{TC} in Table I that ACS is a faster projection method thus could cater better towards time-critical tasks. While other algorithms may achieve better task objective in *Kuka-Pick* (-24.2%), they either fail to guarantee safety nor provide real-time control response. On the contrary, by implicitly predicting the trajectory of the obstacle with its advantage network and bounding it with an adaptive chance constraint, ACS achieves nearly zero safety violation on this task.

Real-world tasks results. We evaluate the performance of ACS against USL, CSC and PPO in terms of success rate, average number of collisions, and number of iterations to derive each action on both tasks. As shown in Table II, results demonstrate that ACS effectively adapts to stochastic real-world tasks and exhibit better success rate (+30%) while preserving low safety violation (-65%) with fewer iterations on both tasks. Since PPO achieves the best task objective via a brute-force path, it severely violates safety constraint. On the contrary, ACS outperforms all other methods in balancing the task and safety considerations.

C. Recovery Capability against Adversarial Policy

To further assess efficiency and robustness, we design an experiment on *Kuka-Reach* where we train an adversarial policy to induce the agent towards unsafe regions with higher cost. During an episode of 1000 steps, the target policy and adversarial policy take turns and alternate every 100 steps.

As shown in Fig. 6(a), ACS demonstrates its effectiveness in quickly decreasing the safety cost to 0 whenever the adversarial policy induces the agent to an unsafe state. In comparison, USL fails to fully recover before the adversary comes to play again. More notably, as shown in Fig. 6(b), ACS quickly recovers even from the worst unsafe states (i.e., $cost = 1$) and remains safe during its vigilance, while the other algorithms keep degenerating under adversarial attacks.

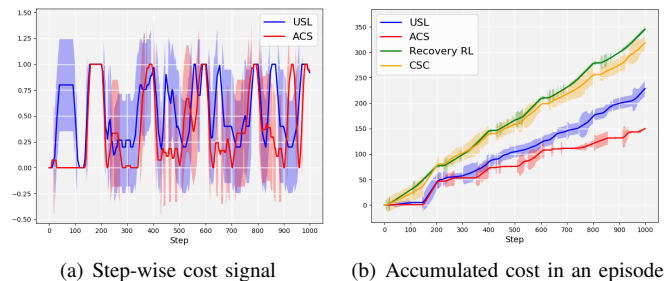


Fig. 6. Recovery capabilities in terms of cost against an adversarial policy.

D. ACS with Different Tolerance Thresholds

Additional experiments demonstrating how different tolerances α affect the performance of ACS is further illustrated in Appendix VI-C. Notably, even when tolerance $\alpha = 1$, indicating that ACS depends solely on the sub-optimal policy layer without any projection, it still outperforms all competing methods. When tolerance $\alpha = 0.2$ the controller can find the best trade-off between task and safety performance.

VI. CONCLUSIONS

In this paper, we propose Adaptive Chance-constraint Safeguard (ACS), a novel safe RL framework utilizing a hierarchical architecture to correct unsafe actions yielded by the upper policy layer via a fast Quasi-Newton method. Through extensive theoretical analysis and experiments on both simulated and real-world tasks, we demonstrate ACS’s superiority in enforcing safety while preserving optimality and robustness across different scenarios.

APPENDIX

A. Safety Probability Approximation via Value Function

In this section, we draw from the Bellman equation [35] and approximate the expected long-term chance-constrained safety probability Ψ in Eq. (3) with a value network $V_C^\pi(x_k)$ updated via RL explorations.

Theorem VI.1. *Let $r_{s_i} = 1\{\bigcap_{\tau \in \mathcal{T}(k)} x_j \in S_C\}$ be a Bernoulli random variable that indicates joint safety constraint satisfaction, and $r_{C_i} = 1 - r_{s_i}$ be the one-shot indicator of the complementary unsafe set, the expected safe possibility at x_k is $\Psi(x_k) = 1 - V_C^\pi(x_k)$.*

Proof. From [35] we have

$$V_C^\pi(x_{k+1}) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_i \mathcal{J}_{C_i} \right] \quad (9a)$$

$$= \frac{\sum_{u \sim \pi_\theta} \sum_x \sum_i (r_{C_i} + \gamma V_C^\pi(x_{k+1}))}{\sum_{u \sim \pi_\theta} \sum_x \sum_i 1} \quad (9b)$$

$$= \sum_{\{\tau: r_{C_i}(\tau_T)=1\}} \gamma^T P(\tau_i) * 1 \quad (9c)$$

$$= 1 - \Psi(x_{k+1}) \quad (9d)$$

Notice that the RHS of Eq. (9c) is essentially the possibility of unsafe trajectory $P(\tau_i)$ weighted by γ^T . Therefore, the expected safe possibility at x_k is $\Psi(x_k) = 1 - V_C^\pi(x_k)$. \square

B. In-training Safety Certificate for Dynamic Policy

In this section, we provide the detail proof of Theorem IV.1. For simplicity, here we consider only one safety constraint (i.e., $|C| = 1$) and omit the subscript i .

Proof. Since $\Psi(x_k)$ can be approximated by the value function of safety cost $V_C^\pi(x_k)$ (Theorem VI.1) and the advantage function $A_C^{\pi_\theta}(x_k, u_k) = \mathbb{E}[V_C(x_{k+1}|x_k, \pi_\theta(x_k))] - V_C^{\pi_\theta}(x_k)$, we first certify safety convergence for a stationary policy π_θ based on forward invariance:

$$\begin{aligned} \mathbb{E}[V_C(x_{k+1})|x_k, \pi_\theta(x_k)] - V_C^{\pi_\theta}(x_k) &\leq \mathcal{F}(\alpha - V_C^{\pi_\theta}(x_k)) \\ &\leq \alpha - V_C^{\pi_\theta}(x_k) \end{aligned} \quad (10a)$$

$$\Rightarrow \mathbb{E}[V_C(x_{k+1})|x_k, \pi_\theta(x_k)] \leq \alpha \quad (10b)$$

where Eq. (10a) is derived by convexity. Since no system dynamics model is available in our settings, it is impossible to guarantee zero in-training safety without any failures [6]. Therefore, we derive an upper-bound for policy updates using the trust-region method [31] to ensure our safety certificate, as previously defined for the stationary policy case, Eq. (10b) holds under dynamic policy updates as well. The trust-region method constrains the update of policy parameters by using the total variation distance $D_{TV} = D_{TV}(\pi_\theta || \pi_{\theta_{old}})$ to ensure the new policy π_θ does not deviate significantly from the old policy $\pi_{\theta_{old}}$. Therefore, following the trust-region derivations in [31] and [1], we start with the following inequality for the safety value function under the new policy:

$$V_C^{\pi_\theta} - V_C^{\pi_{\theta_{old}}} \leq \frac{1}{1-\gamma} \mathbb{E}_{x \sim \rho_{\theta_{old}}, u \sim \pi_{\theta_{old}}} [A_C^{\pi_{\theta_{old}}}] + \beta D_{TV}, \quad (11)$$

where β is a positive coefficient that weighs the total variation distance D_{TV} in the safety cost inequality. Here, $\beta = \frac{2\gamma \max |E_{\tau \sim \pi_\theta} A_C^{\pi_{\theta_{old}}}|}{1-\gamma}$, τ denotes a trajectory resulting from policy π , and γ is the discount factor that controls the expected convergence rate.

Assuming the safety certificate holds for the old policy $\pi_{\theta_{old}}$, which implies that the expected advantage under any state-action pair does not exceed a threshold α , we have the following:

$$\mathbb{E}_{x \sim \rho_{\theta_{old}}, u \sim \pi_{\theta_{old}}} [A_C^{\pi_{\theta_{old}}}] \leq \alpha. \quad (12)$$

To maintain the safety constraint for the updated policy, we derive an upper bound for D_{TV} by rearranging equation Eqn. (11) and considering the safety constraint in Eqn. (12):

$$D_{TV} \leq \frac{(1-\gamma)(\alpha - \mathbb{E}_{x \sim \rho_{\theta_{old}}, u \sim \pi_{\theta_{old}}} [A_C^{\pi_{\theta_{old}}}]}{\beta}. \quad (13)$$

With the upper bound for policy updates D_{TV} in equation Eqn. (13), we can easily derive a bound for the updated safety cost value function in Eqn. (11). Finally we have essentially certificated the updated policy to remain within a safe region defined by the trust-region method and ensure Eqn. (10b) is grounded for the case of dynamic policy update. \square

C. Ablation Study on Tolerance Level α

TABLE III

QUANTITATIVE RESULTS OVER TOLERANCE LEVEL α ON *Kuka-Reach*.

<i>Kuka-Reach</i>	J_r	J_C	J_{TC} (s)
ACS ($\alpha = 1$)	-221.9 \pm 87.40	0.05 \pm 0.01	0
ACS ($\alpha = 0.6$)	-178.4 \pm 85.20	0.02 \pm 0.01	44e-4
ACS ($\alpha = 0.2$)	-138.4 \pm 60.89	0.0032 \pm 0.04	1e-2
ACS ($\alpha = 0.05$)	-196.3 \pm 70.00	0.0028 \pm 0.02	31e-3

ACKNOWLEDGMENT

The authors would like to thank Mahsa Ghasemi, Sidharth Gangadhar, Yorie Nakahira, Weiye Zhao, Changliu Liu for their precious comments and suggestions. Liang Gong would also like to thank JAKA Robotics for providing with the hardware for experimentation in this paper.

REFERENCES

- [1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.
- [2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [3] Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.
- [4] Pranjal Awasthi, Corinna Cortes, Yishay Mansour, and Mehryar Mohri. A theory of learning with competing objectives and user feedback. In *Progress and Challenges in Building Trustworthy Embodied AI*, 2022.
- [5] Dimitri Bertsekas. *Reinforcement learning and optimal control*. Athena Scientific, 2019.
- [6] Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, and Animesh Garg. Conservative safety critics for exploration. *arXiv preprint arXiv:2010.14497*, 2020.
- [7] Steven Bohez, Abbas Abdolmaleki, Michael Neunert, Jonas Buchli, Nicolas Heess, and Raia Hadsell. Value constrained model-free continuous control. *arXiv preprint arXiv:1902.04623*, 2019.

- [8] Yi Chen, Jing Dong, and Zhaoran Wang. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.
- [9] Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3387–3395, 2019.
- [10] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.
- [11] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.
- [12] Priya L Donti, David Rolnick, and J Zico Kolter. Dc3: A learning method for optimization with hard constraints. *arXiv preprint arXiv:2104.12225*, 2021.
- [13] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.
- [14] Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. *arXiv preprint arXiv:1711.06782*, 2017.
- [15] James Ferlez, Mahmoud Elnaggar, Yasser Shoukry, and Cody Fleming. Shieldnn: A provably safe nn filter for unsafe nn controllers. *arXiv preprint arXiv:2006.09564*, 2020.
- [16] Milan Ganai, Zheng Gong, Chenning Yu, Sylvia Herbert, and Sicun Gao. Iterative reachability estimation for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [17] Siddharth Gangadhar, Zhuoyuan Wang, Haoming Jing, and Yorie Nakahira. Adaptive safe control for driving in uncertain environments. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1662–1668. IEEE, 2022.
- [18] Javier Garcia and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.
- [19] Tairan He, Weiye Zhao, and Changliu Liu. Autocost: Evolving intrinsic cost for zero-violation reinforcement learning. *arXiv preprint arXiv:2301.10339*, 2023.
- [20] Haoming Jing and Yorie Nakahira. Probabilistic safety certificate for multi-agent systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 5343–5350. IEEE, 2022.
- [21] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6023–6029. IEEE, 2019.
- [22] Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based model predictive control for safe exploration. In *2018 IEEE conference on decision and control (CDC)*, pages 6059–6066. IEEE, 2018.
- [23] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.
- [24] Qingkai Liang, Fanyu Que, and Eytan Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018.
- [25] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.
- [26] Yongshuai Liu, Avishai Halev, and Xin Liu. Policy learning with constraints in model-free reinforcement learning: A survey. In *The 30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.
- [27] Haitong Ma, Changliu Liu, Shengbo Eben Li, Sifa Zheng, Wen-chao Sun, and Jianyu Chen. Learn zero-constraint-violation policy in model-free constrained reinforcement learning. *arXiv preprint arXiv:2111.12953*, 2021.
- [28] Panagiotis Petsagkourakis, Ilya Orson Sandoval, Eric Bradford, Dongda Zhang, and Ehecatl Antonio del Rio-Chanona. Constrained reinforcement learning for dynamic optimization under uncertainty. *IFAC-PapersOnLine*, 53(2):11264–11270, 2020.
- [29] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.
- [30] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.
- [31] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.
- [32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [33] Mahmoud Selim, Amr Alanwar, Shreyas Kousik, Grace Gao, Marco Pavone, and Karl H Johansson. Safe reinforcement learning using black-box reachability analysis. *IEEE Robotics and Automation Letters*, 7(4):10665–10672, 2022.
- [34] Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. Learning to be safe: Deep rl with a safety critic. *arXiv preprint arXiv:2010.14603*, 2020.
- [35] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [36] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.
- [37] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minh Hwang, Joseph E Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3):4915–4922, 2021.
- [38] Brijen Thananjeyan, Ashwin Balakrishna, Ugo Rosolia, Felix Li, Rowan McAllister, Joseph E Gonzalez, Sergey Levine, Francesco Borrelli, and Ken Goldberg. Safety augmented value estimation from demonstrations (saved): Safe deep model-based rl for sparse cost robotic tasks. *IEEE Robotics and Automation Letters*, 5(2):3612–3619, 2020.
- [39] Nolan C Wagener, Byron Boots, and Ching-An Cheng. Safe reinforcement learning using advantage-based intervention. In *International Conference on Machine Learning*, pages 10630–10640. PMLR, 2021.
- [40] Zhuoyuan Wang, Haoming Jing, Christian Kurniawan, Albert Chern, and Yorie Nakahira. Myopically verifiable probabilistic certificate for long-term safety. In *2022 American Control Conference (ACC)*, pages 4894–4900. IEEE, 2022.
- [41] Tianhao Wei and Changliu Liu. Safe control algorithms using energy functions: A unified framework, benchmark, and new directions. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 238–243. IEEE, 2019.
- [42] Min Wen and Ufuk Topcu. Constrained cross-entropy method for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.
- [43] Linrui Zhang, Qin Zhang, Li Shen, Bo Yuan, Xueqian Wang, and Dacheng Tao. Evaluating model-free reinforcement learning toward safety-critical tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15313–15321, 2023.
- [44] Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 33:15338–15349, 2020.
- [45] Lujie Zhao, Liang Gong, Xudong Li, Chen Yang, Zhaoran Chen, Yixiang Huang, and Chengliang Liu. A bionic arm mechanism design and kinematic analysis of the humanoid traffic police. In *2019 IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 1606–1611. IEEE, 2019.
- [46] Weiye Zhao, Tairan He, Rui Chen, Tianhao Wei, and Changliu Liu. State-wise safe reinforcement learning: A survey. *arXiv preprint arXiv:2302.03122*, 2023.
- [47] Weiye Zhao, Tairan He, and Changliu Liu. Model-free safe control for zero-violation reinforcement learning. In *5th Annual Conference on Robot Learning*, 2021.
- [48] Weiye Zhao, Tairan He, and Changliu Liu. Probabilistic safeguard for reinforcement learning using safety index guided gaussian process models. In *Learning for Dynamics and Control Conference*, pages 783–796. PMLR, 2023.