Safe Reinforcement Learning Using Advantage-Based Intervention

Nolan Wagener [1] **Byron Boots** [2] **Ching-An Cheng** [3]

Abstract

Many sequential decision problems involve finding a policy that maximizes total reward while obeying safety constraints. Although much recent research has focused on the development of safe reinforcement learning (RL) algorithms that produce a safe policy after training, ensuring safety *during* training as well remains an open problem. A fundamental challenge is performing exploration while still satisfying constraints in an unknown Markov decision process (MDP). In this work, we address this problem for the chance-constrained setting. We propose a new algorithm, SAILR, that uses an intervention mechanism based on advantage functions to keep the agent safe throughout training and optimizes the agent's policy using off-the-shelf RL algorithms designed for unconstrained MDPs. Our method comes with strong guarantees on safety during *both* training and deployment (i.e., after training and without the intervention mechanism) and policy performance compared to the optimal safety-constrained policy. In our experiments, we show that SAILR violates constraints far less during training than standard safe RL and constrained MDP approaches and converges to a well-performing policy that can be deployed safely without intervention. Our code is available at https://github.com/nolanwagener/safe_rl.

1. Introduction

Reinforcement learning (RL) (Sutton & Barto, 2018) enables an agent to learn good behaviors with high returns through interactions with an environment of interest. However, in many settings, we want the agent not only to find a high-return policy but also avoid undesirable states as much as possible, even during training. For example, in a bipedal locomotion task, we do not want the robot to fall over and risk damaging itself either during training or deployment. Maintaining safety while exploring an unknown environment is challenging, because venturing into new regions of the state space may carry a chance of a costly failure.

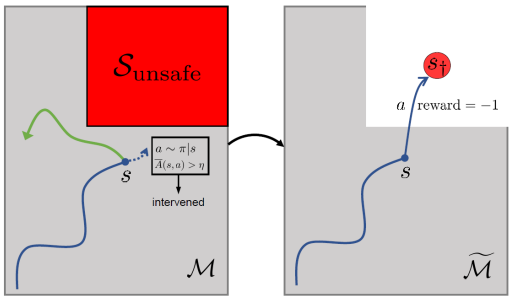


Figure 1. Advantage-based intervention of SAILR and construction of the surrogate MDP $\widetilde{\mathcal{M}}$. In $\mathcal{M}$, whenever the policy π proposes an action a which is disadvantageous (w.r.t. a backup policy μ) in terms of safety, μ intervenes and guides the agent to safety (green path). From the perspective of π, it transitions to an absorbing state $s_\dagger$ and receives a penalizing reward of -1.

Safe reinforcement learning (García & Fernández, 2015; Amodei et al., 2016) studies the problem of designing learning agents for sequential decision-making with this challenge in mind. Most safe RL approaches tackle the safety requirement either by framing the problem as a constrained Markov decision process (CMDP) (Altman, 1999) or by using control-theoretic tools to restrict the actions that the learner can take. However, due to the natural conflict between learning, maximizing long-term reward, and satisfying safety constraints, these approaches make different performance trade-offs.

CMDP-based approaches (Borkar, 2005; Achiam et al., 2017; Le et al., 2019) take inspiration from existing constrained optimization algorithms for non-sequential problems, notably the Lagrangian method (Bertsekas, 2014). The most prominent examples (Chow et al., 2017; Tessler et al., 2018) rely on first-order primal-dual optimization to solve a stochastic nonconvex saddle-point problem. Though they eventually produce a safe policy, such approaches have no guarantees on policy safety during training. Other safe

[1]Institute for Robotics and Intelligent Machines, Georgia Institute of Technology, Atlanta, Georgia, USA [2]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, Washington, USA [3]Microsoft Research, Redmond, Washington, USA. Correspondence to: Nolan Wagener <nolan.wagener@gatech.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

RL approaches (Achiam et al., 2017; Le et al., 2019; Bharadhwaj et al., 2021) conservatively enforce safety constraints on every policy iterate by solving a constrained optimization problem, but they can be difficult to scale due to their high computational complexity. All of the above methods suffer from numerical instability originating in solving the stochastic nonconvex saddle-point problems (Facchinei & Pang, 2007; Lin et al., 2020); consequently, they are less robust than typical unconstrained RL algorithms.

Control-theoretic approaches to safe RL use interventions, projections, or planning (Hans et al., 2008; Wabersich & Zeilinger, 2018; Dalal et al., 2018; Berkenkamp et al., 2017) to enforce safe interactions between the agent and the environment, independent of the policy the agent uses. The idea is to use domain-specific heuristics to decide whether an action proposed by the agent's policy can be safely executed. However, some of these algorithms do not allow the agent to *learn* to be safe after training (Wabersich & Zeilinger, 2018; Hans et al., 2008; Polo & Rebollo, 2011), so they may not be applicable in scenarios where the control mechanism relies on resouces only available during training (such as computationally demanding online planning). It is also often unclear how these policies perform compared to the optimal policy in the CMDP-based approach.

In this work, we propose a new algorithm, SAILR (*Safe Advantage-based Intervention for Learning policies with Reinforcement*), that uses a novel advantage-based intervention rule to enable safe and stable RL for *general* MDPs. Our method comes with strong guarantees on safety during *both* training and deployment (i.e., after training and without the intervention mechanism) and has good on-policy performance compared to the optimal safety-constrained policy. Specifically, SAILR trains the agent's policy by calling an off-the-shelf RL algorithm designed for standard *unconstrained* MDPs. In each iteration, SAILR: *1)* queries the base RL algorithm to get a data-collection policy; *2)* runs the policy in the MDP while utilizing the advantage-based intervention rule to ensure safe interactions (and executes a backup policy upon intervention to ensure safety); *3)* transforms the collected data into experiences in a new unconstrained MDP that penalizes any visits of intervened state-actions (visualized in Fig. 1); and *4)* gives the transformed data to the base RL algorithm to perform policy optimization.

Under very mild assumptions on the MDP and the safety of the backup policy used during the intervention,[1] we prove that running SAILR with *any* RL algorithm for unconstrained MDPs can safely learn a policy that has good performance in the safety-constrained MDP (with a bias propotional to how often the true optimal policy would be overridden by our intervention mechanism). Compared with existing work, SAILR is easier to implement and runs more reliably than the CMDP-based approaches. In addition, since we only rely on estimated advantage functions, our approach is also more generic than the aforementioned control-theoretic approaches which make assumptions on smoothness or ergodicity of the problem.

We also empirically validate our theory by comparing SAILR with several standard safe RL algorithms in simulated robotics tasks. The encouraging experimental results strongly support the theory: SAILR can learn safe policies with competitive performance using a standard unconstrained RL algorithm, PPO (Schulman et al., 2017), while incurring only a small fraction of unsafe training rollouts compared to the baselines.

2. Preliminaries

2.1. Notation

A γ-discounted infinite-horizon Markov decision process (MDP) is denoted as a 5-tuple, $\mathcal{M} \coloneqq (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, $P(s'|s,a)$ is a transition dynamics, $r(s,a) \in [0,1]$ is a reward function, and $\gamma \in [0,1)$ is a discount factor. In this work, $\mathcal{S}$ and $\mathcal{A}$ can be either discrete or continuous. A policy π on $\mathcal{M}$ is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes probability distributions on $\mathcal{A}$. We use the following overloaded notation: For a state distribution $d \in \Delta(\mathcal{S})$ and a function $f : \mathcal{S} \to \mathbb{R}$, we define $f(d) \coloneqq \mathbb{E}_{s\sim d}[f(s)]$; similarly, for a policy π and a function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define $g(s,\pi) \coloneqq \mathbb{E}_{a\sim\pi|s}[g(s,a)]$. We would often omit the random variable in the subscript of expectations, if it is clear.

A policy π induces a trajectory distribution $\rho^\pi(\xi)$, where $\xi = (s_0, a_0, s_1, a_1, \dots)$ denotes a random trajectory. The state-action value function of π is defined as $Q^\pi(s,a) \coloneqq \mathbb{E}_{\xi\sim\rho^\pi|s_0=s,a_0=a}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t)]$ and its state value function as $V^\pi(s) = Q^\pi(s,\pi)$. We denote the optimal stationary policy of $\mathcal{M}$ as π^* and its respective value functions as Q^* and V^*. Let $d_t^\pi(s)$ be the state distribution at time t induced by running π in $\mathcal{M}$ from an initial state distribution d_0 (note that $d_0^\pi = d_0$); then the average state distribution induced by π is $d^\pi(s) \coloneqq (1-\gamma)\sum_{t=0}^\infty \gamma^t d_t^\pi(s)$. For brevity, we overload the notation d^π to also denote the state-action distribution $d^\pi(s,a) \coloneqq d^\pi(s)\pi(a|s)$. Finally, later in the paper we will consider multiple variants of an MDP (specifically, $\mathcal{M}$, $\overline{\mathcal{M}}$, and $\widetilde{\mathcal{M}}$) and will use the decorative symbol on the MDP notation to distinguish similar objects from different MDPs (e.g., V^π and $\overline{V}^\pi$ will denote the state value functions of π in $\mathcal{M}$ and $\overline{\mathcal{M}}$). Throughout this paper, we'll take $\mathbb{E}_{s'|s,a}$ to mean $\mathbb{E}_{s'\sim\mathcal{P}|s,a}$.

[1] We only assume that the unsafe states are absorbing and that the backup policy is safe from the initial state with high probability. We do *not* assume that the backup policy can achieve high rewards.

2.2. Safe Reinforcment Learning

We consider safe RL in a γ-discounted infinite horizon MDP $\mathcal{M}$, where safety means that the probability of the agent entering an unsafe subset $\mathcal{S}_{\text{unsafe}} \subset \mathcal{S}$ is low. We assume that we know the unsafe subset $\mathcal{S}_{\text{unsafe}}$ and the safe subset $\mathcal{S}_{\text{safe}} \coloneqq \mathcal{S} \setminus \mathcal{S}_{\text{unsafe}}$. However, we make no assumption on the knowledge of reward r and dynamics P, except that the reward r is zero on $\mathcal{S}_{\text{unsafe}}$ and that $\mathcal{S}_{\text{unsafe}}$ is absorbing: once the agent enters $\mathcal{S}_{\text{unsafe}}$ in a rollout, it cannot travel back to $\mathcal{S}_{\text{safe}}$ and stays in $\mathcal{S}_{\text{unsafe}}$ for the rest of the rollout.

Objective Our goal is to find a policy π that is safe and has a high return in $\mathcal{M}$, and to do so via a safe data collection process. Specifically, while keeping the agent safe during exploration, we want to solve the following chance-constrained policy optimization problem:

$$\max_{\pi} \quad V^{\pi}(d_0) \tag{1}$$
$$\text{s.t.} \quad (1-\gamma)\sum_{h=0}^{\infty}\gamma^h \text{Prob}(\xi_h \subset \mathcal{S}_{\text{safe}} \mid \pi) \geq 1-\delta,$$

where $\delta \in [0,1]$ is the tolerated failure probability, $\xi_h = (s_0, a_0, \ldots, s_{h-1}, a_{h-1})$ denotes an h-step trajectory segment, and $\text{Prob}(\xi_h \subset \mathcal{S}_{\text{safe}} \mid \pi)$ denotes the probability of ξ_h being safe (i.e., not entering $\mathcal{S}_{\text{unsafe}}$ from time step 0 to $h-1$) under the trajectory distribution ρ^{π} of π on $\mathcal{M}$.[2]

We desire the the agent to provide anytime safety in *both* training and deployment. During training, the agent can interact with the unknown MDP $\mathcal{M}$ to collect data under a training budget, such as the maximum number of environment interactions or allowed unsafe trajectories the agent can generate. Once the budget is used up, training stops, and an approximate solution of (1) needs to be returned.

The constraint in (1) is known as a *chance constraint*. The definition here accords to an exponentially weighted average (based on the discount factor γ) of trajectory safety probabilities of different horizons. This weighted average concept arises naturally in γ-discounted MDPs, because the objective in (1) can also be written as a weighted average of undiscounted expected returns, i.e., $V^{\pi}(d_0) = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h U_h^{\pi}(d_0)$, where $U_h^{\pi}(d_0) \coloneqq \mathbb{E}_{\rho^{\pi}}[\sum_{t=0}^{h} r(s_t, a_t)]$.

CMDP Formulation The chance-constrained policy optimization problem in (1) can be formulated as a constrained Markov decision process (CMDP) problem (Altman, 1999; Chow et al., 2017). For the mathematical convenience of defining and analyzing the equivalence between (1) and a CMDP, instead of treating $\mathcal{S}_{\text{unsafe}}$ as a single meta-absorbing state, without loss of generality we define $\mathcal{S}_{\text{unsafe}} \coloneqq \{s_{\triangleright}, s_{\circ}\}$. The semantics of this set is that when an agent leaves $\mathcal{S}_{\text{safe}}$ and enters $\mathcal{S}_{\text{unsafe}}$, it first goes to $s_{\triangleright}$ and, regardless of which action it takes at $s_{\triangleright}$, it then goes to the absorbing state $s_{\circ}$ and stays there forever. We can view $s_{\triangleright}$ as a meta-state that summarizes the unsafe region in a given RL application (e.g., a biped robot falling on the ground) and $s_{\circ}$ as a fictitious state that captures the absorbing property of $\mathcal{S}_{\text{unsafe}}$.

For an MDP $\mathcal{M}$ with an unsafe set $\mathcal{S}_{\text{unsafe}} \coloneqq \{s_{\triangleright}, s_{\circ}\}$, define the cost $c(s,a) \coloneqq \mathbb{1}\{s = s_{\triangleright}\}$, where $\mathbb{1}$ denotes the indicator function. Then we can define a CMDP $(\mathcal{S}, \mathcal{A}, P, r, c, \gamma)$ using a reward-based MDP $\mathcal{M} \coloneqq (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ and a cost-based MDP $\overline{\mathcal{M}} \coloneqq (\mathcal{S}, \mathcal{A}, P, c, \gamma)$. Using these new definitions, we can write the chance-constrained policy optimization in (1) as a CMDP problem:

$$\max_{\pi} \quad V^{\pi}(d_0) \quad \text{subject to} \quad \overline{V}^{\pi}(d_0) \leq \delta. \tag{2}$$

For completeness, we include a proof of this equivalence in Appendix A.2, which follows from the fact that the unsafe probability can be represented as the expected cumulative cost, i.e., $\text{Prob}(s_{\triangleright} \in \xi^h \mid \pi) = \mathbb{E}_{\rho^{\pi}}[\sum_{t=0}^{h-1} c(s_t, a_t)]$. In other words, the chance-constrained policy optimization problem is a CMDP problem that aims to find a policy that has a high cumulative reward $V^{\pi}(d_0)$ with cumulative cost $\overline{V}^{\pi}(d_0)$ below the allowed failure probability δ.

Challenges This CMDP formulation has been commonly studied to design RL algorithms to find good polices that can be deployed safely (Chow et al., 2017; Achiam et al., 2017; Tessler et al., 2018; Efroni et al., 2020). However, as mentioned in the introduction, these algorithms do not necessarily ensure safety during training and can be numerically unstable. At a high level, this instability stems from the lack of off-the-shelf computationally reliable and efficient solvers for large-scale constrained stochastic optimization.

While several control-theoretic techniques have been proposed to ensure safe data collection (Dalal et al., 2018; Wabersich & Zeilinger, 2018; Perkins & Barto, 2002; Chow et al., 2018; 2019; Berkenkamp et al., 2017; Fisac et al., 2018) and in some cases prevent the need for solving a constrained problem, it is unclear how the learned policy performs in terms of the objective $V^{\pi}(d_0)$ in (2) (i.e., without any interventions). Most of these algorithms also require stronger assumptions on the environment than approaches based on CMDPs (e.g., smoothness or ergodicity).

As we will show, our proposed approach retains the best of both approaches, ensuring safe data collection via interventions while guaranteeing good performance and safety when deployed without the intervention mechanism.

[2] We abuse the notation $\xi_h \subset \mathcal{S}_{\text{safe}}$ to mean that $s_\tau \in \mathcal{S}_{\text{safe}}$ for each s_τ in $\xi_h = (s_0, a_0, \ldots, s_{h-1}, a_{h-1})$.

3. Method

Our safe RL approach, SAILR, finds an approximate solution to the CMDP problem in (2) by using an advantage-based intervention rule for safe data collection and an off-the-self RL algorithm for policy optimization. As we will see, SAILR can ensure safety for *both* training and deployment, when *1)* the intervention rule belongs to an "admissible class" (see Definition 1 in Section 3.1.2); and *2)* the base RL algorithm finds a nearly optimal policy for a new *unconstrained* problem of a surrogate MDP $\widetilde{\mathcal{M}}$ constructed by the *intervention rule* together with $\mathcal{M}$. Moreover, because SAILR can reuse exsiting RL algorithms for unconstrained MDPs to optimize policies, it is easier to implement and is more stable than typical CMDP approaches based on constrained optimization.

Specifically, SAILR optimizes policies iteratively as outlined in Section 3. As input, it takes an RL algorithm $\mathcal{F}$ for unconstrained MDPs and an intervention rule $\mathcal{G} : \pi \mapsto \mathcal{G}(\pi)$, where $\pi' = \mathcal{G}(\pi)$ is a *shielded policy* such that π' runs a *backup policy* $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ instead of π when π proposes "unsafe actions." In every iteration, SAILR first queries the base RL algorithm $\mathcal{F}$ for a data-collection policy π to execute in $\widetilde{\mathcal{M}}$ (line 3). Then it uses the intervention rule $\mathcal{G}$ to modify π into π' (line 4) such that running π' in the original MDP $\mathcal{M}$ can be safe with high probability while effectively simulating execution of π in the surrogate $\widetilde{\mathcal{M}}$. Next, it collects the data $\mathcal{D}$ by running π' in $\mathcal{M}$ and then transforms it into new data $\widetilde{\mathcal{D}}$ of π in $\widetilde{\mathcal{M}}$ (line 5). It then feeds $\widetilde{\mathcal{D}}$ to the base RL algorithm $\mathcal{F}$ for policy optimization (line 6), and optionally uses $\mathcal{D}$ to refine the intervention rule $\mathcal{G}$ (line 7). The process above is repeated until the training budget is used up. When this happens, SAILR terminates and returns the best policy $\hat{\pi}^*$ the base algorithm $\mathcal{F}$ can produce for $\widetilde{\mathcal{M}}$ so far (line 9).

We provide the following informal guarantee for SAILR, which is a corollary of our main result in Theorem 1 presented in Section 3.3.

Proposition 1 (Informal Guarantee)**.** *Let π^* be an optimal policy for $\mathcal{M}$. For* SAILR*, if the intervention rule $\mathcal{G}$ is admissible (Definition 1 in Section 3.1.2)) and the RL algorithm $\mathcal{F}$ learns an ε-suboptimal policy $\hat{\pi}$ for $\widetilde{\mathcal{M}}$, then $\hat{\pi}$ has the following performance and safety guarantees in $\mathcal{M}$:*

$$V^{\pi^*}(d_0) - V^{\hat{\pi}}(d_0) \leq \varepsilon + 2P_{\mathcal{G}}(\pi^*)/(1-\gamma)$$
$$\overline{V}^{\hat{\pi}}(d_0) \leq \varepsilon + \overline{V}^{\mu}(d_0),$$

where μ is the backup policy in $\mathcal{G}$ and $P_{\mathcal{G}}(\pi^*)$ is the probability that π^* visits the intervention set of $\mathcal{G}$ in $\mathcal{M}$.

In other words, if the base algorithm $\mathcal{F}$ used by SAILR can find an ε-suboptimal policy for the surrogate, unconstrained MDP $\widetilde{\mathcal{M}}$, then the policy returned by SAILR is roughly ε-suboptimal in the original MDP $\mathcal{M}$, up to an additional error proportional to the probabilty that the optimal policy π^* would be overridden by the intervention rule $\mathcal{G}$ at some point while running in $\mathcal{M}$. Furthermore, the returned policy $\hat{\pi}$ is as safe as the backup policy μ of the intervention rule $\mathcal{G}$, up to an additional unsafe probability ε arising from the suboptimality in solving $\widetilde{\mathcal{M}}$ with $\mathcal{F}$.

Algorithm 1 SAILR

Input: MDP $\mathcal{M}$, RL algorithm $\mathcal{F}$, Intervention rule $\mathcal{G}$
Output: Optimized safe policy $\hat{\pi}^*$

1: $\mathcal{F}$.`Initialize()`
2: **while** training budget available **do**
3: $\quad\pi \leftarrow \mathcal{F}$.`GetDataCollectionPolicy()`
4: $\quad\pi' \leftarrow$ `DefineShieldedPolicy(`π`,`$\mathcal{G}$`)`
5: $\quad\mathcal{D}, \widetilde{\mathcal{D}} \leftarrow$ `CollectData(`π'`,`$\mathcal{M}$`)`
6: $\quad\mathcal{F}$.`OptimizePolicy(`$\widetilde{\mathcal{D}}$`)`
7: $\quad\mathcal{G} \leftarrow$ `UpdateInterventionRule(`$\mathcal{D}$`)` *(Optional)*
8: **end while**
9: $\hat{\pi}^* \leftarrow \mathcal{F}$.`GetOptimizedPolicy()`

We point out the results above hold without any assumption on the MDP (other than that the unsafe subset $\mathcal{S}_{\text{unsafe}}$ is absorbing and the reward is zero on $\mathcal{S}_{\text{unsafe}}$). To learn a safe policy, SAILR only needs a good *unconstrained* RL algorithm $\mathcal{F}$, a backup policy μ that is safe starting at the *initial state* (not globally), and an advantage function estimate of μ, as we explain later in this section.

The price we pay for keeping the agent safe using an intervention rule $\mathcal{G}$ is a performance bias proportional to $P_{\mathcal{G}}(\pi^*)/(1-\gamma)$. This happens because employing an intervention rule during data collection limits where the agent can explore in $\mathcal{M}$. Thus, if the optimal policy π^* goes to high-reward states which would be cut off by the intervention rule, SAILR (and any other intervention-based algorithm) will suffer in proportion to the intervention probability. Despite the dependency on $P_{\mathcal{G}}(\pi^*)$, we argue that SAILR provides a resonable trade-off for safe RL thanks to its training safety and numerical stability. Moreover, we will discuss how to use data to improve the intervention rule $\mathcal{G}$ to reduce this performance bias.

In the following, we first discuss the design of our advantage-based intervention rules (Section 3.1) and provide details of the new MDP $\widetilde{\mathcal{M}}$ (Section 3.2). Then we state and prove the main result Theorem 1 (Section 3.3). The omitted proofs for the results in this section can be found in Appendix A.

3.1. Advantage-Based Intervention

We propose a family of intervention rules based on *advantage functions*. Each intervention rule $\mathcal{G}$ here is specified by a 3-tuple $(\overline{Q}, \mu, \eta)$, where $\overline{Q} : \mathcal{S}_{\text{safe}} \times \mathcal{A} \to [0, 1]$ is a state-action value function, $\eta \in [0, 1]$ is a threshold, and

$\mu \in \Pi$ is a backup policy. Given an arbitrary policy π, $\mathcal{G} = (\overline{Q}, \mu, \eta)$ constructs a new *shielded* policy π' based on an intervention set $\mathcal{I}$ defined by the advantage-like function $\overline{A}(s,a) \coloneqq \overline{Q}(s,a) - \overline{Q}(s,\mu)$:

$$\mathcal{I} \coloneqq \{(s,a) \in \mathcal{S}_{\text{safe}} \times \mathcal{A} \ : \ \overline{A}(s,a) > \eta\}. \tag{3}$$

When sampling a from $\pi'(\cdot|s)$ at some $s \in \mathcal{S}_{\text{safe}}$, it first samples a_- from $\pi(\cdot|s)$. If $(s,a_-) \notin \mathcal{I}$, it executes $a = a_-$. Otherwise, it samples a according to $\mu(\cdot|s)$. Mathematically, π' is described by the conditional distribution

$$\pi'(a|s) \coloneqq \pi(a|s)\mathbb{1}\{(s,a) \notin \mathcal{I}\} + \mu(a|s)w(s), \tag{4}$$

where $w(s) \coloneqq 1 - \sum_{\tilde{a}:(s,\tilde{a})\in\mathcal{I}} \pi(\tilde{a}|s)$. Note that π' may still take actions in $\mathcal{I}$ when μ has non-zero probability assigned to those actions.

By running the shielded policy contructed by the advantage function $\overline{A}$, SAILR controls the safety relative to the backup policy μ with respect to d_0. As we will show later, if the relative safety for each time step (i.e., advantage) is close to zero, then the relative safety overall is also close to zero (i.e. $\overline{V}^{\pi'}(d_0) \leq \delta$). Note that the sheilded policy π', while satisfying $\overline{V}^{\pi'}(d_0) \leq \delta$, can generally visit (with low probability) the states where $\overline{V}^{\mu}(s) > 0$ (e.g., $= 1$). At these places where μ is useless for safety, we need an intervention rule that naturally deactivates and lets the learner explore. Our advantage-based rule does exactly that. On the contrary, designing an intervention rule direclty based on Q-based functions, as in (Bharadhwaj et al., 2021; Thananjeyan et al., 2020; Eysenbach et al., 2018; Srinivasan et al., 2020), can be overly conservative in this scenario.

3.1.1. Motivating Example

Let us use an example to explain why the advantage-based rule works. Suppose we have a baseline policy μ that is safe starting at the intial state of the MDP $\mathcal{M}$ (i.e., $\overline{V}^{\mu}(d_0)$ is small). We can use μ as the backup policy and construct an intervention rule $\mathcal{G} = (\overline{Q}^{\mu}, \mu, 0)$, where we recall $\overline{Q}^{\mu}$ denotes the state-action value of μ for the cost-based MDP $\overline{\mathcal{M}}$. Because the intervention set in (3) only allows actions that are no more unsafe than than backup policy μ in execution, intuitively we see that the intervenend policy π' will be at least as safe as the baseline policy μ. Indeed, we can quickly verify this by the performance difference lemma (Lemma 3): $\overline{V}^{\pi'}(d_0) = \overline{V}^{\mu}(d_0) + \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi'}}[\overline{A}^{\mu}(s,a)] \leq \overline{V}^{\mu}(d_0)$. Importantly, in this example, we see that the safety of π' is ensured without requiring $\overline{V}^{\mu}(s)$ to be small for any $s \in \mathcal{S}$, but only starting from states sampled from d_0.

3.1.2. General Rules

We now generalize the above motivating example to a class of *admissible* intervention rules.

Definition 1 (σ-Admissible Intervention Rule)**.** We say an intervention rule $\mathcal{G} = (\overline{Q}, \mu, \eta)$ is *σ-admissible* if for all $s \in \mathcal{S}_{\text{safe}}$ and $a \in \mathcal{A}$: for some $\sigma \geq 0$,

$$\overline{Q}(s,a) \in [0, \gamma] \tag{5}$$

$$\overline{Q}(s,a) + \sigma \geq c(s,a) + \gamma\mathbb{E}_{s'|s,a}[\overline{Q}(s',\mu)], \tag{6}$$

where we recall $c(s,a) = \mathbb{1}\{s = s_{\triangleright}\}$. If the above holds with $\sigma = 0$, we say $\mathcal{G}$ is *admissible*.

One can verify that the previous example $\mathcal{G} = (\overline{Q}^{\mu}, \mu, 0)$ is admissible. But more generally, an admissible intervention rule with a backup policy μ can use $\overline{Q} \neq \overline{Q}^{\mu}$. In a sense, admissibility (with $\sigma = 0$) only needs $\overline{Q}$ to be a conservative version of $\overline{Q}^{\mu}$, because $\overline{Q}^{\mu}(s,a) = c(s,a) + \gamma\mathbb{E}_{s'|s,a}[\overline{Q}^{\mu}(s,\mu)]$ and (6) uses an upper bound; the σ term is a slack to allow for non-conservative $\overline{Q}$. More precisely, we have the following relationship.

Proposition 2. *If $\mathcal{G} = (\overline{Q}, \mu, \eta)$ is σ-admissible, then $\overline{Q}^{\mu}(s,a) \leq \overline{Q}(s,a) + \frac{\sigma}{1-\gamma}$ for all $s \in \mathcal{S}_{\text{safe}}$ and $a \in \mathcal{A}$.*

The condition in (6) is also closely related to the concept and theory of improvable heuristics in (Cheng et al., 2021) (i.e., we can view the $\overline{Q}(s,\mu)$ as a heurisitic for safety), where the authors show such $\overline{Q}$ can be constructed by pessimistic offline RL methods.

Examples We discuss several ways to construct admissible intervention rules. From Definition 1, it is clear that if $\mathcal{G} = (\overline{Q}, \mu, \eta)$ is σ-admissible, then $\mathcal{G}$ is also σ'-admissible for any $\sigma' \geq \sigma$ (in particular, $(\overline{Q}, \mu, \eta)$ is γ-admissible if $\overline{Q}(s,a) \in [0,\gamma]$). So we only discuss the minimal σ.

Proposition 3 (Intervention Rules)**.** *The following are true.*

1. ***Baseline policy**: Given a baseline policy μ of $\mathcal{M}$, $\mathcal{G} = (\overline{Q}^{\mu}, \mu, \eta)$ or $\mathcal{G} = (\overline{Q}^{\mu}, \mu^{+}, \eta)$ is admissible, where μ^{+} is the greedy policy that treats $\overline{Q}^{\mu}$ as a cost.*

2. ***Composite intervention**: Given K intervention rules $\{\mathcal{G}_k\}_{k=1}^{K}$, where each $\mathcal{G}_k = (\overline{Q}_k, \mu_k, \eta)$ is σ_k-admissible. Define $\overline{Q}_{\min}(s,a) = \min_k \overline{Q}_k(s,a)$ and let $\mu_{\min}$ be the greedy policy w.r.t. $\overline{Q}_{\min}$, and $\sigma_{\max} = \max_k \sigma_k$. Then, $\mathcal{G} = (\overline{Q}_{\min}, \mu_{\min}, \eta)$ is $\sigma_{\max}$-admissible.*

3. ***Value iteration**: Define $\overline{\mathcal{T}}$ as $\overline{\mathcal{T}}Q(s,a) \coloneqq c(s,a) + \gamma\mathbb{E}_{s'\sim P|s,a}[\min_{a'} Q(s',a')]$. If $\mathcal{G} = (\overline{Q}, \mu, \eta)$ is σ-admissible, then $\mathcal{G}^k = (\overline{\mathcal{T}}^k\overline{Q}, \mu^k, \eta)$ is $\gamma^k\sigma$-admissible, where μ^k is the greedy policy that treats $\overline{\mathcal{T}}^k\overline{Q}$ as a cost.*

4. ***Optimal intervention**: Let $\overline{\pi}^*$ be an optimal policy for $\overline{\mathcal{M}}$, and let $\overline{Q}^*$ be the corresponding state-action value function. Then $\mathcal{G}^* = (\overline{Q}^*, \overline{\pi}^*, \eta)$ is admissible.*

5. ***Approximation**: For σ-admissible $\mathcal{G} = (\overline{Q}, \mu, \eta)$, consider $\hat{Q}$ such that $\hat{Q}(s,a) \in [0,\gamma]$ for all $s \in \mathcal{S}_{\text{safe}}$ and $a \in \mathcal{A}$. If $\|\hat{Q} - \overline{Q}\|_{\infty} \leq \delta$, then $\hat{\mathcal{G}} = (\hat{Q}, \mu, \eta)$ is $(\sigma + (1+\gamma)\delta)$-admissible.*

Proposition 3 provides recipes for constructing σ-admissible intervention rules for safe RL, such as leveraging existing baseline policies in a system (Examples 1 and 2) and performing short-horizon planning (Example 3; namely model-predictive control (Bertsekas, 2017)). Moreover, Proposition 3 hints that we can treat designing intervention rules as finding the optimal state-action value function $\overline{Q}^*$ in the cost-based MDP $\overline{\mathcal{M}}$ (Example 4). Later in Section 3.3.1, we prove that this intuition is indeed correct: among all intervention rules that provide optimal safety, the rule $\mathcal{G}^* = (\overline{Q}^*, \overline{\pi}^*, 0)$ provides the largest free space for data collection (i.e., small $P_{\mathcal{G}}(\pi^*)$ in Proposition 1) among the safest intervention rules. Finally, Proposition 3 shows that an approximation of any σ-admissible intervention rule (such as one learned from data or inferred from an inaccurate model, see (Cheng et al., 2021)) is also a reasonable intervention rule (Example 5). As learning continues in SAILR, we can use the newly collected data from $\mathcal{M}$ to refine our estimate of the ideal $\overline{Q}$, such as by performing additional policy evaluation for μ or policy optimization to find $\overline{Q}^*$ of the cost-based MDP $\overline{\mathcal{M}}$.

General Backup Policies To conclude this section, we briefly discuss how to extend the above results to work with general backup policies that may take actions outside $\mathcal{A}$ (i.e., the actions the learner policy aims to use), as in (Turchetta et al., 2020). For example, such a backup policy can be implemented through an external kill switch in a robotics system. For SAILR's theoretical guarantees to hold in this case, we require one extra assumption: for all $(s, a) \in \mathcal{I}$ that can be reached from d_0 with some policy, there must be some $a' \in \mathcal{A}$ such that $\overline{A}(s, a') = \overline{Q}(s, a') - \overline{Q}(s, \mu) \leq \eta$. In other words, for every state-action we can reach from d_0 that will be overridden, there must an alternative action *in the agent's action space* $\mathcal{A}$ that keeps the agent's policy from being intervened. This condition is a generalization of Definition 2 introduced later for our analysis (a condition we call *partial*), which is essential to the unconstrained policy optimization reduction in SAILR (Section 3.3.2). Note that while this condition holds trivially when backup policy μ takes only actions in $\mathcal{A}$, generally the validity of this condition depends on the details of μ and transition dynamics P.

3.2. Absorbing MDP

SAILR performs policy optimization by running a base RL algorithm $\mathcal{F}$ to solve a new unconstrained MDP $\widetilde{\mathcal{M}}$. In this section, we define $\widetilde{\mathcal{M}}$ and discuss how to simulate experiences of π in $\widetilde{\mathcal{M}}$ by running the shielded policy $\pi' = \mathcal{G}(\pi)$ in the original MDP $\mathcal{M}$.

Given the MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ and the intervention set $\mathcal{I}$ in (3), we define $\widetilde{\mathcal{M}} = (\widetilde{\mathcal{S}}, \mathcal{A}, \widetilde{P}, \widetilde{r}, \gamma)$ as follows: Let $s_\dagger$ denote an absorbing state and $\widetilde{R} \leq 0$ be some problem-independent constant. The new MDP $\widetilde{\mathcal{M}}$ has the state space $\widetilde{\mathcal{S}} = \mathcal{S} \cup \{s_\dagger\}$ and modified dynamics and reward,

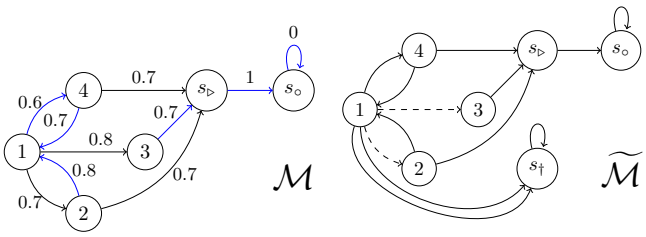


Figure 2. A simple example of the construction of $\widetilde{\mathcal{M}}$ from $\mathcal{M}$ using advantage-based intervention given by some $\mathcal{G} = (\overline{Q}, \mu, \eta)$. In $\mathcal{M}$, the transitions are deterministic, and the blue arrows correspond to actions given by μ. The edge weights correspond to $\overline{Q}$, and $\mathcal{G}$ can be verified to be 0.25-admissible when $\gamma = 0.9$. The surrogate MDP $\widetilde{\mathcal{M}}$ is formed upon intervention with $\eta = 0.05$. The transitions $1 \to 2$ and $1 \to 3$ are replaced with transitions to the absorbing state $s_\dagger$.

$$\widetilde{r}(s, a) = \begin{cases} \widetilde{R}, & (s, a) \in \mathcal{I} \\ 0, & s = s_\dagger \\ r(s, a), & \text{otherwise} \end{cases} \tag{7}$$

$$\widetilde{P}(s'|s, a) = \begin{cases} \mathbb{1}\{s' = s_\dagger\}, & (s, a) \in \mathcal{I} \text{ or } s = s_\dagger \\ P(s'|s, a), & \text{otherwise.} \end{cases} \tag{8}$$

Since $s_\dagger$ is absorbing, given a policy π defined on $\mathcal{M}$, without loss of generality we extend its definition on $\widetilde{\mathcal{M}}$ by setting $\pi(a|s_\dagger)$ to be the uniform distribution over $\mathcal{A}$. A simple example of this construction is shown in Fig. 2.

Compared with the original $\mathcal{M}$, the new MDP $\widetilde{\mathcal{M}}$ has more absorbing state-action pairs and assigns lower rewards to them. When the agent takes some $(s, a) \in \mathcal{I}$ in $\widetilde{\mathcal{M}}$, it goes to an absorbing state $s_\dagger$ and receives a *non-positive* reward. Thus, the new MDP $\widetilde{\mathcal{M}}$ gives larger penalties for taking intervened state-actions than for going into $\mathcal{S}_{\text{unsafe}}$, where we only receive zero reward. This design ensures that any nearly-optimal policy of $\widetilde{\mathcal{M}}$ will (when run in $\mathcal{M}$) have high reward and low probability of visiting intervened state-actions. As we will see, as long as $\mathcal{G}$ provides *safe* shielded polices, solving $\widetilde{\mathcal{M}}$ will lead to a safe policy with potentially good performance in the original MDP $\mathcal{M}$ even after we lift the intervention.

To simulate experiences of a policy π in $\widetilde{\mathcal{M}}$, we simply run $\pi' = \mathcal{G}(\pi)$ in the original MDP $\mathcal{M}$ and collect samples until the intervention triggers (if at all). Specifically, suppose running π' in $\mathcal{M}$ generates a trajectory $\xi = (s_0, a_0, \ldots, s_T, a'_T, \ldots)$, where T is the time step of intervention and a'_T is the first action given by the backup policy μ. Let a_T be the corresponding action from π that was overridden. We construct the trajectory $\widetilde{\xi}$ that would be generated by running π in $\widetilde{\mathcal{M}}$ by setting

$\widetilde{\xi} = (s_0, a_0, \ldots, s_T, a_T, \widetilde{s}_{T+1}, \widetilde{a}_{T+1}, \ldots)$, where $\widetilde{s}_\tau = s_\dagger$ and $\widetilde{a}_\tau$ is arbitrary for any $\tau \geq t+1$. This is valid since the two MDPs $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ share the same dynamics until the intervention happens at time step T.

3.3. Theoretical Analysis

We state the main theoretical result of SAILR, which includes the informal Proposition 1 as a special case.

Theorem 1 (Performance and Safety Guarantee at Deployment)**.** *Let $\widetilde{R} = -1$, $\mathcal{G}$ be σ-admissible, and π^* be an optimal policy for $\mathcal{M}$. If $\hat{\pi}$ is an ε-suboptimal policy for $\widetilde{\mathcal{M}}$, then the following performance and safety guarantees hold for $\hat{\pi}$ in $\mathcal{M}$:*

$$V^*(d_0) - V^{\hat{\pi}}(d_0) \leq \frac{2}{1-\gamma} P_{\mathcal{G}}(\pi^*) + \varepsilon$$

$$\overline{V}^{\hat{\pi}}(d_0) \leq \overline{Q}(d_0, \mu) + \frac{\min\{\sigma + \eta, 2\gamma\}}{1-\gamma} + \varepsilon.$$

where $P_{\mathcal{G}}(\pi^*) \coloneqq (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathrm{Prob}(\xi^h \cap \mathcal{I} \neq \varnothing \mid \pi^*, \mathcal{M})$ is the probability that π^* visits $\mathcal{I}$ in $\mathcal{M}$.

Theorem 1 shows that, when the base RL algorithm $\mathcal{F}$ finds an ε-suboptimal policy $\hat{\pi}$ in $\widetilde{\mathcal{M}}$, this policy $\hat{\pi}$ is also close to ε-suboptimal in the CMDP in (2), as long as running the optimal policy π^* in $\mathcal{M}$ will result in low probability of visiting state-actions that would be intervened by $\mathcal{G}$ (i.e., $P_{\mathcal{G}}(\pi^*)$ is small). In addition, the policy $\hat{\pi}$ is almost as safe as the backup policy μ, since $\overline{Q}(d_0, \mu)$ can be viewed as an upper bound of $\overline{Q}^{\mu}(d_0, \mu)$. The safety deterioration can be made small when the suboptimality ε, intervention threshold η, and imperfect admissibility σ of $\mathcal{G}$ are small. The proof of Theorem 1 follows directly from Theorem 2 and Proposition 7 below, which are main properties of the advantage-based intervention rules and the absorbing MDPs in SAILR. We now discuss these properties in more detail.

3.3.1. Intervention Rules

First, we show that the shielded policy π' produced by a σ-admissible intervention rule $\mathcal{G} = (\overline{Q}, \mu, \eta)$ has a small unsafe cost if backup policy μ has a small cost.

Theorem 2 (Safety of Shielded Policy)**.** *Let $\mathcal{G} = (\overline{Q}, \mu, \eta)$ be σ-admissible as per Definition 1. For any policy π, let $\pi' = \mathcal{G}(\pi)$. Then,*

$$\overline{V}^{\pi'}(d_0) \leq \overline{Q}(d_0, \mu) + \frac{\min\{\sigma + \eta, 2\gamma\}}{1-\gamma}. \tag{9}$$

Next we provide a formal statement that $\mathcal{G}^* = (\overline{Q}^*, \overline{\pi}^*, 0)$ is the optimal intervention rule that gives the largest free space for policy optimization, among the safest intervention rules. The size of the free space provided $\mathcal{G}^*$ is captured as $\mathrm{Supp}_{\mathcal{S}\times\mathcal{A}}(\widetilde{d}^{*,\pi})$, which can be interpreted as the state-actions that $\mathcal{G}^*(\pi)$ can explore before any intervention is triggered.

Proposition 4. *Let $\overline{\pi}^*$ be an optimal policy for $\overline{\mathcal{M}}$, $\overline{Q}^*$ be its state-action value function, and $\overline{V}^*$ be its state value function. Let $G_0 = \{(\overline{Q}, \mu, 0) : (\overline{Q}, \mu, 0)$ is admissible, $\overline{Q}(d_0, \mu) = \overline{V}^*(d_0)\}$. Let $\mathcal{G}^* = (\overline{Q}^*, \overline{\pi}^*, 0) \in G_0$. Consider arbitrary $\mathcal{G} \in G_0$ and policy π. Let $\widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{M}}^*$ be the absorbing MDPs induced by $\mathcal{G}$ and $\mathcal{G}^*$, respectively, and let $\widetilde{d}^{\pi}$ and $\widetilde{d}^{*,\pi}$ be their state-action distributions of π. Then,*

$$\mathrm{Supp}_{\mathcal{S}\times\mathcal{A}}(\widetilde{d}^{\pi}) \subseteq \mathrm{Supp}_{\mathcal{S}\times\mathcal{A}}(\widetilde{d}^{*,\pi}),$$

where $\mathrm{Supp}_{\mathcal{S}\times\mathcal{A}}(d)$ denotes the support of a distribution d when restricted on $\mathcal{S}\times\mathcal{A}$.

Finally, we highlight a property of the intervention set $\mathcal{I}$ of our advantage-based rules, which is crucial for the unconstrained MDP reduction described in the next section.

Definition 2. A set $\mathcal{X} \subset \mathcal{S}_{\mathrm{safe}} \times \mathcal{A}$ is called *partial* if for every $(s,a) \in \mathcal{X}$, there is some $a' \in \mathcal{A}$ such that $(s,a') \notin \mathcal{X}$.

Proposition 5. *If $\eta \geq 0$, then $\mathcal{I}$ in (3) is partial.*

Proof. For $(s,a) \in \mathcal{I}$, define $a' = \arg\min_{a'' \in \mathcal{A}} \overline{Q}(s, a'')$. Because $\overline{A}(s,a') = \overline{Q}(s,a') - \overline{Q}(s,\mu) \leq 0 \leq \eta$, we conclude that $(s,a') \notin \mathcal{I}$. □

3.3.2. Absorbing MDP

As discussed in Section 3.2, the new MDP $\widetilde{\mathcal{M}}$ provides a pessimistic value estimate of $\mathcal{M}$ by penalizing trajectories that trigger the intervention rule $\mathcal{G}$. Precisely, we can show that the amount of pessimism introduced on a policy π is proportional to $P_{\mathcal{G}}(\pi)$ (the probability of triggering the intervention rule $\mathcal{G}$ when running π in $\mathcal{M}$).

Lemma 1. *For every policy π, it holds that*

$$|\widetilde{R}|\, P_{\mathcal{G}}(\pi) \leq V^{\pi}(d_0) - \widetilde{V}^{\pi}(d_0) \leq \left(|\widetilde{R}| + \frac{1}{1-\gamma}\right) P_{\mathcal{G}}(\pi).$$

As a result, one would intuitively imagine that an optimal policy of $\widetilde{\mathcal{M}}$ would never visit the intervention set $\mathcal{I}$ at all. Below we show that this intuition is correct. Importantly, we highlight that this property holds *only* because the intervention set $\mathcal{I}$ used here is *partial* (Proposition 5). If we were to construct an absorbing MDP $\widetilde{\mathcal{M}}'$ described in Section 3.2 using an arbitrary non-partial subset $\mathcal{I}' \subseteq \mathcal{S}_{\mathrm{safe}} \times \mathcal{A}$, then the optimal policy of $\widetilde{\mathcal{M}}'$ can still enter $\mathcal{I}'$ for any $\widetilde{R} > -\infty$, because an optimal policy of $\widetilde{\mathcal{M}}'$ can use earlier rewards to mitigate penalties incurred in $\mathcal{I}'$ (Appendix B.1).

Proposition 6. *If $\widetilde{R}$ is negative and $\mathcal{G}$ induces a partial $\mathcal{I}$, then every optimal policy $\widetilde{\pi}^*$ of $\widetilde{\mathcal{M}}$ satisfies $P_{\mathcal{G}}(\widetilde{\pi}^*) = 0$.*

The partial property of $\mathcal{I}$ enables our unconstrained MDP reduction, which relates the performance and safety of a

policy π in the original MDP $\mathcal{M}$ to the suboptimality in the new MDP $\widetilde{\mathcal{M}}$ and the safety of $\pi' = \mathcal{G}(\pi)$.

Proposition 7 (Suboptimality in $\widetilde{\mathcal{M}}$ to Suboptimality and Safety in $\mathcal{M}$)**.** *Let $\widetilde{R}$ be negative. For any policy π, let π' be the shielded policy defined in (4). Let π^* be an optimal policy for $\mathcal{M}$. Suppose π is ε-suboptimal for $\widetilde{\mathcal{M}}$. Then the following performance and safety guarantees hold in $\mathcal{M}$:*

$$V^{\pi^*}(d_0) - V^{\pi}(d_0) \leq \left(|\widetilde{R}| + \frac{1}{1-\gamma}\right) P_{\mathcal{G}}(\pi^*) + \varepsilon$$
$$\overline{V}^{\pi}(d_0) \leq \overline{V}^{\pi'}(d_0) + \frac{\varepsilon}{|\widetilde{R}|}.$$

4. Related Work

CMDPs (Altman, 1999) have been a popular framework for safe RL as it side-steps the reward design problem for ensuring safety in a standard MDP (Geibel & Wysotzki, 2005; Shalev-Shwartz et al., 2016). Most existing CMDP-based safe RL algorithms closely follow algorithms in the constrained optimization literature (Bertsekas, 2014). They can be classified into either online or offline schemes. Online schemes learn by coupling the iteration of a numerical optimization algorithm (notably primal-dual gradient updates) with data collection (Borkar, 2005; Chow et al., 2017; Tessler et al., 2018; Bohez et al., 2019), and these algorithms have also been studied in the exploration context (Ding et al., 2020; Qiu et al., 2020; Efroni et al., 2020). However, they have no guarantees on policy safety during training. Offline schemes (Achiam et al., 2017; Bharadhwaj et al., 2021; Le et al., 2019; Efroni et al., 2020), on the other hand, separate optimization and data collection. They conservatively enforce safety constraints on every policy iterate but are more difficult to scale up. Many of these constrained algorithms for CMDPs, however, have worse numerical stability compared with typical RL algorithms for MDPs, because of the nonconvex saddle-point of the CMDP (Lee et al., 2017; Chow et al., 2018).

Another line of safe RL research uses control-theoretic techniques to enforce safe exploration, though only few provide guarantees with respect to the CMDP in (2). These methods include restricting the agent to take actions that lead to next-state safety (Dalal et al., 2018; Wabersich & Zeilinger, 2018) or states where a safe backup exists (Hans et al., 2008; Polo & Rebollo, 2011; Li & Bastani, 2020). Other works consider more structured shielding approaches, including those with temporal logic safety rules and backup policies (Alshiekh et al., 2018) and neurosymbolic policies (Anderson et al., 2020) whose safety can be checked easily. Many of these approaches require strong assumptions on the MDP (e.g., taking an action to ensure the next state's safety being sufficient to imply all future states will continue to have such safe actions available). Algorithms based on Lyapunov functions and reachability (Perkins & Barto, 2002; Chow et al., 2018; 2019; Berkenkamp et al., 2017; Fisac et al., 2018) address the long-term feasibility issue, but they are more complicated than common RL algorithms. We note that our admissible intervention rules in (6) can be viewed as a state-action Lyapunov function.

To the best of our knowledge, SAILR is the first unconstrained method that provides formal guarantees with respect to the CMDP objective. The closest work to ours is (Turchetta et al., 2020), which also uses the idea of intervention for training safety and trains the agent in a new MDP that discourages visiting intervened state-actions. However, their algorithm, CISR, is still based on calling CMDP subroutines (Le et al., 2019). They neither specify how the intervention rules can be constructed nor provide performance guarantees. By comparsion, we provide a general recipe of intervention rules and obtain the properties desired in (Turchetta et al., 2020) by simply unconstrained RL.

5. Experiments

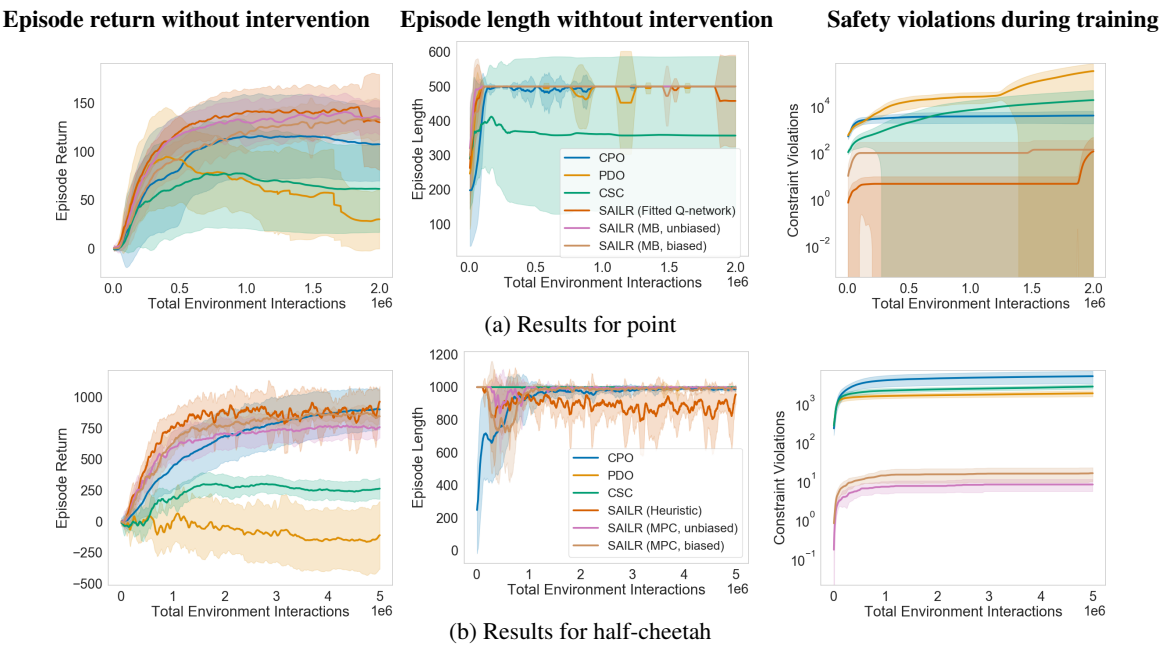
We conduct experiments to corroborate our theoretical analysis of SAILR. We aim to verify whether a properly designed intervention mechanism can drastically reduce the amount of unsafe trajectories generated in training while still resulting in good safety and performance in deployment.

Our experiments consider two different tasks: *1)* A toy point robot based on (Achiam et al., 2017) that gets reward for following a circular path at high speed, but is constrained to stay in a region smaller than the target circle; and *2)* a half-cheetah that gets reward equal to its forward velocity, with one of its links constrained to remain in a given height range, outside of which the robot is deemed to be unsafe. In all experiments, when computing $\overline{Q}$, we opt to use a *shaped* cost function in place of the original sparse indicator cost function to make our intervention mechanism more conservative (and hence the training process safer). In particular, this shaped cost function is a function of the distance to the unsafe set and is an upper bound of the original sparse cost. The appendix includes some additional experiments where the original sparse cost is used.

We implement SAILR by using PPO (Schulman et al., 2017) as the RL subroutine. We also compare our approach to two CMDP-based approaches: CPO (Achiam et al., 2017) and a primal-dual optimization (PDO) algorithm (Chow et al., 2017). For the PDO algorithm, we use PPO as the policy optimization subroutine and dual gradient ascent as the Lagrange multiplier update. We also consider a variant of PDO, called CSC, where a learned conservative critic is used to filter unsafe actions (Bharadhwaj et al., 2021).

5.1. Point Robot

Here SAILR uses the intervention rule $\mathcal{G} = (\mu, \overline{Q}, \eta)$: the baseline policy μ aims to stop the robot by deceleration.



(a) Results for point

(b) Results for half-cheetah

Figure 3. Results of SAILR and baseline CMDP-based methods. Overall SAILR dramatically reduces the amount of safety constraint violations while still having large returns at deployment. Plots in a row share the same legend. All error bars are ± 1 standard deviation over 10 (point robot) or 8 (half-cheetah) random seeds. Any curve not plotted in the third column corresponds to zero safety violations.

The function $\overline{Q}$ is estimated by either querying a fitted Q-network or by rolling out μ on a dynamical model (denoted "MB" in Fig. 3a) of the point robot and querying a shaped cost function. We consider both biased and unbiased models (details in Appendix C.1). Fig. 3a show the main experimental results, with all three instances of SAILR outperforming the baselines on all three metrics. For SAILR, the shielding prevents many safety violations, and the unconstrained approach allows for reliable convergence as opposed to the baselines which rely on elaborate constrained approaches.

5.2. Half-Cheetah

We consider two intervention rules in SAILR: a reset backup policy μ with a simple heuristic $\overline{Q}$ based on the predicted height of the link after taking a proposed action, and a reset backup policy μ based on a sampling-based model predictive control (MPC) algorithm (Williams et al., 2017; Bhardwaj et al., 2021) with a model-based value estimate (i.e., $\overline{Q} \approx \overline{Q}^{\mu}$). The simple heuristic uses a slightly smaller height range for intervention to attempt to construct a *partial* intervention set (Section 3.3). The MPC algorithm optimizes a control sequence over the same cost function. The function $\overline{Q}$ is computed by rolling out this control sequence on the dynamical model and querying the cost function. We also consider model bias in the MPC experiments (details in Appendix C.2).

As with the point environment, SAILR incurs orders of magnitude fewer safety violations than the baselines (right plot of Fig. 3b), with all three instances having comparable deployment performance to that of CPO. Though the heuristic intervention violates no constraints in training, it is consistently unsafe in deployment (middle plot), likely because the resulting intervention set is not partial. On the other hand, MPC-based approaches are consistently safe in deployment, owing to its multi-step lookahead yielding an intervention rule that is likely to be σ-admissible (and therefore give an intervention set that is partial).

6. Conclusion

We presented an intervention-based method for safe reinforcement learning. By utilizing advantage functions for intervention and penalizing an agent for taking intervened actions, we can use unconstrained RL algorithms in the safe learning domain. Our analysis shows that using advantage functions for the intervention decision gives strong guarantees for safety during training and deployment, with the performance only limited by how often the true optimal policy would be intervened. We also discussed ways of synthesizing good intervention rules, such as using value iteration techniques. Finally, our experiments showed that the shielded policy violates few if any constraints during training while the corresponding deployed policy enjoys convergence to a large return.

Acknowledgements

This work was supported in part by ARL SARA CRA W911NF-20-2-0095. We thank Mohak Bhardwaj for providing MPC code used in the half-cheetah experiment. We thank Anqi Li for insightful discussions and Panagiotis Tsiotras for helpful comments on the paper.

References

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

Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., and Topcu, U. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Altman, E. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anderson, G., Verma, A., Dillig, I., and Chaudhuri, S. Neurosymbolic reinforcement learning with formally verified exploration. *arXiv preprint arXiv:2009.12612*, 2020.

Berkenkamp, F., Turchetta, M., Schoellig, A., and Krause, A. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, pp. 908–918, 2017.

Bertsekas, D. P. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena Scientific, Belmont, MA, 4th edition, 2017.

Bharadhwaj, H., Kumar, A., Rhinehart, N., Levine, S., Shkurti, F., and Garg, A. Conservative Safety Critics for Exploration. In *International Conference on Learning Representations*, 2021.

Bhardwaj, M., Choudhury, S., and Boots, B. Blending MPC & Value Function Approximation for Efficient Reinforcement Learning. In *International Conference on Learning Representations*, 2021.

Bohez, S., Abdolmaleki, A., Neunert, M., Buchli, J., Heess, N., and Hadsell, R. Value constrained model-free continuous control. *arXiv preprint arXiv:1902.04623*, 2019.

Borkar, V. S. An actor-critic algorithm for constrained markov decision processes. *Systems & Control Letters*, 54(3):207–213, 2005.

Cheng, C.-A., Kolobov, A., and Agarwal, A. Policy improvement via imitation of multiple oracles. *Advances in Neural Information Processing Systems*, 33, 2020.

Cheng, C.-A., Kolobov, A., and Swaminathan, A. Heuristic-Guided Reinforcement Learning. *arXiv preprint arXiv:2106.02757*, 2021.

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

Chow, Y., Nachum, O., Duenez-Guzman, E., and Ghavamzadeh, M. A Lyapunov-Based Approach to Safe Reinforcement Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8092–8101, 2018.

Chow, Y., Nachum, O., Faust, A., Duenez-Guzman, E., and Ghavamzadeh, M. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019.

Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., and Tassa, Y. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanović, M. R. Provably efficient safe exploration via primal-dual policy optimization. *arXiv preprint arXiv:2003.00534*, 2020.

Efroni, Y., Mannor, S., and Pirotta, M. Exploration-Exploitation in Constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

Eysenbach, B., Gu, S., Ibarz, J., and Levine, S. Leave No Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning. In *International Conference on Learning Representations*, 2018.

Facchinei, F. and Pang, J.-S. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

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

García, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Geibel, P. and Wysotzki, F. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

Hans, A., Schneegaß, D., Schäfer, A. M., and Udluft, S. Safe exploration for reinforcement learning. In *ESANN*, pp. 143–148. Citeseer, 2008.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-Order Methods Almost Always Avoid Saddle Points. *arXiv preprint arXiv:1710.07406*, 2017.

Li, S. and Bastani, O. Robust model predictive shielding for safe reinforcement learning with stochastic dynamics. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7166–7172. IEEE, 2020.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Perkins, T. J. and Barto, A. G. Lyapunov design for safe reinforcement learning. *Journal of Machine Learning Research*, 3(Dec):803–832, 2002.

Polo, F. J. G. and Rebollo, F. F. Safe reinforcement learning in high-risk tasks through policy improvement. In *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 76–83. IEEE, 2011.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. Upper Confidence Primal-Dual Reinforcement Learning for CMDP with Adversarial Loss. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Srinivasan, K., Eysenbach, B., Ha, S., Tan, J., and Finn, C. Learning to be Safe: Deep RL with a Safety Critic. *arXiv preprint arXiv:2010.14603*, 2020.

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

Tessler, C., Mankowitz, D. J., and Mannor, S. Reward Constrained Policy Optimization. In *International Conference on Learning Representations*, 2018.

Thananjeyan, B., Balakrishna, A., Nair, S., Luo, M., Srinivasan, K., Hwang, M., Gonzalez, J. E., Ibarz, J., Finn, C., and Goldberg, K. Recovery RL: Safe reinforcement learning with learned recovery zones. *arXiv preprint arXiv:2010.15920*, 2020.

Turchetta, M., Kolobov, A., Shah, S., Krause, A., and Agarwal, A. Safe reinforcement learning via curriculum induction. *Advances in Neural Information Processing Systems*, 33, 2020.

Wabersich, K. P. and Zeilinger, M. N. Safe exploration of nonlinear dynamical systems: A predictive safety filter for reinforcement learning. *arXiv preprint arXiv:1812.05506*, 2018.

Williams, G., Wagener, N., Goldfain, B., Drews, P., Rehg, J. M., Boots, B., and Theodorou, E. A. Information Theoretic MPC for Model-Based Reinforcement Learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1714–1721. IEEE, 2017.