A. Missing Proofs

A.1. Useful Lemmas

Lemma 2. *For any γ-discounted MDP with reward function r, the identity $V^\pi(d_0) = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h U_h^\pi(d_0)$ holds, where $U_h^\pi(d_0) = \mathbb{E}_{\rho^\pi}[\sum_{t=0}^{h} r(s_t, a_t)]]$ is the undiscounted h-step return.*

Proof. The proof follows from exchanging the order of summations:

$$\begin{aligned}
(1-\gamma)\sum_{h=0}^{\infty}\gamma^h U_h^\pi(d_0) &= (1-\gamma)\sum_{h=0}^{\infty}\gamma^h \mathbb{E}_{\rho^\pi}\left[\sum_{t=0}^{h} r(s_t,a_t)\right] \\
&= (1-\gamma)\mathbb{E}_{\rho^\pi}\left[\sum_{t=0}^{\infty} r(s_t,a_t)\sum_{h=t}^{\infty}\gamma^h\right] \\
&= \mathbb{E}_{\rho^\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] \\
&= V^\pi(d_0)
\end{aligned}$$

□

Lemma 3 (Performance Difference Lemma (Kakade & Langford, 2002; Cheng et al., 2020))**.** *Let $\mathcal{M}$ be an MDP and π be a policy. For any function $f : \mathcal{S} \to \mathbb{R}$ and any initial state distribution d_0, it holds that*

$$V^\pi(d_0) - f(d_0) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}[r(s,a) + \gamma\mathbb{E}_{s'|s,a}[f(s')] - f(s)]$$

Corollary 1. *Let $\mathcal{M}$ and $\widehat{\mathcal{M}}$ be MDPs with common state and action spaces. For any policy π, the difference in value functions in $\mathcal{M}$ and $\widehat{\mathcal{M}}$ satisfies*

$$V^\pi(d_0) - \widehat{V}^\pi(d_0) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}[(\mathcal{D}^\pi\widehat{Q}^\pi)(s,a)]$$

where $\mathcal{D}^\pi$ is the temporal-difference operator of $\mathcal{M}$:

$$(\mathcal{D}^\pi Q)(s,a) \coloneqq (\mathcal{B}^\pi Q)(s,a) - Q(s,a),$$

and $\mathcal{B}^\pi$ is the Bellman operator of $\mathcal{M}$:

$$(\mathcal{B}^\pi Q)(s,a) \coloneqq r(s,a) + \gamma\mathbb{E}_{s'|s,a}[Q(s',\pi)].$$

Proof. Set $f = \widehat{V}^\pi$ and observe that $\widehat{V}^\pi(s) = \widehat{Q}^\pi(s,\pi)$. □

A.2. Proof of Equivalent CMDP Formulation in Section 2

Here we show that (1) and (2) are the same by proving the equivalence

$$(1-\gamma)\sum_{h=0}^{\infty}\gamma^h \mathrm{Prob}(\xi_h \subset \mathcal{S}_{\mathrm{safe}}|\pi) \geq 1-\delta \quad \Longleftrightarrow \quad \overline{V}^\pi(d_0) \leq \delta \tag{10}$$

By the definition of the cost function $c(s,a) = \mathbb{1}\{s = s_\triangleright\}$ and absorbing property of $\mathcal{S}_{\mathrm{unsafe}} = \{s_\triangleright, s_\circ\}$, we can write

$$1 - \mathrm{Prob}(\xi_h \subset \mathcal{S}_{\mathrm{safe}}|\pi) = \mathrm{Prob}(s_\triangleright \in \xi_h|\pi) = \mathbb{E}_{\rho^\pi}\left[\sum_{t=0}^{h} c(s_t,a_t)\right] \tag{11}$$

since $s_\triangleright$ can only appear at most once within ξ_h. Substituting this equality into the negation of the chance constraint,

$$\begin{aligned}1-(1-\gamma)\sum_{h=0}^{\infty}\gamma^h\mathrm{Prob}(\xi_h\subset\mathcal{S}_{\mathrm{safe}}|\pi) &= (1-\gamma)\sum_{h=0}^{\infty}\gamma^h\mathbb{E}_{\rho^\pi}\left[\sum_{t=0}^{h}c(s_t,a_t)\right]\\ &= \mathbb{E}_{\rho^\pi}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right]\\ &= \overline{V}^\pi(d_0)\end{aligned}$$

where the second equality follows from Lemma 2. Therefore, (10) holds.

A.3. Proof for Intervention Rules in Section 3

A.3.1. Admissible Rules and Pessimism

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

Proof of Proposition 2. The proof follows by repeating the inequality of $\overline{Q}$.

$$\begin{aligned}\overline{Q}(s,a) &\ge c(s,a)+\gamma\mathbb{E}_{s'|s,a}[\overline{Q}(s',\mu)]-\sigma\\ &\ge c(s,a)+\gamma\mathbb{E}_{s'|s,a}\left[c(s',\mu),+\gamma\mathbb{E}_{s''|s',\mu}[\overline{Q}^\mu(s'',\mu)]\right]-(1+\gamma)\sigma\\ &\ \ \vdots\\ &\ge \overline{Q}^\mu(s',\mu)-\frac{\sigma}{1-\gamma}.\end{aligned}$$

□

A.3.2. Example Intervention Rules

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 $s\in\mathcal{S}_{\mathrm{safe}}$ and $a\in\mathcal{A}$. If $\|\hat{Q}-\overline{Q}\|_\infty\le\delta$, then $\hat{\mathcal{G}}=(\hat{Q},\mu,\eta)$ is $(\sigma+(1+\gamma)\delta)$-admissible.*

Proof of Proposition 3. We show each intervention rule $\mathcal{G}=(\overline{Q},\mu,\eta)$ below satisfies the admissibility condition

$$\overline{Q}(s,a)+\sigma\ge c(s,a)+\gamma\mathbb{E}_{s'\sim P|s,a}[\overline{Q}(s',\mu)].$$

For convenience, we define the Bellman operator $\overline{\mathcal{B}}^\mu$ as $\overline{\mathcal{B}}^\mu Q(s,a)\coloneqq c(s,a)+\gamma\mathbb{E}_{s'|s,a}[Q(s,\mu)]$. Then the admissibility condition can be written as $\overline{Q}(s,a)+\sigma\ge(\overline{\mathcal{B}}^\mu\overline{Q})(s,a)$ for any $s\in\mathcal{S}_{\mathrm{safe}}$ and $a\in\mathcal{A}$. Also, we write $\overline{Q}\in[0,\gamma]$ on $\mathcal{S}_{\mathrm{safe}}$ if $\overline{Q}(s,a)\in[0,\gamma]$ for all $s\in\mathcal{S}_{\mathrm{safe}}$ and $a\in\mathcal{A}$.

1. **Baseline policy**: We know $\mathcal{G} = (\overline{Q}^{\mu}, \mu, \eta)$ is admissible since $\overline{Q}^{\mu} = \overline{\mathcal{B}}^{\mu}\overline{Q}^{\mu}$. For $\mathcal{G} = (\overline{Q}^{\mu}, \mu^{+}, \eta)$, we have $\overline{Q}^{\mu} \geq \overline{\mathcal{B}}^{\mu^{+}}\overline{Q}^{\mu}$ since μ^{+} is greedy with respect to $\overline{Q}^{\mu}$. Also, by the definition of the cost c and transition dynamics P, we know that $\overline{Q}^{\mu}(s,a) \in [0,1]$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Furthermore, when $s \in \mathcal{S}_{\text{safe}}$, we have $c(s,a)$ and therefore $\overline{Q}^{\mu}(s,a) = \gamma \mathbb{E}_{s'|s,a}[\overline{Q}^{\mu}(s',\mu)] \in [0,\gamma]$.

2. **Composite intervention**: For any $k \in \{1,\dots,K\}$, the following bound holds:

$$\begin{aligned}
(\overline{\mathcal{B}}^{\mu_{\min}}\overline{Q}_{\min})(s,a) &= c(s,a) + \gamma \mathbb{E}_{s'|s,a}[\overline{Q}_{\min}(s',\mu_{\min})] \\
&\leq c(s,a) + \gamma \mathbb{E}_{s'|s,a}[\overline{Q}_{\min}(s',\mu_k)] \\
&\leq c(s,a) + \gamma \mathbb{E}_{s'|s,a}[\overline{Q}_k(s',\mu_k)] \\
&\leq \overline{Q}_k(s,a) + \sigma_k \\
&\leq \overline{Q}_k(s,a) + \sigma_{\max},
\end{aligned}$$

where the first inequality comes from $\mu_{\min}$ being a minimizer of $\overline{Q}_{\min}$, and the second inequality from $\overline{Q}_{\min}$ being a pointwise minimum of $\{\overline{Q}_k\}_{k=1}^{K}$. Since this holds for every k, we conclude:

$$\begin{aligned}
(\overline{\mathcal{B}}^{\mu_{\min}}\overline{Q}_{\min})(s,a) &\leq \min_k \left[\overline{Q}_k(s,a) + \sigma_{\max}\right] \\
&= \min_k \overline{Q}_k(s,a) + \sigma_{\max} \\
&= \overline{Q}_{\min}(s,a) + \sigma_{\max},
\end{aligned}$$

which establishes the Bellman bound holds. Finally, since each $\overline{Q}_k$ satisfies $\overline{Q}_k \in [0,\gamma]$ on $\mathcal{S}_{\text{safe}}$, we conclude that $\overline{Q}_{\min}$ has the same range. Therefore, $\mathcal{G}$ is $\sigma_{\max}$-admissible.

3. **Value iteration**: Define shortcuts $\overline{Q}_k := \overline{\mathcal{T}}^k\overline{Q}$, where $\overline{Q}_0 = \overline{Q}$.

We first show that, by policy improvement, we have $\overline{Q}_k(s,a) \leq \overline{Q}_{k-1}(s,a) + \gamma^{k-1}\sigma$ on $\mathcal{S}_{\text{safe}} \times \mathcal{A}$. We do this by induction. First, we see that:

$$\begin{aligned}
\overline{Q}_1(s,a) &= \overline{\mathcal{T}}\,\overline{Q}_0(s,a) \\
&= c(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[\min_{a'} \overline{Q}_0(s',a')\right] \\
&= c(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[\min_{a'} \overline{Q}(s',a')\right] \\
&\leq c(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[\overline{Q}(s',\mu)\right] \\
&\leq \overline{Q}(s,a) + \sigma \\
&= \overline{Q}_0(s,a) + \sigma.
\end{aligned}$$

Now suppose $\overline{Q}_\kappa(s,a) \leq \overline{Q}_{\kappa-1}(s,a) + \gamma^{\kappa-1}\sigma$ holds on $\mathcal{S}_{\text{safe}} \times \mathcal{A}$ for some κ. Therefore,

$$\begin{aligned}
\overline{Q}_{\kappa+1}(s,a) &= \overline{\mathcal{T}}\,\overline{Q}_\kappa(s,a) \\
&= c(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[\min_{a'} \overline{Q}_\kappa(s',a')\right] \\
&\leq c(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[\min_{a'} \overline{Q}_{\kappa-1}(s',a')\right] + \gamma^{\kappa}\sigma \\
&= \overline{\mathcal{T}}\,\overline{Q}_{\kappa-1}(s,a) + \gamma^{\kappa}\sigma \\
&= \overline{Q}_\kappa(s,a) + \gamma^{\kappa}\sigma.
\end{aligned}$$

Using this inequality, we now show that $\mathcal{G}^k = (\overline{Q}_k, \mu^k, \eta)$ is indeed $\gamma^k\sigma$-admissible:

$$\begin{aligned}
\overline{Q}_k(s,a) &= \overline{\mathcal{T}}\,\overline{Q}_{k-1}(s,a) \\
&= c(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[\min_{a'} \overline{Q}_{k-1}(s',a')\right] \\
&\geq c(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[\min_{a'} \overline{Q}_{k}(s',a')\right] - \gamma^k\sigma \\
&= \overline{\mathcal{T}}\,\overline{Q}_k(s,a) - \gamma^k\sigma \\
&= \overline{\mathcal{B}}^{\mu^k}\overline{Q}_k(s,a) - \gamma^k\sigma,
\end{aligned}$$

where the inequality was used in the third line. This establishes the Bellman bound holds.

We prove that $\overline{Q}_k \in [0,\gamma]$ on $\mathcal{S}_{\text{safe}}$ by induction. Clearly, $\overline{Q}_0 = \overline{Q} \in [0,\gamma]$ on $\mathcal{S}_{\text{safe}}$ since $\mathcal{G}$ is σ-admissible. Now suppose $\overline{Q}_\kappa \in [0,\gamma]$ on $\mathcal{S}_{\text{safe}}$ for some κ. Then, for any $s \in \mathcal{S}_{\text{safe}}$ and $a \in \mathcal{A}$, we have $\overline{Q}_{\kappa+1}(s,a) = \gamma\mathbb{E}_{s'|s,a}[\min_{a'} \overline{Q}_\kappa(s,a)] \in [0,\gamma]$. Therefore, $\mathcal{G}^k$ is $\gamma^k\sigma$-admissible.

4. **Optimal intervention**: This is a special case of case 1.

5. **Approximation**: The following holds on $\mathcal{S}_{\text{safe}} \times \mathcal{A}$:

$$\begin{aligned}
\hat{Q}(s,a) &= \hat{Q}(s,a) - \overline{Q}(s,a) + \overline{Q}(s,a) \\
&\geq -\delta + (\overline{\mathcal{B}}^{\mu}\overline{Q})(s,a) - \sigma \\
&= -\delta - \sigma + c(s,a) + \gamma\mathbb{E}_{s'|s,a}[\overline{Q}(s',\mu)] \\
&\geq -\delta - \sigma + c(s,a) + \gamma\mathbb{E}_{s'|s,a}[\hat{Q}(s',\mu) - \delta] \\
&= -\delta - \sigma - \gamma\delta + \overline{\mathcal{B}}^{\mu}\hat{Q}(s,a).
\end{aligned}$$

That is, $\overline{\mathcal{B}}^{\mu}\hat{Q}(s,a) \leq \hat{Q}(s,a) + \sigma + (1+\gamma)\delta$. Therefore, $\hat{\mathcal{G}} = (\hat{Q}, \mu, \eta)$ is $(\sigma + (1+\gamma)\delta)$-admissible.

□

A.3.3. Safety Guarantee of Shielded Policy

Before proving Theorem 2, we prove two lemmas, one showing that the average advantage of a shielded policy satisfies the intervention threshold (Lemma 4) and the other stating that the cost-value function is equal to the expected occupancy of the unsafe set (Lemma 5).

Lemma 4. *For some policy π and intervention rule $\mathcal{G} = (\overline{Q}, \mu, \eta)$, let $\pi' \coloneqq \mathcal{G}(\pi)$ and $\overline{A}(s,a) \coloneqq \overline{Q}(s,a) - \overline{Q}(s,\mu)$. Then, $\overline{A}(s,\pi') \leq \eta$ for any $s \in \mathcal{S}_{\text{safe}}$.*

Proof. We use the definition of π' (in (4)), the facts that $\overline{A}(s,\mu) = 0$, and that $(s,a) \notin \mathcal{I}$ if and only if $\overline{A}(s,a) \leq \eta$. The following then holds:

$$\begin{aligned}
\overline{A}(s,\pi') &= \sum_{a\in\mathcal{A}} \pi'(a|s)\overline{A}(s,a) \\
&= \sum_{a:(s,a)\notin\mathcal{I}} \pi(a|s)\overline{A}(s,a) + w(s)\sum_{a\in\mathcal{A}} \mu(a|s)\overline{A}(s,a) \\
&\leq \eta \sum_{a:(s,a)\notin\mathcal{I}} \pi(a|s) + w(s)\overline{A}(s,\mu) \\
&\leq \eta \cdot 1 + w(s)\cdot 0 \\
&= \eta.
\end{aligned}$$

□

Lemma 5. *For any policy π,*

$$\mathbb{E}_{s\sim d^\pi}[\mathbb{1}\{s \in \{s_\triangleright, s_\circ\}\}] = \overline{V}^\pi(d_0).$$

Proof. We know from the definition of the cost function that $\overline{V}^\pi(d_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi}[\mathbb{1}\{s = s_\triangleright\}]$. From the absorbing property of $\mathcal{S}_{\text{unsafe}}$, we have $\mathbb{E}_{s\sim d^\pi}[\mathbb{1}\{s = s_\circ\}] = \frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d^\pi}[\mathbb{1}\{s = s_\triangleright\}]$. We can then derive

$$\begin{aligned}
\mathbb{E}_{s\sim d^\pi}[\mathbb{1}\{s \in \{s_\triangleright, s_\circ\}\}] &= \mathbb{E}_{s\sim d^\pi}[\mathbb{1}\{s = s_\triangleright\}] + \mathbb{E}_{s\sim d^\pi}[\mathbb{1}\{s = s_\circ\}] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi}[\mathbb{1}\{s = s_\triangleright\}] \\
&= \overline{V}^\pi(d_0).
\end{aligned}$$

□

We now prove the safety guarantee of the shieled policy π'.

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,*

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

Proof. To prove the upper bounds, we first extend the definition of $\overline{Q}$ onto $\mathcal{S}_{\text{unsafe}}$, where we define

$$\overline{Q}(s_\triangleright, a) = 1 \qquad \text{and} \qquad \overline{Q}(s_\circ, a) = 0 \qquad \text{for all } a \in \mathcal{A}.$$

Define $\overline{V}(s) \coloneqq \overline{Q}(s, \mu)$. Since $c(s,a) + \gamma\mathbb{E}_{s'|s,a}[\overline{V}(s')] = \overline{V}(s)$ when $s \in \{s_\triangleright, s_\circ\}$, we can use the performance difference lemma (Lemma 3) to derive

$$\begin{aligned}
\overline{V}^{\pi'}(d_0) - \overline{Q}(d_0, \mu) &= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi'}}[c(s,a) + \gamma\mathbb{E}_{s'|s,a}[\overline{V}(s')] - \overline{V}(s)] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi'}}\left[\left(c(s,a) + \gamma\mathbb{E}_{s'|s,a}[\overline{V}(s')] - \overline{V}(s)\right)\mathbb{1}\{s \notin \{s_\triangleright, s_\circ\}\}\right] \\
&\le \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi'}}\left[\left(\min\{\sigma, \gamma\} + \overline{Q}(s,a) - \overline{V}(s)\right)\mathbb{1}\{s \notin \{s_\triangleright, s_\circ\}\}\right] \\
&\le \frac{\min\{\sigma,\gamma\} + \min\{\eta,\gamma\}}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}}[\mathbb{1}\{s \notin \{s_\triangleright, s_\circ\}\}] \\
&= \frac{\min\{\sigma,\gamma\} + \min\{\eta,\gamma\}}{1-\gamma}\overline{V}^{\pi'}(d_0),
\end{aligned}$$

where the first inequality comes from $\overline{Q}$ being σ-admissible and γ-admissible, the second inequality from $\overline{A}(s, \pi') \le \eta$ (Lemma 4) and $\overline{A}(s, \pi') \le \gamma$ (Definition 1) for $s \notin \{s_\triangleright, s_\circ\}$, and the last equality from Lemma 5.

Therefore, after some algebraic rearrangement,

$$\begin{aligned}
\overline{V}^{\pi'}(d_0) &\le \frac{(1-\gamma)\overline{Q}(d_0,\mu) + \min\{\sigma,\gamma\} + \min\{\eta,\gamma\}}{1-\gamma+\min\{\sigma,\gamma\}+\min\{\eta,\gamma\}} \\
&\le \overline{Q}(d_0,\mu) + \frac{\min\{\sigma,\gamma\} + \min\{\eta,\gamma\}}{1-\gamma} \\
&\le \overline{Q}(d_0,\mu) + \frac{\min\{\sigma+\eta, 2\gamma\}}{1-\gamma}.
\end{aligned}$$

□

A.3.4. AN OPTIMAL INTERVENTION RULE

First, we show that every state-action pair visited by π' will not have an advantage function lower than that of the optimal policy for $\overline{\mathcal{M}}$.

Lemma 6. *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, $\overline{Q}(d_0, \mu) = \overline{V}^*(d_0)\}$ be a subset of admissible intervention rules with a threshold of zero and average $\overline{Q}$ that matches $\overline{V}^*$. Define $\overline{A}^*(s,a) = \overline{Q}^*(s,a) - \overline{Q}^*(s,\overline{\pi}^*)$ as the advantage function of the optimal policy. For some intervention rule $\mathcal{G} \in G_0$ and policy π, let $\pi' = \mathcal{G}(\pi)$.*

Then, the inequality $\overline{A}(s,a) \geq \overline{A}^*(s,a)$ holds for all $a \in \mathcal{A}$ almost surely over the distribution $d^{\pi'}(s)$.

Proof. First, we show by induction that running π' starting from d_0 results in the agent staying in the subset $\mathcal{S}_\mathcal{G} = \{s \in \mathcal{S} : \overline{Q}(s,\mu) = \overline{V}^*(s)\}$.

For $t = 0$, consider some $s_0 \sim d_0$. We observe from admissibility of $\mathcal{G}$ and Proposition 2 that $\overline{Q}(s,a) \geq \overline{Q}^\mu(s,a) \geq \overline{V}^*(s)$ on $\mathcal{S} \times \mathcal{A}$. Since $\overline{Q}(d_0,\mu) = \overline{V}^*(d_0)$, we conclude that $\overline{Q}(s_0,\mu) = \overline{V}^*(s_0)$. Therefore, $s_0 \in \mathcal{S}_\mathcal{G}$ almost surely over d_0.

Now suppose the agent is in $\mathcal{S}_\mathcal{G}$ with probability one at some time step t. Consider some $s_t \sim d_t^{\pi'}$ (observing that $s_t \in \mathcal{S}_\mathcal{G}$). We assume that $s_t \in \mathcal{S}_{\text{safe}}$ (otherwise, the below is trivially true as there is no intervention outside $\mathcal{S}_{\text{safe}}$). By Lemma 4 and admissibility, we can derive:

$$\begin{aligned}
0 = \eta &\geq \overline{A}(s_t, \pi') \\
&= \overline{Q}(s_t, \pi') - \overline{Q}(s_t, \mu) \\
&\geq c(s_t, \pi') + \gamma \mathbb{E}_{s_{t+1} \sim \mathcal{P}|s_t,\pi'}[\overline{Q}(s_{t+1}, \mu)] - \overline{Q}(s_t, \mu) \\
&= \gamma \mathbb{E}_{s_{t+1}|s_t,\pi'}[\overline{Q}(s_{t+1}, \mu)] - \overline{Q}(s_t, \mu) \\
&= \gamma \mathbb{E}_{s_{t+1}|s_t,\pi'}[\overline{Q}(s_{t+1}, \mu)] - \overline{V}^*(s_t) \\
&= \gamma \mathbb{E}_{s_{t+1}|s_t,\pi'}[\overline{Q}(s_{t+1}, \mu)] - \gamma \mathbb{E}_{s_{t+1}|s_t,\overline{\pi}^*}[\overline{V}^*(s_{t+1})],
\end{aligned}$$

where the second and fourth equalities are due to $s_t \in \mathcal{S}_{\text{safe}}$, and the third equality is due to $s_t \in \mathcal{S}_\mathcal{G}$. Notice also, since $s_t \in \mathcal{S}_{\text{safe}}$, we have

$$\gamma \mathbb{E}_{s_{t+1}|s_t,\pi'}[\overline{V}^*(s_{t+1})] = \overline{Q}^*(s_t, \pi') \geq \overline{Q}^*(s_t, \overline{\pi}^*) = \gamma \mathbb{E}_{s_{t+1}|s_t,\overline{\pi}^*}[\overline{V}^*(s_{t+1})].$$

Therefore, combining the two inequalities above, we have

$$\mathbb{E}_{s_{t+1}|s_t,\pi'}[\overline{V}^*(s_{t+1})] \geq \mathbb{E}_{s_{t+1}|s_t,\pi'}[\overline{Q}(s_{t+1}, \mu)].$$

Since $\overline{Q}(s,a) \geq \overline{V}^*(s)$ on $\mathcal{S} \times \mathcal{A}$, by the same argument we made for s_0, we conclude $\overline{Q}(s_{t+1}, \mu) = \overline{V}^*(s_{t+1})$ with probability one. Therefore, the agent stays in the subset $\mathcal{S}_\mathcal{G}$.

With this property in mind, let $s \sim d^{\pi'}$. Then the following holds for all $a \in \mathcal{A}$:

$$\begin{aligned}
\overline{A}(s,a) &= \overline{Q}(s,a) - \overline{Q}(s,\mu) \\
&= \overline{Q}(s,a) - \overline{Q}^*(s,\overline{\pi}^*) \\
&\geq \overline{Q}^*(s,a) - \overline{Q}^*(s,\overline{\pi}^*) = \overline{A}^*(s,a),
\end{aligned}$$

where the second equality is due to $\overline{Q}(s,\mu) = \overline{V}^*(s) = \overline{Q}^*(s,\overline{\pi}^*)$ on $\mathcal{S}_\mathcal{G}$. □

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,*

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

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

Proof. Let $\xi = (s_0, a_0, s_1, a_1, \dots)$ be any trajectory that has non-zero probabilty in the trajectory distribution $\widetilde{\rho}^{\pi}$ of π on $\widetilde{\mathcal{M}}$. Let $\mathcal{I}$ and $\mathcal{I}^*$ be the intervention sets of $\mathcal{G}$ and $\mathcal{G}^*$, respectively. Suppose for some t that $(s_t, a_t) \in \mathcal{I}$. We know for $\tau \geq t+1$ that $s_\tau = s_\dagger$. In addition, by Lemma 6, we have $\overline{A}^*(s_\tau, a_\tau) \leq \overline{A}(s_\tau, a_\tau) \leq 0$ for any $\tau \in [0, t-1]$, so $(s_\tau, a_\tau) \notin \mathcal{I}^*$. Therefore, the sub-trajectory (s_τ, a_τ) with $\tau \in \{0, 1, \dots, t\}$ also has a non-zero probability in $\widetilde{\mathcal{M}}^*$. By this argument, every sub-trajectory in $\mathcal{S} \times \mathcal{A}$ with non-zero probability in $\widetilde{\mathcal{M}}$ also has non-zero probability in $\widetilde{\mathcal{M}}^*$. The final thesis follows from defining the state-action distributions through averaging the trajectory distributions. □

A.4. Proof for Absorbing MDP in Section 3.3.2

We derive some properties of the Bellman operator of the absorbing MDP.

Lemma 7. *For a policy π, let $(\mathcal{B}^\pi Q)(s,a) \coloneqq r(s,a) + \gamma \mathbb{E}_{s'|s,a}[Q(s', \pi)]$ denote the Bellman operator of π in $\mathcal{M}$; similarly define $\widetilde{\mathcal{B}}^\pi$ for $\widetilde{\mathcal{M}}$. Let $Q : \widetilde{\mathcal{S}} \times \mathcal{A} \to \mathbb{R}$ be some function satisfying $Q(s_\dagger, a) = 0$ for all $a \in \mathcal{A}$.*

1. *The Bellman operator in $\widetilde{\mathcal{M}}$ can be written as*

$$(\widetilde{\mathcal{B}}^\pi Q)(s,a) = \begin{cases} (\mathcal{B}^\pi Q)(s,a) \cdot \mathbb{1}\{(s,a) \notin \mathcal{I}\} + \widetilde{R} \cdot \mathbb{1}\{(s,a) \in \mathcal{I}\}, & (s,a) \in \mathcal{S} \times \mathcal{A} \\ 0, & s = s_\dagger. \end{cases} \tag{12}$$

2. *The following holds when the temporal-difference operator $\widetilde{\mathcal{D}}^\pi$ for $\widetilde{\mathcal{M}}$ is applied to the policy's state-action value function Q^π for $\mathcal{M}$:*

$$\left(\widetilde{R} - \frac{1}{1-\gamma}\right) \mathbb{1}\{(s,a) \in \mathcal{I}\} \leq (\widetilde{\mathcal{D}}^\pi Q^\pi)(s,a) \leq \widetilde{R}\, \mathbb{1}\{(s,a) \in \mathcal{I}\} \quad \textit{for all } (s,a) \in \mathcal{S} \times \mathcal{A} \tag{13}$$

$$(\widetilde{\mathcal{D}}^\pi Q^\pi)(s_\dagger, a) = 0, \tag{14}$$

where the definition of Q^π is extended to $s_\dagger$ as $Q^\pi(s_\dagger, a) = 0$.

Proof. For brevity, let $\Omega(s,a) = \mathbb{1}\{(s,a) \in \mathcal{I}\}$.

1. Since $Q(s_\dagger, \pi) = 0$, the following holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned} (\widetilde{\mathcal{B}}^\pi Q)(s,a) &= \widetilde{r}(s,a) + \gamma \mathbb{E}_{s' \sim \widetilde{P}|s,a}[Q(s', \pi)] \\ &= (1 - \Omega(s,a)) \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P|s,a}[Q(s', \pi)]\right) + \Omega(s,a) \cdot \widetilde{R} \\ &= (1 - \Omega(s,a)) \cdot (\mathcal{B}^\pi Q)(s,a) + \Omega(s,a) \cdot \widetilde{R} \end{aligned}$$

   and

$$(\widetilde{\mathcal{B}}^\pi Q)(s_\dagger, a) = 0 + \gamma Q(s_\dagger, \pi) = 0.$$

2. For (13), using the fact that $(\mathcal{B}^\pi Q)^\pi = Q^\pi$, the following applies on $\mathcal{S} \times \mathcal{A}$:

$$(\widetilde{\mathcal{D}}^\pi Q^\pi)(s,a) = (\widetilde{\mathcal{B}}^\pi Q^\pi)(s,a) - Q^\pi(s,a) = \Omega(s,a) \cdot \left(\widetilde{R} - Q^\pi(s,a)\right).$$

   Since $0 \leq Q^\pi(s,a) \leq \frac{1}{1-\gamma}$, we have

$$\left(\widetilde{R} - \frac{1}{1-\gamma}\right) \Omega(s,a) \leq (\widetilde{\mathcal{D}}^\pi Q^\pi)(s,a) \leq \widetilde{R}\, \Omega(s,a).$$

   For the absorbing state in (14), by the extended definition and the equality $(\widetilde{\mathcal{B}}^\pi Q)(s_\dagger, a) = 0$, we have

$$(\widetilde{\mathcal{D}}^\pi Q^\pi)(s_\dagger, a) = (\widetilde{\mathcal{B}}^\pi Q^\pi)(s_\dagger, a) - Q^\pi(s_\dagger, a) = 0.$$

□

Lemma 8. *For any policy π, $P_{\mathcal{G}}(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi}}[\mathbb{1}\{(s,a)\in\mathcal{I}\}]$.*

Proof. Notice that for any h,

$$\mathrm{Prob}(\xi^h \cap \mathcal{I} \neq \varnothing \mid \pi, \mathcal{M}) = \mathrm{Prob}(\xi^h \cap \mathcal{I} \neq \varnothing \mid \pi, \widetilde{\mathcal{M}}) = \mathbb{E}_{\widetilde{\rho}^{\pi}}\left[\sum_{t=0}^{h-1} \mathbb{1}\{(s_t,a_t)\in\mathcal{I}\}\right].$$

By Lemma 2,

$$\begin{aligned}
\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi}}[\mathbb{1}\{(s,a)\in\mathcal{I}\}] &= \mathbb{E}_{\widetilde{\rho}^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{1}\{(s_t,a_t)\in\mathcal{I}\}\right] \\
&= (1-\gamma)\sum_{h=0}^{\infty}\gamma^h\mathbb{E}_{\widetilde{\rho}^{\pi}}\left[\sum_{t=0}^{h-1}\mathbb{1}\{(s_t,a_t)\in\mathcal{I}\}\right] \\
&= (1-\gamma)\sum_{h=0}^{\infty}\gamma^h\mathrm{Prob}(\xi^h\cap\mathcal{I}\neq\varnothing\mid\pi,\mathcal{M}) \\
&= P_{\mathcal{G}}(\pi).
\end{aligned}$$

□

Using the above results, we can bound the difference between the values of the original and the absorbing MDPs.

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

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

Proof of Lemma 1. First, extend the definition of Q^{π} to $s_{\dagger}$ as $Q^{\pi}(s_{\dagger},a)=0$ for any $a\in\mathcal{A}$. By Corollary 1, we have

$$\widetilde{V}^{\pi}(d_0) - V^{\pi}(d_0) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi}}[(\widetilde{\mathcal{D}}^{\pi}Q^{\pi})(s,a)]$$

By Lemma 7, we can derive

$$\left(\widetilde{R} - \frac{1}{1-\gamma}\right)\frac{\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi}}[\mathbb{1}\{(s,a)\in\mathcal{I}\}]}{1-\gamma} \le \widetilde{V}^{\pi}(d_0) - V^{\pi}(d_0) \le \widetilde{R}\,\frac{\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi}}[\mathbb{1}\{(s,a)\in\mathcal{I}\}]}{1-\gamma}.$$

Finally, substituting the equality from Lemma 8 and negating the inequality concludes the proof. □

Next we derive some lemmas, which will be later to used to show that when the intervention set is partial, the unconstrained reduction is effective.

Lemma 9. *Let $\mathcal{I}\subset\mathcal{S}_{\mathrm{safe}}\times\mathcal{A}$ be partial, and let $\mathcal{F}=(\mathcal{S}_{\mathrm{safe}}\times\mathcal{A})\setminus\mathcal{I}$ be the state-action pairs that are not intervened. For an arbitrary policy π, define*

$$\pi_f(a|s) \coloneqq \pi(a|s)\mathbb{1}\{(s,a)\in\mathcal{F}\} + f(s,a), \tag{15}$$

where $f(s,a)$ is some arbitrary non-negative function which is zero on $\mathcal{I}$ and that ensures $\sum_{a\in\mathcal{A}}\pi_f(a|s)=1$ for all $s\in\mathcal{S}$. Define

$$\widetilde{J}^{\pi}_{+} \coloneqq \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi}}[r(s,a)\cdot\mathbb{1}\{(s,a)\in\mathcal{F}\}]$$
$$\widetilde{J}^{\pi}_{-} \coloneqq \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi}}[\widetilde{R}\cdot\mathbb{1}\{(s,a)\in\mathcal{I}\}]$$

as the expected returns in $\mathcal{F}$ and $\mathcal{I}$, respectively.

The following are true:

1. $\widetilde{V}^{\pi}(d_0) = \widetilde{J}^{\pi}_{+} + \widetilde{J}^{\pi}_{-}$.
2. $\widetilde{d}^{\pi_f}(s,a) \geq \widetilde{d}^{\pi}(s,a)$ *for all* $(s,a) \in \mathcal{F}$.
3. $\widetilde{J}^{\pi_f}_{+} \geq \widetilde{J}^{\pi}_{+}$.
4. $\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi_f}}[\mathbb{1}\{(s,a) \in \mathcal{I}\}] = 0$, *implying* $\widetilde{J}^{\pi_f}_{-} = 0$.
5. $\widetilde{V}^{\pi_f}(d_0) \geq \widetilde{V}^{\pi}(d_0)$ *whenever* $\widetilde{R} \leq 0$. *Furthermore, if* $\widetilde{R} < 0$ *and* $\pi(a|s) > 0$ *for some* $(s,a) \in \mathcal{I}$, *then* $\widetilde{V}^{\pi_f}(d_0) > \widetilde{V}^{\pi}(d_0)$.

Proof. 1. This follows from the definition of $\widetilde{r}$ in (7).

2. Recall $\widetilde{d}^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \widetilde{d}^{\pi}_t(s,a)$. To show the desired result, we show by induction that $\widetilde{d}^{\pi_f}_t(s,a) \geq \widetilde{d}^{\pi}_t(s,a)$ for all $(s,a) \in \mathcal{F}$ and $t \geq 0$. For $t = 0$, by construction of π_f, we have $\pi_f(a|s) \geq \pi(a|s)$ for all $(s,a) \in \mathcal{F}$ and therefore $\widetilde{d}^{\pi_f}_0(s,a) \geq \widetilde{d}^{\pi}_0(s,a)$ for all $(s,a) \in \mathcal{F}$.

   Now suppose that for some $t \geq 0$ the inequality $\widetilde{d}^{\pi_f}_t(s,a) \geq \widetilde{d}^{\pi}_t(s,a)$ holds for all $(s,a) \in \mathcal{F}$. Then, for some $(s,a) \in \mathcal{F}$, we can derive

$$\begin{aligned}
\widetilde{d}^{\pi_f}_{t+1}(s,a) &= \pi_f(a|s) \sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}} \widetilde{P}(s|s_t,a_t)\widetilde{d}^{\pi_f}_t(s_t,a_t) \\
&= \pi_f(a|s) \sum_{(s_t,a_t)\in\mathcal{F}} P(s|s_t,a_t)\widetilde{d}^{\pi_f}_t(s_t,a_t) \\
&\geq \pi(a|s) \sum_{(s_t,a_t)\in\mathcal{F}} P(s|s_t,a_t)\widetilde{d}^{\pi}_t(s_t,a_t) \\
&= \widetilde{d}^{\pi}_{t+1}(s,a),
\end{aligned}$$

   where we use the inductive hypothesis in the inequality. Thus, we have $\widetilde{d}^{\pi_f}(s,a) \geq \widetilde{d}^{\pi}(s,a)$ by summing over each time step.

3. By statement 2, definition of $\widetilde{J}^{\pi}_{+}$, and non-negativity of the reward r, it follows that $\widetilde{J}^{\pi_f}_{+} \geq \widetilde{J}^{\pi}_{+}$.

4. This statement from the construction of π_f and induction. First, we have $\widetilde{d}^{\pi_f}_0(s,a) = 0$ for all $(s,a) \in \mathcal{I}$. Now suppose for some $t \geq 0$ that $\widetilde{d}^{\pi_f}_t(s,a) = 0$ for all $(s,a) \in \mathcal{I}$. We can see that $\widetilde{d}^{\pi_f}_{t+1}(s,a) = 0$ for all $(s,a) \in \mathcal{I}$ since π_f never chooses actions such that $(s,a) \in \mathcal{I}$.

   Therefore, $\widetilde{d}^{\pi_f}(s,a) = 0$ for all $(s,a) \in \mathcal{I}$. By definition of $\widetilde{J}^{\pi}_{-}$, this allows us to conclude that $\widetilde{J}^{\pi}_{-} = 0$.

5. Using statements 3 and 4, we conclude that

$$\widetilde{V}^{\pi_f}(d_0) = \widetilde{J}^{\pi_f}_{+} + \widetilde{J}^{\pi_f}_{-} \geq \widetilde{J}^{\pi}_{+} + \widetilde{J}^{\pi}_{-} = \widetilde{V}^{\pi}(d_0).$$

   The special case follows from observing that $\widetilde{J}^{\pi}_{-} < 0$ whenever $\pi(a|s) > 0$ for some $(s,a) \in \mathcal{I}$.

□

Lemma 10. *Let* $\widetilde{R}$ *be non-positive and* $\widetilde{V}^*$ *denote the optimal value function for* $\widetilde{\mathcal{M}}$.

1. *For any policy* π,
$$\widetilde{V}^*(d_0) \geq \widetilde{J}^{\pi}_{+}.$$
2. *There is an optimal policy* $\widetilde{\pi}^*$ *of* $\widetilde{\mathcal{M}}$ *satisfying*
$$\mathbb{E}_{(s,a)\sim\widetilde{d}^{\widetilde{\pi}^*}}[\mathbb{1}\{(s,a) \in \mathcal{I}\}] = 0. \tag{16}$$

3. *If $\widetilde{R}$ is negative, every optimal policy of $\widetilde{\mathcal{M}}$ satisfies* (16).

Proof of Lemma 10. 1. Let the policy π be arbitrary, and define π_f using (15). The following then holds by Lemma 9:

$$\widetilde{V}^*(d_0) \geq \widetilde{V}^{\pi_f}(d_0) = \widetilde{J}_+^{\pi_f} \geq \widetilde{J}_+^{\pi}.$$

2. Suppose that π is an optimal policy of $\widetilde{\mathcal{M}}$, and define π_f using (15). Because $\widetilde{R}$ is non-positive, we know by Lemma 9 and optimality of π that $\widetilde{V}^{\pi_f}(d_0) = \widetilde{V}^{\pi}(d_0)$. Therefore, we can define an optimal policy as $\widetilde{\pi}^* = \pi_f$ and conclude by Lemma 9 that $\mathbb{E}_{(s,a)\sim\widetilde{d}^{\widetilde{\pi}^*}}[\mathbb{1}\{(s,a) \in \mathcal{I}\}] = 0$.

3. Suppose for the sake of contradiction there is an optimal policy $\widetilde{\pi}^*$ of $\widetilde{\mathcal{M}}$ such that (16) does *not* hold (i.e., it may take some $(s,a) \in \mathcal{I}$). By Lemma 9, we can construct some policy π_f such that $\widetilde{V}^{\pi_f}(d_0) > \widetilde{V}^{\widetilde{\pi}^*}(d_0)$. This contradicts $\widetilde{\pi}^*$ being optimal, so every optimal policy of $\widetilde{\mathcal{M}}$ must satisfy (16).

□

These results show that if the intervention set is partial and the penalty of being intervened is strict, then the optimal policy of the absorbing MDP would not be intervened.

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 $P_{\mathcal{G}}(\widetilde{\pi}^*) = 0$.*

Proof of Proposition 6. This directly follows from Lemmas 8 and 10. □

Below we derive some lemmas to show a near optimal policy of the absorbing MDP is safe. (We already proved above that the optimal policy of the absorbing MDP is safe).

Lemma 11. *Let $\mathcal{I} \subset \mathcal{S} \times \mathcal{A}$ be partial (Definition 2). Given some policy π, let π' be the corresponding shielded policy defined in* (4)*. Then, the following holds for any $h \geq 0$ in $\mathcal{M}$:*

$$\text{Prob}(s_\triangleright \in \xi^h \mid \pi, \mathcal{M}) \leq \text{Prob}(s_\triangleright \in \xi^h \mid \pi', \mathcal{M}) + \text{Prob}(\xi^h \cap \mathcal{I} \neq \varnothing \mid \pi, \mathcal{M}), \tag{17}$$

where $\xi^h = (s_0, a_0, \ldots, s_{h-1}, a_{h-1})$ is an h-step trajectory segment.

Proof. First, we notice that $\pi'(a|s) \geq \pi(a|s)$ when $(s,a) \notin \mathcal{I}$, because $\pi'(a|s) = \pi(a|s) + w(s)\mu(a|s) \geq \pi(a|s)$.

We bound the probability of π violating a constraint in $\mathcal{M}$ by introducing whether π visits the intervention set:

$$\begin{aligned}\text{Prob}(s_\triangleright \in \xi^h \mid \pi, \mathcal{M}) &= \text{Prob}(s_\triangleright \in \xi^h,\ \xi^h \cap \mathcal{I} = \varnothing \mid \pi, \mathcal{M}) + \text{Prob}(s_\triangleright \in \xi^h,\ \xi^h \cap \mathcal{I} \neq \varnothing \mid \pi, \mathcal{M}) \\ &\leq \text{Prob}(s_\triangleright \in \xi^h,\ \xi^h \cap \mathcal{I} = \varnothing \mid \pi, \mathcal{M}) + \text{Prob}(\xi^h \cap \mathcal{I} \neq \varnothing \mid \pi, \mathcal{M}).\end{aligned}$$

We now bound the first term. Let ξ^h satisfy the event "$s_\triangleright \in \xi^h,\ \xi^h \cap \mathcal{I} = \varnothing$", and let T be the time index such that $s_T = s_\triangleright$ in ξ^h. Then, the probability of this trajectory under π and $\mathcal{M}$ is

$$d_0(s_0)\pi(a_0|s_0)P(s_1|s_0,a_0)\cdots\pi(a_{T-1}|s_{T-1})P(s_T|s_{T-1},a_{T-1}).$$

Since each (s_t, a_t) is not in $\mathcal{I}$, we have $\pi(a_t|s_t) \leq \pi'(a_t|s_t)$ for each (s_t, a_t) in ξ^h. Thus, the probability of this trajectory under π and $\mathcal{M}$ is upper bounded by its probability under π' and $\mathcal{M}$. Summing over each trajectory ξ^h satisfying the event then yields:

$$\text{Prob}(s_\triangleright \in \xi^h,\ \xi^h \cap \mathcal{I} = \varnothing \mid \pi, \mathcal{M}) \leq \text{Prob}(s_\triangleright \in \xi^h,\ \xi^h \cap \mathcal{I} = \varnothing \mid \pi', \mathcal{M}).$$

We now complete the original bound:

$$\begin{aligned}\text{Prob}(s_\triangleright \in \xi^h \mid \pi, \mathcal{M}) &\leq \text{Prob}(s_\triangleright \in \xi^h,\ \xi^h \cap \mathcal{I} = \varnothing \mid \pi, \mathcal{M}) + \text{Prob}(\xi^h \cap \mathcal{I} \neq \varnothing \mid \pi, \mathcal{M}) \\ &\leq \text{Prob}(s_\triangleright \in \xi^h,\ \xi^h \cap \mathcal{I} = \varnothing \mid \pi', \mathcal{M}) + \text{Prob}(\xi^h \cap \mathcal{I} \neq \varnothing \mid \pi, \mathcal{M}) \\ &\leq \text{Prob}(s_\triangleright \in \xi^h \mid \pi', \mathcal{M}) + \text{Prob}(\xi^h \cap \mathcal{I} \neq \varnothing \mid \pi, \mathcal{M}).\end{aligned}$$

□

Lemma 12. *For any policy π and $\mathcal{I} \subset \mathcal{S} \times \mathcal{A}$ that is partial, let π' be the corresponding shielded policy. Then, the following safety bound holds:*

$$\overline{V}^{\pi}(d_0) \leq \overline{V}^{\pi'}(d_0) + \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi}}[\mathbb{1}\{(s,a) \in \mathcal{I}\}].$$

Proof of Lemma 12. Using (17) from Lemma 11 and the fact that the probabilities can be expressed as expected sums of indicators:

$$\begin{aligned}
\text{Prob}(s_{\triangleright} \in \xi^h \mid \pi, \mathcal{M}) &= \mathbb{E}_{\rho^{\pi}}\left[\sum_{t=0}^{h-1} \mathbb{1}\{s_t = s_{\triangleright}\}\right] \\
\text{Prob}(s_{\triangleright} \in \xi^h \mid \pi', \mathcal{M}) &= \mathbb{E}_{\rho^{\pi'}}\left[\sum_{t=0}^{h-1} \mathbb{1}\{s_t = s_{\triangleright}\}\right] \\
\text{Prob}(\xi^h \cap \mathcal{I} \neq \varnothing \mid \pi, \mathcal{M}) &= \mathbb{E}_{\widetilde{\rho}^{\pi}}\left[\sum_{t=0}^{h-1} \mathbb{1}\{(s_t, a_t) \in \mathcal{I}\}\right]
\end{aligned}$$

Then, applying Lemma 2 results in the desired inequality. □

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 $\mathcal{M}$:*

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

Proof of Proposition 7. The performance bound follows from Lemma 1.

$$\begin{aligned}
V^{\pi^*}(d_0) - V^{\pi}(d_0) &= V^{\pi^*}(d_0) - \widetilde{V}^{\pi^*}(d_0) + \widetilde{V}^{\pi^*}(d_0) - \widetilde{V}^{\pi}(d_0) + \widetilde{V}^{\pi}(d_0) - V^{\pi}(d_0) \\
&\leq \left(|\widetilde{R}| + \frac{1}{1-\gamma}\right) P_{\mathcal{G}}(\pi^*) + \widetilde{V}^{\pi^*}(d_0) - \widetilde{V}^{\pi}(d_0) - |\widetilde{R}|\, P_{\mathcal{G}}(\pi) \\
&\leq \left(|\widetilde{R}| + \frac{1}{1-\gamma}\right) P_{\mathcal{G}}(\pi^*) + \widetilde{V}^{*}(d_0) - \widetilde{V}^{\pi}(d_0) \\
&\leq \left(|\widetilde{R}| + \frac{1}{1-\gamma}\right) P_{\mathcal{G}}(\pi^*) + \varepsilon.
\end{aligned}$$

For the safety bound, we start with Lemma 12:

$$\overline{V}^{\pi}(d_0) \leq \overline{V}^{\pi'}(d_0) + \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\widetilde{d}^{\pi}}[\mathbb{1}\{(s,a) \in \mathcal{I}\}]$$

We provide an upper bound on the second term on the right hand side above. Using the definition of $\widetilde{J}^{\pi}_{-}$ in Lemma 9, we derive that

$$\begin{aligned}
\frac{\mathbb{E}_{\widetilde{d}^{\pi}}[\mathbb{1}\{(s,a) \in \mathcal{I}\}]}{1-\gamma} &= -\frac{\widetilde{J}^{\pi}_{-}}{|\widetilde{R}|} \\
&= \frac{1}{|\widetilde{R}|}\left(-\widetilde{V}^{\pi}(d_0) + \widetilde{V}^{*}(d_0) + \widetilde{J}^{\pi}_{+} - \widetilde{V}^{*}(d_0)\right) \\
&\leq \frac{1}{|\widetilde{R}|}\left(\widetilde{V}^{*}(d_0) - \widetilde{V}^{\pi}(d_0)\right) = \frac{\varepsilon}{|\widetilde{R}|},
\end{aligned}$$

where the inequality is due to Lemma 10.

Combine everything altogether:

$$\begin{aligned}\overline{V}^{\pi}(d_0) &\leq \overline{V}^{\pi'}(d_0) + \frac{\mathbb{E}_{\widetilde{d}^{\pi}}[\mathbb{1}\{(s,a) \in \mathcal{I}\}]}{1-\gamma} \\ &= \overline{V}^{\pi'}(d_0) + \frac{\varepsilon}{|\widetilde{R}|}.\end{aligned}$$

□

We now prove the main result of the paper.

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 $\mathcal{M}$:*

$$\begin{aligned}V^*(d_0) - V^{\hat{\pi}}(d_0) &\leq \frac{2}{1-\gamma} P_{\mathcal{G}}(\pi^*) + \varepsilon \\ \overline{V}^{\hat{\pi}}(d_0) &\leq \overline{Q}(d_0, \mu) + \frac{\min\{\sigma + \eta, 2\gamma\}}{1-\gamma} + \varepsilon.\end{aligned}$$

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 $\mathcal{M}$.

Proof. This is a direct result of Proposition 7.

The performance suboptimality results from:

$$\begin{aligned}V^{\pi^*}(d_0) - V^{\hat{\pi}}(d_0) &\leq \left(|\widetilde{R}| + \frac{1}{1-\gamma}\right) P_{\mathcal{G}}(\pi^*) + \varepsilon \\ &\leq \left(1 + \frac{1}{1-\gamma}\right) P_{\mathcal{G}}(\pi^*) + \varepsilon \\ &= \frac{2-\gamma}{1-\gamma} P_{\mathcal{G}}(\pi^*) + \varepsilon \\ &\leq \frac{2}{1-\gamma} P_{\mathcal{G}}(\pi^*) + \varepsilon.\end{aligned}$$

For the safety bound,

$$\begin{aligned}\overline{V}^{\hat{\pi}}(d_0) &\leq \overline{V}^{\mathcal{G}(\hat{\pi})}(d_0) + \varepsilon \\ &\leq \overline{Q}(d_0, \mu) + \frac{\min\{\sigma + \eta, 2\gamma\}}{1-\gamma} + \varepsilon,\end{aligned}$$

where the second inequality follows from Theorem 2 and ε-suboptimality of $\hat{\pi}$ in $\widetilde{\mathcal{M}}$. □

B. Additional Discussion of SAILR

B.1. Necessity of the partial property

We highlight that the subset $\mathcal{I}$ being *partial* (Definition 1) is crucial for the unconstrained MDP reduction behind SAILR. If we were to construct an absorbing MDP $\widetilde{\mathcal{M}}'$ described in Section 3.2 using an arbitrary non-partial subset $\mathcal{I}' \subseteq \mathcal{S} \times \mathcal{A}$, then the optimal policy of $\widetilde{\mathcal{M}}'$ can still enter $\mathcal{I}'$ when $\widetilde{R} > -\infty$, because the optimal policy of $\widetilde{\mathcal{M}}'$ can use earlier rewards to make up for the penalty incurred in $\mathcal{I}'$.

To see this, consider the toy MDP $\mathcal{M}$ shown in Fig. 4. Since there is no alternative action available at state 2, the intervention illustrated in $\widetilde{\mathcal{M}}'$ is *not* partial. Suppose $\widetilde{R} > -1/\gamma$. Then, in $\widetilde{\mathcal{M}}'$, a policy choosing to transition from 1 to 3 has a value of 0, and a policy choosing to transition from 1 to 2 has a value of $1 + \gamma\widetilde{R} > 0$. Therefore, the optimal policy will transition

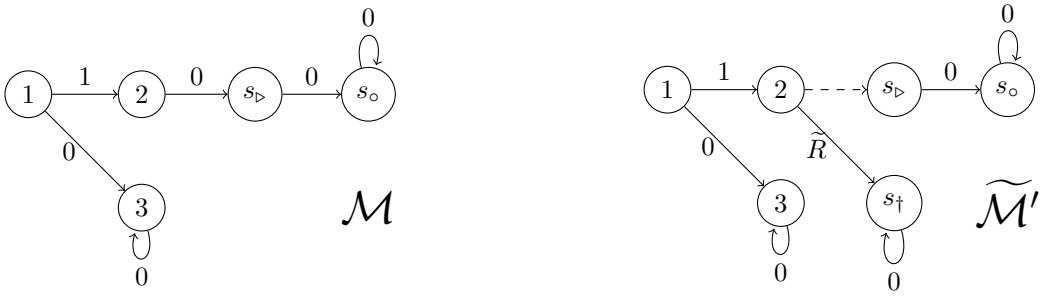


Figure 4. A simple example illustrating a non-partial intervention. Edge weights correspond to rewards. If $\widetilde{R} > -1/\gamma$, the optimal policy in $\widetilde{\mathcal{M}}'$ will always go into the intervention set.

from 1 to 2 and go into the non-partial intervention set $\mathcal{I}'$. Once applied to the original MDP $\mathcal{M}$, this policy will always go into the unsafe set.

One might think generally it is possible to set $\widetilde{R}$ to be negative enough to ensure the optimal policy will never go into the intervention set, which is indeed true for the counterexample above. But we remark that we need to set $\widetilde{R}$ to be arbitrarily large (in the negative direction) for general problems, which can cause high variance issues in return or gradient estimation (Shalev-Shwartz et al., 2016). Because of the discount factor $\gamma < 1$, the negative reward steming from the absorbing state will be at most $\gamma^T \widetilde{R}$, where T is the time step that the system enters $\mathcal{I}'$. For a fixed and finite $\widetilde{R}$, we can then extend the above MDP construction to let the agent go through a long enough chain after transitioning from 1 to 2 so that the resultant value satisfies $1 + \gamma^T \widetilde{R} > 0$. Like the example above, this path would be the only path with positive reward, despite intersecting the intervention set. Therefore, the optimal policy of $\widetilde{\mathcal{M}}'$ will enter $\mathcal{I}'$.

B.2. Bias of SAILR

In Theorem 1, we give a performance guarantee of SAILR

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

It shows that SAILR has a bias $P_{\mathcal{G}}(\pi^*) \in [0, 1]$, which is the probability that the optimal policy π^* would be intervened by the advantage-based intervention rule. Here we discuss special cases where this bias vanishes.

The first special case is when the original problem is unconstrained (i.e., (2) has a trivial constraint with $\delta = 1$). In this case, we can set the threshold $\eta \geq \gamma$ in SAILR to turn off the intervention, and SAILR returns the optimal policy of the MDP $\mathcal{M}$ when the base RL algorithm can find one.

Another case is when π^* is a perfect safe policy, i.e., $\overline{V}^{\pi^*}(d_0) = 0$ and we run SAILR with the intervention rule $\mathcal{G}^* = (\overline{Q}^*, \overline{\pi}^*, 0)$ (Proposition 4). Similar to the proof of Lemma 6, one can show that running π^* would not trigger the intervention rule $\mathcal{G}^*$ and therefore the bias $P_{\mathcal{G}^*}(\pi^*)$ is zero.

However, we note that generally the bias $P_{\mathcal{G}}(\pi^*)$ can be non-zero.

C. Experimental Details

C.1. Point Robot

This environment (Fig. 5) is a simplification of the point environment from (Achiam et al., 2017). The state is $s = (x, y, \dot{x}, \dot{y})$, where (x, y) is the x-y position and $(\dot{x}, \dot{y})$ is the corresponding velocity. The action $a = (a_x, a_y)$ is the force applied to the robot (each component has maximum magnitude $a_{\max}$). The agent has some mass m and can achieve maximum speed $v_{\max}$. The dynamics update (with time increment Δt) is:

$$(x_{t+1}, y_{t+1}) = (x_t, y_t) + (\dot{x}_t, \dot{y}_t)\Delta t + \frac{1}{2m} a_t \Delta t^2$$
$$(\dot{x}_{t+1}, \dot{y}_{t+1}) = \texttt{clip-norm}\left(((\dot{x}_t, \dot{y}_t) + \frac{1}{m} a_t \Delta t,\ v_{\max} \right),$$

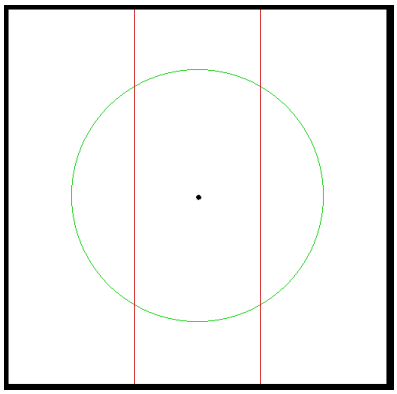


Figure 5. The point environment. The black dot corresponds to the agent, the green circle to the desired path, and the red lines to the constraints on the horizontal position. The vertical constraints are outside of the visualized environment.

where $\texttt{clip-norm}(u, c)$ scales u so that its norm matches c if $\|u\| > c$. The reward corresponds to following a circular path of radius R^* at a high speed and the safe set to staying within desired positional bounds $x_{\max}$ and $y_{\max}$:

$$r(s,a) = \frac{(\dot{x},\dot{y}) \cdot (-y, x)}{1 + |\|(x,y)\| - R^*|}$$
$$\mathcal{S}_{\text{safe}} = \{s \in \mathcal{S} : |x| \leq x_{\max} \text{ and } |y| \leq y_{\max}\}$$

For our experiments, we set these parameters to $m = 1$, $v_{\max} = 2$, $a_{\max} = 1$, $\Delta t = 0.1$, $R^* = 5$, $x_{\max} = 2.5$, and $y_{\max} = 15$.

For this environment, we also consider a *shaped* cost function $\hat{c}(s, a)$ which is a function of the *distance* of the state s to the boundary of the unsafe set, denoted by $\text{dist}(s, \mathcal{S}_{\text{unsafe}})$. Here, $\mathcal{S}_{\text{unsafe}}$ denotes the 2D unsafe region in this environment (i.e., those outside the vertical lines in Fig. 5). Note that in the theoretical analysis $\mathcal{S}_{\text{unsafe}}$ is abstracted into $\{s_\triangleright, s_\dagger\}$.

For the point environment, the distance function is $\text{dist}(s, S_{\text{unsafe}}) = \max\{0, \min\{x_{\max} - x, x_{\max} + x, y_{\max} - y, y_{\max} + y\}\}$. For some constant $\alpha \geq 0$, the cost function is defined as a hinge function of the distance:

$$\hat{c}(s,a) = \begin{cases} \mathbb{1}\{\text{dist}(s, \mathcal{S}_{\text{unsafe}}) = 0\}, & \alpha = 0 \\ \max\left\{0, 1 - \frac{1}{\alpha}\text{dist}(s, \mathcal{S}_{\text{unsafe}})\right\}, & \text{otherwise.} \end{cases} \tag{18}$$

We note that $\hat{c}$ is an upper bound for c if $\alpha > 0$ and $\hat{c} = c$ if $\alpha = 0$. We shape the cost here to make it continuous, so that the effects of approximation bias is smaller than that resulting from a discontinuous cost (i.e., the original indicator function).

Intervention Rule: The backup policy μ applies a decelerating force (with component-wise magnitude up to $a_{\max}$) until the agent has zero velocity. Our experiments consider the following approaches to construct $\overline{Q}$:

- **Neural network approximation**: We construct a dataset of points mapping states and actions to state-action values $\overline{Q}^\mu$ by picking some state and action in the MDP, executing the action from that state, and then continuing the rollout with the backup policy μ to find the empirical state-action value with respect to the shaped cost function $\hat{c}$. Our dataset consists of 10^7 points resulting from a uniform discretization of the state-action space. We apply a similar method to form a dataset for the state values $\overline{V}^\mu$.

  We then train four networks (two to independently approximate $\overline{Q}^\mu$, and two for $\overline{V}^\mu$), where each network has three hidden layers each with 256 neurons and a ReLU activation. The predicted advantage is $\overline{A}(s,a) = \max\{\overline{Q}_1(s,a), \overline{Q}_2(s,a)\} - \min\{\overline{V}_1(s), \overline{V}_2(s)\}$, where we apply the pessimistic approach from (Thananjeyan et al., 2020) to prevent overestimation bias.

- **Model-based evaluation**: Here, we have access to a model of the robot where all parameters match the real environment except possibly the mass $\hat{m}$. We refer to the modeled transition dynamics as $\hat{\mathcal{P}}$ and the resulting trajectory distribution under μ as $\hat{\rho}^\mu$. The function $\overline{Q}$ is then set to be the model-based estimate of $\overline{Q}^\mu$ using the shaped cost function $\hat{c}$ and

dynamics $\hat{\mathcal{P}}$:

$$\overline{Q}(s,a) = \mathbb{E}_{\hat{\rho}^{\mu}|s_0=s,a_0=a}\left[\sum_{t=0}^{\infty}\gamma^t \hat{c}(s_t,a_t)\right].$$

For our experiments, the modeled mass $\hat{m}$ is either 1 (unbiased case) or 0.5 (biased case).

For our experiments, we set the advantage threshold $\eta = 0.08$ when using the neural network approximator and $\eta = 0$ when using the model-based rollouts.

Hyperparameters: All point experiments were run on a 32-core Threadripper machine. The given hyperparameters were found by hand-tuning until good performance was found on all algorithms.

Hyperparameter	**Value**
Epochs	500
Neural Network Architecture	2 hidden layers, 64 neurons per hidden layer, $\tanh$ act.
Batch size	4000
Discount γ	0.99
Entropy bonus	0.001
CMDP threshold δ	0.01
Penalty value $\tilde{R}$	-2
Lagrange multiplier step size (for constrained approaches)	0.05
Cost shaping constant α	0.5
Number of seeds	10

C.2. Half-Cheetah

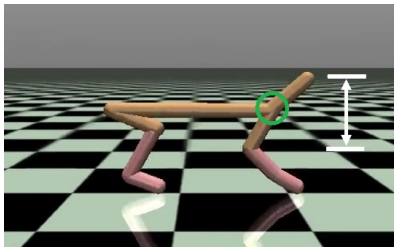


Figure 6. The half-cheetah environment. The green circle is centered on the link of interest, and the white double-headed arrow denotes the allowed height range of the link.

This environment (Fig. 6) comes from OpenAI Gym and has reward equal to the agent's forward velocity. One of the agent's links (denoted by the green circle in Fig. 6) is constrained to lie in a given height range, outside of which the robot is deemed to have fallen over. In other words, if h is the height of the link of interest, $h_{\min}$ is the minimum height, and $h_{\max}$ is the maximum height, the safe set is defined as $\mathcal{S}_{\text{safe}} = \{s \in \mathcal{S} : h_{\min} \leq h \leq h_{\max}\}$. For our experiments, we set $h_{\min} = 0.4$ and $h_{\max} = 1$.

Heuristic Intervention Rule: This intervention rule $\mathcal{G} = (\overline{Q}, \mu, \eta)$ relies on a dynamics model (here, unbiased) to greedily predict whether the safety constraint would be violated at the next time step. In particular, if s is the current state and $\hat{a} \sim \pi(\cdot|s)$ is the proposed action, the agent will be intervened if the height $\hat{h}'$ in the next state $\hat{s}' \sim \mathcal{P}(\cdot|s,\hat{a})$ lies outside the range $[\hat{h}_{\min}, \hat{h}_{\max}]$, where $\hat{h}_{\min}$ and $\hat{h}_{\max}$ can be set to a smaller range than $[h_{\min}, h_{\max}]$ to induce a more conservative intervention. Once intervened, the episode terminates. The reason for using a smaller range $[\hat{h}_{\min}, \hat{h}_{\max}]$ is an attempt to make the intervention rule possess the partial property (see the discussion in Section 3.1.2). If we were to set the range to be the ordinary range $[h_{\min}, h_{\max}]$ that defines the safe subset, the penalty $\widetilde{R}$ would need to be very negative, which would destabilize learning. Furthermore, there is no guarantee that the intervention set for the original range is partial since there may be no available action to keep the agent from being intervened.

MPC-Based Intervention Rule: Similarly with the model-based intervention rule for the point environment, the MPC intervention rule $\mathcal{G} = (\overline{Q}, \mu, \eta)$ uses a model of the half-cheetah. The backup policy μ is a sampling-based model predictive

control (MPC) algorithm based on (Williams et al., 2017). The MPC algorithm has an optimization horizon of $H = 16$ time steps and minimizes the cost function corresponding to an indicator function of the link height being in the range $[0.45, 0.95]$.[3] The function $\overline{Q}$ is defined as:

$$\overline{Q}(s,a) = \mathbb{E}_{\hat{\rho}}\left[\sum_{t=0}^{H}\gamma^t \hat{c}(\hat{s}_t,\hat{a}_t) \;\middle|\; \hat{s}_0 = s, \hat{a}_0 = a, \hat{a}_{1:H} = \mathrm{MPC}(\hat{s}_1)\right],$$

where $\hat{c}(s,a)$ is the hinge-shaped cost function (in (18)) corresponding to the distance function $\mathrm{dist}(s, S_{\mathrm{unsafe}}) = \max\{0, \min\{h - h_{\min}, h_{\max} - h\}\}$.

For our experiments, we set the advantage threshold $\eta = 0.2$. We also use a modeled mass of 16 (unbiased) and 12 (biased) in our experiments.

Hyperparameters: Except for the MPC-based intervention, all half-cheetah experiments were run on a 32-core Threadripper machine. The MPC-based intervention experiments were run on 64-core Azure servers with each run taking 24 hours. The given hyperparameters were found by hand-tuning until good performance was found on all algorithms.

Hyperparameter	**Value**
Epochs	1250
Neural Network Architecture	2 hidden layers, 64 neurons per hidden layer, $\tanh$ act.
Batch size	4000
Discount γ	0.99
Entropy bonus	0.01
CMDP threshold δ	0.01
Penalty value $\tilde{R}$	-0.1
Lagrange multiplier step size (for constrained approaches)	0.05
Heuristic intervention range $[\hat{h}_{\min}, \hat{h}_{\max}]$	$[0.4, 0.9]$
Cost shaping constant α	0.05
Number of seeds	8

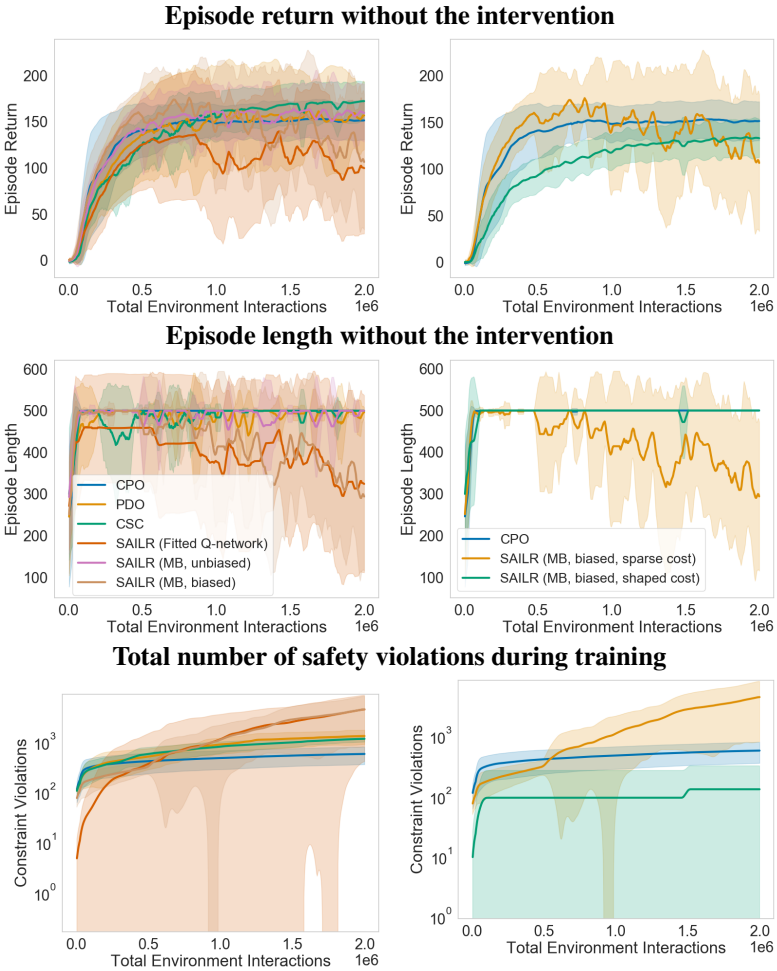
D. Ablations for Point Robot

We run the following two ablations for the point environment, with results shown in Fig. 7:

1. We additionally run all the algorithms with the original sparse cost (Fig. 7a). Here, the baseline algorithms as expected yield high deployment returns while violating many constraints during training. For SAILR, however, only the model-based instance with an unbiased model is able to satisfy the desiderata of high deployment returns while being safe during training. In this case, the sparse cost along with the approximation errors from the other two instances result in the slack σ being large for admissibilty, meaning the safety bounds in Theorems 1 and 2 are loose.

2. We run the model-based instance of SAILR with a biased model and either the sparse cost or the shaped cost (Fig. 7b). Using the sparse cost with the biased model for intervention has deleterious effects in safety and performance. The model mismatch causes a compounding number of safety violations in training (bottom plot) and destabilizes the policy optimization, as observed in the deteriorated returns (top plot) and safety (middle plot), respectively. Shaping the cost function for intervention results in far fewer safety violations and stabilizes the policy optimization.

[3] Observe that this is slightly smaller than the $[0.4, 1]$ height range of the original safety constraint.



(a) Sparse cost

(b) Biased model, either a sparse cost or shaped cost

Figure 7. Ablations for point experiment