Supplementary Material – APPENDIX

A. Proof of Lemma 1

Lemma 1: Given an arbitrary initial policy, iterative application of (9) for policy update will eventually result in a policy π_s^* that satisfies $\mathcal{V}_\infty^{\bar{\gamma}\pi_s^*}(\boldsymbol{s}) \leq \mathcal{V}_\infty^{\bar{\gamma}\pi}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}, \forall \pi, \forall \bar{\gamma} \in [0, 1)$.

Proof: Like in normal RL, the operation in (9) is a contraction mapping. Thus, during policy iteration, the policy will eventually converge to an optimal policy π_s^* whose risk state value is less than or equal to that of any policy. ■

B. Proof of Proposition 1

Proposition 1: For $\pi \in \{\pi \mid \mathcal{V}_\infty^{\bar{\gamma}\pi}(\boldsymbol{s}) < h, \forall \boldsymbol{s} \in \mathcal{S}_s, \forall \bar{\gamma} \in [0, 1)\}$, define $\epsilon = \inf_{\boldsymbol{s}, \bar{\gamma}}\{h - \mathcal{V}_\infty^{\bar{\gamma}\pi}(\boldsymbol{s})\}$ where $\epsilon \in (0, 1]$. When $\bar{\gamma} \geq (1-\epsilon)^{1/(T-1)}$, then we have $\mathcal{V}_\infty^{\pi}(\boldsymbol{s}) \leq h$.

Proof: When $\bar{\gamma} \geq (1-\epsilon)^{1/(T-1)}$, by Assumption 1 and the definition of $\mathcal{V}_N^{\pi}(\boldsymbol{s})$, we have

$$\mathcal{V}_\infty^{\pi}(\boldsymbol{s}) - \mathcal{V}_\infty^{\bar{\gamma}\pi}(\boldsymbol{s}) \overset{(a)}{\leq} 1 - \bar{\gamma}^{T-1} \leq \epsilon \overset{(b)}{\leq} h - \mathcal{V}_\infty^{\bar{\gamma}\pi}(\boldsymbol{s}).$$

For (a), by definition of $\mathcal{V}_N^{\pi}(\boldsymbol{s})$, for a trajectory that does not enter unsafe states, we have $\sum_{k=0}^{\infty} c_{t+k+1} = \sum_{k=0}^{\infty} \bar{\gamma}^k c_{t+k+1} = 0$, so we have $\mathcal{V}_\infty^{\pi}(\boldsymbol{s}) - \mathcal{V}_\infty^{\bar{\gamma}\pi}(\boldsymbol{s}) = 0$. For a trajectory that enters unsafe states, the maximum total cost is 1 (definition of $\mathcal{V}_N^{\pi}(\boldsymbol{s})$), and the minimal discounted total cost is $\bar{\gamma}^{T-1}$ (unsafe state is only in step T). For (b), ϵ is the infimum of $h - \mathcal{V}_\infty^{\bar{\gamma}\pi}(\boldsymbol{s})$. Thus, we have $\mathcal{V}_\infty^{\pi}(\boldsymbol{s}) \leq h$. ■

C. Proof of Lemma 2

Lemma 2: Given an initial policy π_s^0 and $\bar{\gamma} \geq (1-\epsilon)^{1/(T-1)}$, in the obtained policy sequence $\{\pi_s^0, \pi_s^1, ..., \pi_s^*\}$ by applying (9) iteratively, exist a policy $\pi^\dagger$ that satisfies $\mathcal{V}_N^{\pi^\dagger}(\boldsymbol{s}) \leq h, \forall \boldsymbol{s} \in \mathcal{S}_s$.

Proof: By Assumption 2, we have $\exists \tilde{\pi}, \mathcal{V}_\infty^{\bar{\gamma}\tilde{\pi}}(\boldsymbol{s}) \leq \mathcal{V}_\infty^{\tilde{\pi}}(\boldsymbol{s}) < h$. Thus, according to Lemma 1, we can get $\mathcal{V}_\infty^{\bar{\gamma}\pi_s^*}(\boldsymbol{s}) \leq \mathcal{V}_\infty^{\bar{\gamma}\tilde{\pi}}(\boldsymbol{s}) < h, \forall \boldsymbol{s} \in \mathcal{S}_s$. Then, by Proposition 1 we have $\mathcal{V}_N^{\pi_s^*}(\boldsymbol{s}) \leq \mathcal{V}_\infty^{\pi_s^*}(\boldsymbol{s}) \leq h$. Hence, $\exists \pi^\dagger$ in the policy sequence that satisfies $\mathcal{V}_N^{\pi^\dagger}(\boldsymbol{s}) \leq h, \forall \boldsymbol{s} \in \mathcal{S}_s$. ■

D. Proof of Lemma 3

Lemma 3 (Constrained Policy Improvement): If a new policy π' is the optimal feasible solution of the maximization problem defined in (11). Then, we have $V^{\pi}(\boldsymbol{s}) \leq V^{\pi'}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}$.

Proof: According to (11), we can get

$$Q^{\pi}(\boldsymbol{s}, \pi'(\boldsymbol{s})) \geq Q^{\pi}(\boldsymbol{s}, \pi(\boldsymbol{s})) = V^{\pi}(\boldsymbol{s}), \quad \forall \boldsymbol{s} \in \mathcal{S}.$$

Then, we have

$$\begin{aligned}
V^{\pi}(\boldsymbol{s}) &\leq Q^{\pi}(\boldsymbol{s}, \pi'(\boldsymbol{s})) \\
&= \mathbb{E}\left[r_{t+1} + \gamma V^{\pi}\left(\boldsymbol{s}_{t+1}\right) \mid \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \pi'(\boldsymbol{s})\right] \\
&= \mathbb{E}_{\pi'}\left[r_{t+1} + \gamma V^{\pi}\left(\boldsymbol{s}_{t+1}\right) \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&\leq \mathbb{E}_{\pi'}\left[r_{t+1} + \gamma Q^{\pi}\left(\boldsymbol{s}_{t+1}, \pi'\left(\boldsymbol{s}_{t+1}\right)\right) \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&= \mathbb{E}_{\pi'}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 V^{\pi}\left(\boldsymbol{s}_{t+2}\right) \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&\leq \mathbb{E}_{\pi'}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 V^{\pi}\left(\boldsymbol{s}_{t+3}\right) \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&\ \ \vdots \\
&\leq \mathbb{E}_{\pi'}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \cdots \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&= V^{\pi'}(\boldsymbol{s}).
\end{aligned}$$

■

E. Proof of Lemma 4

Lemma 4 (Safety Guarantee): If a new policy π' is the optimal feasible solution of the maximization problem defined in (11). Then, we have $\mathcal{V}_N^{\pi}(\boldsymbol{s}) \geq \mathcal{V}_N^{\pi'}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}$.

Proof: According to (11), we can get, $\mathcal{Q}^{\pi}_{N-t}(\boldsymbol{s},\pi'(\boldsymbol{s})) \leq \mathcal{V}^{\pi}_{N-t}(\boldsymbol{s}), \quad t=0,...,N-1, \quad \forall \boldsymbol{s} \in \mathcal{S}$. Then, we have

$$\begin{aligned}
\mathcal{V}^{\pi}_{N}(\boldsymbol{s}) &\geq \mathcal{Q}^{\pi}_{N}(\boldsymbol{s},\pi'(\boldsymbol{s})) \\
&= \mathbb{E}\left[c_{t+1} + \mathcal{V}^{\pi}_{N-1}\left(\boldsymbol{s}_{t+1}\right) \mid \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \pi'(\boldsymbol{s})\right] \\
&= \mathbb{E}_{\pi'}\left[c_{t+1} + \mathcal{V}^{\pi}_{N-1}\left(\boldsymbol{s}_{t+1}\right) \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&\geq \mathbb{E}_{\pi'}\left[c_{t+1} + \mathcal{Q}^{\pi}_{N-1}\left(\boldsymbol{s}_{t+1}, \pi'\left(\boldsymbol{s}_{t+1}\right)\right) \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&= \mathbb{E}_{\pi'}\left[c_{t+1} + c_{t+2} + \mathcal{V}^{\pi}_{N-2}\left(\boldsymbol{s}_{t+2}\right) \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&\geq \mathbb{E}_{\pi'}\left[c_{t+1} + c_{t+2} + c_{t+3} + \mathcal{V}^{\pi}_{N-3}\left(\boldsymbol{s}_{t+3}\right) \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&\vdots \\
&\geq \mathbb{E}_{\pi'}\left[c_{t+1} + c_{t+2} + ... + c_{t+N} + \mathcal{V}^{\pi}_{0}\left(\boldsymbol{s}_{t+N}\right) \mid \boldsymbol{s}_t = \boldsymbol{s}\right] \\
&= \mathcal{V}^{\pi'}_{N}(\boldsymbol{s}).
\end{aligned}$$

■

F. Proof of Theorem 1

Theorem 1: Starting from an arbitrary policy π_0, the proposed FHCPI algorithm can eventually make the policy converge to a local optimal policy whose safety is guaranteed.

Proof: Starting from an initial policy π_0, according to Lemma 2, we have that a safe policy $\pi^{\dagger}$ that satisfies the constraint can be obtained by Algorithm 1. Then, by Lemma 3 and 4, we know that starting from a safe policy, a new policy obtained by Algorithm 2 will have higher returns, and it still satisfies the constraint. Also, since the policy space is finite, the proposed Algorithm 2 will converge to a policy such that no further improvement is feasible in terms of $V^{\pi}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}$. Thus, a local optimal of the problem defined in (8) is obtained, and the safety of the learned policy is guaranteed. ■

G. Discrete Situation When h=0

When we set $h=0$, the problem defined in (8) can be transformed into the following problem,

$$\min_{\pi} \quad -\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(\boldsymbol{s}_t,\boldsymbol{a}_t)\right], \quad \text{s.t.} \qquad \mathcal{V}^{\pi}_{N}(\boldsymbol{s}) = 0, \quad \forall \boldsymbol{s} \in \mathcal{S}_s. \tag{22}$$

Then, from an initial safe policy $\pi^{\dagger}$, we can modify the way of policy improvement in the second policy iteration stage as,

$$\pi'(\boldsymbol{s}) = \arg\max_{\tilde{\boldsymbol{a}}} Q^{\pi}(\boldsymbol{s},\tilde{\boldsymbol{a}}), \quad \forall \boldsymbol{s} \in \mathcal{X}, \quad \text{s.t.} \qquad \mathcal{Q}^{\pi}_{N}(\boldsymbol{s},\tilde{\boldsymbol{a}}) = 0. \tag{23}$$

Note that according to the definition of the risk value function, when $\mathcal{Q}^{\pi}_{N}(\boldsymbol{s},\tilde{\boldsymbol{a}}) = 0$, we have $\mathcal{Q}^{\pi}_{N-t}(\boldsymbol{s},\tilde{\boldsymbol{a}}) = 0, t=0,...,N-1$.

Definition 2: The safe policy set is defined as, $\Pi_{\text{safe}} = \{\pi \mid \mathcal{V}^{\pi}_{N}(\boldsymbol{s}) = 0, \forall \boldsymbol{s} \in \mathcal{S}_s\}$.

Definition 3: For π, the safe action set under $\boldsymbol{s}$ is defined as, $\mathcal{I}^{\pi}(\boldsymbol{s}) = \{\boldsymbol{a} \mid \mathcal{Q}^{\pi}_{N}(\boldsymbol{s},\boldsymbol{a}) = 0\}$.

Proposition 2: For any policy $\pi \in \Pi_{\text{safe}}$, under the state $\boldsymbol{s}$, the safe action set $\mathcal{I}^{\pi}(\boldsymbol{s})$ is equivalent to the policy-independent set $\mathcal{I}^{*}(\boldsymbol{s}) = \{\boldsymbol{a} \mid P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) = 0, \boldsymbol{s}' \in \mathcal{S}_u\}$, which means $\mathcal{I}^{\pi}(\boldsymbol{s}) \iff \mathcal{I}^{*}(\boldsymbol{s})$.

Proof: **Necessity**: By Definition 3, for the policy $\pi \in \Pi_{\text{safe}}$, under state $\boldsymbol{s} \in \mathcal{S}_s$, for all $\boldsymbol{a} \in \mathcal{I}^{\pi}(\boldsymbol{s})$ we have $\mathcal{Q}^{\pi}_{N}(\boldsymbol{s},\boldsymbol{a}) = 0$. Thus, the next state must be a safe state, so we can get $P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) = 0, \boldsymbol{s}' \in \mathcal{S}_u$.

Sufficiency: Due to $\pi \in \Pi_{\text{safe}}$, according to Definition 2 we can get $\mathcal{V}^{\pi}_{N}(\boldsymbol{s}) = 0, \forall \boldsymbol{s} \in \mathcal{S}_s$, then we have $\mathcal{V}^{\pi}_{N-1}(\boldsymbol{s}) = 0, \forall \boldsymbol{s} \in \mathcal{S}_s$. In addition, because of $P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) = 0, \boldsymbol{s}' \in \mathcal{S}_u$, we can get,

$$\begin{aligned}
\mathcal{Q}^{\pi}_{N}(\boldsymbol{s},\boldsymbol{a}) &= \sum_{\boldsymbol{s}' \in \mathcal{S}} P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})[c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \mathcal{V}^{\pi}_{N-1}(\boldsymbol{s}')] \\
&= \sum_{\boldsymbol{s}' \in \mathcal{S}_u} 0 * [c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \mathcal{V}^{\pi}_{N-1}(\boldsymbol{s}')] + \sum_{\boldsymbol{s}' \in \mathcal{S}_s} P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) * [c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + 0] \\
&= 0 + \sum_{\boldsymbol{s}' \in \mathcal{S}_s} P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) * [0 + 0] = 0.
\end{aligned}$$

■

Definition 4: The constrained Bellman optimality operator $\mathcal{B}^{*}$ is defined as, $\mathcal{B}^{*}V(\boldsymbol{s}) = \max_{\boldsymbol{a} \in \mathcal{I}^{*}(\boldsymbol{s})} \sum_{\boldsymbol{s}'} P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\left[r(\boldsymbol{s},\boldsymbol{a}) + \gamma V(\boldsymbol{s}')\right]$.

Proposition 3: The constrained Bellman optimality operator $\mathcal{B}^{*}$ is a contraction mapping.

Proof: For $\forall \pi_1, \pi_2 \in \Pi_{\text{safe}}$, under the state $\boldsymbol{s} \in \mathcal{S}$, we have,

$$\begin{aligned}
\|\mathcal{B}^* V^{\pi_1} - \mathcal{B}^* V^{\pi_2}\|_\infty &= \sup_{\boldsymbol{s}} \left\{ \left| \max_{\boldsymbol{a}_1 \in \mathcal{I}^*(\boldsymbol{s})} \mathbb{E}_{\boldsymbol{s}' \sim P} \left[r(\boldsymbol{s}, \boldsymbol{a}_1) + \gamma V^{\pi_1}(\boldsymbol{s}') \right] - \max_{\boldsymbol{a}_2 \in \mathcal{I}^*(\boldsymbol{s})} \mathbb{E}_{\boldsymbol{s}' \sim P} \left[r(\boldsymbol{s}, \boldsymbol{a}_2) + \gamma V^{\pi_2}(\boldsymbol{s}') \right] \right| \right\} \\
&\leq \sup_{\boldsymbol{s}} \left\{ \left| \max_{\boldsymbol{a}_1 \in \mathcal{I}^*(\boldsymbol{s})} \left(\mathbb{E}_{\boldsymbol{s}' \sim P} \left[r(\boldsymbol{s}, \boldsymbol{a}_1) + \gamma V^{\pi_1}(\boldsymbol{s}') \right] \right) - \mathbb{E}_{\boldsymbol{s}' \sim P} \left[r(\boldsymbol{s}, \boldsymbol{a}_1) + \gamma V^{\pi_2}(\boldsymbol{s}') \right] \right| \right\} \\
&\overset{(a)}{\leq} \gamma \sup_{\boldsymbol{s}} \left\{ \max_{\boldsymbol{a} \in \mathcal{I}^*(\boldsymbol{s})} \left| \mathbb{E}_{\boldsymbol{s}' \sim P(\cdot|\boldsymbol{s}, \boldsymbol{a})} \left[V^{\pi_1}(\boldsymbol{s}') - V^{\pi_2}(\boldsymbol{s}') \right] \right| \right\} \\
&\leq \gamma \sup_{\boldsymbol{s}} \left\{ \max_{\boldsymbol{a} \in \mathcal{I}^*(\boldsymbol{s})} \mathbb{E}_{\boldsymbol{s}' \sim P(\cdot|\boldsymbol{s}, \boldsymbol{a})} \left[\left| V^{\pi_1}(\boldsymbol{s}') - V^{\pi_2}(\boldsymbol{s}') \right| \right] \right\} \\
&= \gamma \sup_{\boldsymbol{s}} \max_{\boldsymbol{a} \in \mathcal{I}^*(\boldsymbol{s})} \left\{ \mathbb{E}_{\boldsymbol{s}' \sim P(\cdot|\boldsymbol{s}, \boldsymbol{a})} \left[\left| V^{\pi_1}(\boldsymbol{s}') - V^{\pi_2}(\boldsymbol{s}') \right| \right] \right\} \\
&\leq \gamma \max_{\boldsymbol{s}'} \left| V^{\pi_1}(\boldsymbol{s}') - V^{\pi_2}(\boldsymbol{s}') \right| \\
&= \gamma \| V^{\pi_1} - V^{\pi_2} \|_\infty .
\end{aligned}$$

For (a), this is because,

$$\begin{aligned}
|\max_x f(x) - g(x)| &= |f(x^*) - g(x^*)|, \quad (\text{let } x^* = \arg\max_x f(x)) \\
&\leq \max_x |f(x) - g(x)|.
\end{aligned}$$

■

Theorem 2: When $h = 0$, the proposed two-stage FHCPI algorithm can get a global optimal policy π^* for the problem defined in (8) that $V^{\pi^*}(\boldsymbol{s}) \geq V^\pi(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}, \quad \forall \pi \in \{\pi \mid \mathcal{V}_N^\pi(\boldsymbol{s}) = 0, \forall \boldsymbol{s} \in \mathcal{S}_s\}$.

Proof: By Proposition 3, $\mathcal{B}^*$ is a contraction mapping. According to Banach's fixed-point theorem, we have that π^* obtained by Algorithm 2 must be the global optimal solution for the problem shown in (22), and the corresponding value function V^{π^*} is unique. In addition, when $h = 0$, the problem defined in (22) is equivalent to the problem presented in (8). Thus, the proposed two-stage fixed-horizon policy iteration algorithm can get a global optimal policy π^* for the problem shown in (8). ■

Therefore, when we set $h = 0$, the learned policy π^* via the two-stage fixed-horizon policy iteration algorithm satisfies the safety constraint in problem defined in (8), i.e., $\mathbb{E}_{\boldsymbol{s} \sim \phi(\cdot)}[\mathcal{V}_N^\pi(\boldsymbol{s})] = 0$, and it is the global optimal policy in terms of the discounted cumulative reward.

H. Proof of Lemma 5

Proposition 4: When the distribution between two policies is close, i.e., the KL-divergence between the two policies is sufficiently small $\mathbb{E}_{\boldsymbol{s}}[D(\pi'(\cdot|\boldsymbol{s})||\pi(\cdot|\boldsymbol{s}))] \leq \delta$, we have, $\mathbb{E}_{\pi'} \left[\sum_{k=0}^{N-2} \mathcal{V}_N^\pi(\boldsymbol{s}_{t+k+1}) \middle| \boldsymbol{s}_t = \boldsymbol{s} \right] \approx \mathbb{E}_{\pi'} \left[\sum_{k=0}^{N-1} \mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+k+1}) \middle| \boldsymbol{s}_t = \boldsymbol{s} \right]$.

Proof:

$$\begin{aligned}
\mathbb{E}_{\pi'} \left[\sum_{k=0}^{N-2} \mathcal{V}_N^\pi(\boldsymbol{s}_{t+k+1}) \middle| \boldsymbol{s}_t = \boldsymbol{s} \right] &= \mathbb{E}_{\pi'} \left[\sum_{k=0}^{N-2} \mathbb{E}_\pi \left[c(\boldsymbol{s}_{t+k+1}, \boldsymbol{a}_{t+k+1}) + \mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+k+2}) \middle| \boldsymbol{s}_{t+k+1} \right] \middle| \boldsymbol{s}_t = \boldsymbol{s} \right] \\
&= \mathbb{E}_{\pi'} \left[\sum_{k=0}^{N-2} \mathbb{E}_\pi \left[c(\boldsymbol{s}_{t+k+1}, \boldsymbol{a}_{t+k+1}) \middle| \boldsymbol{s}_{t+k+1} \right] + \sum_{k=0}^{N-2} \mathbb{E}_\pi \left[\mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+k+2}) \middle| \boldsymbol{s}_{t+k+1} \right] \middle| \boldsymbol{s}_t = \boldsymbol{s} \right] \\
&\overset{*}{\approx} \mathbb{E}_{\pi'} \left[\sum_{k=0}^{N-2} \mathbb{E}_\pi \left[c(\boldsymbol{s}_{t+k+1}, \boldsymbol{a}_{t+k+1}) \middle| \boldsymbol{s}_{t+k+1} \right] + \sum_{k=0}^{N-2} \mathbb{E}_{\pi'} \left[\mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+k+2}) \middle| \boldsymbol{s}_{t+k+1} \right] \middle| \boldsymbol{s}_t = \boldsymbol{s} \right] \\
&\overset{\dagger}{\approx} \mathbb{E}_{\pi'} \left[\mathbb{E}_\pi \left[\sum_{k=0}^{N-2} c(\boldsymbol{s}_{t+k+1}, \boldsymbol{a}_{t+k+1}) \middle| \boldsymbol{s}_{t+1} \right] + \sum_{k=0}^{N-2} \mathbb{E}_{\pi'} \left[\mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+k+2}) \middle| \boldsymbol{s}_{t+k+1} \right] \middle| \boldsymbol{s}_t = \boldsymbol{s} \right] \\
&= \mathbb{E}_{\pi'} \left[\mathbb{E}_\pi \left[\sum_{k=0}^{N-2} c(\boldsymbol{s}_{t+k+1}, \boldsymbol{a}_{t+k+1}) \middle| \boldsymbol{s}_{t+1} \right] + \sum_{k=1}^{N-1} \mathbb{E}_{\pi'} \left[\mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+k+1}) \middle| \boldsymbol{s}_{t+k} \right] \middle| \boldsymbol{s}_t = \boldsymbol{s} \right] \\
&= \mathbb{E}_{\pi'} \left[\mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+1}) + \sum_{k=1}^{N-1} \mathbb{E}_{\pi'} \left[\mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+k+1}) \middle| \boldsymbol{s}_{t+k} \right] \middle| \boldsymbol{s}_t = \boldsymbol{s} \right] \\
&= \mathbb{E}_{\pi'} \left[\mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+1}) + \sum_{k=1}^{N-1} \mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+k+1}) \middle| \boldsymbol{s}_t = \boldsymbol{s} \right] \\
&= \mathbb{E}_{\pi'} \left[\sum_{k=0}^{N-1} \mathcal{V}_{N-1}^\pi(\boldsymbol{s}_{t+k+1}) \middle| \boldsymbol{s}_t = \boldsymbol{s} \right].
\end{aligned}$$

$*$: π is close to π'.

$\dagger$: The distribution of $\{c(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}), c(\boldsymbol{s}_{t+2}, \boldsymbol{a}_{t+2}), \cdots, c(\boldsymbol{s}_{t+N-1}, \boldsymbol{a}_{t+N-1})\}$ under π, π' is similar, because π is close to π'. ■

Lemma 5 (Approximate Performance Difference): If we let $\mathbb{E}_{\boldsymbol{s}}[D(\pi'(\cdot|\boldsymbol{s})||\pi(\cdot|\boldsymbol{s}))] \leq \delta$, where δ is a small positive real number that is close to 0, then we can have that $\mathcal{V}_N^{\pi'}(\boldsymbol{s}) - \mathcal{V}_N^\pi(\boldsymbol{s}) \approx \mathbb{E}_{\pi'} \left[\sum_{k=0}^{N-1} \mathcal{A}_N^\pi(\boldsymbol{s}_{t+k}, \boldsymbol{a}_{t+k}) \middle| \boldsymbol{s}_t = \boldsymbol{s} \right]$, where $\mathcal{A}_N^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathcal{Q}_N^\pi(\boldsymbol{s}, \boldsymbol{a}) - \mathcal{V}_N^\pi(\boldsymbol{s})$, and $D(\cdot||\cdot)$ is some kind of distance metric.

Proof:

$$\begin{aligned}
\mathcal{V}_N^{\pi'}(\boldsymbol{s}) - \mathcal{V}_N^{\pi}(\boldsymbol{s}) &= \mathbb{E}_{\pi'}\left[\sum_{k=0}^{N-1} c(\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k})\middle|\boldsymbol{s}_t=\boldsymbol{s}\right] - \mathcal{V}_N^{\pi}(\boldsymbol{s}) \\
&= \mathbb{E}_{\pi'}\left[\sum_{k=0}^{N-1} \left(c(\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k}) + \mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k}) - \mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k})\right)\middle|\boldsymbol{s}_t=\boldsymbol{s}\right] - \mathcal{V}_N^{\pi}(\boldsymbol{s}) \\
&= \mathbb{E}_{\pi'}\left[\sum_{k=0}^{N-1} c(\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k}) + \sum_{k=0}^{N-1}\mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k}) - \sum_{k=0}^{N-1}\mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k}))\middle|\boldsymbol{s}_t=\boldsymbol{s}\right] - \mathcal{V}_N^{\pi}(\boldsymbol{s}) \\
&= \mathbb{E}_{\pi'}\left[\sum_{k=0}^{N-1} c(\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k}) + \sum_{k=0}^{N-2}\mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k+1}) - \sum_{k=0}^{N-1}\mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k}))\middle|\boldsymbol{s}_t=\boldsymbol{s}\right] \\
&\approx \mathbb{E}_{\pi'}\left[\sum_{k=0}^{N-1} c(\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k}) + \sum_{k=0}^{N-1}\mathcal{V}_{N-1}^{\pi}(\boldsymbol{s}_{t+k+1}) - \sum_{k=0}^{N-1}\mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k})\middle|\boldsymbol{s}_t=\boldsymbol{s}\right] \text{(Prop.4)} \\
&= \mathbb{E}_{\pi'}\left[\sum_{k=0}^{N-1} \left(c(\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k}) + \mathcal{V}_{N-1}^{\pi}(\boldsymbol{s}_{t+k+1}) - \mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k})\right)\middle|\boldsymbol{s}_t=\boldsymbol{s}\right] \\
&= \mathbb{E}_{\pi'}\left[\sum_{k=0}^{N-1} \left(c(\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k}) + \mathbb{E}[\mathcal{V}_{N-1}^{\pi}(\boldsymbol{s}_{t+k+1})|\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k}] - \mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k})\right)\middle|\boldsymbol{s}_t=\boldsymbol{s}\right] \\
&= \mathbb{E}_{\pi'}\left[\sum_{k=0}^{N-1} \left(\mathcal{Q}_N^{\pi}(\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k}) - \mathcal{V}_N^{\pi}(\boldsymbol{s}_{t+k})\right)\middle|\boldsymbol{s}_t=\boldsymbol{s}\right] \\
&= \mathbb{E}_{\pi'}\left[\sum_{k=0}^{N-1} \mathcal{A}_N^{\pi}(\boldsymbol{s}_{t+k},\boldsymbol{a}_{t+k})\middle|\boldsymbol{s}_t=\boldsymbol{s}\right].
\end{aligned}$$

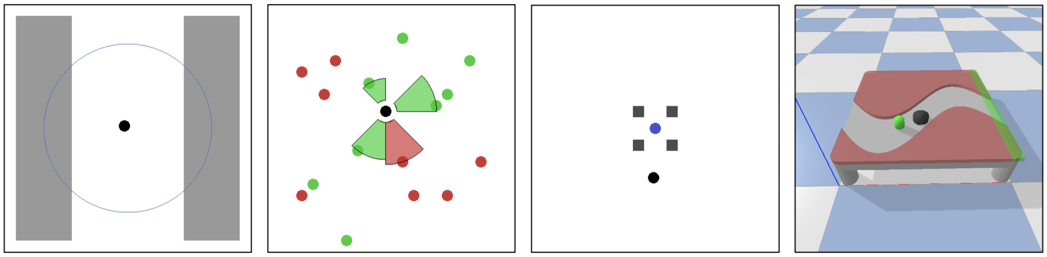
■

I. Risk-Sensitive Environment Description

The four risk-sensitive environments used in the continuous experiments are illustrated in Fig. 7. **Circle**: This environment is mainly based on [18]. In a 2-dimensional space with a size of 20×20, the agent is aimed at moving along a circle while avoiding entering unsafe areas. To achieve this target, the reward is designed as

$$r = \frac{-dx \cdot (y-10) + dy \cdot (x-10)}{1 + |\sqrt{(x-10)^2 + (y-10)^2} - 7.5|}, \tag{24}$$

where (x, y) is the position of the agent in the environment, and the center of the coordinate system is taken at position $(10, 10)$. The cost is defined as $\mathbb{1}\{|x-10| \geq 5\}$. **Gather**: This environment is also based on [18]. In a 2-dimensional space with a size of 20×20, the agent aims to collect as many green points as possible while avoiding collecting red points. To achieve this target, the agent is rewarded for collecting a green point (+10) and punished by collecting a red point (-10). The cost is defined as $\mathbb{1}\{\text{collect a red point in current step}\}$. To further evaluate the performance of the proposed algorithm, we build two additional environments with greater risk: Safe-Reach and Safe-Push. **Safe-Reach**: In a 2-dimensional space with a size of 20×20, the agent aims to reach the target while avoiding entering unsafe areas. To achieve this target, the agent is rewarded for reaching the target (+100) and is punished by moving forward $(-0.5 \times \text{movement length})$. The episode length is 100. **Safe-Push**: The agent aims to push the car into the target area while ensuring that the car is not pushed into unsafe areas. To achieve this, the agent is rewarded for pushing the car to the right $(+1000 \times \Delta x)$ and making the car reach the target $(+1000)$. The cost is defined as $\mathbb{1}\{\text{the car is in unsafe areas}\}$.



(a) Circle (b) Gather (c) Safe-Reach (d) Safe-Push

Fig. 7. Four continuous environments. (a) Circle: the agent is rewarded for moving along a circle, but it needs to avoid entering into unsafe areas (gray regions). (b) Gather: the agent is rewarded for gathering green points, and it should avoid collecting red points. (c) Safe-Reach: the agent is rewarded for reaching the target (blue), but it should avoid entering unsafe areas (gray squares). (d) Safe-Push: the agent (green) is rewarded for pushing the car (black cylinder) into the goal (green areas), but it must ensure that the car is not pushed into unsafe areas (red).

J. Parameters of Continuous Experiments

The details parameters of the experiment and the algorithms are shown in Table II and Table III.

TABLE II
PARAMETERS OF EMPIRICAL CONTINUOUS EXPERIMENTS OF DIFFERENT ALGORITHMS.

Parameter	FHCPO	PPO [17]	CPO [18]	IPO [30]	PCPO [31]
Fixed-Horizon N	5	N/A	N/A	N/A	N/A
Discounted factor γ	0.99	0.99	0.99	0.99	0.99
GAE λ^{GAE}	0.97	0.97	0.97	0.97	0.97
Activation function	Tanh	Tanh	Tanh	Tanh	Tanh
Number of hidden layers	2	2	2	2	2
Number of hidden nodes	64	64	64	64	64
Learning rate of the policy net	N/A	3e-4	N/A	3e-4	N/A
Learning rate of the reward critic net	1e-3	1e-3	1e-3	1e-3	1e-3
Learning rate of the cost critic net	1e-3	N/A	1e-3	1e-3	1e-3
Clip ratio	N/A	0.2	N/A	0.2	N/A
Penalty factor	N/A	N/A	N/A	1e-8	N/A
Max KL divergence δ	0.01	0.01	0.01	0.01	See Table III

TABLE III
PARAMETERS OF EMPIRICAL EXPERIMENTS IN CONTINUOUS ENVIRONMENTS.

Parameter	Circle	Gather	Safe-Reach	Safe-Push
Epochs	1000	1000	500	5000
Steps per epoch	500	300	1000	1000
Safe threshold h	0.1	0.005	0.1	0.1
Cost threshold C	4	0.1	2	10
Max KL divergence δ for PCPO *	1e-4	1e-4	5e-4	5e-4

* Due to the instability training of PCPO, different δ are used for different environments.

Basic requirements and rules for the selection of the three important parameters (N, h, and γ) are as follows:

- The fixed-horizon N: This parameter determines how many time steps the agent considers in the future. If N is large, risks in the long distant future will be taken into account by the agent, which makes the agent's actions conservative. If N is small, the agent only considers risks for a short period of time in the future, which will cause the agent to behave aggressively. Therefore, N is usually set according to the specific industrial problem. In general, for problems with long-delayed or sparse costs, the value of N should be larger, such as a water conservancy dispatch problem, in which an action to open the floodgates will take a long time to affect the amount of water downstream. Conversely, for low-speed robot manipulation problems where collisions are considered, N can be set to a relatively small value. In summary, the value of N depends on the characteristics of the industrial problem, e.g., the dynamics of the environment or the setting of the costs.
- Safe threshold h: This value determines the tolerance for risk occurrence in different industrial scenarios. It limits the probability that risks will occur in the next N time steps in the future. Usually, its value should be close to 0 (e.g., 0.001).
- The discount factor γ: To make the future rewards fully considered, γ is commonly set to a value close to 1. In addition, to ensure that the Bellman operator is a contraction mapping, the value of γ cannot be taken directly as 1. Therefore, the value of γ is usually set to 0.99.

Other parameters such as the learning rate can be set as default parameters in the optimizer.