Quantile Constrained Reinforcement Learning: A Reinforcement Learning Framework Constraining Outage Probability

Whiyoung Jung, Myungsik Cho, Jongeui Park, Youngchul Sung*
School of Electrical Engineering, KAIST
Daejeon 34141, Republic of Korea
{wy.jung, ms.cho, jongeui.park, ycsung}@kaist.ac.kr

Abstract

Constrained reinforcement learning (RL) is an area of RL whose objective is to find an optimal policy that maximizes expected cumulative return while satisfying a given constraint. Most of the previous constrained RL works consider expected cumulative sum cost as the constraint. However, optimization with this constraint cannot guarantee a target probability of outage event that the cumulative sum cost exceeds a given threshold. This paper proposes a framework, named Quantile Constrained RL (QCRL), to constrain the quantile of the distribution of the cumulative sum cost that is a necessary and sufficient condition to satisfy the outage constraint. This is the first work that tackles the issue of applying the policy gradient theorem to the quantile and provides theoretical results for approximating the gradient of the quantile. Based on the derived theoretical results and the technique of the Lagrange multiplier, we construct a constrained RL algorithm named Quantile Constrained Policy Optimization (QCPO). We use distributional RL with the Large Deviation Principle (LDP) to estimate quantiles and tail probability of the cumulative sum cost for the implementation of QCPO. The implemented algorithm satisfies the outage probability constraint after the training period.

1 Introduction

Reinforcement learning (RL) has been developed in the direction of finding an optimal policy that maximizes expected cumulative return for a given environment. Thus, most of the works in RL consider only rewards given by the environment to optimize the policy. However, many real-world control problems impose constraints on the behavior of a policy. Constrained RL is an area of RL whose objective is to find an optimal policy that maximizes expected cumulative return while satisfying a certain constraint on the cumulative cost. A conventional constrained RL problem (ExpCP) can be written as

$$\begin{array}{ll} \text{Maximize} & V^\pi(s_0) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] \\ \text{subject to} & C^\pi(s_0) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right] \le d_{th}, \end{array} \tag{ExpCP}$$

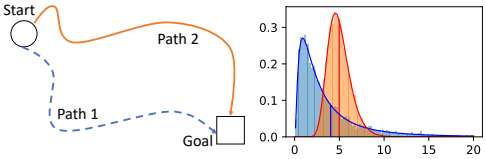
where the cost constraint is that the expectation of the sum of costs is less than or equal to a threshold parameter d_{th}. Note that the threshold d_{th} is set on the average (i.e., expectation) of the cumulative sum cost to avoid undesired high-cost events in this formulation. If we do not want any event causing a positive cost, d_{th} should be set as a sufficiently small value, i.e., $d_{th} \approx 0$. On the other hand, if we can afford events with low costs, we can set d_{th} properly as we desire. Most of the previous constrained RL works solved the problem (ExpCP) [1, 12, 14, 24, 25, 28, 29] partly because the

*Corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

constraint on the expectation of the cumulative sum cost in (ExpCP) is well fit with the objective given by the expectation of the sum reward, and this makes the problem amenable. However, solving the problem (ExpCP) may have an undesirable outcome for real environments that typically need a constrained behavior on the event that the cost exceeds the threshold d_{th}. For example, in the case of an autonomous driving car, what we want for our sure safety is to control and limit the probability of accident itself. In the case of a telecommunication system, what we want to control is the probability of packet loss through the communication system. These probabilities are called 'outage probability' in general. Thus, in many real-world systems, the system requires a constraint on the outage probability, i.e., the probability of critical or unsafe events. In this case, a constraint on the expectation of critical events, as in (ExpCP), cannot guarantee the desired target probability of critical events. To illustrate this, let us consider the following example.

Consider a simple two-path environment, as shown in Fig. 1a. The objective of the environment is for an agent to reach the goal by driving a car. The environment gives a reward when the agent reaches the goal, and gives costs until the agent reaches the goal. There exist two paths to reach the goal, and each path has a different cost distribution. Fig. 1b shows the distribution of the cumulative sum cost (curve) and its mean (vertical line) for each path: blue for path 1 and red for path 2. As we can see in Fig. 1b, the distribution for path 1 has a lower average than that for path 2, but has a longer right tail than that for path 2. If we use the expectation of cost as a constraint, then following path 1 is a better choice, but in this case, the probability of a high cumulative sum cost (e.g. > 10.0 in Fig. 1b) is higher than following path 2. If the threshold of 10 represents a catastrophic event, we should constrain the probability of events exceeding 10 to a small value. Thus, solving the problem with the expectation constraint (ExpCP) does not necessarily have precise control over the target outage probability. When the event that the cumulative sum cost exceeds d_{th} is a critical unsafe event, this means that such critical event can occur in high probability even if we solve the constrained problem (ExpCP). Therefore, in this paper, we aim to solve the following constrained RL problem with an outage probability constraint:



(a) Two-path environment (b) Distribution of cost

Figure 1: A simple example of environment.

$$\begin{aligned} &\text{Maximize} && V^{\pi}(s_0) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\ &\text{Subject to} && \Pr\left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t) > d_{th}\right] \leq \epsilon_0 \\ &&& \text{for } S_0 = s_0, A_t \sim \pi(\cdot|S_t), S_{t+1} \sim M(\cdot|S_t, A_t). \end{aligned} \tag{ProbCP}$$

Our approach to this problem is first to convert the outage probability constraint in (ProbCP) into a quantile constraint $q^{\pi}_{1-\epsilon_0}(s_0) := \inf\{x \mid \Pr\left(\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t) \leq x\right) \geq 1-\epsilon_0\} \leq d_{th}$ which is equivalent to the outage probability constraint (See Fig. 2), and then to solve the optimization:

$$\min_{\lambda \geq 0} \max_{\pi} \; \{V^{\pi}(s_0) - \lambda \left(q^{\pi}_{1-\epsilon_0}(s_0) - d_{th}\right)\}, \tag{1}$$

where λ is the Lagrange multiplier, based on policy gradient with the parameterized policy. However, we note that the policy gradient theorem, which is the most basic theorem for on-policy RL, cannot be applied directly to compute the gradient of (1) with respect to (w.r.t.) the policy parameter due to the quantile term $q^{\pi}_{1-\epsilon_0}(s_0)$. Therefore, we derive theoretical results for approximating the gradient of the quantile (in Section 3). Then, based on the derived theoretical results and the technique of the Lagrange multiplier, we construct our algorithm named Quantile Constrained Policy Optimization (QCPO) to solve the outage probability constrained RL problem (in Section 4). Here, we use distributional RL with the Large Deviations Principle (LDP) to estimate quantiles and tail probability of the cumulative sum cost for implementation of QCPO. The implemented algorithm satisfies the outage probability constraint after the training period. To the best of our knowledge, this is the first work that tackles the issue of applying the policy gradient theorem to the quantile and obtains an (approximate) policy gradient for the quantile, and this is one of the main contributions of this paper, together with the QCPO algorithm.

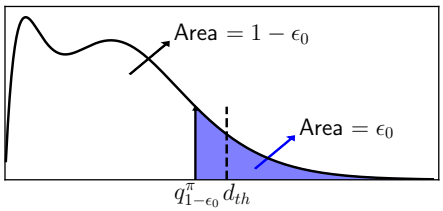


Figure 2: Equivalence between the outage probability constraint and the quantile constraint

2 Background and Related Works

Constrained RL A constrained Markov decision process (CMDP) is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, r, c, M, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ is the cost function, $M : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability, and γ is the discount factor.

Constrained RL or Safe RL is an area of RL whose objective is to find an optimal policy for a given CMDP that maximizes the expected return $\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$ while satisfying a constraint on the cumulative sum cost $\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)$. If one wants to constrain the average cumulative sum cost of the policy π, then constraint $C^\pi(s_0) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] \leq d_{th}$ can be considered. On the other hand, if one wants to constrain the outage behavior of the policy π, then constraint $\Pr\left(X^\pi(s_0) > d_{th}\right) \leq \epsilon_0$ should be considered, where $X^\pi(s_0)$ is a random variable defined as $X^\pi(s_0) := \sum_{t=0}^{\infty} \gamma^t c(S_t, A_t)$ with $S_0 = s_0$, $A_t \sim \pi(\cdot|S_t)$, $S_{t+1} \sim M(\cdot|S_t, A_t)$ for $t = 0, 1, \cdots$.

Most of the previous constrained RL works considered a constraint on the expectation of the cumulative sum cost: $C^\pi(s_0) \leq d_{th}$ [1, 12, 14, 24, 25, 28, 29]. In order to solve the constrained optimization problem with this expectation-based constraint, researchers considered the Lagrangian multiplier method [1, 21], Lyapunov-based methods [6, 7], projection-based methods [28], safety-layer methods [10].

Table 1: Comparison of constrained RL with probabilistic constraints

Papers	Algorithm Type	Distribution Modeling	Theory	Deep RL
Risk-Const. RL [5]	On-policy (Trajectory-based)	No (need only trajectory samples)	Yes	No
WCSAC [27]	Off-policy	Gaussian (on all range of distribution)	No	Yes
QCPO (This paper)	On-policy (State-based)	LDP with Weibull (only on tail)	Yes	Yes

Quantile (i.e., Value at Risk, VaR) and *Conditional Value at Risk (CVaR)* are two well-known techniques to manage undesirable events in the domain of finance[18]. In the context of RL, the definitions of the quantile and the CVaR for the distribution of the cumulative sum cost for a given π are given by $q_u^\pi(s_0) := \inf\{x \mid \Pr(X^\pi(s_0) \leq x) \geq u\}$ and $\text{CVaR}_u^\pi(s_0) := \mathbb{E}_\pi \left[X^\pi(s_0) \mid X^\pi(s_0) \geq q_u^\pi(s_0)\right]$, respectively.

The CVaR was previously used in RL to constrain undesirable events and the problem with a CVaR constraint is explicitly formulated as

$$\begin{array}{ll} \text{Maximize} & \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\ \text{Subject to} & \text{CVaR}_{1-\epsilon_0}^\pi(s_0) \leq d_{th}. \end{array} \tag{CVaR-CP}$$

Note that the $(1 - \epsilon_0)$-CVaR denoted as $\text{CVaR}_{1-\epsilon_0}^\pi(s)$ is always greater than or equal to the $(1 - \epsilon_0)$-quantile denoted as $q_{1-\epsilon_0}^\pi(s)$ for all $s \in \mathcal{S}$ because of the definition of the CVaR. Therefore, satisfying the CVaR constraint $\text{CVaR}_{1-\epsilon_0}^\pi(s_0) \leq d_{th}$ in (CVaR-CP) is a sufficient condition for satisfying the probabilistic constraint $\Pr\left(X^\pi(s_0) > d_{th}\right) \leq \epsilon_0$ in (ProbCP), and hence (CVaR-CP) is a stricter problem than (ProbCP). Therefore, algorithms proposed to solve (CVaR-CP) can be used for solving (ProbCP), and this should satisfy the probabilistic constraint in theory.

Chow et al. [5] proposed a trajectory-based CVaR method and provided convergence for their method. They used trajectory-based policy gradient to their Lagrangian, and it is simple to compute. Like most trajectory-based RL algorithms, however, it suffers from sample inefficiency since it collects a number of trajectories and updates its parameter once. Recently, Yang et al. [27] proposed an off-policy algorithm to solve (CVaR-CP). They only estimated the mean and variance of the cost distribution using a technique in distributional RL and computed $\text{CVaR}_u^\pi(s)$[2] as the CVaR of the Gaussian distribution of the estimated mean and variance. However, the distribution of $X^\pi(s)$ is not Gaussian in general, and the Gaussian approximation has limited capability to capture the decay rate of the tail probability because a Gaussian probability density function (PDF) has the form of $\exp(-\beta x^2)$ with fixed rate function x^2. Therefore, this algorithm can yield a poor estimation

[2]They actually consider the CVaR $\text{CVaR}_u^\pi(s, a)$ of the cumulative sum cost $X^\pi(s, a)$ for a given (s, a) pair.

of the CVaR of the tail, especially for small tail probability, and cannot guarantee to satisfy the CVaR constraint (see Section 5). Furthermore, note that the CVaR and the quantile are two different measures for undesirable events, and the choice between the two depends on what we desire. For example, an insurance company prefers the CVaR of undesirable events to determine an insurance premium. On the other hand, a company developing an autonomous driving car system needs the quantile of undesirable events to guarantee the accident probability for safety. Thus, in the context of safe learning, our work focuses directly on *the constraint on the quantile*, which is an equivalent (i.e., necessary and sufficient) constraint to the outage probability constraint in (ProbCP). To the best of our knowledge, this is the first work that provides a state-based policy gradient for the quantile and required theoretical results regarding the quantile-constrained RL problem. Moreover, our implementation approximates the tail distribution of $X^\pi(s)$ with a Weibull distribution (a particular case of generalized Gamma distribution), which is general enough to capture various rates of decay of the tail probability. Table 1 summarizes the previous constrained RL methods and this paper.

Large Deviation Principle (LDP) Large deviation principle (LDP) [11] is a technique for estimating the limiting behavior of a sequence of distributions, especially on the tail. A simple example is the empirical mean $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ of i.i.d. random variables X_i. We say that a sequence $\{\bar{X}_n\}$ satisfies LDP if the sequence of its log probability distribution $\frac{1}{n}\log \Pr\left(\bar{X}_n \in \Gamma\right)$ satisfies the following condition $\frac{1}{n}\log \Pr\left(\bar{X}_n \in \Gamma\right) \overset{n\to\infty}{\longrightarrow} -\inf_{x\in\Gamma} I(x)$ for some function $I(x)$. The function $I(x)$ satisfying such limiting behavior is called the rate function of $\bar{X}_n$. The rate function $I(x)$ is also related to the cumulative distribution function (CDF) $F_{\bar{X}_n}(x)$ since $1 - F_{\bar{X}_n}(x_0) = \Pr\left(\bar{X}_n \in [x_0, \infty)\right) \approx \exp\left(-n \inf_{x\in[x_0,\infty)} I(x)\right)$ for some $x_0 > \mathbb{E}[X]$ and sufficiently large n. LDP can be applied to finite-state Markov chains, and there exists a rate function for a given Markov chain [11]. In this paper, we consider the tail probability of the distribution of the cumulative sum cost $X^\pi(s_0) = \sum_{t=0}^\infty \gamma^t c(s_t, a_t)$. Finding its analytic rate function is hard. Therefore, we instead approximate the rate function directly as $I_{X^\pi(s)}(x) \approx (x/\beta(s))^{\alpha(s)}$ with learnable parameters $\alpha(s)$ and $\beta(s)$, which results in a Weibull distribution: $1 - F_{X^\pi(s)}(x) = \exp\left\{-(x/\beta(s))^{\alpha(s)}\right\}$. We use this distribution to approximate the tail probability of $p_{X^\pi(s)}(x)$ of $X^\pi(s)$.

3 Quantile Constrained RL

In this section, we explain an equivalent form of (ProbCP) that we use to learn an optimal constrained policy under the outage probability constraint and then explain the difficulty of applying the policy gradient theorem to optimize the Lagrangian of the equivalent problem. Finally, we provide theoretical results that circumvent this difficulty in Section 3.2.

3.1 Motivation: Problem of Applying Policy Gradient Theorem to Quantile

Solving (ProbCP) with a direct approach is too hard in making a loss function for π based on the outage probability. Thus, we convert the probability constrained problem to an equivalent form of a quantile constrained problem:

$$\begin{array}{ll} \text{Maximize} & \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)\right] \\ \text{Subject to} & q^\pi_{1-\epsilon_0}(s_0) \le d_{th}, \end{array} \tag{QuantCP}$$

where $q^\pi_u(s) = \inf\{x \mid \Pr(X^\pi(s) \le x) \ge u\}$ is the u-quantile of the random variable $X^\pi(s)$ of the cumulative sum cost: $X^\pi(s) = \sum_{t=0}^\infty \gamma^t c(S_t, A_t)$ with $S_0 = s$, $A_t \sim \pi(\cdot|S_t)$, $S_{t+1} \sim M(\cdot|S_t, A_t)$, $t = 0, 1, 2, \cdots$. Note that $q^\pi_{1-\epsilon_0}(s_0) \le d_{th}$ is equivalent to $\Pr\left[X^\pi(s_0) > d_{th}\right] \le \epsilon_0$ due to the definition of the quantile. We propose a direct approach to solve the equivalent problem (QuantCP) instead of (ProbCP). Although we have an equivalent form of (ProbCP), it is still difficult to solve (QuantCP). We explain what makes solving the problem (QuantCP) still hard below.

In the case of (ExpCP), the Lagrange-based optimization of (ExpCP) is given by $\min_{\lambda\ge 0}\max_\pi L_{exp}(\pi, \lambda) := V^\pi(s_0) - \lambda\left(C^\pi(s_0) - d_{th}\right)$, where $V^\pi(s_0) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)\right]$ and $C^\pi(s_0) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t c(s_t, a_t)\right]$. Then, due to the form of $C^\pi(s_0)$, the policy gradient theorem [22] can directly be applied, and the gradient of the Lagrangian w.r.t. π is given by the expectation form:

$$\nabla_\pi L_{exp}(\pi, \lambda) = \sum_s \rho^\pi(s) \sum_a \nabla\pi(a|s)\left\{A^\pi_r(s,a) - \lambda A^\pi_c(s,a)\right\}, \tag{2}$$

where $\rho^\pi(s) := \sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s|s_0, \pi)$, $A_r^\pi(s,a) := r(s,a) + \gamma \mathbb{E}_{s'\sim M}\left[V^\pi(s')\right] - V^\pi(s)$, and $A_c^\pi(s,a) := c(s,a) + \gamma \mathbb{E}_{s'\sim M}\left[C^\pi(s')\right] - C^\pi(s)$. However, the gradient of the Lagrangian of the problem (QuantCP)

$$\min_{\lambda\geq 0} \max_{\pi} L_{quant}(\pi, \lambda) := V^\pi(s_0) - \lambda\left(q_{1-\epsilon_0}^\pi(s_0) - d_{th}\right) \tag{3}$$

w.r.t. the policy π cannot be expressed as an expectation form:

$$\nabla_\pi L_{quant}(\pi, \lambda) \neq \mathbb{E}_\pi\left[\nabla \log \pi(a|s)\left\{A_r^\pi(s,a) - \lambda \bar{A}_{1-\epsilon_0}^\pi(s,a)\right\}\right] \tag{4}$$

where $\bar{A}_u^\pi(s,a) := c(s,a) + \gamma \mathbb{E}_{s'\sim M}\left[q_u^\pi(s')\right] - q_u^\pi(s)$. This is because the u-quantile $q_u^\pi(s)$ is not the expectation of the cumulative sum cost. However, if the u-quantile $q_u^\pi(s)$ can be written as

$$q_u^\pi(s_0) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \left\{c(s_t, a_t) + \tilde{c}_u(s_t, a_t)\right\}\right] \tag{5}$$

for some function $\tilde{c}_u(s,a)$ that is independent of the policy π, we can apply the policy gradient theorem by defining the advantage function for the quantile term:

$$A_u^\pi(s,a) := c(s,a) + \tilde{c}_u(s,a) + \gamma \mathbb{E}_{s'\sim M}\left[q_u^\pi(s')\right] - q_u^\pi(s). \tag{6}$$

This fact motivates us to search for such $\tilde{c}_u(s,a)$. For this, under mild assumptions, we first show the existence of a policy-dependent additional cost $\tilde{c}_u^\pi(s,a)$ and then show that the additional cost can be approximated by a cost $\tilde{c}_u^{\pi'}(s,a)$ for some fixed π' independent of π except the requirement $\max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \leq \delta$.

3.2 Theoretical Results

We here provide theoretical results showing the existence of an additional cost $\tilde{c}_u^\pi(s,a)$ and showing that this can be approximated as another cost $\tilde{c}_u^{\pi'}(s,a)$ for a base policy π' independent of π, only requiring $\max_s \mathrm{KL}\left(\pi'(\cdot|s) \,\middle\|\, \pi(\cdot|s)\right) \leq \delta$ for some $\delta > 0$. These theoretical results make the quantile constrained policy optimization tractable by enabling application of the policy gradient theorem. For the theoretical results, we assume that the CDF $F_{X^\pi(s)}(x)$ is strictly increasing on $[0, \infty)$, and it is continuously differentiable for all $s \in \mathcal{S}$. The proofs of the theoretical results are in Appendix B.

We begin with deriving the temporal-difference (TD) relation between the u-quantiles of $X^\pi(s)$ at s_t and s_{t+1}. Theorem 1 states the TD relation for the u-quantile under the following assumptions of boundness of quantile difference and smoothness of CDF of $X^\pi(s)$.

Assumption 1 (Boundness of quantile difference)**.** *For a given policy π, the following two quantities are bounded*

$$|c(s,a) + \gamma q_u^\pi(s') - q_u^\pi(s)| \leq \gamma R \tag{7}$$

$$\left|q_u^\pi(s) - F_{X^\pi(s)}^{-1}\left(F_{X^\pi(s')}\left(\frac{q_u^\pi(s) - c(s,a)}{\gamma}\right)\right)\right| \leq R \tag{8}$$

for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ such that $\pi(a|s) \cdot M(s'|s,a) > 0$.

Assumption 2 (Smoothness of CDF of $X^\pi(s)$)**.** *For each state s, the average slope of $F_{X^\pi(s)}(x)$ between $q_u^\pi(s)$ and $y \in [q_u^\pi(s) - R, q_u^\pi(s) + R]$ is bounded by*

$$\frac{1}{1+\epsilon} \cdot p_{X^\pi(s)}\left(q_u^\pi(s)\right) \leq \frac{F_{X^\pi(s)}\left(q_u^\pi(s)\right) - F_{X^\pi(s)}\left(y\right)}{q_u^\pi(s) - y} \leq \frac{1}{1-\epsilon} \cdot p_{X^\pi(s)}\left(q_u^\pi(s)\right) \tag{9}$$

for small $0 < \epsilon < \frac{1}{2}$.

Theorem 1. *Under Assumptions 1 and 2, the u-quantile of the random variable $X^\pi(s_t)$ satisfies the following temporal-difference (TD) relation. For some constant R and small $\epsilon > 0$,*

$$\left|\mathbb{E}_\pi\left[\mu_u^\pi\left(s_t, a_t, s_{t+1}\right)\left\{c(s_t, a_t) + \gamma q_u^\pi(s_{t+1}) - q_u^\pi(s_t)\right\}\right]\right| \leq \frac{\epsilon}{1-\epsilon} R, \tag{10}$$

where $\mu_u^\pi\left(s, a, s'\right) = p_{X^\pi(s')}\left(\frac{q_u^\pi(s) - c(s,a)}{\gamma}\right) \,/\, \gamma p_{X^\pi(s)}\left(q_u^\pi(s)\right)$. Here, the expectation is for the action $a_t \sim \pi(\cdot|s_t)$ and the next state $s_{t+1} \sim M(\cdot|s_t, a_t)$. ($s_t$ is given.)

Note that for the expectation of the cumulative sum cost $C^\pi(s)$ considered in (ExpCP), the expectation of TD under the policy π follows

$$\mathbb{E}_\pi\left[c(s_t,a_t)+\gamma C^\pi(s_{t+1})-C^\pi(s_t)\right]=0 \tag{11}$$

by the Bellman equation. The TD relation (11) for expectation has a similar form to that for the u-quantile (10), but the difference is that (10) is the weighted expectation of the TD ($c(s_t,a_t)+\gamma q_u^\pi(s_{t+1})-q_u^\pi(s_t)$). The numerator $p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)$ of the weight $\mu_u^\pi(s_t,a_t,s_{t+1})$ in (10) involves two quantities: 1) a target quantile $\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}$ and 2) the PDF of the sum of costs $X^\pi(s_{t+1})=\sum_{k=0}^\infty\gamma^k c(s_{t+k+1},a_{t+k+1})$ starting from state s_{t+1}. Here, the value $\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}$ is the target value of the sum of costs $\sum_{k=0}^\infty\gamma^k c(s_{t+k+1},a_{t+k+1})$ from the next state s_{t+1} such that the sum of costs $\sum_{k=0}^\infty\gamma^k c(s_{t+k},a_{t+k})$ for a given pair (s_t,a_t) at t is the u-quantile $q_u^\pi(s_t)$. Thus, the numerator $p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)$ of the weight $\mu_u^\pi(s_t,a_t,s_{t+1})$ in (10) is the probability of the event that the cumulative sum cost starting from (s_t,a_t) becomes the u-quantile $q_u^\pi(s_t)$ at s_t from the perspective of the next state s_{t+1}. Based on Theorem 1, we obtain the following corollary:

Corollary 1. *Under Assumptions 1 and 2, the u-quantile $q_u^\pi(s_t)$ of the random variable $X^\pi(s_t)$ is bounded as*

$$\left|q_u^\pi(s_t)-\mathbb{E}_\pi\left[\mu_u^\pi\left(s_t,a_t,s_{t+1}\right)\left\{c(s_t,a_t)+\gamma q_u^\pi(s_{t+1})\right\}\right]\right|\le\frac{\epsilon}{1-\epsilon}R. \tag{12}$$

Proof: Note that the term $q_u^\pi(s_t)$ can go outside the expectation in (10) since the expectation is over (a_t,s_{t+1}). From eq. (26) in Appendix A.1, the expectation of the numerator of $\mu_u^\pi\left(s_t,a_t,s_{t+1}\right)$ is the same as the the denominator of the weight, i.e., $\mathbb{E}_\pi\left[p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)\right]=\gamma p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)$, and this leads to $\mathbb{E}_\pi[\mu_u^\pi(s_t,a_t,s_{t+1})]=1$. So, we have the claim. □

As seen in Corollary 1, the u-quantile at s_t can be approximated as a weighted expectation of $c(s_t,a_t)+\gamma q_u^\pi(s_{t+1})$, and the weight is proportional to $p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)$. This means that the more probable is the pair (s_t,a_t,s_{t+1}) to achieve $q_u^\pi(s_t)$, the higher weight is multiplied to $c(s_t,a_t)+\gamma q_u^\pi(s_{t+1})$ for approximating $q_u^\pi(s_t)$. Furthermore, if we assume that the transition dynamics of CMDP are deterministic, i.e., $s_{t+1}=h(s_t,a_t)$ as in many real-world control problem, we can approximate the u-quantile $q_u^\pi(s_0)$ at s_0 as the expectation of the sum of costs under a distorted policy $\tilde{\pi}_u$, as stated in the following lemma:

Lemma 1. *Suppose that the state transition dynamics are deterministic, i.e., $s_{t+1}=h(s_t,a_t)$. Then, under Assumptions 1 and 2, the u-quantile $q_u^\pi(s_0)$ of the random variable $X^\pi(s_0)$ is expressed as*

$$\left|q_u^\pi(s_0)-\mathbb{E}_{\tilde{\pi}_u}\left[\sum_{t=0}^\infty\gamma^t c(s_t,a_t)\right]\right|\le\frac{\epsilon R}{(1-\epsilon)(1-\gamma)}, \tag{13}$$

where $\tilde{\pi}_u(a|s)=\pi(a|s)\cdot\mu_u^\pi\left(s,a,h(s,a)\right)\propto\pi(a|s)\cdot p_{X^\pi(h(s,a))}\left(\frac{q_u^\pi(s)-c(s,a)}{\gamma}\right)$.

Now, plugging (13) into the quantile term in the Lagrangian (3) of the problem (QuantCP), we may apply the policy gradient theorem based on the chain rule since the u-quantile $q_u^\pi(s_0)$ is expressed as the expectation of the sum of costs. However, the gradient of $\tilde{\pi}_u$ w.r.t. π for chain rule is too complicated due to the μ_u^π term in Lemma 1. Thus, we find another expectation form of $q_u^\pi(s_0)$ using an additional cost function $\tilde{c}_u^\pi(s,a)$, as stated in the following theorem:

Theorem 2. *Under deterministic dynamics $s_{t+1}=h(s_t,a_t)$ and Assumptions 1 and 2, $q_u^\pi(s)$ can be expressed as*

$$\left|q_u^\pi(s_0)-\mathbb{E}_\pi\left[\sum_{t=0}^\infty\gamma^t\left\{c(s_t,a_t)+\tilde{c}_u^\pi(s_t,a_t)\right\}\right]\right|\le\frac{\epsilon R}{(1-\epsilon)(1-\gamma)}, \tag{14}$$

where $\tilde{c}_u^\pi(s,a)=\left(\mu_u^\pi\left(s,a,h(s,a)\right)-1\right)\cdot\left[c(s,a)+\gamma q_u^\pi(h(s,a))\right]$.

Note that the additional cost $\tilde{c}_u^\pi(s_t,a_t)$ in Theorem 2 is a policy-dependent cost function. Under an additional mild assumption, we can find an upper bound of (14) which replaces the policy-dependent cost function $\tilde{c}_u^\pi(s,a)$ with another cost $\tilde{c}_u^{\pi'}(s,a)$ for some fixed π' independent of π, only requiring $\max_s \mathrm{KL}\left(\pi'(\cdot|s)\,\middle\|\,\pi(\cdot|s)\right)\le\delta$ for some $\delta>0$. The additional assumption is as follows:

Assumption 3 (Lipschitz continuity of $\tilde{c}_u^\pi(s,a)$ over π)**.** *For any given fixed $u \in (0,1)$ and any policies π and π', there exists a coefficient C_u such that*

$$\left|\tilde{c}_u^{\pi'}(s,a) - \tilde{c}_u^{\pi}(s,a)\right| \leq C_u \cdot \max_{s'} KL\left(\pi'(\cdot|s') \,\middle\|\, \pi(\cdot|s')\right), \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{15}$$

Basically, Assumption 3 is that the function $\tilde{c}_u^\pi$ as a function of π is continuous, which is expected to be satisfied if there is no abrupt change in the associated distributions. With Assumption 3 and Theorem 2, we obtain an expression for the quantile $q_u^\pi(s_0)$ as a form of desired expected sum:

Theorem 3. *Under deterministic dynamics $s_{t+1} = h(s_t, a_t)$ and Assumptions 1, 2, and 3, the u-quantile $q_u^\pi(s_0)$ is expressed as the expectation of the sum of actual cost and a π-independent additional cost $\tilde{c}_u^{\pi'}(s,a)$ for π' satisfying $\max_s KL(\pi'(\cdot|s) \,\|\, \pi(\cdot|s)) \leq \delta$:*

$$\left|q_u^\pi(s_0) - \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \left\{c(s_t,a_t) + \tilde{c}_u^{\pi'}(s_t,a_t)\right\}\right]\right| \leq \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma}\delta. \tag{16}$$

By Theorem 3, we can approximate the u-quantile $q_u^\pi(s_0)$ as the expectation of the sum of costs plus π-independent additional costs $\tilde{c}_u^{\pi'}(s,a)$ for a base policy π', and this approximation is tighter when the distance between the current policy π and the base policy π' is smaller. Theorem 3 can be interpreted the other way around. As in the case of PPO [20], if we first simply set π' as the policy before the update, denoted as π_{old}, then the updated π is near from the base policy $\pi_{old} = \pi'$, and we can compute the corresponding KL distance between π_{old} (= π') and π. Then, still, the inequality (16) holds for $\delta = \max_s KL\left(\pi_{old}(\cdot|s)||\pi(\cdot|s)\right)$. Now, this result enables us to solve the quantile constrained problem (QuantCP) by applying the policy gradient theorem.

4 Quantile Constrained Policy Optimization

Using the theoretical results in Section 3, we now construct an algorithm named quantile constrained policy optimization (QCPO) to solve (QuantCP) based on an on-policy RL algorithm: PPO [20]. The QCPO is a direct method to constrain the outage probability and consists of three parts: 1) estimation of the u-quantile $q_u^\pi(s)$ of $X^\pi(s)$ for a given policy π, 2) estimation of PDF of $X^\pi(s)$ to compute the additional cost $\tilde{c}_u^{\pi'}(s,a)$, and 3) updating method of the Lagrange multiplier to control the outage probability. We first explain the overall structure of QCPO and the base loss function for the policy, which has a similar form to that in [19]. Then, we provide a condition for policy improvement for the proposed method. The implementation of the proposed algorithm is based on the implementation of [21], and details of the implementation, including the network structure, the loss functions, the Lagrangian multiplier update method, and the hyper-parameters, are in Appendix E. The implementation code of QCPO is available at github.com/wyjung0625/QCPO.

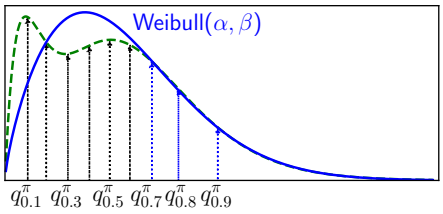


Figure 3: Illustration of the quantile approximation and tail-probability approximation. Green-dash curve is the unknown PDF $p_{X^\pi(s)}(x)$ and up-arrow points are the estimated u-quantiles $q_u^\pi(s)$. We approximate the PDF $p_{X^\pi(s)}(x)$ on the right tail by a Weibull distribution (blue curve) and this is approximated using the 4 rightmost quantile points (blue arrow).

4.1 Overall Structure of QCPO

The agent of QCPO uses function approximators for the policy π, the value function $V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)\right]$ and the quantile function $q^\pi(s) = \left[q_{u_1}^\pi(s), q_{u_2}^\pi(s), \ldots, q_{u_{n_q}}^\pi(s)\right]$ of $X^\pi(s) = \sum_{t=0}^{\infty} \gamma^t c(s_t,a_t)$ for the policy π. These functions are parameterized by deep neural networks with parameters θ, ϕ, and ψ, respectively. We denote θ_{old}, ϕ_{old}, and ψ_{old} as their old parameters. Note that the quantile function $q_\psi(s)$ outputs n_q values, and the i-th value represents an estimate of the u_i-quantile $q_{u_i}^\pi(s)$ of $X^\pi(s)$ for fixed target CDF values $\left[u_1, u_2, \ldots, u_{n_q}\right]$ with $u_1 < u_2 < \cdots < u_{n_q}$.

In addition to these parameterized functions, we need another function that approximates the PDF of $X^\pi(s)$ on the right tail. (Here, the right-tail probability most matters since the target outage probability is typically small.)

As aforementioned in Section 2, we know that $X^\pi(s)$ follows LDP with a rate function $I_{X^\pi(s)}(x)$. However, finding the rate function $I_{X^\pi(s)}(x)$ in an analytic approach is hard. Therefore, in QCPO, the agent approximates the rate function of the form of $I_{X^\pi(s)}(x) \approx (x/\beta(s))^{\alpha(s)}$ and learns the state-dependent parameters $\alpha(s)$ and $\beta(s)$ by using the quantiles on the right tail approximated by its quantile function for the right tail: $q^\pi_{u_{n_q-k+1}}(s), \ldots, q^\pi_{u_{n_q}}(s)$. This approximation of the rate function results in approximation probability on the right tail as a Weibull distribution, whose tail distribution is $1 - F_{X^\pi(s)}(x) = e^{-(x/\beta(s))^{\alpha(s)}}$. In order to obtain the state-dependent parameters $\alpha(s)$ and $\beta(s)$ for right-tail distribution approximation, we again parameterize them by neural networks with parameters ξ and ζ, respectively. Fig. 3 shows both the quantile approximation and right-tail approximation of QCPO. Note that our approach actually learns the rate function governing the tail-probability decay rate, whereas the previous Gaussian approximation [27] on the PDF of $X^\pi(s)$ fixes the rate function as quadratic x^2, which is not the correct rate function in general.

The overall procedure of QCPO is as follows: 1) estimate the value function $V^\pi(s)$ for return and estimate the quantile function $q^\pi_u(s)$, $u \in \left[u_1, u_2, \ldots, u_{n_q}\right]$ for the cumulative sum cost, 2) approximate tail distribution $p_{X^\pi(s)}(x)$ on the right tail using a Weibull distribution with parameters $\alpha(s)$, $\beta(s)$, 3) compute the additional cost for the base policy π', $\tilde{c}^{\pi'}_{1-\epsilon_0}(s,a)$ for the quantile advantage $A^\pi_{1-\epsilon_0}(s,a) := c(s,a) + \tilde{c}^{\pi'}_{1-\epsilon_0}(s,a) + \gamma q^\pi_{1-\epsilon_0}(s') - q^\pi_{1-\epsilon_0}(s)$, 4) take policy gradient using the sum of the value advantage and the quantile advantage $A_r(s,a) - \lambda A^\pi_{1-\epsilon_0}(s,a)$, 5) update the Lagrange multiplier λ. Since QCPO is based on PPO [20], the loss functions for the policy and the value function are similar to those of PPO [20]. Please see Appendix E.2 and E.3 for detail.

4.2 Policy Loss Function and Policy Improvement Condition

Let us consider the policy loss function of QCPO to solve (QuantCP). The basic loss function of QCPO for a given Lagrange multiplier λ is given by

$$L^{\pi_{old}}(\pi_\theta) - \tilde{C}_1 \max_s \mathrm{KL}(\pi_{old}(\cdot|s) \parallel \pi_\theta(\cdot|s)) \tag{17}$$

where

$$L^{\pi_{old}}(\pi_\theta) = \left(V^{\pi_{old}}(s_0) - \lambda q^{\pi_{old}}_{1-\epsilon_0}(s_0)\right) + \mathbb{E}_{s\sim\rho^{\pi_{old}}, a\sim\pi_\theta}\left[A^{\pi_{old}}_r(s_t,a) - \lambda A^{\pi_{old}}_{1-\epsilon_0}(s_t,a)\right] \tag{18}$$

$$A^{\pi_{old}}_r(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim M(\cdot|s,a)}\left[V^{\pi_{old}}(s')\right] - V^{\pi_{old}}(s) \tag{19}$$

$$A^{\pi_{old}}_{1-\epsilon_0}(s,a) = c(s,a) + \tilde{c}^{\pi_{old}}_{1-\epsilon_0}(s,a) + \gamma \mathbb{E}_{s'\sim M(\cdot|s,a)}\left[q^{\pi_{old}}_{1-\epsilon_0}(s')\right] - q^{\pi_{old}}_{1-\epsilon_0}(s), \tag{20}$$

and $\pi_{old} := \pi_{\theta_{old}}$ is the policy that collects the most recent batch of samples, $\rho^{\pi_{old}}(s) := \sum_{t=0}^\infty \gamma^t \Pr(S_t = s|s_0, \pi_{old})$ is the stationary state distribution under π_{old}, $\tilde{C}_1$ is a constant (see Appendix B.5), Now, we consider the relationship between the actually-desired maximization objective $L_{quant}(\pi,\lambda) = V^\pi(s_0) - \lambda\left(q^\pi_{1-\epsilon_0}(s_0) - d_{th}\right)$ in (3) and the practical QCPO objective $L^{\pi_{old}}(\theta) - \tilde{C}_1 \max_s \mathrm{KL}(\pi_{old}(\cdot|s) \parallel \pi_\theta(\cdot|s))$ in (17). The relationship between the two is given by the following theorem.

Theorem 4. *Let $\pi_{new} := \pi_{\theta_{new}}$ be the solution of the problem of maximizing*

$$L^{\pi_{old}}(\pi_\theta) - \tilde{C}_1 \max_s KL(\pi_{old}(\cdot|s) \parallel \pi_\theta(\cdot|s)) \tag{21}$$

for some constant $\tilde{C}_1 > 0$. Then, under deterministic dynamics $s_{t+1} = h(s_t, a_t)$ and Assumptions 1, 2, and 3, the following inequality holds:

$$L_{quant}(\pi_{new},\lambda) - L_{quant}(\pi_{old},\lambda) \tag{22}$$

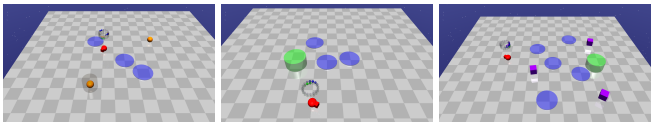
$$\geq L^{\pi_{old}}(\pi_{new}) - L^{\pi_{old}}(\pi_{old}) - \tilde{C}_1 KL_{max}(\pi_{old}||\pi_{new}) - \underbrace{\tilde{C}_2 \frac{\epsilon}{1-\epsilon}}_{\textit{approximation loss}} \tag{23}$$

for a given Lagrange multiplier $\lambda > 0$, some constant $\tilde{C}_2$ and small $\epsilon > 0$.

Note that the term $\tilde{C}_2 \frac{\epsilon}{1-\epsilon}$ in (23) is due to our approximation of the quantile as an expected sum to apply policy gradient. Therefore, by Theorem 4, when the improvement $L^{\pi_{old}}(\pi_{new}) - L^{\pi_{old}}(\pi_{old}) - \tilde{C}_1 \mathrm{KL}_{max}(\pi_{old}||\pi_{new})$ (> 0) by the policy update from the QCPO loss function is large enough to compensate for the approximation loss, the desired quantity will also be improved by our policy update. That is, the Lagrangian for the quantile constrained problem for π_{new} will be higher than that for π_{old}.

5 Experiments

5.1 Environments



(a) SimpleButtonEnv (b) DynamicEnv (c) GremlinEnv

Figure 4: The considered environments

We examined the performance of the proposed QCPO and compared it to that of WCSAC, which uses the CVaR constraint. The environments we considered are SimpleButtonEnv, DynamicEnv [27], and GremlinEnv, which are based on Safety Gym [17], MuJoCo [23], and OpenAI Gym [4]. The environments can be considered as simplified versions of a real environment of an automatic serving robot and are illustrated in Fig. 4. The goal of these environments is for a robot (red sphere) to reach a goal (orange sphere wrapped by a grey translucent pillar, or green pillar) while avoiding the non-goal button (orange sphere), hazards (blue circle) or moving gremlins (purple box). Once the robot reaches the current goal, the environments generate the next goal deterministically (SimpleButtonEnv) or randomly (DynamicEnv, GremlinEnv), so the task complexity increases in the order of SimpleButtonEnv, DynamicEnv, and GremlinEnv. When the robot performs an action at time t, it receives a reward $\left\{\|p_{t+1} - p_{\text{goal}}\|_2 - \|p_t - p_{\text{goal}}\|_2\right\} + 1_{\text{goal reached}}$, where p_t is the position (x, y) of the robot at time t and p_{goal} is the current goal position at time t. It also receives a cost $+1$ if the robot touches a non-goal object (the non-goal button, a hazard, or a gremlin) and 0 otherwise. Thus, for the robot, it receives a higher return when the robot touches more goals in maximum timesteps $T = 1000$, and a higher sum of costs when the robot touches one of the other objects more often. A more detailed explanation of the environments is in Appendix C.

5.2 Empirical Results

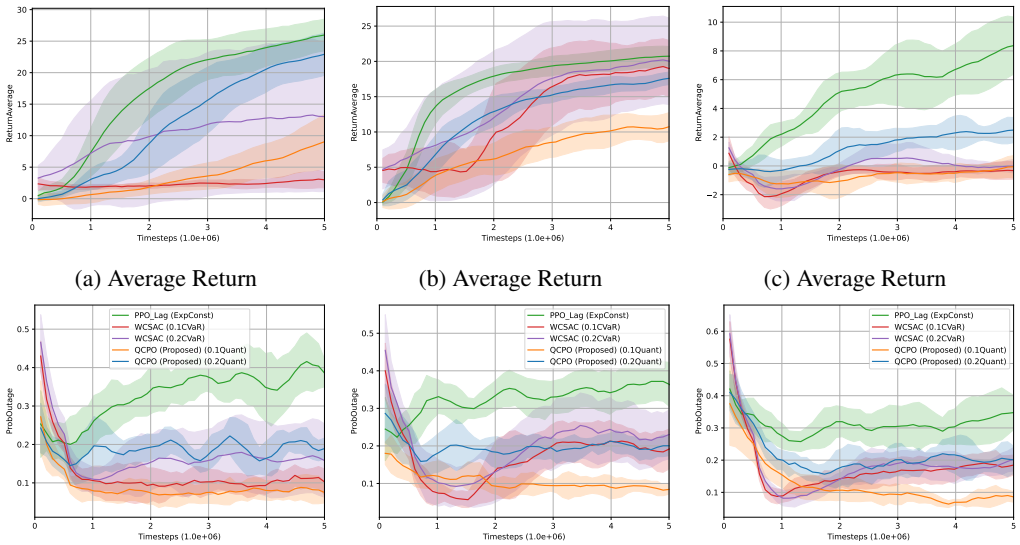
We compared the performance of the proposed algorithm (QCPO) with that of PPO with the Lagrangian multiplier method (PPO_Lag)[3] for (ExpCP) and that of WCSAC [27][4] for (CVaR-CP) which is a stricter problem than (ProbCP). We set the threshold $d_{th} = 15$ in (ExpCP), (CVaR-CP), and (QuantCP) and the target outage probability $\epsilon_0 = 0.1, 0.2$ in (CVaR-CP) and (QuantCP).

Fig. 5 shows the results of the considered algorithms on SimpleButtonEnv, DynamicEnv, and GremlinEnv. All experiments were done with 10 different random seeds, and the real line and the shaded area represent the average and average $\pm$ standard deviation, respectively. PPO with the Lagrangian multiplier method for (ExpCP) (green) keeps the average of the sum cost around the threshold $d_{th} = 15$ well (please see the graph in Appendix D.3), and its outage probability becomes around 0.35 as we can observe in Fig. 5d, 5e, and 5f. As aforementioned, the CVaR approach (WCSAC) should satisfy a sufficient condition for satisfying the outage probability constraint in (ProbCP). It is seen that WCSAC ($\epsilon_0 = 0.2$ (purple), $\epsilon_0 = 0.1$ (red)) achieves a lower or similar outage probability to the threshold ϵ_0 in Fig. 5d, but the algorithm does not satisfy the outage probability constraint exactly in Fig. 5e and 5f. This means that the Gaussian distribution approximation of the distribution of $X^\pi(s)$ has limited capability to capture the decay rate of the tail probability. On the other hand, the proposed QCPO ($\epsilon_0 = 0.2$ (blue), $\epsilon_0 = 0.1$ (orange)) maintains the outage probability around the desired target outage probability very well, as shown in Fig. 5d, 5e, and 5f.

Now consider the average return of these algorithms. In constrained RL, in general, if an algorithm is allowed to have a higher sum of costs, then it has a higher return. Thus, as seen in Fig. 5d, 5e, and 5f, PPO_Lag induces the highest outage probability, so it has the highest average return, as shown in Fig. 5a, 5b, and 5c. The direct comparison between WCSAC and QCPO is less meaningful in

[3]We used the implementation code in `https://github.com/astooke/rlpyt/tree/master/rlpyt/projects/safe`, (MIT License)

[4]We used the github code that the authors of the paper uploaded: `https://github.com/AlgTUDelft/WCSAC`, (MIT License)



(a) Average Return (b) Average Return (c) Average Return

(d) Outage Probability (e) Outage Probability (f) Outage Probability

Figure 5: (left) SimpleButtonEnv, (middle) DynamicEnv, and (right) GremlinEnv: (upper row) average return and (lower row) outage probability of the most current 100 episodes.

DynamicEnv and GremlinEnv, since WCSAC does not satisfy the outage probability constraint, but it is fair in SimpleButtonEnv because both algorithms satisfy the outage probability constraint. As seen in Fig. 5a, QCPO achieves a higher average return than WCSAC for the same target probability constraint $\epsilon_0 = 0.1, 0.2$. This is because QCPO satisfies the target outage probability exactly, i.e., uses the given cost budget fully for a higher return. We provided more results in Appendix D.

6 Conclusion

We have proposed the framework of quantile-constrained RL to constrain the outage probability by adopting a constraint on the quantile, which is equivalent to the outage probability constraint. We have investigated issues in applying the policy gradient theorem to the Lagrangian of the quantile-constrained RL problem and have converted the quantile into an additive form of costs so that the application of the policy gradient theorem is feasible. Based on our derivation, we have constructed the QCPO algorithm, which uses distributional RL techniques to learn the u-quantile of the cumulative sum cost $X^\pi(s)$, and Weibull distribution to approximate the tail distribution of $X^\pi(s)$. We also proved the policy improvement condition for QCPO and showed that there exists an approximation loss due to our approximation of the quantile. Empirical results show that QCPO constrains the outage probability well as the desired target value. The meaning of such exact satisfaction of the outage probability is two-fold: First, the constraint on the outage probability is satisfied to control the probability of unsafe events, and second, the exact satisfaction of the cost constraint enables us to exploit the cost budget fully and obtain a higher return. Empirical results demonstrated the effectiveness of the proposed scheme.

Acknowledgments and Disclosure of Funding

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00469, Development of Core Technologies for Task-oriented Reinforcement Learning for Commercialization of Autonomous Drones, 50%) and by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities, 50%)

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.

[2] Gabriel Barth-Maron, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. In *International Conference on Learning Representations*, Apr 2018. URL `https://openreview.net/forum?id=SyZipzbCb`.

[3] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[4] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[5] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[6] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, Dec 2018.

[7] Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019.

[8] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning*, volume 80, pages 1096–1105. PMLR, Jul 2018.

[9] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

[11] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 1998.

[12] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

[13] Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, volume 119, pages 5556–5566. PMLR, 2020.

[14] Yongshuai Liu, Jiaxin Ding, and Xin Liu. IPO: Interior-point policy optimization under constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4940–4947, 2020.

[15] Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[16] Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In *International Conference on Machine Learning*, pages 4424–4434. PMLR, 2019.

[17] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. 2019.

[18] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

[19] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, volume 37, pages 1889–1897. PMLR, 2015.

[20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[21] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *International Conference on Machine Learning*, volume 119, pages 9133–9143. PMLR, 2020.

[22] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[23] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[24] Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, and Alekh Agarwal. Safe reinforcement learning via curriculum induction. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

[26] Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. In *Advances in neural information processing systems*, volume 32, pages 6193–6202, 2019.

[27] Qisong Yang, Thiago D. Simão, Simon H Tindemans, and Matthijs T. J. Spaan. WCSAC: Worst-case soft actor critic for safety-constrained reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10639–10646, 2021.

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

[29] Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. Convergent policy optimization for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 3127–3139, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We mentioned theoretical bound and approximation loss in Section 3 and 4.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] I have read the guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Please see Appendix B.
   (b) Did you include complete proofs of all theoretical results? [Yes] Please see Appendix B.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The implementation code and instructions are uploaded in github. Please see github address in page 7.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see Appendix E.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Experimental results show mean and standard deviation with 10 random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see Appendix C.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Please see footnotes in page 9.
   (b) Did you mention the license of the assets? [Yes] Please see footnotes in page 9.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Please see github address in page 7.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A More Backgrounds

A.1 Distributional RL

Distributional RL [2, 3, 8] is an area of RL that considers the distribution of the cumulative return $Z^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)$ for $S_0 = s$, $A_0 = a$, $S_{t+1} \sim M(\cdot|S_t, A_t)$, $A_{t+1} \sim \pi(\cdot|S_t)$, $t = 0, 1, \cdots$, instead of the expectation of the cumulative return $Q^\pi(s,a) = \mathbb{E}_\pi[Z^\pi(s,a)] = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$ to optimize a policy π. In distributional RL, the distribution of the cumulative return $Z^\pi(s,a)$ is computed by the distributional Bellman equation [3], defined as

$$Z^\pi(s,a) \overset{D}{=} r(s,a) + Z^\pi(S', A') \tag{24}$$

for $S' \sim M(\cdot|s,a), A' \sim \pi(\cdot|S')$, where $\overset{D}{=}$ means that the random variable in the left-hand side (LHS) has the same distribution to that in the right-hand side (RHS). So, the following holds[15]:

$$\begin{aligned} F_{Z^\pi(s,a)}(z) &= \mathbb{E}_{s'\sim M, a'\sim\pi}\left[F_{r(s,a)+\gamma Z^\pi(s',a')}(z)\right] \\ &= \mathbb{E}_{s'\sim M, a'\sim\pi}\left[F_{Z^\pi(s',a')}\left(\frac{z - r(s,a)}{\gamma}\right)\right] \end{aligned} \tag{25}$$

$$p_{Z^\pi(s,a)}(z) = \frac{1}{\gamma}\mathbb{E}_{s'\sim M, a'\sim\pi}\left[p_{Z^\pi(s',a')}\left(\frac{z - r(s,a)}{\gamma}\right)\right], \tag{26}$$

where $F_X(x)$ and $P_X(x)$ denote the cumulative distribution function (CDF) and PDF of a random variable X, respectively, and (26) is obtained by taking derivative of (25). To train the distribution of the cumulative return $Z^\pi(s,a)$, the p-Wasserstein distance $W_p(X,Y)$ is typically used, which can be written explicitly as

$$W_p(X,Y) = \left(\int_0^1 \left|F_X^{-1}(u) - F_Y^{-1}(u)\right|^p\right)^{1/p} \tag{27}$$

for $p < \infty$, where $F_X^{-1}(u) = \inf\{x \mid F_X(x) \geq u\} =: Q_X(u)$ is the quantile function (inverse CDF) of the random variable X. Dabney et al. [8, 9], Mavrin et al. [16], Kuznetsov et al. [13], Yang et al. [26] used quantile regression to learn the quantile of the cumulative return $Z^\pi(s,a)$. The quantile regression loss is given by $L_{quant,u}(q) = \mathbb{E}_X\left[l_{quant,u}(X - q)\right]$, where

$$l_{quant,u}(x) = \left(u - 1_{\{x<0\}}\right) \cdot x. \tag{28}$$

To smooth the gradient, they used the quantile Huber loss function $L_{Huber,u}(q) = \mathbb{E}_X\left[l_{Huber,u}(X - q)\right]$ for a given $\kappa > 0$, where

$$\begin{aligned} l_{Huber,u}(x) &= \left|u - 1_{\{x<0\}}\right| \frac{L_\kappa(x)}{\kappa}, \\ L_\kappa(x) &= \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \leq \kappa \\ \kappa\left(|x| - \frac{1}{2}\kappa\right), & \text{otherwise.} \end{cases} \end{aligned} \tag{29}$$

In this paper, we estimate the quantiles of the cumulative sum cost using the quantile loss, and use them to solve the constrained optimization problem (QuantCP).

$$\begin{array}{ll} \text{Maximize} & \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\ \text{Subject to} & q_{1-\epsilon_0}^\pi(s_0) \leq d_{th}, \end{array} \tag{QuantCP}$$

A.2 Large Deviation Principle (LDP)

Large deviation principle (LDP) [11] is a technique for estimating the limiting behavior of a sequence of distributions. A simple example is the empirical mean $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ of i.i.d. random variables X_i. We say that a sequence $\{\bar{X}_n\}$ satisfies LDP if the sequence of its log probability distribution $\frac{1}{n}\log \Pr\left(\bar{X}_n \in \Gamma\right)$ satisfies the following condition $\frac{1}{n}\log \Pr\left(\bar{X}_n \in \Gamma\right) \overset{n\to\infty}{\longrightarrow} -\inf_{x\in\Gamma} I(x)$ for some function $I(x)$. The function $I(x)$ satisfying such limiting behavior is called the rate function of $\bar{X}_n$. The rate function $I(x)$ is also related to the cumulative distribution function $F_{\bar{X}_n}(x)$ since $1 - F_{\bar{X}_n}(x_0) = \Pr\left(\bar{X}_n \in [x_0, \infty)\right) \approx \exp\left(-n \inf_{x\in[x_0,\infty)} I(x)\right)$ for some $x_0 > \mathbb{E}[X]$ and sufficiently large n.

LDP can be applied to finite state Markov chains [11]. Let $Y_k \in \mathcal{Y} = \{y^1, \dots y^m\}$ be random variables that follows the Markov property: $\Pr(Y_1 = y_1, \dots, Y_n = y_n) = p_0(y_1) \prod_{i=1}^{n} M(y_{i+1}|y_i)$. Then, the sequence of empirical means $Z_n := \frac{1}{n}\sum_{k=0}^{n} X_k$, where $X_k = f(Y_k)$ for some function $f : \mathcal{Y} \to \mathbb{R}^d$, satisfies LDP and the rate function is given by $I(z) = \sup_{\lambda \in \mathbb{R}^d} \{\langle \lambda, z\rangle - \log \rho(\Pi_\lambda)\}$, where $\rho(\Pi)$ is the Perron-Frobenius eigenvalue of a given matrix Π, and Π_λ is the matrix whose (i, j)-th element is $M(y^j|y^i) \exp \langle \lambda, f(y^j)\rangle$.

In this paper, we consider the tail probability of the distribution of the cumulative sum cost $X^\pi(s_0) = \sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)$. Finding its analytic rate function is hard. Therefore, we instead approximate the rate function directly as $I_{X^\pi(s)}(x) \approx (x/\beta(s))^{\alpha(s)}$ with learnable parameters $\alpha(s)$ and $\beta(s)$, which results in a Weibull distribution: $1 - F_{X^\pi(s)}(x) = \exp\left\{-(x/\beta(s))^{\alpha(s)}\right\}$. We use this distribution to approximate the tail probability of $p_{X^\pi(s)}(x)$ of $X^\pi(s)$.

A.3 The Considered Constrained Problems

In this subsection, we list the problems for constrained RL. The first constrained problem is a common problem used in many previous constrained RL papers.

$$\begin{array}{ll} \text{Maximize} & V^\pi(s_0) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\ \text{subject to} & C^\pi(s_0) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] \le d_{th}, \end{array} \tag{ExpCP}$$

In (ExpCP), the cost constraint is that the expectation of the sum of costs is less than or equal to a threshold parameter d_{th}. Note that the threshold d_{th} is set on the average (i.e., expectation) of the cumulative sum cost to avoid undesired high-cost events in this formulation. However, solving the problem (ExpCP) may have undesirable outcomes for real environments that typically need constrained behavior on the event that the cost exceeds the threshold d_{th}.

There are two well-known techniques, called Value at Risk (VaR, or Quantile) and Conditional Value at Risk (CVaR), to manage undesirable events in the domain of finance[18]. In the context of RL, the definitions of the quantile and the CVaR for the distribution of the cumulative sum cost for a given π are given by $q_u^\pi(s_0) := \inf\{x \mid \Pr(X^\pi(s_0) \le x) \ge u\}$ and $\text{CVaR}_u^\pi(s_0) := \mathbb{E}_\pi\left[X^\pi(s_0) \mid X^\pi(s_0) \ge q_u^\pi(s_0)\right]$, respectively. Note that the CVaR and the quantile are two different measures for undesirable events, and the choice between the two depends on what we desire. For example, an insurance company prefers the CVaR of undesirable events to determine an insurance premium. On the other hand, a company developing an autonomous driving car system needs the quantile of undesirable events to guarantee the accident probability for safety.

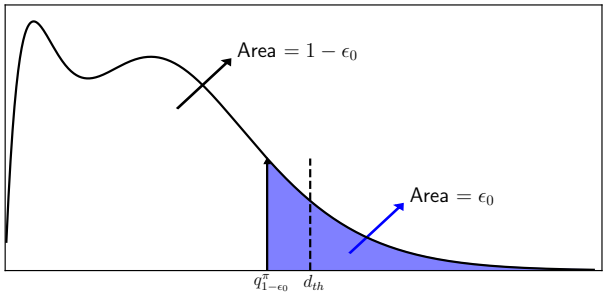


Figure 6: Equivalence between the outage probability constraint and the quantile constraint

The CVaR constrained problem to constrain undesirable events was previously used in RL [5, 27], and the problem is explicitly formulated as

$$\begin{array}{ll} \text{Maximize} & \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\ \text{Subject to} & \text{CVaR}_{1-\epsilon_0}^\pi(s_0) \le d_{th}, \end{array} \tag{CVaR-CP}$$

In this paper, we focus on constraining the probability of undesirable events that the cost exceeds the threshold d_{th}. Thus we can consider a constrained problem with a probabilistic constraint as follows:

$$\begin{array}{ll} \text{Maximize} & V^\pi(s_0) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\ \text{Subject to} & \Pr\left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t) > d_{th}\right] \le \epsilon_0 \\ & \quad \text{for } S_0 = s_0, A_t \sim \pi(\cdot|S_t), S_{t+1} \sim M(\cdot|S_t, A_t). \end{array} \tag{ProbCP}$$

Our approach to this problem is first to convert the outage probability constraint in (ProbCP) into a quantile constraint $q^{\pi}_{1-\epsilon_0}(s_0) \leq d_{th}$, which is equivalent to the original probabilistic constraint (See Fig. 6), and then to solve the equivalent optimization:

$$\begin{array}{ll} \text{Maximize} & \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\ \text{Subject to} & q^{\pi}_{1-\epsilon_0}(s_0) \leq d_{th}, \end{array} \tag{QuantCP}$$

Note that the $(1-\epsilon_0)$-quantile denoted as $q^{\pi}_{1-\epsilon_0}(s)$ is always less or equal to than the $(1-\epsilon_0)$-CVaR denoted as $\text{CVaR}^{\pi}_{1-\epsilon_0}(s)$ for all $s \in \mathcal{S}$ because of the definition of the CVaR. Therefore, satisfying the CVaR constraint is a sufficient condition for satisfying the probabilistic constraint, and hence this problem is a stricter problem than (ProbCP) or (QuantCP). Therefore, the algorithms proposed to solve (CVaR-CP) can be used for solving (ProbCP), and this should satisfy the probabilistic constraint in theory.

B Proofs

In the following proofs, we used text color so that readers can follow the proof easily.

B.1 Proof of Theorem 1

Assumption 1 (Boundness of quantile difference)**.** *For a given policy π, the following two quantities are bounded*

$$|c(s,a) + \gamma q_u^\pi(s') - q_u^\pi(s)| \leq \gamma R \tag{30}$$

$$\left| q_u^\pi(s) - F^{-1}_{X^\pi(s)}\left(F_{X^\pi(s')}\left(\frac{q_u^\pi(s) - c(s,a)}{\gamma}\right)\right)\right| \leq R \tag{31}$$

for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ such that $\pi(a|s) \cdot M(s'|s,a) > 0$.

Note that for finite MDPs, which are assumed for many RL proofs, this assumption definitely holds with a finite cost function.

Assumption 2 (Smoothness of CDF of $X^\pi(s)$)**.** *For each state s, the average slope of $F_{X^\pi(s)}(x)$ between $q_u^\pi(s)$ and $y \in [q_u^\pi(s) - R, q_u^\pi(s) + R]$ is bounded by*

$$\frac{1}{1+\epsilon} \cdot p_{X^\pi(s)}\left(q_u^\pi(s)\right) \leq \frac{F_{X^\pi(s)}\left(q_u^\pi(s)\right) - F_{X^\pi(s)}\left(y\right)}{q_u^\pi(s) - y} \leq \frac{1}{1-\epsilon} \cdot p_{X^\pi(s)}\left(q_u^\pi(s)\right) \tag{32}$$

for small $0 < \epsilon < \frac{1}{2}$.

This assumption holds when discrete masses are not present in the PDF and the CDF is continuous.

Theorem 1. *Under Assumptions 1 and 2, the u-quantile of the random variable $X^\pi(s_t)$ satisfies the following temporal-difference(TD) relation. For some constant R and small $\epsilon > 0$,*

$$\left| \mathbb{E}_\pi \left[\frac{p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma}\right)}{\gamma p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)} \left\{c(s_t,a_t) + \gamma q_u^\pi(s_{t+1}) - q_u^\pi(s_t)\right\} \right] \right| \leq \frac{\epsilon}{1-\epsilon} R, \tag{33}$$

Here, the expectation is for the action $a_t \sim \pi(\cdot|s_t)$ and the next state $s_{t+1} \sim M(\cdot|s_t,a_t)$. ($s_t$ is given.)

Proof. Note that from (25),

$$F_{X^\pi(s)}\left(x\right) = \mathbb{E}_\pi\left[F_{X^\pi(s')}\left(\frac{x - c(s,a)}{\gamma}\right)\right], \tag{34}$$

for all x. If $x = q_u^\pi(s)$, then this becomes

$$u = F_{X^\pi(s)}\left(q_u^\pi(s)\right) = \mathbb{E}_\pi\left[F_{X^\pi(s')}\left(\frac{q_u^\pi(s) - c(s,a)}{\gamma}\right)\right]. \tag{35}$$

Using (35), we can obtain

$$\mathbb{E}_\pi \left[\frac{p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma}\right)}{\gamma p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)} \left\{c(s_t,a_t) + \gamma q_u^\pi(s_{t+1}) - q_u^\pi(s_t)\right\} \right] \tag{36}$$

$$= \frac{\mathbb{E}_\pi\left[p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma}\right)\left\{c(s_t,a_t) + \gamma q_u^\pi(s_{t+1}) - q_u^\pi(s_t)\right\}\right]}{\gamma p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)} \tag{37}$$

$$= \frac{1}{\gamma p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)} \times \mathbb{E}_\pi\Bigg[\gamma \underbrace{\left(F_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma}\right) - u\right)}_{=0 \text{ by (35)}}$$

$$+ p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma}\right)\left\{c(s_t,a_t) + \gamma q_u^\pi(s_{t+1}) - q_u^\pi(s_t)\right\}\Bigg] \tag{38}$$

$$= \mathbb{E}_\pi \left[\frac{\left(F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) - u \right) + p_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \left\{ q_u^\pi(s_{t+1}) - \frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right\}}{p_{X^\pi(s_t)} \left(q_u^\pi(s_t) \right)} \right] \tag{39}$$

$$= \mathbb{E}_\pi \left[\frac{\left(u - F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \right)}{p_{X^\pi(s_t)} \left(q_u^\pi(s_t) \right)} \times \frac{\left(F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) - u \right) + p_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \left\{ q_u^\pi(s_{t+1}) - \frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right\}}{\left(u - F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \right)} \right] \tag{40}$$

$$= \mathbb{E}_\pi \left[\frac{\left(u - F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \right)}{p_{X^\pi(s_t)} \left(q_u^\pi(s_t) \right)} \times \left\{ \frac{p_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \left\{ q_u^\pi(s_{t+1}) - \frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right\}}{\left(u - F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \right)} - 1 \right\} \right] \tag{41}$$

Then, by Cauchy-Schwarz inequality, we can obtain a bound such that

$$\mathbb{E}_\pi \left[\frac{p_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right)}{\gamma p_{X^\pi(s_t)} \left(q_u^\pi(s_t) \right)} \left\{ c(s_t,a_t) + \gamma q_u^\pi(s_{t+1}) - q_u^\pi(s_t) \right\} \right]^2 \tag{42}$$

$$= \mathbb{E}_\pi \left[\frac{\left(u - F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \right)}{p_{X^\pi(s_t)} \left(q_u^\pi(s_t) \right)} \times \left\{ \frac{p_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \left\{ q_u^\pi(s_{t+1}) - \frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right\}}{\left(u - F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \right)} - 1 \right\} \right]^2 \tag{43}$$

$$\leq \underbrace{\mathbb{E}_\pi \left[\left(\frac{u - F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right)}{p_{X^\pi(s_t)} \left(q_u^\pi(s_t) \right)} \right)^2 \right]}_{(a)} \times \underbrace{\mathbb{E}_\pi \left[\left(1 - \frac{p_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \cdot \left\{ q_u^\pi(s_{t+1}) - \frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right\}}{u - F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right)} \right)^2 \right]}_{(b)} \tag{44}$$

Now we find upper bounds of (a) and (b).

- First, consider an upper bound of (a).

$$\frac{u - F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right)}{p_{X^\pi(s_t)} \left(q_u^\pi(s_t) \right)} \tag{45}$$

$$= \frac{u - F_{X^\pi(s_t)} \left(F^{-1}_{X^\pi(s_t)} \left(F_{X^\pi(s_{t+1})} \left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma} \right) \right) \right)}{p_{X^\pi(s_t)} \left(q_u^\pi(s_t) \right)} \tag{46}$$

$$= \frac{u - F_{X^\pi(s_t)} \left(\bar{q}_u^\pi(s_t, a_t, s_{t+1}) \right)}{p_{X^\pi(s_t)} \left(q_u^\pi(s_t) \right)} \tag{47}$$

$$= \frac{F_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right) - F_{X^\pi(s_t)}\left(\bar{q}_u^\pi(s_t,a_t,s_{t+1})\right)}{p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)} \tag{48}$$

$$= \frac{\frac{F_{X^\pi(s_t)}(q_u^\pi(s_t)) - F_{X^\pi(s_t)}(\bar{q}_u^\pi(s_t,a_t,s_{t+1}))}{q_u^\pi(s_t) - \bar{q}_u^\pi(s_t,a_t,s_{t+1})}}{p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)} \cdot \{q_u^\pi(s_t) - \bar{q}_u^\pi(s_t,a_t,s_{t+1})\} \tag{49}$$

where $\bar{q}_u^\pi(s_t,a_t,s_{t+1}) = F^{-1}_{X^\pi(s_t)}\left(F_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)\right)$. Note that

$$\frac{F_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right) - F_{X^\pi(s_t)}\left(\bar{q}_u^\pi(s_t,a_t,s_{t+1})\right)}{q_u^\pi(s_t) - \bar{q}_u^\pi(s_t,a_t,s_{t+1})}$$

is the average slope of $F_{X^\pi(s_t)}(x)$ between $q_u^\pi(s_t)$ and $\bar{q}_u^\pi(s_t,a_t,s_{t+1})$, and $p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)$ is the slope of $F_{X^\pi(s_t)}(x)$ at $x = q_u^\pi(s_t)$. Therefore by Assumption 1 and 2, we can obtain an upper bound of (a) as follows:

$$\mathbb{E}_\pi\left[\left(\frac{u - F_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)}{p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)}\right)^2\right] \tag{50}$$

$$= \mathbb{E}_\pi\left[\left(\frac{\frac{F_{X^\pi(s_t)}(q_u^\pi(s_t)) - F_{X^\pi(s_t)}(\bar{q}_u^\pi(s_t,a_t,s_{t+1}))}{q_u^\pi(s_t) - \bar{q}_u^\pi(s_t,a_t,s_{t+1})}}{p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)} \cdot \{q_u^\pi(s_t) - \bar{q}_u^\pi(s_t,a_t,s_{t+1})\}\right)^2\right] \tag{51}$$

$$= \mathbb{E}_\pi\left[\left(\frac{\frac{F_{X^\pi(s_t)}(q_u^\pi(s_t)) - F_{X^\pi(s_t)}(\bar{q}_u^\pi(s_t,a_t,s_{t+1}))}{q_u^\pi(s_t) - \bar{q}_u^\pi(s_t,a_t,s_{t+1})}}{p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)} \times \left\{q_u^\pi(s_t) - F^{-1}_{X^\pi(s_t)}\left(F_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t) - c(s_t,a_t)}{\gamma}\right)\right)\right\}\right)^2\right] \tag{52}$$

$$\leq \frac{1}{(1-\epsilon)^2}R^2 \tag{53}$$

- Next, we consider an upper bound of (b). By Assumption 1 and 2,

$$\mathbb{E}_\pi\left[\left(1 - \frac{p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right) \cdot \left\{q_u^\pi(s_{t+1}) - \frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right\}}{u - F_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)}\right)^2\right] \tag{54}$$

$$= \mathbb{E}_\pi\left[\left(1 - \frac{p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)}{\frac{u - F_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)}{\left\{q_u^\pi(s_{t+1}) - \frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right\}}}\right)^2\right] \tag{55}$$

$$\leq \epsilon^2 \tag{56}$$

Therefore by combining two upper bounds, we can conclude the theorem.

$$\left|\mathbb{E}_\pi\left[\frac{p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)}{\gamma p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)}\{c(s_t,a_t) + \gamma q_u^\pi(s_{t+1}) - q_u^\pi(s_t)\}\right]\right| \tag{57}$$

$$\leq \left(\underbrace{\mathbb{E}_\pi\left[\left(\frac{u - F_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)}{p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)}\right)^2\right]}_{(a)}\right)^{\frac{1}{2}} \tag{58}$$

$$\times \left(\underbrace{\mathbb{E}_\pi \left[\left(1 - \frac{p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right) \cdot \left\{ q_u^\pi(s_{t+1}) - \frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma} \right\}}{u - F_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)} \right)^2 \right]}_{(b)} \right)^{\frac{1}{2}} \tag{59}$$

$$\leq \frac{\epsilon}{1-\epsilon} R \tag{60}$$

□

Corollary 1. *Under Assumptions 1 and 2, the u-quantile $q_u^\pi(s_t)$ of the random variable $X^\pi(s_t)$ is bounded as*

$$\left| q_u^\pi(s_t) - \mathbb{E}_\pi \left[\mu_u^\pi\left(s_t, a_t, s_{t+1}\right) \left\{ c(s_t,a_t) + \gamma q_u^\pi(s_{t+1}) \right\} \right] \right| \leq \frac{\epsilon}{1-\epsilon} R. \tag{61}$$

where

$$\mu_u^\pi\left(s_t, a_t, s_{t+1}\right) := \frac{p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)}{\gamma p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)} \tag{62}$$

Proof. Note that the term $q_u^\pi(s_t)$ can go outside the expectation in (33) since the expectation is over (a_t, s_{t+1}). From eq. (26) in Appendix A.1, the expectation of the numerator of $\mu_u^\pi\left(s_t, a_t, s_{t+1}\right)$ is the same as the the denominator of the weight, i.e., $\mathbb{E}_\pi\left[p_{X^\pi(s_{t+1})}\left(\frac{q_u^\pi(s_t)-c(s_t,a_t)}{\gamma}\right)\right] = \gamma p_{X^\pi(s_t)}\left(q_u^\pi(s_t)\right)$ and this leads to $\mathbb{E}_\pi[\mu_u^\pi(s_t, a_t, s_{t+1})] = 1$. So, we have the claim. □

B.2 Proof of Lemma 1

Lemma 1. *Suppose that the state transition dynamics are deterministic, i.e., $s_{t+1} = h(s_t, a_t)$. Then, under Assumptions 1 and 2, the u-quantile $q_u^\pi(s_0)$ of the random variable $X^\pi(s_0)$ is expressed as*

$$\left| q_u^\pi(s_0) - \mathbb{E}_{\tilde{\pi}_u}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \right] \right| \leq \frac{\epsilon R}{(1-\epsilon)(1-\gamma)}, \tag{63}$$

where

$$\tilde{\pi}_u(a|s) = \pi(a|s) \cdot \frac{p_{X^\pi(h(s,a))}\left(\frac{q_u^\pi(s)-c(s,a)}{\gamma}\right)}{\gamma p_{X^\pi(s)}\left(q_u^\pi(s)\right)} \propto \pi(a|s) \cdot p_{X^\pi(h(s,a))}\left(\frac{q_u^\pi(s)-c(s,a)}{\gamma}\right). \tag{64}$$

Proof. Remind that $\mu_u^\pi\left(s, a, s'\right)$ is defined in (62) as

$$\mu_u^\pi\left(s, a, s'\right) := \frac{p_{X^\pi(s')}\left(\frac{q_u^\pi(s)-c(s,a)}{\gamma}\right)}{\gamma p_{X^\pi(s)}\left(q_u^\pi(s)\right)}$$

Consider $\mathbb{E}_\pi\left[\mu_u^\pi\left(s, a, s'\right)\left\{c(s,a) + \gamma q_u^\pi(s')\right\}\right]$.

$$\mathbb{E}_\pi\left[\mu_u^\pi\left(s, a, s'\right)\left\{c(s,a) + \gamma q_u^\pi(s')\right\}\right] \tag{65}$$

$$= \sum_a \sum_{s'} \pi(a|s) \cdot M(s'|s,a) \cdot \mu_u^\pi\left(s, a, s'\right)\left\{c(s,a) + \gamma q_u^\pi(s')\right\} \tag{66}$$

$$= \sum_a \sum_{s'} \tilde{\pi}_u(a|s) \cdot \tilde{M}_u(s'|s,a)\left\{c(s,a) + \gamma q_u^\pi(s')\right\} \tag{67}$$

for some distorted policy $\tilde{\pi}_u$ and some distorted state transition dynamics $\tilde{M}_u$. This is because

$$\mathbb{E}_\pi\left[\mu_u^\pi\left(s, a, s'\right)\right] = \sum_a \sum_{s'} \pi(a|s) \cdot M(s'|s,a) \cdot \frac{p_{X^\pi(s')}\left(\frac{q_u^\pi(s)-c(s,a)}{\gamma}\right)}{\gamma p_{X^\pi(s)}\left(q_u^\pi(s)\right)} = 1. \tag{68}$$

where the last equation holds from (26). Now, under the assumption that the state transition dynamics is deterministic $s' = h(s,a)$, i.e., $p(s'|s,a) = \delta_{h(s,a)}(s')$, the distorted transition dynamics are the same as the original transition dynamics and the only difference is the distorted policy:

$$\tilde{M}_u(s'|s,a) = \frac{\pi(a|s)M(s'|a,s)\mu_u^\pi(s,a,s')}{\sum_{\tilde{s}} \pi(a|s)M(\tilde{s}|a,s)\mu_u^\pi(s,a,\tilde{s})} \tag{69}$$

$$= \frac{\delta_{h(s,a)}(s')\pi(a|s)\mu_u^\pi(s,a,s')}{\sum_{\tilde{s}} \delta_{h(s,a)}(\tilde{s})\pi(a|s)\mu_u^\pi(s,a,\tilde{s})} \tag{70}$$

$$= \frac{\delta_{h(s,a)}(s')\mu_u^\pi(s,a,h(s,a))}{\mu_u^\pi(s,a,h(s,a))} \tag{71}$$

$$= \delta_{h(s,a)}(s') \tag{72}$$

$$= M(s'|s,a) \tag{73}$$

$$\tilde{\pi}_u(a|s) = \pi(a|s)\frac{\sum_{s'} M(s'|a,s)\mu_u^\pi(s,a,s')}{\sum_{\tilde{a},s'} \pi(\tilde{a}|s)M(s'|\tilde{a},s)\mu_u^\pi(s,\tilde{a},s')} \tag{74}$$

$$= \pi(a|s)\frac{\mu_u^\pi(s,a,h(s,a))}{\sum_{\tilde{a}} \pi(\tilde{a}|s)\mu_u^\pi(s,\tilde{a},h(s,\tilde{a}))} \tag{75}$$

$$\overset{(a)}{=} \pi(a|s)\mu_u^\pi(s,a,h(s,a)) \tag{76}$$

$$= \pi(a|s)\frac{p_{X^\pi(h(s,a))}\left(\frac{q_u^\pi(s)-c(s,a)}{\gamma}\right)}{\gamma p_{X^\pi(s)}(q_u^\pi(s))} \tag{77}$$

$$\propto \pi(a|s)\cdot p_{X^\pi(h(s,a))}\left(\frac{q_u^\pi(s)-c(s,a)}{\gamma}\right). \tag{78}$$

Here the equality (a) holds from (68). Thus, from Corollary 1, we can obtain the following approximation:

$$\mathbb{E}_{a_t\sim\tilde{\pi}_u}\left[c(s_t,a_t)+\gamma q_u^\pi(s_{t+1})\right] - \frac{\epsilon R}{1-\epsilon} \le q_u^\pi(s_t) \le \mathbb{E}_{a_t\sim\tilde{\pi}_u}\left[c(s_t,a_t)+\gamma q_u^\pi(s_{t+1})\right] + \frac{\epsilon R}{1-\epsilon}. \tag{79}$$

Therefore, we obtain

$$q_u^\pi(s_0) \le \mathbb{E}_{a_0\sim\tilde{\pi}_u}\left[c(s_0,a_0)+\gamma q_u^\pi(s_1)\right] + \frac{\epsilon R}{1-\epsilon} \tag{80}$$

$$\le \mathbb{E}_{a_0,a_1\sim\tilde{\pi}_u}\left[c(s_0,a_0)+\gamma c(s_1,a_1)+\gamma^2 q_u^\pi(s_2)\right] + \frac{\epsilon R}{1-\epsilon}(1+\gamma) \tag{81}$$

$$\le \cdots \tag{82}$$

$$\le \mathbb{E}_{\tilde{\pi}_u}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)}, \tag{83}$$

$$q_u^\pi(s_0) \ge \mathbb{E}_{a_0\sim\tilde{\pi}_u}\left[c(s_0,a_0)+\gamma q_u^\pi(s_1)\right] - \frac{\epsilon R}{1-\epsilon} \tag{84}$$

$$\ge \mathbb{E}_{a_0,a_1\sim\tilde{\pi}_u}\left[c(s_0,a_0)+\gamma c(s_1,a_1)+\gamma^2 q_u^\pi(s_2)\right] - \frac{\epsilon R}{1-\epsilon}(1+\gamma) \tag{85}$$

$$\ge \cdots \tag{86}$$

$$\ge \mathbb{E}_{\tilde{\pi}_u}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)}. \tag{87}$$

□

B.3 Proof of Theorem 2

Theorem 2. *Under deterministic dynamics $s_{t+1} = h(s_t,a_t)$ and Assumptions 1 and 2, $q_u^\pi(s)$ can be expressed as*

$$\left| q_u^\pi(s_0) - \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t)+\tilde{c}_u^\pi(s_t,a_t)\right\}\right]\right| \le \frac{\epsilon R}{(1-\epsilon)(1-\gamma)}, \tag{88}$$

where

$$\tilde{c}_u^\pi(s,a) = \left(\frac{p_{X^\pi(s')}\left(\frac{q_u^\pi(s)-c(s,a)}{\gamma}\right)}{\gamma p_{X^\pi(s)}\left(q_u^\pi(s)\right)} - 1 \right) \cdot \left[c(s,a) + \gamma q_u^\pi(h(s,a))\right].$$

Proof. From (79), we have

$$q_u^\pi(s_t) \leq \mathbb{E}_{a_t \sim \tilde{\pi}_u}\left[c(s_t,a_t) + \gamma q_u^\pi(s_{t+1})\right] + \frac{\epsilon R}{1-\epsilon} \tag{89}$$

$$q_u^\pi(s_t) \geq \mathbb{E}_{a_t \sim \tilde{\pi}_u}\left[c(s_t,a_t) + \gamma q_u^\pi(s_{t+1})\right] - \frac{\epsilon R}{1-\epsilon} \tag{90}$$

The expectation $\mathbb{E}_{a_t \sim \tilde{\pi}_u}\left[c(s_t,a_t) + \gamma q_u^\pi(s_{t+1})\right]$ can be rewritten as

$$\mathbb{E}_{a_t \sim \tilde{\pi}_u}\left[c(s_t,a_t) + \gamma q_u^\pi(s_{t+1})\right] = \mathbb{E}_{a_t \sim \pi}\left[\frac{\tilde{\pi}_u(a_t|s_t)}{\pi(a_t|s_t)} \cdot \left\{c(s_t,a_t) + \gamma q_u^\pi(h(s_t,a_t))\right\}\right] \tag{91}$$

$$= \mathbb{E}_{a_t \sim \pi}\left[c(s_t,a_t) + \tilde{c}_u^\pi(s_t,a_t) + \gamma q_u^\pi(h(s_t,a_t))\right] \tag{92}$$

$$= \mathbb{E}_{a_t \sim \pi}\left[c(s_t,a_t) + \tilde{c}_u^\pi(s_t,a_t) + \gamma q_u^\pi(s_{t+1})\right] \tag{93}$$

where

$$\tilde{c}_u^\pi(s,a) := \left(\frac{\tilde{\pi}_u(a|s)}{\pi(a|s)} - 1\right) \cdot \left\{c(s,a) + \gamma q_u^\pi(h(s,a))\right\} \tag{94}$$

$$= \left(\frac{p_{X^\pi(h(s,a))}\left(\frac{q_u^\pi(s)-c(s,a)}{\gamma}\right)}{\gamma p_{X^\pi(s)}\left(q_u^\pi(s)\right)} - 1 \right) \cdot \left\{c(s,a) + \gamma q_u^\pi(h(s,a))\right\}. \tag{95}$$

Then, using (93), we obtain

$$q_u^\pi(s_0) \leq \mathbb{E}_{a_0 \sim \tilde{\pi}_u}\left[c(s_0,a_0) + \gamma q_u^\pi(s_1)\right] + \frac{\epsilon R}{1-\epsilon} \tag{96}$$

$$= \mathbb{E}_{a_0 \sim \pi}\left[c(s_0,a_0) + \tilde{c}_u^\pi(s_0,a_0) + \gamma q_u^\pi(s_1)\right] + \frac{\epsilon R}{1-\epsilon} \tag{97}$$

$$\leq \mathbb{E}_{a_0,a_1 \sim \pi}\left[\left\{c(s_0,a_0) + \tilde{c}_u^\pi(s_0,a_0)\right\} + \gamma\left\{c(s_1,a_1) + \tilde{c}_u^\pi(s_1,a_1)\right\} + \gamma^2 q_u^\pi(s_2)\right] \tag{98}$$

$$+ \frac{\epsilon R}{1-\epsilon}(1+\gamma) \tag{99}$$

$$\leq \cdots \tag{100}$$

$$\leq \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t \left\{c(s_t,a_t) + \tilde{c}_u^\pi(s_t,a_t)\right\}\right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{101}$$

$$q_u^\pi(s_0) \geq \mathbb{E}_{a_0 \sim \tilde{\pi}_u}\left[c(s_0,a_0) + \gamma q_u^\pi(s_1)\right] - \frac{\epsilon R}{1-\epsilon} \tag{102}$$

$$= \mathbb{E}_{a_0 \sim \pi}\left[c(s_0,a_0) + \tilde{c}_u^\pi(s_0,a_0) + \gamma q_u^\pi(s_1)\right] - \frac{\epsilon R}{1-\epsilon} \tag{103}$$

$$\geq \mathbb{E}_{a_0,a_1 \sim \pi}\left[\left\{c(s_0,a_0) + \tilde{c}_u^\pi(s_0,a_0)\right\} + \gamma\left\{c(s_1,a_1) + \tilde{c}_u^\pi(s_1,a_1)\right\} + \gamma^2 q_u^\pi(s_2)\right] \tag{104}$$

$$- \frac{\epsilon R}{1-\epsilon}(1+\gamma) \tag{105}$$

$$\geq \cdots \tag{106}$$

$$\geq \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t \left\{c(s_t,a_t) + \tilde{c}_u^\pi(s_t,a_t)\right\}\right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{107}$$

$$\tag{108}$$

□

B.4 Proof of Theorem 3

Assumption 3 (Lipschitz continuity of $\tilde{c}_u^\pi(s,a)$ over π)**.** *For any given fixed $u \in (0,1)$ and any policies π and π', there exists a coefficient C_u such that*

$$\left|\tilde{c}_u^{\pi'}(s,a) - \tilde{c}_u^{\pi}(s,a)\right| \leq C_u \cdot \max_{s'} KL\left(\pi'(\cdot|s') \,\middle\|\, \pi(\cdot|s')\right) \tag{109}$$

for all $s \in \mathcal{S}$, $a \in \mathcal{A}$.

Basically, Assumption 3 is that the function $\tilde{c}_u^\pi$ as a function of π is continuous, which is expected to be satisfied if there is no abrupt change in the associated distributions.

Theorem 3. *Under deterministic dynamics $s_{t+1} = h(s_t, a_t)$ and Assumptions 1, 2, and 3, the u-quantile $q_u^\pi(s_0)$ is expressed as the expectation of the sum of actual cost and a π-independent additional cost $\tilde{c}_u^{\pi'}(s,a)$ for π' satisfying $\max_s KL(\pi'(\cdot|s) \,\|\, \pi(\cdot|s)) \leq \delta$:*

$$\left|q_u^\pi(s_0) - \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t) + \tilde{c}_u^{\pi'}(s_t,a_t)\right\}\right]\right| \leq \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma}\delta. \tag{110}$$

Proof. From Assumption 3, the additional cost $\tilde{c}_u^\pi(s,a)$ is bounded as follows

$$\begin{aligned}
\tilde{c}_u^\pi(s,a) &\leq \tilde{c}_u^{\pi'}(s,a) + C_u \cdot \max_s \mathrm{KL}(\pi'(\cdot|s) \,\|\, \pi(\cdot|s)) && (111)\\
&\leq \tilde{c}_u^{\pi'}(s,a) + C_u \cdot \delta && (112)\\
\tilde{c}_u^\pi(s,a) &\geq \tilde{c}_u^{\pi'}(s,a) - C_u \cdot \max_s \mathrm{KL}(\pi'(\cdot|s) \,\|\, \pi(\cdot|s)) && (113)\\
&\geq \tilde{c}_u^{\pi'}(s,a) - C_u \cdot \delta && (114)
\end{aligned}$$

for π' satisfying $\max_s \mathrm{KL}(\pi'(\cdot|s) \,\|\, \pi(\cdot|s)) \leq \delta$. Thus, from (88), we can obtain the following bounds

$$\begin{aligned}
q_u^\pi(s_0) &\leq \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t) + \tilde{c}_u^{\pi}(s_t,a_t)\right\}\right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} && (115)\\
&\leq \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t) + \tilde{c}_u^{\pi'}(s_t,a_t) + C_u\cdot\delta\right\}\right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} && (116)\\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t) + \tilde{c}_u^{\pi'}(s_t,a_t)\right\}\right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma}\delta && (117)\\
q_u^\pi(s_0) &\geq \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t) + \tilde{c}_u^{\pi}(s_t,a_t)\right\}\right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} && (118)\\
&\geq \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t) + \tilde{c}_u^{\pi'}(s_t,a_t) - C_u\cdot\delta\right\}\right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} && (119)\\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t) + \tilde{c}_u^{\pi'}(s_t,a_t)\right\}\right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} - \frac{C_u}{1-\gamma}\delta && (120)
\end{aligned}$$

□

B.5 Proof of Policy Improvement Condition

Lemma 2 (Telescoping Lemma for u-quantile)**.** *Under deterministic dynamics $s_{t+1} = h(s_t, a_t)$ and Assumption 1 and 2, the following holds for any two policies π and π':*

$$\left|q_u^\pi(s_0) - \left\{q_u^{\pi'}(s_0) + \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left(c(s_t,a_t) + \tilde{c}_u^\pi(s_t,a_t) + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t)\right)\right]\right\}\right| \tag{121}$$

$$\leq \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{122}$$

Proof. With Assumption 1 and 2, we have the following inequality by Theorem 2:

$$q_u^\pi(s_0) \leq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^\pi(s_t, a_t) \right\} \right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{123}$$

$$q_u^\pi(s_0) \geq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^\pi(s_t, a_t) \right\} \right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{124}$$

Then note that

$$q_u^{\pi'}(s_0) = -\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right]. \tag{125}$$

Therefore,

$$q_u^\pi(s_0) - q_u^{\pi'}(s_0) \tag{126}$$

$$\leq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^\pi(s_t, a_t) \right\} \right] + \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] \tag{127}$$

$$+ \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{128}$$

$$= \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^\pi(s_t, a_t) + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{129}$$

$$q_u^\pi(s_0) - q_u^{\pi'}(s_0) \tag{130}$$

$$\geq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^\pi(s_t, a_t) \right\} \right] + \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] \tag{131}$$

$$- \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{132}$$

$$= \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^\pi(s_t, a_t) + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{133}$$

□

Next, we can obtain the following corollary.

Corollary 2. *Under deterministic dynamics $s_{t+1} = h(s_t, a_t)$ and Assumptions 1, 2, and 3, the following holds for any two policies π and π' :*

$$\left| q_u^\pi(s_0) - \left\{ q_u^{\pi'}(s_0) + \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(c(s_t, a_t) + \tilde{c}_u^{\pi'}(s_t, a_t) + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right) \right] \right\} \right| \tag{134}$$

$$\leq \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma} \max_s \mathit{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{135}$$

Proof. Let denote $\delta = \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s))$ for simplicity in this proof. If Assumption 3 holds, this can be rewritten as

$$q_u^\pi(s_0) - q_u^{\pi'}(s_0) \tag{136}$$

$$\leq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^\pi(s_t, a_t) + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{137}$$

$$\leq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^{\pi'}(s_t, a_t) + C_u \cdot \delta + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{138}$$

$$= \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^{\pi'}(s_t, a_t) + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma}\delta \tag{139}$$

$$q_u^{\pi}(s_0) - q_u^{\pi'}(s_0) \tag{140}$$

$$\geq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^{\pi}(s_t, a_t) + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{141}$$

$$\geq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^{\pi'}(s_t, a_t) - C_u \cdot \delta + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{142}$$

$$= \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left\{ c(s_t, a_t) + \tilde{c}_u^{\pi'}(s_t, a_t) + \gamma q_u^{\pi'}(s_{t+1}) - q_u^{\pi'}(s_t) \right\} \right] - \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} - \frac{C_u}{1-\gamma}\delta \tag{143}$$

□

To prove improvement theorem, we need a definition of α-coupled policy and several lemmas similar to [19].

Definition 1 (From [19])**.** *The two policies π and π' are α-coupled if Pr $(a \neq a') \leq \alpha$, $(a, a') \sim (\pi(a|s), \pi'(a'|s))$ for all s.*

For the u-quantile, we define an advantage function $A_u^\pi(s, a)$ using the additional cost function $\tilde{c}_u^\pi(s, a)$ as

$$A_u^\pi(s, a) := c(s, a) + \tilde{c}_u^\pi(s, a) + \gamma \mathbb{E}_{s'} \left[q_u^\pi(s') \right] - q_u^\pi(s) \tag{144}$$

Lemma 3 (Similar to Lemma 2 in [19])**.** *Under deterministic dynamics $s_{t+1} = h(s_t, a_t)$ and Assumptions 1 and 2, α-coupled policies π and π' satisfy the following inequality*

$$\left| \mathbb{E}_\pi \left[A_u^{\pi'}(s, a) \right] \right| \leq 2\alpha \max_{s,a} \left| A_u^{\pi'}(s, a) \right| + \frac{\epsilon R}{(1-\epsilon)} \tag{145}$$

for all s.

Proof. (Similar to the proof of Lemma 2 in [19]) First we note that the following holds by (93) and Theorem 1:

$$\left| \mathbb{E}_{a \sim \pi'} \left[A_u^{\pi'}(s, a) \right] \right| = \left| \mathbb{E}_{a \sim \pi'} \left[c(s, a) + \tilde{c}_u^{\pi'}(s, a) + \gamma q_u^{\pi'}(s') - q_u^{\pi'}(s) \right] \right| \tag{146}$$

$$= \left| \mathbb{E}_{a \sim \pi'} \left[\frac{p_{X^{\pi'}(s')} \left(\frac{q_u^{\pi'}(s) - c(s,a)}{\gamma} \right)}{\gamma p_{X^{\pi'}(s)} \left(q_u^{\pi'}(s) \right)} \left\{ c(s, a) + \gamma q_u^{\pi'}(s') - q_u^{\pi'}(s) \right\} \right] \right| \tag{147}$$

$$\leq \frac{\epsilon R}{(1-\epsilon)} \tag{148}$$

Therefore,

$$\left| \mathbb{E}_\pi \left[A_u^{\pi'}(s, a) \right] \right| \overset{(a)}{\leq} \left| \mathbb{E}_{a \sim \pi} \left[A_u^{\pi'}(s, a) \right] - \mathbb{E}_{a' \sim \pi'} \left[A_u^{\pi'}(s, a') \right] \right| + \left| \mathbb{E}_{\pi'} \left[A_u^{\pi'}(s, a) \right] \right| \tag{149}$$

$$\overset{(b)}{\leq} \left| \mathbb{E}_{a \sim \pi} \left[A_u^{\pi'}(s, a) \right] - \mathbb{E}_{a' \sim \pi'} \left[A_u^{\pi'}(s, a') \right] \right| + \frac{\epsilon R}{(1-\epsilon)} \tag{150}$$

$$= \left| \mathbb{E}_{(a,a') \sim (\pi, \pi')} \left[A_u^{\pi'}(s, a) - A_u^{\pi'}(s, a') \right] \right| + \frac{\epsilon R}{(1-\epsilon)} \tag{151}$$

$$= \Bigg| \Pr(a = a') \, \mathbb{E}_{(a,a') \sim (\pi, \pi')|_{a=a'}} \left[A_u^{\pi'}(s, a) - A_u^{\pi'}(s, a') \right] \tag{152}$$

$$+ \Pr(a \neq a') \, \mathbb{E}_{(a,a') \sim (\pi,\pi')|_{a \neq a'}} \left[A_u^{\pi'}(s,a) - A_u^{\pi'}(s,a') \right] \Big| + \frac{\epsilon R}{(1-\epsilon)} \tag{153}$$

$$= \Pr(a \neq a') \left| \mathbb{E}_{(a,a') \sim (\pi,\pi')|_{a \neq a'}} \left[A_u^{\pi'}(s,a) - A_u^{\pi'}(s,a') \right] \right| + \frac{\epsilon R}{(1-\epsilon)} \tag{154}$$

$$\overset{(c)}{\leq} 2\alpha \max_{s,a} \left| A_u^{\pi'}(s,a) \right| + \frac{\epsilon R}{(1-\epsilon)} \tag{155}$$

where (a) holds by the triangular inequality, (b) holds by (148), and (c) holds since π and π' are α-coupled policies. □

Lemma 4 (Similar to Lemma 3 in [19])**.** *Under deterministic dynamics $s_{t+1} = h(s_t, a_t)$ and Assumptions 1 and 2, the following holds for α-coupled policies π and π'*

$$\left| \mathbb{E}_{s_t \sim \pi} \left[\mathbb{E}_{a \sim \pi} \left[A_u^{\pi'}(s_t,a) \right] \right] - \mathbb{E}_{s_t \sim \pi'} \left[\mathbb{E}_{a \sim \pi} \left[A_u^{\pi'}(s_t,a) \right] \right] \right| \tag{156}$$

$$\leq \left(1 - (1-\alpha)^t\right) \left\{ 4\alpha \max_{s,a} \left| A_u^{\pi'}(s,a) \right| + \frac{2\epsilon R}{(1-\epsilon)} \right\} \tag{157}$$

Proof. (Similar to the proof of Lemma 3 in [19]) For α-coupled policies π and π', first we consider trajectories drawn from each policy, i.e., $\tau = (s_0, a_0, s_1, a_1, \ldots) \sim \pi$ and $\tau' = (s_0, a'_0, s'_1, a'_1, \ldots) \sim \pi'$. We consider the timestep t and observe the advantage of π' over π. Let define n_t as the number of times that mismatched actions occurs, $a_i \neq a'_i$ for $i < t$. Then

$$\mathbb{E}_{s_t \sim \pi} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] \tag{158}$$

$$= P(n_t = 0) \cdot \mathbb{E}_{s_t \sim \pi | n_t = 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right]$$
$$+ P(n_t > 0) \cdot \mathbb{E}_{s_t \sim \pi | n_t > 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] \tag{159}$$

$$\mathbb{E}_{s_t \sim \pi'} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] \tag{160}$$

$$= P(n_t = 0) \cdot \mathbb{E}_{s_t \sim \pi' | n_t = 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right]$$
$$+ P(n_t > 0) \cdot \mathbb{E}_{s_t \sim \pi' | n_t > 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] \tag{161}$$

For the case $n_t = 0$,

$$\mathbb{E}_{s_t \sim \pi | n_t = 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] = \mathbb{E}_{s_t \sim \pi' | n_t = 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] \tag{162}$$

Thus by subtracting (161) and (159), we can obtain

$$\mathbb{E}_{s_t \sim \pi} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] - \mathbb{E}_{s_t \sim \pi'} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] \tag{163}$$

$$= P(n_t > 0) \cdot \left(\mathbb{E}_{s_t \sim \pi | n_t > 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] - \mathbb{E}_{s_t \sim \pi' | n_t > 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] \right) \tag{164}$$

From the definition of α-coupled policy, we get

$$P(n_t = 0) \geq (1-\alpha)^t, \qquad P(n_t > 0) \leq 1 - (1-\alpha)^t \tag{165}$$

Then note that

$$\left| \mathbb{E}_{s_t \sim \pi | n_t > 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] - \mathbb{E}_{s_t \sim \pi' | n_t > 0} \left[\mathbb{E}_{a \sim \pi(\cdot|s_t)} \left[A_u^{\pi'}(s_t,a) \right] \right] \right| \tag{166}$$

$$\overset{(a)}{\leq} \left| \mathbb{E}_{s_t \sim \pi | n_t > 0} \left[\mathbb{E}_{a \sim \pi} \left[A_u^{\pi'}(s_t,a) \right] \right] \right| + \left| \mathbb{E}_{s_t \sim \pi' | n_t > 0} \left[\mathbb{E}_{a \sim \pi} \left[A_u^{\pi'}(s_t,a) \right] \right] \right| \tag{167}$$

$$\leq 2 \max_{s} \left| \mathbb{E}_{a \sim \pi} \left[A_u^{\pi'}(s,a) \right] \right| \tag{168}$$

$$\overset{(b)}{\leq} 4\alpha \max_{s,a} \left| A_u^{\pi'}(s,a) \right| + \frac{2\epsilon R}{(1-\epsilon)} \tag{169}$$

where (a) holds by the triangular inequality, and (b) holds by Lemma 3. Therefore using (164), (165), and (169), we can conclude

$$\mathbb{E}_{s_t\sim\pi}\left[\mathbb{E}_{a\sim\pi(\cdot|s_t)}\left[A_u^{\pi'}(s_t,a)\right]\right]-\mathbb{E}_{s_t\sim\pi'}\left[\mathbb{E}_{a\sim\pi(\cdot|s_t)}\left[A_u^{\pi'}(s_t,a)\right]\right] \tag{170}$$

$$\leq\left(1-(1-\alpha)^t\right)\left\{4\alpha\max_{s,a}\left|A_u^{\pi'}(s,a)\right|+\frac{2\epsilon R}{(1-\epsilon)}\right\} \tag{171}$$

□

Now we define $L_u^{\pi'}(\pi)$ as

$$L_u^{\pi'}(\pi):=q_u^{\pi'}(s_0)+\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right] \tag{172}$$

$$=q_u^{\pi'}(s_0)+\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{a\sim\pi}\left[c(s_t,a)+\tilde{c}_u^{\pi'}(s_t,a)+\gamma q_u^{\pi'}(s_{t+1})-q_u^{\pi'}(s_t)\right]\right] \tag{173}$$

Then note that

$$L_u^{\pi'}(\pi')=q_u^{\pi'}(s_0)+\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t)+\tilde{c}_u^{\pi'}(s_t,a_t)+\gamma q_u^{\pi'}(s_{t+1})-q_u^{\pi'}(s_t)\right\}\right] \tag{174}$$

$$=\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t)+\tilde{c}_u^{\pi'}(s_t,a_t)\right\}\right] \tag{175}$$

Then from Theorem 2,

$$\left|q_u^{\pi'}(s_0)-L_u^{\pi'}(\pi')\right|=\left|q_u^{\pi'}(s_0)-\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t)+\tilde{c}_u^{\pi'}(s_t,a_t)\right\}\right]\right| \tag{176}$$

$$\leq\frac{\epsilon R}{(1-\epsilon)(1-\gamma)}. \tag{177}$$

Therefore, we get

$$q_u^{\pi'}(s_0)\geq L_u^{\pi'}(\pi')-\frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{178}$$

Proposition 1. *Under deterministic dynamics $s_{t+1}=h(s_t,a_t)$ and Assumptions 1, 2, and 3, the following holds*

$$q_u^{\pi}(s_0)\leq L_u^{\pi'}(\pi)+C_1\max_s KL(\pi'(\cdot|s)\parallel\pi(\cdot|s))+C_2\frac{\epsilon}{1-\epsilon} \tag{179}$$

where

$$C_1=\left(\frac{4\gamma\max_{s,a}\left|A_u^{\pi'}(s,a)\right|+\gamma R}{(1-\gamma)^2}+\frac{C_u}{1-\gamma}\right),\qquad C_2=\frac{R}{(1-\gamma)^2} \tag{180}$$

Proof. Let define $B=\max_{s,a}\left|A_u^{\pi'}(s,a)\right|$. Remind that the definition of the advantage for the u-quantile (144) $A_u^{\pi'}(s,a):=c(s,a)+\tilde{c}_u^{\pi'}(s,a)+\gamma\mathbb{E}_{s'}\left[q_u^{\pi'}(s')\right]-q_u^{\pi'}(s)$, Corollary 2

$$\left|q_u^{\pi}(s_0)-\left\{q_u^{\pi'}(s_0)+\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t\underbrace{\left(c(s_t,a_t)+\tilde{c}_u^{\pi'}(s_t,a_t)+\gamma q_u^{\pi'}(s_{t+1})-q_u^{\pi'}(s_t)\right)}_{=A_u^{\pi'}(s_t,a_t)}\right]\right\}\right|$$

$$\leq\frac{\epsilon R}{(1-\epsilon)(1-\gamma)}+\frac{C_u}{1-\gamma}\max_s\mathrm{KL}(\pi'(\cdot|s)\parallel\pi(\cdot|s)) \tag{181}$$

and the definition of $L_u^{\pi'}(\pi)$ in (172)

$$L_u^{\pi'}(\pi) := q_u^{\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right] \tag{182}$$

Then we can obtain

$$\left|q_u^{\pi}(s_0) - L_u^{\pi'}(\pi)\right| \tag{183}$$

$$\overset{(a)}{\leq} \left|q_u^{\pi}(s_0) - \left\{q_u^{\pi'}(s_0) + \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t A_u^{\pi'}(s_t,a_t)\right]\right\}\right| \tag{184}$$

$$+ \left|\left\{q_u^{\pi'}(s_0) + \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t A_u^{\pi'}(s_t,a_t)\right]\right\} - L_u^{\pi'}(\pi)\right| \tag{185}$$

$$\overset{(b)}{\leq} \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma}\max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{186}$$

$$+ \left|\left\{q_u^{\pi'}(s_0) + \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t A_u^{\pi'}(s_t,a_t)\right]\right\} - L_u^{\pi'}(\pi)\right| \tag{187}$$

$$= \left|\left\{q_u^{\pi'}(s_0) + \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t A_u^{\pi'}(s_t,a_t)\right]\right\} - \left\{q_u^{\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right]\right\}\right| \tag{188}$$

$$+ \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma}\max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{189}$$

$$= \underbrace{\left|\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right] - \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right]\right|}_{(c)} \tag{190}$$

$$+ \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma}\max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{191}$$

where (a) holds by the triangular inequality, (b) holds from (181). The term (c) can be written as

$$\left|\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right] - \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right]\right| \tag{192}$$

$$\overset{(a)}{\leq} \sum_{t=0}^{\infty}\gamma^t \left|\mathbb{E}_{s_t\sim\pi}\left[\mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s,a)\right]\right] - \mathbb{E}_{s_t\sim\pi'}\left[\mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s,a)\right]\right]\right| \tag{193}$$

$$\overset{(b)}{\leq} \sum_{t=0}^{\infty}\gamma^t\left(1-(1-\alpha)^t\right)\left\{4\alpha\max_{s,a}\left|A_u^{\pi'}(s,a)\right| + \frac{2\epsilon R}{(1-\epsilon)}\right\} \tag{194}$$

$$= \left(4\alpha\max_{s,a}\left|A_u^{\pi'}(s,a)\right| + \frac{2\epsilon R}{(1-\epsilon)}\right)\left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha)}\right) \tag{195}$$

$$= \left(4\alpha\max_{s,a}\left|A_u^{\pi'}(s,a)\right| + \frac{2\epsilon R}{(1-\epsilon)}\right)\frac{\alpha\gamma}{(1-\gamma)(1-\gamma(1-\alpha))} \tag{196}$$

$$\overset{(c)}{\leq} \left(4\alpha\max_{s,a}\left|A_u^{\pi'}(s,a)\right| + \frac{2\epsilon R}{(1-\epsilon)}\right)\frac{\alpha\gamma}{(1-\gamma)^2} \tag{197}$$

where (a) holds by the triangular inequality, (b) holds by Lemma 4, and (c) holds by $\alpha < 1$ (α is for α-coupled policy). Therefore by putting (197) into term (c) in (191), we can obtain

$$\left|q_u^{\pi}(s_0) - L_u^{\pi'}(\pi)\right| \tag{198}$$

$$= \underbrace{\left|\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right] - \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right]\right|}_{(c)} \tag{199}$$

$$+ \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma} \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{200}$$

$$\leq \left(4\alpha \max_{s,a} \left| A_u^{\pi'}(s,a) \right| + \frac{2\epsilon R}{(1-\epsilon)} \right) \frac{\alpha\gamma}{(1-\gamma)^2} + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma} \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{201}$$

$$= \left(\frac{4\gamma \max_{s,a} \left| A_u^{\pi'}(s,a) \right|}{(1-\gamma)^2} \right) \alpha^2 + \left(2\alpha \frac{\epsilon}{1-\epsilon} \right) \cdot \frac{\gamma R}{(1-\gamma)^2} \tag{202}$$

$$+ \left(\frac{(1-\gamma)R}{(1-\epsilon)(1-\gamma)^2} \right) \epsilon + \frac{C_u}{1-\gamma} \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{203}$$

$$\overset{(a)}{\leq} \left(\frac{4\gamma \max_{s,a} \left| A_u^{\pi'}(s,a) \right|}{(1-\gamma)^2} \right) \alpha^2 + \left\{ \alpha^2 + \left(\frac{\epsilon}{1-\epsilon} \right)^2 \right\} \cdot \frac{\gamma R}{(1-\gamma)^2} \tag{204}$$

$$+ \left(\frac{(1-\gamma)R}{(1-\epsilon)(1-\gamma)^2} \right) \epsilon + \frac{C_u}{1-\gamma} \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{205}$$

$$= \left(\frac{4\gamma \max_{s,a} \left| A_u^{\pi'}(s,a) \right| + \gamma R}{(1-\gamma)^2} \right) \alpha^2 + \frac{C_u}{1-\gamma} \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{206}$$

$$+ \left(\frac{(1-\gamma)R}{(1-\gamma)^2} \right) \left(\frac{\epsilon}{1-\epsilon} \right) + \frac{\gamma R}{(1-\gamma)^2} \left(\frac{\epsilon}{1-\epsilon} \right)^2 \tag{207}$$

$$\overset{(b)}{\leq} \left(\frac{4\gamma \max_{s,a} \left| A_u^{\pi'}(s,a) \right| + \gamma R}{(1-\gamma)^2} \right) \alpha^2 + \frac{C_u}{1-\gamma} \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{208}$$

$$+ \left(\frac{(1-\gamma)R}{(1-\gamma)^2} \right) \left(\frac{\epsilon}{1-\epsilon} \right) + \frac{\gamma R}{(1-\gamma)^2} \left(\frac{\epsilon}{1-\epsilon} \right) \tag{209}$$

$$= \frac{R}{(1-\gamma)^2} \left(\frac{\epsilon}{1-\epsilon} \right) + \left(\frac{4\gamma \max_{s,a} \left| A_u^{\pi'}(s,a) \right| + \gamma R}{(1-\gamma)^2} \right) \alpha^2 + \frac{C_u}{1-\gamma} \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{210}$$

where (a) holds by the inequality of arithmetic and geometric means, and (b) holds from the definition of $0 < \epsilon < \frac{1}{2}$ in Assumption 2. Like [19], if we take α as the maximum of the total variation of two policies π and π', i.e., $\alpha = \max_s D_{TV}(\pi'(\cdot|s)||\pi(\cdot|s))$, then these policies are α-coupled. Since $D_{TV}(\pi'(\cdot|s)||\pi(\cdot|s))^2 \leq \mathrm{KL}(\pi'(\cdot|s)||\pi(\cdot|s))$, eq. (210) becomes

$$\left| q_u^{\pi}(s_0) - L_u^{\pi'}(\pi) \right| \leq C_1 \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) + C_2 \frac{\epsilon}{1-\epsilon} \tag{211}$$

where

$$C_1 = \left(\frac{4\gamma \max_{s,a} \left| A_u^{\pi'}(s,a) \right| + \gamma R}{(1-\gamma)^2} + \frac{C_u}{1-\gamma} \right), \qquad C_2 = \frac{R}{(1-\gamma)^2} \tag{212}$$

□

Together with Proposition 1 above, and Theorem 1 in [19], we can obtain the Theorem 4 for policy improvement condition.

Theorem 4. *Let $\pi_{new} := \pi_{\theta_{new}}$ be the solution of the problem of maximizing*

$$L^{\pi_{old}}(\pi_\theta) - \tilde{C}_1 \max_s KL(\pi_{old}(\cdot|s) \parallel \pi_\theta(\cdot|s)), \tag{213}$$

where

$$L^{\pi_{old}}(\pi_\theta) = L_r^{\pi_{old}}(\pi_\theta) - \lambda L_{1-\epsilon_0}^{\pi_{old}}(\pi_\theta) \tag{214}$$

$$= \left(V^{\pi_{old}}(s_0) - \lambda q_{1-\epsilon_0}^{\pi_{old}}(s_0)\right) + \mathbb{E}_{\pi_{old}}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi_\theta}\left[A_r^{\pi_{old}}(s_t,a) - \lambda A_{1-\epsilon_0}^{\pi_{old}}(s_t,a)\right]\right], \tag{215}$$

and some constant $\tilde{C}_1 > 0$. Then, under deterministic dynamics $s_{t+1} = h(s_t, a_t)$ and Assumptions 1, 2, and 3, the following inequality holds:

$$L_{quant}(\pi_{new}, \lambda) - L_{quant}(\pi_{old}, \lambda) \tag{216}$$

$$\geq L^{\pi_{old}}(\pi_{new}) - L^{\pi_{old}}(\pi_{old}) - \tilde{C}_1 KL_{max}(\pi_{old}||\pi_{new}) - \underbrace{\tilde{C}_2 \frac{\epsilon}{1-\epsilon}}_{\textit{approximation loss}} \tag{217}$$

for a given Lagrange multiplier $\lambda > 0$, some constant $\tilde{C}_2$ and small $\epsilon > 0$.

Remind that Theorem 1 of [19] with our notation:

Theorem 5 (Theorem 1 of [19])**.**

$$V^\pi(s_0) \geq \underbrace{V^{\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_r^{\pi'}(s_t,a)\right]\right]}_{=:L_r^{\pi'}(\pi)} - C_3 \max_s KL(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{218}$$

$$= L_r^{\pi'}(\pi) - C_3 \max_s KL(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{219}$$

where

$$C_3 = \frac{4\gamma \max_{s,a}\left|A_r^{\pi'}(s,a)\right|}{(1-\gamma)^2} \tag{220}$$

$$A_r^{\pi'}(s,a) := r(s,a) + \gamma \mathbb{E}_{s'}\left[V^{\pi'}(s')\right] - V^{\pi'}(s) \tag{221}$$

We omit the proof of Theorem 5. Please see [19] for the proof. Note that Theorem 5 holds for any two policies π and π' as we can see in the appendix of the original paper [19].

Finally now we prove Theorem 4.

Proof of Theorem 4. From Proposition 1, we have

$$q_u^\pi(s_0) \leq \underbrace{q_u^{\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right]}_{L_u^{\pi'}(\pi)} + C_1 \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) + C_2 \frac{\epsilon}{1-\epsilon} \tag{222}$$

for

$$A_u^{\pi'}(s,a) := c(s,a) + \tilde{c}_u^{\pi'}(s,a) + \gamma \mathbb{E}_{s'}\left[q_u^{\pi'}(s')\right] - q_u^{\pi'}(s) \tag{223}$$

$$C_1 = \left(\frac{4\gamma \max_{s,a}\left|A_u^{\pi'}(s,a)\right| + \gamma R}{(1-\gamma)^2} + \frac{C_u}{1-\gamma}\right) \tag{224}$$

$$C_2 = \frac{R}{(1-\gamma)^2}, \tag{225}$$

and from Theorem 1 in [19] (or Theorem 5 in this appendix), we have

$$V^\pi(s_0) \geq \underbrace{V^{\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_r^{\pi'}(s_t,a)\right]\right]}_{=:L_r^{\pi'}(\pi)} - C_3 \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{226}$$

where

$$A_r^{\pi'}(s,a) := r(s,a) + \gamma \mathbb{E}_{s'}\left[V^{\pi'}(s')\right] - V^{\pi'}(s) \tag{227}$$

$$C_3 = \frac{4\gamma \max_{s,a}\left|A_r^{\pi'}(s,a)\right|}{(1-\gamma)^2}. \tag{228}$$

For a given $\lambda > 0$, by subtracting $\lambda\times$ (222) from (226), then we have

$$\begin{aligned}
&V^{\pi}(s_0) - \lambda q_u^{\pi}(s_0) \\
&\geq L_r^{\pi'}(\pi) - \lambda L_u^{\pi'}(\pi) - (\lambda C_1 + C_3)\max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) - \lambda C_2 \frac{\epsilon}{1-\epsilon} && (229)\\
&= V^{\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_r^{\pi'}(s_t,a)\right]\right] - \lambda\left\{q_u^{\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_u^{\pi'}(s_t,a)\right]\right]\right\} \\
&\quad - (\lambda C_1 + C_3)\max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) - \lambda C_2 \frac{\epsilon}{1-\epsilon} && (230)\\
&= \underbrace{\left(V^{\pi'}(s_0) - \lambda q_u^{\pi'}(s_0)\right) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a\sim\pi}\left[A_r^{\pi'}(s_t,a) - \lambda A_u^{\pi'}(s_t,a)\right]\right]}_{=:L^{\pi'}(\pi)} && (231)\\
&\quad - (\lambda C_1 + C_3)\max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) - \lambda C_2 \frac{\epsilon}{1-\epsilon} && (232)\\
&= L^{\pi'}(\pi) - (\lambda C_1 + C_3)\max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) - \lambda C_2 \frac{\epsilon}{1-\epsilon} && (233)
\end{aligned}$$

Therefore now we have

$$\begin{aligned}
L_{quant}(\pi,\lambda) &= V^{\pi}(s_0) - \lambda\left(q_u^{\pi}(s_0) - d_{th}\right) && (234)\\
&\geq L^{\pi'}(\pi) + \lambda \cdot d_{th} - (\lambda C_1 + C_3)\max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) - \lambda C_2 \frac{\epsilon}{1-\epsilon} && (235)
\end{aligned}$$

Note that

$$\begin{aligned}
L_r^{\pi'}(\pi') &= V^{\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a'\sim\pi'}\left[A_r^{\pi'}(s_t,a')\right]\right] && (236)\\
&= V^{\pi'}(s_0) && (237)\\
L_u^{\pi'}(\pi') &= q_u^{\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{a'\sim\pi'}\left[A_u^{\pi'}(s_t,a')\right]\right] && (238)\\
&\overset{(a)}{\leq} q_u^{\pi'}(s_0) + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} && (239)
\end{aligned}$$

where (a) holds by (178). Therefore,

$$\begin{aligned}
L^{\pi'}(\pi') &= L_r^{\pi'}(\pi') - \lambda L_u^{\pi'}(\pi') && (240)\\
&\geq V^{\pi'}(s_0) - \lambda\left(q_u^{\pi'}(s_0) + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)}\right) && (241)\\
&= V^{\pi'}(s_0) - \lambda\left(q_u^{\pi'}(s_0) - d_{th}\right) - \lambda\left(d_{th} + \frac{\epsilon R}{(1-\epsilon)(1-\gamma)}\right) && (242)\\
&= L_{quant}(\pi',\lambda) - \lambda \cdot d_{th} - \lambda\frac{\epsilon R}{(1-\epsilon)(1-\gamma)}. && (243)
\end{aligned}$$

By rearranging this, we get

$$-L_{quant}(\pi',\lambda) \geq -L^{\pi'}(\pi') - \lambda \cdot d_{th} - \lambda\frac{\epsilon R}{(1-\epsilon)(1-\gamma)}. \tag{244}$$

Therefore by adding (244) and (235), we can conclude

$$L_{quant}(\pi, \lambda) - L_{quant}(\pi', \lambda) \geq L^{\pi'}(\pi) - L^{\pi'}(\pi') - (\lambda C_1 + C_3) \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) \tag{245}$$

$$- \lambda C_2 \frac{\epsilon}{1-\epsilon} - \lambda \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} \tag{246}$$

$$= L^{\pi'}(\pi) - L^{\pi'}(\pi') - \tilde{C}_1 \max_s \mathrm{KL}(\pi'(\cdot|s) \parallel \pi(\cdot|s)) - \tilde{C}_2 \frac{\epsilon}{1-\epsilon} \tag{247}$$

where

$$\tilde{C}_1 = \lambda C_1 + C_3 \tag{248}$$

$$= \lambda \left(\frac{4\gamma \max_{s,a} \left| A_u^{\pi'}(s,a) \right| + \gamma R}{(1-\gamma)^2} + \frac{C_u}{1-\gamma} \right) + \frac{4\gamma \max_{s,a} \left| A_r^{\pi'}(s,a) \right|}{(1-\gamma)^2} \tag{249}$$

$$= \frac{4\gamma \left(\max_{s,a} \left| A_r^{\pi'}(s,a) \right| + \lambda \max_{s,a} \left| A_u^{\pi'}(s,a) \right| \right)}{(1-\gamma)^2} + \lambda \frac{\gamma R}{(1-\gamma)^2} + \lambda \frac{C_u}{1-\gamma} \tag{250}$$

$$\tilde{C}_2 = \lambda C_2 + \lambda \frac{R}{1-\gamma} \tag{251}$$

$$= \lambda \left(\frac{R}{(1-\gamma)^2} + \frac{R}{1-\gamma} \right) \tag{252}$$

$$= \lambda \frac{(2-\gamma)R}{(1-\gamma)^2} \tag{253}$$

□

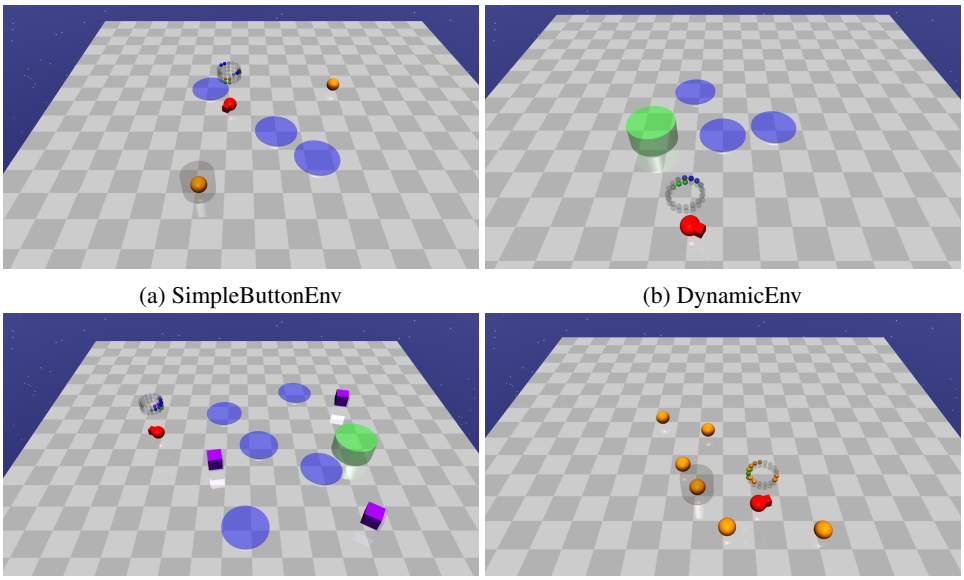
C Detailed Explanation of The Environments

The considered environments are SimpleButtonEnv, DynamicEnv [27], GremlinEnv, and DynamicButtonEnv, which are based on Safety Gym [17], MuJoCo [23], and OpenAI Gym [4]. The experiments are performed on a server with Intel(R) Xeon(R) Gold 6240R CPU @2.40GHz, and each experiment takes $8 \sim 10$ hours. The environments are illustrated in Fig. 7. The goal of these environments is for a robot (red sphere) to reach a goal (the orange sphere wrapped by a grey translucent pillar for SimpleButtonEnv and DynamicButtonEnv, and the green pillar for DynamicEnv and GremlinEnv), while avoiding hazards (blue circles) or the non-goal button (the orange sphere). Once the robot reaches the current goal, the environments generate the next goal deterministically (SimpleButtonEnv) or randomly (DynamicEnv, GremlinEnv, DynamicButtonEnv). When the robot performs an action at time step t, it receives a reward $\left\{\|p_{t+1} - p_{\text{goal}}\|_2 - \|p_t - p_{\text{goal}}\|_2\right\} + 1_{\text{goal reached}}$, where p_t is the position (x, y) of the robot at time step t and p_{goal} is the current goal position at time step t. It also receives a cost $+1$ if the robot touches non-goal objects (a hazard or the non-goal button), and 0 otherwise. Hence, for the robot, it receives a higher return when the robot touches more goals in a maximum timesteps $T = 1000$, and causes a higher sum of costs when the robot touches the other objects more often.

SimpleButtonEnv: This environment consists of a robot (the red sphere), three hazards (blue pillars), a goal button (the orange sphere wrapped by a grey translucent pillar), and a non-goal button (the orange sphere). When it starts a new episode, it locates the robot randomly in in a restricted region $[x_{min}, x_{max}, y_{min}, y_{max}] = [-1.5, 1.5, -1.5, 1.5]$ and the other objects in a fixed position. When the robot reaches the current goal, it sets the next goal as the non-goal button. Thus, the objective of this environment is to touch two buttons many times iteratively in a fixed maximum timesteps.

DynamicEnv: This environment consists of a robot (the red sphere), three hazards (blue pillars), and a goal (the green pillar). When it starts a new episode, it locates these objects randomly in a restricted region $[x_{min}, x_{max}, y_{min}, y_{max}] = [-1.5, 1.5, -1.5, 1.5]$. When the robot reaches the current goal, the next goal is generated at a random position.

GremlinEnv: This environment consists of a robot (the red sphere), five hazards (blue pillars), three gremlins (purple moving cubes), and a goal (the green pillar). This is similar to DynamicEnv except the gremlins and higher complexity of the task. Each gremlin goes around in a circle, and when the agent touches a gremiln, it receives a cost. When it starts a new episode, it locates these objects randomly in a restricted region $[x_{min}, x_{max}, y_{min}, y_{max}] = [-2, 2, -2, 2]$. When the robot reaches the current goal, the next goal is generated at a random position.



(a) SimpleButtonEnv (b) DynamicEnv

(c) GremlinEnv (d) DynamicButtonEnv

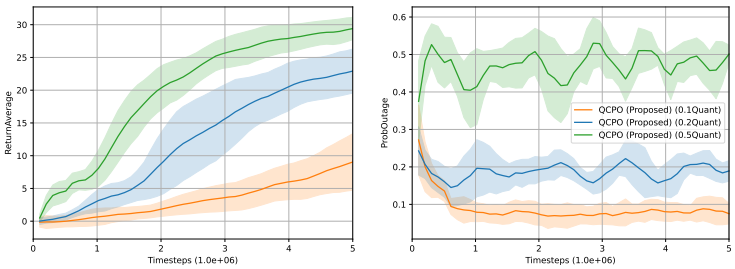
Figure 7: The considered environments

DynamicButtonEnv: This environment consists of a robot (the red sphere), and goal button (the orange sphere wrapped by a grey translucent pillar), and five non-goal buttons (the orange sphere). When it starts a new episode, it locates these objects randomly in a restricted region $[x_{min}, x_{max}, y_{min}, y_{max}] = [-1.5, 1.5, -1.5, 1.5]$. When the robot reaches the current goal, it sets the next goal randomly among non-goal buttons. This environment is similar to DynamicEnv but the hazards are the non-goal buttons.

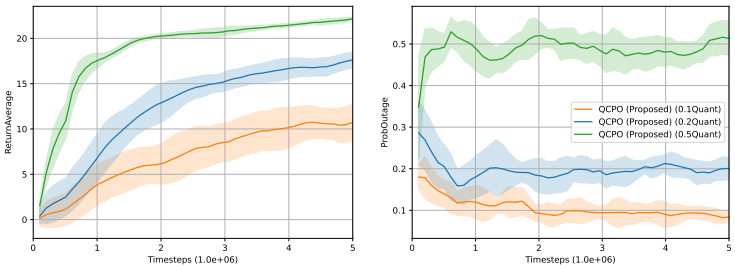
Observation Space: The observation in these environments is sensor values (accelerometer, velocimeter, gyro, and magnetometer) plus lidar values which measure the distance between the robot and the other objects. There are 16 lidar sensors for each object (a goal, hazards, buttons, gremlins) and these are located around the robot. Each lidar sensor for an object measures the distance between the robot and the object located in its corresponding direction. Gathering all these sensor values, the environment gives these values to the agent as an observation at the current time. The dimensions of the observation spaces are 44 (DynamicEnv, DynamicButtonEnv) and 60 (SimpleButtonEnv and GremlinEnv).

D More Results

D.1 QCPO with Various Target Outage Probability ϵ_0

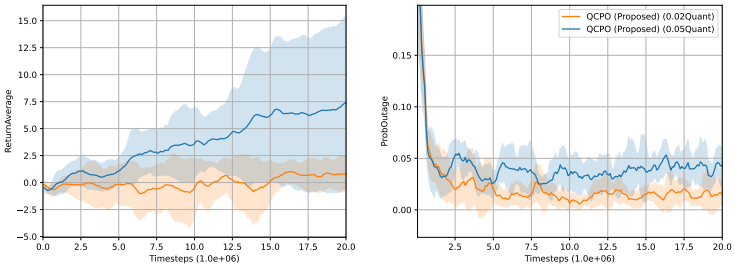


(a) Average Return (b) Outage Probability

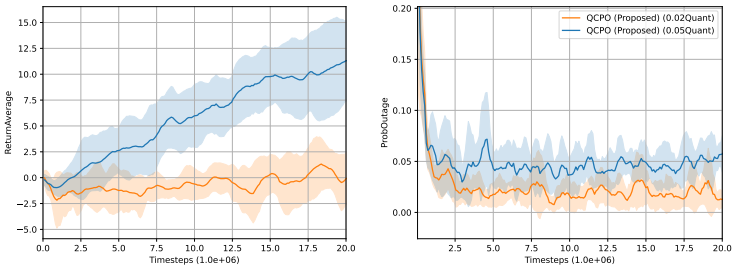


(c) Average Return (d) Outage Probability

Figure 8: Results of QCPO with $\epsilon_0 = 0.5$ (green), 0.2 (blue) and 0.1 (orange) on SimpleButtonEnv (1st row), DynamicEnv (2nd row): (left) average return of the most current 100 episodes and (right) outage probability of the most current 100 episodes.



(a) Average Return (b) Outage Probability



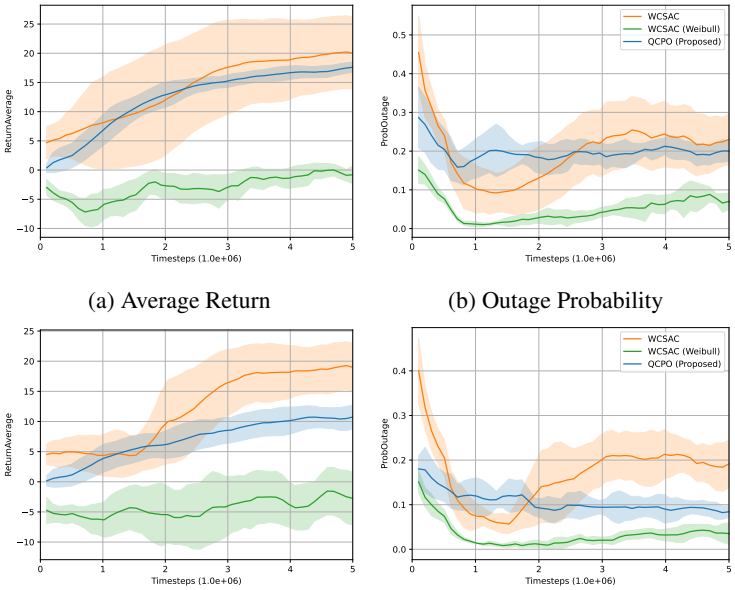
(c) Average Return (d) Outage Probability

Figure 9: Results of QCPO with $\epsilon_0 = 0.05$ (blue) and 0.02 (orange) on SimpleButtonEnv (1st row) and DynamicEnv (2nd row): (left) average return of the most current 100 episodes and (right) outage probability of the most current 100 episodes.

In this subsection, we provide results of QCPO with various target outage probabilities $\epsilon_0 = 0.5, 0.2, 0.1, 0.05$ and 0.02. Fig. 8 shows the average return and the outage probability of QCPO with $\epsilon_0 = 0.5, 0.2$ and 0.1. It is seen that QCPO satisfies the outage probability constraint after some initial time and then tries to increase the return while satisfying the outage probability constraint. Fig. 9 shows the average return and the outage probability of QCPO with $\epsilon_0 = 0.05$ and 0.02. In Fig. 9, it is again seen that QCPO satisfies the outage probability constraint after some initial time and then tries to increase the return while satisfying the outage probability constraint. However, it seems that more initial time steps are required than in the case of $\epsilon_0 = 0.5, 0.2$ and 0.1 to satisfy the target outage probability.

D.2 WCSAC with Weibull distribution approximation

In this subsection, we provide results of QCPO, WCSAC[27], and WCSAC with Weibull distribution approximation. In Fig. 10, it is seen that WCSAC with Weibull distribution approximation satisfies the outage probability constraint, while the original WCSAC with Gaussian distribution approximation does not. These results can imply that Weibull distribution approximation can estimate the true underlying distribution of the cumulative sum cost better than Gaussian distribution, and this is due to the limited capability of Gaussian distribution to capture the decay rate of the tail probability.



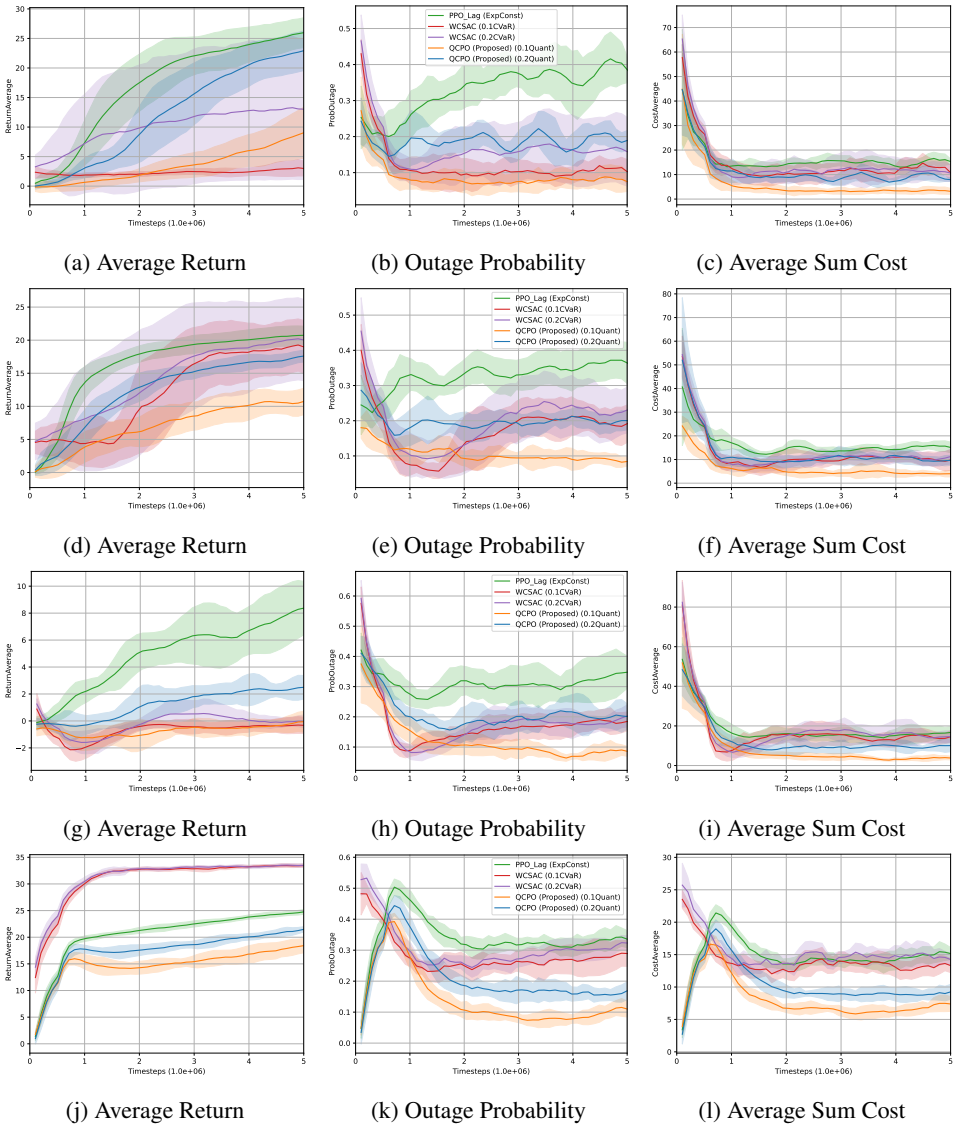
(a) Average Return (b) Outage Probability

(c) Average Return (d) Outage Probability

Figure 10: Results of QCPO (blue), WCSAC[27] (orange) and WCSAC with Weibull distribution approximation (green) on DynamicEnv: (1st row) $\epsilon_0 = 0.2$, (2nd row) $\epsilon_0 = 0.1$, (left) average return of the most current 100 episodes and (right) outage probability of the most current 100 episodes.

D.3 Performance Comparison

Fig. 11 shows the results of the considered algorithms on SimpleButtonEnv, DynamicEnv, GremlinEnv, and DynamicButtonEnv explained in Appendix C. All experiments were done with 10 different random seeds, and the real line and the shaded area represent the average and average $\pm$ standard deviation, respectively. PPO with the Lagrangian multiplier method for (ExpCP) (green) keeps the average of the sum cost around the threshold $d_{th} = 15$ well (see Fig. 11c, 11f, 11i, and 11l), and its outage probability is around 0.35 on SimpleButtonEnv and DynamicEnv (Fig. 11b and 11e), and 0.3 on GremlinEnv and DynamicButtonEnv (Fig. 11h and 11k). Note that the CVaR approach (WCSAC) should satisfy a sufficient condition for satisfying the outage probability constraint in (ProbCP). It is seen that WCSAC ($\epsilon_0 = 0.2$ (purple), $\epsilon_0 = 0.1$ (red)) achieves a lower or similar outage probability to the threshold ϵ_0 in Fig. 11b, but the algorithm does not satisfy the outage probability constraint exactly in Fig. 11e, 11h, and 11k. This means that the Gaussian distribution approximation of the



(a) Average Return (b) Outage Probability (c) Average Sum Cost

(d) Average Return (e) Outage Probability (f) Average Sum Cost

(g) Average Return (h) Outage Probability (i) Average Sum Cost

(j) Average Return (k) Outage Probability (l) Average Sum Cost

Figure 11: Results on SimpleButtonEnv (1st row), DynamicEnv (2nd row), GremlinEnv (3rd row), and DynamicButtonEnv (4th row): (1st column) average return of the most current 100 episodes, (2nd column) outage probability of the most current 100 episodes, and (3rd column) average sum of costs of the most current 100 episodes.

distribution of $X^\pi(s)$ has limited capability to capture the decay rate of the tail probability. On the other hand, the proposed QCPO ($\epsilon_0 = 0.2$ (blue), $\epsilon_0 = 0.1$ (orange)) maintains the outage probability around the desired target outage probability very well, as shown in Fig. 11b, 11e, 11h, and 11k.

Now consider the average return of these algorithms. In constrained RL, in general, if an algorithm is allowed to have a higher sum of costs, then it has a higher return. Thus, as seen in Fig. 11b, 11e, 11h, and 11k, PPO_Lag induces the highest outage probability, so it has the highest average return, as shown in Fig. 11a, 11d, and 11g. (For DynamicButtonEnv, WCSAC outperforms PPO_Lag, and this is because that SAC, the base algorithm of WCSAC, is a better algorithm than PPO, the base algorithm of PPO_Lag, on most unconstrained environments.) The direct comparison between WCSAC and QCPO is less meaningful in DynamicEnv, GremlinEnv, and DynamicButtonEnv since WCSAC does not satisfy the outage probability constraint, but it is fair in SimpleButtonEnv because both algorithms satisfy the outage probability constraint. As seen in Fig. 11a, QCPO achieves a higher average return than WCSAC for the same target probability constraint $\epsilon_0 = 0.1, 0.2$. This is

because QCPO satisfies the target outage probability exactly, i.e., uses the given cost budget fully for a higher return.

E Implementation Details

The implementation of the proposed algorithm[5] is based on the implementation of [21][6]

E.1 Network structures

Since the proposed algorithm is based on PPO [20], the network structure is similar to the network structure of PPO. The networks for the value function, the quantile function, the policy, and the Weibull distribution parameters have a common shared network to extract a feature of its observation. The common network has two MLP layers of size 512 with the tanh activation function, and an LSTM layer of size 512 with tanh activation function. The current observation changes to its feature through the two MLP layers, then concatenates this feature of the current observation, the previous action, the previous reward, and the previous cost to input the LSTM network. Thus, the common network outputs a feature of all previous information in the current trajectory. The output of the LSTM layer is then used as the input of the uncommon parts of the functions. The value function for reward has a linear MLP layer of size 1, and the quantile function for cost has a MLP layer of size n_q (number of quantile estimates) with exponential activation $\exp(x)$. Thus, the feature computed by the common feature network goes through these MLP networks to compute its value $V^\pi(s)$ and its quantile $q_u^\pi(s)$ for $u \in \{u_1, u_2, \ldots, u_{n_q}\}$. The policy network has a linear MLP layer of size 1, which outputs the mean parameter of Gaussian distribution, and a variable which indicates state-independent log standard deviation for Gaussian distribution. For the Weibull distribution parameters, there are two networks, one for $\alpha(s)$ and the other for $\beta(s)$, having a MLP layer of size 1. For $\alpha(s)$, the network has 4 * sigmoid activation function, and for $\beta(s)$, the network has the exponential activation function.

E.2 Loss Functions

The parameters are updated by minimizing their own loss functions. The loss function of the value parameter ϕ is

$$L(\phi) = \frac{1}{2}\hat{\mathbb{E}}_{s\sim\rho^\pi}\left[\|V_\phi(s) - R\|^2\right], \tag{254}$$

where R is a sampled return at s, and $\hat{\mathbb{E}}$ is the sample mean for s drawn from ρ^π. This loss function is the same as that of PPO. For the quantile function, the loss function is composed of two losses. The first one is the value parameter loss for cost, defined as

$$L_{value}(\psi) = \frac{1}{2}\hat{\mathbb{E}}_{s\sim\rho^\pi}\left[\left\|\frac{1}{n_q}\sum_{i=1}^{n_q} q_{\psi,u_i}(s) - C\right\|^2\right], \tag{255}$$

where C is the sampled cumulative sum cost at s. Note that value function for cost is computed as $C^\pi(s) := \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t c(s_t, a_t)\right] = \int_0^1 q_u^\pi(s)\, du \approx \frac{1}{n_q}\sum_{i=1}^{n_q} q_{\psi,u_i}(s)$. The second loss for the quantile function is the quantile loss $l_{Huber,u_i}(x)$ (for definition, please see Appendix A.1) with the Huber loss $L_\kappa(x)$, defined as

$$L_{quant}(\psi) = \frac{1}{n_q^2}\sum_{i,j=1}^{n_q}\hat{\mathbb{E}}_{(s,a,s')\sim\pi}\left[l_{Huber,u_i}(\delta_{ij}(s,a,s'))\right]$$
$$\delta_{ij}(s,a,s') = c(s,a) + \gamma q_{\psi_{old},u_j}(s') - q_{\psi,u_i}(s),$$

where $(s,a,s') \sim \pi$ means that $s \sim \rho^\pi(\cdot)$, $a \sim \pi(\cdot|s)$, and $s' \sim M(\cdot|s,a)$, ψ_{old} is a copied parameter of ψ which does not update when ψ updates. Thus, the loss function for the quantile function parameter ψ is given by

$$L(\psi) = L_{value}(\psi) + L_{quant}(\psi). \tag{256}$$

The parameters ξ and ζ for estimating the Weibull distribution parameters $\alpha_\xi(s)$ and $\beta_\zeta(s)$ at state s are updated by minimizing the following loss function:

$$L(\xi,\zeta) = \hat{\mathbb{E}}_{s\sim\rho^\pi}\left[\frac{1}{k}\sum_{i=n_q-k+1}^{n_q}\frac{1}{2}\left\|\log\beta_\zeta(s) + \frac{\log c_{u_i}}{\alpha_\xi(s)} - \log q_{\psi,u_i}(s)\right\|^2\right], \tag{257}$$

[5] `https://github.com/wyjung0625/QCPO`, (MIT License)

[6] `https://github.com/astooke/rlpyt/tree/master/rlpyt/projects/safe`, (MIT License)

where $c_u = -\log(1-u)$. Note that the u-quantile of Weibull distribution with parameters α and β is $\beta \cdot (c_u)^{1/\alpha}$. Thus, (257) is the mean square error of log-scale of the u-quantile for $u \in \{u_{n_q-k+1}, \ldots, u_{n_q}\}$.

E.3 Policy Loss Function

As aforementioned in Section 4, the basic policy loss function of QCPO for a given Lagrange multiplier λ is

$$L^{\pi_{old}}(\pi_\theta) - \tilde{C}_1 \max_s \mathrm{KL}(\pi_{old}(\cdot|s) \parallel \pi_\theta(\cdot|s)) \tag{258}$$

where

$$\begin{aligned}
L^{\pi_{old}}(\pi_\theta) &= \left(V^{\pi_{old}}(s_0) - \lambda q^{\pi_{old}}_{1-\epsilon_0}(s_0)\right) + \mathbb{E}_{\pi_{old}}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{a\sim\pi_\theta}\left[A^{\pi_{old}}_r(s_t,a) - \lambda A^{\pi_{old}}_{1-\epsilon_0}(s_t,a)\right]\right] \\
&= \left(V^{\pi_{old}}(s_0) - \lambda q^{\pi_{old}}_{1-\epsilon_0}(s_0)\right) + \mathbb{E}_{s\sim\rho^{\pi_{old}}, a\sim\pi_\theta}\left[A^{\pi_{old}}_r(s_t,a) - \lambda A^{\pi_{old}}_{1-\epsilon_0}(s_t,a)\right] \quad (259) \\
A^{\pi_{old}}_r(s,a) &= r(s,a) + \gamma \mathbb{E}_{s'\sim M(\cdot|s,a)}\left[V^{\pi_{old}}(s')\right] - V^{\pi_{old}}(s) \quad (260) \\
A^{\pi_{old}}_{1-\epsilon_0}(s,a) &= c(s,a) + \tilde{c}^{\pi_{old}}_{1-\epsilon_0}(s,a) + \gamma \mathbb{E}_{s'\sim M(\cdot|s,a)}\left[q^{\pi_{old}}_{1-\epsilon_0}(s')\right] - q^{\pi_{old}}_{1-\epsilon_0}(s), \quad (261)
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}_{\pi_{old}}\left[A^{\pi_{old}}_{1-\epsilon_0}(s,a)\right] &= \mathbb{E}_{\pi_{old}}\left[c(s,a) + \tilde{c}^{\pi_{old}}_{1-\epsilon_0}(s,a) + \gamma q^{\pi_{old}}_{1-\epsilon_0}(s') - q^{\pi_{old}}_{1-\epsilon_0}(s)\right] \quad (262) \\
&= \mathbb{E}_{\pi_{old}}\left[\frac{p_{X^{\pi_{old}}(s')}\left(\frac{q^{\pi_{old}}_{1-\epsilon_0}(s) - c(s,a)}{\gamma}\right)}{\gamma \cdot p_{X^{\pi_{old}}(s)}\left(q^{\pi_{old}}_{1-\epsilon_0}(s)\right)}\left\{c(s,a) + \gamma q^{\pi_{old}}_{1-\epsilon_0}(s') - q^{\pi_{old}}_{1-\epsilon_0}(s)\right\}\right] \quad (263)
\end{aligned}$$

since

$$\tilde{c}^{\pi_{old}}_u(s,a) = \left(\frac{p_{X^{\pi_{old}}(s')}\left(\frac{q^{\pi_{old}}_{1-\epsilon_0}(s) - c(s,a)}{\gamma}\right)}{\gamma \cdot p_{X^{\pi_{old}}(s)}\left(q^{\pi_{old}}_{1-\epsilon_0}(s)\right)} - 1\right)\left\{c(s,a) + \gamma q^{\pi}_{1-\epsilon_0}(s')\right\} \tag{264}$$

$$1 = \mathbb{E}_{\pi_{old}}\left[\frac{p_{X^{\pi_{old}}(s')}\left(\frac{q^{\pi_{old}}_{1-\epsilon_0}(s) - c(s,a)}{\gamma}\right)}{\gamma \cdot p_{X^{\pi_{old}}(s)}\left(q^{\pi_{old}}_{1-\epsilon_0}(s)\right)}\right] \tag{265}$$

Therefore we use $\frac{p_{X^{\pi_{old}}(s')}\left(\frac{q^{\pi_{old}}_{1-\epsilon_0}(s) - c(s,a)}{\gamma}\right)}{\gamma \cdot p_{X^{\pi_{old}}(s)}\left(q^{\pi_{old}}_{1-\epsilon_0}(s)\right)}\left\{c(s,a) + \gamma q^{\pi_{old}}_{1-\epsilon_0}(s') - q^{\pi_{old}}_{1-\epsilon_0}(s)\right\}$ as the advantage for the $(1-\epsilon_0)$-quantile.

$$A^{\pi_{old}}_{1-\epsilon_0}(s,a) = \frac{p_{X^{\pi_{old}}(s')}\left(\frac{q^{\pi_{old}}_{1-\epsilon_0}(s) - c(s,a)}{\gamma}\right)}{\gamma \cdot p_{X^{\pi_{old}}(s)}\left(q^{\pi_{old}}_{1-\epsilon_0}(s)\right)}\left\{c(s,a) + \gamma q^{\pi_{old}}_{1-\epsilon_0}(s') - q^{\pi_{old}}_{1-\epsilon_0}(s)\right\} \tag{266}$$

Finally QCPO is based on PPO[20], the actual policy loss function is as follows:

$$L(\theta) = -\hat{\mathbb{E}}_{\pi_{\theta_{old}}}\left[\min\left\{\hat{A}_1(s,a,s'), \hat{A}_2(s,a,s')\right\}\right] \tag{267}$$

where

$$\hat{A}_1(s,a,s') = \mathrm{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}, 1 - r_{clip}, 1 + r_{clip}\right) \times \hat{A}(s,a,s') \tag{268}$$

$$\hat{A}_2(s,a,s') = \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}\hat{A}(s,a,s') \tag{269}$$

$$\hat{A}(s,a,s') := \hat{A}_r(s,a,s') - \lambda \hat{A}_{q,1-\epsilon}(s,a,s') \tag{270}$$

$$\hat{A}_r(s,a,s') := r(s,a) + \gamma V_{\phi_{old}}(s') - V_{\phi_{old}}(s) \tag{271}$$

$$\hat{A}_{1-\epsilon_0}(s,a,s') = \frac{p_{\alpha_{old}(s'),\beta_{old}(s')}\left(\frac{q_{\psi_{old},1-\epsilon_0}(s)-c(s,a)}{\gamma}\right)}{\gamma \cdot p_{\alpha_{old}(s'),\beta_{old}(s')}\left(q_{\psi_{old},1-\epsilon_0}(s)\right)} \left\{c(s,a) + \gamma q_{\psi_{old},1-\epsilon_0}(s') - q_{\psi_{old},1-\epsilon_0}(s)\right\} \tag{272}$$

Here,

$$p_{\alpha_{old}(s'),\beta_{old}(s')}(x) = \frac{\alpha_{\xi_{old}}(s')}{\beta_{\zeta_{old}}(s')}\left(\frac{x}{\beta_{\zeta_{old}}(s')}\right)^{\alpha_{\xi_{old}}(s')-1} \exp\left(-\left(\frac{x}{\beta_{\zeta_{old}}(s')}\right)^{\alpha_{\xi_{old}}(s')}\right) \tag{273}$$

is the probability density function (PDF) of the approximated weibull distribution with parameter $\alpha_{\xi_{old}}(s')$ and $\beta_{\zeta_{old}}(s')$.

However, the variance of the ratio $\frac{p_{\alpha_{old}(s'),\beta_{old}(s')}\left(\frac{q_{\psi_{old},1-\epsilon_0}(s)-c(s,a)}{\gamma}\right)}{\gamma \cdot p_{\alpha_{old}(s'),\beta_{old}(s')}\left(q_{\psi_{old},1-\epsilon_0}(s)\right)}$ is large with actual samples. Hence, for implementation, using the Taylor series $\log x = (x-1) + \frac{1}{2}(x-1)^2 + \cdots$ around $x = 1$, we smooth the weight as

$$\begin{aligned}
&\frac{p_{\alpha_{old}(s'),\beta_{old}(s')}\left(\frac{q_{\psi_{old},1-\epsilon_0}(s)-c(s,a)}{\gamma}\right)}{\gamma \cdot p_{\alpha_{old}(s'),\beta_{old}(s')}\left(q_{\psi_{old},1-\epsilon_0}(s)\right)} \\
&\approx \left(1 + \text{clip}\left(\log \frac{p_{\alpha_{old}(s'),\beta_{old}(s')}\left(\frac{q_{\psi_{old},1-\epsilon_0}(s)-c(s,a)}{\gamma}\right)}{\gamma \cdot p_{\alpha_{old}(s'),\beta_{old}(s')}\left(q_{\psi_{old},1-\epsilon_0}(s)\right)}, -c_{\text{clip}}, c_{\text{clip}}\right)\right)
\end{aligned} \tag{274}$$

and apply (274) into (272) so the actual advantage estimate we used is

$$\begin{aligned}
&\hat{A}_{1-\epsilon_0}(s,a,s') \\
&= \left(1 + \text{clip}\left(\log \frac{p_{\alpha_{old}(s'),\beta_{old}(s')}\left(\frac{q_{\psi_{old},1-\epsilon_0}(s)-c(s,a)}{\gamma}\right)}{\gamma \cdot p_{\alpha_{old}(s'),\beta_{old}(s')}\left(q_{\psi_{old},1-\epsilon_0}(s)\right)}, -c_{\text{clip}}, c_{\text{clip}}\right)\right) \\
&\quad \times \left\{c(s,a) + \gamma q_{\psi_{old},1-\epsilon_0}(s') - q_{\psi_{old},1-\epsilon_0}(s)\right\}
\end{aligned} \tag{275}$$

E.4 Lagrange Multiplier for Quantile Constraint

We also need the Lagrange multiplier λ in (270) for the policy loss function to satisfy the quantile constraint. The Lagrange multiplier is updated to minimize the Lagrange form of (QuantCP) $L_{quant}(\pi,\lambda) := V^\pi(s_0) - \lambda\left(q^\pi_{1-\epsilon}(s_0) - d_{th}\right)$ to satisfy the quantile constraint $q^\pi_{1-\epsilon}(s_0) \leq d_{th}$. Thus, the update rule of the Lagrange multiplier λ is $\lambda \leftarrow \max\{\lambda + \eta(q^\pi_{1-\epsilon}(s_0) - d_{th}), 0\}$, where η is a learning rate. To constrain the outage probability of the sum of costs in a trajectory, we collect 100 trajectories, and compute the $(1-\epsilon)$-quantile of them to replace $q^\pi_{1-\epsilon}(s_0)$ in the Lagrange update rule.

E.5 Hyper-parameters

For the quantile network, we used $n_q = 25$, and $u_i = \frac{2i-1}{2n_q} = \frac{2i-1}{50}$, $i = 1, \ldots, n_q (= 25)$ for $\{u_1, u_2, \ldots, u_{n_q}\}$. For training the Weibull network, we used the rightmost k-quantiles among n_q, and the k is 8 ($\approx$ 30% of n_q quantiles). The discount factor γ is 0.99, and all learning rates for Adam optimizers for all parameters are 10^{-4}. The η for updating the Lagrange multiplier is 0.1. The r_{clip} in (268) for updating the policy parameter is 0.1, and the c_{clip} in (275) for computing $\tilde{c}^{\pi'}_u(s,a)$ is 0.5. Since PPO is an on-policy algorithm, it first collects 12000 samples by interaction with its environment. Then, it reshapes these samples by 120 sub-trajectories of length 100, and uses all sub-trajectories to update its parameters (this is because we use LSTM for the feature extraction network). This update is performed 8 times for the same collected sub-trajectories, then we remove them and collect new 12000 samples by interaction with the environment. This procedure is performed until the maximum training timesteps 5×10^6.