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

Anonymous Author(s)
Affiliation
Address
email

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 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 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 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 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

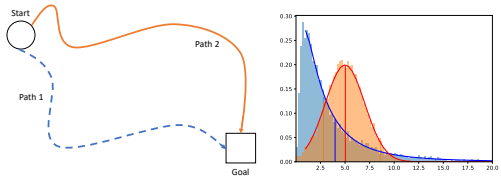
$$\begin{array}{ll} \text{Maximize} & V^\pi(s_0) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\ \text{subject to} & C^\pi(s_0) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] \leq d_{th}, \end{array} \tag{ExpCP}$$

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 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 cost, we can set d_{th} properly as we desire. Most of the previous constrained RL works solved the problem (ExpCP) [1, 21, 20, 9, 17, 8, 18] partly because the constraint on the expectation of the 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

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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 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 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 the probability of 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.

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

Our approach to this problem is first to convert the outage probability constraint in (ProbCP) into a quantile constraint $q^{\pi}_{1-\epsilon_0}(s_0) \leq d_{th}$, which is equivalent to the original probabilistic constraint (details are in Section 3), and then to solve the equivalent 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 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.) policy parameter due to the quantile term $q^{\pi}_{1-\epsilon_0}(s_0)$. Therefore, we derive theoretical results to approximate 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 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 quantile and obtains an (approximate) policy gradient for quantile, and this is one of the main contributions of this paper together with the QCPO algorithm.

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 of cost $\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)$. If one wants to constrain the average cost of the policy π, then the constraint $C^\pi(s_0) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right] \le d_{th}$ can be considered. On the other hand, if one wants to constrain the outage behavior of the policy π, then the constraint $\Pr\left(X^\pi(s_0) > d_{th}\right) \le \epsilon_0$ 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 cumulative sum cost: $C^\pi(s_0) \le d_{th}$ [1, 21, 20, 9, 17, 8, 18]. In order to solve the constrained optimization problem with this expectation-based constraint, researchers considered Lagrangian multiplier methods [1, 14], Lyapunov-based methods [4, 5], projection-based methods [20], safety-layer methods [6].

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[11]. In the context of RL, the definitions of quantile and 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 $\text{CVaR}_u^\pi(s_0) := \mathbb{E}_\pi\left[X^\pi(s_0) \mid X^\pi(s_0) > q_u^\pi(s_0)\right]$, respectively. CVaR was previously used in RL to constrain undesirable events. Chow et al. [3] 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. [19] proposed an off-policy algorithm to satisfy a CVaR constraint, and we refer to the constrained problem as (CVaR-CP). The explicit formulation of the CVaR-CP problem is given in Appendix A.3. They only estimated the mean and variance of the cost distribution using a technique in distributional RL and compute $\text{CVaR}_u^\pi(s)$[1] as CVaR of 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 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 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 CVaR and 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 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 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 the decay of the tail probability. Table 1 summarizes the previous constrained RL methods and this paper.

Table 1: Comparison of constrained RL with probabilistic constraints

Papers	Algorithm Type	Distribution Modeling	Theory	Deep RL
Risk-Const. RL [3]	On-policy (Trajectory-based)	No (need only trajectory samples)	Yes	No
WCSAC [19]	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

Large Deviation Principle (LDP) Large deviation principle (LDP) [7] is a technique for estimating the limiting behavior of a sequence of distributions especially on 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 $I(x)$.

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

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 [7]. In this paper, we consider the tail probability of the distribution of the cumulative sum of costs $X^\pi(s_0) = \sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)$. 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 cumulative sum costs: $X^\pi(s) = \sum_{t=0}^{\infty} \gamma^t c(S_t, A_t)$ 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 $\Pr\left[X^\pi(s_0) > d_{th}\right] < \epsilon_0$ due to the definition of 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 the cost, the policy gradient theorem [15] can directly be applied, and the gradient 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 $\rho^\pi(s) := \sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s | s_0, \pi)$, $A^\pi_r(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim M}\left[V^\pi(s')\right] - V^\pi(s)$, and $A^\pi_c(s, a) := c(s, a) + \gamma \mathbb{E}_{s' \sim M}\left[C^\pi(s')\right] - C^\pi(s)$. However, the gradient of the Lagrangian in the equivalent problem of (QuantCP)

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

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

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

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

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

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:

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

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

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 under a mild assumption, 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. We assume that the CDF $F_{X^\pi(s)}(x)$ is strictly increasing on $[0,\infty)$ and it is continuously differentiable for all $s \in \mathcal{S}$, and 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.

Theorem 1. *Under Assumptions 2 and 3 in Appendix B.1 stating the boundedness of quantile and smoothness of the CDF of $X^\pi(s_t)$, the u-quantile of random variable $X^\pi(s_t)$ satisfies the following temporal-difference(TD) relation. For some constant R and small $\epsilon > 0$,*

$$\left|\mathbb{E}_\pi\left[\mu_u^\pi\left(s_t,a_t,s_{t+1}\right)\left\{c(s_t,a_t)+\gamma q_u^\pi(s_{t+1})-q_u^\pi(s_t)\right\}\right]\right| \leq \frac{\epsilon}{1-\epsilon}R, \tag{7}$$

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

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

$$\mathbb{E}_\pi\left[c(s_t,a_t)+\gamma C^\pi(s_{t+1})-C^\pi(s_t)\right]=0 \tag{8}$$

by the Bellman equation for the sum cost. The TD relation (8) for expectation has a similar form to that for the u-quantile (7), but the difference is that (7) 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 (7) involves two quantities: 1) a target u-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 (7) is the probability of the event that the 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 the same assumptions as Theorem 1, the u-quantile $q_u^\pi(s_t)$ of random variable $X^\pi(s_t)$ is bounded as*

$$\left|q_u^\pi(s_t)-\mathbb{E}_\pi\left[\mu_u^\pi\left(s_t,a_t,s_{t+1}\right)\left\{c(s_t,a_t)+\gamma q_u^\pi(s_{t+1})\right\}\right]\right| \leq \frac{\epsilon}{1-\epsilon}R. \tag{9}$$

Proof: Note that the term $q_u^\pi(s_t)$ can go outside the expectation in (7) since the expectation is over (a_t,s_{t+1}). From eq. (3) in Appendix A, the expectation of the numerator of $\mu_u^\pi(s_t,a_t,s_{t+1})$ 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 the same assumptions as Theorem 1, the u-quantile $q_u^\pi(s_0)$ of random variable $X^\pi(s_0)$ is*

expressed as

$$\left| q_u^\pi(s_0) - \mathbb{E}_{\tilde{\pi}_u}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right]\right| \leq \frac{\epsilon R}{(1-\epsilon)(1-\gamma)}, \tag{10}$$

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

Now, plugging (10) 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 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. *Suppose that the state transition dynamics are deterministic, i.e.,* $s_{t+1} = h(s_t,a_t)$. *Then, the same assumptions as Theorem 1,* $q_u^\pi(s)$ *can be expressed as*

$$\left| q_u^\pi(s_0) - \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t)+\tilde{c}_u^\pi(s_t,a_t)\right\}\right]\right| \leq \frac{\epsilon R}{(1-\epsilon)(1-\gamma)}, \tag{11}$$

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

Note that the additional cost $\tilde{c}_u^\pi(s_t,a_t)$ in Theorem 2 is a policy-dependent cost function. Under a mild assumption, we can find an upper bound of (11) 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 $\max_s \mathrm{KL}\left(\pi'(\cdot|s)\,\middle\|\,\pi(\cdot|s)\right) \leq \delta$ for some $\delta > 0$. The assumption is as follows:

Assumption 1. *For any given fixed* $u \in (0,1)$ *and any policies* π *and* π'*, there exists a coefficient* C_u *such that*

$$\left|\tilde{c}_u^{\pi'}(s,a) - \tilde{c}_u^\pi(s,a)\right| \leq C_u \cdot \max_{s'} KL\left(\pi'(\cdot|s')\,\middle\|\,\pi(\cdot|s')\right), \quad \forall s\in\mathcal{S}, a\in\mathcal{A}. \tag{12}$$

Basically, Assumption 1 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 1 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* $s_{t+1} = h(s_t,a_t)$*, Assumption 1, and the same assumptions as Theorem 1, 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* $\max_s KL(\pi'(\cdot|s)\,\|\,\pi(\cdot|s)) \leq \delta$*:*

$$\left| q_u^\pi(s_0) - \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left\{c(s_t,a_t)+\tilde{c}_u^{\pi'}(s_t,a_t)\right\}\right]\right| \leq \frac{\epsilon R}{(1-\epsilon)(1-\gamma)} + \frac{C_u}{1-\gamma}\delta. \tag{13}$$

By Theorem 3, we can approximate the u-quantile $q_u^\pi(s_0)$ as the expectation of 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. 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 [13]. 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 policy which has a similar form to that in [12]. Then, we provide a condition for policy improvement for the proposed method. The implementation of the proposed algorithm is based on the implementation of [14] and details of the implementation including network structure, loss function, Lagrange update method, and hyper-parameters are in Appendix E.

4.1 Overall Structure of QCPO

The agent of QCPO uses function approximators for policy π, value function $V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$ and quantile function $q^\pi(s) = \left[q^\pi_{u_1}(s), q^\pi_{u_2}(s), \ldots, q^\pi_{u_{n_q}}(s)\right]$ of $X^\pi(s) = \sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)$ for 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 u_i-quantile $q^\pi_{u_i}(s)$ 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 $I_{X^\pi(s)}(x) \approx (x/\beta(s))^{\alpha(s)}$ 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 with parameters ξ and ζ, respectively. Fig. 2 shows both 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 [19] on the PDF of $X^\pi(s)$ fixes the rate function as quadratic x^2, which is not the correct rate function in general.

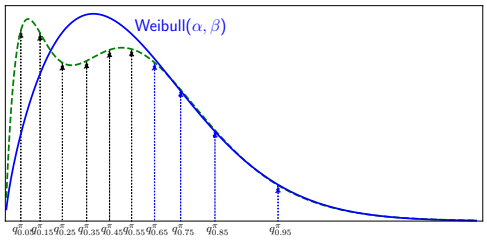


Figure 2: Illustration of quantile approximation and tail-probability approximation. Green-dash curve is the unknown PDF $p_{X^\pi(s)}(x)$ and up-arrow points are the estimated u-quantiles $q^\pi_u(s)$. We approximate the PDF $p_{X^\pi(s)}(x)$ on the right tail by a Weibull distribution (blue curve) and this is approximated using the 4 rightmost quantile points (blue arrow).

The overall procedure of QCPO is as follows: 1) estimate value function $V^\pi(s)$ for return and estimate quantile function $q^\pi_u(s)$, $u \in \left[u_1, u_2, \ldots, u_{n_q}\right]$ for cumulative sum costs, 2) approximate tail distribution $p_{X^\pi(s)}(x)$ on the right tail using weibull distribution with parameter $\alpha(s)$, $\beta(s)$, 3) compute additional cost for the base policy π', $\tilde{c}^{\pi'}_{1-\epsilon_0}(s,a)$ for 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 value advantage and quantile advantage $A_r(s,a) - \lambda A^\pi_{1-\epsilon_0}(s,a)$, 4) the Lagrange multiplier λ update. Since QCPO is based on PPO [13], the loss functions for policy and value function are similar to those of PPO [13]. Please see in 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

$$L^{\pi_{old}}(\pi_\theta) - \tilde{C}_1 \max_s \mathrm{KL}(\pi_{old}(\cdot|s) \parallel \pi_\theta(\cdot|s)) \tag{14}$$

where

$$L^{\pi_{old}}(\pi_\theta) = \left(V^{\pi_{old}}(s_0) - \lambda q^{\pi_{old}}_{1-\epsilon_0}(s_0)\right) + \mathbb{E}_{s\sim\rho^{\pi_{old}}, a\sim\pi_\theta}\left[A^{\pi_{old}}_r(s_t,a) - \lambda A^{\pi_{old}}_{1-\epsilon_0}(s_t,a)\right] \tag{15}$$

$$A^{\pi_{old}}_r(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim M(\cdot|s,a)}\left[V^{\pi_{old}}(s')\right] - V^{\pi_{old}}(s) \tag{16}$$

$$A^{\pi_{old}}_{1-\epsilon_0}(s,a) = c(s,a) + \tilde{c}^{\pi_{old}}_{1-\epsilon_0}(s,a) + \gamma\mathbb{E}_{s'\sim M(\cdot|s,a)}\left[q^{\pi_{old}}_{1-\epsilon_0}(s')\right] - q^{\pi_{old}}_{1-\epsilon_0}(s), \tag{17}$$

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 (14). 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*

$$L^{\pi_{old}}(\pi_\theta) - \tilde{C}_1 \max_s KL(\pi_{old}(\cdot|s) \parallel \pi_\theta(\cdot|s)) \tag{18}$$

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

$$L_{quant}(\pi_{new}, \lambda) - L_{quant}(\pi_{old}, \lambda) \tag{19}$$

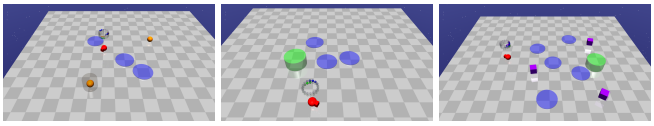
$$\geq L^{\pi_{old}}(\pi_{new}) - L^{\pi_{old}}(\pi_{old}) - \tilde{C}_1 KL_{max}(\pi_{old}||\pi_{new}) - \underbrace{\tilde{C}_2 \frac{\epsilon}{1-\epsilon}}_{\textit{approximation loss}} \tag{20}$$

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

Note that the term $\tilde{C}_2 \frac{\epsilon}{1-\epsilon}$ in (20) 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 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 3: The considered environments

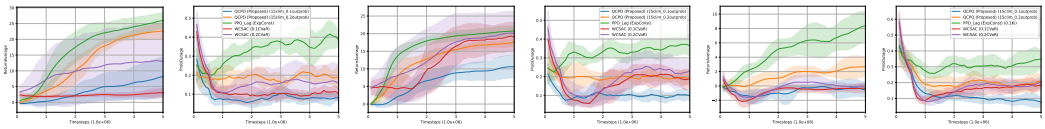
We examined the performance of the proposed QCPO and compared it to that of WCSAC which uses the CVaR constraint. The environments we considered are SimpleButtonEnv, DynamicEnv [19] and GremlinEnv, which are based on Safety Gym [10], MuJoCo [16], and OpenAI Gym [2]. The environments can be considered as simplified versions of a real environment of automatic serving robot, and are illustrated in Fig. 3. The goal of these environments are 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 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 higher return when the robot touches more goals in maximum timesteps $T = 1000$, and higher sum of costs when the robot touches one of the other objects more often. 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 Lagrangian multiplier method (PPO_Lag)[2] for (ExpCP) and that of WCSAC [19][3] for (CVaR-CP) which is

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

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



(a) Avg. Return (b) Outage Prob. (c) Avg. Return (d) Outage Prob. (e) Avg. Return (f) Outage Prob.

Figure 4: (a) and (b) SimpleButtonEnv, (c) and (d) DynamicEnv, and (e) and (f) GremlinEnv: (a), (c), and (e) - average return of the most current 100 episodes, and (b), (d), (f) - outage probability of the most current 100 episodes

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. 4 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 the threshold $d_{th} = 15$ well (please see the graph in Appendix D), and its outage probability becomes around 0.35 as we can observe in Fig. 4b, 4d, and 4f. 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 lower or similar outage probability to the threshold ϵ_0 in Fig 4b, but the algorithm does not satisfy the outage probability constraint exactly in Fig. 4d and 4f. 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 ($\epsilon_0 = 0.2$ (orange), $\epsilon_0 = 0.1$ (blue)) maintains the outage probability around the desired target outage probability very well, as shown in Fig. 4b, 4d, and 4f.

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. 4b, 4d, and 4f, PPO_Lag induces the highest outage probability, so it has the highest average return, as shown in Fig. 4a, 4c, and 4e. The direct comparison between WCSAC and QCPO is less meaningful in 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. 4a, QCPO achieves 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 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 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] 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.

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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`.

[21] 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 the supplemental material.
   (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 6 and 8.
   (b) Did you mention the license of the assets? [Yes] Please see footnotes in page 6 and 8.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] New assets are in the supplemental material.
   (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]