TOWARDS SAFE REINFORCEMENT LEARNING VIA CONSTRAINING CONDITIONAL VALUE-AT-RISK

Anonymous authors
Paper under double-blind review

ABSTRACT

Though deep reinforcement learning (DRL) has obtained substantial success, it may encounter catastrophic failures due to the intrinsic uncertainty caused by stochasticity in both environments and policies. Existing safe reinforcement learning methods are often based on transforming the optimization criterion and adopting the variance of the return as a measure of uncertainty. However, the return variance introduces a bias for penalizing both positive and negative risk equally, deviated from the purpose of safe reinforcement learning to penalize negative ones only. To address this issue, we propose to use the conditional value-at-risk (CVaR) as an assessment of risk, which guarantees that the probability for reaching a catastrophic state is below a desired threshold. Furthermore, we present a novel reinforcement learning framework of CVaR-Proximal-Policy-Optimization (CPPO) which formalizes the risk-sensitive constrained optimization problem by keeping its CVaR under a given threshold. To evaluate the robustness of policies, we theoretically prove that performance degradation under observation disturbance and transition disturbance depends on the gap of value function between the best state and the worst state. We also show that CPPO can generate more robust policies under disturbance. Experimental results show that CPPO achieves higher cumulative reward and exhibits stronger robustness against observation disturbance and transition disturbance on a series of continuous control tasks in MuJoCo.

1 INTRODUCTION

Deep reinforcement learning (DRL) has achieved enormous success on a variety of tasks, ranging from playing Atari games (Mnih et al., 2013; 2015; 2016), Go (Silver et al., 2016) to manipulating complex robotics in the real world (Kendall et al., 2019). However, due to the intrinsic stochasticity in both environments and policies, these methods may result in catastrophic failures (Heger, 1994; Coraluppi & Marcus, 1999) and the agent may receive significantly negative outcomes. Several factors can be associated with this phenomenon. One is that traditional DRL only aims at cumulative reward maximization without considering the stochasticity of the environment (Garcıa & Fernández, 2015), which may lead to serious consequences with a certain probability and expose our policies to risk. This can be illustrated briefly in the case of self-driving, where the agent might try to achieve the highest reward by acting dangerously, e.g. agents may drive along the edge of a curve for reaching the end with a short time ignoring the danger in driving. Also, the usage of deep neural networks to construct complicated mappings from a high-dimensional state space $\mathcal{S}$ to an action space $\mathcal{A}$ in DRL algorithms can make them vulnerable to adversarial attacks (Huang et al., 2017).

Various efforts have been made on safe reinforcement learning (safe RL) (Heger, 1994; Coraluppi & Marcus, 1999; Garcıa & Fernández, 2015). Garcıa & Fernández (2015) conduct a comprehensive survey on safe RL and argue that an array of methods in this area are based on transforming the optimization criterion by considering the risk of the return. For example, $\hat{Q}$-$Learning$ (Heger, 1994) uses a lower bound to estimate the Q-target in Q-$Learning$ for avoiding the risk; and Geibel & Wysotzki (2005) propose the *expected-value-minus-variance-criterion* that subtracts the variance of the return from the cumulative reward. However, due to the consideration of the worst-case outcomes (Heger, 1994), one major drawback of those transformed optimization criteria is that they may lead to overly pessimistic policies, which will focus too much on the worst case and own poor average performance. Moreover, although variance is a standard measure of the risk of the policy (Gosavi, 2009; Tamar et al., 2012), it does not distinguish between positive and negative risk

and penalizes both equally (Szegö, 2002), deviated form the purpose of safe RL to only penalize negative ones.

To address the shortcomings of the worst-case outcomes as well as the variance of the return used in previous objective-modification methods of safe RL (Garcıa & Fernández, 2015; Geibel & Wysotzki, 2005; Heger, 1994), we propose to use conditional value-at-risk (CVaR) for evaluating the risk of policies. CVaR is a well established metric in economic uncertainty analysis (Alexander & Baptista, 2004; Alexander et al., 2006) and captures the expectation of the random variable to be an outlier with a given threshold. Unlike variance, CVaR can only capture negative trajectories with relatively low return. Considering to use CVaR to capture the respectively low return of the trajectory, we naturally propose to improve the robustness of the on-policy algorithms by CVaR. By integrating CVaR with Proximal Policy Optimization (PPO) (Schulman et al., 2017), we present a new algorithm called CVaR-Proximal-Policy-Optimization (CPPO) and notionally analyze policies' robustness against different kinds of disturbance. We show that although the observation disturbance and transition disturbance are structurally different, the performance degradation resulted from each of them is theoretically dependent on the *Value Function Range* (VFR), which is introduced as the value function gap between the best and worst states in this paper. We further show that CPPO can improve the robustness of policies against observation disturbance and transition disturbance since CVaR can control the value function of states with relatively low value and further control the VFR value. Empirically, we compare CPPO to multiple on-policy baselines as well as some previous CVaR-based methods on various continuous control tasks in MuJoCo (Todorov et al., 2012). Our results show that CPPO achieves higher cumulative reward in the training stage and exhibits stronger robustness when we apply perturbations to these environments.

In summary, our contributions are:

- We analyze the advantages of choosing CVaR as the metric for evaluating the risk of policy compared with the worst-case outcome as well as the variance of the return. Furthermore, we propose a constrained optimization problem in order to maximize the cumulative reward as well as controlling the risk, which can be solved by our CPPO algorithm;
- We theoretically analyze the performance of trained policies under observation and transition disturbance, and build a theoretical connection of these two types of structurally different disturbance. This analysis indicates that our CPPO can improve the robustness of policies;
- We empirically demonstrate that our method exhibits stronger robustness under observation/transition perturbations than other common on-policy RL algorithms and previous CVaR-based RL algorithms in MuJoCo simulator.

2 BACKGROUND

In this section, we briefly introduce safe reinforcement learning (safe RL) and conditional value-at-risk (CVaR), which motivate us to adopt CVaR as a metric of risk in safe RL.

2.1 SAFE RL

In standard RL setting, the agent interacts with an unknown environment and learns to achieve the highest long-term return. The task is modeled as a Markov decision process (MDP) of $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ represent the state space and the action space, respectively; $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ denotes the transition probability that captures the dynamics of the environment; $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$ represents the reward function; and γ is a discount factor. We use π_θ to represent the policy of the agent with parameter θ, which is a mapping from $\mathcal{S}$ to $\mathcal{A}$. At any time step t, the agent perceives current state $s_t \in \mathcal{S}$, chooses its action $a_t \in \mathcal{A}$ sampled from the distribution $\pi_\theta(\cdot|s_t)$ and obtains a reward r_t. All these timesteps consist of a trajectory $\tau = (s_0, a_0, r_i, s_1, a_1, ...)$. Given an MDP $\mathcal{M}$, the goal of RL is to find the optimal policy π_{θ^*} with the highest expected cumulative reward as

$$\max_\theta J(\pi_\theta) \triangleq \mathbb{E}\left[D(\pi_\theta) \triangleq \sum_{t=1}^{\infty} \gamma^t r_t \Big| \pi_\theta\right], \tag{1}$$

where $D(\pi_\theta)$ represents the return of the policy π_θ, and $J(\pi_\theta)$ is the expectation of $D(\pi_\theta)$.

However, problem (1) only focuses on cumulative reward without considering the risk of the policy, which may cause catastrophic results (Heger, 1994; Coraluppi & Marcus, 1999). To address this problem, an array of safe RL methods tend to change the objective in problem (1) in order to eliminate the uncertainty and avoid the danger. In general, uncertainty can be categorized into two types, namely, inherent uncertainty and parameter uncertainty (Garcıa & Fernández, 2015). The inherent uncertainty of RL refers to the transition dynamics in MDP. For example the agent might end up in completely different situations when repeating its actions from the same starting state. Previous works (Heger, 1994; Gaskett, 2003) choose the worst-case criterion to address the issue as

$$\max_{\theta} J_{inh}(\pi_\theta) \triangleq \max_{\theta} \min_{\tau \sim \pi_\theta} \left[D(\tau) \triangleq \sum_{t=1}^{\infty} \gamma^t r_t \right]. \tag{2}$$

As a counterpart of Q-Learning, Heger (1994) proposes $\hat{Q}$-$Learning$ with the implementation of (2), and Gaskett (2003) presents β-$pessimistic\ Q$-$Learning$, which adds a parameter β to control the pessimistic level.

There are also various studies that assess the effectiveness of variance for acquiring safe policies (Sato et al., 2001; Gosavi, 2009; Tamar et al., 2012). Some previous work (Howard & Matheson, 1972) considers exponential utility function and formalizes it as the combination of cumulative reward and the variance of the return $Var(D(\pi_\theta))$ as

$$\max_{\theta} \delta^{-1} \log \mathbb{E}_\pi \left[\exp(\delta D(\pi_\theta)) \right] = \max_{\theta} \left[J(\pi_\theta) + \frac{\delta}{2} Var(D(\pi_\theta)) + O(\delta^2) \right].$$

As for the parameter uncertainty of RL, it denotes scenarios where the parameters of the MDP are unknown or there is a gap between the training and testing environments. Studies conducted by Nilim & El Ghaoui (2005) and Tamar et al. (2013) assume that the actual transition belongs to a set $\hat{\mathcal{P}}$ and consider the following problem as

$$\max_{\theta} \min_{\mathcal{P} \in \hat{\mathcal{P}}} J_{par}(\pi_\theta, \mathcal{P}) \triangleq \mathbb{E} \left[D(\pi_\theta) \triangleq \sum_{t=1}^{\infty} \gamma^t r_t \middle| \pi_\theta \right]. \tag{3}$$

However, previous safe RL methods suffer from serious drawbacks. First, both (2) and (3) are *max-min problems*, which do not have general effective solutions and usually have a high computational complexity. Second, focusing on the worst trajectories may cause over-pessimistic behaviors. For example, $\hat{Q}$-$Learning$ aims to improve the performance under the worst scenario, which can lead to extremely conservative actions (Heger, 1994). Finally, the direct usage of variance to evaluate risk is another potential concern because it will penalize not only the possibility of particularly bad trajectories, but also the good ones, yielding a drop in the agent's performance (Szegö, 2002).

2.2 CVaR

Value-at-risk (VaR) and conditional value-at-risk (CVaR) are well-established metrics for measuring risk in economy (Alexander & Baptista, 2004; Alexander et al., 2006). First, we will give the definition of VaR and CVaR (Chow & Ghavamzadeh, 2014):

Definition 1 (VaR and CVaR). *For a bounded-mean random variable Z, the value-at-risk (VaR) of confidence level $\alpha \in (0, 1)$ is defined as:*

$$\mathrm{VaR}_\alpha(Z) = \min\{z | F(z) \geq \alpha\}, \tag{4}$$

where $F(z) = P(Z \leq z)$ is the cumulative distribution function (CDF); and the condition value-at-risk (CVaR) of confidence level α is defined as the expectation of the α-tail distribution of Z as

$$\mathrm{CVaR}_\alpha(Z) = \mathbb{E}_{z \sim Z}\{z | z \geq \mathrm{VaR}_\alpha(Z)\}. \tag{5}$$

It is easy to prove that (Chow et al., 2015):

$$\lim_{\alpha \to 1^-} \mathrm{CVaR}_\alpha(Z) = \max(Z). \tag{6}$$

Previous works have attempted to use CVaR to analyze the risk-MDP, which considers cost function $\mathcal{C}$ rather than reward function $\mathcal{R}$. Chow & Ghavamzadeh (2014) and Chow et al. (2017) propose gradient-based methods like policy gradient and actor critic to optimize loss of MDP as well as keeping the CVaR under certain value. They also propose methods based on value iteration and Bellman equation to deal with the optimization of risk-MDP with CVaR (Chow et al., 2015). However, these works ignore the reward in MDP and thus cannot be directly used in RL settings.

3 METHODOLOGY

We now present our method that maximizes the expected reward while restricting the risk of the policy. We focus on increasing the agent's performance on relatively worse trajectories, which loosens the *max-min problem* to an constrained optimization problem. Moreover, we can make our policy less conservative by modifying the parameter α in CVaR. Compared with variance, CVaR is a better metric for measuring risk, because it, by definition, captures only bad trajectories.

3.1 PROBLEM FORMULATION

In standard RL, what we receive is the reward signal rather than the risk signal, thus we can only evaluate the risk of a trajectory by its return. For simplicity, we suppose there exists a decreasing smoothing function $f : \mathbb{R} \to \mathbb{R}$ with its inverse function f^{-1} and the risk of a trajectory τ is $f(D(\tau))$. For example, the most simple case is that we can use the opposite number of the return to define the risk of the trajectory, i.e. $f(D(\tau)) = -D(\tau)$.

First we propose Theorem 1 as below to calculate VaR and CVaR of $f(D(\tau))$:

Theorem 1. *For any given policy π_θ and its cumulative reward $D(\pi_\theta)$, we have:*

$$\mathrm{VaR}_\alpha(f(D(\pi_\theta))) = \min\{z|F_{D(\pi_\theta)}(f^{-1}(z)) \leq 1-\alpha\}$$
$$\mathrm{CVaR}_\alpha(f(D(\pi_\theta))) = \mathbb{E}_{z\sim D(\pi_\theta)}\{f(z)|f(z) \geq \mathrm{VaR}_\alpha(f(D(\pi_\theta))\}.$$

Specially, we consider to take the opposite number of the return of a trajectory as its risk, i.e. $f(D(\pi_\theta)) = -D(\pi_\theta)$ and we can prove that

$$-\mathrm{VaR}_\alpha(-D(\pi_\theta)) = \max\{z|F_{D(\pi_\theta)}(z) \leq 1-\alpha\},$$
$$-\mathrm{CVaR}_\alpha(-D(\pi_\theta)) = \mathbb{E}_{z\sim D(\pi_\theta)}\{z|z \leq -\mathrm{VaR}_\alpha(-D(\pi_\theta))\}.$$

Based on equation (6), we have:

$$\lim_{\alpha\to 1^-} -\mathrm{CVaR}_\alpha(-D(\pi_\theta)) = \min(D(\pi_\theta)), \tag{7}$$

and if we assume that $-\mathrm{CVaR}_\alpha(-D(\pi_\theta)) \geq \beta$, then we have:

$$P(D(\pi_\theta) \leq \beta) \leq 1-\alpha.$$

The proof of Theorem 1 is in Appendix B.1. By this theorem, we can use $-\mathrm{CVaR}_\alpha(-D(\pi_\theta))$ to represent the expected reward of the trajectories generated by π_θ with relatively lower reward.

As mentioned in Section 2.1, some safe RL objectives, such as problems (2) and (3), are intractable *max-min problems*. However, with the property in Eq. (7) of CVaR, we can equally transform problem (2) as

$$\max_\theta J_{inh}(\pi_\theta) = \max_\theta \lim_{\alpha\to 1^-} \left[-\mathrm{CVaR}_\alpha(-D(\pi_\theta))\right]. \tag{8}$$

We can further loosen problem (8) by assigning α a fixed value, which reforms the original *max-min problem* into a solvable optimization problem. Furthermore, to address the pessimism in safe RL, we balance between the standard RL objective (1) and the safe RL objective (8) after relaxation, which yields the constrained optimization problem as

$$\begin{aligned} &\max_\theta J(\pi_\theta) \\ &s.t. -\mathrm{CVaR}_\alpha(-D(\pi_\theta)) \geq \beta, \end{aligned} \tag{9}$$

where α, β are hyper-parameters and we denote the best policy of problem (9) as $\pi_c(\alpha, \beta)$.

Now we discuss some properties of $\pi_c(\alpha, \beta)$. Since $\pi_c(\alpha, \beta)$ is the optimal solution of (9) and satisfies the constraints. By Theorem 1, we naturally have

$$P(D(\pi_c(\alpha, \beta)) \leq \beta) \leq 1-\alpha,$$

which means that we can guarantee the probability of policy $\pi_c(\alpha, \beta)$ generating low-reward trajectories is below a desired threshold.

Compared with the best policy π_s of the standard RL problem (1), $\pi_c(\alpha, \beta)$ is the policy that maximizes the expected total reward in a restricted region related to hyper-parameters α, β. Obviously we have $J(\pi_c(\alpha, \beta)) \leq J(\pi_s)$. However, we can also give a lower bound of $J(\pi_c(\alpha, \beta))$ as follows:

Theorem 2. *Assume there exists a constant $M > 0$ and every trajectory $\tau = (S_0, A_0, R_1, S_1, A_1, ...)$ satisfies $\sum_{t=1}^{\infty} \gamma^t R_t \leq M$, we have*

$$J(\pi_c(\alpha, \beta)) \geq \frac{J(\pi_s) - \alpha M}{1 - \alpha}.$$

The key of the proof is to consider whether π_s satisfies our constraint and the detailed proof of Theorem 2 can be found in Appendix B.2. Therefore, although $\pi_c(\alpha, \beta)$ is in a restricted region, its expected cumulative reward will be no worse than the lower bound we prove in Theorem 2.

3.2 Optimization and Algorithm

We now simplify the constrained problem (9) to an unconstrained one. First, with properties of CVaR, we can equivalently reformulate problem (9) as

$$\begin{aligned} &\min_{\theta, \nu} -J(\pi_\theta) \\ &s.t. \ -\nu + \frac{1}{1-\alpha}\mathbb{E}[(-D(\pi_\theta) + \nu)^+] \leq -\beta. \end{aligned} \tag{10}$$

The deviation is provided in Appendix B.3. Then, by using Lagrangian relaxation method (Bertsekas, 1997), we need to solve the saddle point of the function $L(\theta, \nu, \lambda)$ as

$$\max_{\lambda \geq 0} \min_{\theta, \nu} L(\theta, \nu, \lambda) \triangleq -J(\pi_\theta) + \lambda \left(-\nu + \frac{1}{1-\alpha}\mathbb{E}[(-D(\pi_\theta) + \nu)^+] + \beta \right). \tag{11}$$

For solving problem (11), we will extend Proximal Policy Optimization (PPO) (Schulman et al., 2017) with CVaR and propose our algorithm named CVaR Proximal Policy Optimization (CPPO). In particular, the key point of Policy Gradient methods is to evaluate the gradient (Sutton et al., 2000) of the objective. Here, we use methods in (Chow & Ghavamzadeh, 2014) to compute the gradient of our objective function (11) with respected to ν, θ, λ as below:

$$\partial_\nu L(\theta, \nu, \lambda) = -\lambda + \frac{\lambda}{1-\alpha}\mathbb{E}_{\xi \sim \pi_\theta} \mathbf{1}\{\nu \geq D(\xi)\}) \tag{12}$$

$$\nabla_\theta L(\theta, \nu, \lambda) = -\mathbb{E}_{\xi \sim \pi_\theta}(\nabla_\theta \log P_\theta(\xi)) \left(D(\xi) - \frac{\lambda}{1-\alpha}(-D(\xi) + \nu)^+ \right) \tag{13}$$

$$\nabla_\lambda L(\theta, \nu, \lambda) = -\nu + \frac{1}{1-\alpha}\mathbb{E}_{\xi \sim \pi_\theta}(-D(\xi) + \nu)^+ + \beta. \tag{14}$$

The key of the deviation is to deform the objective in problem (11) as the integration of trajectories and the detailed calculation is in Appendix B.4. Moreover, with the increasing of policies' performance during training, it's unreasonable to fix β to constrain the risk of the policy. Thus we consider to modify β as a function of the risk of trajectories in the current epoch. Based on PPO and the algorithm by Chow & Ghavamzadeh (2014), we can use the gradient given above to develop an on-policy algorithm called CPPO (see Algorithm 1 in Appendix A).

4 Theoretical Analysis

In this section, we analyze the robustness of policies against observation and transition perturbations, and explain why CVaR can improve the robustness of policies.

4.1 Performance against observation Disturbance

For any MDP $\mathcal{M}$ and given policy π, we denote its expected cumulative reward and value function as $J_{\mathcal{M}}(\pi)$ and $V_{\mathcal{M},\pi}$, respectively. We define the *Value Function Range* (VFR) to capture the gap of the value function between the best state and the worst state as following.

Definition 2 (Value Function Range)**.** *For MDP $\mathcal{M}$, we define the Value Function Range (VFR) of policy π as*

$$\hat{V}_{\mathcal{M},\pi} = \max_s V_{\mathcal{M},\pi}(s) - \min_s V_{\mathcal{M},\pi}(s), \tag{15}$$

where $V_{\mathcal{M},\pi}$ is the value function (Sutton & Barto, 2018) of policy π in MDP $\mathcal{M}$.

Moreover, for every state $s \in \mathcal{M}$, we can define its discounted future state distribution as $d^{\pi}_{\mathcal{M}}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s|\pi, \mathcal{M})$. First, we will consider the situation of observation disturbance. Similar to the setting of SA-MDP (Zhang et al., 2020), we introduce adversary $\nu : \mathcal{S} \to \mathcal{S}$ to describe the disturbance of state and denote the policy disturbed by adversary ν as $\hat{\pi}_\nu$, which means $\hat{\pi}_\nu(\cdot|s) = \pi(\cdot|\nu(s))$. We can theoretically calculate and bound the difference of performance between π and $\hat{\pi}_\nu$ in Theorem 3 as below:

Theorem 3. *For any policy π and any adversary ν, the reduction of expected cumulative reward of π against the observation disturbance of ν is:*

$$\begin{aligned} J_{\mathcal{M}}(\pi) - J_{\mathcal{M}}(\hat{\pi}_\nu) =& \frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d^{\hat{\pi}_\nu}_{\mathcal{M}}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left(1 - \frac{\pi(a|s)}{\pi(a|\nu(s))}\right)\mathbb{E}_{s'\sim P}V_{\mathcal{M},\pi}(s') \\ &+ \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\hat{\pi}_\nu}_{\mathcal{M}}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left(1 - \frac{\pi(a|s)}{\pi(a|\nu(s))}\right)R(s,a). \end{aligned} \tag{16}$$

Furthermore, an upper bound of it is as follows:

$$\begin{aligned} |J_{\mathcal{M}}(\pi) - J_{\mathcal{M}}(\hat{\pi}_\nu)| \leq& \frac{\gamma}{1-\gamma}\max_s D_{TV}(\pi(\cdot|s), \pi(\cdot|\nu(s)))\hat{V}_{\mathcal{M},\pi} \\ &+ \frac{2}{1-\gamma}\max_s D_{TV}(\pi(\cdot|s), \pi(\cdot|\nu(s))\max_{s,a}|R(s,a)|. \end{aligned} \tag{17}$$

The key of the proof is to analyze the relations of $V_{\mathcal{M},\hat{\pi}_\nu} - V_{\mathcal{M},\pi}$ with different states and the complete proof of Theorem 3 is in Appendix B.5, resembling the proof by Kakade & Langford (2002). Moreover, for the upper bound, Theorem 3 provides a structurally homologous, but tighter bound than the bound provided in Zhang et al. (2020) since our VFR can be bounded by $\max_{s,a}|R(s,a)|$, which is also proven in Appendix B.5. Compared with the victim policy π for given MDP $\mathcal{M}$, the factors mainly affect the performance of the disturbed policy π_ν are Total Variation distance $\max_s D_{TV}(\pi(\cdot|s), \pi(\cdot|\nu(s))$ and the VFR $\hat{V}_{\mathcal{M},\pi}$. The former one, TV distance, depends on the victim policy π as well as the disturbance ν and reflects the robustness of the victim policy and the adversarial ability of the adversary both. However, independent of the adversary, the latter one, VFR of the policy, only depends on the value functions of π in $\mathcal{M}$, reflecting the robustness of the victim policy. Thus we can improve the robustness under observation disturbance of the policy by controlling VFR of the policy.

4.2 Performance against transition Disturbance

Now, we consider the situation of transition disturbance. We assume that the transition $\mathcal{P}$ is disturbed to $\hat{\mathcal{P}}$ and attempt to evaluate the reduction of cumulative reward against the disturbance. Similar to Theorem 3, we can also theoretically show a similar result as below:

Theorem 4. *For any policy π in MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ and any disturbed environment $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, \mathcal{R}, \gamma)$, the reduction of cumulative reward against the transition disturbance is:*

$$J_{\mathcal{M}}(\pi) - J_{\hat{\mathcal{M}}}(\pi) = \frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{\hat{\mathcal{M}}}}\mathbb{E}_{a\sim\pi}\mathbb{E}_{s'\sim\hat{P}}\left(1 - \frac{P(s'|s,a)}{\hat{P}(s'|s,a)}\right)V_{\mathcal{M},\pi}(s'). \tag{18}$$

Furthermore, an upper bound of the reduction is:

$$J_{\mathcal{M}}(\pi) - J_{\hat{\mathcal{M}}}(\pi) \leq \frac{2\gamma}{1-\gamma}\max_{s,a} D_{TV}(P(\cdot|s,a), \hat{P}(\cdot|s,a))\hat{V}_{\mathcal{M},\pi}. \tag{19}$$

The proof of Theorem 4 is similar to that of Theorem 3 and is also deferred to Appendix B.5. Similarly, compared with the victim policy π for a given MDP $\mathcal{M}$, the factors that mainly affect the performance of π in disturbed environment $\hat{\mathcal{M}}$ are TV distance $\max_{s,a} D_{TV}(P(\cdot|s,a), \hat{P}(\cdot|s,a))$ and the VFR $\hat{V}_{\mathcal{M},\pi}$. The former one, TV distance, depends on the range of transition disturbance and reflect the adversarial ability of the adversary, which cannot be controlled by safe RL. Nevertheless, the latter one VFR only depends on the value functions of π in $\mathcal{M}$ and is an intrinsic property of the victim policy. Therefore, we can improve the robustness of the victim under transition disturbance policy by controlling $\hat{V}_{\mathcal{M},\pi}$.

4.3 CONNECTION BETWEEN THE OBSERVATION AND TRANSITION DISTURBANCE

Observation disturbance and transition disturbance are structurally different, as they affect observation of the policy and MDP respectively. Although existing literature usually considers them separately, by Theorem 3 and Theorem 4, we can find out that the effects of them on cumulative reward are similarly depending on the VFR $\hat{V}_{\mathcal{M},\pi}$, which is an inherent property of π and independent of the adversary. Thus we can improve the robustness of the policy under observation disturbance as well as transition disturbance by controlling its VFR.

Moreover, we will discuss the connection between controlling the VFR $\hat{V}_{\mathcal{M},\pi}$ and CVaR-based RL. For controlling $\hat{V}_{\mathcal{M},\pi}$, it's more reasonable to maximize $\min_s V_{\mathcal{M},\pi}(s)$ rather than minimize $\max_s V_{\mathcal{M},\pi}(s)$. However, as mentioned in Sec 3.1, directly maximizing the value function of the worst state may cause our policy to be over conservative. Thus it's more reasonable to loosen $\min_s V_{\mathcal{M},\pi}(s)$ to $-\mathrm{CVaR}_\alpha(-V(s))$, here $s \sim \mu(\cdot)$ obeys the initial distribution of the environment. Our CVaR-based objective (9) imposes a constraint on $-\mathrm{CVaR}_\alpha(-D(\tau))$ since we can prove that

Theorem 5. *For any $\alpha \in [0,1]$, We can prove that $-\mathrm{CVaR}_\alpha(-D(\tau))$ is a lower bound of $-\mathrm{CVaR}_\alpha(-V(s))$, i.e.*

$$-\mathrm{CVaR}_\alpha(-D(\tau)) \leq -\mathrm{CVaR}_\alpha(-V(s)) \tag{20}$$

The proof of Theorem 5 is deferred to Appendix B.6. Therefore, our CVaR-based methods consider to constrain $-\mathrm{CVaR}_\alpha(-D(\tau))$ for improving VFR of the policy and further improve the robustness of the policy against observation disturbance as well as transition disturbance.

5 EXPERIMENTS

In this section, we empirically evaluate the performance and the robustness under observation disturbance and transition disturbance of our method CPPO in a series of continuous control tasks in MuJoCo (Todorov et al., 2012) against other common on-policy RL algorithms.

5.1 EXPERIMENT SETUP

Environments. We choose MuJoCo (Todorov et al., 2012) as our experiments environment. As a robotic locomotion simulator, MuJoCo has an array of different continuous control tasks such as Ant, Walker2d, HalfCheetah, Hopper, Swimmer and so on, which are widely used for the evaluation of RL algorithms.

Baselines and Codes. We will compare our algorithm with the common on-policy algorithms and previous CVaR-based algorithms. For the former, we choose Vanilla Policy Gradient (VPG) (Sutton et al., 2000), Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) and PPO (Schulman et al., 2017). For the latter, we implement PG-CMDP (Chow & Ghavamzadeh, 2014) with deep neural network. And we use Adam (Kingma & Ba, 2015) to optimize all the parameters. The implementation of all codes, including CPPO and baselines, are based on SpinningUp (Achiam, 2018).

Evaluations. First, we compare the cumulative reward of each algorithm in the training process and their performance after convergence. For the trained models, in order to measure their robustness and safety, we compare their performance under transition disturbance and observation disturbance respectively. For observation disturbance, we apply Gaussian disturbance to the agent's observation to study the relationship between the agent's performance and the magnitude of the disturbance. For transition perturbation, since MuJoCo is a physical simulation engine and its transition is depend on its physics parameters, we choose to modify the mass of the agent to change the transition dynamics, and study the relationship between the the agent's performance and the mass of the agent.

5.2 PERFORMANCE IN TRAINING STAGE

In this part, we compare the performance of our CPPO against common on-policy algorithms as well as the previous CVaR-based algorithm in MuJoCo environments such as Ant, Halfcheetah, Walker2d, Swimmer and Hopper. For each algorithm in each task, we train 10 policies with different random seeds since the environment and environments and policies are stochastic. Table 1 shows the mean

Method \ Env	Ant-v3	HalfCheetah-v3	Walker2d-v3	Swimmer-v3	Hopper-v3
VPG	12.8± 0.0	896.9± 531.1	628.6± 229.4	48.3± 11.3	888.4± 209.5
TRPO	1625.4± 356.4	2073.8± 741.3	2005.6± 398.7	101.2± 29.3	2391.4± 455.3
PPO	3372.2± 301.4	3245.4± 947.3	2946.3± 944.3	122.0± 7.9	2726.0± 886.0
PG-CMDP	7.4 ± 3.6	928.7± 562.9	596.7± 219.9	55.4± 18.8	1039.2± 21.1
CPPO(ours)	**3514.7± 247.2**	**3680.5± 1121.3**	**3194.0± 648.2**	**182.5± 46.0**	**3144.6± 158.4**

Table 1: The cumulative reward (mean ± one std) of best policy trained by VPG, TRPO, PPO and CPPO in different MuJoCo games. In each column we **bold** the best performance over all algorithms.

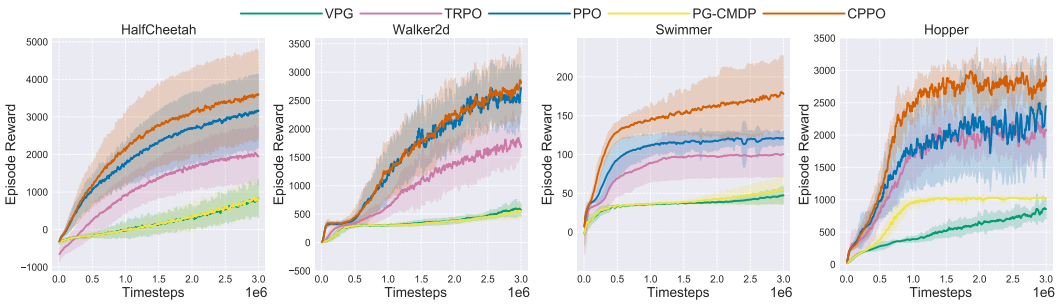

Figure 1: Cumulative reward curves for VPG, TRPO, PPO and our CPPO. The x-axes indicate the number of steps interacting with the environment, and the y-axes indicate the performance of the agent, including average rewards with standard deviations.

and variance of the cumulative reward of 10 policies trained by each algorithm in each environment and we **bold** the highest cumulative reward over all algorithms. For each algorithm in each task, we also plot the mean and variance of the ten policies as a function of timesteps in the training stage as shown in Figure 1. The four subgraphs represent the experimental results on Halfcheetah, Walker2d, Swimmer and Hopper respectively. The solid line represents the average reward of 10 strategies, and the part with lighter color represents the variance of them. As we can see from the figure, CPPO represented by pink has achieved significant performance improvement on HalfCheetah, Swimmer and Hopper against all baselines. We can also find that our CPPO gain higher cumulative reward of the worst-case outcome than other baselines on Walker2d.

5.3 Robustness against observation Disturbance in Test Stage

Trained agents may failed in the test stage because of the gap between the observation and the true state. Consequently, for evaluating the robustness of each algorithm, we add standard Gaussian disturbance to the observation in the test stage. For this purpose, we plot the performance of the trained policies under observation disturbance in Figure 2. In each subfigure, the solid line and the part with lighter color represent the average reward and the variance of 10 strategies respectively. From the figure, we can found that the performance degradation is positively related to the size of disturbance, which is shown in Theorem 3. Moreover, since the value function of all states in these policies are relatively low and VFR of these policies is low, we can discover that VPG and PG-CMDP stay robustness under observation disturbance, which is shown in Theorem 3. As shown in the figure, CPPO has made significant progress in Swimmer and Hopper than baselines. Therefore, our CPPO enables to keep robustness under observation disturbance.

5.4 Robustness against transition Disturbance in Test Stage

Trained agents may also fail in testing stage because of the transition gap between the simulator and the true environment. Therefore, we evaluate the performance of all algorithms under the transition disturbance for measuring their robustness and safety. Since MuJoCo is a physics simulator modeled on the physical world, we can disturb the transition by modifying environment parameters. For this purpose, we choose to modify the mass of the robot and the default mass of environment HalfCheetah,

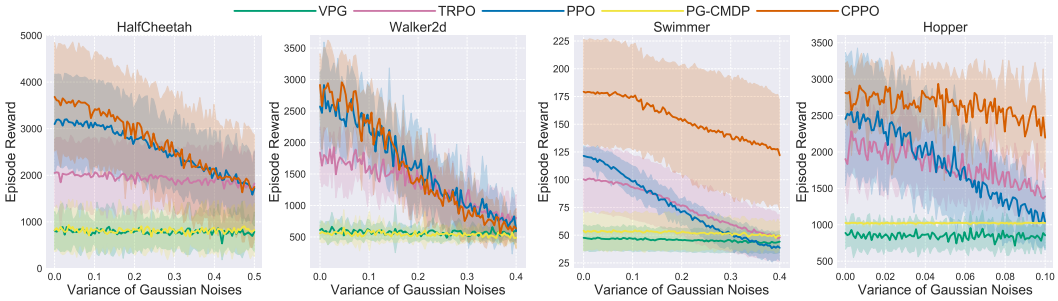

Figure 2: Cumulative reward curves for VPG, TRPO, PPO and our CPPO under observation disturbance. The x-axes indicate the range of the disturbance, and the y-axes indicate the average performance of the algorithm under the state disturbance.

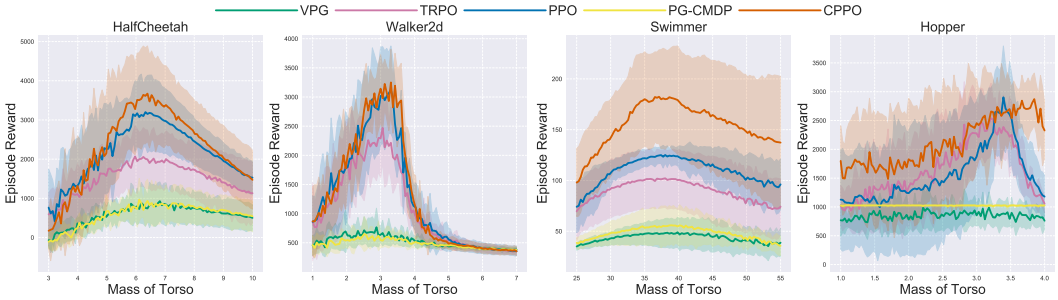

Figure 3: Cumulative reward curves for VPG, TRPO, PPO and our CPPO under transition disturbance. The x-axes indicate the mass of the agent, and the y-axes indicate the average performance of the algorithm when the mass changes.

Walker2d, Swimmer and Hopper are 6.36, 3.53, 34.6 and 3.53 respectively. Therefore, we draw Figure 3 to describe the results of agents, which are trained under standard mass condition and tested under different mass conditions. The solid line represents the average reward of 10 strategies, and the part with lighter color represents the variance of them. As seen in this figure, the performance of all algorithms decreases to a certain extent with the change of agent quality (whether it becomes larger or smaller) and the degree of decline is positively correlated with the quality change, which is consistent with our theoretical analysis in Theorem 4, that is, the upper bound of the performance difference of the algorithm is related to the size of the transition disturbance. Similar to the result under observation disturbance, we can can discover that VPG and PG-CMDP stay robustness under transition disturbance since their VFR is low, which is also shown in Theorem 4. At the same time, we can also see that CPPO achieve higher outcome in different tasks, specially in Swimmer and Hopper. It indicates that our method can improve the robustness of policies under transition disturbance.

6 Conclusions

In this paper, we analyze the advantages of CVaR for evaluating the risk of policy compared with the worst-case outcome as well as the variance of the return. Furthermore, we consider a risk-sensitive optimization objective and propose CPPO to solve it. Moreover, we provide theoretical connection of policies' robustness against observation disturbance and transition disturbance, which are structurally different. By introducing the notion of value function range (VFR), we indicate that our CPPO can improve the robustness of policies. Finally, we evaluate our algorithms in various MuJoCo tasks and show that CPPO obtains better performance as well as stronger robustness than various strong competitors.

REPRODUCIBILITY STATEMENT

We ensure the reproducibility of our paper from two aspects. (1) Experiment: The implementation and result of our experiment are described in Sec. 5. (2) Theory and Method: We provide the pseudo code of our algorithm in Appendix A. We also provide complete proofs of all the theoretical results mentioned in the paper in Appendix B.

ETHICS STATEMENT

Deep reinforcement learning may encounter catastrophic failures due to the stochasticity. It is very imperative to develop safe reinforcement learning algorithms. This paper proposes a CPPO method to improve the robustness under observation and transition disturbance. Also, this paper provides theoretical analysis of the connection between observation and transition disturbance. It may promote the development of safe and reliable reinforcement learning algorithms in the future.

REFERENCES

Joshua Achiam. Spinning Up in Deep Reinforcement Learning. 2018. 5.1

Gordon J Alexander and Alexandre M Baptista. A comparison of var and cvar constraints on portfolio selection with the mean-variance model. *Management Science*, 50(9):1261–1273, 2004. 1, 2.2

Siddharth Alexander, Thomas F Coleman, and Yuying Li. Minimizing cvar and var for a portfolio of derivatives. *Journal of Banking & Finance (JBF)*, 30(2):583–605, 2006. 1, 2.2

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society (JORS)*, 48(3):334–334, 1997. 3.2

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3509–3517, 2014. 2.2, 2.2, 3.2, 3.2, 5.1, B.4

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in Neural Information Processing Systems (NeurIPS)*, 28:1522–1530, 2015. 2.2, 2.2, B.3

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research (JMLR)*, 18(1):6070–6120, 2017. 2.2

Stefano P Coraluppi and Steven I Marcus. Risk-sensitive and minimax control of discrete-time, finite-state markov decision processes. *Automatica*, 35(2):301–309, 1999. 1, 2.1

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 16(1):1437–1480, 2015. 1, 2.1

Chris Gaskett. Reinforcement learning under circumstances beyond its control. *Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation*, 2003. 2.1, 2.1

Peter Geibel and Fritz Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research (JAIR)*, 24:81–108, 2005. 1

Abhijit Gosavi. Reinforcement learning for model building and variance-penalized control. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pp. 373–379. IEEE, 2009. 1, 2.1

Matthias Heger. Consideration of risk in reinforcement learning. In *Machine Learning Proceedings 1994*, pp. 105–111. Elsevier, 1994. 1, 2.1, 2.1, 2.1

Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972. 2.1

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017. 1

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning (ICML)*. Citeseer, 2002. 4.1

Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8248–8254. IEEE, 2019. 1

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6980. 5.1

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 1

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. URL https://doi.org/10.1038/nature14236. 1

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning (ICML)*, pp. 1928–1937. PMLR, 2016. 1

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. 2.1

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. B.3

Makoto Sato, Hajime Kimura, and Shibenobu Kobayashi. Td algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence (JSAI)*, 16(3):353–362, 2001. 2.1

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning (ICML)*, pp. 1889–1897. PMLR, 2015. 5.1

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1, 3.2, 5.1

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. URL https://doi.org/10.1038/nature16961. 1

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 2

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems (NeurIPS)*, pp. 1057–1063, 2000. 3.2, 5.1

Giorgio Szegö. Measures of risk. *Journal of Banking & finance (JBF)*, 26(7):1253–1272, 2002. 1, 2.1

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML)*, pp. 1651–1658, 2012. 1, 2.1

Aviv Tamar, Huan Xu, and Shie Mannor. Scaling up robust mdps by reinforcement learning. *arXiv preprint arXiv:1306.6189*, 2013. 2.1

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*, pp. 5026–5033. IEEE, 2012. 1, 5, 5.1

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21024–21037, 2020. 4.1, 4.1, B.5

A PSEUDO CODE OF CPPO

Algorithm 1 CVaR Proximal Policy Optimization(CPPO)

Require: confidence level α and reward tolerance β, learning rate $lr_\eta, lr_\theta, lr_\lambda, lr_\phi$

Ensure: θ of parameterized policy π_θ(always be random policy), ϕ of parameterized value function V_ϕ.

for $k = 1, 2, ..., N_{iter}$ **do**

Generate N trajectories $\mathcal{D}_k = \{\xi_i\}_{i=1}^N$ by following the current policy π_θ.

Compute reward $\hat{R}_i^t$ of each state $s_{i,t}$ in each trajectory ξ_i and the cumulative reward $D(\xi_i)$.

Compute advantage estimates $\hat{A}_i^t$ of each state $s_{i,t}$ in each trajectory ξ_i.

Update parameters respectively:

$$\eta \leftarrow \eta - lr_\eta \left(-\lambda + \frac{\lambda}{N(1-\alpha)} \sum_{i=1}^{N} \mathbf{1}\{\eta \geq D(\xi_i)\}) \right)$$

$$\theta \leftarrow \theta + lr_\theta \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=0}^{T} \nabla_\theta \min \left(\frac{\pi_\theta(a_i^t|s_i^t)}{\pi_{\theta_k}(a_i^t|s_i^t)} \hat{A}_i^t, g(\epsilon, \hat{A}_i^t) \right)$$
$$- lr_\theta \frac{1}{N} \sum_{i=1}^{N} (\nabla_\theta \log P_\theta(\xi_i)) \frac{\lambda}{1-\alpha} (-D(\xi_i) + \eta) \mathbf{1}\{\eta \geq D(\xi_i)\}$$

$$\lambda \leftarrow \lambda + lr_\lambda \left(-\eta + \frac{\sum_{i=1}^{N} (-D(\xi_i) + \eta)^+}{N(1-\alpha)} + \beta \right)$$

$$\phi \leftarrow \phi + lr_\phi \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=0}^{T} 2(V_\phi(s_{i,t}) - \hat{R}_i^t) \nabla_\phi V_\phi(s_{i,t}) \right)$$

Modify β as a function of the return of current trajectories:

$$\beta \leftarrow g(\xi_1, \xi_2, ..., \xi_N)$$

end for

B PROOFS OF THEOREMS

In this section, we will provide the proofs of theorems proposed in the paper.

B.1 THE PROOF OF THEOREM 1

Proof. By definition of VaR and CVaR, we have:

$$F_{f(D(\pi_\theta))}(z) = P(f(D(\pi_\theta)) \leq z) = P(D(\pi_\theta) \geq f^{-1}(z)) = 1 - F_{D(\pi_\theta)}(f^{-1}(z)),$$

$$\begin{aligned} \mathrm{VaR}_\alpha(f(D(\pi_\theta))) &= \min\{z | F_{f(D(\pi_\theta))}(z) \geq \alpha\} \\ &= \min\{z | 1 - F_{D(\pi_\theta)}(f^{-1}(z)) \geq \alpha\} \\ &= \min\{z | F_{D(\pi_\theta)}(f^{-1}(z)) \leq 1 - \alpha\}, \end{aligned}$$

$$\begin{aligned} \mathrm{CVaR}_\alpha(f(D(\pi_\theta))) &= \mathbb{E}_{w \sim f(D(\pi_\theta))}\{w | w \geq \mathrm{VaR}_\alpha(f(D(\pi_\theta)))\} \\ &= \mathbb{E}_{z \sim D(\pi_\theta)}\{f(z) | f(z) \geq \mathrm{VaR}_\alpha(f(D(\pi_\theta))\}. \end{aligned}$$

When we set $f(Z) = -Z$, we can naturally prove that

$$\begin{aligned}
-\mathrm{VaR}_\alpha(-Z) &= -\min\{z|F_Z(-z) \leq 1-\alpha\} \\
&= -\min\{z|1-F_Z(-z) \geq \alpha\} \\
&= \max\{-z|1-F_Z(-z) \geq \alpha\} \\
&= \max\{z|F_Z(z) \leq 1-\alpha\},
\end{aligned}$$

$$\begin{aligned}
-\mathrm{CVaR}_\alpha(-Z) &= -\mathbb{E}_{z\sim Z}\{-z|-z \geq \mathrm{VaR}_\alpha(-Z)\}. \\
&= \mathbb{E}_{z\sim Z}\{z|-z \geq \mathrm{VaR}_\alpha(-Z)\} \\
&= \mathbb{E}_{z\sim Z}\{z|z \leq -\mathrm{VaR}_\alpha(-Z)\}.
\end{aligned}$$

If we assume that $-\mathrm{CVaR}_\alpha(-Z) \geq \beta$, then we have:

$$\begin{aligned}
P(Z \leq \beta) &\leq P(Z \leq -\mathrm{CVaR}_\alpha(-Z)) \\
&= P(Z \leq \mathbb{E}_{w\sim Z}\{w|w \leq -\mathrm{VaR}_\alpha(-Z)\}) \\
&= P(Z \leq -\mathrm{VaR}_\alpha(-Z)) \\
&= P(Z \leq \max\{z|F_Z(z) \leq 1-\alpha\}) \\
&= 1-\alpha.
\end{aligned}$$

So we have proven it. □

B.2 The Proof of Theorem 2

Proof. Since we assume M is the upper bound of the total reward of every trajectory, we have $J(\pi_s) \leq M$. We consider two scenarios.

In the first case, if π_s satisfies that $-CVaR_\alpha(-D(\pi_s)) \geq \beta$. Obviously, we have $\pi_c(\alpha,\beta) = \pi_s$, thus

$$J(\pi_c(\alpha,\beta)) = J(\pi_s) \geq \frac{J(\pi_s) - \alpha M}{1-\alpha}.$$

Otherwise, we assume that $-\mathrm{CVaR}_\alpha(-D(\pi_s)) < \beta$, Since $-\mathrm{CVaR}_\alpha(-D(\pi_c(\alpha,\beta))) \geq \beta$, we set $B = -VaR_\alpha(-D(\pi_c(\alpha,\beta)))$ and have:

$$\begin{aligned}
J(\pi_c(\alpha,\beta)) &= \int_{\tau\sim\pi_c(\alpha,\beta)} p(\tau)D(\tau)d\tau \\
&= \int_{D(\tau)\leq B} p(\tau)D(\tau)d\tau + \int_{D(\tau)>B} p(\tau)D(\tau)d\tau \\
&\geq -\alpha CVaR(-D(\pi_c(\alpha,\beta))) + \int_{D(\tau)>B} p(\tau)Bd\tau \\
&\geq A\alpha + A(1-\alpha) \\
&= \beta.
\end{aligned}$$

By the similar way, we set:

$$A = -\mathrm{VaR}_\alpha(-D(\pi_\theta)) = \max\{z|F_{D(\pi_\theta)}(z) \leq 1-\alpha\},$$

thus

$$\begin{aligned}
J(\pi_s) &= \int_{\tau\sim\pi_s} p(\tau)D(\tau)d\tau \\
&= \int_{D(\tau)\leq A} p(\tau)D(\tau)d\tau + \int_{D(\tau)>A} p(\tau)D(\tau)d\tau \\
&= \int_{D(\tau)\leq A} p(\tau)Ad\tau + \int_{D(\tau)>A} p(\tau)Md\tau \\
&= A(1-\alpha) + M\alpha \\
&< \beta(1-\alpha) + M\alpha \\
&\leq J(\pi_c(\alpha,\beta))(1-\alpha) + M\alpha.
\end{aligned}$$

So we have proven $J(\pi_c(\alpha,\beta)) \geq \frac{J(\pi_s)-\alpha M}{1-\alpha}$. □

B.3 THE PROOF OF EQUIVALENTLY DEFORMING PROBLEM (9)

In this part, we will equivalently deforming problem (9) as

$$\begin{aligned}
&\max_{\theta} J(\pi_\theta) \quad s.t. -\mathrm{CVaR}_\alpha(-D(\pi_\theta)) \geq \beta \\
\Leftrightarrow &\min_{\theta} -J(\pi_\theta) \quad s.t. \mathrm{CVaR}_\alpha(-D(\pi_\theta)) \leq -\beta \\
\overset{1}{\Leftrightarrow} &\min_{\theta} -J(\pi_\theta) \quad s.t. \min_{\nu \in R}\{\nu + \frac{1}{1-\alpha}\mathbb{E}[(-D(\pi_\theta)-\nu)^+]\} \leq -\beta \\
\Leftrightarrow &\min_{\theta} -J(\pi_\theta) \quad s.t. \min_{\nu \in R}\{-\nu + \frac{1}{1-\alpha}\mathbb{E}[(-D(\pi_\theta)+\nu)^+]\} \leq -\beta \\
\Leftrightarrow &\min_{\theta,\nu} -J(\pi_\theta) \quad s.t. -\nu + \frac{1}{1-\alpha}\mathbb{E}[(-D(\pi_\theta)+\nu)^+] \leq -\beta.
\end{aligned}$$

Here we derive a formula 1 since CVaR owns the property (Rockafellar et al.; Chow et al., 2015):

$$\mathrm{CVaR}_\alpha(Z) = \min_{\eta \in R}\left\{\eta + \frac{1}{1-\alpha}\mathbb{E}[(Z-\eta)^+]\right\}. \tag{21}$$

So we have proven it. □

B.4 CALCULATING THE GRADIENT OF $L(\theta, \nu, \lambda)$

In this part, we will calculate the gradient $\partial_\nu L(\theta,\nu,\lambda)$, $\bigtriangledown_\theta L(\theta,\nu,\lambda)$ and $\bigtriangledown_\lambda L(\theta,\nu,\lambda)$ of the function $L(\theta,\nu,\lambda)$ by using the methods in (Chow & Ghavamzadeh, 2014):

$$L(\theta,\nu,\lambda) = -J(\pi_\theta) + \lambda(-\nu + \frac{1}{1-\alpha}\mathbb{E}[(-D(\pi_\theta)+\nu)^+] + \beta).$$

First we can expand the expectation as

$$\begin{aligned}
L(\theta,\nu,\lambda) =& -J(\pi_\theta) + \lambda(-\nu + \frac{1}{1-\alpha}\mathbb{E}[(-D(\pi_\theta)+\nu)^+] + \beta) \\
=& -\sum_\xi P_\theta(\xi)D(\xi) - \lambda\nu + \frac{\lambda}{1-\alpha}\sum_\xi P_\theta(\xi)(-D(\xi)+\nu)^+ + \lambda\beta.
\end{aligned}$$

We can see that $P_\theta(\xi)$ will only depend on θ and ξ, so we have easily calculate the gradient of λ as

$$\begin{aligned}
\nabla_\lambda L(\theta,\nu,\lambda) =& -\nu + \frac{1}{1-\alpha}\sum_\xi P_\theta(\xi)(-D(\xi)+\nu)^+ + \beta \\
=& -\nu + \frac{1}{1-\alpha}\mathbb{E}_{\xi\sim\pi_\theta}(-D(\xi)+\nu)^+ + \beta.
\end{aligned}$$

Then we calculate the gradient of ν. Since $(D(\xi)-\nu)^+$ isn't differentiable to ν at the point of $\nu = D(\xi)$, so we consider its semi gradient as

$$\partial_\nu(-D(\xi)+\nu)^+ = \begin{cases} 0 & \nu < D(\xi) \\ q(0 \leq q \leq 1) & \nu = D(\xi) \\ 1 & \nu > D(\xi) \end{cases}$$

And we can calculate the gradient of ν as below:

$$\begin{aligned}
\partial_\nu L(\theta,\nu,\lambda) =& -\lambda + \frac{\lambda}{1-\alpha}\sum_\xi P_\theta(\xi)\partial_\nu(-D(\xi)+\nu)^+ \\
=& -\lambda + \frac{\lambda}{1-\alpha}\sum_\xi P_\theta(\xi)\mathbf{1}\{\nu > D(\xi)\} + \frac{\lambda q}{1-\alpha}\sum_\xi P_\theta(\xi)\mathbf{1}\{\nu = D(\xi)\} \\
=& -\lambda + \frac{\lambda}{1-\alpha}\sum_\xi P_\theta(\xi)\mathbf{1}\{\nu \geq D(\xi)\} \quad (let\ q = 1) \\
=& -\lambda + \frac{\lambda}{1-\alpha}\mathbb{E}_{\xi\sim\pi_\theta}\mathbf{1}\{\nu \geq D(\xi)\}).
\end{aligned}$$

Finally, we will calculate the gradient of θ as

$$\begin{aligned}
\nabla_\theta L(\theta,\nu,\lambda) =& -\sum_\xi \nabla_\theta P_\theta(\xi)D(\xi) + \frac{\lambda}{1-\alpha}\sum_\xi \nabla_\theta P_\theta(\xi)(-D(\xi)+\nu)^+ \\
=& \sum_\xi \nabla_\theta P_\theta(\xi)(-D(\xi)+\frac{\lambda}{1-\alpha}(-D(\xi)+\nu)\mathbf{1}\{\nu \geq D(\xi)\}) \\
=& \sum_\xi (\nabla_\theta \log P_\theta(\xi))P_\theta(\xi)(-D(\xi)+\frac{\lambda}{1-\alpha}(-D(\xi)+\nu)\mathbf{1}\{\nu \geq D(\xi)\}) \\
=& -\sum_\xi (\nabla_\theta \log P_\theta(\xi))P_\theta(\xi)D(\xi) + \sum_\xi (\nabla_\theta \log P_\theta(\xi))P_\theta(\xi)\frac{\lambda(\nu-D(\xi))}{1-\alpha}\mathbf{1}\{\nu \geq D(\xi)\} \\
=& -\mathbb{E}_{\xi\sim\pi_\theta}(\nabla_\theta \log P_\theta(\xi))\left(D(\xi) - \frac{\lambda}{1-\alpha}(-D(\xi)+\nu)^+\right).
\end{aligned}$$

So we have calculated these three gradient. □

B.5 The Proof of Theorem 3 and Theorem 4

Before proving Theorem 3 and Theorem 4, we first examine a property of $d^\pi_{\mathcal{M}}$:

Lemma 1. *For any state $s \in \mathcal{S}$, we have:*

$$d^\pi_{\mathcal{M}}(s) = (1-\gamma)P(s_0 = s) + \gamma \sum_{s'} d^\pi_{\mathcal{M}}(s') \sum_a \pi(a|s)P(s'|s,a). \tag{22}$$

Proof. Here we'll prove this lemma. By the definition of $d^\pi_{\mathcal{M}}(s)$, we have:

$$\begin{aligned}
& d^\pi_{\mathcal{M}}(s) - (1-\gamma)P(s_0 = s) \\
=& (1-\gamma)\sum_{t=1}^{\infty}\sum_{s'} \gamma^t P(s_{t-1}=s', s_t = s|\pi,\mathcal{M}) \\
=& (1-\gamma)\sum_{t=0}^{\infty}\sum_{s'} \gamma^{t+1} P(s_t = s'|\pi,\mathcal{M})P(s_{t+1}=s|s_t=s',\pi,\mathcal{M}) \\
=& \gamma \sum_{s'}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t=s'|\pi,\mathcal{M})\right] P(s_1 = s|s_0=s',\pi,\mathcal{M}) \\
=& \gamma \sum_{s'} d^\pi_{\mathcal{M}}(s')P(s_1=s|s_0=s',\pi,\mathcal{M}) \\
=& \gamma \sum_{s'} d^\pi_{\mathcal{M}}(s')\sum_a \pi(a|s)P(s'|s,a).
\end{aligned} \tag{23}$$

Thus we have proven it. □

Now we will prove Theorem 3.

Theorem 3. *For any policy π and any adversary ν, we can calculate the reduction of expected cumulative reward of π against the observation disturbance of ν as*

$$\begin{aligned}
J_{\mathcal{M}}(\pi) - J_{\mathcal{M}}(\hat{\pi}_\nu) =& \frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d^{\hat{\pi}_\nu}_{\mathcal{M}}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left(1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right)\mathbb{E}_{s'\sim P}V_{\mathcal{M},\pi}(s') \\
&+\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\hat{\pi}_\nu}_{\mathcal{M}}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left(1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right)R(s,a).
\end{aligned} \tag{24}$$

Furthermore, we can give an upper bound of it:

$$\begin{aligned}
|J_{\mathcal{M}}(\pi) - J_{\mathcal{M}}(\hat{\pi}_\nu)| \leq& \frac{\gamma}{1-\gamma}\max_s D_{TV}(\pi(\cdot|s),\pi(\cdot|\nu(s)))\hat{V}_{\mathcal{M},\pi} \\
&+\frac{2}{1-\gamma}\max_s D_{TV}(\pi(\cdot|s),\pi(\cdot|\nu(s))\max_{s,a}|R(s,a)|.
\end{aligned} \tag{25}$$

Proof. Considering the bellman equation of value function of $\pi, \hat{\pi}_\nu$ in $\mathcal{M}$, we have:

$$V_{\mathcal{M},\pi}(s) = \sum_a \pi(a|s)[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_{\mathcal{M},\pi}(s')],$$
$$V_{\mathcal{M},\hat{\pi}_\nu}(s) = \sum_a \pi(a|\nu(s))[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_{\mathcal{M},\hat{\pi}_\nu}(s')].$$

By subtracting two value functions, we can deduce:

$$\begin{aligned} V_{\mathcal{M},\hat{\pi}_\nu}(s) - V_{\mathcal{M},\pi}(s) =& \gamma \sum_a (\pi(a|\nu(s)) - \pi(a|s)) \sum_{s'} P(s'|s,a) V_{\mathcal{M},\pi}(s') \\ &+ \gamma \sum_a \pi(a|\nu(s)) \sum_{s'} P(s'|s,a)(V_{\mathcal{M},\hat{\pi}_\nu}(s') - V_{\mathcal{M},\pi}(s')) \\ &+ \sum_a [\pi(a|\nu(s)) - \pi(a|s)] R(s,a). \end{aligned} \tag{26}$$

Since equation (26) satisfies for every state s, thus we calculate the expectation of equation (26) for $s \sim d_{\mathcal{M}}^{\hat{\pi}_\nu}$:

$$\begin{aligned} &\sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s)[V_{\mathcal{M},\hat{\pi}_\nu}(s) - V_{\mathcal{M},\pi}(s)] \\ =& \gamma \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a (\pi(a|\nu(s)) - \pi(a|s)) \sum_{s'} P(s'|s,a) V_{\mathcal{M},\pi}(s') \\ &+ \gamma \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a \pi(a|\nu(s)) \sum_{s'} P(s'|s,a)(V_{\mathcal{M},\hat{\pi}_\nu}(s') - V_{\mathcal{M},\pi}(s')) \\ &+ \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a [\pi(a|\nu(s)) - \pi(a|s)] R(s,a) \\ =& \gamma \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a (\pi(a|\nu(s)) - \pi(a|s)) \sum_{s'} P(s'|s,a) V_{\mathcal{M},\pi}(s') \\ &+ \sum_{s'} (V_{\mathcal{M},\hat{\pi}_\nu}(s') - V_{\mathcal{M},\pi}(s')) \left[\gamma \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a \pi(a|\nu(s)) P(s'|s,a) \right] \\ &+ \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a [\pi(a|\nu(s)) - \pi(a|s)] R(s,a). \end{aligned} \tag{27}$$

By Lemma 1, we have:

$$\begin{aligned} &\sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s)[V_{\mathcal{M},\hat{\pi}_\nu}(s) - V_{\mathcal{M},\pi}(s)] \\ =& \gamma \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a (\pi(a|\nu(s)) - \pi(a|s)) \sum_{s'} P(s'|s,a) V_{\mathcal{M},\pi}(s') \\ &+ \sum_{s'} (V_{\mathcal{M},\hat{\pi}_\nu}(s') - V_{\mathcal{M},\pi}(s')) \left[d_{\mathcal{M}}^{\hat{\pi}_\nu}(s') - (1-\gamma) P(s_0 = s') \right] \\ &+ \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a [\pi(a|\nu(s)) - \pi(a|s)] R(s,a). \end{aligned} \tag{28}$$

By moving the second term of the right part in (28) to the left part, we can deduce:

$$\begin{aligned} &(1-\gamma) \sum_{s'} (V_{\mathcal{M},\hat{\pi}_\nu}(s') - V_{\mathcal{M},\pi}(s')) P(s_0 = s') \\ =& \gamma \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a (\pi(a|\nu(s)) - \pi(a|s)) \sum_{s'} P(s'|s,a) V_{\mathcal{M},\pi}(s') \\ &+ \sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s) \sum_a [\pi(a|\nu(s)) - \pi(a|s)] R(s,a), \end{aligned} \tag{29}$$

thus:

$$\begin{aligned}(1-\gamma)(J_{\mathcal{M}}(\hat{\pi}_\nu)-J_{\mathcal{M}}(\pi)) =&(1-\gamma)\sum_{s'}(V_{\mathcal{M},\hat{\pi}_\nu}(s')-V_{\mathcal{M},\pi}(s'))P(s_0=s')\\ =&\gamma\sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s)\sum_a(\pi(a|\nu(s))-\pi(a|s))\sum_{s'}P(s'|s,a)V_{\mathcal{M},\pi}(s')\\ &+\sum_s d_{\mathcal{M}}^{\hat{\pi}_\nu}(s)\sum_a[\pi(a|\nu(s))-\pi(a|s)]R(s,a)\\ =&\gamma\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left(1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right)\mathbb{E}_{s'\sim P(\cdot|s,a)}V_{\mathcal{M},\pi}(s')\\ &+\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left(1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right)R(s,a).\end{aligned}$$

And we can prove:

$$\begin{aligned}J_{\mathcal{M}}(\pi)-J_{\mathcal{M}}(\hat{\pi}_\nu) =&\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left(1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right)\mathbb{E}_{s'\sim P(\cdot|s,a)}V_{\mathcal{M},\pi}(s')\\ &+\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left(1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right)R(s,a).\end{aligned}\tag{30}$$

Since $\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left(1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right)=0$, we can subtract a benchmark, which will not affect its value. Specially, we consider VFR $\hat{V}_{\mathcal{M},\pi}=\max_{s'}V_{\mathcal{M},\pi}(s')-\min_{s'}V_{\mathcal{M},\pi}(s')$ and we have $|V_{\mathcal{M},\pi}(s)-\hat{V}_{\mathcal{M},\pi}|\leq\frac{\hat{V}_{\mathcal{M},\pi}}{2}$ for every state s, thus we can prove that

$$\begin{aligned}|J_{\mathcal{M}}(\pi)-J_{\mathcal{M}}(\hat{\pi}_\nu)| \leq&\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left|1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right|\left|\mathbb{E}_{s'\sim P(\cdot|s,a)}V_{\mathcal{M},\pi}(s')-\hat{V}_{\mathcal{M},\pi}\right|\\ &+\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left|1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right||R(s,a)|\\ \leq&\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left|1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right|\frac{\hat{V}_{\mathcal{M},\pi}}{2}\\ &+\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\mathbb{E}_{a\sim\pi(\cdot|\nu(s))}\left|1-\frac{\pi(a|s)}{\pi(a|\nu(s))}\right|\max_{s,a}|R(s,a)|\\ \leq&\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\sum_a|\pi(a|\nu(s))-\pi(a|s)|\frac{\hat{V}_{\mathcal{M},\pi}}{2}\\ &+\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\sum_a|\pi(a|\nu(s))-\pi(a|s)|\max_{s,a}|R(s,a)|\\ =&\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\max_s D_{TV}(\pi(\cdot|s),\pi(\cdot|\nu(s)))\hat{V}_{\mathcal{M},\pi}\\ &+\frac{2}{1-\gamma}\mathbb{E}_{s\sim d_{\mathcal{M}}^{\hat{\pi}_\nu}}\max_s D_{TV}(\pi(\cdot|s),\pi(\cdot|\nu(s))\max_{s,a}|R(s,a)|.\end{aligned}\tag{31}$$

Thus we have proven Theorem 3. □

Furthermore, we will prove that our bound is tighter than the bound in Zhang et al. (2020):

$$\begin{aligned}
|J_{\mathcal{M}}(\pi) - J_{\mathcal{M}}(\hat{\pi}_\nu)| \leq & \frac{\gamma}{1-\gamma} \max_s D_{TV}(\pi(\cdot|s), \pi(\cdot|\nu(s))) \hat{V}_{\mathcal{M},\pi} \\
& + \frac{2}{1-\gamma} \max_s D_{TV}(\pi(\cdot|s), \pi(\cdot|\nu(s)) \max_{s,a} |R(s,a)| \\
\leq & \frac{2\gamma}{1-\gamma} \max_s D_{TV}(\pi(\cdot|s), \pi(\cdot|\nu(s))) \max_s |V_{\mathcal{M},\pi}(s)| \\
& + \frac{2}{1-\gamma} \max_s D_{TV}(\pi(\cdot|s), \pi(\cdot|\nu(s)) \max_{s,a} |R(s,a)| \\
\leq & \left(\frac{2\gamma}{(1-\gamma)^2} + \frac{2}{1-\gamma} \right) \max_s D_{TV}(\pi(\cdot|s), \pi(\cdot|\nu(s))) \max_{s,a} |R(s,a)|.
\end{aligned} \tag{32}$$

Finally, we will prove Theorem 4 by using the similar method of Theorem 3.

Theorem 4. *For any policy π in MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ and any disturbed environment $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, \mathcal{R}, \gamma)$, the reduction of cumulative reward against the transition disturbance is*

$$J_{\mathcal{M}}(\pi) - J_{\hat{\mathcal{M}}}(\pi) = \frac{\gamma}{1-\gamma} \mathbb{E}_{s\sim d^{\pi}_{\hat{\mathcal{M}}}} \mathbb{E}_{a\sim\pi} \mathbb{E}_{s'\sim\hat{P}} \left(1 - \frac{P(s'|s,a)}{\hat{P}(s'|s,a)} \right) V_{\mathcal{M},\pi}(s'). \tag{33}$$

Furthermore, we can give a upper bound of the reduction is therefore

$$J_{\mathcal{M}}(\pi) - J_{\hat{\mathcal{M}}}(\pi) \leq \frac{2\gamma}{1-\gamma} \max_{s,a} D_{TV}(P(\cdot|s,a), \hat{P}(\cdot|s,a)) \hat{V}_{\mathcal{M},\pi}. \tag{34}$$

Proof. Similarly, considering the bellman equation of value function of π in $\mathcal{M}, \hat{\mathcal{M}}$, we have

$$\begin{aligned}
V_{\mathcal{M},\pi}(s) &= \sum_a \pi(a|s)[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_{\mathcal{M},\pi}(s')], \\
V_{\hat{\mathcal{M}},\pi}(s) &= \sum_a \pi(a|s)[R(s,a) + \gamma \sum_{s'} \hat{P}(s'|s,a) V_{\hat{\mathcal{M}},\pi}(s')].
\end{aligned} \tag{35}$$

By subtracting them, we have

$$\begin{aligned}
V_{\hat{\mathcal{M}},\pi}(s) - V_{\mathcal{M},\pi}(s) =& \gamma \sum_a \pi(a|s) \sum_{s'} (\hat{P}(s'|s,a) - P(s'|s,a)) V_{\mathcal{M},\pi}(s') \\
& + \gamma \sum_a \pi(a|s) \sum_{s'} \hat{P}(s'|s,a)(V_{\hat{\mathcal{M}},\pi}(s') - V_{\mathcal{M},\pi}(s')).
\end{aligned} \tag{36}$$

Since equation (26) satisfies for every state s, thus we calculate the expectation of equation (26) for $s \sim d^{\hat{\pi}_\nu}_{\hat{\mathcal{M}}}$ and use Lemma 1:

$$\begin{aligned}
& \sum_s d^{\pi}_{\hat{\mathcal{M}}}(s)[V_{\hat{\mathcal{M}},\pi}(s) - V_{\mathcal{M},\pi}(s)] \\
=& \gamma \sum_s d^{\pi}_{\hat{\mathcal{M}}}(s) \sum_a \pi(a|s) \sum_{s'} (\hat{P}(s'|s,a) - P(s'|s,a)) V_{\mathcal{M},\pi}(s') \\
& + \gamma \sum_s d^{\pi}_{\hat{\mathcal{M}}}(s) \sum_a \pi(a|s) \sum_{s'} \hat{P}(s'|s,a)(V_{\hat{\mathcal{M}},\pi}(s') - V_{\mathcal{M},\pi}(s')) \\
=& \gamma \sum_s d^{\pi}_{\hat{\mathcal{M}}}(s) \sum_a \pi(a|s) \sum_{s'} (\hat{P}(s'|s,a) - P(s'|s,a)) V_{\mathcal{M},\pi}(s') \\
& + \sum_{s'} (V_{\hat{\mathcal{M}},\pi}(s') - V_{\mathcal{M},\pi}(s')) \left[\gamma \sum_s d^{\pi}_{\hat{\mathcal{M}}}(s) \sum_a \pi(a|s) \hat{P}(s'|s,a) \right] \\
=& \gamma \sum_s d^{\pi}_{\hat{\mathcal{M}}}(s) \sum_a \pi(a|s) \sum_{s'} (\hat{P}(s'|s,a) - P(s'|s,a)) V_{\mathcal{M},\pi}(s') \\
& + \sum_{s'} (V_{\hat{\mathcal{M}},\pi}(s') - V_{\mathcal{M},\pi}(s')) \left[d^{\pi}_{\hat{\mathcal{M}}}(s') - (1-\gamma) P(s_0 = s') \right].
\end{aligned} \tag{37}$$

Similarly, by moving the second term of the right part in (37) to the left part, we can deduce that

$$\begin{aligned}&(1-\gamma)\sum_{s'}(V_{\hat{\mathcal{M}},\pi}(s')-V_{\mathcal{M},\pi}(s'))P(s_0=s')\\=&\gamma\sum_s d^{\pi}_{\hat{\mathcal{M}}}(s)\sum_a\pi(a|s)\sum_{s'}(\hat{P}(s'|s,a)-P(s'|s,a))V_{\mathcal{M},\pi}(s'),\end{aligned}\tag{38}$$

thus:

$$\begin{aligned}(1-\gamma)(J_{\hat{\mathcal{M}}}(\pi)-J_{\mathcal{M}}(\pi))=&(1-\gamma)\sum_{s'}(V_{\hat{\mathcal{M}},\pi}(s')-V_{\mathcal{M},\pi}(s'))P(s_0=s')\\=&\gamma\sum_s d^{\pi}_{\hat{\mathcal{M}}}(s)\sum_a\pi(a|s)\sum_{s'}(\hat{P}(s'|s,a)-P(s'|s,a))V_{\mathcal{M},\pi}(s')\\=&\gamma\mathbb{E}_{s\sim d^{\pi}_{\hat{\mathcal{M}}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim\hat{P}(\cdot|s,a)}\left(1-\frac{P(s'|s,a)}{\hat{P}(s'|s,a)}\right)V_{\mathcal{M},\pi}(s').\end{aligned}\tag{39}$$

Thus we have proven:

$$J_{\mathcal{M}}(\pi)-J_{\hat{\mathcal{M}}}(\pi)=\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{\hat{\mathcal{M}}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim\hat{P}(\cdot|s,a)}\left(1-\frac{P(s'|s,a)}{\hat{P}(s'|s,a)}\right)V_{\mathcal{M},\pi}(s').\tag{40}$$

Similarly, we consider VFR $\hat{V}_{\mathcal{M},\pi}=\max_{s'}V_{\mathcal{M},\pi}(s')-\min_{s'}V_{\mathcal{M},\pi}(s')$ and we have $|V_{\mathcal{M},\pi}(s)-\hat{V}_{\mathcal{M},\pi}|\leq\frac{\hat{V}_{\mathcal{M},\pi}}{2}$ for every state s, thus we can prove:

$$\begin{aligned}|J_{\mathcal{M}}(\pi)-J_{\mathcal{M}}(\hat{\pi}_\nu)|\leq&\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{\hat{\mathcal{M}}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim\hat{P}(\cdot|s,a)}\left|1-\frac{P(s'|s,a)}{\hat{P}(s'|s,a)}\right|\left|V_{\mathcal{M},\pi}(s')-\hat{V}_{\mathcal{M},\pi}\right|\\\leq&\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{\hat{\mathcal{M}}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim\hat{P}(\cdot|s,a)}\left|1-\frac{P(s'|s,a)}{\hat{P}(s'|s,a)}\right|\frac{\hat{V}_{\mathcal{M},\pi}}{2}\\=&\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{\hat{\mathcal{M}}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\sum_{s'}\left|\hat{P}(s'|s,a)-P(s'|s,a)\right|\frac{\hat{V}_{\mathcal{M},\pi}}{2}\\=&\frac{\gamma}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{\hat{\mathcal{M}}}}\mathbb{E}_{a\sim\pi(\cdot|s)}D_{TV}(P(\cdot|s,a),\hat{P}(\cdot|s,a))\hat{V}_{\mathcal{M},\pi}.\end{aligned}\tag{41}$$

Thus we have proven Theorem 4.

B.6 The Proof of Theorem 5

By Theorem 1, we have

$$\begin{aligned}-\mathrm{CVaR}_\alpha(-V(s))&=\mathbb{E}_{z\sim V(s)}\{z|z\leq-\mathrm{VaR}_\alpha(-V(s))\}\\&=\mathbb{E}_\tau\{D(\tau)|V(s_0)\leq-\mathrm{VaR}_\alpha(-V(s))\}\\&\overset{1}{\geq}\mathbb{E}_\tau\{D(\tau)|D(\tau)\leq-\mathrm{VaR}_\alpha(-D(\tau))\}\\&=-\mathrm{CVaR}_\alpha(-D(\tau)).\end{aligned}\tag{42}$$

Here the inequality holds since $P(V(s_0)\leq-\mathrm{VaR}_\alpha(-V(s)))=P(D(\tau)\leq-\mathrm{VaR}_\alpha(-D(\tau)))=\alpha$. Thus we have proven Theorem 5.

□