

---

# Safe Distributional Reinforcement Learning

---

Jainyi Zhang<sup>1</sup>

Paul Weng<sup>1,2</sup>

<sup>1</sup>UM-SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai, China

<sup>2</sup>Department of Automation, Shanghai Jiao Tong University, Shanghai, China

## Abstract

Safety in reinforcement learning (RL) is a key property in both training and execution in many domains such as autonomous driving or finance. In this paper, we formalize it with a constrained RL formulation in the distributional RL setting. Our general model accepts various definitions of safety (e.g., bounds on expected performance, CVaR, variance, or probability of reaching bad states). To ensure safety during learning, we extend a safe policy optimization method to solve our problem. The distributional RL perspective leads to a more efficient algorithm while additionally catering for natural safe constraints. We empirically validate our propositions on artificial and real domains against appropriate state-of-the-art safe RL algorithms.

## 1 INTRODUCTION

Reinforcement learning (RL) has shown great promise in various applications [Silver et al., 2017, Jin, 2017]. As such techniques start to be deployed in real applications, safety in RL [Garcia and Fernandez, 2015] starts to be recognized as a key consideration both during learning, but also during execution after training. Indeed, in many domains from medical applications to autonomous driving to finance, the actions chosen by an RL agent can have disastrous consequences and therefore the corresponding risks need to be controlled both during training, but also during execution.

While traditional RL does not take safety into account, recent work has started to study it more actively. Safety takes various definitions in the literature. In its simplest sense, it means avoiding bad states [Geibel and Wysotzky, 2005], but it can take more general meaning such as decision-theoretic risk aversion [Borkar, 2010], or risk constraints [Prashanth and Ghavamzadeh, 2016], satisfaction of logic specifications [Alshiekh et al., 2018], but also simple bounds on

expected cumulated costs [Yu et al., 2019].

For a given definition of safety, one may want to learn a policy that satisfies it, without constraining the training process. Such approach would provide a safe policy for deployment after training. In contrast, recent work in safe RL aims at enforcing safety during learning as well, which is a difficult task as the RL agent needs to explore.

This paper follows this latter trend and safety is formulated as the satisfaction of a set of general constraints on distributions of costs or rewards. Thus, a safe policy is defined as a policy that respects some constraints in expectation or in probability. Our goal is to learn among safe policies one that optimizes the usual expected discounted sum of rewards. Furthermore, we also require safe learning, i.e., the safety constraints shall be satisfied during training as well.

To that aim, we first propose a general framework that accepts various safety formulations from bounds on CVaR to variance, to probability of reaching bad states. This general framework is made amenable by formulating the problem in the distributional RL setting, where distributions of returns are learned in contrast to their expectations. Based on this general distributional formulation, we extend an existing safe RL algorithm, Interior-Point Policy Optimization (IPO) [Liu et al., 2020], to the distributional setting, for which we formulate a performance bound.

**Contributions** Our contributions are threefold: (1) We propose a general framework for safe RL where safety is expressed as the satisfaction of risk constraints, which is enforced during and after training. A risk constraint can be expressed as any (sub)differentiable function of a random variable representing a cumulative reward or cost. (2) In order to obtain a practical algorithm, we formulate our problem and solution method in the distributional RL setting. (3) Our proposition, called SDPO, is empirically validated on multiple domains with various risk constraints against relevant state-of-the-art algorithms.

Algorithm	PD	CPO	IPO	PCPO	SDPO
Expectation	✓	✓	✓	✓	✓
Variance	✓	✗	✗	✗	✓
CVaR	✓	✗	✗	✗	✓
(Sub)differentiable fun.	✗	✗	✗	✗	✓
Safe learning	✗	✓	✓	✓	✓
Safe execution	✓	✓	✓	✓	✓

Table 1: Summary of related algorithms: which constraints are accepted, whether safety is guaranteed during learning/execution. PD (primal-dual) actually corresponds to several algorithms.

## 2 RELATED WORK

Safe RL is becoming an important research direction in RL [Garcia and Fernandez, 2015]. In this paper, we distinguish three main non-exclusive aspects for safe RL: policy safety, algorithmic safety, and exploration safety in exploration.

Policy safety corresponds to the goal of learning a safe policy such that its execution would avoid/limit the occurrence of bad outcomes (e.g., probability of reaching bad states or bound on performance). Safety can be modeled as additional constraints or penalization. In that sense, safe RL is related to risk-sensitive RL [Borkar, 2010, Chow and Ghavamzadeh, 2014, Chow et al., 2015] where the goal is to learn a policy that optimizes a risk-sensitive objective function, constrained RL [Achiam et al., 2017, Tessler et al., 2019, Miryoosefi et al., 2019, Liu et al., 2020] where the goal is to learn a policy that optimizes some constraints, and risk-constrained RL [Geibel and Wysotzky, 2005, Borkar and Jain, 2014, Prashanth and Ghavamzadeh, 2016, Chow et al., 2017, Brazdil et al., 2020], which in some sense combines the previous settings. The works in those three areas, with a few exceptions (CPO [Achiam et al., 2017], IPO [Liu et al., 2020], PCPO [Yang et al., 2020]), do not provide any safety guarantee during learning. They are based on a primal-dual approach (PD). For the exceptions, they can only accept simple constraints on expected discounted total costs. Notably, our algorithm, called Safe Distributional Policy Optimization (SDPO), builds on IPO [Liu et al., 2020] and extends it to the distributional RL setting, which then allows the formulation of sophisticated constraints. See Table 1 for a summary.

Algorithmic safety corresponds to the idea that running a safe RL algorithm should also guarantee some safety property, e.g., continuous policy improvement [Pirota et al., 2013], convergence to stationary point [Yu et al., 2019], satisfaction of logic specifications [Alshiekh et al., 2018], satisfaction of constraints [Achiam et al., 2017, Yang et al., 2020] during learning. However, none of those propositions can take into account sophisticated safety constraints (e.g., on risk measure).

Exploration safety focuses on an important aspect of safe RL: the exploration problem during learning in order to limit/avoid selecting dangerous actions.. In this context, safety is generally modeled as avoiding bad states. One main line of work [Turchetta et al., 2016, Berkenkamp et al., 2017, Wachi et al., 2018, Cheng et al., 2019] tries to prevent the choice of a bad action by learning a model. Other directions have been explored, for instance, by using a verification method [Fulton and Platzer, 2018] or by correcting a chosen action [Dalal et al., 2018]. However, this type of approaches requires the assumption that the environment is deterministic.

Although research has been active in safe RL, to the best of our knowledge, no efficient algorithm has been proposed for the general framework that we propose. In particular, our proposition can learn a risk-constrained policy while ensuring the satisfaction of the risk constraint during learning. Our proposition is based on distributional RL [Bellemare et al., 2017], which has demonstrated that estimating distributions of returns instead of their expectations can ensure better overall performance of RL algorithms. Most work [Dabney et al., 2018, Yang et al., 2019] in this area focuses on value-based methods, extending mostly the DQN algorithm [Mnih et al., 2015]. However, one recent work has also investigated the extension of the distributional setting to policy optimization [Barth-Maron et al., 2018]. Our work is based on the IQN algorithm [Dabney et al., 2018] instead of more recent propositions (e.g., [Yang et al., 2019]) because of its simplicity and because it perfectly fits our purposes. Note that in IQN, the authors consider optimizing a risk-sensitive objective function, but they do not consider constraints, as we do.

## 3 BACKGROUND

In this section, we present the notations, recall the definition of a Markov Decision Process (MDP) as well as its extension to Constrained Markov Decision Process (CMDP), and review the notions (e.g., CVaR) and the related deep RL algorithms, which we use to formulate our method.

**Notations** For any set  $X$ ,  $\Delta(X)$  denotes the set of probability distributions (or densities if  $X$  is continuous) over  $X$ . For any function  $f : Y \rightarrow \Delta(X)$  and any  $(x, y) \in X \times Y$ ,  $f(x | y)$  denotes the probability (or density value if  $X$  is continuous) of obtaining  $x$  according to  $f(y)$ . For any  $n \in \mathbb{N}$ ,  $[n]$  denotes  $\{1, 2, \dots, n\}$ . Vectors (resp. matrix) will be denoted in bold lowercase (resp. uppercase) with their components in normal font face with indices. For instance,  $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$  or  $\mathbf{M} = (m_{ij})_{i \in [n], j \in [m]} \in \mathbb{R}^{n \times m}$ .

**MDP Model** A *Markov Decision Process* (MDP) [Sutton and Barto, 2018] is defined as a tuple  $(\mathcal{S}, \mathcal{A}, P, r, \boldsymbol{\mu}, \gamma)$ , where  $\mathcal{S}$  is a set of states,  $\mathcal{A}$  is a set of actions,  $P$  :

$\mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$  is a transition function,  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is a reward function,  $\boldsymbol{\mu} \in \Delta(\mathcal{S})$  is a distribution over initial states, and  $\gamma \in [0, 1]$  is a discount factor. In this model, a policy  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$  is defined as a mapping from states to distributions over actions. We also use notation  $\pi_{\boldsymbol{\theta}}$  to emphasize that the policy is parameterized by  $\boldsymbol{\theta}$  (e.g., parameters of neural network). In the remaining, we identify  $\pi_{\boldsymbol{\theta}}$  to its parameter  $\boldsymbol{\theta}$  for ease of notation. The usual goal in an MDP is to search for a policy that maximizes the expected discounted total reward:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\mu}, P, \pi_{\boldsymbol{\theta}}} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] \quad (1)$$

where  $\mathbb{E}_{\boldsymbol{\mu}, P, \pi_{\boldsymbol{\theta}}}$  is the expectation with respect to the distribution  $\boldsymbol{\mu}$ , the transition function  $P$ , and  $\pi_{\boldsymbol{\theta}}$ . We define the (*state*) *value function* of a policy  $\pi_{\boldsymbol{\theta}}$  for state  $s$  as:

$$V^{\boldsymbol{\theta}}(s) = \mathbb{E}_{P, \pi_{\boldsymbol{\theta}}} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s] \quad (2)$$

where  $\mathbb{E}_{P, \pi_{\boldsymbol{\theta}}}$  is the expectation with respect to the transition function  $P$  and  $\pi_{\boldsymbol{\theta}}$ . The (*action*) *value function* is defined as follows:

$$Q^{\boldsymbol{\theta}}(s, a) = \mathbb{E}_{P, \pi_{\boldsymbol{\theta}}} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a] \quad (3)$$

and the *advantage function* is defined as:  $A^{\boldsymbol{\theta}}(s, a) = Q^{\boldsymbol{\theta}}(s, a) - V^{\boldsymbol{\theta}}(s)$ . As there is no risk of ambiguity, to avoid clutter we drop  $\boldsymbol{\mu}$  and  $P$  in the notation of the expectation from now on.

Reinforcement learning (RL) is based on MDP, but in RL, the transition and reward functions are not assumed to be known. Thus, in (online) RL, an optimal policy needs to be learned by trial and error.

**CMDP Model** The MDP model can be extended to the *Constrained MDP* (CMDP) setting [Altman, 1999] in order to handle constraints. In a CMDP,  $m$  cost functions  $c_i : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  for  $i \in [m]$  are introduced in addition to the original rewards. For each cost function  $c_i$ , the corresponding value functions can be defined. They are denoted with a subscript, e.g.,  $J_{c_i}$ ,  $V_{c_i}$ , or  $Q_{c_i}$ . For a CMDP, the goal is to find a policy that maximizes the expected discounted total reward while satisfying constraints on the expected costs  $J_{c_i}(\boldsymbol{\theta})$ :

$$\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \text{ s.t. } J_{c_i}(\boldsymbol{\theta}) \leq d_i \quad \forall i \in [m], \quad (4)$$

where  $\mathbf{d} = (d)_{i \in [m]} \in \mathbb{R}^m$  is a fixed vector constraint bound.

**Proximal Policy Optimization** The family of policy gradient methods constitutes the standard approach for tackling an RL problem when considering parametrized policies. Such a method iteratively updates a policy parameter in the direction of a gradient given by [Sutton and Barto, 2018]:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{(s, a) \sim \pi_{\boldsymbol{\theta}}} [A^{\boldsymbol{\theta}}(s, a) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a | s)]$$

where the expectation is taken with the respect to the state-action visitation distribution of  $\pi_{\boldsymbol{\theta}}$ . One issue in applying

a policy gradient method is the difficulty of estimating  $A^{\boldsymbol{\theta}}$  online. This issue motivates the use of an actor-critic scheme where an actor ( $\pi_{\boldsymbol{\theta}}$ ) and a critic (e.g.,  $A^{\boldsymbol{\theta}}$  or  $V^{\boldsymbol{\theta}}$  depending on the specific algorithm) are simultaneously learned. Learning the value function can help the policy update, such as reducing the gradient variance.

Proximal Policy Optimization (PPO) [Schulman et al., 2017] is a state-of-the-art actor-critic algorithm, which optimizes instead a clipped surrogate objective function  $J_{PPO}(\boldsymbol{\theta})$  defined by:

$$\sum_{t=0}^{\infty} \min(\omega_t(\boldsymbol{\theta}) A^{\bar{\boldsymbol{\theta}}}(s_t, a_t), \text{clip}(\omega_t(\boldsymbol{\theta}), \epsilon) A^{\bar{\boldsymbol{\theta}}}(s_t, a_t)), \quad (5)$$

where  $\bar{\boldsymbol{\theta}}$  is the current policy parameter,  $\omega_t(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\theta}}(a_t | s_t)}{\pi_{\bar{\boldsymbol{\theta}}}(a_t | s_t)}$ , and  $\text{clip}(\cdot, \epsilon)$  is the function to clip between  $[1 - \epsilon, 1 + \epsilon]$ . This surrogate function was motivated as an approximation of that used in TRPO [Schulman et al., 2015], which was introduced to ensure monotonic improvement after a policy parameter update. Although PPO is more heuristic than TRPO, its advantages are its simplicity and lower sample complexity.

**Distributional Reinforcement Learning** The key idea in distributional RL [Bellemare et al., 2017] is to learn a random variable to represent the discounted return  $\mathbf{Z}^{\boldsymbol{\theta}}(s, a) = \sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t$  where  $\mathbf{r}_t$  is the random variable representing the immediate reward received at time step  $t$  when applying action  $a$  in state  $s$  and policy  $\pi_{\boldsymbol{\theta}}$  thereafter. In contrast, standard RL algorithms directly estimate the expectation of  $\mathbf{Z}^{\boldsymbol{\theta}}(s, a)$ , since  $Q^{\boldsymbol{\theta}}(s, a) = \mathbb{E}_{\mathbf{Z}^{\boldsymbol{\theta}}} [\mathbf{Z}^{\boldsymbol{\theta}}(s, a)]$  where the expectation is with respect to the distribution of  $\mathbf{Z}^{\boldsymbol{\theta}}(s, a)$ .

Recall that any real random variable  $Z$  can be represented by its cumulative distribution denoted  $F_Z(z) = \mathbb{P}(Z \leq z) \in [0, 1]$ , or equivalently by its quantile function (inverse cumulative distribution) denoted  $F_Z^{-1}(p) = \inf\{z \in \mathbb{R} \mid p \leq F_Z(z)\}$  for any  $p \in [0, 1]$ . For ease of notation,  $Z_p$  denotes the  $p$ -quantile  $F_Z^{-1}(p)$ . In the *Implicit Quantile Network* (IQN) algorithm, Dabney et al. [2018] proposed to approximate the quantile function of  $\mathbf{Z}(s, a)$  with a neural network and to learn it using quantile regression [Koenker, 2005].

Concretely, the quantile function of  $\mathbf{Z}(s, a)$  can be learned as follows. Denote  $\hat{\mathbf{Z}}(s, a)$  the approximated random variable whose quantile function is given by a neural network  $\Psi(s, \tau)$ , which takes as input a state  $s$  and a probability  $\tau \in [0, 1]$  and returns the corresponding  $\tau$ -quantile  $\hat{\mathbf{Z}}_{\tau}(s, a)$  for each action  $a$ . After observing a transition  $(s, a, r, s')$ ,  $\Psi$  can be trained by sampling  $2N$  values  $\boldsymbol{\tau} = (\tau_1, \dots, \tau_N)$  and  $\boldsymbol{\tau}' = (\tau'_1, \dots, \tau'_N)$  with the uniform distribution on  $[0, 1]$ . By inverse transform sampling, sampling  $\boldsymbol{\tau}$  amount to sampling  $N$  values from  $\hat{\mathbf{Z}}(s, a)$  corresponding to  $\hat{\mathbf{Z}}_{\tau_1}(s, a), \dots, \hat{\mathbf{Z}}_{\tau_N}(s, a)$ , and similarly for  $\boldsymbol{\tau}'$  and sampling from  $\hat{\mathbf{Z}}(s', \pi(s'))$  where  $\pi$  is the current policy. Those samples define  $N^2$  TD errors in the distributional setting:

$$\delta_{ij} = r + \gamma \hat{\mathbf{Z}}_{\tau'_j}(s', \pi(s')) - \hat{\mathbf{Z}}_{\tau_i}(s, a) \quad (6)$$

Following quantile regression, the following loss function for training the neural network  $\Psi$  in  $(s, a, r, s')$  is given by:

$$L_{IQN} = \frac{1}{N} \sum_{i \in [N]} \sum_{j \in [N]} \xi_{\tau_i}^{\kappa}(\delta_{ij}) \quad (7)$$

where for any  $\tau \in (0, 1]$ ,  $\xi_{\tau}^{\kappa}(\delta_{ij}) = |\tau - \mathbb{I}(\delta_{ij} < 0)| \frac{L_{\kappa}(\delta_{ij})}{\kappa}$  is the quantile Huber loss with threshold  $\kappa$  with  $L_{\kappa}(\delta) = \frac{1}{2} \delta^2$  for  $|\delta| \leq \kappa$  or  $\kappa(|\delta| - \frac{1}{2} \kappa)$  otherwise.

**Interior-Point Policy Optimization** In the CMDP setting, Interior-point Policy Optimization (IPO) [Liu et al., 2020] is a recent RL algorithm to maximize an expected discounted total rewards while satisfying constraints on some expected discounted total costs. To deal with a constraint, IPO augments PPO’s objective function with a logarithmic barrier function applied to it, which provides a smooth approximation of the indicator function. The constrained problem then becomes an unconstrained one with an augmented objective function:

$$\max_{\theta} J_{IPO}(\theta) = J_{PPO}(\theta) + \sum_{i \in [m]} \frac{\ln(d_i - J_{c_i}(\theta))}{\eta}, \quad (8)$$

where  $\eta$  is a hyper-parameter. As  $\eta$  tends to  $\infty$ , the solution of (8) tends to that of the original constrained problem. The objective  $J_{IPO}$  is differentiable, therefore, we can apply a gradient-based optimization method to update the policy.

## 4 PROBLEM FORMULATION

Let  $\Delta(\mathbb{R})$  denote the set of real random variables. Therefore,  $\mathbf{Z} \in \Delta(\mathbb{R})^{\mathcal{S}}$  denotes a function from states to random variables. Given an (unknown) CMDP, the problem tackled in this paper can be expressed as a constrained optimization problem formulated in the distributional RL setting:

$$\max_{\theta} \mathbb{E}_{s_0 \sim \mu, \mathbf{Z}^{\theta}}[\mathbf{Z}^{\theta}(s_0)] \quad (9)$$

$$\text{s.t. } \rho_i(\mathbf{Y}_i^{\theta}) \leq d_i \quad \forall i \in [m] \quad (10)$$

where  $\mathbf{Z}^{\theta}(s)$  corresponds to the return distribution generated by policy  $\pi_{\theta}$  from the reward function, for all  $i \in [m]$ ,  $\mathbf{Y}_i^{\theta}(s)$  corresponds to the cumulated cost distribution from cost function  $c_i$ , and  $\rho_i : \Delta(\mathbb{R})^{\mathcal{S}} \rightarrow \mathbb{R}$  is a (sub)differentiable function. Note that this formulation is strictly more general than problem (4) thanks to the possibly non-linear functions  $\rho_i$ ’s.

We recall a few common cases for  $\rho_i$  in Table 2. The expectation is a simple example. For episodic MDPs with absorbing bad states, another simple example is the probability of bad states, which is defined like the expectation, but applied to a undiscounted cost equal to 1 for a bad state and 0 otherwise. CVaR is a widely-used risk measure in finance. In this context, the  $\alpha$ -CVaR of a portfolio is intuitively its expected return in the worst  $\alpha \times 100\%$  cases. Here,

$\rho_i$	Definition
Expectation	$\mathbb{E}_{s_0 \sim \mu, \mathbf{Y}}[\mathbf{Y}(s_0)]$
Prob. of bad states	$\mathbb{E}_{s_0 \sim \mu, \mathbf{Y}}[\mathbf{Y}(s_0)]$
$\alpha$ -CVaR of rewards	$\mathbb{E}_{s_0 \sim \mu}[\frac{1}{\alpha} \int_0^{\alpha} Y_{\zeta}(s_0) d\zeta]$
Variance	$\mathbb{E}_{s_0 \sim \mu}[\mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(s_0)^2] - \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}(s_0)]^2]$

Table 2: Common examples for  $\rho_i$ .

we adapted the definition to rewards (instead of costs). Naturally, a CVaR of an additional cost would also be possible. In contrast to previous methods, our framework can accept any (sub)differentiable definitions for  $\rho_i$  (e.g., coherent risk measures).

Note that we chose to take the mean (over initial states) of the CVaRs instead of the CVaR of the mean. The latter would have been possible as well, but because CVaR is a convex risk measure, our definition is an upperbound of the CVaR of the mean, which means that our formulation is more conservative and in that sense, safer. The same trick applies if  $\rho_i$  were defined based on any other coherent risk measure, of which CVaR is only one instance. Similarly, for the variance, we use the mean (over initial states) of variances instead of the other way around. Since the initial states are sampled in an independent way, the  $\mathbf{Y}(s_0)$ ’s are independent. This means that our definition upperbounds the variance of the mean of the  $\mathbf{Y}(s_0)$ ’s, leading to a more cautious formulation, which is more desirable for safe RL.

In this paper, we define a *safe policy* as a policy satisfying constraints (10). Our goal is to learn a policy maximizing an expected discounted total rewards (9) among all safe policies (i.e., safe execution). Besides, we require that any policy used during learning be safe (i.e., safe learning).

The formulation of (9)-(10) in the distributional RL setting serves two purposes. First, as observed in distributional RL, estimating the distributions of the cumulated rewards improves the overall performance. Second, many safety constraints (10), such as CVaR, become natural and simple to express in the distributional setting.

## 5 PROPOSED METHOD

To solve problem (9)-(10) in the safe RL setting, we extend IPO to the distributional RL setting and combine it with an adaptation of IQN. Next, we explain the general principle of our approach, and then discuss some techniques to obtain a concrete efficient implementation.

## 5.1 GENERAL PRINCIPLE

To adapt IPO, we rewrite the surrogate objective function used in PPO in the distributional setting:

$$J_{PPO}(\boldsymbol{\theta}) = \sum_{t=0}^{\infty} \min(\omega_t(\boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{Z}^{\bar{\boldsymbol{\theta}}}(s_t, a_t) - \mathbf{Z}^{\bar{\boldsymbol{\theta}}}(s_t)], \text{clip}(\omega_t(\boldsymbol{\theta}), \varepsilon) \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{Z}^{\bar{\boldsymbol{\theta}}}(s_t, a_t) - \mathbf{Z}^{\bar{\boldsymbol{\theta}}}(s_t)]). \quad (11)$$

Problem (9)-(10) can then be tackled by iteratively solving the following problem with this surrogate function:

$$\max_{\boldsymbol{\theta}} J_{PPO}(\boldsymbol{\theta}) \text{ s.t. } \rho_i(\mathbf{Y}_i^{\boldsymbol{\theta}}) \leq d_i \quad \forall i \in [m]. \quad (12)$$

Now, following IPO, using the log barrier function, we reformulate problem (12) as an unconstrained problem:

$$\max_{\boldsymbol{\theta}} J_{PPO}(\boldsymbol{\theta}) + \sum_{i \in [m]} \frac{\ln(d_i - \rho_i(\mathbf{Y}_i^{\boldsymbol{\theta}}))}{\eta_i}. \quad (13)$$

In contrast to convex optimization [Boyd and Vandenberghe, 2004], we introduce a constraint-dependent hyperparameter  $\eta_i$  to better control the satisfaction of each constraint separately.

Finally, we propose to solve problem (13) with an actor-critic architecture where both the actor and the critic are approximated with neural networks. For the critic, we adapt the approach proposed for IQN [Dabney et al., 2018] to learn random returns  $\mathbf{Z}$  and random cumulated costs  $\mathbf{Y}_i$ 's. For the actor, parameter  $\boldsymbol{\theta}$  of policy  $\pi_{\boldsymbol{\theta}}(a|s)$  is updated in the direction of the gradient of the objective function defined in (13):

$$\nabla_{\boldsymbol{\theta}} J_{PPO}(\boldsymbol{\theta}) - \sum_{i \in [m]} \frac{1}{\eta_i} \frac{\nabla_{\boldsymbol{\theta}} \rho_i(\mathbf{Y}_i^{\boldsymbol{\theta}})}{d_i - \rho_i(\mathbf{Y}_i^{\boldsymbol{\theta}})}. \quad (14)$$

This gradient raises one difficulty regarding the computation of  $\nabla_{\boldsymbol{\theta}} \rho_i(\mathbf{Y}_i^{\boldsymbol{\theta}})$ , which corresponds to the gradient of a critic with respect to the parameters of the actor. When  $\rho_i$  is linear (i.e., for expectation constraints), the policy gradient theorem [Sutton et al., 2000] applies and specifies how to compute  $\nabla_{\boldsymbol{\theta}} \rho_i(\mathbf{Y}_i^{\boldsymbol{\theta}})$ . However, when  $\rho_i$  is non-linear (i.e., for more sophisticated risk constraints), the gradient in (14) cannot be obtained easily. To solve this issue, we propose a simple and generic solution, which consists in connecting the actor network to any critic network with a non-linear  $\rho_i$  (see Figure 1 for an illustration where only one critic corresponding to non-linear  $\rho_i$  is displayed). Using this construct, the exact gradient of  $\rho_i(\mathbf{Y}_i^{\boldsymbol{\theta}})$  can be computed by automatic differentiation if  $\rho_i$  is (sub)differentiable and  $\mathbf{Y}_i^{\boldsymbol{\theta}}$  is approximated with a neural network, as we assume. Note that in previous work, Dabney et al. [2018] who proposed to optimize a risk measure in IQN did not face this gradient issue because their algorithm is based on DQN [Mnih et al., 2015] and therefore does not have an actor network. As a side note,

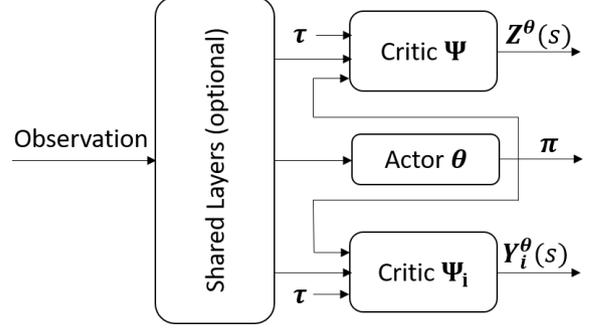


Figure 1: Architecture of SDPO where critic  $\Psi$  corresponds to the objective function and critic  $\Psi_i$  corresponds to constraint  $i$ . Both critics outputs a distribution.

this construct could be used to deal with a more general problem than (9)-(10) where a non-linear transformation is also applied on the objective function. For instance, one may want to optimize the CVaR of some rewards subject to some other risk constraints, which is as far as we know a completely novel problem. We leave this investigation to future work.

Like any interior point method, an initial feasible (i.e., safe) solution is needed. This requirement is actually not as strong as it seems. In many MDPs (or CMDPs), there is a known safe action for every state. For instance, in navigation problem, the action of not moving is safe if the current state is safe. In finance, investing in cash or a risk-free asset is safe. For many problems, a dummy action that does not have any effect can be added to define an initial safe action. More generally, when such a simple safe policy cannot be defined, an expert could possibly provide this initial safe policy or it could be obtained by pretraining with an imperfect simulator.

## 5.2 TECHNIQUES FOR EFFICIENT IMPLEMENTATION

In this section, to simplify notations, we do not write the superscript  $\boldsymbol{\theta}$  for the random variables  $\mathbf{Z}$  and  $\mathbf{Y}_i$ 's.

To make our final algorithm more efficient, we propose to learn  $\mathbf{Z}(s)$  only, instead of  $\mathbf{Z}(s, a)$  as it is the usual practice in distributional RL. This serves two purposes: (1) a state-dependent distribution is easier to learn, and (2) the advantage function can be easily estimated from a state value function alone. Note that for the constraints only  $\mathbf{Y}_i(s)$  is needed for any  $i \in [m]$ . Recall that the two random variables  $\mathbf{Z}(s)$  and  $\mathbf{Z}(s, a)$  are related by the following equation:

$$\mathbf{Z}(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)}[\mathbf{Z}(s')] \quad (15)$$

Following IQN, random variable  $\mathbf{Z}(s)$  is approximated by a random variable  $\hat{\mathbf{Z}}$ , which is represented by a neural network. The expectation of  $\mathbf{Z}(s)$  can then be approximated by that

---

**Algorithm 1** SDPO
 

---

**Require:** Constraint bound  $\mathbf{d}$ , Initial policy network  $\pi_{\theta_0}$ , Initial IQN network  $\Psi_0$ , Hyperparameters  $\varepsilon$  for PPO clip rate and  $\eta_i$  for each logarithmic barrier function.

```

1: for  $k = 0, 1, \dots$  do
2:    $\mathcal{B} \leftarrow$  run policy  $\pi_{\theta}$  for  $N$  trajectories
3:   # update the IQN network
4:   Sample  $\tau_1 < \dots < \tau_N$  from  $\mathcal{U}[0, 1]$ 
5:   # quantile regression
6:   for  $i, j \in [N]$  do
7:      $\delta_{ij} = r + \gamma \hat{\mathbf{Z}}_{\tau_j}'(s', \pi(s')) - \hat{\mathbf{Z}}_{\tau_i}(s, a)$ 
8:   end for
9:   Update  $\Psi_{k+1}$  with  $\nabla L_{IQN}$  (see (7)) using  $\mathcal{B}$ 
10:  Update  $\theta_{k+1}$  with  $\nabla J(\theta_k)$  defined in (14) using  $\mathcal{B}$ 
11: end for

```

---

of  $\hat{\mathbf{Z}}(s)$  with  $\tau$  randomly uniformly sampled in  $[0, 1]$ :

$$\mathbb{E}_{\theta}[\mathbf{Z}(s)] \approx \sum_{i=1}^N (\tau_i - \tau_{i-1}) \hat{\mathbf{Z}}_{\tau_i}(s). \quad (16)$$

setting  $\tau_0 = 0$  by convention and assuming  $0 < \tau_1 < \tau_2 < \dots < \tau_N < 1$ .

The exact handling of the constraints depend on the definition of  $\rho_i$ . As illustrative examples, we explain how they can be computed for some concrete cases. If  $\rho_i$  is simply defined as an expectation, it can be dealt with like the objective function. For CVaR, it can be estimated as follows for a random variable  $\mathbf{Y}(s_0)$ :

$$c_{\alpha}(\mathbf{Y}) \approx c_{\alpha}(\hat{\mathbf{Y}}) = \frac{1}{\alpha} \sum_{i|\tau_i \leq \alpha} (\tau_i - \tau_{i-1}) \hat{\mathbf{Y}}_{\tau_i}(s_0) \quad (17)$$

Here, in contrast to the standard expectation (e.g., (16)), an implementation trick consists in sampling  $\tau$  in  $[0, \alpha]$  such as  $\tau_1 < \tau_2 < \dots < \tau_N = \alpha$  since (17) corresponds to the expectation conditioned on event " $\hat{\mathbf{Y}} \leq \hat{\mathbf{Y}}_{\alpha}$ ". For the variance,  $\rho_i(\mathbf{Y})$  can be estimated by:

$$\sum_{i=1}^N (\tau_i - \tau_{i-1}) \hat{\mathbf{Y}}_{\tau_i}(s_0)^2 - \left( \sum_{i=1}^N (\tau_i - \tau_{i-1}) \hat{\mathbf{Y}}_{\tau_i}(s_0) \right)^2 \quad (18)$$

The pseudo code of our method is shown in Algorithm 1.

### 5.3 PERFORMANCE GUARANTEE BOUND

For fixed  $\eta$ , solving (13) instead of (12) may incur a performance loss, which can be bounded under natural conditions, which we discuss below. Since this result uses weak Lagrange duality, we first recall the definition of the Lagrangian of (12):

$$\mathcal{L}(\theta, \lambda) = J_{PPO}(\theta) + \sum_{i \in [m]} \lambda_i (d_i - \rho_i(\mathbf{Y}_i^{\theta}))$$

and its dual function:  $g(\lambda) = \max_{\theta} \mathcal{L}(\theta, \lambda)$ . The following bound can be proven:

**Theorem 1.** *If  $\theta_1^*$  is an optimal solution of (12),  $\theta_2^*$  is the strictly feasible optimal solution of (13) and the unique stationary point of  $\mathcal{L}(\cdot, \lambda^*)$  with  $\lambda_i^* = \frac{1}{\eta_i (d_i - \rho_i(\mathbf{Y}_i^{\theta_2^*}))}$  then:*

$$J_{PPO}(\theta_1^*) - J_{PPO}(\theta_2^*) \leq \sum_{i \in [d]} \frac{1}{\eta_i} \quad (19)$$

*Proof.* This result generalizes Theorem 1 of [Liu et al., 2020], whose proof implicitly uses convexity (which does not hold in deep RL) and follows from the discussion in page 566 of [Boyd and Vandenberghe, 2004].

We adapt the proof to our more general setting. We have:

$$J_{PPO}(\theta_1^*) \leq g(\lambda^*) \quad (20)$$

$$= J_{PPO}(\theta_2^*) + \sum_{i \in [m]} \lambda_i^* (d_i - \rho_i(\mathbf{Y}_i^{\theta_2^*})) \quad (21)$$

$$= J_{PPO}(\theta_2^*) + \sum_{i \in [m]} \frac{1}{\eta_i} \quad (22)$$

Step (20) holds by weak duality because  $\lambda_i^* \geq 0$  for all  $i \in [m]$  (since  $\theta_2^*$  is strictly feasible). Step (21) holds because we have by definition of  $\theta_2^*$ :

$$\nabla_{\theta} J_{PPO}(\theta_2^*) - \sum_{i \in [m]} \frac{\nabla_{\theta} \rho_i(\mathbf{Y}_i^{\theta_2^*})}{\eta_i (d_i - \rho_i(\mathbf{Y}_i^{\theta_2^*}))} = 0 \quad (23)$$

which implies that  $\theta_2^*$  maximizes  $\mathcal{L}(\cdot, \lambda^*)$  since  $\theta_2^*$  is assumed to be its unique stationary point. Step (22) holds by definition of  $\lambda^*$ .  $\square$

The conditions in this theorem are natural. In order to apply an interior point method, the constrained problem needs to be strictly feasible. The condition on the stationarity of  $\theta_2^*$  is reasonable and can be controlled by setting  $\varepsilon$  (used in the clipping function of  $J_{PPO}$ ) small enough.

As a direct corollary, this result implies that if (13) could be solved exactly, the error made by algorithm SDPO is controllable via setting appropriate  $\eta_i$ 's. Naturally, in the on-line RL setting, this assumption does not hold perfectly, but this result still provides some theoretical foundation to our proposition. In the next section, we validate the algorithm using various experimental settings.

## 6 EXPERIMENTAL RESULTS

The experiments are carried out in three different domains to validate our algorithm: random CMDPs, safety gym, as well as financial investment. See Appendix A for details about hyperparameter settings.

Random CMDPs are CMDPs with  $N$  states and  $M$  actions, where transition probabilities  $P(s' | s, a)$  are randomly assigned with  $\lceil \ln N \rceil$  positive values for each pair of state-action, and rewards are sampled from a uniform distribution,

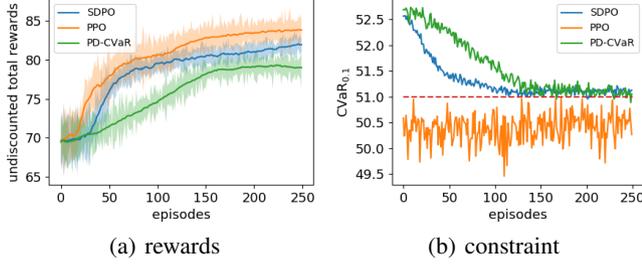


Figure 2: 2(a): Average performance over 10 runs of PPO, SDPO and PD-CVaR under the random CMDP for  $N = 1000$ . 2(b): 0.1-CVaR bounded by 51. Both SDPO and PD-CVaR converge to the level indicated by the dashed line.

i.e.,  $r(s, a) \sim \mathcal{U}[0, 1]$ . In the experiments, we set  $N = 1000$  and  $M = 10$ . We consider two cases: a bound over the variance or a bound over the CVaR, both over the distribution of discounted total rewards.

Safety gym [Ray et al., 2019] includes a set of environments designed for evaluating safe exploration in RL. They all correspond to navigation problems where an agent (i.e., Point, Car, Doggo) moves to some random goal positions to perform some tasks (i.e., Goal, Button, Push) while avoiding entering dangerous hazards or bumping into fragile vases. Each task has two difficulty levels (i.e., 1 or 2). See Appendix A for more details. For space reasons, we only present a selection of results in this domain in the main paper. More experimental results in these Mujoco environments are shown in Appendix B.

The third domain is the financial stock market. The RL agent can observe the close prices of the stocks in one day, i.e., the observation  $o_t = \mathbf{p}_t = (1, p_{1,t}, \dots, p_{N,t})$  for  $N$  selected stocks where the first component corresponds to cash. We further assume that all transactions are dealt at these prices. The action of the agent is defined by a portfolio vector, which corresponds to allocation weights over cash and stocks, i.e.,  $a_t = \mathbf{w}_{t+1} = (w_{0,t+1}, \dots, w_{N,t+1})$ ,  $w_0$  (resp.  $w_i$  for  $i \in [N]$ ) is the weight for cash (resp. stock  $i$ ) and  $\sum_{i=0}^N w_{i,t} = 1$ . Naturally, for each stock, we want to maximize the profit. Thus, with reward function  $r_t = \ln \sum_{i=0}^N w_{i,t} \frac{p_{i,t}}{p_{i,t-1}}$ , optimizing the undiscounted cumulative rewards can maximize the profit. We set the CVaR boundary  $d_1 = 0$  to avoid any possible loss. Detailed settings of the experiment are listed in Appendix A.

In all our experiments, all the agents are initialized so that they are in a feasible region at the beginning. In practice, an initial safe policy can be defined using domain knowledge or by an expert, e.g., in Mujoco domain, the agent can be initialized to stay and doing nothing. For fairness, the PPO agent is also initialized with the same safe policy as all other agents. Two policy gradient algorithms with CVaR and variance constraints respectively, PD-CVaR [Chow and Ghavamzadeh, 2014] and PD-VAR, which is modified from

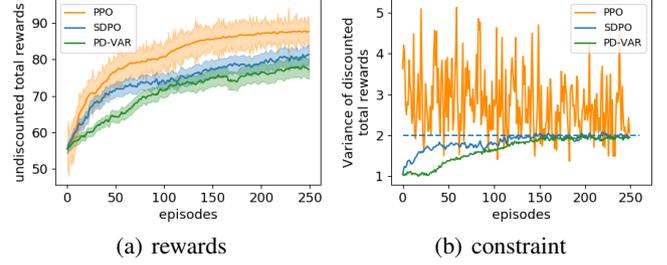


Figure 3: 3(a): Average performance over 10 runs of PPO, SDPO and PD-VAR under the random CMDP for  $N = 1000$ . 3(b): Variance bounded by 2. Both SDPO and PD-VAR converge to the level indicated by the dashed line.

Algorithm 2 in [Prashanth and Ghavamzadeh, 2016] are used as baselines in the first domain. SDPO is compared with CPO [Achiam et al., 2017], PCPO [Yang et al., 2020], and IPO [Liu et al., 2020] in the second domain. PPO [Schulman et al., 2017] is evaluated on all domains to serve as a non-safe RL method. Note that in contrast to our architecture SDPO, none of those algorithms can tackle the problem defined in (9)-(10) in its most general form.

The experiments are designed to evaluate SDPO in a variety of domains with various risk constraints and to answer the following questions: **(A)** How does SDPO compare with methods based on Lagrangian relaxation? **(B)** How does SDPO compare with other safe RL algorithms? Does the distributional formulation of SDPO help compared to IPO? **(C)** How does SDPO perform with multiple constraints (cumulative cost and probability of reaching a bad states)? **(D)** How does SDPO perform on a real domain? How does the constraint stringency impact the performance of SDPO?

**Question (A)** To answer (A), we perform some experiments on the first domain, random CMDPs, with either a constraint on CVaR or a constraint on variance. Both are based on the rewards. Therefore, the first needs to be lower-bounded, while the second needs to be upper-bounded. The confidence level is fixed to  $\alpha = 0.1$  and the bound for CVaR is set to 51 and that for the variance is set to 2. The bounds were chosen so that they are not too restrictive.

From the results in Figure 2, as expected, PPO without constraint achieves the best total rewards and converges faster than the constrained ones. When the CVaR value is bounded, PD-CVaR and SDPO both converge to a slightly worse but safe policy, however SDPO converges faster. From the results in Figure 3, similar observations can be drawn for PPO, PD-VAR, and SDPO. With regards to safety, we can again conclude that SDPO is superior.

**Question (B)** To answer (B), we perform some experiments on the second domain, Safety gym, which is a much more difficult domain than random CMDPs. For this domain,

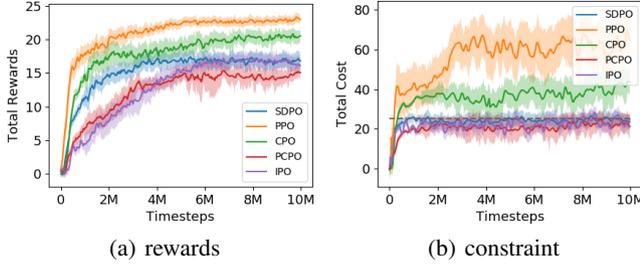


Figure 4: Average performance over 10 runs of PPO, SDPO, CPO, PCPO and IPO under Point-Goal1. They are bounded by the dashed line  $d_1 = 25$  in 4(b).

we did not evaluate the methods based on Lagrangian relaxation: since they do not use a critic, they would not be competitive. In Safety gym, the agent is penalized by receiving a cost  $c_1 = 1$  when touching a fragile vase. With a constraint on expected total cost  $\rho_1(\mathbf{Y}_1) = \mathbb{E}_{s_0 \sim \mu, \mathbf{Y}_1}[\mathbf{Y}_1(s_0)] \leq d_1$ , we are able to compare SDPO with other safe RL algorithms like CPO, PCPO and IPO.

We only show the results for Point-Goal1 in Figure 4. For other tasks, please refer to Appendix B. According to Figure 4, SDPO, PCPO and IPO can explore safely, while CPO cannot satisfy the constraint well. This latter observation regarding CPO may be surprising since CPO was designed to solve CMDPs, but similar results were also reported in previous work [Ray et al., 2019]. Among these three latter algorithms, SDPO and IPO performs the best. In Figure 4 and in all the Safety-gym environments (see Appendix B), SDPO dominates IPO in terms of either returns or convergence rates (and sometimes both), which confirms the positive contribution of the distributional critics.

**Question (C)** To demonstrate that SDPO can satisfy multiple constraints, the safety gym environment is used again, but with a variation. We modify the hazard area to be end states where an agent receives a cost  $c_2 = 1$ , and the episode is terminated. In addition to the previous constraint, another one is enforced:  $\rho_2(\mathbf{Y}_2) = \mathbb{E}_{s_0 \sim \mu, \mathbf{Y}_2}[\mathbf{Y}_2(s_0)] \leq d_2$ , where  $\mathbf{Y}_2$  is the undiscounted cumulative cost distribution from cost function  $c_2$ . Here, we set the bounds:  $d_1 = 10$  and  $d_2 = 0.1$ .

From Figures 5(a) and 5(b), PPO without constraints achieves much more goals, but at the cost of violating all the constraints. For constraint  $\rho_2$ , both SDPO and IPO agents can avoid entering into hazards during training. For constraint  $\rho_1$ , SDPO converges faster than IPO because of the adaption to distributional RL.

**Question (D)** To answer (D), we switch to the finance domain, where the stock market data of year 2019 is used. We run SDPO with a constraint on CVaR defined over rewards using different confidence levels  $\alpha$ . Note that since the CVaR is defined over rewards, it needs to be lower-

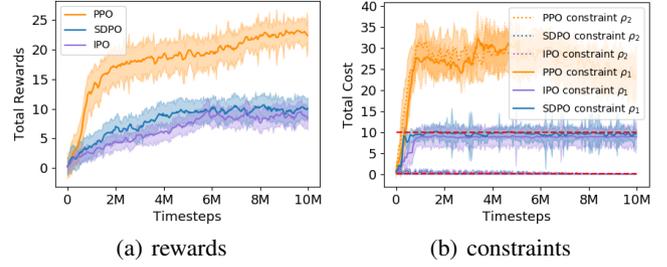


Figure 5: 5(a): Average performance over 5 runs of PPO, SDPO and IPO under Point-Goal2. 5(b): Average costs of PPO, SDPO and IPO under Point-Goal2.

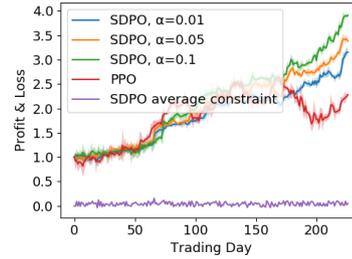


Figure 6: Average performance over 10 runs of PPO and SDPO with confidence level  $\alpha = 0.01, 0.05, 0.1$ .

bounded. We also run PPO as a baseline to show the performance without any constraints.

From Figure 6, all agents manage to make profits. With tighter constraint on risk (smaller  $\alpha$ ), the SDPO agent makes less profit. While PPO does not satisfy the constraint as expected, the curves for the constraint satisfaction of all SDPO agents are all similar. We therefore plot their average directly in Figure 6. PPO without constraint cannot avoid risk and thus suffers from fluctuation and loss at some time point. Interestingly, all the SDPO agents eventually perform better than PPO, which demonstrates that enforcing safety does not necessarily prevent good performance. Finally, SDPO with  $\alpha = 0.1$  performs best.

## 7 CONCLUSION

We presented a general framework for safe RL that encompasses many previous propositions. The novelty of our approach is the exploitation of a distributional RL formulation that allows us to deal with sophisticated risk constraints in a natural and efficient way for policy optimization. Our algorithm, SDPO, is shown to perform well in diverse environments and is competitive with previous algorithms in situations when they can be applied. However, SDPO can cover a larger range of safety formulations.

## References

- J. Achiam, D. Held, and A. Tamar et al. Constrained policy optimization. In *ICML*, 2017.
- M. Alshiekh, R. Bloem, R. R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. Safe reinforcement learning via shielding. In *AAAI*, 2018.
- E. Altman. *Constrained Markov Decision Processes*. CRC Press, 1999.
- G. Barth-Maron, M. W Hoffman, and D. et al. Budden. Distributed distributional deterministic policy gradients. *ICLR*, 2018.
- M. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. *ICML*, 2017.
- F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause. Safe model-based reinforcement learning with stability guarantees. In *NeurIPS*, 2017.
- V. Borkar and R. Jain. Risk-constrained Markov decision processes. *IEEE Transactions on Automatic Control*, 59(9):2574–2579, 2014.
- V. S. Borkar. Learning algorithms for risk-sensitive control. In *International Symposium on Mathematical Theory of Networks and Systems*, 2010.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.
- T. Brazdil, K. Chatterjee, P. Novotny, and J. Vahala. Reinforcement learning of risk-constrained policies in Markov decision processes. *AAAI*, 2020.
- R. Cheng, G. Orosz, R. M Murray, and J. W Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *AAAI*, 2019.
- Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *NeurIPS*, 2014.
- Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. In *NeurIPS*, 2015.
- Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *JMLR*, 18(1), 2017.
- W. Dabney, G. Ostrovski, D. Silver, and R. Munos. Implicit quantile networks for distributional reinforcement learning. *ICML*, 2018.
- G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa. Safe exploration in continuous action spaces. *CoRR*, 2018.
- N. Fulton and A. Platzer. Safe reinforcement learning via formal methods. In *AAAI*, 2018.
- J. Garcia and F. Fernandez. A comprehensive survey on safe reinforcement learning. *JMLR*, 16(1437–1480), 2015.
- P. Geibel and F. Wysotzky. Risk-sensitive reinforcement learning applied to control under constraints. *JAIR*, 24: 81–108, 2005.
- R. Jin. Deep learning at Alibaba. In *IJCAI*, 2017. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/2.
- R. Koenker. *Quantile Regression*. Cambridge university press, 2005.
- Y. Liu, J. Ding, and X. Liu. IPO: Interior-point policy optimization under constraints. *AAAI*, 2020.
- S. Miryoosefi, K. Brantley, H. Daume III, M. Dudik, and R. Schapire. Reinforcement learning with convex constraints. In *NeurIPS*, 2019.
- V. Mnih, K. Kavukcuoglu, and D. Silver et al. Human-level control through deep reinforcement learning. *Nature*, 2015.
- M. Pirodda, M. Restelli, A. Pecorino, and D. Calandriello. Safe policy iteration. In *ICML*, 2013.
- L. Prashanth and M. Ghavamzadeh. Variance-constrained actor-critic algorithms for discounted and average reward MDPs. *Machine Learning*, 2016.
- A. Ray, J. Achiam, and D. Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. 2019.
- J. Schulman, S. Levine, P. Abbeel, M.I. Jordan, and P. Moritz. Trust region policy optimization. In *ICML*, 2015.
- J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR*, 2017. URL <http://arxiv.org/abs/1707.06347>.
- D. Silver, J. Schrittwieser, and K. Simonyan et al. Mastering the game of go without human knowledge. *Nature*, 2017.
- R. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NeurIPS*, 2000.
- C. Tessler, D. Mankowitz, and S. Mannor. Reward constrained policy optimization. *ICLR*, 2019.
- M. Turchetta, F. Berkenkamp, and A. Krause. Safe exploration in finite Markov decision processes with gaussian processes. In *NeurIPS*, 2016.

- A. Wachi, Y. Sui, Y. Yue, and M. Ono. Safe exploration and optimization of constrained MDPs using gaussian processes. In *AAAI*, 2018.
- D. Yang, L. Zhao, Z. Lin, T. Qin, J. Bian, and T. Liu. Fully parameterized quantile function for distributional reinforcement learning. In *NeurIPS*. 2019.
- T. Yang, J. Rosca, K. Narasimhan, and P. Ramadge. Projection-based constrained policy optimization. In *ICLR*, 2020.
- M. Yu, Z. Yang, M. Kolar, and Z. Wang. Convergent policy optimization for safe reinforcement learning. In *NeurIPS*, 2019.

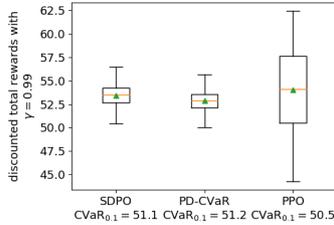


Figure 7: Evaluation of  $Z^{\pi_\theta}$  of trained policy over 1000 runs.

## A MORE DETAILS ON EXPERIMENTS

### A.1 RANDOM CMDP

In the random CMDP domain, we constructed it with  $N = 1000$  states and 10 actions. The number of randomly-chosen possible successor states is  $\lceil \ln N \rceil = 7$ . An episode in this CMDP is terminated after 100 time-steps. To achieve the results in Figures 2 and 3, we trained the agents 5 times for both random seeds 5 and 10. The ADAM optimizer is used. The hyper-parameters are listed in Table 3.

	PPO	SDPO	PD-CVaR	PD-VAR
discount factor $\gamma$	0.99	0.99	0.99	0.99
batch size	1000	1000	1000	1000
learning rate (actor)	1e-4	1e-4	1e-4	1e-4
learning rate (critic)	1e-3	1e-3	/	/
hidden sizes	(64,64)	(64,64)	(64,64)	(64,64)
GAE factor $\lambda$	0.9	0.9	/	/
clip range $\epsilon$	0.2	0.2	/	/
$\eta$	/	20	/	/
# quantile atoms $N$	/	128	/	/
quantile dimension <sup>1</sup>	/	256	/	/

Table 3: Hyperparameters for experiments on random CMDP.

To evaluate the converged policies in 2(a) with CVaR constraint, we run them for 1000 episodes each. Figure 7 indicates that PPO without constraint can reach the highest result but suffers from the risk of getting the lowest reward. Both SDPO and PD-CVaR receives a lower mean reward but much higher CVaR values, which indicates lower risk.

### A.2 STOCK TRANSACTION

For the experiment on stock market, we use Quandl<sup>1</sup> in Python to load all market data. The trading agent is assumed to have zero market impact and zero transaction cost. When conducting this experiment, We choose 9 stocks in SP500 (AAPL, CSCO, DOW, GE, GS, JNJ, JPM, MMM, MSFT). The agents are initialized with a safe policy that always holding cash, and then trained in a rolling bias in year 2019 to evaluate the offline performance, i.e., at time step  $t$ , prices from  $t - 15$  to  $t$  are used for training. The shared part of the actor and critic network is implemented as an LSTM network. The hyper-parameters are listed in Table 4.

### A.3 MUJOCO SIMULATOR

For the parameters and other settings of the Point-Goal2 domain we used the default values set in the source code of Safety-Gym (see line 108 in [safety-gym/safety\\_gym/envs/suite.py](https://github.com/GoogleCloudPlatform/robotics/blob/master/safety-gym/safety_gym/envs/suite.py)). The hyper-parameters are listed in Table 5.

<sup>1</sup><https://quandl.com>

<sup>1</sup>refer to Equation (4) in [Dabney et al., 2018]

	PPO	SDPO
discount factor $\gamma$	0.99	0.99
batch size	1280	1280
learning rate (actor)	1e-4	1e-4
learning rate (critic)	1e-3	1e-3
hidden sizes	(64,64)	(64,64)
GAE factor $\lambda$	0.9	0.9
clip range $\epsilon$	0.1	0.1
$\eta$	/	60
# quantile atoms $N$	/	128
quantile dimension	/	256

Table 4: Hyperparameters for experiments on stock transaction.

	PPO	SDPO	IPO
discount factor $\gamma$	0.99	0.99	0.99
discount factor for constraints $\gamma_1, \gamma_2$	1	1	1
batch size	30000	30000	30000
learning rate (actor)	1e-4	1e-4	1e-4
learning rate (critic)	1e-3	1e-3	1e-3
hidden sizes (actor)	(256,256)	(256,256)	(256,256)
hidden sizes (critic)	(256,256)	(256,256)	(256,256)
GAE factor $\lambda$	0.9	0.9	0.9
clip range $\epsilon$	0.1	0.1	0.1
$\eta_1$	/	40	60
$\eta_2$	/	60	60
# quantile atoms $N$	/	128	/
quantile dimension	/	256	/

Table 5: Hyperparameters for experiments on Point-Goal2.

## B ADDITIONAL EXPERIMENTS

We further compare our method to Constrained Policy Optimization (CPO) [Achiam et al., 2017] and Projection-based Constrained Policy Optimization (PCPO) [Yang et al., 2020]. PCPO is a two-step approach. In the first step, the policy is updated in the direction to improve the objective function in the trust region. In the second step, PCPO projects the potentially infeasible policy back to the constraint set.

To demonstrate the performance of SDPO, compared with PCPO, IPO and CPO, we conduct the experiment in the safety gym environment with constraint  $\rho_1$ . Three tasks, Goal, Button and Push with two levels of difficulties are tested in the experiments with agent point, car and dpggo. Task Goal is to move the agent to a series of goal positions, while task button is to press a series of goal buttons, and task push is to move a box to a series of goal positions. For detailed explanation of these tasks, please refer to [Ray et al., 2019]. The hyper-parameters in the experiment are the same as Table 5, except  $\eta = 30$  for SDPO and IPO.

From the results in Figures 8 to 10, SDPO, IPO and PCPO can explore the environment safely, but CPO may failed to learn a safe policy as the tasks go harder (i.e., with the car agent). In experiments, when the CPO agents violate the constraints, Equation (14) in [Achiam et al., 2017] failed to purely decrease the constraint value, and thus learn an unsafe policy. For the other three agents, SDPO converges faster to a slightly better policy than IPO and PCPO. An interesting future work would be to extend PCPO to the distributional setting as well.

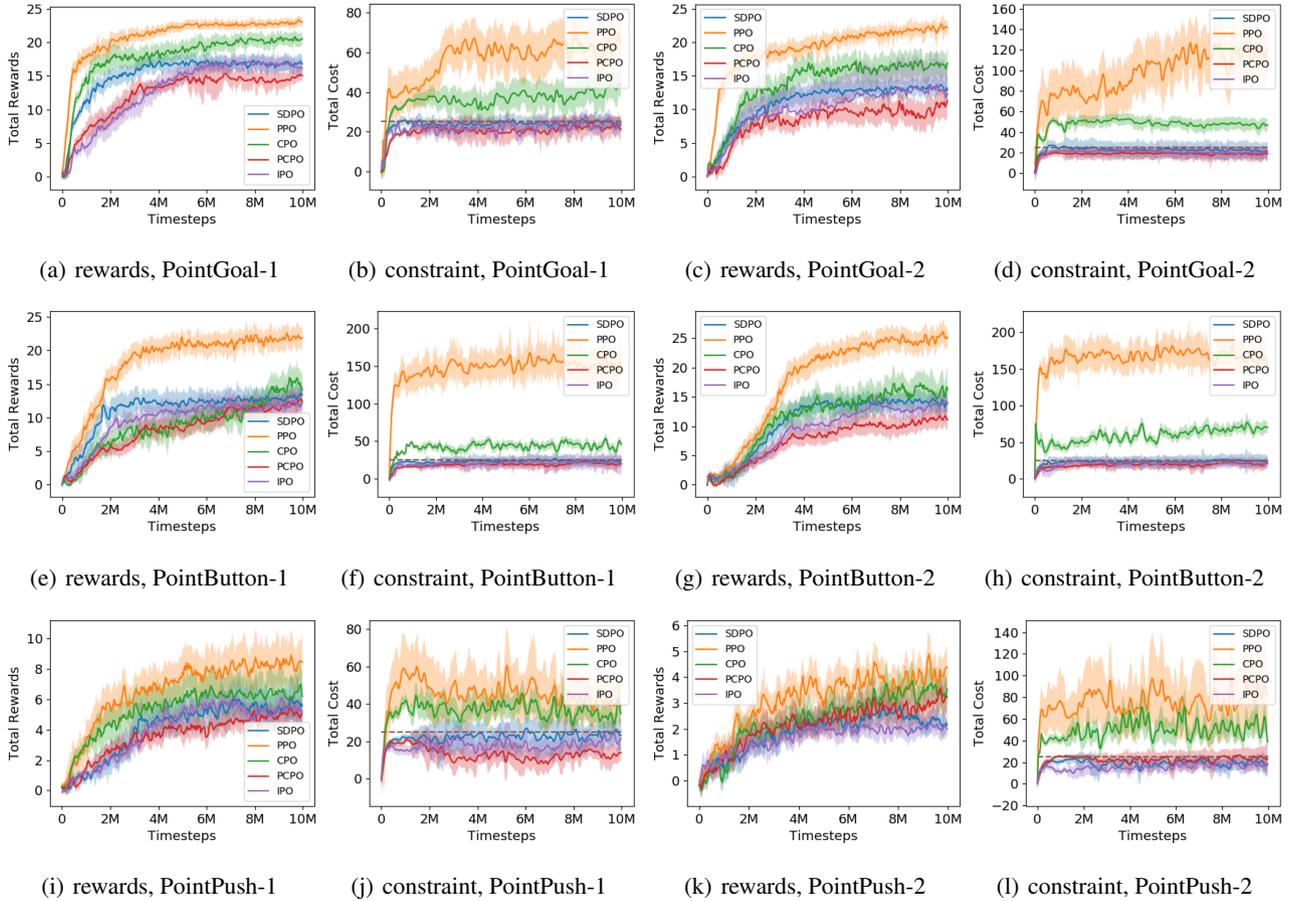


Figure 8: Average performance of the point agent over 10 runs of PPO, SDPO, PCPO and IPO under Safety-Gym. Both SDPO, PCPO and IPO converge to the level indicated by the dashed line.

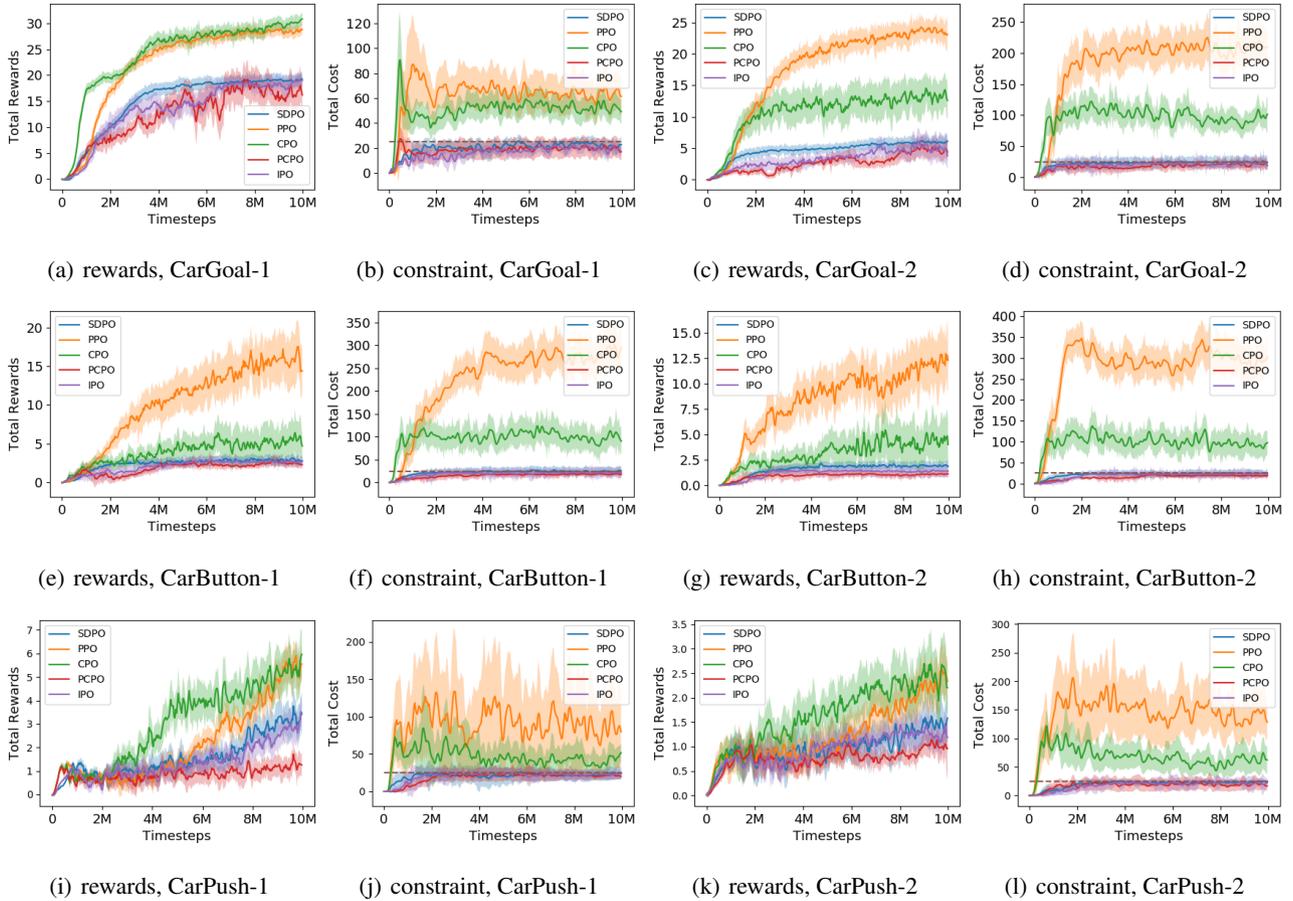


Figure 9: Average performance of the car agent over 10 runs of PPO, SDPO, PCPO and IPO under Safety-Gym. Both SDPO, PCPO and IPO converge to the level indicated by the dashed line.

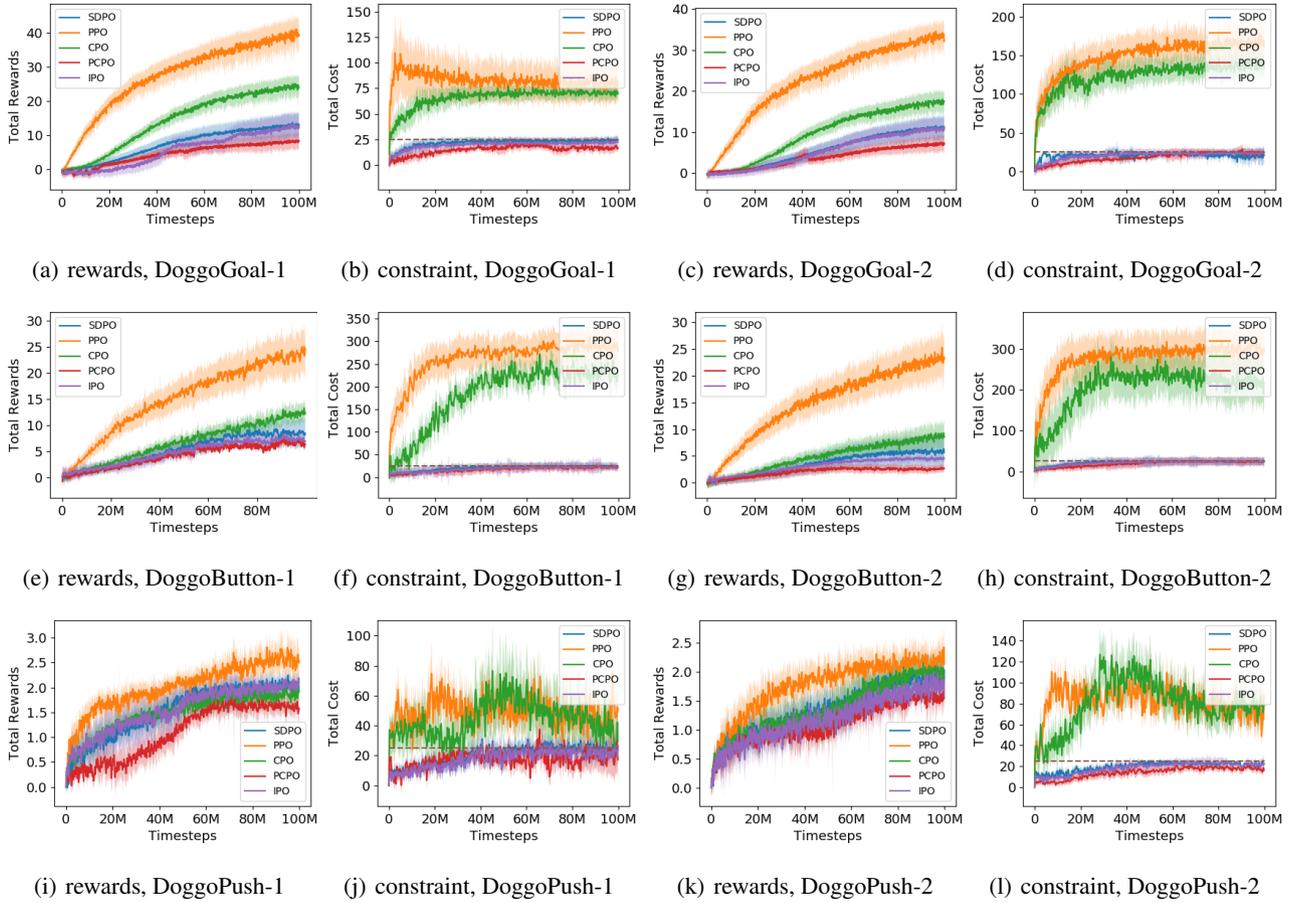


Figure 10: Average performance of the doggo agent over 10 runs of PPO, SDPO, PCPO and IPO under Safety-Gym. Both SDPO, PCPO and IPO converge to the level indicated by the dashed line.