Refined Risk Management in Safe Reinforcement Learning with a Distributional Safety Critic

Yang, Q.; Simão, T. D.; Tindemans, Simon H.; Spaan, M.T.J.

Publication date
2022

Document Version
Final published version

Published in
Safe RL Workshop at IJCAI 2022

Refined Risk Management in Safe Reinforcement Learning with a Distributional Safety Critic

Qisong Yang[1] , **Thiago D. Simão**[1,2] , **Simon H. Tindemans**[1] and **Matthijs T. J. Spaan**[1]
[1]Delft University of Technology, Delft, The Netherlands
[2]Radboud University, Nijmegen, The Netherlands
q.yang@tudelft.nl, thiago.simao@ru.nl, {s.h.tindemans, m.t.j.spaan}@tudelft.nl

Abstract

Safety is critical to broadening the real-world use of reinforcement learning (RL). Modeling the safety aspects using a safety-cost signal separate from the reward is becoming standard practice, since it avoids the problem of finding a good balance between safety and performance. However, the total safety-cost distribution of different trajectories is still largely unexplored. In this paper, we propose an actor critic method for safe RL that uses an implicit quantile network to approximate the distribution of accumulated safety-costs. Using an accurate estimate of the distribution of accumulated safety-costs, in particular of the upper tail of the distribution, greatly improves the performance of risk-averse RL agents. The empirical analysis shows that our method achieves good risk control in complex safety-constrained environments.

1 Risk-Averse Constrained RL

Traditional expectation-based safe RL methods maximize the return under the premise that the average performance is safe. In this way, RL agents are not aware of the potential risks because of the randomness in cost-return, which is generated by the stochastic policy and the dynamics of the environment. In safety-critical domains, the optimal policies are expected to be more robust, i.e., to have a lower risk of hazardous events even for stochastic or heavy-tailed cost-return.

In this paper, we consider the discounted *return* and discounted *cost-return*, accumulated discounted rewards and costs, respectively, from (s, a) as

$$\begin{aligned} Z^r_\pi(s,a) &= \sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) \mid s_0 = s, a_0 = a, \text{ and} \\ Z^c_\pi(s,a) &= \sum_{t=0}^{\infty} \gamma^t c(s_t,a_t) \mid s_0 = s, a_0 = a. \end{aligned} \tag{1}$$

We will refer to the cost-return $Z^c_\pi(s,a)$ as C whenever π, s and a are clear from the context. So, we have $Q^r_\pi(s,a) = \mathbb{E}[Z^r_\pi(s,a)]$, and $Q^c_\pi(s,a) = \mathbb{E}[Z^c_\pi(s,a)] = \mathbb{E}[C]$.

Considering the probability distribution of cost-returns $p^\pi(C)$ induced by the aleatoric uncertainty of the environment and the policy π, we model the safety-constrained RL problem in a more risk-averse way than the traditional expectation-based formulation. We focus on the α-percentile $F^{-1}_C(1-\alpha)$, where F_C is the CDF of $p^\pi(C \mid s,a)$, so we can get the Conditional Value-at-Risk (CVaR) [Rockafellar and Uryasev, 2000]:

$$\Gamma_\pi(s,a,\alpha) \doteq \text{CVaR}^\alpha_\pi(C) = \mathop{\mathbb{E}}_{p^\pi}[C \mid C \geq F^{-1}_C(1-\alpha)], \tag{2}$$

where a positive scalar $\alpha \in (0,1]$ is used to define the risk level. A smaller α ($\alpha \to 0$) is expected to be more pessimistic and risk-averse. Conversely, a larger value of α leads to a less risk-averse behavior, with $\alpha = 1$ corresponding to the risk-neutral case. The following definition gives us a new constraint to learn risk-averse policies, which differs from the traditional constraint.

Definition 1 (Safety based on CVaR)**.** *Given the risk level α, a policy π is safe if it satisfies $\Gamma_\pi(s_t,a_t,\alpha) \leq d \quad \forall t$, where $(s_t,a_t) \sim \mathcal{T}_\pi$ and $s_0 \sim \iota$.*

Now we can generalize the maximum entropy RL with the above risk-sensitive safety constraints. That is, the optimal policy in a constrained RL problem might be stochastic therefore it is reasonable to seek a policy with some entropy. So, the policy is optimized to satisfy

$$\max_\pi \mathbb{E}\left[Z^r_\pi\right] \text{ s.t. } \begin{cases} \text{CVaR}^\alpha_\pi(C) \leq d \\ \mathop{\mathbb{E}}_{(s_t,a_t)\sim\mathcal{T}_\pi}[-\log(\pi_t(a_t \mid s_t))] \geq h \quad \forall t. \end{cases} \tag{3}$$

With (3) it is possible to solve safe RL problems using the Soft Actor Critic (SAC) framework, maintaining a minimum expected entropy [Haarnoja *et al.*, 2018].

2 Worst-Case Soft Actor Critic

The *risk-averse constrained RL* problem (3) can be solved by Worst-Case Soft Actor Critic (WCSAC) algorithm [Yang *et al.*, 2021]. WCSAC generalizes SAC-Lag [Ha *et al.*, 2020], regarded as WCSAC with $\alpha = 1$, such that $\Gamma_\pi(s,a,1) = Q^c_\pi(s,a)$ (2). WCSAC use a separate Gaussian safety critic (parallel to the reward critic for the return) to estimate the distribution of C instead of computing a point estimate of the expected cost-return, as the SAC-Lag algorithm. We will refer to the WCSAC with a Gaussian safety critic as WCSAC-GS in the following parts of the paper. To obtain the cost-return distribution, $p^\pi(C \mid s,a)$ is approximated with a Gaussian,

$$Z^c_\pi(s,a) \sim \mathcal{N}(Q^c_\pi(s,a), V^c_\pi(s,a)), \tag{4}$$

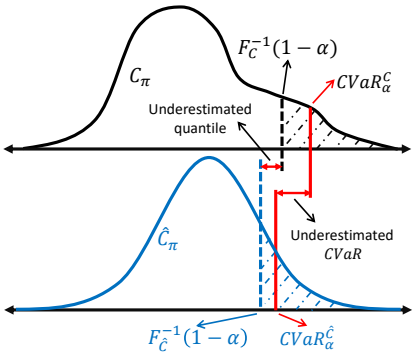

Figure 1: Unreliability of Gaussian approximation. The top curve depicts the true cost distribution, while the bottom curve depicts the estimated Gaussian distribution, based on the correct mean and standard deviation. In this case, the $(1-\alpha)$-quantile and corresponding CVaR are underestimated.

where $V_\pi^c(s,a) = \mathbb{E}_{p^\pi}[C^2 \mid s,a] - (Q_\pi^c(s,a))^2$ is the variance of the cost-return.

At each iteration, $Q_\pi^c(s,a)$ and $V_\pi^c(s,a)$ can be estimated. Since the Gaussian distribution results in a closed form estimation for CVaR [Khokhlov, 2016; Tang *et al.*, 2020], the new safety measure for risk level α is computed by

$$\Gamma_\pi(s,a,\alpha) \doteq Q_\pi^c(s,a) + \alpha^{-1}\phi(\Phi^{-1}(\alpha))\sqrt{V_\pi^c(s,a)}, \quad (5)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability distribution function (PDF) and the cumulative distribution function (CDF) of the standard normal distribution.

For a certain risk level α, WCSAC optimizes the policy π until it satisfies the safety criterion $\Gamma_\pi(s_t,a_t,\alpha) \le d \quad \forall t$ according to Definition 1. In the policy improvement step, the policy is updated towards the exponential of the new policy evaluation $X_{\alpha,\omega}^\pi(s,a) = Q_\pi^r(s,a) - \omega\Gamma_\pi(s,a,\alpha)$, where an adaptive safety weight ω is used to manage a trade-off between safety and performance. The role of safety changes over the training process. As the policy becomes safe, the influence of the safety term wanes, then the return optimization will play a greater role in our formulation.

3 Safety Critic With Quantile Regression

Although the Gaussian approximation leverages distributional information to attain more risk-averse policies, only an additional variance is estimated compared to regular constrained RL methods. This means the information of the experiences collected are only used to a limited extent. Thus, the Gaussian approximation does not possess the general advantages of distributional RL algorithms.

Besides, it is not always appropriate to approximate the cost-return by a Gaussian distribution, as shown in Figure 1, since the contribution from the tail of the cost distribution might be underestimated. In this case, the agent might converge to an unsafe policy, according to (3). In this section, we present a distributional safety critic modeled by an *implicit quantile network* (IQN) [Dabney *et al.*, 2018], which provides a more precise estimate of the upper tail part of the

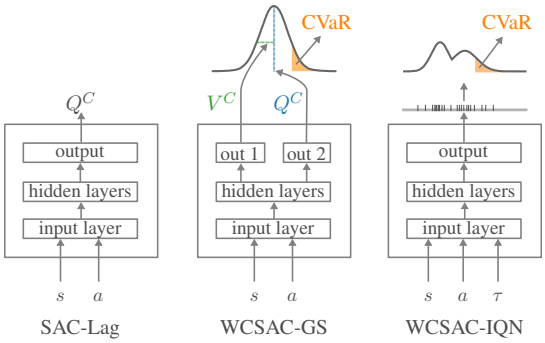

Figure 2: Overview of the safety critics. The traditional safety critic only estimates the average of the cost-return distribution Q^C, while the critics of the WCSAC-GS algorithms keep track of the full distribution by a Gaussian approximation. WCSAC-IQN models the distributional safety critic by IQN.

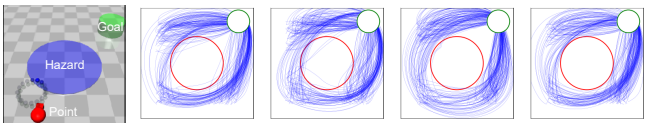

Figure 3: Trajectory analysis in StaticEnv [Ray *et al.*, 2019; Yang *et al.*, 2021]. With a higher risk level α (IQN-0.9), WCSAC-IQN can attain risk-neutral performance similar to expectation-based (IQN-1.0) methods. WCSAC-IQN algorithms can become more risk-averse by setting lower risk level α.

distribution. Henceforth, we refer to WCSAC with a safety critic modeled by IQN as WCSAC-IQN. This is an extended abstract of our journal paper [Yang *et al.*, 2022].

We model the cost-return distribution by an IQN (safety-IQN), regarded as the safety critic. Safety-IQN maps the samples from a base distribution (usually $\tau \sim U([0,1])$) to the corresponding quantile values of the cost-return distribution. In theory, by adjusting the capacity of the neural network, safety-IQN can fit the cost-return distribution with arbitrary precision, which is essential for safety-critical problems.

We denote $F_C^{-1}(\tau)$ as the quantile function for the cost-return C and, for clarity of exposition, we define $C^\tau = F_C^{-1}(\tau)$. We use θ_C to parameterize the safety-IQN. The approximation is implemented as $\hat{C}^\tau(s,a) \leftarrow f_{IQN}(s,a,\tau \mid \theta_C)$, which also takes the quantile fraction τ as the input of the model, so that it uses the neural network to fit the entire continuous distribution. When training f_{IQN}, two quantile fraction samples $\tau,\tau' \sim U([0,1])$ at time step t are used to get the sampled TD error:

$$\delta_t^{\tau,\tau'} = c_t + \gamma C^{\tau'}(s_{t+1},a_{t+1}) - C^\tau(s_t,a_t). \quad (6)$$

The quantile values of safety-IQN are learned based on the Huber quantile regression loss [Huber, 1964]:

$$\rho_\tau^\kappa(\delta) = |\tau - \mathbb{I}\{\delta<0\}|\frac{\mathcal{L}_\kappa(\delta)}{\kappa}, \quad (7)$$

where

$$\mathcal{L}_\kappa(\delta) = \begin{cases} \frac{1}{2}\delta^2, & \text{if } |\delta \le \kappa| \\ \kappa(|\delta - \frac{1}{2}\kappa|), & \text{otherwise} \end{cases}, \quad (8)$$

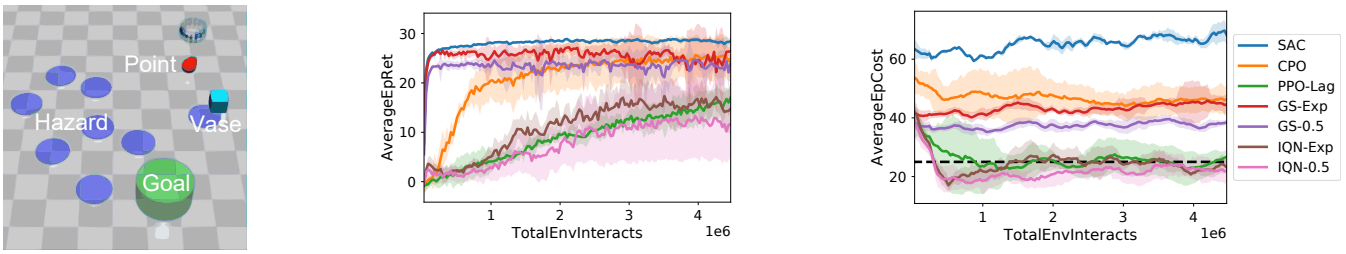

Figure 4: Performance of the algorithms during training in terms of mean (solid lines) and ± 1 standard deviation (shaded area) of the runs within an epoch. The black dashed lines indicate the safety thresholds. In the *Safety Gym* environment [Ray *et al.*, 2019], both WCSAC-GS and WCSAC-IQN get more risk-averse performance with lower risk-level α. With benefits from the quantile regression to enhance exploration and avoid overfitting, WCSAC-IQN has the best performance in safety compared to all the baselines.

where κ is the threshold to make the loss within an interval $[-\kappa, \kappa]$ quadratic but a regular quantile loss if outside the interval. Then, we can get the loss function for safety-IQN, i.e.,

$$J_C(\theta_C) = \mathop{\mathbb{E}}_{(s_t,a_t,c_t,s_{t+1})\sim\mathcal{D}} \mathcal{I}_C(s_t, a_t, c_t, s_{t+1} \mid \theta_C), \tag{9}$$

where

$$\begin{aligned}
&\mathcal{I}_C(s_t, a_t, c_t, s_{t+1} \mid \theta_C) \\
&\underset{(a)}{=} \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\mathcal{B}^\pi C}[\rho_{\tau_i}^{\kappa}(\mathcal{B}^\pi C(s_t, a_t) - C^{\tau_i}(s_t, a_t))] \\
&\underset{(b)}{=} \sum_{i=1}^{N} \mathop{\mathbb{E}}_{C}[\rho_{\tau_i}^{\kappa}(c_t + \gamma C(s_{t+1}, a_{t+1}) - C^{\tau_i}(s_t, a_t))] \\
&\underset{(c)}{\doteq} \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \rho_{\tau_i}^{\kappa}(c_t + \gamma C^{\tau'_j}(s_{t+1}, a_{t+1}) - C^{\tau_i}(s_t, a_t)) \\
&\underset{(d)}{=} \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \rho_{\tau_i}^{\kappa}(\delta_t^{\tau_i,\tau'_j}).
\end{aligned} \tag{10}$$

In (10): (a) indicates that the total loss of all the target quantiles $\tau_i, i = 1, \cdots, N$ is computed at once, and applies the distributional Bellman operator $\mathcal{B}$ [Bellemare *et al.*, 2017], (b) expands the Bellman operator, taking an action for the next state sampled from the current policy $a_{t+1} \sim \pi(\cdot \mid s_{t+1})$, (c) introduces τ_j to estimate the TD target, and (d) uses (6). Since we base our estimate of the distribution of cost-return on a quantile-parameterized approximation, we approximate the CVaR based on the expectation over the values of the quantile τ as $\Gamma_\pi(s, a, \alpha) \doteq \mathbb{E}_{\tau\sim U([1-\alpha,1])}[C_\pi^\tau(s, a)]$. This allows us to estimate $\Gamma_\pi(s, a, \alpha)$ at each update step using K i.i.d. samples of $\widetilde{\tau} \sim U([1 - \alpha, 1])$:

$$\Gamma_\pi(s, a, \alpha) \doteq \frac{1}{K} \sum_{k=1}^{K} C_\pi^{\widetilde{\tau}_k}(s, a). \tag{11}$$

Our method efficiently estimates the CVaR using a sampling approach. This can attain higher accuracy due to the quantile regression framework. We also highlight that this method still estimates the full distribution, sampling τ, τ' from $U([0, 1])$ to compute the safety critic loss. We use (11) only when estimating the CVaR to compute the Lagrangian safety loss. Based on the new safety measure, we refer the reader to [Yang *et al.*, 2021] for the policy optimization.

References

[Bellemare *et al.*, 2017] Marc G Bellemare, Will Dabney, and Rémi Munos. A Distributional Perspective on Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[Dabney *et al.*, 2018] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit Quantile Networks for Distributional Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1096–1105, 2018.

[Ha *et al.*, 2020] Sehoon Ha, Peng Xu, Zhenyu Tan, Sergey Levine, and Jie Tan. Learning to Walk in the Real World with Minimal Human Effort. arXiv preprint arxiv:2002.08550, 2020.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft Actor-Critic Algorithms and Applications. arXiv preprint arxiv:1812.05905, 2018.

[Huber, 1964] Peter J Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[Khokhlov, 2016] Valentyn Khokhlov. Conditional value-at-risk for elliptical distributions. *Evropskỳ časopis ekonomiky a managementu*, 2(6):70–79, 2016.

[Ray *et al.*, 2019] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning, 2019. https://cdn.openai.com/safexp-short.pdf.

[Rockafellar and Uryasev, 2000] R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2(3):21–41, 2000.

[Tang *et al.*, 2020] Yichuan Charlie Tang, Jian Zhang, and Ruslan Salakhutdinov. Worst Cases Policy Gradients. In *3rd Annual Conference on Robot Learning*, pages 1078–1093. PMLR, 2020.

[Yang *et al.*, 2021] Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs T J Spaan. WCSAC: Worst-Case Soft Actor Critic for Safety-Constrained Reinforcement Learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

[Yang *et al.*, 2022] Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs T J Spaan. Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 2022. In press.