Extreme Risk mitigation in Reinforcement Learning

Anonymous Authors[1]

1. Proof of convergence of the Bellman Operator for EVT-RL

Definition 1: For any two random variables $J_1(s,a)$ and $J_2(s,a)$ with distributions $Z_1(s,a)$ and $Z_2(s,a)$ with inverse CDF functions $F^{-1}_{J_1(s,a)}$ and $F^{-1}_{J_2(s,a)}$ respectively, the Wasserstein distance d_p is defined as:

$$d_p\Big(F_{J_1(s,a)}, F_{J_2(s,a)}\Big) = \Big(\int_0^1 |F^{-1}_{J_1(s,a)}(u) - F^{-1}_{J_2(s,a)}(u)|^p\Big)^{1/p}$$

Equivalently, the maximal Wasserstein distance $\bar{d}_p$ is defined as:

$$\bar{d}_p(F_{J_1}, F_{J_2}) = \sup_{s,a} d_p\Big(F_{J_1(s,a)}, F_{J_2(s,a)}\Big)$$

Property 1: **For a scalar constant r, the shifted random variables $J_1(s,a)+r$ and $J_2(s,a)+r$ have**

$$d_p\Big(F_{J_1(s,a)+r}, F_{J_2(s,a)+r}\Big) = d_p\Big(F_{J_1(s,a)}, F_{J_2(s,a)}\Big)$$

Property 2: **For a real constant scaling factor γ, the scaled random variables $\gamma J_1(s,a)$ and $\gamma J_2(s,a)$ have**

$$d_p\Big(F_{\gamma J_1(s,a)}, F_{\gamma J_2(s,a)}\Big) \leq \gamma d_p\Big(F_{J_1(s,a)}, F_{J_2(s,a)}\Big)$$

Theorem 1: **Let $\mathcal{Z}$ denote the space of all state action value distributions. For the state action value distribution $Z(s,a) = Z_L(s,a) + (1-\eta)Z_H(s,a)$, where Z_L represents the non-tail region of Z and Z_H represents the tail region of Z, the Bellman operator $T^\pi : \mathcal{Z} \times \mathcal{Z}$, is a γ contraction under the maximal Wasserstein distance metric $\bar{d}_p$.**

Note: For notational convenience, we express $d_p\Big(F_{J_1(s,a)}, F_{J_2(s,a)}\Big)$ as $d_p\Big(Z_1(s,a), Z_2(s,a)\Big)$.

Proof:

$$\begin{aligned}
&d_p(T^\pi Z_1, T^\pi Z_2) \qquad (1)\\
&= d_p\Big(r(s,a) + \gamma\Big[Z_{L_1}(s',a') + (1-\eta)Z_{H_1}(s',a')\Big], r(s,a) + \gamma\Big[Z_{L_2}(s',a') + (1-\eta)Z_{H_2}(s',a')\Big]\Big)\\
&= d_p\Big(\gamma\Big[Z_{L_1}(s',a') + (1-\eta)Z_{H_1}(s',a')\Big], \gamma\Big[Z_{L_2}(s',a') + (1-\eta)Z_{H_2}(s',a')\Big]\Big) \qquad \because \textbf{Property 1}\\
&\leq \gamma d_p\Big(\Big[Z_{L_1}(s',a') + (1-\eta)Z_{H_1}(s',a')\Big], \Big[Z_{L_2}(s',a') + (1-\eta)Z_{H_2}(s',a')\Big]\Big) \qquad \because \textbf{Property 2}\\
&= \gamma d_p\Big(Z_1(s',a'), Z_2(s',a')\Big)
\end{aligned}$$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

$$\begin{aligned}\bar{d_p}(T^\pi Z_1, T^\pi Z_2) &= \sup_{s,a} d_p\Big(T^\pi Z_1(s,a), T^\pi Z_2(s,a)\Big) \\ &\leq \gamma \sup_{s,a} d_p\Big(Z_1(s,a), Z_2(s,a)\Big) \qquad \because \text{From Eqn. 1} \\ &= \gamma \bar{d_p}\Big(Z_1(s,a), Z_2(s,a)\Big)\end{aligned} \tag{2}$$

From the above equation, we prove that the T^π operator is a contraction and that the Bellman update with the GPD tail distribution converges.

2. MLE Estimation of the parameters of the GPD Distribution

The CDF of the GPD distribution $F_{\xi,\sigma}(x)$ is given by:

$$\begin{cases} 1 - \Big(1 + \frac{\xi x}{\sigma}\Big)^{-1/\xi} \text{ for } \xi \neq 0 \\ 1 - \exp(-\frac{x}{\sigma}) \text{ for } \xi = 0 \end{cases}$$

The log-density function (log-PDF) of the same GPD distribution is given by:

$$\log f_{\xi,\sigma}(x) = \begin{cases} -\log(\sigma) + \Big(-1/\xi - 1\Big)\log\Big(1 + \xi x/\sigma\Big) \text{ for } \xi \neq 0 \\ -\log(\sigma) - x/\sigma \text{ for } \xi = 0 \end{cases} \tag{3}$$

$$\frac{\partial \log f_{\xi,\sigma}}{\partial \xi} = \Big(-1/\xi - 1\Big)\Big(\frac{1}{1+\xi x/\sigma}\Big)\cdot\frac{x}{\sigma} + \log\Big(1+\xi x/\sigma\Big)\Big(1/\xi^2\Big) = 0$$

$$\frac{\partial \log f_{\xi,\sigma}}{\partial \sigma} = -\frac{1}{\sigma} + \Big(-1/\xi - 1\Big)\cdot\frac{1}{1+\xi x/\sigma}\cdot\xi x = 0$$

However when $\xi = 0$; we are left with the MLE estimation of the parameter σ of the exponential distribution

$$\frac{\partial \log f_{0,\sigma}}{\partial \sigma} = -\frac{1}{\sigma} + \frac{x}{\sigma^2} = 0 \tag{4}$$

In our experiments we assume the limiting case of $\xi = 0$ so that the MLE estimation is done for the exponential distribution parameter σ. We make the GPD parameter $\sigma(s,a)$ be parameterized and represent it by $\sigma_\theta(s,a)$. Given a batch of transition tuples (s_i, a_i, r_i, s_i') ; $i = 1 \to B$, where B is the batch size, we source samples from the GPD distribution by sampling K samples x_k from $Z_H^\pi(s,a)$, the constructed GPD distribution. We then compute the empirical log-likelihood loss $\mathcal{L}$ that needs to be maximized.

$$\mathcal{L}_\theta = \frac{1}{B}\sum_{i=1}^{B}\cdot\frac{1}{K}\sum_{k=1}^{K}\Big[-\log\Big(\sigma_\theta(s_i,a_i)\Big) - \frac{x_k}{\sigma_\theta(s_i,a_i)}\Big] \tag{5}$$

where $x_k \sim Z_H^\pi(s,a)$, K is the number of samples sampled from the GPD distribution $Z_H^\pi(s,a)$. We set $K = 100$ and $B = 128$ for all experimentation. We perform gradient ascent on the parameters θ by using the empirical loss:

$$\theta : \theta + \alpha \cdot \nabla_\theta \mathcal{L}_\theta \tag{6}$$

where α is the learning rate.

3. Performance (Mean reward) of the risk averse agents

We also probed the agents to understand how the agents perform with respect to obtaining rewards while being risk averse.

Denoting $R(s, a)$ as the original non-penalized reward, we define the stochastic penalized reward $r(s, a)$ to be:

HalfCheetah-v3:

$$r(s, a) = R(s, a) + \mathbf{1}_{v>\alpha} L.\mathcal{B}_p$$

where $L = -100$, $\alpha = 0.8$ and $p = 0.1$ or $p = 0.05$.

Hopper-v3:

$$r(s, a) = R(s, a) + \mathbf{1}_{|\theta|>\alpha} L.\mathcal{B}_p$$

where $L = -50$, $\alpha = 0.03$ and $p = 0.1$ or $p = 0.05$

Walker2d-v3:

$$r(s, a) = R(s, a) + \mathbf{1}_{|\theta|>\alpha} L.\mathcal{B}_p$$

where $L = -30$, $\alpha = 0.2$,$p = 0.1$ or $p = 0.05$ and $\mathcal{B}_p$ is the Bernoulli random variable with parameter p.

We let the maximum episode length to be 200.

We indicate the cumulative reward as the cumulative reward for one run during inference. The mean $\pm$ standard deviation is shown for 5 trained agents as before. The 5 agents were trained for 30,000 time steps, with episode extending to a maximum of 200 time steps i.e. (max episode length = 200)

Method	Environment	Mean Overshoot	Percentage Failure	Cumulative Reward
DDPG RAAC	Half-Cheetah	0.4 ± 0.12	0.3 ± 0.18	-471.03 $\pm$ 251.18
	Hopper	0.0035 ± 0.005	0.12 ± 0.17	45.95 $\pm$ 111.07
	Walker	0.13 ± 0.18	0.18 ± 0.25	92.87 $\pm$ 101.58
DDPG EVT (Ours)	Half-Cheetah	0.22 ± 0.24	0.01 ± 0.0	**-114.59** $\pm$ 27.6
	Hopper	0.01 ± 0.0	0.06 ± 0.06	**187.93** $\pm$ 69.26
	Walker	0.0 ± 0.0	0.0 ± 0.0	**134.41** $\pm$ 98.17

Table 1. Table showing the risk averse behavior of various training algorithms for $p = 0.1$ penalization rate

We also monitor the cumulative mean reward for $p = 0.05$.

Method	Environment	Mean Overshoot	Percentage Failure	Cumulative Reward
TD3 RAAC	Half-Cheetah	0.88 ± 0.21	0.8 ± 0.06	-616.48 $\pm$ 403.43
	Hopper	0.02 ± 0.01	0.37 ± 0.2	72.79 $\pm$ 138.6
	Walker	0.12 ± 0.06	0.25 ± 0.13	193.79 $\pm$ 76.94
TD3 EVT (Ours)	Half-Cheetah	0.08 ± 0.05	0.01 ± 0.01	**-116.99** $\pm$ 5.3
	Hopper	0.04 ± 0.05	0.11 ± 0.16	**135.7** $\pm$ 81.21
	Walker	0.01 ± 0.02	0.03 ± 0.07	**167.03** $\pm$ 68.0

Table 2. Table showing the risk averse behavior of various training algorithms for $p = 0.05$ penalization rate

From Table 1 and Table2 we see that in all the above cases the EVT based agents show better performance in comparison to the baseline agents. We believe that this is the case as the EVT based agents rarely cross over the

threshold α and a result accumulate less stochastic penalties over the episode of operation. However, in agents that cross the threshold α more frequently (like DDPG-RAAC and TD3-RAAC), we see that the cumulative reward accumulates more stochastic penalty leading to the lower cumulative rewards on the whole.

4. Risk Awareness and Performance with changing Penalization rate p

In this section we change the penalization rate p of the reward for the Hopper-v3 environment $r(s,a) = R(s,a)+\mathbf{1}_{|\theta|>\alpha}L.\mathcal{B}_p$. We move from a very rare reward penalty of $p = 0.03$ to fairly frequent penalty rate of $p = 0.1$. For our experimentation purpose, we fix the threshold quantile $\eta = 0.95$ for both TD3-RAAC and TD3-EVT. We observe that the for higher values of p, both agents perform well in terms of risk avoidance and also accmulate fairly high cumulative rewards. **From Figure1, and Table 3 however, as the penalization rate p decreases (rare risky events), the EVT based agent shows very small percentage failure while the cumulative reward is still very high.**

Method	Penalization rate (p)	Percentage Failure	Cumulative Reward
TD3 RAAC	0.03	0.55	101.6
	0.05	0.67	146.04
	0.07	0.3	76.7
	0.1	0.0	203.6
TD3 EVT	0.03	0.07	222.27
	0.05	0.0	200.16
	0.07	0.0	228.06
	0.1	0.0	197.6

Table 3. Table showing the the percentage failure and also the cumulative reward as the penalization rate p is varied for the Hopper-v3 environment (with fixed threshold quantile $\eta = 0.95$).The TD3-EVT agent shows very small percentage failure for even extremely rare penalizaton rate while achieving high cumulative rewards.

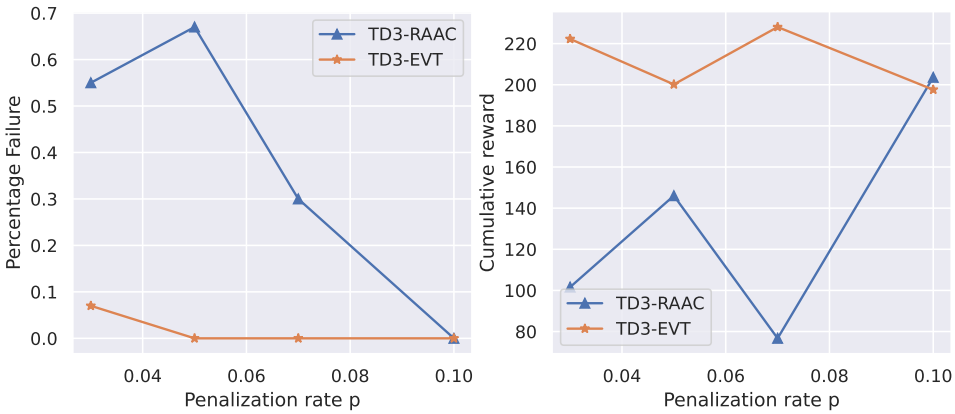

Figure 1. Percentage failure and cumulative reward as a function of varying penalization rate p with fixed threshold quantile $\eta = 0.95$.

5. Risk Awareness and Performance with changing threshold quantile parameter η

We next address the sensitivity to risk awareness when the threshold quantile parameter η is changed for the Hopper-v3 environment. We fix the penalization rate $p = 0.05$ while changing η continuously.

The value of η indicates risk aversion. Higher values of η imply better risk aversion.

We observe that the choice of the threshold quantile η plays an important role in risk aversion. If η is relatively small like $\eta = 0.9$, both the non-EVT and EVT based agents do not exhibit good risk aversion. **From Figure2, and Table 4 however,**

Method	η	Percentage Failure	Cumulative Reward
TD3 RAAC	0.97	0.25	171.56
	0.95	0.32	167.12
	0.92	0.24	192.17
	0.9	0.55	81.76
TD3 EVT	0.97	0.0	202.45
	0.95	0.0	200.37
	0.92	0.43	162.59
	0.9	0.4	192.53

Table 4. Table showing the the percentage failure and also the cumulative reward as the threshold quantile η is varied for the Hopper-v3 environment(with fixed $p = 0.05$) .The TD3-EVT agent shows zero percentage failure for higher values of η unlike the TD3-RAAC agent, while achieving high cumulative rewards.

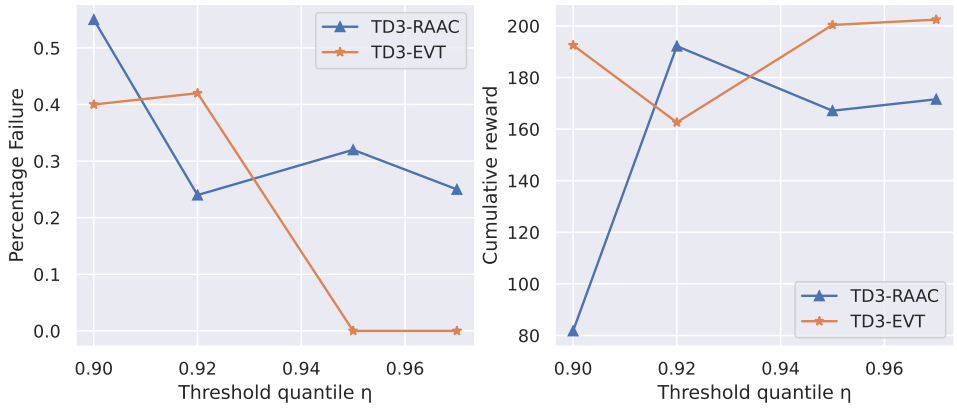

Figure 2. Percentage failure and cumulative reward as a function of varying the threshold quantile η with fixed penalization rate $p = 0.05$.

as η increases, the EVT based agents become very risk averse while producing good cumulative rewards. However, for the non-EVT agent (TD3-RAAC), even larger η values do not aid in risk aversion for the same level of penalization rate $p = 0.05$. This underscores the effectiveness of using EVT theory for risk aversion in reinforcement learning.

6. Clarification on Eqn.7 in the manuscript

The CVaR (Conditional value at risk), is a risk measure, that is optimized in risk averse RL applications. The CVaR denotes the average worst case performance by integrating the quantiles of the state action value distribution between two quantile levels x_1 and x_2. The conditional value at risk (CVaR) is given by:

$$\text{CVaR}^{\pi}(s) = \frac{1}{x_2 - x_1} \int_{\tau=x_1}^{x_2} Z^{\pi}\Big(s, \pi(s)\Big)\Big|_{\tau} \tag{7}$$

Since we model $Z^{\pi}(s, a)$ to be a quantile state action value function, we can query it at any quantile level τ i.e. $Z^{\pi}(s, a)|_{\tau}$. Since the rewards are negated, we consider the right tail of the state action value distribution $Z^{\pi}(s, a)$ for risk aversion which is why the values of x_1 are closer to 1.0.

The integration of the $Z^{\pi}(s, a)$ for larger values of τ indicates the worst case realization of the original non-negated state action value function.

Thus, the optimal policy $\pi^*(s)$ is given by:

$$\pi^* = \arg\min_{\pi} \frac{1}{x_2 - x_1} \int_{\tau=x_1}^{x_2} Z\Big(s, \pi(s)\Big)\Big|_{\tau} \tag{8}$$

We set the values of $x_1 = 0.95$ and $x_2 = 1.0$ for all experiments of the RAAC and EVT agents.

We will change (Eqn.7 in the main manuscript) to include an $\arg\min$ instead of $\arg\max$ in the revised version.

7. Corrected Typo for Figure 2 in the manuscript

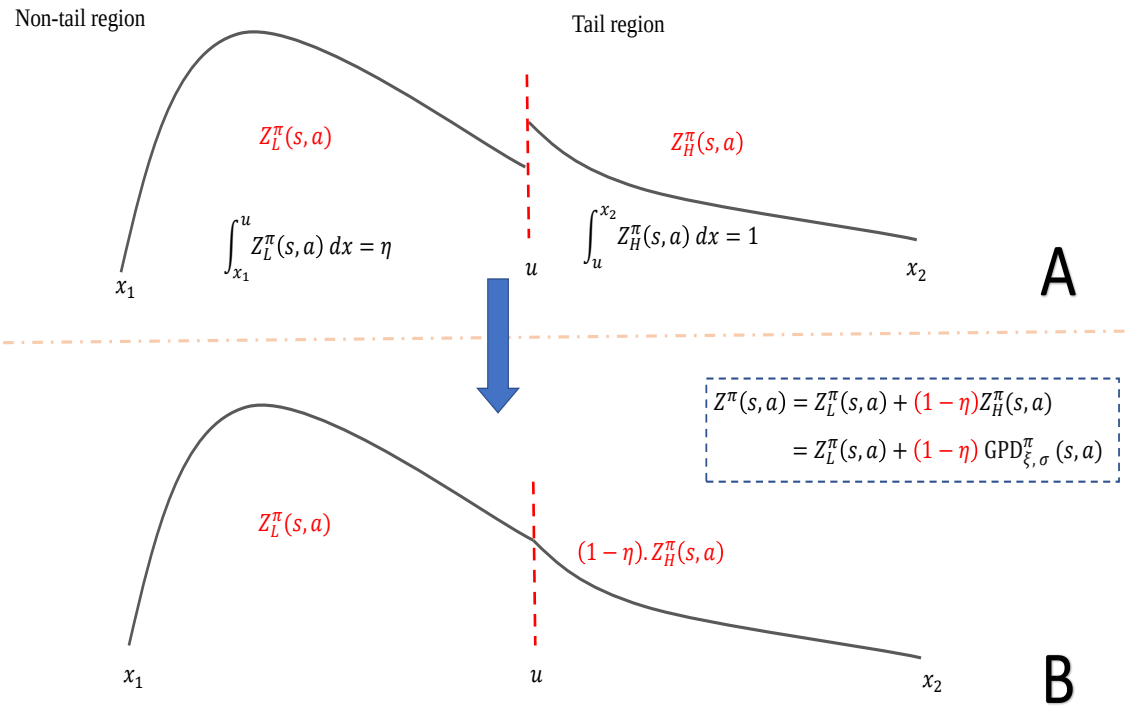

Figure 3. Modeling the tail and non-tail distributions of the state action value function. The area under the non-tail distribution $Z_L^\pi(s,a)$ is η. The area under the tail region of the distribution $Z_H^\pi(s,a)$ is 1.

The figure above is corrected to indicate that the area under $Z_L^\pi(s,a)$ is η and the area under $Z_H^\pi(s,a)$ is 1.

8. Some experimental details

We use an actor critic framework , where the critic is a quantile critic. Both the actor and critic have 3 layers with hidden size being 128.

During training and inference, the max episode length of the agent is set to 200. During training, the agents were trained for 30,000 time steps on the whole. The bacth size $B = 128$ and we set K, the number of samples sampled from the GPD distribution to 100. We set the learning rates for the actor and critic to 0.001 in all cases. The discount factor $\gamma = 0.99$ for all cases too. The soft update paramter $\tau = 0.02$ for all our experiments on the Hopper-v3 and Walker2d-v3, while $\tau = 0.01$ for the HalfCheetah-v3 environment.

The probabilistic proxy reward model, $\hat{P}_R(s,a)$ is constructed as a quantile regression model. We follow the same architecture of the critic with 3 hidden layers and hidden layer size being 128. We train the proxy reward model by the quantile regression loss function with rewards stored in the replay buffer. The learning rate here is also set to 0.001.

9. Better representation for Algorithm 1

Algorithm 1 EVT RL

Input: Initialize $\sigma(s,a), \xi(s,a), Z(s,a)$
Select a threshold quantile level η
POLICY ITERATION:
For tuple $(s, a, r, s, a' = \pi(s'))$
$x \sim GPD\Big(\xi(s',a'), \sigma(s',a')\Big)$
Define $Z'_H = Z(s,a)|_{\tau=\eta} + x$
Define $Z'_L = Z(s,a)|_{\tau=\tau_0}$; where $\tau_0 \sim \mathrm{Unif}(0,\eta)$
Sample $p \sim \mathrm{Ber}_\eta$
Define Bellman target:
$Z_T = r + \gamma[\mathbb{1}_{p=0} Z'_H + \mathbb{1}_{p=1} Z'_L]$
MLE Estimation for $\xi(s,a), \sigma(s,a)$
$y \sim Z(s,a)|_{\tau>\eta}$
$\xi(s,a), \sigma(s,a) = \mathrm{MLE}[\mathrm{GPD}(y - Z(s,a)|_{\tau=\eta}]$
POLICY IMPROVEMENT:
Update policy π according to Eqn.8.