Risk-Sensitive Policy Optimization via Predictive CVaR Policy Gradient

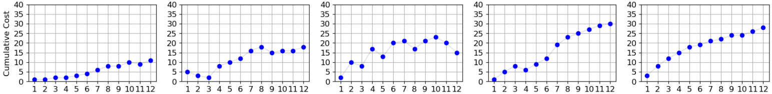
Ju-Hyun Kim Seungki Min

{vgb8111, skmin}@kaist.ac.kr
Department of Industrial and Systems Engineering
KAIST

ICML 2024, Jul 2024

Motivating Example

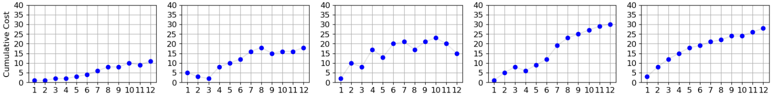
- Given a fixed policy π, consider 5 sample trajectories. For a risk-neutral RL, we can utilize the whole sample trajectories.



$$\mathbb{E}[C_{1:T}] \approx \frac{1}{N} \sum_{i \in [N]} C_{1:T}^{(i)}$$

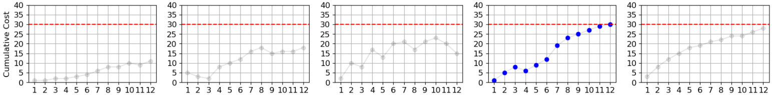
Motivating Example

- Given a fixed policy π, consider 5 sample trajectories. For a risk-neutral RL, we can utilize the whole sample trajectories.



$$\mathbb{E}[C_{1:T}] \approx \frac{1}{N} \sum_{i \in [N]} C_{1:T}^{(i)}$$

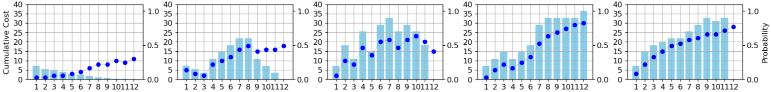
- For a CVaR RL, one may utilize **the worst q fraction** among the whole sample trajectories. $\rightarrow$ High variance and low sample efficiency



$$\mathrm{CVaR}_q[C_{1:T}] = \mathbb{E}[C_{1:T} | C_{1:T} \geq \mathrm{VaR}_q[C_{1:T}]] \approx \frac{1}{qN} \sum_{qNworst} C_{1:T}^{(i)}$$

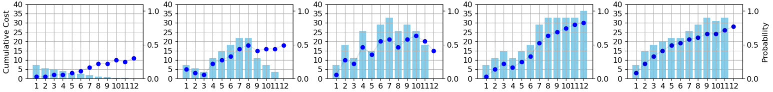
Main Idea

- We introduce a "*predictive tail probability process*" $Q^\pi = (Q_t^\pi)_{t\in[T]}$.
  - In each period, it predicts the probability that the current sample path ends up being one of the worst q fraction of outcomes.

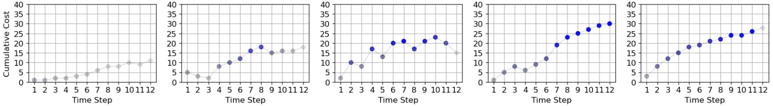


Main Idea

- We introduce a "*predictive tail probability process*" $Q^{\pi} = (Q^{\pi}_t)_{t \in [T]}$.
  - In each period, it predicts the probability that the current sample path ends up being one of the worst q fraction of outcomes.



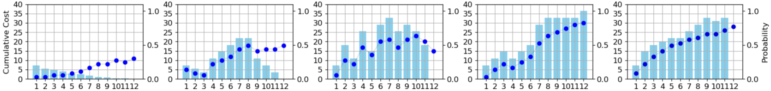
- We reformulate CVaR as an expectation of reweighted cost realization by the "*predictive tail probability*".



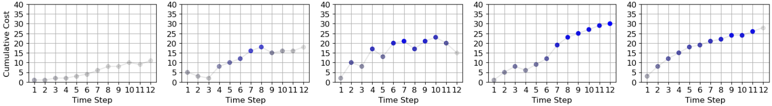
$$\mathrm{CVaR}_q[C_{1:T}] = \frac{1}{q}\mathbb{E}[\sum_{t \in [T]} Q_t C_t] \approx \frac{1}{qN} \sum_{i \in [N]} \sum_{t \in [T]} Q^{(i)}_t C^{(i)}_t$$

Main Idea

- We introduce a "*predictive tail probability process*" $Q^{\pi} = (Q_t^{\pi})_{t \in [T]}$.
  - In each period, it predicts the probability that the current sample path ends up being one of the worst q fraction of outcomes.



- We reformulate CVaR as an expectation of reweighted cost realization by the "*predictive tail probability*".



$$\mathrm{CVaR}_q[C_{1:T}] = \frac{1}{q}\mathbb{E}[\sum_{t \in [T]} Q_t C_t] \approx \frac{1}{qN} \sum_{i \in [N]} \sum_{t \in [T]} Q_t^{(i)} C_t^{(i)}$$

- For CVaR RL, we can utilize **all sample trajectories** via reweighting.
  $\rightarrow$ Low variance, high sample efficiency

Predictive CVaR Policy Gradient (PCVaR)

- Our goal is to find an optimal policy π^* solving

$$\min_{\pi \in \Pi^\Theta} \left\{ J_q(\pi) := q \cdot \mathrm{CVaR}_q^\pi \left[C_{1:T} \right] \right\}. \tag{$*$}$$

- Using reformulated CVaR objective, we change the optimization $(*)$ into adjusted optimization problem with parameters (θ, η, ϕ) as

$$\min_{\theta \in \Theta} \left\{ J(\theta, \eta, \phi) \mid \eta, \phi \ s.t \ \ldots \right\},$$

where

$$J(\theta, \eta, \phi) := \mathbb{E}\left[\sum_{t \in [T]} \hat{Q}_t C_t \right], \hat{Q}_t = f^\phi(X_{t+1}, C_{1:t} - \eta).$$

- Compared to risk-neutral PG objective, we just replace C_t as $\hat{Q}_t C_t$ (even for policy learning process).

Predictive CVaR Policy Gradient (PCVaR)

- The objective $J_q(\pi)$ in $(*)$ can be rewritten as

$$J_q(\pi) = \min_{\eta \in \mathbb{R}} \mathbb{E}^{\pi}[q\eta + (C_{1:T} - \eta)^+]. \tag{1}$$

- Given $\pi \in \Pi^{\mathcal{H}}$, optimal solution η^{π} of (1) is $\mathrm{VaR}_q(C_{1:T})$.

Predictive CVaR Policy Gradient (PCVaR)

- The objective $J_q(\pi)$ in $(*)$ can be rewritten as

$$J_q(\pi) = \min_{\eta \in \mathbb{R}} \mathbb{E}^{\pi}[q\eta + (C_{1:T} - \eta)^{+}]. \tag{1}$$

- Given $\pi \in \Pi^{\mathcal{H}}$, optimal solution η^{π} of (1) is $\mathrm{VaR}_q(C_{1:T})$.

Definition: Predictive tail probability process $Q^{\pi,\eta}$

Given $\pi \in \Pi^{\mathcal{H}}$ and $\eta \in \mathbb{R}$, $Q^{\pi,\eta} = (Q_t^{\pi,\eta})_{t \in \{0,\cdots,T\}}$ is defined as follow:

$$Q_t^{\pi,\eta} := \mathbb{P}(C_{1:T} \geq \eta | H_{t+1})$$

Predictive CVaR Policy Gradient ($\textsc{PCVaR}$)

- The objective $J_q(\pi)$ in $(*)$ can be rewritten as

$$J_q(\pi) = \min_{\eta \in \mathbb{R}} \mathbb{E}^\pi[q\eta + (C_{1:T} - \eta)^+]. \tag{1}$$

  - Given $\pi \in \Pi^{\mathcal{H}}$, optimal solution η^π of (1) is $\mathrm{VaR}_q(C_{1:T})$.

Definition: Predictive tail probability process $Q^{\pi,\eta}$

Given $\pi \in \Pi^{\mathcal{H}}$ and $\eta \in \mathbb{R}$, $Q^{\pi,\eta} = (Q_t^{\pi,\eta})_{t\in\{0,\cdots,T\}}$ is defined as follow:

$$Q_t^{\pi,\eta} := \mathbb{P}(C_{1:T} \geq \eta | H_{t+1})$$

Proposition 3.2: Reformulation of CVaR objective

If $\sum_{t\in[T]} C_t^\pi$ has no probability mass at $\eta^\pi = \mathrm{VaR}_q^\pi[C_{1:T}]$,

$$J_q(\pi) = \mathbb{E}^\pi\left[\sum_{t\in[T]} Q_t^{\pi,\eta^\pi} C_t\right].$$

Predictive CVaR Policy Gradient (PCVaR)

- We decompose the CVaR policy optimization $(*)$ into *three* optimization problems with reformulated objective.

$$\min_{\theta\in\Theta} \left\{ J(\theta,\eta,\phi) \;\middle|\; \begin{array}{l} \eta \in \arg\min_{\eta'\in\mathbb{R}} L(\theta,\eta'), \\ \phi \in \arg\min_{\phi'\in\Phi} M(\theta,\eta,\phi') \end{array} \right\},$$

where

$$J(\theta,\eta,\phi) := \mathbb{E}\left[\sum_{t\in[T]} \hat{Q}_t C_t\right], L(\theta,\eta) := \mathbb{E}\left[q\eta + (C_{1:T}-\eta)^+\right],$$

$$M(\theta,\eta,\phi) := \mathbb{E}\left[\sum_{t\in[T]} \left(\mathbb{I}\{C_{1:T}\geq\eta\} - \hat{Q}_t\right)^2\right].$$

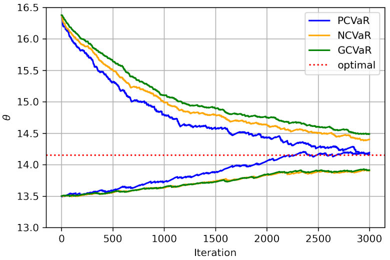
- Update θ (risk-neutral PG with $\hat{Q}_t C_t$), η (simple SGD), ϕ (typical supervised learning) in parallel.

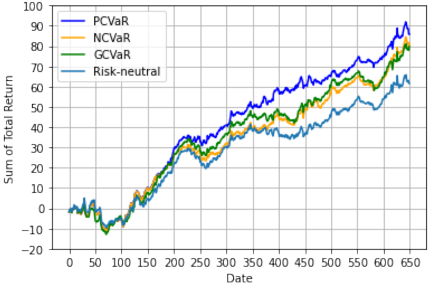
Theoretical Analysis

- Consistency of the estimators related to ϕ and η (respectively **Proposition 4.1** and **Proposition 4.2**)
- Unbiasedness of the gradient estimators of $J(\theta, \eta, \phi)$ (**Theorem 4.3**)
- Variance reduction in the gradient estimation of $\mathrm{PCV{\scriptsize A}R}$ (**Proposition 4.5**)

Numerical Experiments

- Continuous Blackjack (synthetic data)
- Pair trading (real-world data)





Evaluation point	PCVaR	NCVaR	GCVaR
$\theta = \arg\max_\theta \mathbb{E}[R_{1:T}]$	33.66	97.59	96.92
$\theta = \arg\max_\theta \mathrm{CVaR}_q[R_{1:T}]$	6.71	24.01	24.67

Contribution

- We suggest "*Predictive CVaR Policy Gradient* (PCVaR)" , relying on

$$q \cdot \mathrm{CVaR}_q\left[C_{1:T}\right] = \mathbb{E}\left[\sum_{t\in[T]} Q_t C_t\right],$$

where $\boxed{Q_t = \mathbb{P}(\text{current sample} \in \text{the worst q fractions} \mid \text{history}).}$

conditional $\mathbb{E}$ $\rightarrow$ risk-neutral $\mathbb{E}$

- PCVaR utilizes all sample trajectories.
  $\rightarrow$ Improves sample efficiency and then accelerates learning.

- PCVaR can be applied on top of any risk-neutral policy gradient algorithm.

- Its effectiveness is demonstrated with theoretical analyses and numerical experiments.