Percentile Criterion Optimization in Offline RL

Elita Lobo [1] Cyrus Cousins [1] Marek Petrik [2] Yair Zick [1]

[1]University of Massachusetts Amherst [2]University of New Hampshire

Abstract

Motivation

- In model-based offline RL, transition probabilities must be estimated from limited data.
- The *Percentile Criterion* (PC) is used to learn policies that are robust to uncertainty in transition probabilities.
- PC is optimized by constructing a special *uncertainty set* and optimizing the policy against the worst model in this set.
- Existing works use *Bayesian credible regions* as uncertainty sets which can be unnecessarily large and result in overly conservative policies!

Our Contributions:

- A dynamic programming framework that optimizes a lower bound on the percentile criterion without explicitly constructing ambiguity sets.
- Finite-sample and asymptotic analysis of the performance loss of VaR framework.
- Theoretical Comparison of VaR ambiguity sets with Bayesian credible regions-based ambiguity sets.
- Empirically analysis of the the efficacy of the framework in several domains.

Problem

Notation:

- MDP: Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r, p_0, \gamma)$
- $\tilde{P}$: Random transition probabilities with posterior distribution f.
- $\boldsymbol{v}$: Value Function
- $\boldsymbol{w}_{s,a} = \boldsymbol{r}_{s,a} + \gamma \boldsymbol{v}$: 1-step discounted value

Percentile Criterion: Computes a policy π that maximizes the returns corresponding to the worst δ-percentile model.

$$\arg\max_{\pi \in \Pi,\, y \in \mathbb{R}} \left\{ y \,\middle|\, \mathbb{P}_{\tilde{P}\sim f}\left[\rho(\pi, \tilde{P}) \geq y\right] > 1 - \delta \right\}. \tag{1}$$

Percentile Criterion is non-convex. Difficult to solve!

Existing solution: Solve it using Robust MDPs with Bayesian Credible Regions-based ambiguity sets

$$\arg\max_{\pi \in \Pi} \min_{P \in \mathcal{P}^{\mathrm{BCR}}} \rho(\pi, P).$$

$$\forall s \in \mathcal{S}, a \in \mathcal{A},\ \mathcal{P}^{\mathrm{BCR}}_{s,a} = \mathcal{P}_{s,a}(\boldsymbol{w}, \psi, q) = \left\{ \boldsymbol{p} \in \Delta^S \middle| \|\boldsymbol{p}_{s,a} - \tilde{\boldsymbol{p}}_{s,a}\|_{q,\boldsymbol{w}} \leq \psi_{s,a} \right\} \ ,$$

where weights $\boldsymbol{w}$ and size $\psi_{s,a} \forall s \in \mathcal{S}, a \in \mathcal{A}$ are optimized to minimize span of ambiguity sets while satisfying $\mathbb{P}\left[\rho(\pi, \tilde{P}) \geq \min_{P \in \mathcal{P}^{\mathrm{BCR}}} \rho(\pi, P)\right] \geq 1 - \delta$.

These sets can be unnecessarily large and result in conservative policies!

Solution: VaR Framework

Define the VaR Bellman optimality operator $\mathcal{T}_{VaR}$ as

$$(\mathcal{T}_{VaR}\boldsymbol{v})(s) = \max_{a \in \mathcal{A}} VaR_\alpha \left[\tilde{\boldsymbol{p}}^\top_{s,a} \boldsymbol{w}_{s,a}\right] \ .$$

where VaR_α of a bounded random variable X with a CDF function $F : \mathbb{R} \to [0, 1]$ is defined as $VaR_\alpha[\tilde{X}] = \inf\{z \in \mathbb{R} | F(z) > \alpha\}$.

Theoretically Proven Properties of VaR Framework

1. $\mathcal{T}_{VaR}$ is a $\gamma-$Contraction mapping when reward is independent of next-state and $\sqrt{\gamma}$-Contraction mapping otherwise.
2. VaR Framework optimizes a lower bound on the percentile-criterion.
3. The optimal VaR policy $\hat{\pi}$ solves

$$\max_{\pi \in \Pi^D} \min_{P \in \mathcal{P}^{VaR, \hat{\boldsymbol{v}}^\pi}} \rho(\pi, P) \ .$$

where for any value function $\boldsymbol{v}$, the VaR ambiguity set $\mathcal{P}^{VaR,\boldsymbol{v}}$ is defined as

$$\mathcal{P}^{VaR,\boldsymbol{v}} = \bigtimes_{s \in \mathcal{S}, a \in \mathcal{A}} \mathcal{P}^{VaR,\boldsymbol{v}}_{s,a} \text{ where } \mathcal{P}^{VaR,\boldsymbol{v}}_{s,a} = \left\{ \boldsymbol{p}_{s,a} \in \Delta^{\mathcal{S}} \middle| \boldsymbol{p}^\top_{s,a} \boldsymbol{v} \geq VaR_\alpha \left[\tilde{\boldsymbol{p}}^\top_{s,a} \boldsymbol{v}\right] \right\},$$

and $\hat{\boldsymbol{v}}^\pi$ is the fixed point of the VaR Bellman evaluation operator $\mathcal{T}^\pi_{VaR}$ for each $\pi \in \Pi^D$.

Performance Error bounds for VaR Framework

- $\hat{\boldsymbol{v}}$: fixed point of the VaR Bellman optimality operator $\mathcal{T}_{VaR}$
- $\pi^\star$: optimal policy in (1)
- $\rho^\star = VaR_\alpha \left[\rho(\pi^\star, \tilde{P})\right]$: optimal percentile returns
- $\delta \in (0, 1)$ - confidence level
- $\alpha : \delta/(2SA)$ in $\mathcal{T}_{VaR}$
- $I(\boldsymbol{p}^\star_{s,a})^{-1}$: Fisher Information matrix of the true transition probabilities $\boldsymbol{p}^\star_{s,a}$
- $\sigma_{\max} = \max_{s \in \mathcal{S}, a \in \mathcal{A}} \sqrt{\hat{\boldsymbol{w}}^\top_{s,a} I(\boldsymbol{p}^\star_{s,a})^{-1} \hat{\boldsymbol{w}}_{s,a}}$: maximum asymptotic standard deviation of the returns estimate $\tilde{\boldsymbol{p}}^\top_{s,a} \hat{\boldsymbol{w}}_{s,a}$ for state-action pair (s, a)

Theorem 1: [Performance Error] With probability at least $1 - \delta$, the performance loss with respect to $\rho^\star$ is

$$\rho^\star - \hat{\rho} \leq \frac{1}{1-\gamma} \max_{s \in \mathcal{S}} \max_{a \in \mathcal{A}} (\mathrm{VaR}_{1-\alpha}[\tilde{\boldsymbol{p}}^\top_{s,a} \hat{\boldsymbol{w}}_{s,a}] - VaR_\alpha[\tilde{\boldsymbol{p}}^\top_{s,a} \hat{\boldsymbol{w}}_{s,a}]) \ . \tag{2}$$

Theorem 2: [Asymptotic Performance Error] With probability at least $1 - \delta$, the asymptotic performance of the VaR framework $\hat{\rho}$ w.r.t. the optimal percentile returns $\rho^\star$ satisfies

$$\lim_{N \to \infty} \sqrt{N}(\rho^\star - \hat{\rho}) \leq \frac{1}{1-\gamma} \left(2\Phi^{-1}(1-\alpha)\sigma_{\max}\right) \leq \frac{1}{1-\gamma} \sqrt{8 \ln(1/\alpha)} \sigma_{\max} \ .$$

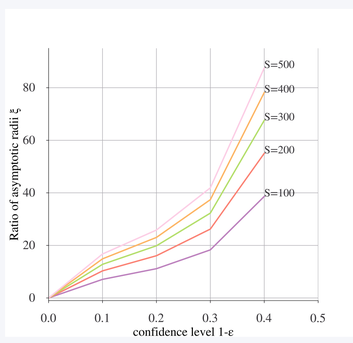
Comparison with Bayesian Credible Regions

- $\bar{P} = (\bar{\boldsymbol{p}}_{s,a})_{s \in \mathcal{S}, a \in \mathcal{A}}$: Maximum Likelihood Estimate (MLE) of transition probabilities computed from data $\mathcal{D}$
- $\Sigma = (I(\boldsymbol{p}^\star_{s,a})^{-1})_{s \in \mathcal{S}, a \in \mathcal{A}}$: covariance matrix of $\tilde{P}$.

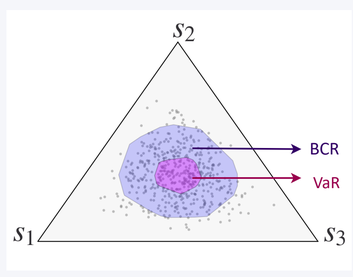
Theorem 3: [Asymptotic radius of VaR Ambiguity Sets] For all $s \in \mathcal{S}, a \in \mathcal{A}$,

$$\lim_{N \to \infty} \sqrt{N}(\mathcal{P}^{VaR}_{s,a} - \bar{\boldsymbol{p}}_{s,a}) = \left\{ \boldsymbol{p}_{s,a} \in \Delta^{\mathcal{S}} \middle| \|\boldsymbol{p}_{s,a} - \bar{\boldsymbol{p}}_{s,a}\|_{\boldsymbol{\Sigma}^{-1}_{s,a}} \leq \Phi^{-1}(1-\alpha) \right\} - \bar{\boldsymbol{p}}_{s,a} \ .$$

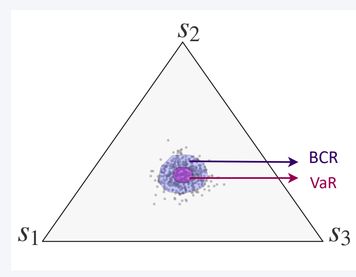
Theorem 4: [Asymptotic radius of Bayesian Credible Region] For any state s and action a, if $\mathcal{P}^{\mathrm{BCR}}_{s,a}$ be any Bayesian credible region and $\xi < \sqrt{\chi^2_{S,1-\alpha}}/\Phi^{-1}(1-\alpha)$, then, $\forall s \in \mathcal{S}, a \in \mathcal{A}$, $\lim_{N\to\infty} \sqrt{N}(\mathcal{P}^{\mathrm{BCR}}_{s,a} - \bar{\boldsymbol{p}}_{s,a}) \not\subseteq \lim_{N\to\infty} \sqrt{N}\xi(\mathcal{P}^{VaR}_{s,a} - \bar{\boldsymbol{p}}_{s,a})$.



(a) Asymptotic Radii

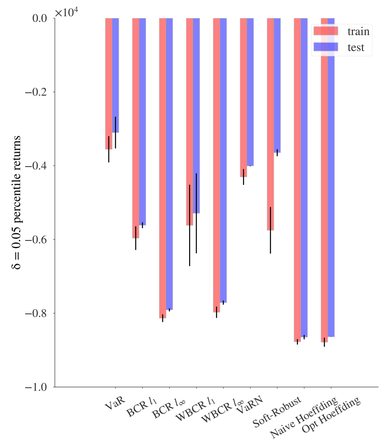


(b) Dirichlet(5,5,5)

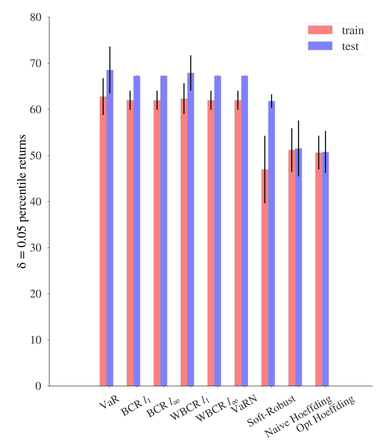


(c) Dirichlet(30,30,30)

Empirical Results on Population and Riverswim Domains



(a) Population-Small



(b) Riverswim

References

Bahram Behzadian, Reazul Hasan Russel, Marek Petrik, and Chin Pang Ho. Optimizing percentile criterion using robust MDPs. In *24th*, pages 1009--1017, 2021.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, 1970.