Quantile MDP

September 25, 2025

Quantile

- Quantile = Value-at-Risk (VaR)

$$Q_\tau(Y) = \inf\{\lambda \in \mathbb{R} : \mathbb{P}[Y \leqslant \lambda] \geqslant \tau\}$$

- However, Quantile MDP: Non-econ $\neq$ Econ
- Hence, they do not cite each other

MDP

- State: $x \in \mathsf{X}$
- Action: $a \in \mathsf{A}$
- Reward function: $r : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$
- Transition: $P(x, a, x')$
- Goal: Looking for the optimal **decision rule**

Value of a Decision Rule

- Under discounted reward criterion

$$v_\pi(x) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \beta^t r(X_t, A_t) | X_0 = x \right]$$

- Under steady-state reward criterion

$$v_\pi(x) = \lim_{t \to \infty} \mathbb{E}_\pi \left[r(X_t, A_t) | X_0 = x \right]$$

Motivation from Applications

DM wants to optimize a quantile of rewards instead of their expectation

- A physician determines the optimal drug regime by maximizing the 0.10 quantile of improvement in health

  (want it to work with at least 90% probability for the patient)

- Amazon determines the optimal cloud service by maximizing the 0.01 quantile of customers' satisfaction

  (provide service that satisfies 99% of its customers)

- Basel Accord requires banks to hold capital reserves to cover at least their 10-day 99% VaR of their loss distribution

  (in the worst 1% of cases, losses could be worse than this number)

Xia and Pan (2025)

Under steady-state reward criterion and consider all RS π

$$\mathrm{VaR}^* = \sup_{\pi \in \Pi} \mathrm{VaR}_\tau^\pi$$
$$\mathrm{VaR}_\tau^\pi = \inf \left\{ \lambda \in \mathbb{R} : \mathbb{P}_\pi [R_\infty^\pi \leqslant \lambda] \geqslant \tau \right\}$$

where $\mathbb{P}_\pi$ is the limiting distribution of the Markov chain under π

Assume each π exactly has a $\mathbb{P}_\pi$, which is independent of initial state

Transform

They transforms

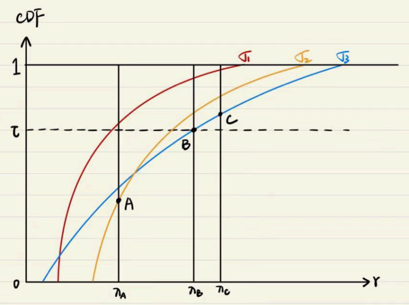
$$\mathrm{VaR}^* = \sup_{\pi \in \Pi} \mathrm{VaR}_\tau^\pi$$

to

$$\mathrm{VaR}^* = \min \left\{ \lambda : \min_{\sigma \in \Sigma} \left[\mathbb{P}_\sigma \{ R_\infty^\sigma \leqslant \lambda \} \right] \geqslant \tau \right\}$$

where Σ is all DS σ

$$\mathrm{VaR}^* = \min\left\{\lambda : \min_{\sigma}\left[\mathbb{P}_\sigma\{R_\infty^\sigma \leqslant \lambda\}\right] \geqslant \tau\right\}$$



- Consider $\Sigma = \{\sigma_1, \sigma_2, \sigma_3\}$, the figure shows their limiting distribution of reward
- λ_B and λ_C satisfy the condition since the minimum points B and C are not less than τ
- $\lambda_B \leqslant \lambda_C$, so it is VaR^*

Li et al. (2022)

Under discounted reward criterion

$$v(x) = \max_{\pi \in \Pi} Q_\tau^\pi \left[\sum_{t=0}^{\infty} \beta^t r(X_t, A_t) | X_0 = x \right]$$

where

- $\pi = \{\mu_0, \mu_1, \cdots\}$
- μ_t maps historical information $h_t = (x_0, a_0, \cdots, x_{t-1}, a_{t-1}, x_t)$ to a feasible action $a_t \in \mathsf{A}$

Challenge: it is non-additive and non-Markovian

(or, in Econ literature, not dynamically consistent)

de Castro and Galvao (2019)

$$
\begin{aligned}
&V_1^{Q_\tau}(h, x, z^t) \\
&= u(x_1^h, x_2^h, z_1) + \beta Q_\tau[u(x_2^h, x_3^h, z_2) + \beta Q_\tau[V_2^{Q_\tau}(h, x, z^t) | Z_2 = z_2] | Z_1 = z] \\
&= Q_\tau[Q_\tau[u(x_1^h, x_2^h, z_1) + \beta u(x_2^h, x_3^h, z_2) + \beta^2 V_2^{Q_\tau}(h, x, z^t) | Z_2 = z_2] | Z_1 = z] \\
&= Q_\tau\left[Q_\tau\left[Q_\tau\left[\sum_{t=1}^{3} \beta^{t-1} u(x_t^h, x_{t+1}^h, z_t) + \beta^3 V_3^{Q_\tau}(h, x, z^t) \middle| Z_3 = z_3 \right] \middle| Z_2 = z_2 \right] \middle| Z_1 = z \right] \\
&= Q_\tau\left[\cdots Q_\tau\left[\sum_{t=1}^{n} \beta^{t-1} u(x_t^h, x_{t+1}^h, z_t) + \beta^n V_n^{Q_\tau}(h, x, z^t) \middle| Z_n = z_n \right] \middle| \cdots \middle| Z_1 = z \right],
\end{aligned}
$$

Take quantile each time!

Their Contribution

- Dynamic Consistency

$$v(x) = \sup_{a \in \Gamma(x)} \{ r(x,a) + \beta Q_\tau [v(x')|(x,a)] \}$$

- Advantages of quantile preferences
  - Capture heterogeneity (τ as a parameter)
  - Separate risk aversion and elasticity of intertemporal substitution

de Castro et al. (2025)

$$(Tv)(x,z) = \sup_{a \in \Gamma(x,z)} \{ r(x,z,a) + \beta \, Q_\tau \left[v(f(x,a,z'), z') | z \right] \}$$

where $z' \sim P(z,\cdot)$

If the following conditions hold

1. Z is either connected or finite
2. $\forall z \in \mathsf{Z}$ and $\varepsilon \in (0,1)$, $\exists$ compact $B \subset \mathsf{Z}$ such that $P(z,B) > 1-\varepsilon$
3. $\forall$ compact $B \subset \mathsf{Z}$, the map $z \mapsto P(z,B)$ is continuous
4. $\forall$ nonempty and open $A \subset \mathsf{Z}$ and $z \in \mathsf{Z}$, $P(z,A) > 0$

then Q_τ is a self-map on $bc(\mathsf{X} \times \mathsf{Z})$

de Castro, L. and A. F. Galvao (2019): "Dynamic quantile models of rational behavior," *Econometrica*, 87, 1893–1939.

de Castro, L., A. F. Galvao, and D. Nunes (2025): "Dynamic economics with quantile preferences," *Theoretical Economics*, 20, 353–425.

Li, X., H. Zhong, and M. L. Brandeau (2022): "Quantile Markov decision processes," *Operations research*, 70, 1428–1447.

Xia, L. and J. Pan (2025): "Markov Decision Processes with Value-at-Risk Criterion," .