Reinforcement Learning Methods for Risk Averse Sequential Decision Making

Erick Delage
Department of Decision Sciences
HEC MONTRÉAL

(joint work with Saeed Marzban (HEC Montréal), Jonathan Y. Li (U. of Ottawa), Jia Lin Hau, Marek Petrik (U. of New Hampshire), Mohammad Ghavamzadeh (Amazon), Esther Derman (U. of Montréal), Weikai Wang (HEC Montréal))

17th Conference on Stochastic Programming

Friday, August 1st, 2025

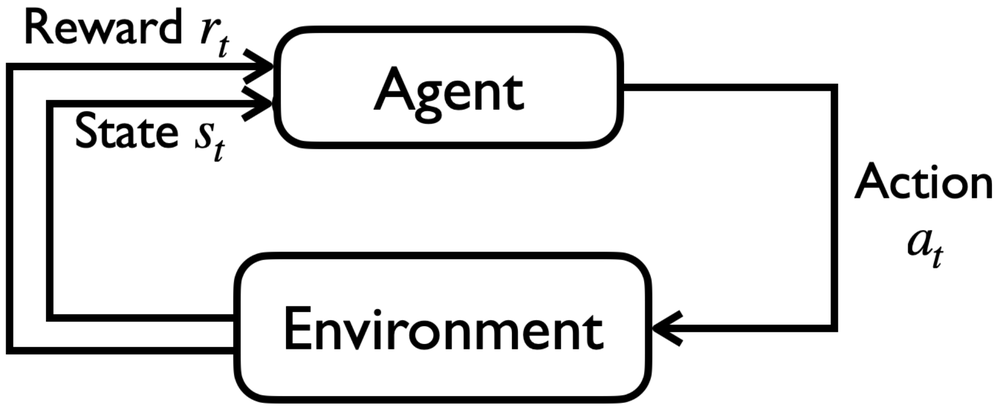
GERAD Mila IVADO Canada Research Chairs / Chaires de recherche du Canada Canada

Sequential decision making using MDPs

- Consider a finite horizon MDP $(\mathcal{S}, \mathcal{A}, r, P, \gamma, s_0)$
- Given a policy $\pi : \mathcal{S} \times [T] \to \mathcal{A}$, we are interested in the risk related to the sum of cumulative discounted reward:

$$\tilde{R}_T(\pi) := \sum_{t=0}^{T-1} \gamma^t r(\tilde{s}_t, \tilde{a}_t)$$

where $\{\tilde{s}_t\}_{t=0}^{T}$ is a trajectory traversed using π_t, i.e. $\tilde{a}_t \sim \pi_t(\tilde{s}_t)$, starting from s_0.



Risk neutral sequential decision making

- Traditional form considers a risk neutral (RN) attitude:

$$\min_{\pi} \quad \mathbb{E}[-\tilde{R}_T]$$

- Different forms of objectives:
  - Finite horizon: $\mathbb{E}[-\tilde{R}_T(\pi)]$
  - Infinite horizon ($T = \infty$): $\lim_{T\to\infty} \mathbb{E}[-\tilde{R}_T(\pi)]$ with $\gamma < 1$
  - Average expected reward: $\lim_{T\to\infty} (1/T)\,\mathbb{E}[-\tilde{R}_T(\pi)]$ with $\gamma = 1$
- Different forms of policy:
  - History dependent: $\pi_t : \mathcal{S}^t \times \mathcal{A}^{t-1} \to \mathcal{A}$
  - Markovian : $\pi_t : \mathcal{S} \to \mathcal{A}$
  - Stationary: $\pi_t = \pi$, for all t

The role of MDPs in stochastic programming

- Consider the following multi-stage stochastic program:

$$\min_{x_0, \{x_t(\cdot)\}_{t=1}^{T-1}} \quad \mathbb{E}[c_0(x_0, z_0) + \sum_{t=1}^{T-1} \gamma^t c_t(x_t(\tilde{z}_{1:t}), \tilde{z}_t)]$$

$$\text{s.t.} \;\; d_{0j}(x_0, z_0) + \sum_{t=1}^{T-1} d_{tj}(x_t(\tilde{z}_{1:t}), \tilde{z}_t) \leq 0, \quad \forall\, j = 1, \ldots, J, \text{ a.s.}$$

with Markov $\tilde{z}$, i.e. $\tilde{z}_{t+1:T-1} \perp \tilde{z}_{1:t-1} \,|\, \tilde{z}_t$ for all t

- An equivalent risk neutral MDP takes the form:
  - $s_t := [z_t^\top \quad \bar{d}_t^\top \quad t]^\top$ where $\bar{d}_{tj} := d_{0j}(x_0, z_0) + \sum_{t'=1}^{t-1} d_{t'j}(x'_t(\tilde{z}_{1:t'}), \tilde{z}'_t)$
  - $a_t := x_t$
  - $r(s, a) := \begin{cases} -\infty & \text{if } t = T \;\&\; \max_j \bar{d}_j > 0 \\ -c_t(x, z) & \text{otherwise} \end{cases}$

The rise of deep reinforcement learning

- 1991: TD-Gammon learns to play backgammon and surpasses some of the best human players (Tesauro [1995]).
- 2015: DeepMind trains an agent that achieves human level performance on Atari games (Mnih et al. [2015]).
- 2016: DeepMing's AlphaGo defeats world champion Lee Sedol in 4 out of 5 games (Silver et al. [2016]).
- 2022: ChatGPT uses DRL to fine-tune its LLM to account for human feedback (Ooyang et al. [2022]).

Q-learning for inf. horizon RN MDPs

- When $T = \infty$, RL methods to solve RN MDPs rely on solution of Bellman equations:

$$Q^*(s,a) = \mathbb{E}\Big[-r(s,a) + \gamma \min_{a'} Q^*(s',a) \Big| s,a \Big], \forall (s,a)$$

which gives $\pi_t^*(s) := \arg\min_{a \in \mathcal{A}} Q^*(s,a)$.

- In tabular setting, Q-learning is a model-free solution scheme, i.e. based on $\{s_k, a_k, s_k'\}_{k=1}^{\infty}$:

$$Q^k(s_k,a_k) \leftarrow Q^{k-1}(s_k,a_k) + \alpha(k) \cdot \left(-r(s_k,a_k) + \gamma \min_{a'} Q^{k-1}(s_k',a') - Q^{k-1}(s_k,a_k) \right)$$

$$Q^k(s,a) \leftarrow Q^{k-1}(s,a), \forall (s,a) \neq (s_k,a_k)$$

It is guaranteed to converge to Q^* if each (s,a) is visited infinitely often and learning rate satisfies Robbins-Monro conditions.

Deep RL for risk neutral MDPs with continuous $\mathcal{S}$ and $\mathcal{A}$

Algorithm Deep Deterministic Policy Gradient (DDPG)

Initialize the main actor θ_π and critic θ_Q networks , the target actor, $\bar{\theta}_\pi$, and critic, $\bar{\theta}_Q$, networks
for $j = 1 : \#Episodes$ **do**
  Initialize a random process $\mathcal{N}$ for action exploration;
  Initialize state to s_0 and effective horizon $\tilde{T}$
  for $t = 0 : \tilde{T} - 1$ **do**
    Select action $a_t = \pi_{\theta_\pi}(s_t) + \mathcal{N}_t$
    Execute a_t and store transition (s_t, a_t, r_t, s'_t)
    Sample a minibatch of N transitions $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$
    Set $y_i := -r_i + \gamma Q_{\bar{\theta}_Q}(s'_i, \pi_{\bar{\theta}_\pi}(s'_i))$
    Update the main critic network:

$$\theta_Q \leftarrow \theta_Q + \alpha \frac{1}{N} \sum_{i=1}^{N} (y_i - Q_{\theta_Q}(s_i, a_i)) \nabla_{\theta_Q} Q_{\theta_Q}(s_i, a_i)$$

    Update the main actor network :

$$\theta_\pi \leftarrow \theta_\pi - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q_{\theta_Q}(s_i, a)\,|_{a=\pi_{\theta_\pi}(s_i)} \nabla_{\theta_\pi} \pi_{\theta_\pi}(s_i)$$

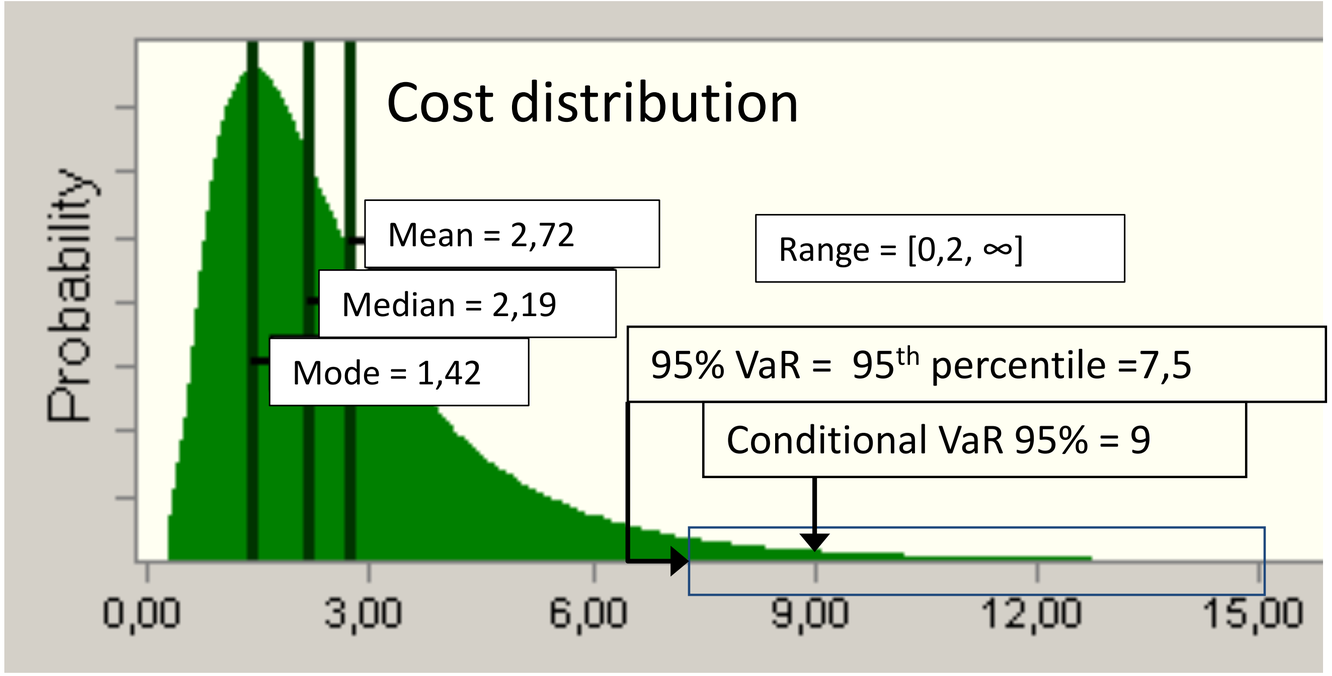
    Update the target networks: $(\bar{\theta}_\pi, \bar{\theta}_Q) \leftarrow (1 - \alpha)(\bar{\theta}_\pi, \bar{\theta}_Q) + \alpha(\theta_\pi, \theta_Q)$
  end for
end for

Moving beyond the RN MDPs

- Two popular approaches for handling risk aversion:
  1. Static law-invariant risk measure (SRM):

$$\min_{\pi} \bar{\rho}(-\tilde{R}(\pi)) := \bar{\varrho}(F_{-\tilde{R}(\pi)})$$

  - E.g. : $\mathbb{E}[-\tilde{R}(\pi)]$, VaR$(-\tilde{R}(\pi))$, CVaR$(-\tilde{R}(\pi))$
  - Pros: Easy to interpret
  - Cons: Can violate dynamic consistency



Moving beyond the RN MDPs

- Two popular approaches for handling risk aversion:
  1. Static law-invariant risk measure (SRM):

$$\min_{\pi} \bar{\rho}(-\tilde{R}(\pi)) := \bar{\varrho}(F_{-\tilde{R}(\pi)})$$

  2. Dynamic law-invariant risk measure (DRM):

$$\min_{\pi} \rho(-\tilde{R}(\pi)) := \bar{\rho}_0(\bar{\rho}_1(\ldots\bar{\rho}_{T-1}(-\tilde{R}(\pi)\,|\,\tilde{a}_{0:T-2}, \tilde{s}_{1:T-1})\cdots\,|\,\tilde{a}_0, \tilde{s}_1))$$

  - E.g.: $\mathbb{E}[-\tilde{R}(\pi)]$, $\text{VaR}(\text{VaR}(\ldots\text{VaR}(-\tilde{R}(\pi)\,|\,\tilde{a}_{0:T-2}, \tilde{s}_{1:T-1})\ldots\,|\,\tilde{a}_0, \tilde{s}_1))$, $\text{CVaR}(\text{CVaR}(\ldots\text{CVaR}(-\tilde{R}(\pi)\,|\,\tilde{a}_{0:T-2}, \tilde{s}_{1:T-1})\ldots\,|\,\tilde{a}_0, \tilde{s}_1))$
  - Pros: Satisfies dynamic consistency, associated to Bellman equation
  - Cons: Can be hard to interpret

Outline

- Introduction

- Q-learning with Dynamic Expectile Risk Measure

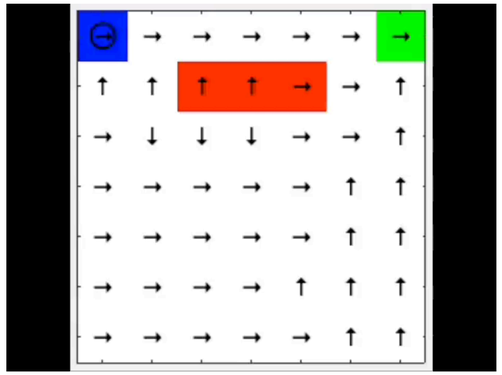
- Q-learning with Static Quantile Measure

- Q-learning for Average Risk-aware MDP

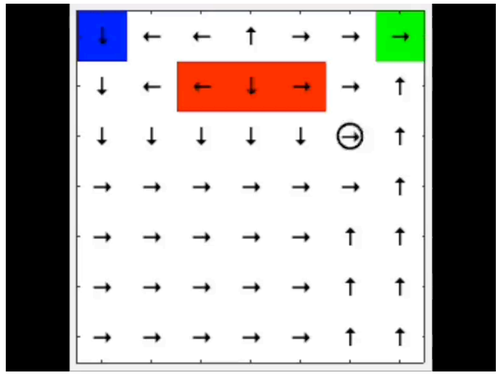
- Conclusion

Q-learning with Dynamic Expectile Risk Measure

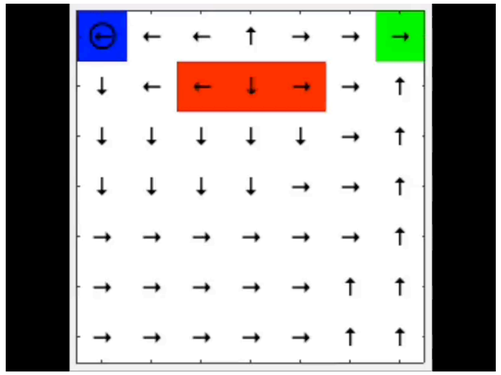
Risk seeking agent

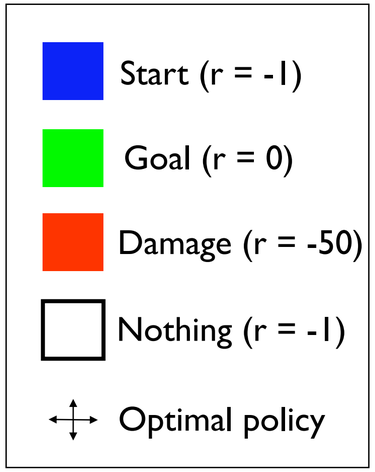


Risk neutral agent



Risk averse agent





Saeed Marzban, D, Jonathan Y. Li, Deep Reinforcement Learning for Equal Risk Pricing and Hedging under Dynamic Expectile Risk Measures, Quantitative Finance, 2023.

Coherent risk measure [Artzner et al. 1999]

- Definition:

  A risk measure is said to be **coherent** if it satisfies the following properties:
  - Monotone: $\forall \tilde{X}, \tilde{Y}$ such that $\tilde{X} \geq \tilde{Y}$ a.s., we have $\rho(\tilde{X}) \geq \rho(\tilde{Y})$
  - Translation invariant: $\forall \tilde{X}$ and t, we have $\rho(\tilde{X} + t) = \rho(\tilde{X}) + t$
  - Positive homogeneous: $\forall \tilde{X}$ and $\alpha \geq 0$, we have $\rho(\alpha\tilde{X}) = \alpha\rho(\tilde{X})$
  - Subadditive: $\forall \tilde{X}, \tilde{Y}$, we have $\rho(\tilde{X} + \tilde{Y}) \leq \rho(\tilde{X}) + \rho(\tilde{Y})$
  - Furthermore, it can be
    - Law-invariant: $\forall \tilde{X}, \tilde{Y}$ such that $\tilde{X} = \tilde{Y}$ in distribution, we have $\rho(\tilde{X}) = \rho(\tilde{Y})$

- Examples:
  - Expected value: $\rho(\tilde{X}) := \mathbb{E}[\tilde{X}]$
  - Conditional Value-at-Risk: $\rho(\tilde{X}) := \mathbb{E}[\tilde{X} \,|\, \tilde{X} \geq F_X^{-1}(\alpha)]$

Elicitable risk measure [Bellini and Bignozzi, 2015]

- Definition:

  A risk measure is said to be **elicitable** if it can be expressed as the unique minimizer of a certain scoring function.

$$\bar{\rho}(\tilde{X}) := \arg\min_{q} \mathbb{E}\left[S(q, \tilde{X})\right] .$$

- We focus on cases where $S(q, x) := \ell(q - x)$:
  - Expected value: $\ell(y) := (1/2)y^2$
  - Quantile: $\ell_\tau(y) := (1-\tau)\max(y,0) + \tau\max(-y,0)$
  - Expectile: $\ell_\tau(y) := (1-\tau)\max(y,0)^2 + \tau\max(-y,0)^2$
- If $\ell'(\,\cdot\,)$ is concave, then $\bar{\rho}(\tilde{X})$ is a *utility-based shortfall risk measure*

Expectile risk measure

- Definition:

  The τ-expectile of a random liability $\tilde{X}$ is defined as:

$$\bar{\rho}(\tilde{X}) := \arg\min_{q} \mathbb{E}\left[(1-\tau)\max(q-\tilde{X},0)^2 + \tau\max(\tilde{X}-q)^2\right]$$

- Examples:
  - $\tau = 0 \Rightarrow \bar{\rho}(\tilde{X}) = \text{ess inf}[\tilde{X}]$, i.e. best-case scenario
  - $\tau = 0.5 \Rightarrow \bar{\rho}(\tilde{X}) = \mathbb{E}[\tilde{X}]$, i.e. risk neutral
  - $\tau = 1 \Rightarrow \bar{\rho}(\tilde{X}) = \text{ess sup}[\tilde{X}]$, i.e. worst-case scenario

- Expectile with $\tau \in [0.5, 1]$ is the class of all elicitable coherent risk measures [Bellini and Bignozzi, 2015]

Dynamic expectile risk measure (DERM)

- Definition:

  A dynamic expectile risk measure takes the form:

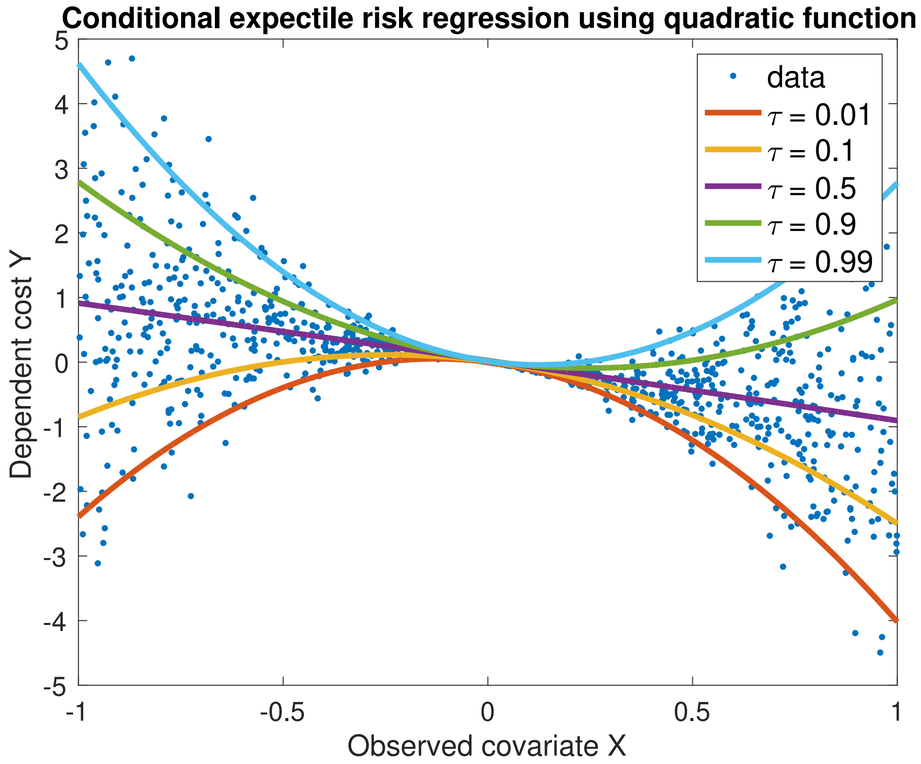
$$\rho(-\tilde{R}(\pi)) := \bar{\rho}_0(\bar{\rho}_1(\ldots\bar{\rho}_{T-1}(-\tilde{R}(\pi)\,|\,\tilde{a}_{0:T-2}, \tilde{s}_{1:T-1})\ldots|\,\tilde{a}_0, \tilde{s}_1))\,,$$

  where each $\bar{\rho}_t(\,\cdot\,|\,\tilde{a}_{0:t-1}, \tilde{s}_{1:t})$ is an expectile risk measure that employs the conditional distribution given $(\tilde{a}_{0:t-1}, \tilde{s}_{1:t})$.

Data-driven conditional risk estimation

- When using elicitable risk measures, conditional risk can be estimated based on i.i.d. data $\{x_i, y_i\}_{i=1}^{M}$ using regression:

$$\theta^* = \arg\min_{\theta} \frac{1}{M} \sum_{i=1}^{M} \ell(h_\theta(x_i) - y_i) \;\Rightarrow\; \bar{\rho}(\tilde{Y} \,|\, \tilde{X}) \approx h_{\theta^*}(\tilde{X})$$



Bellman equations for DERM-MDP

(Ruszczynski [2010], Shen et al. [2013], Pichler and Shapiro [2018], Bäuerle and Glauber [2022])

- With DERMs, one can exploit TI, PH, monotonicity, and mixture quasi-concavity to obtain Bellman equations.
- For example, when $T = 3$:

$$\rho(-\tilde{R}(\pi)) = \bar{\rho}_0\left(\bar{\rho}_1\left(\bar{\rho}_2(-\sum_{t=0}^{2}\gamma^t r(\tilde{s}_t,\tilde{a}_t)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,\middle|\,\tilde{a}_0,\tilde{s}_1\right)\right)$$

$$= \bar{\rho}_0\left(-r(s_0,\tilde{a}_0)+\bar{\rho}_1\left(-\gamma r(\tilde{s}_1,\tilde{a}_1)+\bar{\rho}_2(-\gamma^2 r(\tilde{s}_2,\tilde{a}_2)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,\middle|\,\tilde{a}_0,\tilde{s}_1\right)\right)$$

$$= \bar{\rho}_0\left(-r(s_0,\tilde{a}_0)+\gamma\bar{\rho}_1\left(-r(\tilde{s}_1,\tilde{a}_1)+\gamma\bar{\rho}_2(-r(\tilde{s}_2,\tilde{a}_2)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,\middle|\,\tilde{a}_0,\tilde{s}_1\right)\right)$$

Bellman equations for DERM-MDP

(Ruszczynski [2010], Shen et al. [2013], Pichler and Shapiro [2018], Bäuerle and Glauber [2022])

- With DERMs, one can exploit TI, PH, monotonicity, and mixture quasi-concavity to obtain Bellman equations.
- For example, when $T = 3$:

$$\rho(-\tilde{R}(\pi)) = \bar{\rho}_0\left(\bar{\rho}_1\left(\bar{\rho}_2(-\sum_{t=0}^{2}\gamma^t r(\tilde{s}_t,\tilde{a}_t)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,\middle|\,\tilde{a}_0,\tilde{s}_1\right)\right)$$

$$= \bar{\rho}_0\left(-r(s_0,\tilde{a}_0)+\gamma\bar{\rho}_1\left(-r(\tilde{s}_1,\tilde{a}_1)+\gamma\bar{\rho}_2(-r(\tilde{s}_2,\tilde{a}_2)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,\middle|\,\tilde{a}_0,\tilde{s}_1\right)\right)$$

$$\geq \min_{a_0}\bar{\rho}_0\left(-r(s_0,a_0)+\gamma\min_{a_1}\bar{\rho}_1\left(-r(\tilde{s}_1,a_1)+\gamma\min_{a_2}\bar{\rho}_2(-r(\tilde{s}_2,a_2)\,|\,a_{0:1},\tilde{s}_{1:2})\,\middle|\,a_0,\tilde{s}_1\right)\right)$$

Bellman equations for DERM-MDP

(Ruszczynski [2010], Shen et al. [2013], Pichler and Shapiro [2018], Bäuerle and Glauber [2022])

- With DERMs, one can exploit TI, PH, monotonicity, and mixture quasi-concavity to obtain Bellman equations.
- For example, when $T = 3$:

$$\rho(-\tilde{R}(\pi)) = \bar{\rho}_0\Big(\bar{\rho}_1\Big(\bar{\rho}_2(-\sum_{t=0}^{2}\gamma^t r(\tilde{s}_t,\tilde{a}_t)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,\Big|\,\tilde{a}_0,\tilde{s}_1\Big)\Big)$$

$$\geq \min_{a_0}\bar{\rho}_0\Big(-r(s_0,a_0)+\gamma\min_{a_1}\bar{\rho}_1\Big(-r(\tilde{s}_1,a_1)+\gamma\min_{a_2}\bar{\rho}_2(-r(\tilde{s}_2,a_2)\,|\,a_{0:1},\tilde{s}_{1:2})\,\Big|\,a_0,\tilde{s}_1\Big)\Big)$$

$$= \min_{a_0}\bar{\rho}_0\Big(-r(s_0,a_0)+\gamma\min_{a_1}\bar{\rho}_1\Big(-r(\tilde{s}_1,a_1)+\gamma\min_{a_2}\bar{\rho}_2(-r(\tilde{s}_2,a_2)\,|\,\tilde{s}_2)\,\Big|\,\tilde{s}_1\Big)\Big)$$

Bellman equations for DERM-MDP

(Ruszczynski [2010], Shen et al. [2013], Pichler and Shapiro [2018], Bäuerle and Glauber [2022])

- With DERMs, one can exploit TI, PH, monotonicity, and mixture quasi-concavity to obtain Bellman equations.
- For example, when $T = 3$:

$$\rho(-\tilde{R}(\pi)) = \bar{\rho}_0\Big(\bar{\rho}_1\Big(\bar{\rho}_2(-\sum_{t=0}^{2}\gamma^t r(\tilde{s}_t,\tilde{a}_t)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,\Big|\,\tilde{a}_0,\tilde{s}_1\Big)\Big)$$

$$\geq \min_{a_0}\bar{\rho}_0\Big(-r(s_0,a_0)+\gamma\min_{a_1}\bar{\rho}_1\Big(-r(\tilde{s}_1,a_1)+\gamma\min_{a_2}\bar{\rho}_2(-r(\tilde{s}_2,a_2)\,|\,\tilde{s}_2)\,\Big|\,\tilde{s}_1\Big)\Big)$$

$$= \bar{\rho}_0\Big(-r(s_0,\pi_0^*(s_0))+\gamma\bar{\rho}_1\Big(-r(\tilde{s}_1,\pi_1^*(\tilde{s}_1))+\gamma\bar{\rho}_2\Big(-r(\tilde{s}_2,\pi_2^*(\tilde{s}_2))\,\Big|\,\tilde{s}_2\Big)\,\Big|\,\tilde{s}_1\Big)\Big)$$

Bellman equations for DERM-MDP

(Ruszczynski [2010], Shen et al. [2013], Pichler and Shapiro [2018], Bäuerle and Glauber [2022])

- With DERMs, one can exploit TI, PH, monotonicity, and mixture quasi-concavity to obtain Bellman equations.
- For example, when $T = 3$:

$$\rho(-\tilde{R}(\pi)) = \bar{\rho}_0\Big(\bar{\rho}_1\Big(\bar{\rho}_2(-\sum_{t=0}^{2}\gamma^t r(\tilde{s}_t,\tilde{a}_t)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,\Big|\,\tilde{a}_0,\tilde{s}_1\Big)\Big)$$

$$\geq \bar{\rho}_0\Big(-r(s_0,\pi_0^*(s_0)) + \gamma\bar{\rho}_1\Big(-r(\tilde{s}_1,\pi_1^*(\tilde{s}_1)) + \gamma\bar{\rho}_2\Big(-r(\tilde{s}_2,\pi_2^*(\tilde{s}_2))\,\Big|\,\tilde{s}_2\Big)\,\Big|\,\tilde{s}_1\Big)\Big)$$

$$= \bar{\rho}_0\Big(\bar{\rho}_1\Big(\bar{\rho}_2(-\sum_{t=0}^{2}\gamma^t r(\tilde{s}_t,\pi_t^*(\tilde{s}_t))\,|\,\tilde{s}_{1:2})\,\Big|\,\tilde{s}_1\Big)\Big) = \rho(-\tilde{R}(\pi^*)),$$

where

$$\pi_2^*(s) \in \arg\min_a Q_2^*(s,a) := \bar{\rho}_2(-r(s,a)\,|\,\tilde{s}_2 = s) = -r(s,a)$$

$$\pi_1^*(s) \in \arg\min_a Q_1^*(s,a) := \bar{\rho}_1(-r(s,a) + \gamma\bar{\rho}_2(-r(\tilde{s}_2,\pi_2^*(\tilde{s}_2))\,|\,\tilde{s}_2)\,|\,\tilde{s}_1 = s)$$

$$= \bar{\rho}_1(-r(s,a) + \gamma\min_{a'} Q_2^*(\tilde{s}_2,a')\,|\,\tilde{s}_1 = s)$$

$$\pi_0^*(s) \in \arg\min_a Q_0^*(s,a) := \bar{\rho}_0(-r(s,a) + \gamma\min_{a'} Q_1^*(\tilde{s}_1,a'))$$

Bellman equations for DERM-MDP

(Ruszczynski [2010], Shen et al. [2013], Pichler and Shapiro [2018], Bäuerle and Glauber [2022])

Theorem:

For general T,

$$\min_{\pi} \rho(-\tilde{R}(\pi)) = \rho(-\tilde{R}(\pi^*)) = \min_{a_0} Q_0^*(s_0, a_0)$$

where

$$Q_t^*(s,a) := \bar{\rho}_t\Big(-r(s,a) + \gamma \min_{a'} Q_{t+1}^*(\tilde{s}_{t+1}, a') \,\Big|\, \tilde{s}_t = s \Big)$$

and $Q_T^*(s,a) := 0$ while $\pi_t^*(s) \in \arg\min_a Q_t^*(s,a)$.

Converting Bellman equations to Q-learning

- Exploiting the elicitability property, we get

$$Q_t^*(s,a) = \bar{\rho}_t\Big(-r(s,a) + \gamma \min_{a_{t+1}} Q_{t+1}^*(\tilde{s}_{t+1}, a_{t+1}) \,\Big|\, \tilde{s}_t = s \Big)$$
$$= \arg\min_q \mathbb{E}\Big[\ell\Big(q - (-r(s,a) + \gamma \min_{a_{t+1}} Q_{t+1}(\tilde{s}_{t+1}, a_{t+1})) \Big) \,\Big|\, \tilde{s}_t = s \Big]$$

- This gives rise to a stochastic gradient algorithm that learns from sample $s' \sim P(\,\cdot\,|\,s,a)$:

$$Q_t(s,a) \leftarrow Q_t(s,a) - \alpha \cdot \ell'\Big(Q_t(s,a) - (-r(s,a) + \gamma \min_{a'} Q_{t+1}^*(s',a')) \Big)$$

- This generalizes the Q-learning update for RN case, where $\ell(y) := (1/2)y^2$ and $\ell'(y) = y$:

$$Q_t(s,a) \leftarrow Q_t(s,a) - \alpha \cdot \Big(Q_t(s,a) - (-r(s,a) - \gamma \min_{a'} Q_{t+1}(s',a')) \Big)$$

Convergence of risk-sensitive Q-learning

([Shen et al. [2014], Hau et al. [2025])

Theorem (finite horizon):

In tabular setting, let $\tau \in (0,1)$. Assume that $\alpha(k)$ and $\{(t_k, s_k, a_k, s'_k)\}_{k=0}^{\infty}$ used in

$$Q_{t_k}^k(s_k, a_k) \leftarrow Q_{t_k}^{k-1}(s_k, a_k) - \alpha(k) \cdot \ell' \left(Q_{t_k}^{k-1}(s_k, a_k) + r(s_k, a_k) - \gamma \min_{a'} Q_{t_k+1}^{k-1}(s'_k, a') \right)$$
$$Q_t^k(s, a) \leftarrow Q_t^{k-1}(s, a), \quad \forall (t, s, a) \neq (t_k, s_k, a_k)$$

satisfy the Robbins-Monro conditions:

$$\sum_{k:(t_k,s_k,a_k)=(t,s,a)} \alpha(k) = \infty, \qquad \sum_{k:(t_k,s_k,a_k)=(t,s,a)} \alpha(k)^2 < \infty, \ \forall (t, s, a) \quad \text{a.s.}$$

then, the sequence $\{Q^k\}_{k=0}^{\infty}$ converges almost surely to Q^*.

Convergence of risk-sensitive Q-learning
([Shen et al. [2014], Hau et al. [2025])

Theorem (infinite horizon):

In tabular setting, let $\tau \in (0,1)$. Assume that $\alpha(k)$ and $\{(s_k, a_k, s'_k)\}_{k=0}^{\infty}$ used in

$$Q^k(s_k, a_k) \leftarrow Q^{k-1}(s_k, a_k) - \alpha(k) \cdot \ell' \left(Q^{k-1}(s_k, a_k) + r(s_k, a_k) - \gamma \min_{a'} Q^{k-1}(s'_k, a') \right)$$
$$Q^k(s, a) \leftarrow Q^{k-1}(s, a), \quad \forall (s, a) \neq (s_k, a_k)$$

satisfy the Robbins-Monro conditions:

$$\sum_{k:(s_k,a_k)=(s,a)} \alpha(k) = \infty, \quad \sum_{k:(s_k,a_k)=(s,a)} \alpha(k)^2 < \infty, \forall (s, a) \quad \text{a.s.}$$

then, the sequence $\{Q^k\}_{k=0}^{\infty}$ converges almost surely to Q^*.

Deep risk averse RL using DERMs

- In Marzban et al. [2023], we extend the deep deterministic policy gradient (DDPG) algorithm to solve dynamic problems formulated based on dynamic expectile risk measures:

$$Q^*(s,a) = \bar{\rho}\Big(-r(s,a) + \gamma \min_{a'} Q^*(s',a') \,\Big|\, s \Big)$$

Algorithm Traditional RN DDPG

Initialize the main actor θ_π and critic θ_Q networks
Initialize the target actor, $\bar{\theta}_\pi$, and critic, $\bar{\theta}_Q$, networks
for $j = 1 : \#Episodes$ **do**
  Initialize a random process $\mathcal{N}$ for action exploration;
  Receive initial observation state s_0 and horizon $\tilde{T}$
  for $t = 0 : \tilde{T} - 1$ **do**
    Select action $a_t = \pi_{\theta_\pi}(s_t) + \mathcal{N}_t$
    Execute a_t and store transition (s_t, a_t, r_t, s'_t)
    Sample a minibatch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$
    Set $y_i := -r_i + Q_{\bar{\theta}_Q}(s'_i, \pi_{\bar{\theta}_\pi}(s'_i))$
    Update the main critic network:

$$\theta_Q \leftarrow \theta_Q + \alpha \frac{1}{N} \sum_{i=1}^{N} (y_i - Q_{\theta_Q}(s_i, a_i)) \nabla_{\theta_Q} Q_{\theta_Q}(s_i, a_i)$$

    Update the main actor network :

$$\theta_\pi \leftarrow \theta_\pi - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q_{\theta_Q}(s_i, a)\big|_{a=\pi_{\theta_\pi}(s_i)} \nabla_{\theta_\pi} \pi_{\theta_\pi}(s_i)$$

    Update the target networks $(\bar{\theta}_Q, \bar{\theta}_\pi)$
  end for
end for

Deep risk averse RL using DERMs

- In Marzban et al. [2023], we extend the deep deterministic policy gradient (DDPG) algorithm to solve dynamic problems formulated based on dynamic expectile risk measures:

$$Q^*(s,a) = \bar{\rho}\Big(-r(s,a) + \gamma \min_{a'} Q^*(s',a') \,\Big|\, s \Big)$$

Algorithm Traditional RN DDPG

Initialize the main actor θ_π and critic θ_Q networks
Initialize the target actor, $\bar{\theta}_\pi$, and critic, $\bar{\theta}_Q$, networks
for $j = 1 : \#Episodes$ **do**
  Initialize a random process $\mathcal{N}$ for action exploration;
  Receive initial observation state s_0 and horizon $\tilde{T}$
  for $t = 0 : \tilde{T} - 1$ **do**
    Select action $a_t = \pi_{\theta_\pi}(s_t) + \mathcal{N}_t$
    Execute a_t and store transition (s_t, a_t, r_t, s'_t)
    Sample a minibatch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$
    Set $y_i := -r_i + Q_{\bar{\theta}_Q}(s'_i, \pi_{\bar{\theta}_\pi}(s'_i))$
    Update the main critic network:

$$\theta_Q \leftarrow \theta_Q + \alpha \frac{1}{N} \sum_{i=1}^{N} \ell'(Q_{\theta_Q}(s_i,a_i) - y_i) \nabla_{\theta_Q} Q_{\theta_Q}(s_i,a_i)$$

    Update the main actor network :

$$\theta_\pi \leftarrow \theta_\pi - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q_{\theta_Q}(s_i,a)\big|_{a=\pi_{\theta_\pi}(s_i)} \nabla_{\theta_\pi} \pi_{\theta_\pi}(s_i)$$

    Update the target networks $(\bar{\theta}_Q, \bar{\theta}_\pi)$
  end for
end for

Deep risk averse RL using DERMs

- In Marzban et al. [2023], we extend the deep deterministic policy gradient (DDPG) algorithm to solve dynamic problems formulated based on dynamic expectile risk measures:

$$Q^*(s,a) = \bar{\rho}\Big(-r(s,a) + \gamma \min_{a'} Q^*(s',a') \,\Big|\, s \Big)$$

Algorithm Risk averse DDPG

Initialize the main actor θ_π and critic θ_Q networks
Initialize the target actor, $\bar{\theta}_\pi$, and critic, $\bar{\theta}_Q$, networks
for $j = 1 : \#Episodes$ **do**
  Initialize a random process $\mathcal{N}$ for action exploration;
  Receive initial observation state s_0 and horizon $\tilde{T}$
  for $t = 0 : \tilde{T} - 1$ **do**
    Select action $a_t = \pi_{\theta_\pi}(s_t) + \mathcal{N}_t$
    Execute a_t and store transition (s_t, a_t, r_t, s'_t)
    Sample a minibatch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$
    Set $y_i := -r_i + Q_{\bar{\theta}_Q}(s'_i, \pi_{\bar{\theta}_\pi}(s'_i))$
    Update the main critic network:

$$\theta_Q \leftarrow \theta_Q + \alpha \frac{1}{N} \sum_{i=1}^{N} \ell'(Q_{\theta_Q}(s_i,a_i) - y_i) \nabla_{\theta_Q} Q_{\theta_Q}(s_i,a_i)$$

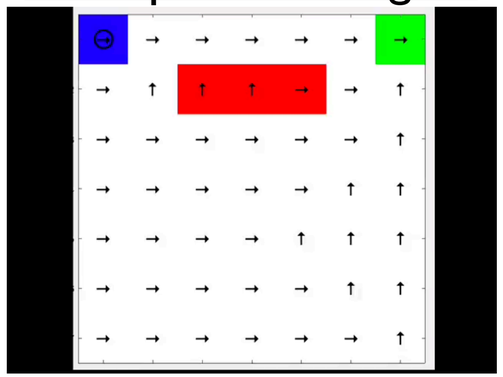
    where ~~$\ell(\Delta) := (1/2)\Delta^2$~~
$\ell(\Delta) := (1-\tau)\max(0,\Delta)^2 + \tau \max(0,-\Delta)^2$
    Update the main actor network :

$$\theta_\pi \leftarrow \theta_\pi - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q_{\theta_Q}(s_i,a)\big|_{a=\pi_{\theta_\pi}(s_i)} \nabla_{\theta_\pi} \pi_{\theta_\pi}(s_i)$$

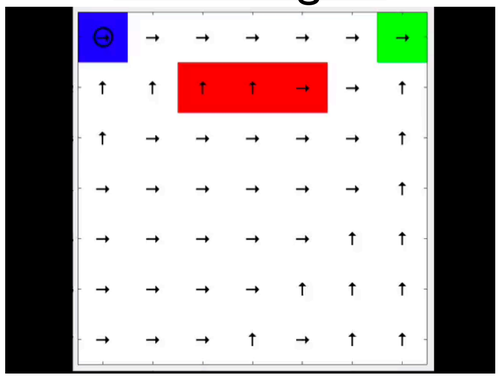
    Update the target networks $(\bar{\theta}_Q, \bar{\theta}_\pi)$
  end for
end for

Q-learning with Static Quantile Measure

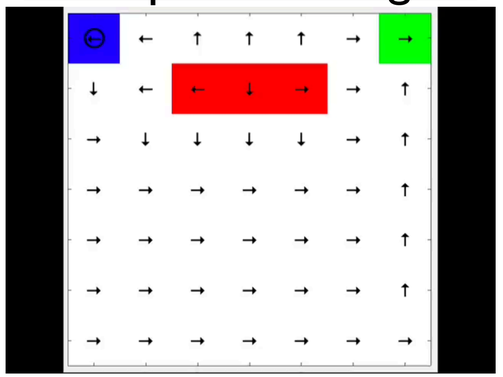
10-th percentile agent



Median agent



90-th percentile agent



Jia Lin Hau, D, Esther Derman, Mohammad Ghavamzadeh, Marek Petrik, Q-learning for Quantile MDPs: A Decomposition, Performance, and Convergence Analysis, AISTATS 2025.

Forms of Quantile MDPs

- Epistemic uncertainty: Considers that there is uncertainty about the MDP model $(\tilde{r}, \tilde{P})$, and policy must optimize:

$$\min_{\pi} \text{Quant.}_{\tau} \left(\mathbb{E}[\tilde{R}_T(\pi) \mid \tilde{r}, \tilde{P}] \right)$$

  - E.g.: D and Mannor [2010], Russel and Petrik [2019], Lobo et al. [2023]

- Aleatoric uncertainty: Considers that the model is determined but policy should control the distribution of total reward

$$\min_{\pi} \text{Quant.}_{\tau}(-\tilde{R}_T(\pi))$$

  - E.g. Filar et al. [1995], Gilbert et al. [2016], Li et al [2022b]

A decomposition for quantile risk

- We focus on value-at-risk:

$$\mathrm{VaR}_\tau(\tilde{X}) := \mathfrak{q}^-(\tilde{X}) = \min\{z \,|\, \mathbb{P}(\tilde{X} \le z) \ge \tau\}$$

- Li et al. [2022b]'s decomposition:

$$\mathrm{VaR}_\tau(\tilde{X}) = \inf_{\xi:\mathscr{Y}\to[0,1]} \left\{ \operatorname{ess\,sup} \left[\mathrm{VaR}_{\xi(\tilde{Y})}(\tilde{X} \,|\, \tilde{Y}) \right] \middle| \, \mathbb{E}[\xi(\tilde{Y})] = \tau \right\}$$

- Our's:

$$\mathrm{VaR}_\tau(\tilde{X}) = \mathrm{VaR}_\tau(\mathrm{VaR}_{\tilde{u}}(\tilde{X} \,|\, \tilde{Y})) \text{ with } \tilde{u} \sim U([0,1])$$

- Sketch of proof:
  $\mathbb{P}(\tilde{X} \le z)$

Hence, $\tilde{X} = \mathrm{VaR}_{\tilde{u}}(\tilde{X} \,|\, Y)$ in distribution.

Bellman equations for Quantile MDP

- Similarly as before, when $T = 3$:

$$\mathsf{VaR}_{\tau_0}(-\tilde{R}(\pi)) = \mathsf{VaR}_{\tau_0}(\mathsf{VaR}_{\tilde{u}_1}(\mathsf{VaR}_{\tilde{u}_2}(-\sum_{t=0}^{2}\gamma^t r(\tilde{s}_t,\tilde{a}_t)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,|\,\tilde{a}_0,\tilde{s}_1))$$

$$= \mathsf{VaR}_{\tau_0}(-r(s_0,\tilde{a}_0)+\gamma\mathsf{VaR}_{\tilde{u}_1}(-r(\tilde{s}_1,\tilde{a}_1)+\gamma\mathsf{VaR}_{\tilde{u}_2}(-r(\tilde{s}_2,\tilde{a}_2)\,|\,\tilde{a}_{0:1},\tilde{s}_{1:2})\,|\,\tilde{a}_0,\tilde{s}_1))$$

$$\geq \min_{a_0}\mathsf{VaR}_{\tau_0}(-r(s_0,a_0)+\gamma\min_{a_1}\mathsf{VaR}_{\tilde{u}_1}(-r(\tilde{s}_1,a_1)+\gamma\min_{a_2}\mathsf{VaR}_{\tilde{u}_2}(-r(\tilde{s}_2,a_2)\,|\,a_{0:1},\tilde{s}_{1:2})\,|\,\tilde{a}_0,\tilde{s}_1))$$

$$= \min_{a_0}\mathsf{VaR}_{\tau_0}(-r(s_0,a_0)+\gamma\min_{a_1}\mathsf{VaR}_{\tilde{u}_1}(-r(\tilde{s}_1,a_1)+\gamma\min_{a_2}\mathsf{VaR}_{\tilde{u}_1}(-r(\tilde{s}_2,a_2)\,|\,\tilde{s}_2)\,|\,\tilde{s}_1))$$

$$= \mathsf{VaR}_{\tau_0}(-r(s_0,\pi_0^*(s_0))+\gamma\mathsf{VaR}_{\tilde{u}_1}(-r(\tilde{s}_1,\pi_1^*(\tilde{s}_1,\tilde{u}_1))+\gamma\mathsf{VaR}_{\tilde{u}_2}(-r(\tilde{s}_2,\pi_2^*(\tilde{s}_2,\tilde{u}_2))\,|\,\tilde{s}_2)\,|\,\tilde{s}_1))$$

$$= \mathsf{VaR}_{\tau_0}(-r(s_0,\bar{\pi}_0^*(s_0))+\gamma\mathsf{VaR}_{\tilde{u}_1}(-r(\tilde{s}_1,\bar{\pi}_1^*(\tilde{s}_1))+\gamma\mathsf{VaR}_{\tilde{u}_2}(-r(\tilde{s}_2,\bar{\pi}_2^*(\tilde{s}_{1:2}))\,|\,\tilde{s}_{1:2})\,|\,\tilde{s}_1))$$

$$= \mathsf{VaR}_{\tau_0}(\mathsf{VaR}_{\tilde{u}_1}(\mathsf{VaR}_{\tilde{u}_2}(-\sum_{t=0}^{2}\gamma^t r(\tilde{s}_t,\bar{\pi}_t^*(\tilde{s}_{1:t}))\,|\,\tilde{s}_{1:2})\,|\,\tilde{s}_1)) = \mathsf{VaR}_{\tau_0}(-\tilde{R}(\bar{\pi}^*)),$$

where

$$\pi_t^*(s,\tau) \in \arg\min_a Q_t^*(s,\tau,a) := \mathsf{VaR}_\tau(-r(s,a)+\gamma\min_a Q_{t+1}^*(s,\tilde{u})\,|\,\tilde{s}_t = s)$$

$$\bar{\pi}_t^*(s_{1:t}) := \pi_t(s_t,\tau_t),\ \text{with}\ \tau_t := \sup\{\tau : \min_a Q_0^*(s_0,\tau_0,a) + \sum_{t'=0}^{t-1}\gamma^{t'} r(s_{t'},\pi_{t'}(s_{1:t'})) \geq \min_a Q_t^*(s_t,\tau_t,a)\}$$

Bellman equations for Quantile MDP

Theorem:

For general T,

$$\min_{\pi} \mathsf{VaR}_{\tau_0}(-\tilde{R}(\pi)) = \mathsf{VaR}_{\tau_0}(-\tilde{R}(\bar{\pi}^*)) = \min_{a_0} Q_0^*(s_0, \tau_0, a_0)$$

where

$$Q_t^*(s, \tau, a) := \mathsf{VaR}_{\tau}\Big(-r(s,a) + \gamma \min_{a'} Q_{t+1}^*(\tilde{s}_{t+1}, \tilde{u}, a') \,\Big|\, \tilde{s}_t = s \Big),$$

and $Q_T^*(s, \tau, a) := 0$, while

$$\bar{\pi}_t^*(s_{1:t}) := \arg\min_{a} Q_t^*(s, f(s_{1:t}), a)$$

with

$$f(s_{1:t}) := \sup \left\{ \tau : \min_{a} Q_0^*(s_0, \tau_0, a) + \sum_{t'=0}^{t-1} \gamma^{t'} r(s_{t'}, \pi_{t'}(s_{1:t'})) \geq \min_{a} Q_t^*(s_t, \tau_t, a) \right\}$$

Converting Bellman equations to Q-learning

- Exploiting the elicitability property of quantiles, we get

$$Q_t^*(s,\tau,a) = \mathsf{VaR}_\tau\Big(-r(s,a) + \gamma \min_{a_{t+1}} Q_{t+1}^*(\tilde{s}_{t+1},\tilde{u}_{t+1},a_{t+1}) \,\Big|\, \tilde{s}_t = s \Big)$$
$$= \arg\min_q \mathbb{E}\Big[\ell_\tau\Big(q - (-r(s,a) + \gamma \min_{a_{t+1}} Q_{t+1}(\tilde{s}_{t+1},\tilde{u}_{t+1},a_{t+1})) \Big) \,\Big|\, \tilde{s}_t = s \Big]$$

- This gives rise to a stochastic gradient algorithm that learns from sample $s' \sim P(\,\cdot\,|\,s_k,a_k)$ and $\tau' \sim U([0,1])$:

$$Q_t(s_k,\tau_k,a_k) \leftarrow Q_t(s_k,\tau_k,a_k) - \alpha(k)\ell'_{\tau_k}\Big(Q_t(s_k,\tau_k,a_k) - (-r(s_k,a_k) + \gamma \min_{a'} Q_{t+1}(s',\tau',a')) \Big)$$

with $\ell'_\tau(y) = (1-\tau)1\{y \geq 0\} + \tau 1\{y < 0\}$ as a subgradient

Convergence of risk-sensitive Q-learning

Theorem:

In tabular setting, let finite set $\mathscr{T} \subset (0,1)$. Assume that $\alpha(k)$ and $\{(t_k, s_k, \tau_k, a_k, s'_k, \tau'_k)\}_{k=0}^{\infty}$, with $\tau_k \in \mathscr{T}$ and $\tau'_k \sim U(\mathscr{T})$, used in

$$Q_{t_k}^k(s_k, \tau_k, a_k) \leftarrow Q_{t_k}^{k-1}(s_k, \tau_k, a_k) - \alpha(k) \cdot \hat{\ell}'_{\tau_k}\left(Q_{t_k}^{k-1}(s_k, \tau_k, a_k) + r(s_k, a_k) - \gamma \min_{a'} Q_{t_k+1}^{k-1}(s'_k, \tau'_k, a')\right)$$

$$Q_t^k(s, \tau, a) \leftarrow Q_t^{k-1}(s, \tau, a), \quad \forall (t, s, \tau, a) \neq (t_k, s_k \,.\, \tau_k, a_k)$$

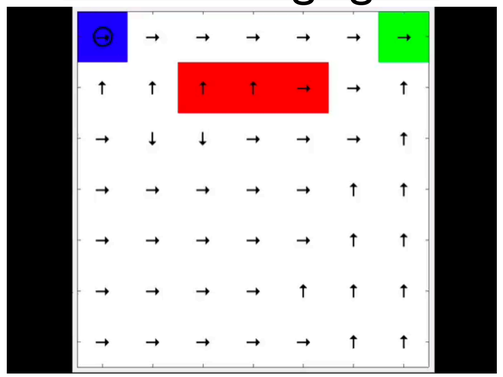
satisfy the Robbins-Monro conditions:

$$\sum_{k:(t_k,s_k,a_k)=(t,s,a)} \alpha(k) = \infty, \qquad \sum_{k:(t_k,s_k,a_k)=(t,s,a)} \alpha(k)^2 < \infty, \ \forall (t, s, a) \quad \text{a.s.}$$

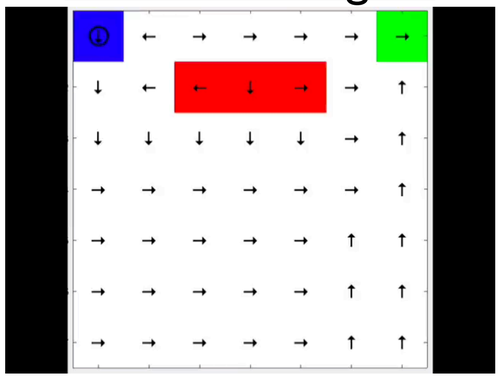
then $Q^k \to Q^\infty \approx Q^*$.

Q-learning for Average Risk-aware MDP

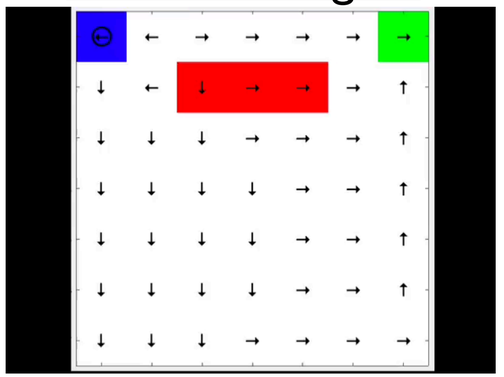
Risk seeking agent



Risk neutral agent



Risk averse agent



Weikai Wang, D, Planning and Learning in Average Risk-aware MDPs, working draft.

Average Cost MDP problem

- Consider the infinite horizon average cost problem:

$$\max_{\pi} \lim_{T \to \infty} (1/T)\mathbb{E}[\tilde{R}_T(\pi)]$$

- Such models are useful in continuing tasks:
  - Supply chain management (Pontrandolfo et al. [2002])
  - Queueing control (van Leeuwen and Nunez-Queija [2017])
  - Ambulance dispatching (Jagtenberg et al. [2017])
  - Traffic control (Haijema et al. [2017])
  - Lot scheduling (van Foreest and Wijngaard [2017])
  - Etc.

Q-learning for RN Average Cost MDPs

- For MDP that is **unichain** under all π, any Q^* and g^* satisfying :

$$Q(s,a) = \bar{\rho}\mathbb{E}[-r(s,a) + \min_{a'} Q(s',a')] - g$$????

gives $\pi^*(s) := \arg\min_a Q^*(s,a)$ achieving optimal value g^*.

- RN Relative Q-value Iteration (Abounadi et al. [2001]):

$$Q^{k+1}(s,a) = \bar{\rho}\mathbb{E}[-r(s,a) + \min_{a'} Q^k(s',a')] - f(Q^k)$$????

with for example $f(q) := \max_{s,a} q(s,a)$, converges to optimal $(Q^*, f(Q^*))$.

- RN Q-learning based on $\{s_k, a_k, s'_k\}_{k=1}^{\infty}$:

$$Q^k(s_k,a_k) \leftarrow Q^{k-1}(s_k,a_k) + \alpha(k)\,\ell'\left(-r(s_k,a_k) + \min_{a'} Q^{k-1}(s'_k,a') - f(Q^k) - Q^{k-1}(s_k,a_k)\right)$$????

$$Q^k(s,a) \leftarrow Q^{k-1}(s,a), \quad \forall (s,a) \neq (s_k,a_k)$$

also converges to optimal $(Q^*, f(Q^*))$

Average Risk MDPs

- Consider the risk averse problem:

$$\min_{\pi} \lim_{T\to\infty} (1/T)\, \mathbb{E}[-\tilde{R}_T(\pi)]$$

$$\bar{\rho}_0\Big(\bar{\rho}_1\Big(\ldots\bar{\rho}_{T-1}(-\tilde{R}_T(\pi)\,|\,\tilde{a}_{0:T-2},\tilde{s}_{1:T-1})\ldots\,\Big|\,\tilde{a}_0,\tilde{s}_1\Big)\Big)$$

- With "proper" MDP and $\bar{\rho}$, any Q^* and g^* satisfying (Shen et al. [2013]):

$$Q(s,a) = \bar{\rho}(-r(s,a) + \min_{a'} Q(s',a')) - g$$

gives $\pi^*(s) := \arg\min_a Q^*(s,a)$ achieving optimal value g^*.

- Risk averse Relative Q-value Iteration (Wang and D [2025]):

$$Q^{k+1}(s,a) = \bar{\rho}(-r(s,a) + \min_{a'} Q^k(s',a') - f(Q^k))$$

converges to optimal $(Q^*, f(Q^*))$.

Q-learning for Average Risk MDPs (I)

- If $\bar{\rho}$ is elicitable, based on risk averse Relative Q-value Iteration:

$$\begin{aligned} Q^{k+1}(s,a) &= \bar{\rho}(-r(s,a) + \min_{a'} Q^k(s',a') - f(Q^k)) \\ &= \arg\min_q \mathbb{E}[\ell(q - (-r(s,a) + \min_{a'} Q^k(s',a') - f(Q^k)))] \end{aligned}$$

- This gives rise to the stochastic gradient algorithm (UBSR Q-learning):

$$Q^k(s_k,a_k) \leftarrow Q^{k-1}(s_k,a_k) - \alpha(k) \cdot \ell' \left(Q^{k-1}(s_k,a_k) - (-r(s_k,a_k) + \min_{a'} Q^{k-1}(s',a') - f(Q^k)) \right)$$

$$Q^k(s,a) \leftarrow Q^{k-1}(s,a),\ \forall (s,a) \neq (s_k,a_k)$$

based on a **single** sample $s' \sim P(\cdot \,|\, s,a)$.

- In risk neutral setting, i.e. $\ell(y) := (1/2)y^2$, reduces to Q-learning proposed by Abounadi et al. [2001]
- Converges empirically but unfortunately no theoretical guarantees yet

Q-learning for Average Risk MDPs (II)

- If one has access to a **simulator**, the risk averse Relative Q-value Iteration:

$$Q^*(s,a) = \bar{\rho}(-r(s,a) + \min_{a'} Q^*(s',a') - f(Q^*))$$

can motivate a different Q-learning algorithm using Robbins-Munro algorithm:

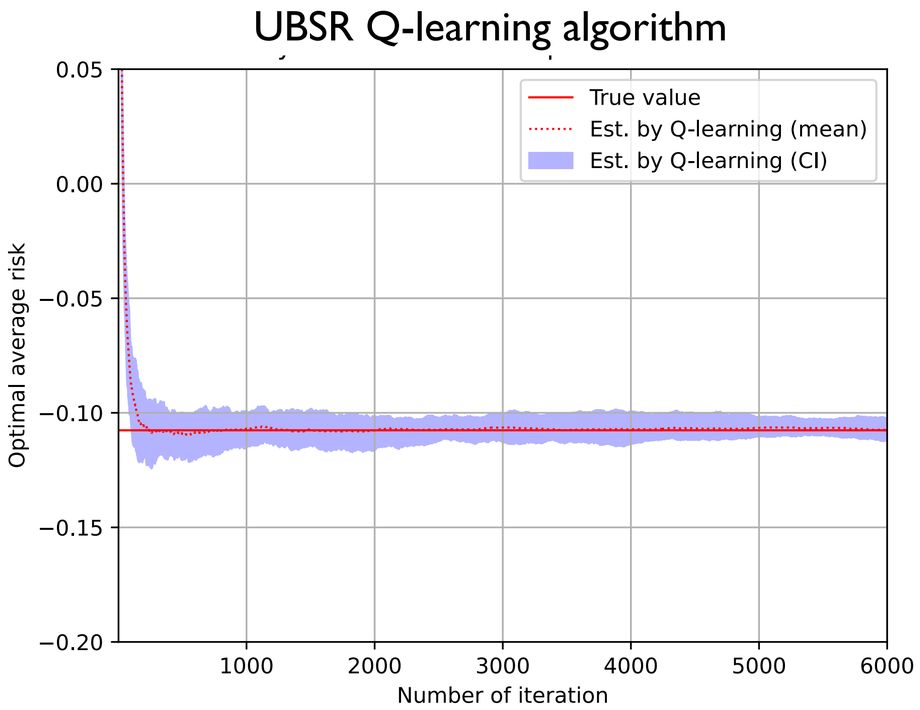
$$Q^k(s_k,a_k) \leftarrow Q^{k-1}(s_k,a_k) + \alpha(k) \cdot \left(\hat{\rho}_N(-r(s_k,a_k) + \min_{a'} Q^{k-1}(s',a') - f(Q^k) - Q^{k-1}(s_k,a_k)) \right)$$

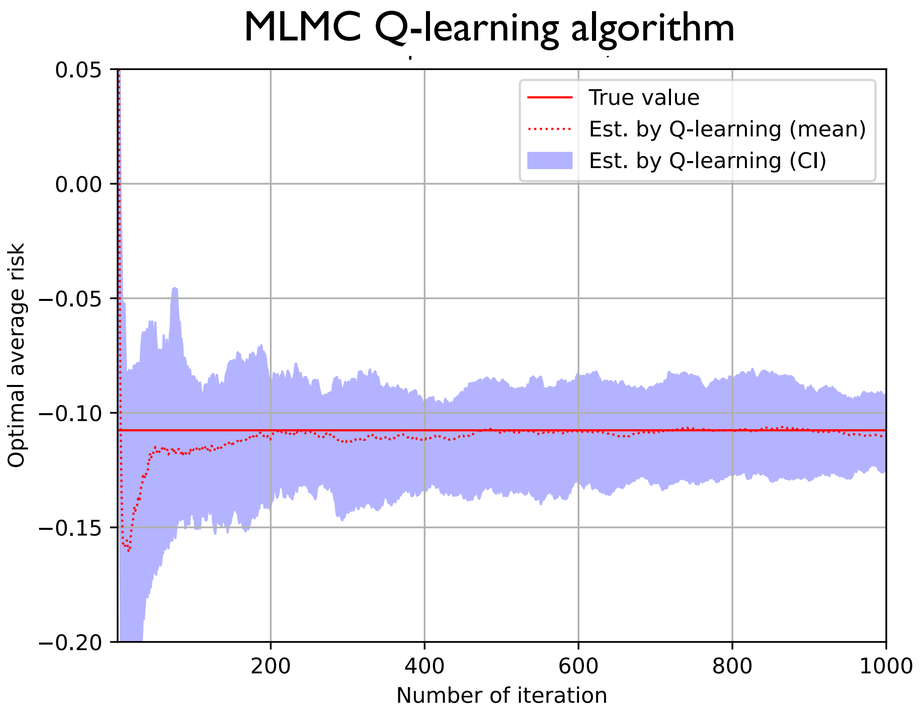
$$Q^k(s,a) \leftarrow Q^{k-1}(s,a),\ \forall (s,a) \neq (s_k,a_k)$$

where $\hat{\rho}_N(X)$ is an unbiased sample-based estimator of $\bar{\rho}(X)$

- We prove convergence to optimal $(Q^*, f(Q^*))$ under the conditions:
  - MDP has a reset state: $P(\bar{s} \,|\, s,a) > 0,\ \forall (s,a)$
  - $\bar{\rho}$ is ε-strictly monotone: $\forall X \geq Y,\ \bar{\rho}(X) \geq \bar{\rho}(Y) + \varepsilon \mathbb{E}[X - Y]$
  - $\hat{\rho}_N$ is unbiased and has bounded variance (e.g. MLMC)
  - Robbins-Monro step size

Comparison of MLMC and UBSR Q-learning





Take-away messages

- Elicitability allows one to adapt model-free reinforcement learning methods to risk aware setting.
- Different types of risk measures can be used:
  - Dynamic risk measures
  - Static risk measures
- Different types of problems:
  - Finite, infinite discounted, infinite average risk
- By developing Deep Reinforcement Learning algorithms that are based on these Q-learning results, one can potentially identify risk aware policies in real world large-scale sequential decision making problems.
- Many potential applications !

References

- Jinane Abounadi, Dimitri P. Bertsekas, and Vivek S. Borkar. Learning algorithms for Markov decision processes with average cost. SIAM Journal on Control and Optimization, 40(3):681–698, 2001.
- Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. Mathematical finance, 9(3):203–228, 1999.
- Fabio Bellini and Valeria Bignozzi. On elicitable risk measures. Quantitative Finance, 15(5):725–733, 2015.
- Nicole Bäuerle and Alexander Glauner. Markov Decision Processes with Recursive Risk Measures. European Journal of Operational Research, 296(3):953–966, 2022.
- Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. ICML, pages 1096–1105, 2018.
- Erick Delage and Shie Mannor. Percentile Optimization for Markov Decision Processes with Parameter Uncertainty. Operations Research, 58(1):203–213, 2010. ISSN 0030-364X, 1526-5463.
- Jerzy A. Filar, Dmitry Krass, and Keith W. Ross. Percentile Performance Criteria For Limiting Average Markov Decision Processes. IEEE Transactions on Automatic Control, 40(1):2–10, 1995.
- Hugo Gilbert, Paul Weng, and Yan Xu. Optimizing Quantiles in Preference-based Markov Decision Processes, arXiv:1612.00094, 2016.
- Rene Haijema, Eligius M. T. Hendrix, and Jan van der Wal. Dynamic control of traffic lights. In Markov Decision Processes in Practice, pages 371–386. Springer, 2017.
- Jia Lin Hau, Erick Delage, Esther Derman, Mohammad Ghavamzadeh, and Marek Petrik. Q-learning for Quantile MDPs: A Decomposition, Performance, and Convergence Analysis. AISTATS, 2025.
- Caroline J. Jagtenberg, Sandjai Bhulai, and Robert D. van der Mei. Optimal Ambulance Dispatching. In Richard J. Boucherie and Nico M. van Dijk, editors, Markov Decision Processes in Practice, pages 269–291. Springer International Publishing, 2017.
- Xiaocheng Li, Huaiyang Zhong, and Margaret L. Brandeau. Quantile Markov Decision Processes. Operations Research, 70(3):1428–1447, 2022.
- Elita A. Lobo, Cyrus Cousins, Yair Zick, and Marek Petrik. Percentile criterion optimization in offline reinforcement learning. NeurIPS, 2023.
- Saeed Marzban, Erick Delage, Jonathan Y. Li, Deep Reinforcement Learning for Equal Risk Pricing and Hedging under Dynamic Expectile Risk Measures, Quantitative Finance, 23(10):1411-1430, 2023.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. Human-Level Control through Deep Reinforcement Learning. Nature 518(7540): 529–33, 2015.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. Training Language Models to Follow Instructions with Human Feedback. NeurIPS, 2022.
- Alois Pichler and Alexander Shapiro. Risk averse stochastic programming: time consistency and optimal stopping. arXiv:1808.10807, 2018.
- Pierpaolo Pontrandolfo, Abhijit Gosavi, O. Geoffrey Okogbaa, , and Tapas K. Das. Global supply chain management: A reinforcement learning approach. International Journal of Production Research, 40(6):1299–1317, 2002.
- Reazul Hasan Russel and Marek Petrik. Beyond confidence regions: tight Bayesian ambiguity sets for robust MDPs. NeurIPS, 2019.
- Andrzej Ruszczynski. Risk-averse dynamic programming for Markov decision processes. Mathematical Programming, 125(2):235–261, 2010.
- Yun Shen, Wilhelm Stannat, and Klaus Obermayer. Risk-sensitive Markov control processes. SIAM Journal on Control and Optimization, 51(5):3652–3672, 2013.
- Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. Neural Computation, 26(7):1298–1328, 2014.
- David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature 529(7587): 484–89, 2016.
- Gerald Tesauro. Temporal Difference Learning and TD-Gammon. Communication of the ACM 38(3): 58–68, 1995.
- Nicky D. van Foreest and Jacob Wijngaard. Analysis of a Stochastic Lot Scheduling Problem with Strict Due-Dates. In Richard J. Boucherie and Nico M. van Dijk, editors, Markov Decision Processes in Practice, pages 407–423. Springer International Publishing, 2017.
- Daphne van Leeuwen and Rudesindo Nunez-Queija. Near-Optimal Switching Strategies for a Tandem Queue. Richard J. Boucherie and Nico M. van Dijk, editors, Markov Decision Processes in Practice, pages 439–459. Springer International Publishing, Cham, 2017.
- Weikai Wang and Erick Delage. Planning and Learning in Average Risk-aware MDPs, arXiv:2503.17629, 2025.