Reinforcement Learning Methods for Risk Averse Sequential Decision Making

(LinkedIn)

Erick Delage
Department of Decision Sciences
HEC MONTRÉAL

(slides)

(joint work with Saeed Marzban (EY), Jonathan Y. Li (U. of Ottawa), Jia Lin Hau (JP Morgan), Marek Petrik (U. of New Hampshire), Mohammad Ghavamzadeh (Amazon), Esther Derman (Mila))

Finance Innovations & AI Seminar at CIBC
Wednesday, February 11th, 2026

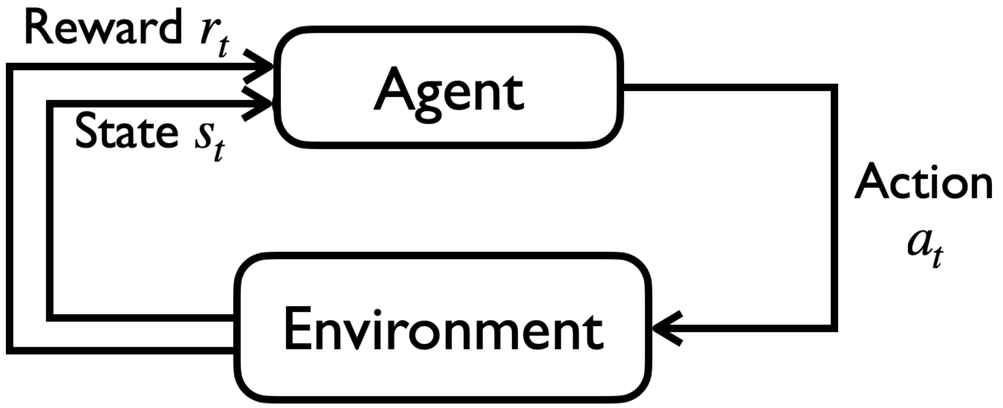
GERAD Mila IVADO Canada Research Chairs Chaires de recherche du Canada Canada

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 quantitative finance

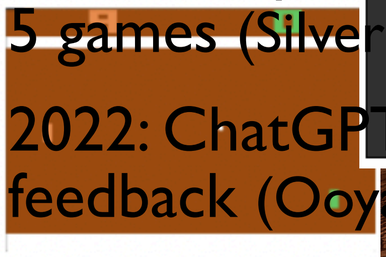
- MDPs of different forms are used in a wide range of quantitative finance applications:
  - Portfolio management
    (Moody et al. [1998], Park et al. [2020], Marzban et al. [2023])
  - **Option pricing and hedging**
    (Longstaff & Schwartz [2001], Halperin [2019], Cao et al. [2021])
  - Algorithmic trading
    (Nevmyvaka [2006], Shen et al. [2014], Lin & Beling [2020])
  - Market making
    (Vadim & Linetsky [2021])
  - Robo-advising
    (Alsabah et al. [2021])
  - Etc.

The rise of deep reinforcement learning

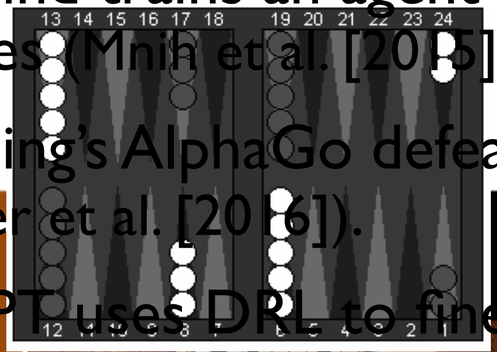
- 1991: TD-Gammon learns to play backgammon and surpasses some of the best human players (Tesauro [1995]).
- 2015: DeepMind trains an agent that achieves human level performance on Atari games (Mnih et al. [2015]).
- 2016: DeepMing's AlphaGo defeats world champion Lee Sedol in 4 out of 5 games (Silver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

The rise of deep reinforcement learning

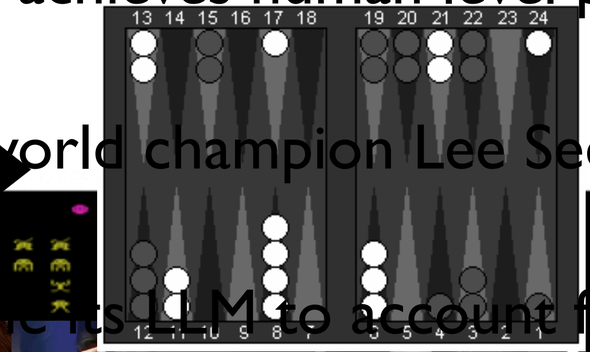
- 1991: TD-Gammon learns to play backgammon and surpasses some of the best human players (Tesauro [1995]).
- 2015: DeepMind trains an agent that achieves human level performance on Atari games (Mnih et al. [2015]).
- 2016: DeepMing's AlphaGo defeats world champion Lee Sedol in 4 out of 5 games (Silver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



Pong

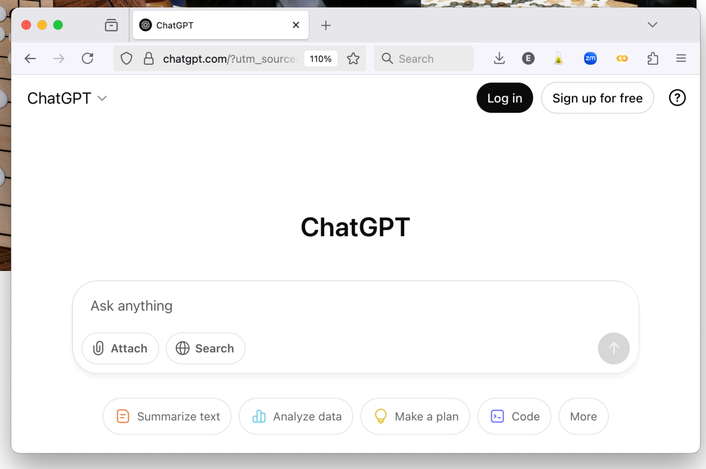


Backgammon opening position



TD-Gamm… superior position…

Beam Rider



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) (Lillicrap et al. [2019])

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 :

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

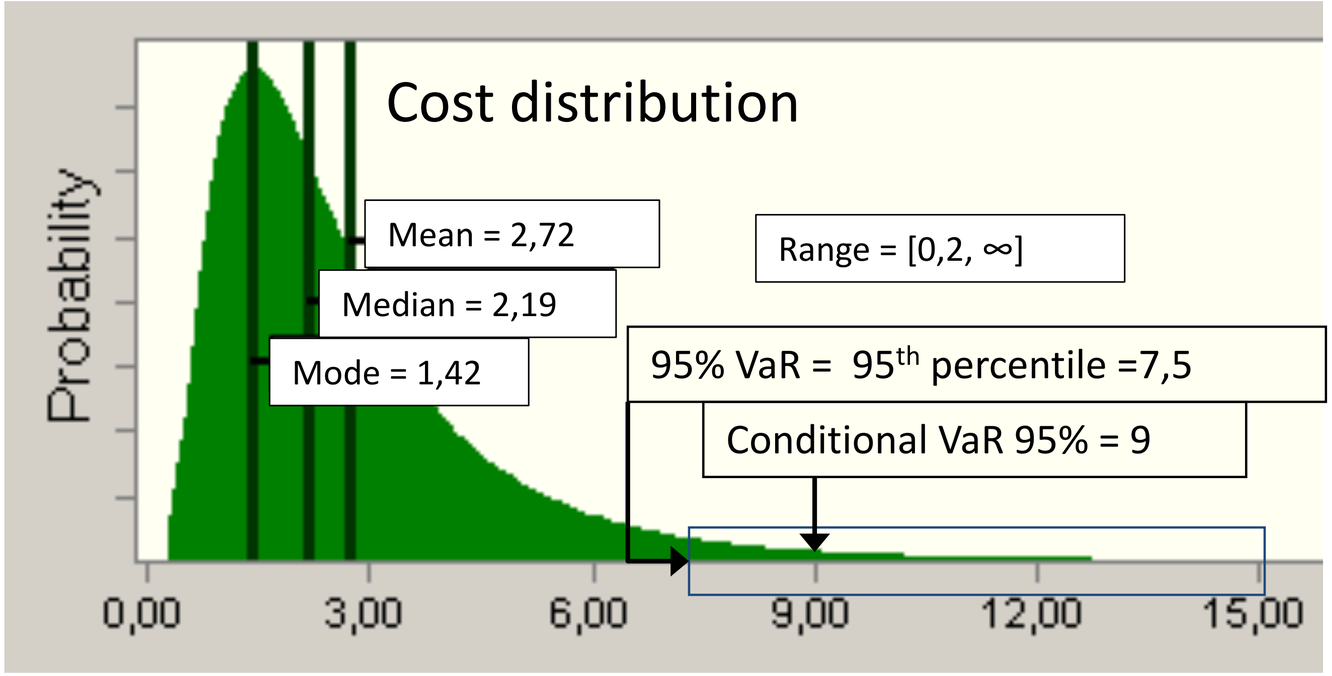
    Update the target networks: $(\bar{\theta}_\pi, \bar{\theta}_Q) \leftarrow (1 - \alpha)(\bar{\theta}_\pi, \bar{\theta}_Q) + \alpha(\theta_\pi, \theta_Q)$
  end for
end for

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

- What are Elicitable Risk Measures?

- Q-learning when using a Static Quantile Measure

- Q-learning when using a Dynamic Expectile Risk Measure

- Option Hedging and Pricing using Risk Averse DDPG

- Conclusion

What are Elicitable Risk Measures?

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: $\rho(\tilde{X}) := \mathbb{E}[\tilde{X} \,|\, \tilde{X} \geq F_X^{-1}(\alpha)]$

  ✗ Quantile: $\rho(\tilde{X}) := \text{Quant.}_\tau(\tilde{X}) = F_X^{-1}(\tau)$ (violates subadditivity)

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**: $\ell_\tau(y) := (1-\tau)\max(y,0)^2 + \tau\max(-y,0)^2$

Expectile risk measure

- Definition:

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

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

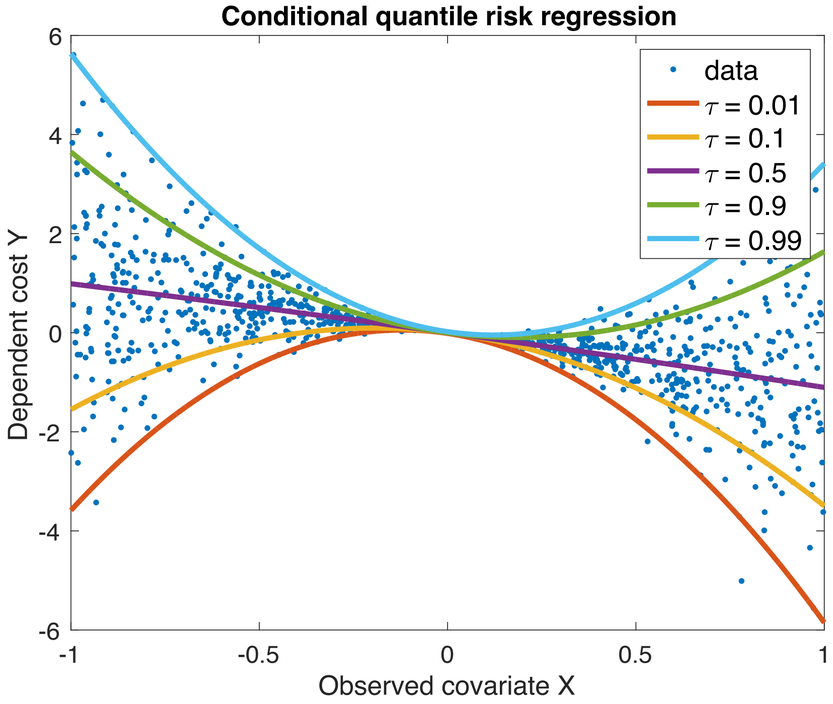
- Examples:
  - $\tau = 0 \Rightarrow \bar{\rho}(\tilde{X}) = \text{ess inf}[\tilde{X}]$, i.e. best-case scenario
  - $\tau = 0.5 \Rightarrow \bar{\rho}(\tilde{X}) = \mathbb{E}[\tilde{X}]$, i.e. risk neutral
  - $\tau = 1 \Rightarrow \bar{\rho}(\tilde{X}) = \text{ess sup}[\tilde{X}]$, i.e. worst-case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

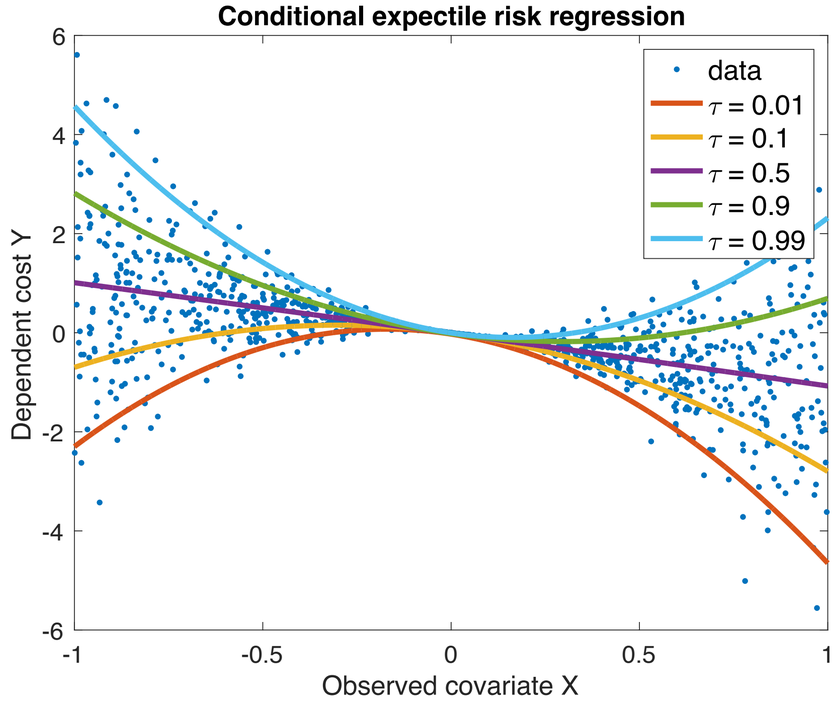
- Expectile with $\tau \in [0.5, 1]$ is the class of all elicitable coherent risk measures [Bellini and Bignozzi, 2015]

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:

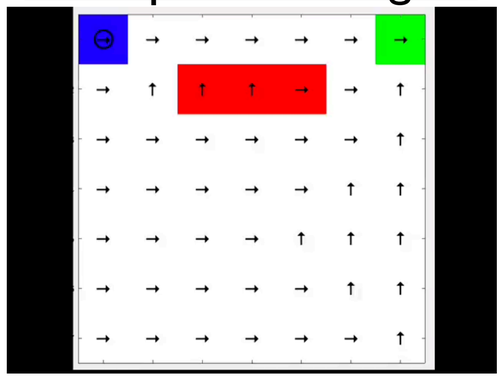
$$\theta^* = \arg\min_{\theta} \frac{1}{M} \sum_{i=1}^{M} \ell(h_\theta(x_i) - y_i) \;\Rightarrow\; \bar{\rho}(\tilde{Y} \,|\, \tilde{X}) \approx h_{\theta^*}(\tilde{X})$$



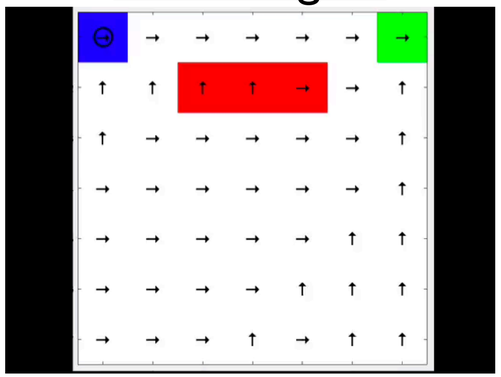


Q-learning when using a Static Quantile Measure

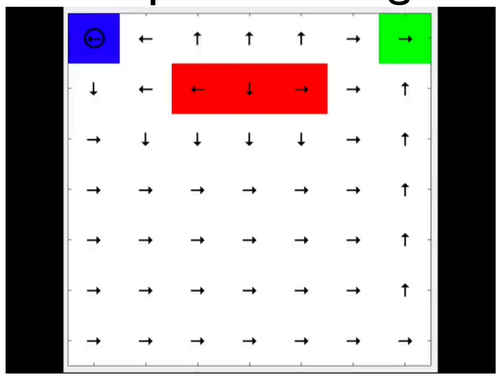
10-th percentile agent



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:

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

  - 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:

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

- 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:

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

Bellman equations for Quantile MDP

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

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

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

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

Bellman equations for Quantile MDP

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

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

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

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

Bellman equations for Quantile MDP

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

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

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

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

Bellman equations for Quantile MDP

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

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

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

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

where

$$\pi_2^*(s,\tau)\in\arg\min_a Q_2^*(s,\tau,a) := \mathsf{VaR}_\tau(-r(s,a)\,|\,\tilde{s}_2=s) = -r(s,a)$$

$$\pi_1^*(s,\tau)\in\arg\min_a Q_1^*(s,\tau,a) := \mathsf{VaR}_\tau(-r(s,a)+\gamma\min_{a'}\mathsf{VaR}_{\tilde{u}_2}(-r(\tilde{s}_2,a')\,|\,\tilde{s}_2)\,|\,\tilde{s}_1=s)$$

$$= \mathsf{VaR}_\tau(-r(s,a)+\gamma\min_{a'}Q_2^*(\tilde{s}_2,\tilde{u}_2,a')\,|\,\tilde{s}_1=s)$$

$$\pi_0^*(s,\tau)\in\arg\min_a Q_0^*(s,\tau,a) := \mathsf{VaR}_\tau(-r(s,a)+\gamma\min_{a'}Q_1^*(\tilde{s}_1,\tilde{u}_1,a'))$$

Bellman equations for Quantile MDP

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

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

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

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

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

where

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

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

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

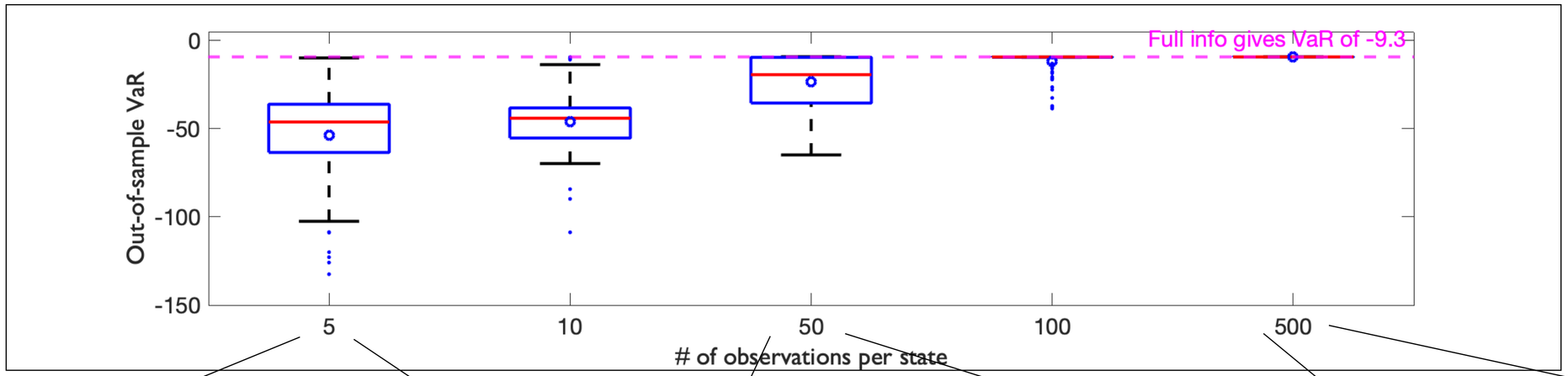
$$Q_t^k(s, \tau, a) \leftarrow Q_t^{k-1}(s, \tau, a), \quad \forall (t, s, \tau, a) \neq (t_k, s_k \,.\, \tau_k, a_k)$$

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

Learning optimal VaR policy in the gridworld



Start (r = -1)

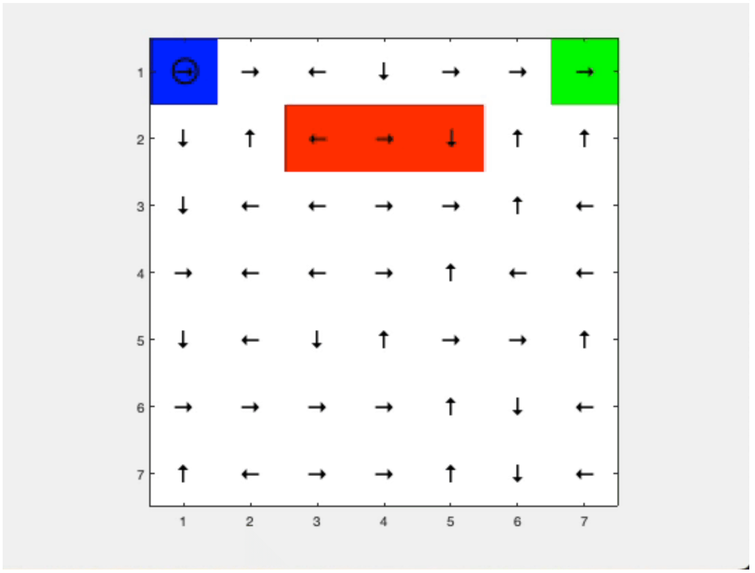
Goal (r = 0)

Damage (r = -50)

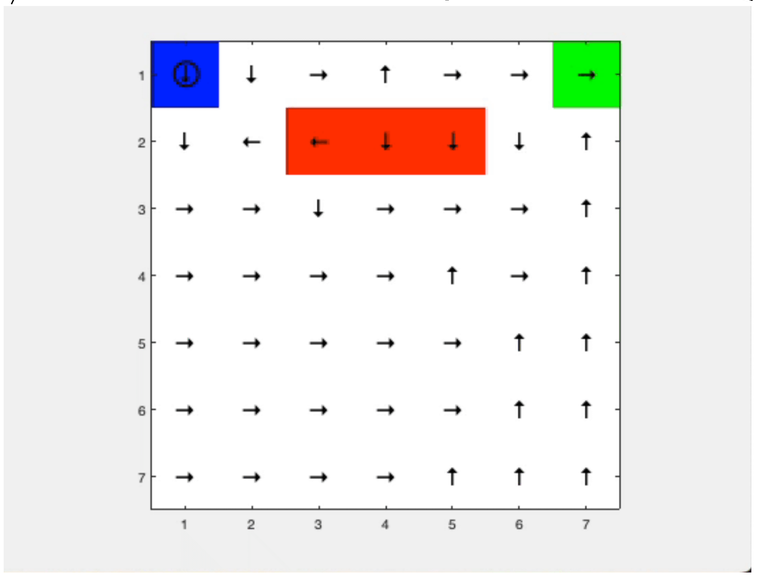
Nothing (r = -1)

Optimal policy

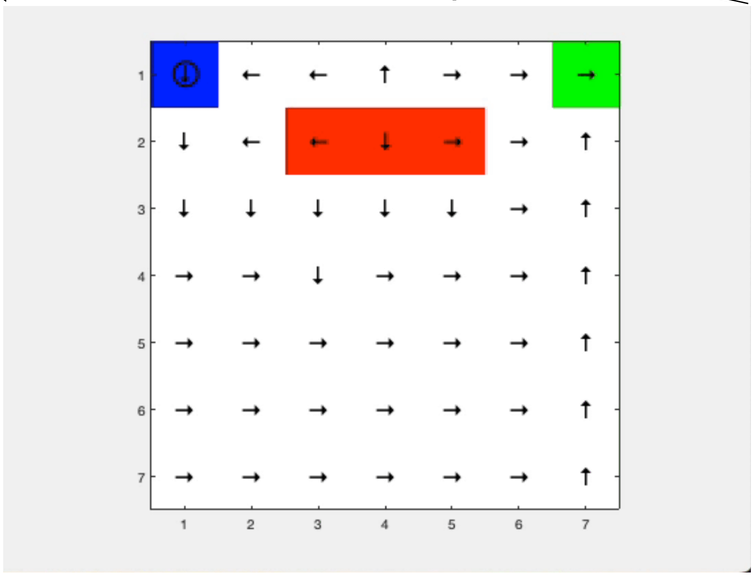
5 observations per state



50 observations per state

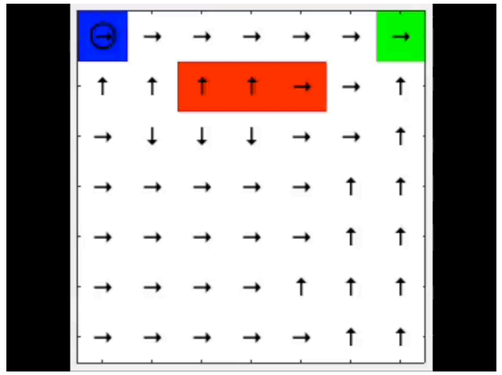


500 observations per state

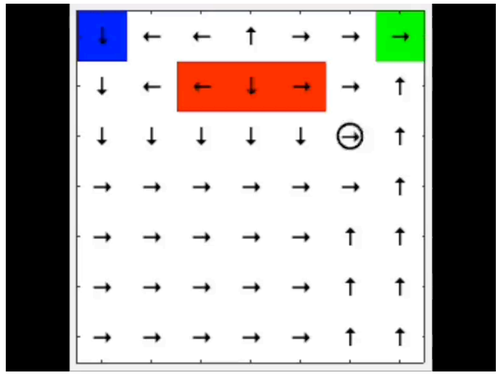


Q-learning when using a Dynamic Expectile Risk Measure

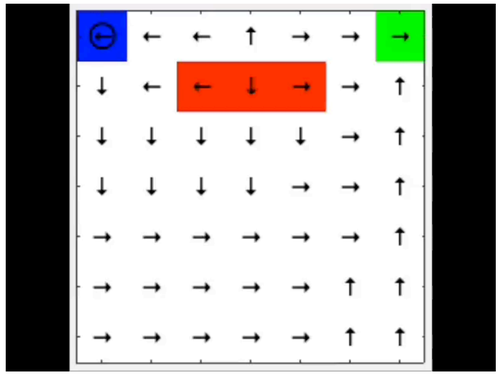
Risk seeking agent

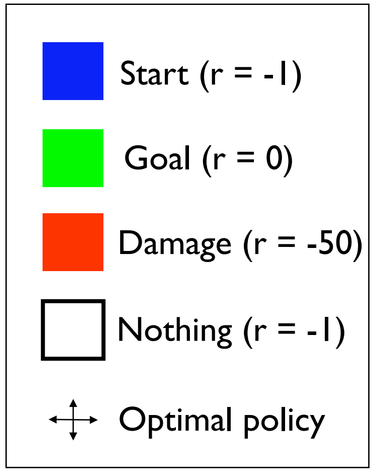


Risk neutral agent



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.

Dynamic expectile risk measure (DERM)

- Definition:

  A dynamic expectile risk measure takes the form:

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

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

- Similar steps as for our quantile risk measure lead, when $T = 3$ to:

$$\begin{aligned}
\min_{\pi}\ \rho(-\tilde{R}(\pi)) &= \min_{\pi}\ \bar{\rho}_0\Big(\,\bar{\rho}_1\Big(\bar{\rho}_2(\,-\sum_{t=0}^{2}\gamma^t r(\tilde{s}_t, \tilde{a}_t)\,|\,\tilde{a}_{0:1}, \tilde{s}_{1:2})\,\Big|\,\tilde{a}_0, \tilde{s}_1\Big)\Big)\\
&= \min_{a_0}\bar{\rho}_0\Big(-r(s_0, a_0) + \gamma\min_{a_1}\bar{\rho}_1\Big(-r(\tilde{s}_1, a_1) + \gamma\min_{a_2}\bar{\rho}_2(-r(\tilde{s}_2, a_2)\,|\,\tilde{s}_2)\,\Big|\,\tilde{s}_1\Big)\Big)\\
&= \bar{\rho}_0\Big(\,-r(s_0, \pi_0^*(s_0)) + \gamma\bar{\rho}_1\Big(\,-r(\tilde{s}_1, \pi_1^*(\tilde{s}_1)) + \gamma\bar{\rho}_2\Big(\,-r(\tilde{s}_2, \pi_2^*(\tilde{s}_2))\,\Big|\,\tilde{s}_2\Big)\,\Big|\,\tilde{s}_1\Big)\Big)
\end{aligned}$$

  where

$$\begin{aligned}
\pi_2^*(s) &\in \arg\min_a Q_2^*(s,a) := -\,r(s,a)\\
\pi_1^*(s) &\in \arg\min_a Q_1^*(s,a) := \bar{\rho}_1(-r(s,a) + \gamma\min_{a'} Q_2^*(\tilde{s}_2, a')\,|\,\tilde{s}_1 = s)\\
\pi_0^*(s) &\in \arg\min_a Q_0^*(s,a) := \bar{\rho}_0(-r(s,a) + \gamma\min_{a'} Q_1^*(\tilde{s}_1, a'))
\end{aligned}$$

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

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

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

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

- This generalizes the Q-learning update for RN case, where $\ell_{0.5}(y) := (1/2)y^2$ and $\ell'_{0.5}(y) = y$:

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

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

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

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

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

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 (RA-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

Option Hedging and Pricing using Risk Averse DDPG

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

What is an option ?

- An option is a type of security that provides the owner with the right to trade a fixed number of shares of an asset at a fixed price (strike price) at a time on or before a given date (maturity)
  [Cox et al., 1979]

A European call option example: $Payoff(S_T) := \max(0, S_T - K)$

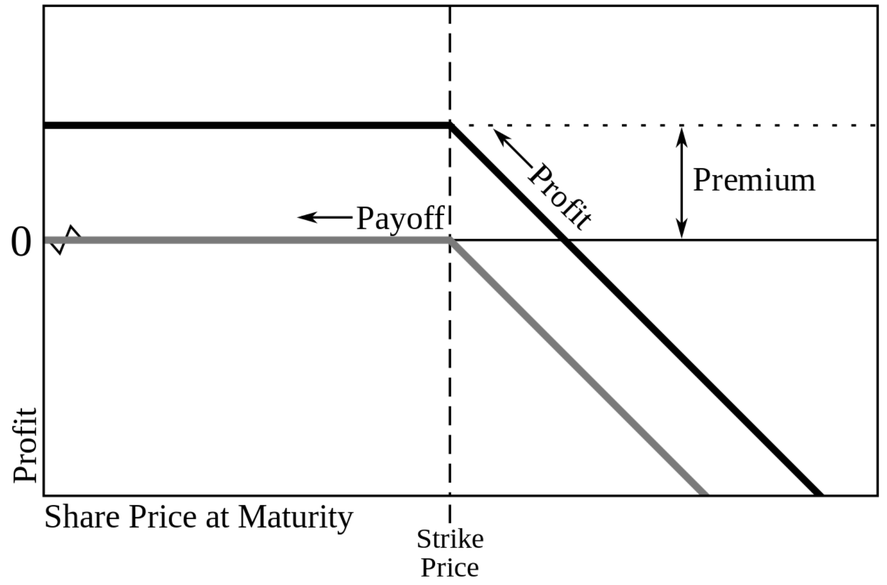


Figure: Profits from writing a call option

Hedging an option in a complete market

- Cox et al. [1979] presents a perfect hedging approach for "complete markets":

$$\text{Asset: } S \longrightarrow \begin{cases} \omega_1 : uS & \mathbb{P}(\omega_1) = q, \\ \omega_2 : dS & \mathbb{P}(\omega_2) = 1 - q, \end{cases}$$

$$\text{Option: } w_0 \longrightarrow \begin{cases} \omega_1 : P_u = \max\{0, uS - K\} & \mathbb{P}(\omega_1) = q, \\ \omega_2 : P_d = \max\{0, dS - K\} & \mathbb{P}(\omega_2) = 1 - q, \end{cases}$$

$$\begin{matrix}\text{Replicating} \\ \text{portfolio}\end{matrix} : \xi S + \zeta \longrightarrow \begin{cases} \omega_1 : \xi uS + \zeta & \mathbb{P}(\omega_1) = q, \\ \omega_2 : \xi dS + \zeta & \mathbb{P}(\omega_2) = 1 - q \end{cases}$$

$$\begin{matrix} \omega_1 : \xi^* uS + \zeta^* = P_u, \\ \omega_2 : \xi^* dS + \zeta^* = P_d \end{matrix} \Rightarrow \begin{matrix} \xi^* = \frac{P_u - P_d}{(u-d)S}, \\ \zeta^* = \frac{uP_d - dP_u}{(u-d)}, \end{matrix}$$

Replicating portfolio cost : $\xi^* S + \zeta^*$

- An arbitrage emerges if price is different from replicating portfolio cost

Hedging an option in an incomplete market

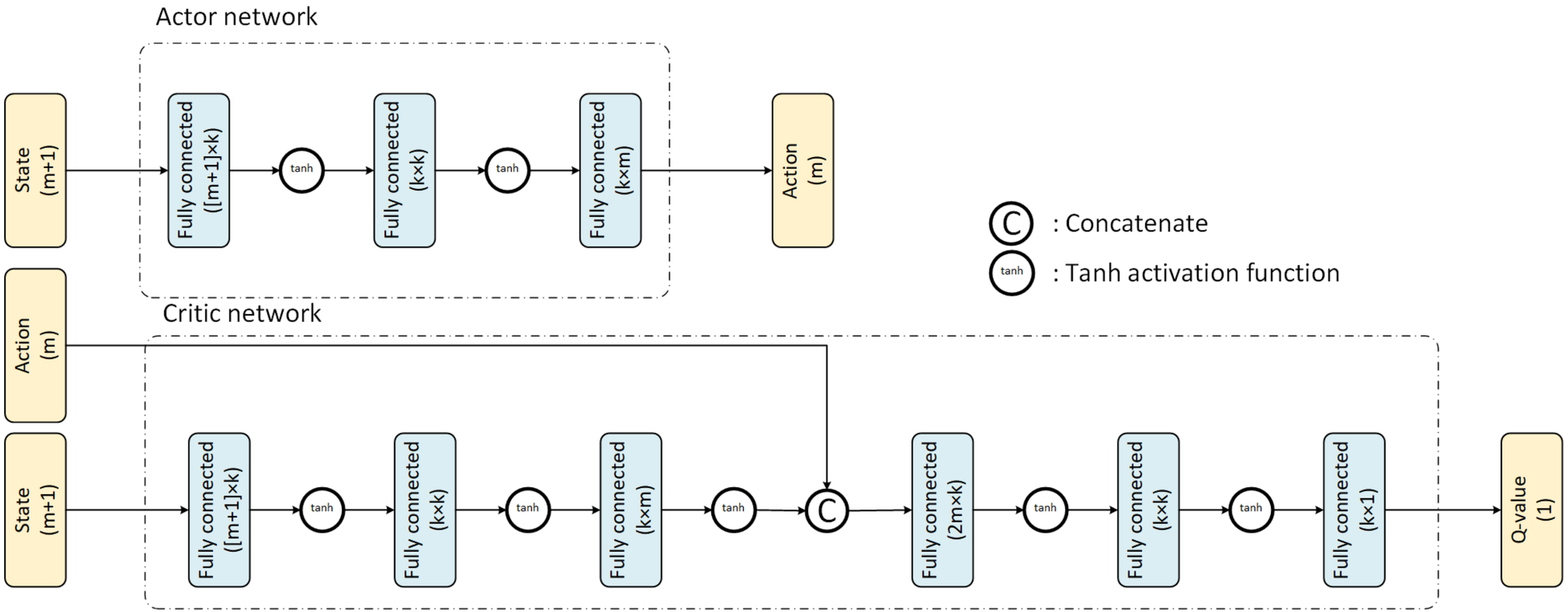
- The problem is when the market is incomplete, it is impossible to perfectly replicate the option.
- All strategies expose the writer of the option to a risk of losses.
- Risk-averse deep RL method:

$$\min_{\pi} \rho(-\tilde{R}(\pi)) = \rho\left(\begin{array}{c}\text{Self-financed portfolio return} \\ -\text{Option payoff}\end{array}\right)$$

where

- State s keeps track of the asset values S and state of the Markov process modelling the market dynamics
- Action $a_t \in [-1,1]^m$ composes the portfolio
- Reward: $r(s_t, a_t, s_{t+1}) := \begin{cases} a_t^\top (S_{t+1} - S_t) & \text{if } t < T \\ -Payoff(S_t) & \text{if } t = T \end{cases}$

Network architecture used in RA-DDPG



Empirical study set-up

- Vanilla option set-up:
  - Brownian motion market model (calibrated on AAPL)
  - At-the-money strike price
  - Maturity = 12 months
  - Monthly rebalancing
- Basket option set-up:
  - Multi-asset geometric Brownian motion (calibrated on AAPL, AMZN, FB, JPM, and GOOGL)
  - Payoff on average asset price
  - At-the-money strike price
  - Maturity = 12 months
  - Monthly rebalancing
- We compare three hedging schemes:
  - Static 90% expectile RL (RA-SRM)
  - Dynamic 90% expectile RL (RA-DDPG)
  - Dynamic 90% expectile with dynamic programming (RA-DP)
- Option pricing: the estimated risk can be interpreted as minimum acceptable price for the option

Vanilla option: Precision of risk-averse RL solution

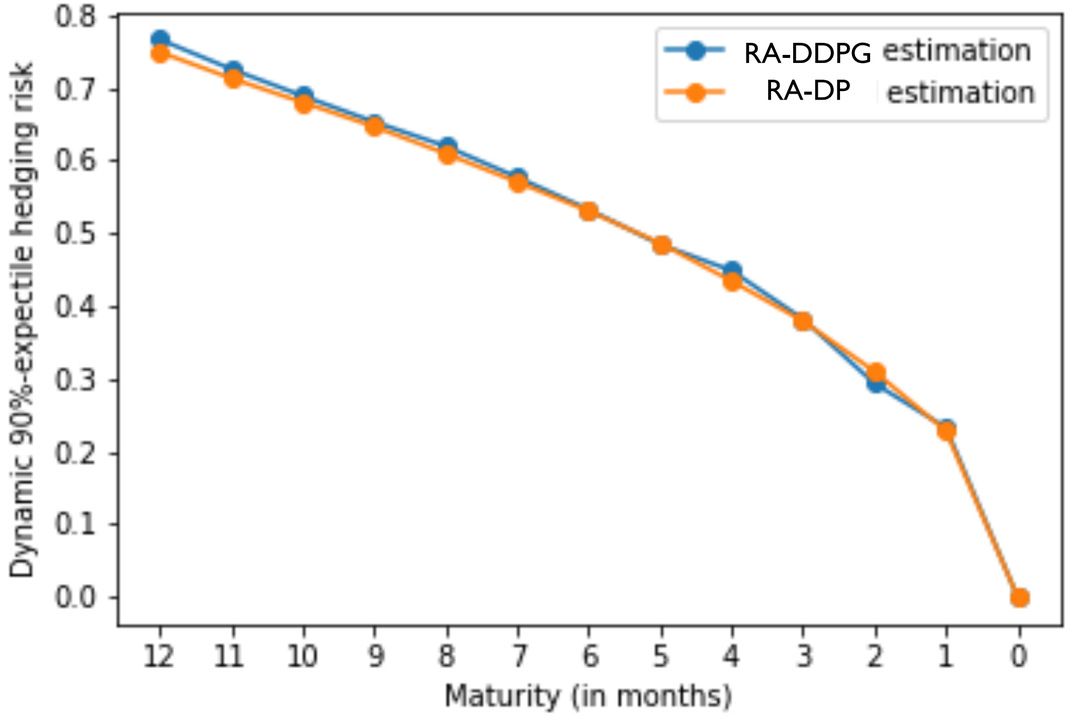


Figure: Dynamic risk of hedging a vanilla at-the-money call option over APPL under the RA-DDPG policy trained for a 12 months maturity.

Observations:

- RA-DDPG algorithm learns accurate hedging policy
- Option coverage cost decreases with maturity

Vanilla option: Static risk exposure

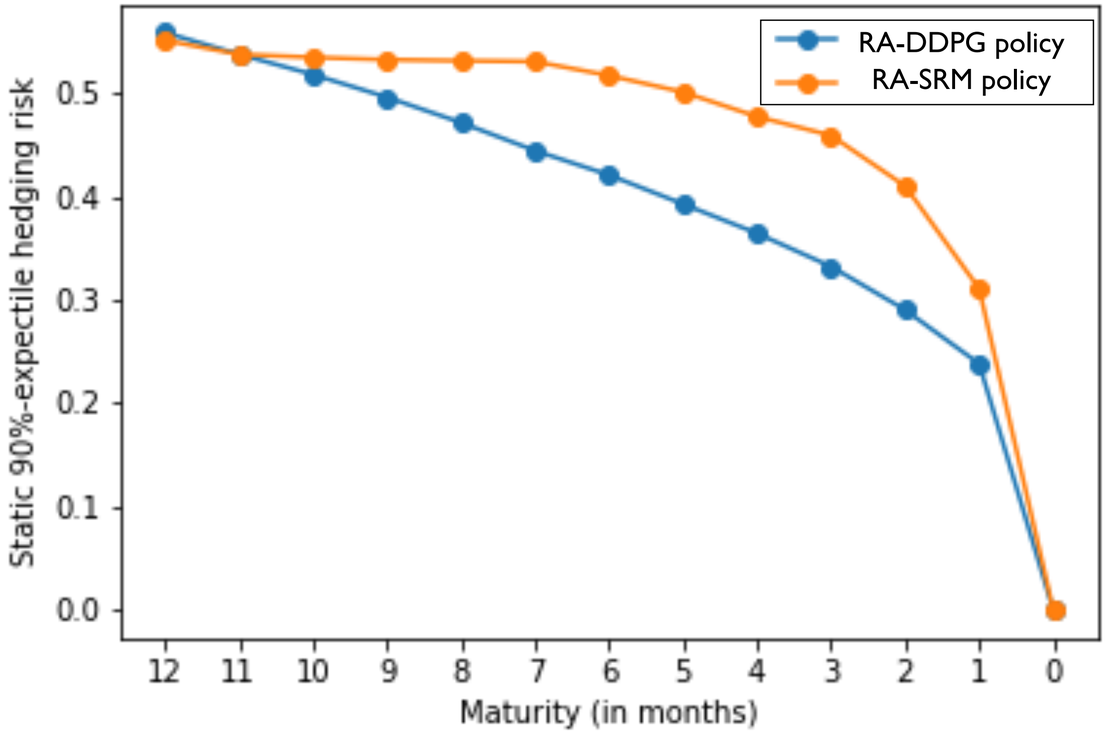


Figure: Static risk for hedging a vanilla at-the-money call option over APPL (with maturity ranging from 12 months to 0 months) under the DERM (RA-DDPG) and Static Risk Measure (RA-SRM) policies trained for a 12 months maturity.

Observations:

- RA-DDPG policy achieves comparable performance to RA-SRM at initial maturity
- RA-DDPG policy offers better protection as time passes compared to RA-SRM

Basket Option: Static risk exposure

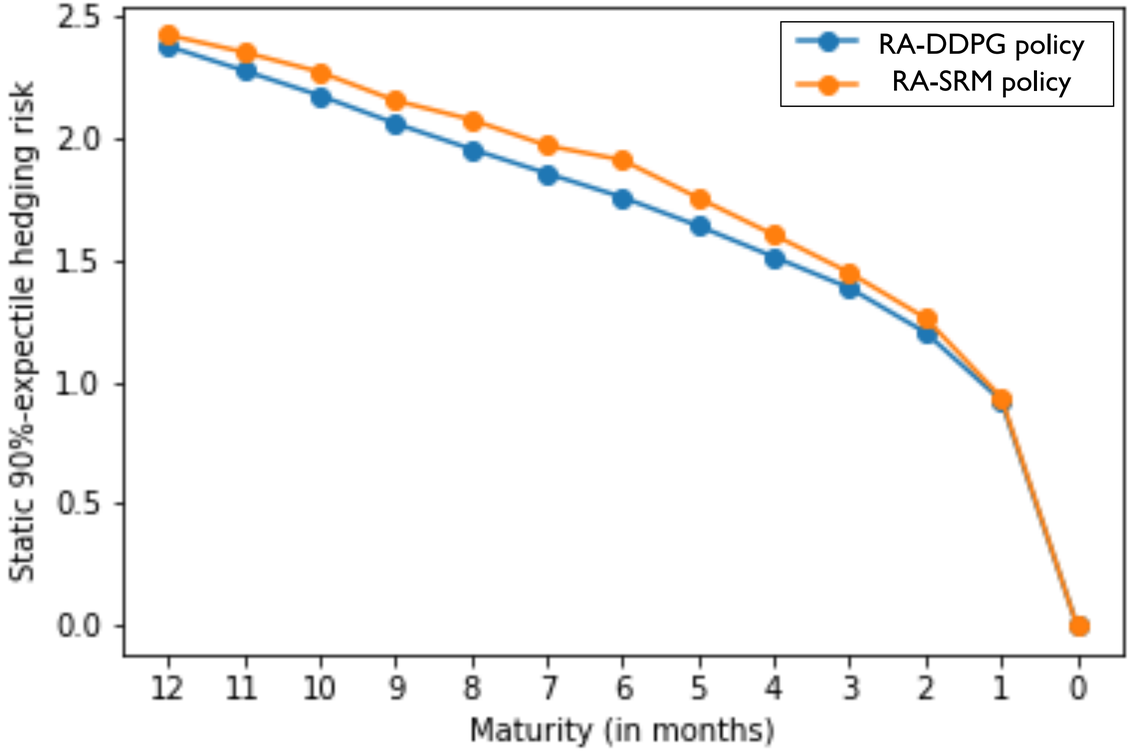


Figure: Static risk for hedging at-the-money call basket option over 5 assets (with maturity ranging from 12 months to 0 months) under the DERM (RA-DDPG) and Static Risk Measure (RA-SRM) policies trained for a 12 months maturity.

Observations:

- RA-DDPG policy achieves comparable performance RA-SRM in terms of static risk
- Unlike RA-SRM policy, the RA-DDPG policy continues to remain consistent as time passes by

Take-away messages

- Elicitability allows one to adapt deep RL methods to risk aware setting (e.g. DDPG).
- Different types of risk measures can be used:
  - Static risk measures
  - Dynamic risk measures
- Different types of problems:
  - Finite, infinite discounted, infinite average risk 
- Risk-averse deep RL algorithms can potentially identify risk aware policies in real world large-scale sequential decision making problems.
- Many potential applications in quantitative finance!

References

- Humoud Alsabah, Agostino Capponi, Ocatvio Ruiz Lacedelli, and Matt. Stern, Robo-advising: Learning investors' risk preferences via portfolio choices, Journal of Financial Econometrics, 19:369–392, 2021.
- 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.
- Jay Cao, Jacky Chen, John Hull, Zissis Poulos, Deep hedging of derivatives using reinforcement learning, The Journal of Financial Data Science, 3:10–27, 2021.
- John C. Cox, Stephen A. Ross, and Mark Rubinstein, Option pricing: A simplified approach. Journal of Financial Economics, 7(3): 229-263, 1979.
- 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.
- Igor Halperin, The QLBS Q-learner goes NuQlear: Fitted Q iteration, inverse RL, and option portfolios, Quantitative Finance, 19:1543–1553, 2019.
- 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.
- Xiaocheng Li, Huaiyang Zhong, and Margaret L. Brandeau. Quantile Markov Decision Processes. Operations Research, 70(3):1428–1447, 2022.
- Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous Control with Deep Reinforcement Learning. arXiv:1509.02971, 2019.
- Siyu Lin and Peter A. Beling, An end-to-end optimal trade execution framework based on proximal policy optimization, IJCAI, 2020.
  Elita A. Lobo, Cyrus Cousins, Yair Zick, and Marek Petrik. Percentile criterion optimization in offline reinforcement learning. NeurIPS, 2023.
- Francis A. Longstaff and Eduardo S. Schwartz, Valuing American Options by Simulation: A Simple Least-Squares Approach, The Review of Financial Studies, 14(1):113–147, 2001.
- 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.
- Saeed Marzban, Erick Delage, Jonathan Y. Li, Jeremie Desgagne-Bouchard, Carl Dussault, WaveCorr: Deep Reinforcement Learning with Permutation Invariant Policy Networks for Portfolio Management, Operations Research Letters, 51(6):680-686, 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.
- John Moody, Lizhong Wu, Yuansong Liao, and Matthew Saffell, Performance functions and reinforcement learning for trading systems and portfolios. Journal of Forecasting, 17: 441-470, 1998.
- Yuriy Nevmyvaka, Yi Feng, and Michael Kearns, Reinforcement learning for optimized trade execution, ICML, 2006.
- 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.
- Hyungjun Park, Min Kyu Sim, and Dong Gu Choi, An intelligent financial portfolio trading strategy using deep Q-learning, Expert Systems with Applications, 158, 2020.
- Alois Pichler and Alexander Shapiro. Risk averse stochastic programming: time consistency and optimal stopping. arXiv:1808.10807, 2018.
- 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, Ruihong Huang, Chang Yan and Klaus Obermayer, Risk-averse reinforcement learning for algorithmic trading, IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), 2014.
- 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.
- Muchen Zhao and Vadim Linetsky, High frequency automated market making algorithms with adverse selection risk control via reinforcement learning, in Proceedings of the Second ACM International Conference on AI in Finance, 2021.