Deep Reinforcement Learning for Risk Averse Multi-stage Decision Making Problems

Erick Delage,
Department of Decision Sciences
HEC MONTRÉAL

(joint work with Saeed Marzban, Jonathan Y. Li (U. of Ottawa))

May 29, 2023

GERAD

Canada Research Chairs
Chaires de recherche du Canada
Canada

Risk aversion in multistage decision making

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

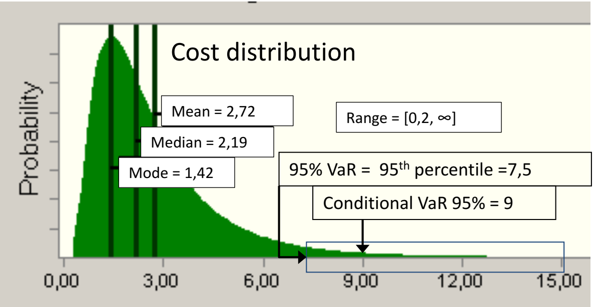
$$\tilde{R}(\pi) := \sum_{t=0}^{T-1} r_t(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})$$

where $\{\tilde{s}_t\}_{t=0}^{T}$ is the random state trajectory traversed when drawing actions from policy π_t, i.e. $\tilde{a}_t \sim \pi_t(\tilde{s}_t)$. We assume that s_0 is deterministic.

Risk aversion in multistage decision making

Risk aversion can be handled using two approaches:

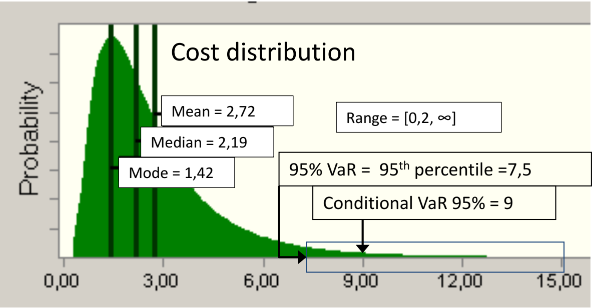
1. Static law-invariant risk measure (SRM):
   $\min_\pi \bar{\rho}(-\tilde{R}(\pi)) := \bar{\varrho}(F_{\tilde{R}(\pi)})$
   - E.g. : $-\mathbb{E}[\tilde{R}]$, $-\mathbb{E}[u(\tilde{R})]$, $\mathrm{VaR}(-\tilde{R})$, $\mathrm{CVaR}(-\tilde{R})$



Risk aversion in multistage decision making

Risk aversion can be handled using two approaches:

1. Static law-invariant risk measure (SRM): $\min_\pi \bar{\rho}(-\tilde{R}(\pi)) := \bar{\varrho}(F_{\tilde{R}(\pi)})$
   - E.g. : $-\mathbb{E}[\tilde{R}]$, $-\mathbb{E}[u(\tilde{R})]$, $\mathrm{VaR}(-\tilde{R})$, $\mathrm{CVaR}(-\tilde{R})$
   - Pros: Easy to interpret
   - Cons: Can violate dynamic consistency
   - Pro or Con ?: Does not distinguish between two policies that have the same $F_{\tilde{R}(\pi)}$



Risk aversion in multistage decision making

Risk aversion can be handled using two approaches:

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):
   $\max_\pi \rho(-\tilde{R}(\pi)) :=$
   $\bar{\rho}_0(\bar{\rho}_1(\ldots \bar{\rho}_{T-1}(-\tilde{R}(\pi)|\tilde{a}_{0:T-1}, \tilde{s}_{1:T}) \cdots |\tilde{a}_0, \tilde{s}_1))$
   - E.g.: $\mathbb{E}[-\tilde{R}]$, $-\mathbb{E}[u(\tilde{R})]$,
     $\mathrm{VaR}(\mathrm{VaR}(\ldots \mathrm{VaR}(-\tilde{R}|\tilde{a}_{0:T-1}, \tilde{s}_{1:T}) \ldots |\tilde{a}_0, \tilde{s}_1))$,
     $\mathrm{CVaR}(\mathrm{CVaR}(\ldots \mathrm{CVaR}(-\tilde{R}|\tilde{a}_{0:T-1}, \tilde{s}_{1:T}) \ldots |\tilde{a}_0, \tilde{s}_1))$

Risk aversion in multistage decision making

Risk aversion can be handled using two approaches:

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):
   $\max_\pi \rho(-\tilde{R}(\pi)) :=$
   $\bar{\rho}_0(\bar{\rho}_1(\ldots \bar{\rho}_{T-1}(-\tilde{R}(\pi)|\tilde{a}_{0:T-1}, \tilde{s}_{1:T}) \cdots |\tilde{a}_0, \tilde{s}_1))$
   - E.g.: $\mathbb{E}[-\tilde{R}]$, $-\mathbb{E}[u(\tilde{R})]$,
     $\mathrm{VaR}(\mathrm{VaR}(\ldots \mathrm{VaR}(-\tilde{R}|\tilde{a}_{0:T-1}, \tilde{s}_{1:T}) \ldots |\tilde{a}_0, \tilde{s}_1))$,
     $\mathrm{CVaR}(\mathrm{CVaR}(\ldots \mathrm{CVaR}(-\tilde{R}|\tilde{a}_{0:T-1}, \tilde{s}_{1:T}) \ldots |\tilde{a}_0, \tilde{s}_1))$
   - Pros: Satisfies dynamic consistency, associated to Bellman equation
   - Cons: Can be hard to interpret
   - Pro or Con ?: Unclear how it handles two policies that have the same $F_{\tilde{R}(\pi)}$

Outline

Introduction

Deep RL for dynamic elicitable risk measure

Equal Risk Option Pricing

Outline

Introduction

Deep RL for dynamic elicitable risk measure

Equal Risk Option Pricing

DEEP RL FOR DYNAMIC RISK MEASURES

- Tamar et al. [2015] exploits risk measure supremum representation to obtain robust MDP reformulation. Policy gradient obtained by simulating the trajectory using reweighted transitions.
- Huang et al. [2021] modifies policy gradient for on-policy learning but requires up to 5 function approximators.
- **Marzban et al. [2023] proposes a simple modification to Deep Deterministic Policy Gradient (DDPG) algorithm to handle dynamic elicitable risk measures**.
- Coache et al. [2022] proposes an on-policy actor-critic approach for conditionally elicitable risk measures.

Elicitable risk measure [Bellini and Bignozzi, 2015]

Definition 1

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

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

- Examples:
  - Expected value: $\ell(y) := y^2$
  - Quantile value: $\ell_\tau(y) := (1-\tau)\max(y,0) + \tau\max(-y,0)$

ELICITABLE RISK MEASURE [BELLINI AND BIGNOZZI, 2015]

Definition 1

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

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

- Examples:
  - Expected value: $\ell(y) := y^2$
  - Quantile value: $\ell_\tau(y) := (1-\tau)\max(y,0) + \tau\max(-y,0)$
- Elicitability implies that if we have i.i.d. samples $\{x_i, y_i\}_{i=1}^M$ then we can estimate conditional risk using regression:

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

Expectile risk measure

Definition 2
The τ-expectile of a random liability $\tilde{X}$ is defined as:

$$\bar{\rho}(\tilde{X}) := \arg\min_q \mathbb{E}\left[(1-\tau)(q-\tilde{X})_+^2 + \tau(q-\tilde{X})_-^2\right].$$

- $\tau = 0.5 \Rightarrow \bar{\rho}(\tilde{X}) = \mathbb{E}[\tilde{X}]$, i.e. risk neutral
- $\tau = 1 \Rightarrow \bar{\rho}(\tilde{X}) = \operatorname{ess\,sup}[\tilde{X}]$, i.e. worst-case scenario
- Expectile is the only elicitable coherent risk measure

Dynamic expectile risk measure (DERM)

Definition 3

A dynamic recursive expectile risk measure takes the form:

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

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

$$\bar{\rho}_t(\tilde{V}_{t+1}|\tilde{a}_{0:t-1}, \tilde{s}_{1:t}) := \arg\min_q \mathbb{E}\left[\tau(q - \tilde{V}_{t+1})_+^2 + (1-\tau)(q - \tilde{V}_{t+1})_+^2 |\tilde{a}_{0:t-1}, \tilde{s}_{1:t}\right]$$

where for example

$$\tilde{V}_{t+1} := \bar{\rho}_{t+1}(\bar{\rho}_{t+2}(\ldots \bar{\rho}_{T-1}(-\tilde{R}|\tilde{a}_{0:T-1}, \tilde{s}_{1:T}) \ldots |\tilde{a}_{0:t+1}, \tilde{s}_{1:t+2}))$$

can be the random "risk-to-go" observable at $t+1$.

Bellman Equations for DRM-MDP

With dynamic recursive risk measures in an MDP, $\min_\pi \bar{\rho}(-\tilde{R}(\pi)) \equiv \min_\pi V_0^\pi(s_0)$ where

$$V_t^\pi(s_t) := \bar{\rho}_t(-r_t(s_t, \tilde{a}_t, \tilde{s}_{t+1}) + V_{t+1}^\pi(\tilde{s}_{t+1}) | \tilde{s}_t = s_t)$$

with $\tilde{a}_t \sim \pi_t(\tilde{s}_t)$ and $V_T^\pi(s_T) := 0$.

With interchangeability property and mixture quasi-concavity of $\bar{\rho}_t$, we have $\min_\pi \bar{\rho}(-\tilde{R}(\pi)) \equiv \min_{a_0} Q_0^*(s_0, a_0)$ where

$$Q_t^*(s_t, a_t) := \bar{\rho}_t(-r_t(s_t, a_t, \tilde{s}_{t+1}) + \min_{a_{t+1}} Q_{t+1}^*(\tilde{s}_{t+1}, a_{t+1}) | \tilde{s}_t = s_t)$$

and $Q_T^*(s_T, a_T) := 0$.

DEEP RISK AVERSE RL USING DERMS

- We show how to extend the popular deep deterministic policy gradient (DDPG) algorithm to solve dynamic problems formulated based on time-consistent dynamic expectile risk measures ?

$$Q_t^*(s_t, a_t) := \bar{\rho}_t\Big(- r_t(s_t, a_t, \tilde{s}_{t+1}) + \max_{a_{t+1}} Q_{t+1}^*(\tilde{s}_{t+1}, a_{t+1}) \Big| s_t\Big)$$

Algorithm Traditional DDPG ($\bar{\rho}_t = \mathbb{E}$)

Initialize the main actor θ^π and critic θ^Q networks
Initialize the target actor, $\theta^{\pi'}$, and critic, $\theta^{Q'}$, networks
Initialize replay buffers R
for $j = 1 : \#Episodes$ **do**
  Initialize a random process $\mathcal{N}$ for action exploration;
  Receive initial observation state s_0
  for $t = 0 : T-1$ **do**
    Select action $a_t = \pi_t(s_t|\theta^\pi) + \mathcal{N}_t$
    Execute a_t and store transition (s_t, a_t, r_t, s_{t+1})
    Sample a minibatch of N transitions
    Set $y_i := -r_i + Q(s_{i+1}, \pi(s_{i+1}|\theta^{\pi'})|\theta^{Q'})$
    Update the main critic network:

$$\theta^Q \leftarrow \theta^Q \boldsymbol{+ \alpha \frac{1}{N} \sum_{i=1}^{N} \partial \ell(Q(s_i, a_i|\theta^Q) - y_i) \nabla_{\theta^Q} Q(s_i, a_i|\theta^Q)}$$

    where $\boldsymbol{\ell(\Delta) := \Delta^2}$

    Update the main actor network :

$$\theta^\pi \leftarrow \theta^\pi - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q(s_j^i, a|\theta^Q)|_{a=\pi(s_j^i|\theta^\pi)} \nabla_{\theta^\pi} \pi(s_j^i|\theta^\pi);$$

    Update the target networks
  end for
end for

Deep Risk Averse RL using dynamic risk measures

- We show how to extend the popular deep deterministic policy gradient (DDPG) algorithm to solve dynamic problems formulated based on time-consistent dynamic expectile risk measures ?

$$Q_t^*(s_t, a_t) := \bar{\rho}_t\Big(-r_t(s_t, a_t, \tilde{s}_{t+1}) + \max_{a_{t+1}} Q_{t+1}^*(\tilde{s}_{t+1}, a_{t+1}) \Big| s_t \Big)$$

Algorithm Risk averse DDPG (ACRL)

Initialize the main actor θ^π and critic θ^Q networks
Initialize the target actor, $\theta^{\pi'}$, and critic, $\theta^{Q'}$, networks
Initialize replay buffers R
for $j = 1 : \#Episodes$ **do**
  Initialize a random process $\mathcal{N}$ for action exploration;
  Receive initial observation state s_0
  for $t = 0 : T-1$ **do**
    Select action $a_t = \pi_t(s_t|\theta^\pi) + \mathcal{N}_t$
    Execute a_t and store transition (s_t, a_t, r_t, s_{t+1})
    Sample a minibatch of N transitions
    Set $y_i := -r_i + Q(s_{i+1}, \pi(s_{i+1}|\theta^{\pi'})|\theta^{Q'})$
    Update the main critic network:

$$\theta^Q \leftarrow \theta^Q + \boldsymbol{\alpha \frac{1}{N}\sum_{i=1}^{N} \partial \ell(Q(s_i, a_i|\theta^Q) - y_i) \nabla_{\theta^Q} Q(s_i, a_i|\theta^Q)}$$

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

$$\theta^\pi \leftarrow \theta^\pi - \alpha \frac{1}{N}\sum_{i=1}^{N} \nabla_a Q(s_j^i, a|\theta^Q)|_{a=\pi(s_j^i|\theta^\pi)} \nabla_{\theta^\pi} \pi(s_j^i|\theta^\pi);$$

    Update the target networks
  end for
end for

OUTLINE

Introduction

Deep RL for dynamic elicitable risk measure

Equal Risk Option Pricing

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 call option example:

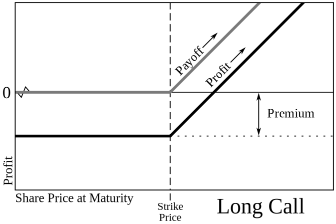


Figure: Profits from buying a call option

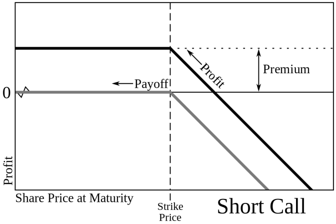


Figure: Profits from writing a call option

Call option payoff: $F(S_T) = \max\{0, S_T - K\}$

Graphs are from: https://en.wikipedia.org/wiki/Call_option

How to Price an Option in a Complete Market?

Cox et al. [1979] presents an option pricing formula that works based on the principle of no-arbitrage:

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

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

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

$$\begin{array}{c} \omega_1 : \xi^* uS + \zeta^* = F_u, \\ \omega_2 : \xi^* dS + \zeta^* = F_d \end{array} \Rightarrow \begin{array}{c} \xi^* = \frac{F_u - F_d}{(u-d)S}, \\ \zeta^* = \frac{uF_d - dF_u}{(u-d)}, \end{array} \Rightarrow w_0 = \xi^* S + \zeta^*$$

Any other price leads to arbitrage.

This approach extends to multi-periods and continuous time in so called "complete markets".

How to Price an Option in an Incomplete market?

- The problem is when the market is incomplete, i.e. it is impossible to perfectly replicate the option.
- Any given price exposes one or both parties in the trade to some risk.
- Equal Risk Pricing [Guo and Zhu, 2017] suggests choosing the price that exposes both parties to the same amount of risk.

DERM-MDP reformulation for ERP

Proposition 1

When the asset process is Markovian and risk aversion is modeled using DERM, the equal risk price is equal to

$$ERP(F) = (\min_{\pi^w} \bar{\rho}(-\tilde{R}_F^w(\pi^w)) + \min_{\pi^b} \bar{\rho}(-\tilde{R}_F^b(\pi^b)))/2$$

where both the writer and buyer seek to hedge the risk related to their position with the option using a portfolio of assets.

Namely,

- $\mathcal{S}$ keeps track of the asset values ξ and state of the MC
- $a_t \in [-1, 1]^m$ composes the portfolio
- $$r_t(s_t, a_t, s_{t+1}) := \begin{cases} a_t^\top(\xi_{t+1} - \xi_t) & t < T \\ F(\xi_t)(1 - 2 \cdot \mathbf{1}\{\text{agent=writer}\}) & t = T \end{cases}$$

Actor and critic network architectures

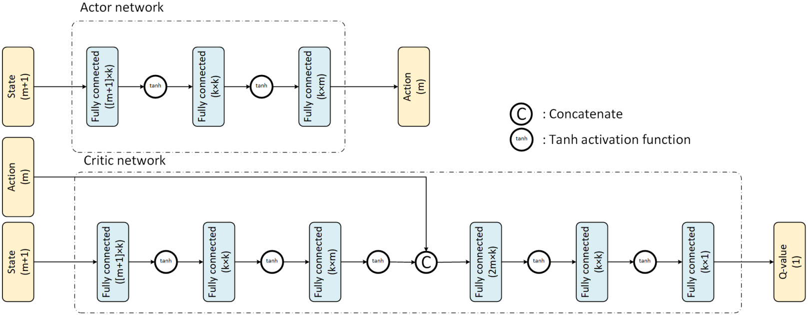
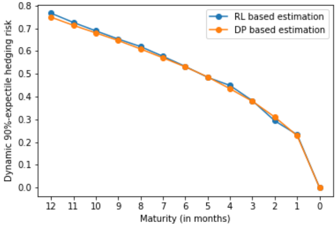
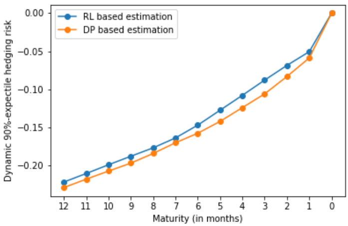


Figure: The architecture of the actor and critic networks in ACRL algorithm

Precision of the ACRL solution



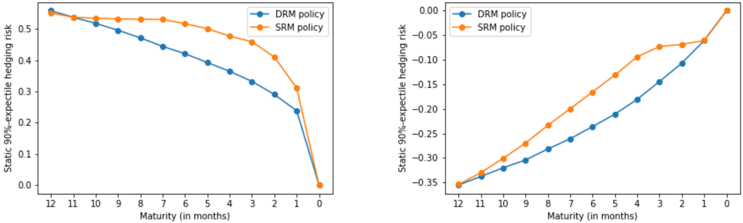
(a) Writer, $\tau = 90\%$



(b) Buyer, $\tau = 90\%$

Figure: The out-of-sample dynamic risk imposed to the two sides of a vanilla at-the-money call option over APPL (with maturity ranging from 12 months to 0 months) under the DERM policy trained for a 12 months maturity and at the risk level $\tau = 90\%$.

STATIC RISK EXPOSURE OF DIFFERENT APPROACHES



(a) Writer, $\tau = 90\%$ (b) Buyer, $\tau = 90\%$

Figure: The out-of-sample static risk imposed to the two sides of a vanilla at-the-money call option over APPL (with maturity ranging from 12 months to 0 months) under the DERM and Static Risk Measure (SRM) policies trained for a 12 months maturity and at the risk level $\tau = 90\%$.

Bibliography

Fabio Bellini and Valeria Bignozzi. On elicitable risk measures. Quantitative Finance, 15(5):725–733, 2015.

Anthony Coache, Sebastian Jaimungal, and Álvaro Cartea. Conditionally elicitable dynamic risk measures for deep reinforcement learning. arXiv preprint arXiv:2206.14666, 2022.

John C Cox, Stephen A Ross, and Mark Rubinstein. Option pricing: A simplified approach. Journal of financial Economics, 7(3):229–263, 1979.

Ivan Guo and Song-Ping Zhu. Equal risk pricing under convex trading constraints. Journal of Economic Dynamics and Control, 76:136–151, 2017.

Audrey Huang, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. On the convergence and optimality of policy gradient for Markov coherent risk. arXiv preprint arXiv:2103.02827, 2021.

Saeed Marzban, Erick Delage, and Jonathan Yu-Meng Li. Deep reinforcement learning for option pricing and hedging under dynamic expectile risk measures. accepted in Quantitative Finance, 2023.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc., 2015.