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

Erick Delage,
Department of Decision Sciences

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

May 29, 2023
Consider a finite horizon MDP \((S, \mathcal{A}, r, P)\). Given a policy \(\pi : S \times [T] \to \mathcal{A}\), 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 \(\pi_t\), 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):
   \[ \min_{\pi} \tilde{\rho}(-\tilde{R}(\pi)) := \tilde{\varrho}(F_{\tilde{R}(\pi)}) \]
   ➤ E.g.: \(-\mathbb{E}[\tilde{R}], -\mathbb{E}[u(\tilde{R})], \text{VaR}(-\tilde{R}), \text{CVaR}(-\tilde{R})\)

Cost distribution

- Mean = 2.72
- Median = 2.19
- Mode = 1.42
- Range = [0, 2, \infty]
- 95% VaR = 95th percentile = 7.5
- Conditional VaR 95% = 9
**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.: \(-\mathbb{E}[\tilde{R}], -\mathbb{E}[u(\tilde{R})], \text{VaR}(\tilde{R}), \text{CVaR}(\tilde{R})\)
   - Pros: Easy to interpret
   - Cons: Can violate dynamic consistency
   - Pro or Con?: Does not distinguish between two policies that have the same \(F_{\tilde{R}(\pi)}\)

![Cost distribution diagram](image)

- **Mean** = 2.72
- **Median** = 2.19
- **Mode** = 1.42
- **Range** = \([0, 2, \infty]\)
- **95% VaR** = 95th percentile = 7.5
- **Conditional VaR 95%** = 9
RISK AVERSION IN MULTISTAGE DECISION MAKING

Risk aversion can be handled using two approaches:

1. Static law-invariant risk measure (SRM):
   \[
   \min_{\pi} \bar{\rho}(-\tilde{R}(\pi)) := \bar{\rho}(F_{\tilde{R}(\pi)})
   \]

2. Dynamic law-invariant risk measure (DRM):
   \[
   \max_{\pi} \rho(-\tilde{R}(\pi)) := \bar{\rho}_0(\bar{\rho}_1(\ldots \bar{\rho}_{T-1}(-\tilde{R}(\pi)|a_{0:T-1}, s_{1:T}) \ldots |a_0, s_1))
   \]
   ▶ E.g.: \( E[-\tilde{R}], -E[u(\tilde{R})], \)
   \( \text{VaR}(\text{VaR}(\ldots \text{VaR}(-\tilde{R}|a_{0:T-1}, s_{1:T}) \ldots |a_0, s_1)), \)
   \( \text{CVaR}(\text{CVaR}(\ldots \text{CVaR}(-\tilde{R}|a_{0:T-1}, s_{1:T}) \ldots |a_0, s_1)) \)
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):
   \[
   \max_{\pi} \rho(-\tilde{R}(\pi)) := \\
   \bar{\rho}_0(\bar{\rho}_1(\ldots \bar{\rho}_{T-1}(-\tilde{R}(\pi)|\tilde{a}_{0:T-1}, \tilde{s}_{1:T}) \ldots |\tilde{a}_0, \tilde{s}_1))
   \]
   - E.g.: \(\mathbb{E}[-\tilde{R}], -\mathbb{E}[u(\tilde{R})],\)
   \(\text{VaR(VaR(\ldots \text{VaR}(\ldots \text{VaR}(-\tilde{R}|\tilde{a}_{0:T-1}, \tilde{s}_{1:T}) \ldots |\tilde{a}_0, \tilde{s}_1)\ldots |\tilde{a}_0, \tilde{s}_1)))\)
   - 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 Bigozzi, 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:

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

Definition 2

The $\tau$-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
- $\tau = 1 \Rightarrow \bar{\rho}(\tilde{X}) = \text{ess sup}[\tilde{X}]$, 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, 
\[ \min_{\pi} \bar{\rho}(-\tilde{R}(\pi)) \equiv \min_{\pi} V^\pi_0(s_0) \] 
where

\[ V^\pi_t(s_t) := \bar{\rho}_t(-r_t(s_t, \tilde{a}_t, \tilde{s}_{t+1}) + V^\pi_{t+1}(\tilde{s}_{t+1})|_{\tilde{s}_t = s_t}) \]

with \( \tilde{a}_t \sim \pi_t(\tilde{s}_t) \) and \( V^\pi_T(s_T) := 0. \)

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 \( Q^*_T(s_T, a_T) := 0. \)
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?

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

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

- Initialize the main actor $\theta^\pi$ and critic $\theta^Q$ 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:
      
      \[
      \theta^Q \leftarrow \theta^Q + \alpha \frac{1}{N} \sum_{i=1}^{N} \partial \ell(Q(s_i, a_i | \theta^Q) - y_i) \nabla_{\theta^Q} Q(s_i, a_i | \theta^Q)
      \]
      
      where $\ell(\Delta) := \Delta^2$
    - Update the main actor network:
      
      \[
      \theta^\pi \leftarrow \theta^\pi - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta^\pi} \pi(s_i' | \theta^\pi) \nabla_{\theta^\pi} \pi(s_i' | \theta^\pi)
      \]
    - 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.

\[
Q^*_t(s_t, a_t) := \bar{\rho}_t \left( - r_t(s_t, a_t, \tilde{s}_{t+1}) + \max_{a_{t+1}} Q^*_{t+1}(\tilde{s}_{t+1}, a_{t+1}) \right| s_t \right)
\]

**Algorithm** Risk averse DDPG (ACRL)

Initialize the main actor \( \theta^\pi \) and critic \( \theta^Q \) 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:

\[
\theta^Q \leftarrow \theta^Q + \alpha \frac{1}{N} \sum_{i=1}^{N} \partial \mathcal{L}(Q(s_i, a_i | \theta^Q) - y_i) \nabla_{\theta^Q} Q(s_i, a_i | \theta^Q)
\]

where \( \mathcal{L}(\Delta) := \Delta^2 \)

\( \mathcal{L}(\Delta) := (1 - \tau) \max(0, \Delta)^2 + \tau \max(0, -\Delta)^2 \)

Update the main actor network:

\[
\theta^\pi \leftarrow \theta^\pi - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q(s^i_j, a | \theta^Q) \big|_{a = \pi(s^i_j | \theta^\pi)} \nabla_{\theta^\pi} \pi(s^i_j | \theta^\pi)
\]

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:**

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

*Graphs are from:* [https://en.wikipedia.org/wiki/Call_option](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:

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

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

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

\[ \omega_1 : \xi^* uS + \zeta^* = F_u, \quad \omega_2 : \xi^* dS + \zeta^* = F_d \quad \Rightarrow \quad \xi^* = \frac{F_u - F_d}{(u-d)S}, \quad \zeta^* = \frac{uF_d - dF_u}{(u-d)}, \quad \Rightarrow \quad w_0 = \xi^* S + \zeta^* \]

Any other price leads to arbitrage.

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

16 / 22
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

\[
ERP(F) = \left( \min_{\pi^w} \rho(-\tilde{R}_F^w(\pi^w)) + \min_{\pi^b} \rho(-\tilde{R}_F^b(\pi^b)) \right) / 2
\]

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

Namely,

- $S$ keeps track of the asset values $\xi$ and state of the MC
- $a_t \in [-1, 1]^m$ composes the portfolio

\[
r_t(s_t, a_t, s_{t+1}) := \begin{cases} 
a_t^\top (\xi_{t+1} - \xi_t) & t < T \\
F(\xi_t)(1 - 2 \cdot 1\{\text{agent=writer}\}) & t = T \end{cases}
\]
**Actor and Critic Network Architectures**

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

- **Actor Network**:
  - State $(m+1)$
  - Fully connected $(m+1 \times k)$
  - Fully connected $(k \times k)$
  - Fully connected $(k \times m)$
  - Action $(m)$

- **Critic Network**:
  - State $(m+1)$
  - Fully connected $(m+1 \times k)$
  - Fully connected $(k \times k)$
  - Fully connected $(k \times m)$
  - Concatenate ($C$)
  - Tanh activation function
  - Fully connected $(2 \times k)$
  - Fully connected $(k \times k)$
  - Fully connected $(k \times k)$
  - Q-value $(1)$

- **Symbols**:
  - $C$: Concatenate
  - $\tanh$: Tanh activation function
**Precision of the ACRL Solution**

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 $\tau = 90\%$. 

(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 $\tau = 90\%$. 

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

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


