Deep Reinforcement Learning for Risk Averse Sequential Decision Making Problems

Erick Delage
Department of Decision Sciences
HEC MONTRÉAL

(joint work with Saeed Marzban, Jonathan Y. Li (U. of Ottawa), Jia Lin Hau, Marek Petrik (U. of New Hampshire), Mohammad Ghavamzadeh (Google Research))

June 4, 2023

GERAD IVADO 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 $\pi : \mathcal{S} \times [T] \to \mathcal{A}$, we are interested in the risk related to the sum of cumulative reward:

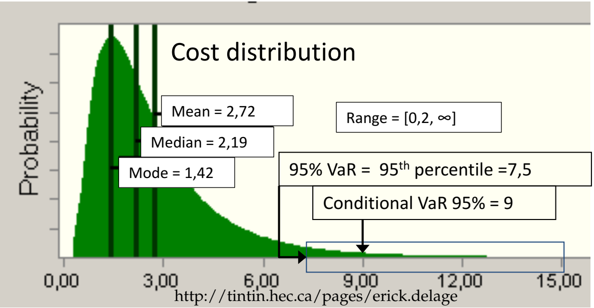
$$\tilde{R}(\pi) := \sum_{t=0}^{T-1} r_t(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})$$

where $\{\tilde{s}_t\}_{t=0}^{T}$ is the random state trajectory traversed when drawing actions from policy 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:

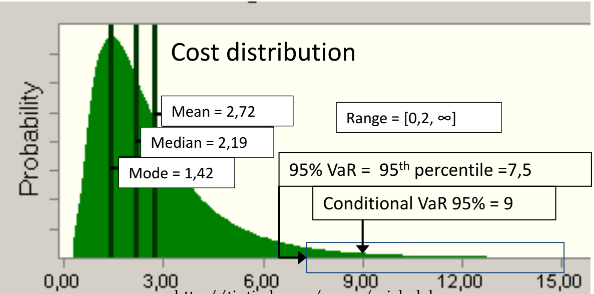
1. Static law-invariant risk measure (SRM):
   $\min_\pi \bar{\rho}(-\tilde{R}(\pi)) := \bar{\varrho}(F_{\tilde{R}(\pi)})$
   - E.g. : $-\mathbb{E}[\tilde{R}]$, $-\mathbb{E}[u(\tilde{R})]$, $\mathrm{VaR}(-\tilde{R})$, $\mathrm{CVaR}(-\tilde{R})$



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

DRL with Static Risk Measure

OUTLINE

Introduction

Deep RL for dynamic elicitable risk measure

DRL with Static Risk Measure

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:

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

    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:

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

    where $\cancel{\boldsymbol{\ell(\Delta) := \Delta^2}}$
$\boldsymbol{\ell(\Delta) := (1-\tau)\max(0, \Delta)^2 + \tau\max(0, -\Delta)^2}$
    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

DRL with Static Risk Measure

PRIMAL-DUAL REPRESENTATIONS OF SRMS

- Value-at-risk [Follmer and Schied, 2016]:

$$\begin{aligned}\mathrm{VaR}_\alpha\left(\tilde{X}\right) &= \inf\ \left\{z \in \mathbb{R} \mid \mathbb{P}(\tilde{X} > z) \leq \alpha\right\} \\ &= \sup\ \left\{z \in \mathbb{R} \mid \mathbb{P}(\tilde{X} \geq z) > \alpha\right\}.\end{aligned}$$

- Conditional Value-at-Risk:

$$\begin{aligned}\mathrm{CVaR}_\alpha\left(\tilde{X}\right) &= \inf_{z\in\mathbb{R}}\ \left(z + \alpha^{-1}\mathbb{E}\left[\tilde{X} - z\right]_+\right) \\ &= \sup_{\xi:\Omega\to\mathbb{R}}\ \left\{\mathbb{E}[\xi\tilde{X}]\middle|\mathbb{E}[\xi] = 1,\, \mathbb{P}(\alpha\xi \leq 1) = 1\right\},\end{aligned}$$

- Entropic Value-at-Risk [Ahmadi-Javid, 2012]:

$$\begin{aligned}\mathrm{EVaR}_\alpha\left(\tilde{X}\right) &= \inf_{\beta>0}\ \beta^{-1}\left(\log(\alpha^{-1}\mathbb{E}[\exp(\beta\tilde{X})])\right) \\ &= \sup_{\xi:\Omega\to\mathbb{R}}\ \left\{\mathbb{E}[\xi\tilde{X}]\middle|\mathbb{E}[\xi] = 1,\, \mathbb{E}[\xi\log(\xi)] \leq -\log(\alpha)\right\},\end{aligned}$$

DEEP RL FOR STATIC RISK MEASURES

- Filar et al. [1995], Wu and Lin [1999], Lin et al. [2003], Boda et al. [2004], Bäuerle and Ott [2011], Xu and Mannor [2011], Chow and Ghavamzadeh [2014], Hau et al. [2023b]:
  exploit the infimum representation of risk measures to define a risk neutral MDP on a lifted state-space, which keeps track of cumulated rewards.

DEEP RL FOR STATIC RISK MEASURES

- Filar et al. [1995], Wu and Lin [1999], Lin et al. [2003], Boda et al. [2004], Bäuerle and Ott [2011], Xu and Mannor [2011], Chow and Ghavamzadeh [2014], Hau et al. [2023b]:
  exploit the infimum representation of risk measures to define a risk neutral MDP on a lifted state-space, which keeps track of cumulated rewards.
- Chow et al. [2015], Chapman et al. [2019], Stanko and Macek [2019], Rigter et al. [2021], Ding and Feinberg [2022]:
  exploit the supremum representation of risk measures to define a robust MDP on a lifted state-space, which keeps track of current risk-level.

DEEP RL FOR STATIC RISK MEASURES

- Filar et al. [1995], Wu and Lin [1999], Lin et al. [2003], Boda et al. [2004], Bäuerle and Ott [2011], Xu and Mannor [2011], Chow and Ghavamzadeh [2014], Hau et al. [2023b]:
  exploit the infimum representation of risk measures to define a risk neutral MDP on a lifted state-space, which keeps track of cumulated rewards.
- Chow et al. [2015], Chapman et al. [2019], Stanko and Macek [2019], Rigter et al. [2021], Ding and Feinberg [2022]:
  exploit the supremum representation of risk measures to define a robust MDP on a lifted state-space, which keeps track of current risk-level.
- Hau et al. [2023a]:
  - Robust MDP approach is inexact in general!
  - For VaR, Li et al. [2022] needs corrections.

Primal DP decomposition for SRM

With Static CVaR MDP,

$$\min_{\pi} \bar{\rho}(-\tilde{R}(\pi)) \equiv \min_{z,\pi} z + \alpha^{-1}\mathbb{E}[(-\tilde{R}(\pi) - z)^+]$$

Primal DP decomposition for SRM

With Static CVaR MDP,

$$\min_{\pi} \bar{\rho}(-\tilde{R}(\pi)) \equiv \min_{z,\pi} z + \alpha^{-1}\mathbb{E}[(-\tilde{R}(\pi) - z)^{+}]$$

Hence,

$$\min_{\pi} \bar{\rho}(-\tilde{R}(\pi)) \equiv \min_{z,a_0} \; z + \alpha^{-1} Q_0^*(s_0, z, a_0)$$

where

$$Q_t^*(s_t, z_t, a_t) := \mathbb{E}[\min_{a_{t+1}} Q_{t+1}^*(\tilde{s}_{t+1}, z_t + r_t(s_t, a_t, \tilde{s}_{t+1}), a_{t+1}) \,|\, \tilde{s}_t = s_t\,]$$

and $Q_T^*(s_T, z_t, a_T) := \max(0, -z_t)$.

Dual DP decomposition for SRM

With Static CVaR MDP, Pflug and Pichler [2016]:

$$\bar{\rho}(-\tilde{R}(\pi)) = \sup_{\xi:\mathcal{A}^T\times\mathcal{S}^T\to\mathbb{R}} \left\{\mathbb{E}\left[-\xi\tilde{R}(\pi)\right]\middle|\mathbb{E}[\xi]=1,\,\mathbb{P}(\alpha\xi\leq 1)=1\right\}$$

$$= \sup_{\xi:\mathcal{A}\times\mathcal{S}\to\mathbb{R}} \left\{\mathbb{E}\left[\xi\,\mathrm{CVaR}_{\alpha\xi}\left(-\tilde{R}(\pi)\middle|\tilde{a}_0,\tilde{s}_1\right)\right]\middle|\mathbb{E}[\xi]=1,\,\mathbb{P}(\alpha\xi\leq 1)=1\right\}$$

Dual DP decomposition for SRM

With Static CVaR MDP, Pflug and Pichler [2016]:

$$\bar{\rho}(-\tilde{R}(\pi)) = \sup_{\xi:\mathcal{A}^T\times\mathcal{S}^T\to\mathbb{R}} \left\{\mathbb{E}\left[-\xi\tilde{R}(\pi)\right]\middle|\mathbb{E}[\xi]=1,\ \mathbb{P}(\alpha\xi\leq 1)=1\right\}$$

$$= \sup_{\xi:\mathcal{A}\times\mathcal{S}\to\mathbb{R}} \left\{\mathbb{E}\left[\xi\,\mathrm{CVaR}_{\alpha\xi}\left(-\tilde{R}(\pi)\middle|\tilde{a}_0,\tilde{s}_1\right)\right]\middle|\mathbb{E}[\xi]=1,\ \mathbb{P}(\alpha\xi\leq 1)=1\right\}$$

Hence,

$$\min_\pi \bar{\rho}(-\tilde{R}(\pi)) \equiv \min_\pi V_0^\pi(s_0,\alpha)$$

where $V_t^\pi(s_t,\alpha_t) :=$

$$\sup_{\xi:\mathcal{A}\times\mathcal{S}\to\mathbb{R}} \{\mathbb{E}[\xi(-r_t(s_t,\tilde{a}_t,\tilde{s}_{t+1})+V_{t+1}^\pi(\tilde{s}_{t+1},\alpha_t\xi))|\tilde{s}_t=s_t]$$

$$|\mathbb{E}[\xi]=1,\ \mathbb{P}(\alpha_t\xi\leq 1)=1\}$$

DUAL DP DECOMPOSITION FOR SRM II

Chow et al. [2015] claims,

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

where $Q_t^*(s_t, \alpha_t, a_t) :=$

$$\sup_{\xi:\mathcal{S}\to\mathbb{R}} \left\{ \mathbb{E}[\xi(-r_t(s_t, a_t, \tilde{s}_{t+1}) + \min_{a_{t+1}} Q_{t+1}^*(\tilde{s}_{t+1}, \alpha_t\xi, a_{t+1}))|\tilde{s}_t = s_t] \middle| \mathbb{E}[\xi] = 1,\, \mathbb{P}(\alpha_t\xi \leq 1) = 1 \right\}$$

Dual DP decomposition for SRM

In fact, Hau et al. [2023a] shows:

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

where $Q_t^*(s_t, \alpha_t, a_t) :=$

$$\sup_{\xi:\mathcal{S}\to\mathbb{R}} \left\{ \mathbb{E}[\xi(-r_t(s_t, a_t, \tilde{s}_{t+1}) + \min_{a_{t+1}} Q_{t+1}^*(\tilde{s}_{t+1}, \alpha_t \xi, a_{t+1})) | \tilde{s}_t = s_t] \right.$$
$$\left. \Big| \mathbb{E}[\xi] = 1,\ \mathbb{P}(\alpha_t \xi \leq 1) = 1 \right\}$$

New DP decomposition for Static VaR

Inspired by Li et al. [2022], we derive a new decomposition for Static VaR:

$$\bar{\rho}(-\tilde{R}(\pi))$$

$$= \inf_{\xi:\mathcal{A}^T\times\mathcal{S}^T\to\mathbb{R}} \left\{ \operatorname{ess\,sup} \left[-\tilde{R}(\pi) \middle| \xi < 1 \right] \middle| \mathbb{E}[\xi] = 1,\ \mathbb{P}(\alpha\xi \leq 1) = 1 \right\}$$

$$= \inf_{\xi:\mathcal{A}\times\mathcal{S}\to\mathbb{R}} \left\{ \operatorname{ess\,sup} \left[\mathrm{VaR}_{\alpha\xi}\left(-\tilde{R}(\pi) | \tilde{a}_0, \tilde{s}_1 \right) \middle| \xi < 1 \right] \middle| \mathbb{E}[\xi] = 1,\ \mathbb{P}(\alpha\xi \leq 1) = 1 \right\}$$

We also show that,

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

where $Q_t^*(s_t, \alpha_t, a_t) := \inf_{\xi:\mathcal{S}\to\mathbb{R}} \{$

$$\operatorname{ess\,sup}[-r_t(s_t, a_t, \tilde{s}_{t+1}) + \min_{a_{t+1}} Q_{t+1}^*(\tilde{s}_{t+1}, \alpha_t\xi, a_{t+1}))\tilde{s}_t = s_t | \xi < 1]$$

$$|\mathbb{E}[\xi] = 1,\ \mathbb{P}(\alpha_t\xi \leq 1) = 1\}$$

Questions & Comments ...

... Thank you!

BIBLIOGRAPHY

A. Ahmadi-Javid. Entropic Value-at-Risk: A new coherent risk measure. Journal of Optimization Theory and Applications, 155(3):1105–1123, 2012.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. Mathematical Methods of Operations Research, 74(3): 361–379, 2011.

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

Kang Boda, Jerzy A Filar, Yuanlie Lin, and Lieneke Spanjers. Stochastic target hitting time and the problem of early retirement. IEEE Transactions on Automatic Control, 49(3):409–419, 2004.

Margaret P Chapman, Jonathan Lacotte, Aviv Tamar, Donggun Lee, Kevin M Smith, Victoria Cheng, Jaime F Fisac, Susmit Jha, Marco Pavone, and Claire J Tomlin. A risk-sensitive finite-time reachability approach for safety of stochastic dynamic systems. In 2019 American Control Conference (ACC), pages 2958–2963. IEEE, 2019.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In Neural Information Processing Systems (NIPS), 2014.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: A CVaR optimization approach. In Neural Information Processing Systems (NIPS), 2015.

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

Rui Ding and Eugene Feinberg. CVaR optimization for MDPs: Existence and computation of optimal policies. ACM SIGMETRICS Performance Evaluation