An Introduction to Risk-Aware RL with Dynamic Risk Measures

Anthony Coache (University of Toronto)

anthonycoache.ca

Joint work with

Sebastian Jaimungal (University of Toronto & Oxford-Man Institute)

and

Álvaro Cartea (Oxford-Man Institute & University of Oxford)

Graduate Student Research Day ⋆ April 27, 2023 ⋆ University of Toronto

Table of contents

Reinforcement Learning (RL)

Subfield of machine learning

- Model-agnostic framework for learning-based control
- Learning optimal behaviors from interactions to minimize a cost signal
- Classic trade-off between exploration and exploitation

Applications of interest:

- Portfolio allocation
- Hedging and pricing financial instruments
- Robot control
- Board games and video games
- etc.

Reinforcement Learning (RL)

Subfield of machine learning

- Model-agnostic framework for learning-based control
- Learning optimal behaviors from interactions to minimize a cost signal
- Classic trade-off between exploration and exploitation

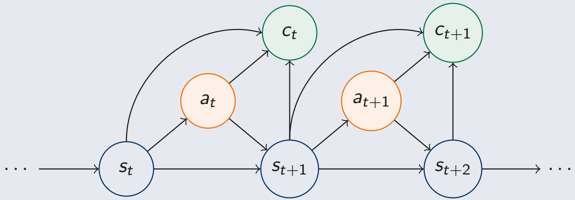
Applications of interest:

- Portfolio allocation
- Hedging and pricing financial instruments
- Robot control
- Board games and video games
- etc.

RL Notation

Markov Decision Process $(\mathcal{S}, \mathcal{A}, \pi, \mathbb{P}, c)$

- $\mathcal{S}$ – State space
- $\mathcal{A}$ – Action space
- $\pi^\theta(a_t|s_t)$ – Randomized policy characterized by θ
- $\mathbb{P}(s_0), \mathbb{P}(s_{t+1}|s_t, a_t)$ – Transition probability distribution
- $c_t(s_t, a_t, s_{t+1}) \in \mathcal{C}$ – Cost function



Risk-Aware RL

RL aim at minimizing problems of the form

$$\min_{\theta} J\Big(\{c_t^{\theta}\}_t \,\Big|\, s_0 = s\Big).$$

- Standard RL: $J(Y) = \mathbb{E}[Y]$, where $Y = \sum_t c_t^{\theta}$

Risk-aware RL uses risk measures as the criterion, e.g.

- Expected utility [Nass et al., 2019]: $\min_{\theta} \mathbb{E}[U(Y)]$
- Risk-constrained [Di Castro et al., 2019]: $\min_{\theta} \mathbb{E}[Y]$ subj. to. $\rho(Y) \leq c^*$
- Coherent risk measure [Tamar et al., 2016] such as expected-shortfall:

$$\min_{\theta} \frac{1}{1-\alpha} \int_{[\alpha,1]} \mathrm{VaR}_u(Y)\mathrm{d}u$$

Risk-Aware RL

RL aim at minimizing problems of the form

$$\min_{\theta} J\left(\{c_t^{\theta}\}_t \,\middle|\, s_0 = s\right).$$

- Standard RL: $J(Y) = \mathbb{E}[Y]$, where $Y = \sum_t c_t^{\theta}$

Risk-aware RL uses risk measures as the criterion, e.g.

- Expected utility [Nass et al., 2019]: $\min_{\theta} \mathbb{E}[U(Y)]$
- Risk-constrained [Di Castro et al., 2019]: $\min_{\theta} \mathbb{E}[Y]$ subj. to. $\rho(Y) \leq c^*$
- Coherent risk measure [Tamar et al., 2016] such as expected-shortfall:

$$\min_{\theta} \frac{1}{1-\alpha} \int_{[\alpha,1]} \mathrm{VaR}_u(Y) \mathrm{d}u$$

Examples of (Static) Risk Measures

Consider $\mathcal{Y} := \mathcal{L}_p(\Omega, \mathcal{F}, P)$ – p-integrable, $\mathcal{F}$-measurable random variables

Convex risk measure $\rho : \mathcal{Y} \to \mathbb{R}$ [Föllmer and Schied, 2002]

- *monotone:* $Y_1 \leq Y_2$ implies $\rho(Y_1) \leq \rho(Y_2)$
- *translation invariant:* $\rho(Y + m) = \rho(Y) + m, \ \forall m \in \mathbb{R}$
- *convex:* $\rho(\lambda Y_1 + (1-\lambda)Y_2) \leq \lambda\rho(Y_1) + (1-\lambda)\rho(Y_2)$

Spectral risk measure $\rho^\varphi : \mathcal{Y} \to \mathbb{R}$ [Kusuoka, 2001]

$$\rho^\varphi(Y) = \int_{[0,1]} \mathrm{CVaR}_\alpha(Y)\varphi(\mathrm{d}\alpha) \quad \text{with} \quad \mathrm{CVaR}_\alpha(Y) = \frac{1}{1-\alpha}\int_{[\alpha,1]} \mathrm{VaR}_u(Y)\mathrm{d}u,$$

where φ is a nonnegative, nonincreasing measure such that $\int_{[0,1]} \varphi(\mathrm{d}\alpha) = 1$.

Examples of (Static) Risk Measures

Consider $\mathcal{Y} := \mathcal{L}_p(\Omega, \mathcal{F}, P)$ – p-integrable, $\mathcal{F}$-measurable random variables

Convex risk measure $\rho : \mathcal{Y} \to \mathbb{R}$ [Föllmer and Schied, 2002]

- *monotone:* $Y_1 \leq Y_2$ implies $\rho(Y_1) \leq \rho(Y_2)$
- *translation invariant:* $\rho(Y + m) = \rho(Y) + m, \ \forall m \in \mathbb{R}$
- *convex:* $\rho(\lambda Y_1 + (1-\lambda)Y_2) \leq \lambda\rho(Y_1) + (1-\lambda)\rho(Y_2)$

Spectral risk measure $\rho^\varphi : \mathcal{Y} \to \mathbb{R}$ [Kusuoka, 2001]

$$\rho^\varphi(Y) = \int_{[0,1]} \mathrm{CVaR}_\alpha(Y)\varphi(\mathrm{d}\alpha) \quad \text{with} \quad \mathrm{CVaR}_\alpha(Y) = \frac{1}{1-\alpha}\int_{[\alpha,1]} \mathrm{VaR}_u(Y)\mathrm{d}u,$$

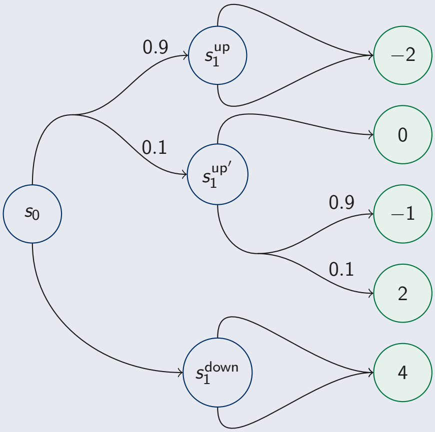
where φ is a nonnegative, nonincreasing measure such that $\int_{[0,1]} \varphi(\mathrm{d}\alpha) = 1$.

Time-Consistency Issue

Let us minimize $\mathrm{CVaR}_{0.9}$ of the terminal cost.

- *Optimal actions at s_0:* Move up, then down
- *Optimal actions at $s_1^{up'}$:* Move up

Contradiction with the state/time dependent strategy...

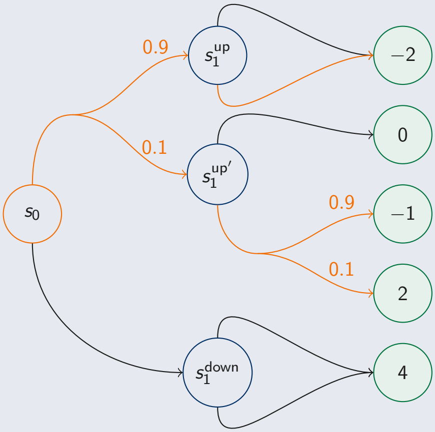


Time-Consistency Issue

Let us minimize $\mathrm{CVaR}_{0.9}$ of the terminal cost.

- *Optimal actions at s_0*: Move up, then down
- *Optimal actions at $s_1^{up'}$*: Move up

Contradiction with the state/time dependent strategy...

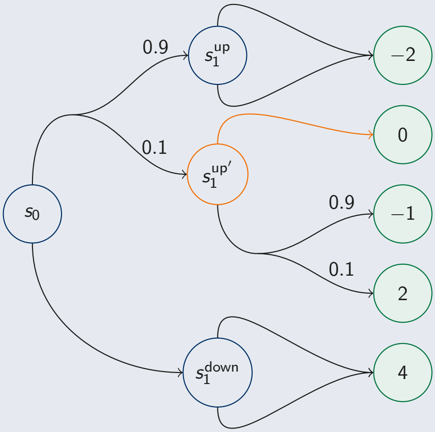


Time-Consistency Issue

Let us minimize $\text{CVaR}_{0.9}$ of the terminal cost.

- *Optimal actions at s_0*: Move up, then down
- *Optimal actions at $s_1^{up'}$*: Move up

Contradiction with the state/time dependent strategy...



Risk-Aware RL with Dynamic Risk

Optimizing static risk measures leads to optimal precommitment policies

Recent approaches to overcome this issue:

- DP equations for Kusuoka-type *conditional risk mappings* with latent costs and random actions [Cheng and Jaimungal, 2022]
- Bayesian approach to account for model uncertainty with *recursive risk filters* and unobserved costs [Bielecki et al., 2022]
- Policy iteration algorithms for *recursive coherent risk measures* [Bäuerle and Glauner, 2022]
- Deep Q-learning algorithm for *dynamic expectile risk measures* [Marzban et al., 2021]
- etc.

Risk-Aware RL with Dynamic Risk

Optimizing static risk measures leads to optimal precommitment policies

Recent approaches to overcome this issue:

- DP equations for Kusuoka-type *conditional risk mappings* with latent costs and random actions [Cheng and Jaimungal, 2022]
- Bayesian approach to account for model uncertainty with *recursive risk filters* and unobserved costs [Bielecki et al., 2022]
- Policy iteration algorithms for *recursive coherent risk measures* [Bäuerle and Glauner, 2022]
- Deep Q-learning algorithm for *dynamic expectile risk measures* [Marzban et al., 2021]
- etc.

Dynamic Risk Measures

Consider

- $\mathcal{T} := \{0, \dots, T\}$
- $\mathcal{F}_0 \subseteq \cdots \subseteq \mathcal{F}_T$ – Filtration on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \in \mathcal{T}}, \mathbb{P})$
- $\mathcal{Y}_t := \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ – p-integrable, $\mathcal{F}_t$-measurable random variables
- $\mathcal{Y}_{t_1,t_2} := \mathcal{Y}_{t_1} \times \cdots \times \mathcal{Y}_{t_2}$ – Sequence of random variables

Dynamic risk measure $\{\rho_{t,T}\}_{t \in \mathcal{T}}$

Sequence of conditional risk measures $\rho_{t,T} : \mathcal{Y}_{t,T} \to \mathcal{Y}_t$ where

$$\rho_{t,T}(Y) \leq \rho_{t,T}(Z), \text{ for all } Y, Z \in \mathcal{Y}_{t,T} \text{ such that } Y \leq Z$$

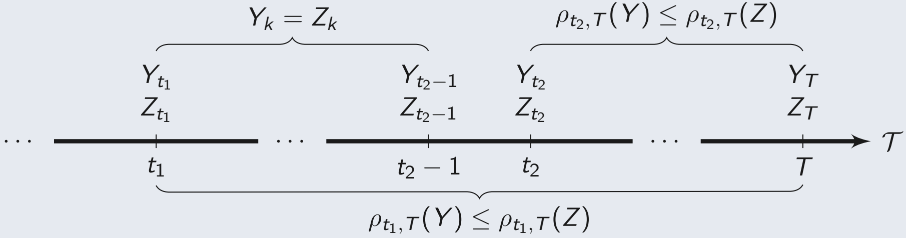
Time-Consistency

Strong time-consistency

$\{\rho_{t,T}\}_t$ is *strongly time-consistent* iff for any $Y, Z \in \mathcal{Y}_{t_1,T}$ and $0 \leq t_1 < t_2 \leq T$, we have

$$Y_k = Z_k,\ \forall k = t_1, \ldots, t_2 - 1 \ \text{ and } \ \rho_{t_2,T}(Y_{t_2}, \ldots, Y_T) \leq \rho_{t_2,T}(Z_{t_2}, \ldots, Z_T)$$

implies that $\rho_{t_1,T}(Y_{t_1}, \ldots, Y_T) \leq \rho_{t_1,T}(Z_{t_1}, \ldots, Z_T)$.



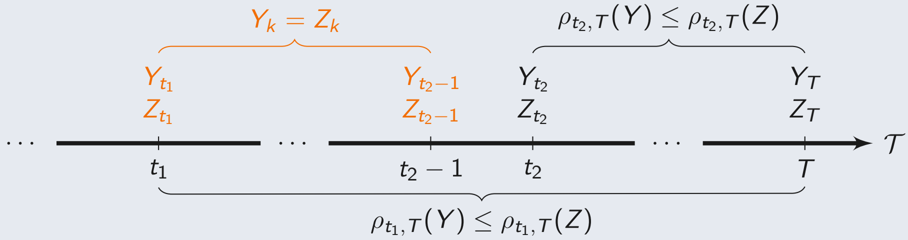
Time-Consistency

Strong time-consistency

$\{\rho_{t,T}\}_t$ is *strongly time-consistent* iff for any $Y, Z \in \mathcal{Y}_{t_1,T}$ and $0 \leq t_1 < t_2 \leq T$, we have

$$Y_k = Z_k,\ \forall k = t_1, \ldots, t_2 - 1 \quad \text{and} \quad \rho_{t_2,T}(Y_{t_2}, \ldots, Y_T) \leq \rho_{t_2,T}(Z_{t_2}, \ldots, Z_T)$$

implies that $\rho_{t_1,T}(Y_{t_1}, \ldots, Y_T) \leq \rho_{t_1,T}(Z_{t_1}, \ldots, Z_T)$.



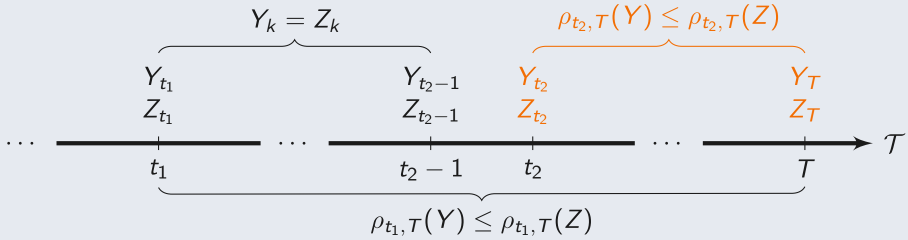
Time-Consistency

Strong time-consistency

$\{\rho_{t,T}\}_t$ is *strongly time-consistent* iff for any $Y, Z \in \mathcal{Y}_{t_1,T}$ and $0 \le t_1 < t_2 \le T$, we have

$$Y_k = Z_k,\ \forall k = t_1, \dots, t_2 - 1 \quad \text{and} \quad \rho_{t_2,T}(Y_{t_2}, \dots, Y_T) \le \rho_{t_2,T}(Z_{t_2}, \dots, Z_T)$$

implies that $\rho_{t_1,T}(Y_{t_1}, \dots, Y_T) \le \rho_{t_1,T}(Z_{t_1}, \dots, Z_T)$.



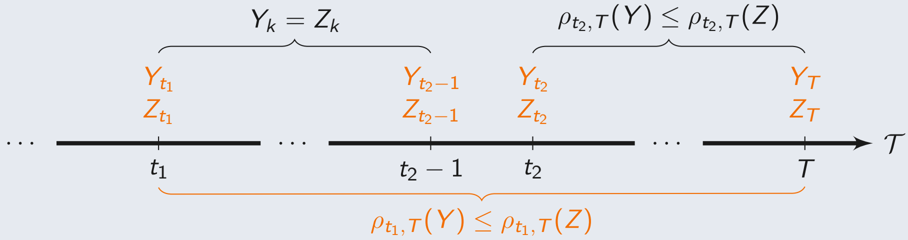
Time-Consistency

Strong time-consistency

$\{\rho_{t,T}\}_t$ is *strongly time-consistent* iff for any $Y, Z \in \mathcal{Y}_{t_1,T}$ and $0 \leq t_1 < t_2 \leq T$, we have

$$Y_k = Z_k,\ \forall k = t_1, \dots, t_2 - 1 \ \text{ and } \ \rho_{t_2,T}(Y_{t_2}, \dots, Y_T) \leq \rho_{t_2,T}(Z_{t_2}, \dots, Z_T)$$

implies that $\rho_{t_1,T}(Y_{t_1}, \dots, Y_T) \leq \rho_{t_1,T}(Z_{t_1}, \dots, Z_T)$.



Time-Consistency

[Thm. 1, Ruszczyński, 2010]
Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic risk measure satisfying for any $Y \in \mathcal{Y}_{t,T},\ t \in \mathcal{T}$

- $\rho_{t,T}(Y_t, Y_{t+1}, \dots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \dots, Y_T)$;
- $\rho_{t,T}(0, \dots, 0) = 0$;
- $\rho_{t_1,t_2}(\mathbb{1}_A Y) = \mathbb{1}_A \rho_{t_1,t_2}(Y)$ for any $A \in \mathcal{F}_{t_1}$.

Then $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ is time-consistent iff for any $0 \le t_1 \le t_2 \le T$ and $Y \in \mathcal{Y}_{0,T}$, we have

$$\rho_{t_1,T}(Y_{t_1}, \dots, Y_T) = \rho_{t_1,t_2}\Big(Y_{t_1}, \dots, Y_{t_2-1}, \rho_{t_2,T}(Y_{t_2}, \dots, Y_T)\Big)$$

Time-Consistency

[Thm. 1, Ruszczyński, 2010]
Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic risk measure satisfying for any $Y \in \mathcal{Y}_{t,T},\ t \in \mathcal{T}$

- $\rho_{t,T}(Y_t, Y_{t+1}, \ldots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \ldots, Y_T)$;
- $\rho_{t,T}(0, \ldots, 0) = 0$;
- $\rho_{t_1,t_2}(\mathbb{1}_A Y) = \mathbb{1}_A \rho_{t_1,t_2}(Y)$ for any $A \in \mathcal{F}_{t_1}$.

Then $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ is time-consistent iff for any $0 \le t_1 \le t_2 \le T$ and $Y \in \mathcal{Y}_{0,T}$, we have

$$\rho_{t_1,T}(Y_{t_1}, \ldots, Y_T) = \rho_{t_1,t_2}\Big(Y_{t_1}, \ldots, Y_{t_2-1}, \rho_{t_2,T}(Y_{t_2}, \ldots, Y_T)\Big)$$

Time-Consistency

[Thm. 1, Ruszczyński, 2010]
Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic risk measure satisfying for any $Y \in \mathcal{Y}_{t,T},\ t \in \mathcal{T}$

- $\rho_{t,T}(Y_t, Y_{t+1}, \dots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \dots, Y_T)$;
- $\rho_{t,T}(0, \dots, 0) = 0$;
- $\rho_{t_1,t_2}(\mathbb{1}_A Y) = \mathbb{1}_A \rho_{t_1,t_2}(Y)$ for any $A \in \mathcal{F}_{t_1}$.

Then $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ is time-consistent iff for any $0 \le t_1 \le t_2 \le T$ and $Y \in \mathcal{Y}_{0,T}$, we have

$$\rho_{t_1,T}(Y_{t_1}, \dots, Y_T) = \rho_{t_1,t_2}\Big(Y_{t_1}, \dots, Y_{t_2-1}, \rho_{t_2,T}(Y_{t_2}, \dots, Y_T)\Big)$$

Time-Consistency

Recursive relationship for time-consistent dynamic risk

Let *one-step conditional risk measures* $\rho_t : \mathcal{Y}_{t+1} \to \mathcal{Y}_t$ satisfy $\rho_t(Y) = \rho_{t,t+1}(0, Y)$ for any $Y \in \mathcal{Y}_{t+1}$. Then

$$\rho_{t,T}(Y_t, \ldots, Y_T) = Y_t + \rho_t\Big(Y_{t+1} + \rho_{t+1}\big(Y_{t+2} + \cdots + \rho_{T-1}(Y_T) \cdots\big)\Big).$$

Additional assumed properties for ρ_t:

- Either axioms of convex risk, coherent risk, form of spectral risk, etc.

Problem Setup

Problems of the form

$$\min_{\theta} \rho_{0,T}\left(\{c_t^\theta\}_{t\in\mathcal{T}}\right) = \min_{\theta} \rho_0\left(c_0^\theta + \rho_1\left(c_1^\theta + \cdots + \rho_{T-1}\left(c_{T-1}^\theta + \rho_T\left(c_T^\theta\right)\right)\cdots\right)\right)$$

where $c_t^\theta := c(s_t^\theta, a_t^\theta, s_{t+1}^\theta)$ are $\mathcal{F}_{t+1}$-measurable random costs.

DP equations for the *value function*, i.e. running risk-to-go, for $s \in \mathcal{S}$:

$$V_t(s;\theta) = \rho_t\left(\underbrace{c_t^\theta}_{\text{current cost}} + \underbrace{V_{t+1}(s_{t+1}^\theta;\theta)}_{\text{one-step ahead risk-to-go}} \;\middle|\; s_t = s\right),$$

under transition probabilities $\mathbb{P}^\theta(a, s'|s_t = s) = \mathbb{P}(s'|s,a)\pi^\theta(a|s_t = s)$

Problem Setup

Problems of the form

$$\min_{\theta} \rho_{0,T}\Big(\{c_t^\theta\}_{t\in\mathcal{T}}\Big) = \min_{\theta} \rho_0\Bigg(c_0^\theta + \rho_1\bigg(c_1^\theta + \cdots + \rho_{T-1}\Big(c_{T-1}^\theta + \rho_T\big(c_T^\theta\big)\Big)\cdots\bigg)\Bigg)$$

where $c_t^\theta := c(s_t^\theta, a_t^\theta, s_{t+1}^\theta)$ are $\mathcal{F}_{t+1}$-measurable random costs.

DP equations for the *value function*, i.e. running risk-to-go, for $s \in \mathcal{S}$:

$$V_t(s;\theta) = \rho_t\Bigg(\underbrace{c_t^\theta}_{\text{current cost}} + \underbrace{V_{t+1}(s_{t+1}^\theta;\theta)}_{\text{one-step ahead risk-to-go}} \;\Bigg|\; s_t = s\Bigg),$$

under transition probabilities $\mathbb{P}^\theta(a, s'|s_t = s) = \mathbb{P}(s'|s,a)\pi^\theta(a|s_t = s)$

Policy Gradient

- We wish to optimize the value function over policies θ via a policy gradient method:

$$\theta \leftarrow \theta - \eta \nabla_\theta V(\cdot;\theta)$$

[Gradient of V, Coache and Jaimungal, 2021]

Under some assumptions on the form of the risk envelope, the gradient of the value function at any period $t \in \mathcal{T}$ and any state $s \in \mathcal{S}$ for dynamic convex risk measures is

$$\nabla_\theta V_t(s;\theta) = \mathbb{E}_t^{\xi^*}\left[\left(c(s, a_t^\theta, s_{t+1}^\theta) + V_{t+1}(s_{t+1}^\theta;\theta) - \lambda^*\right)\nabla_\theta \log \pi^\theta(a_t^\theta|s) + \nabla_\theta V_{t+1}(s_{t+1}^\theta;\theta)\right] - \nabla_\theta \rho_t^*(\xi^*)$$

Actor-critic style algorithm composed of two interleaved procedures:

- *Critic* calculates the value function given a policy
- *Actor* updates the policy given a value function
- We parametrize policy and value function by ANNs

Policy Gradient

- We wish to optimize the value function over policies θ via a policy gradient method:

$$\theta \leftarrow \theta - \eta \nabla_\theta V(\cdot; \theta)$$

[Gradient of V, Coache and Jaimungal, 2021]

Under some assumptions on the form of the risk envelope, the gradient of the value function at any period $t \in \mathcal{T}$ and any state $s \in \mathcal{S}$ for dynamic convex risk measures is

$$\nabla_\theta V_t(s;\theta) = \mathbb{E}_t^{\xi^*}\left[\left(c(s, a_t^\theta, s_{t+1}^\theta) + V_{t+1}(s_{t+1}^\theta;\theta) - \lambda^* \right) \nabla_\theta \log \pi^\theta(a_t^\theta | s) + \nabla_\theta V_{t+1}(s_{t+1}^\theta;\theta) \right] - \nabla_\theta \rho_t^*(\xi^*)$$

Actor-critic style algorithm composed of two interleaved procedures:

- *Critic* calculates the value function given a policy
- *Actor* updates the policy given a value function
- We parametrize policy and value function by ANNs

Policy Gradient

- We wish to optimize the value function over policies θ via a policy gradient method:

$$\theta \leftarrow \theta - \eta \nabla_\theta V(\cdot;\theta)$$

[Gradient of V, Coache and Jaimungal, 2021]

Under some assumptions on the form of the risk envelope, the gradient of the value function at any period $t \in \mathcal{T}$ and any state $s \in \mathcal{S}$ for dynamic convex risk measures is

$$\nabla_\theta V_t(s;\theta) = \mathbb{E}_t^{\xi^*}\left[\left(c(s, a_t^\theta, s_{t+1}^\theta) + V_{t+1}(s_{t+1}^\theta;\theta) - \lambda^* \right) \nabla_\theta \log \pi^\theta(a_t^\theta|s) + \nabla_\theta V_{t+1}(s_{t+1}^\theta;\theta) \right] - \nabla_\theta \rho_t^*(\xi^*)$$

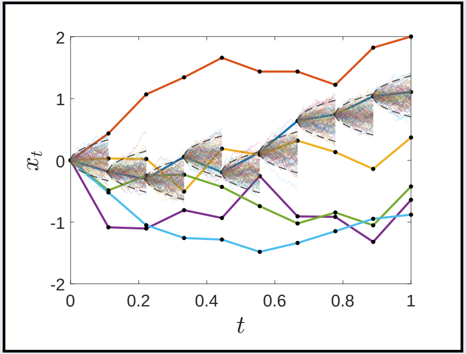
Actor-critic style algorithm composed of two interleaved procedures:

- *Critic* calculates the value function given a policy
- *Actor* updates the policy given a value function
- We parametrize policy and value function by ANNs

Estimation of V

Nested simulation approach
[Coache and Jaimungal, 2021]

- Generate (outer) trajectories and (inner) transitions for every visited state
- Class of *dynamic convex risk measures*
- Computationally expensive



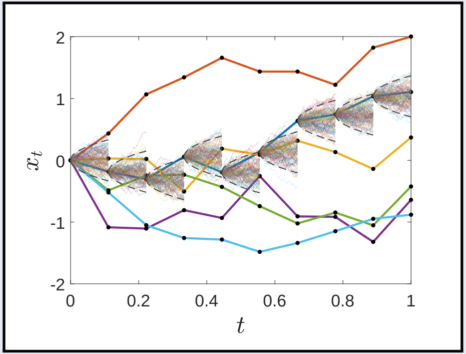
Elicitable approach [Coache et al., 2022]

- *Conditional elicitability* of dynamic spectral risk measures
- Avoids nested simulations, *memory efficient*

Estimation of V

Nested simulation approach [Coache and Jaimungal, 2021]

- Generate (outer) trajectories and (inner) transitions for every visited state
- Class of *dynamic convex risk measures*
- Computationally expensive



Elicitable approach [Coache et al., 2022]

- *Conditional elicitability* of dynamic spectral risk measures
- Avoids nested simulations, *memory efficient*

Elicitability

Elicitable risk measure [Gneiting, 2011]

ρ is elicitable iff there exists a scoring function $S : \mathbb{R} \times \mathbb{Y} \to \mathbb{R}$ s.t.

$$\rho(Y) = \arg\min_{\mathfrak{a} \in \mathbb{R}} \mathbb{E}_{Y \sim F_Y} \left[S(\mathfrak{a}, Y) \right].$$

$\rho(Y)$	Mean	Median	VaR_α	CVaR_α
$S(\mathfrak{a}, y)$	$(\mathfrak{a} - y)^2$	$\lvert\mathfrak{a} - y\rvert$	$\mathbb{1}_{\mathfrak{a} \leq y} - \alpha$	$\emptyset$

Elicitability

Elicitable risk measure [Gneiting, 2011]

ρ is elicitable iff there exists a scoring function $S : \mathbb{R} \times \mathbb{Y} \to \mathbb{R}$ s.t.

$$\rho(Y) = \arg\min_{\mathfrak{a} \in \mathbb{R}} \mathbb{E}_{Y \sim F_Y}\Big[S(\mathfrak{a}, Y)\Big].$$

$\rho(Y)$	Mean	Median	VaR_α	CVaR_α
$S(\mathfrak{a}, y)$	$(\mathfrak{a} - y)^2$	$\lvert\mathfrak{a} - y\rvert$	$\mathbb{1}_{\mathfrak{a} \leq y} - \alpha$	$\emptyset$

Conditional Elicitability

Non-elicitable mappings can be components of an elicitable vector-valued mapping:

- Spectral risk measures are not elicitable
- [Fissler and Ziegel, 2016] proved that a class of spectral risk measures is conditionally elicitable
- they characterized the scoring function S

Example (CVaR_α): the pair $(\mathrm{VaR}_\alpha(Y), \mathrm{CVaR}_\alpha(Y))$ is elicitable, i.e.

$$\Big(\mathrm{VaR}_\alpha(Y), \mathrm{CVaR}_\alpha(Y)\Big) = \underset{(a_1,a_2)\in\mathbb{R}^2}{\arg\min}\ \mathbb{E}_{Y\sim F_Y}\Big[S(a_1, a_2, Y)\Big]$$

where

$$S(a_1, a, y) = \Big(\mathbb{1}_{y\le a_1} - \alpha\Big)\Big(G_1(a_1) - G_1(y)\Big) - G_2(a_2) + G_2(y) \\ + G_2'(a_2)\left[a_2 + \frac{1}{1-\alpha}\Big(a_1\Big(\mathbb{1}_{y>a_1} - (1-\alpha)\Big) - y\,\mathbb{1}_{y>a_1}\Big)\right].$$

Conditional Elicitability

Non-elicitable mappings can be components of an elicitable vector-valued mapping:

- Spectral risk measures are not elicitable
- [Fissler and Ziegel, 2016] proved that a class of spectral risk measures is conditionally elicitable
- they characterized the scoring function S

Example (CVaR_α): the pair $(\mathrm{VaR}_\alpha(Y), \mathrm{CVaR}_\alpha(Y))$ is elicitable, i.e.

$$\Big(\mathrm{VaR}_\alpha(Y), \mathrm{CVaR}_\alpha(Y)\Big) = \underset{(\mathfrak{a}_1,\mathfrak{a}_2)\in\mathbb{R}^2}{\arg\min}\ \mathbb{E}_{Y\sim F_Y}\Big[S(\mathfrak{a}_1, \mathfrak{a}_2, Y)\Big]$$

where

$$\begin{aligned} S(\mathfrak{a}_1, \mathfrak{a}, y) = {} & \Big(\mathbb{1}_{y\le \mathfrak{a}_1} - \alpha\Big)\Big(G_1(\mathfrak{a}_1) - G_1(y)\Big) - G_2(\mathfrak{a}_2) + G_2(y) \\ & + G_2'(\mathfrak{a}_2)\left[\mathfrak{a}_2 + \frac{1}{1-\alpha}\Big(\mathfrak{a}_1\Big(\mathbb{1}_{y>\mathfrak{a}_1} - (1-\alpha)\Big) - y\,\mathbb{1}_{y>\mathfrak{a}_1}\Big)\right]. \end{aligned}$$

Conditional Elicitability

Example (CVaR$_\alpha$, cont'd): In our RL problems, the random variable Y (i.e. costs) are supported by observed features $s \in \mathcal{S}$ (i.e. states)

$$\rho(Y \mid s_t = s) = \underset{h:\mathcal{S}\to\mathbb{R}}{\arg\min}\, \mathbb{E}_{Y\sim F_Y}\Big[S(h(s), Y)\Big].$$

- Model $V_t(s;\theta)$ with ANNs $H_t^\psi(s), V_t^\phi(s)$
- Use empirical estimates based on observed data
- Lead to the following loss for the update of H_t^ψ, V_t^ϕ:

$$\underset{\psi,\phi}{\arg\min} \sum_{t\in\mathcal{T}} \sum_{i=1}^{n} S\Big(\underbrace{H_t^\psi(s^{(i)})}_{\text{VaR}_\alpha}, \underbrace{V_t^\phi(s^{(i)})}_{\text{CVaR}_\alpha}, \underbrace{c_t^{(i)} + V_{t+1}^\phi(s_{t+1}^{(i)})}_{\text{random costs}} \Big)$$

[Approximation of V, Coache et al., 2022]

Suppose π^θ is a fixed policy, with its corresponding value function $V_t(s;\theta)$. Then there exist ANNs such that we can approximate $V_t(s;\theta)$ to any arbitrary accuracy for any $t \in \mathcal{T}$ using the framework devised here.

Conditional Elicitability

Example (CVaR$_\alpha$, cont'd): In our RL problems, the random variable Y (i.e. costs) are supported by observed features $s \in \mathcal{S}$ (i.e. states)

$$\rho(Y \mid s_t = s) = \underset{h:\mathcal{S}\to\mathbb{R}}{\arg\min}\, \mathbb{E}_{Y\sim F_Y}\Big[S(h(s), Y)\Big].$$

- Model $V_t(s;\theta)$ with ANNs $H_t^\psi(s), V_t^\phi(s)$
- Use empirical estimates based on observed data
- Lead to the following loss for the update of H_t^ψ, V_t^ϕ:

$$\underset{\psi,\phi}{\arg\min} \sum_{t\in\mathcal{T}} \sum_{i=1}^{n} S\Big(\underbrace{H_t^\psi(s^{(i)})}_{\text{VaR}_\alpha}, \underbrace{V_t^\phi(s^{(i)})}_{\text{CVaR}_\alpha}, \underbrace{c_t^{(i)} + V_{t+1}^\phi(s_{t+1}^{(i)})}_{\text{random costs}}\Big)$$

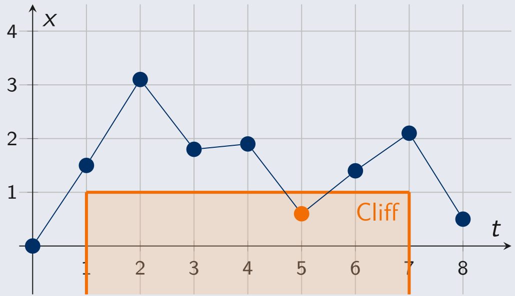
[Approximation of V, Coache et al., 2022]

Suppose π^θ is a fixed policy, with its corresponding value function $V_t(s;\theta)$. Then there exist ANNs such that we can approximate $V_t(s;\theta)$ to any arbitrary accuracy for any $t \in \mathcal{T}$ using the framework devised here.

Cliff Walking

Consider an autonomous rover that:

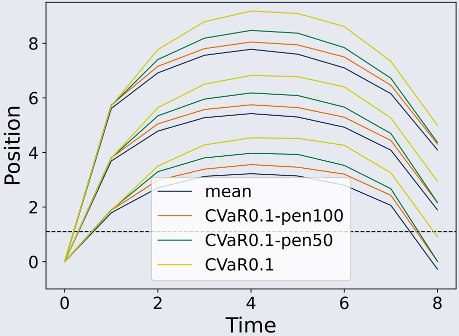
- starts at $(0, 0)$, wants to go at $(T, 0)$
- takes actions $a_t^\theta \sim \pi^\theta = \mathcal{N}(\mu^\theta, \sigma)$
- moves from (t, x_t) to $(t + 1, x_t + a_t)$
- receives penalties when stepping into the cliff and landing away from (T, x)

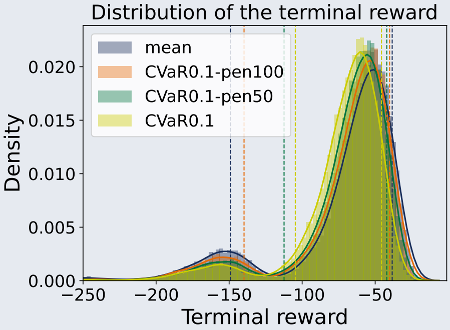


Cliff Walking

Consider an autonomous rover that:

- starts at $(0, 0)$, wants to go at $(T, 0)$
- takes actions $a_t^\theta \sim \pi^\theta = \mathcal{N}(\mu^\theta, \sigma)$
- moves from (t, x_t) to $(t+1, x_t + a_t)$
- receives penalties when stepping into the cliff and landing away from (T, x)

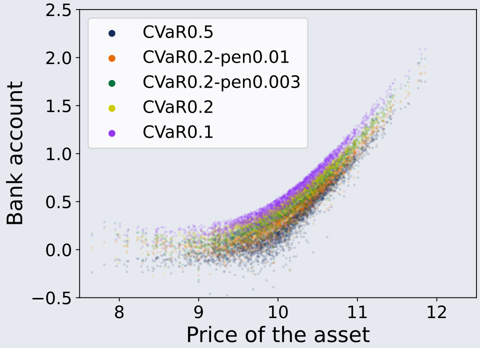




Option Hedging

Consider a call option where underlying asset dynamics follow an Heston model. An agent:

- sells the call option, aims to hedge it trading solely the asset
- observes its previous position, its bank account, the price of the asset
- trades in a market with transaction costs (per share)
- receives a cost that affects its wealth



Portfolio Allocation

Consider a market with d assets. An agent

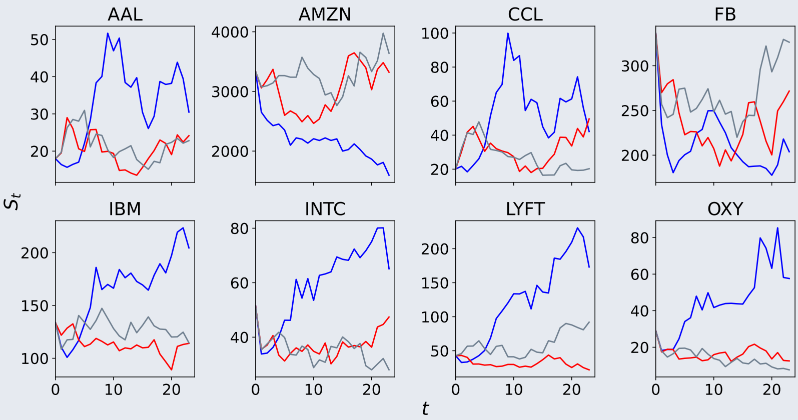
- observes the time t and asset prices $\{S_t^{(i)}\}_{i=1,\dots,d}$
- decides on the proportion of its wealth $\pi_t^{(i)}$ to invest in asset i
- receives feedback from P&L differences $y_t - y_{t+1}$, where its wealth y_t varies according to

$$\mathrm{d}y_t = y_t \left(\sum_{i=1}^{d} \pi_t^{(i)} \frac{\mathrm{d}S_t^{(i)}}{S_t^{(i)}} \right), \quad y_0 = 1.$$

We assume a null interest rate, no leveraging nor short-selling.

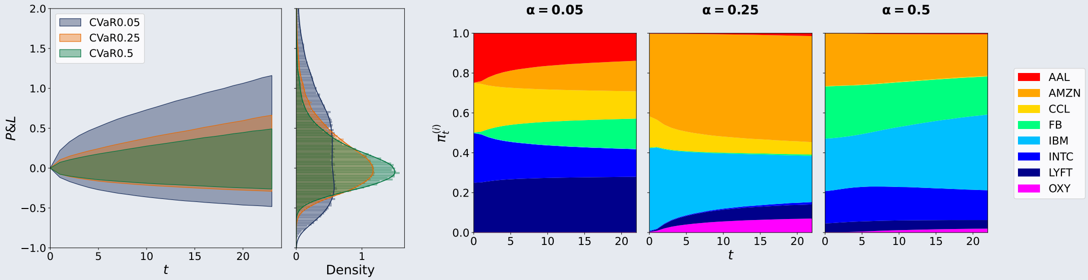
Portfolio Allocation

Co-integration model using daily data from eight different stocks listed on the NASDAQ exchange between September 31, 2020 and December 31, 2021.



Portfolio Allocation

Co-integration model using daily data from eight different stocks listed on the NASDAQ exchange between September 31, 2020 and December 31, 2021.



Account for Model Uncertainty

- Training experience should reflect events similar to those likely to occur during the testing phase
- What if there is model uncertainty?

Robustification of RL approaches:

- Deep RL algorithm to solve problems where the agent minimizes a (static) RDEU of random variables lying within a *Wasserstein ball* [Jaimungal et al., 2022]
- Distributionally robust RL algorithm, restricting to policies having a *KL divergence within a given* ϵ of a reference action probability distribution [Smirnova et al., 2019]
- Bayesian approach to *account for model uncertainty* with recursive risk filters and unobserved costs [Bielecki et al., 2022]
- etc.

Account for Model Uncertainty

- Training experience should reflect events similar to those likely to occur during the testing phase
- What if there is model uncertainty?

Robustification of RL approaches:

- Deep RL algorithm to solve problems where the agent minimizes a (static) RDEU of random variables lying within a *Wasserstein ball* [Jaimungal et al., 2022]
- Distributionally robust RL algorithm, restricting to policies having a *KL divergence within a given ϵ* of a reference action probability distribution [Smirnova et al., 2019]
- Bayesian approach to *account for model uncertainty* with recursive risk filters and unobserved costs [Bielecki et al., 2022]
- etc.

Robustifying Static Risk Measures

Let $\breve{F}_Y$ be the quantile function of Y

- 2-Wasserstein distance: $d_2[Y, Z] = \left(\int_0^1 \left| \breve{F}_Y(u) - \breve{F}_Z(u) \right|^2 \mathrm{d}u \right)^{1/2}$
- Distortion risk measure: $\rho^{\gamma}(Y) = \mathbb{E}\Big[Y \, \gamma(F_Y(Y)) \Big] = \int_0^1 \gamma(u) \breve{F}_Y(u) \mathrm{d}u$

We work with a class of 2-Wasserstein-robust distortion risk measures

$$\rho^{\gamma,\epsilon}(Y) = \sup_{Y^{\phi} \in \varphi_Y^{\epsilon}} \mathbb{E}\Big[Y^{\phi} \, \gamma\Big(F_{Y^{\phi}}(Y^{\phi})\Big) \Big], \quad \text{where} \quad \varphi_Y^{\epsilon} = \Big\{ Y^{\phi} : d_2[Y^{\phi}, Y] \le \epsilon \Big\}$$

- takes into account the uncertainty
- allows risk-averse and risk-seeking behaviors
- are (conditionally) elicitable

Robustifying Static Risk Measures

Let $\breve{F}_Y$ be the quantile function of Y

- 2-Wasserstein distance: $d_2[Y,Z] = \left(\int_0^1 \left| \breve{F}_Y(u) - \breve{F}_Z(u) \right|^2 \mathrm{d}u \right)^{1/2}$
- Distortion risk measure: $\rho^{\gamma}(Y) = \mathbb{E}\Big[Y\, \gamma(F_Y(Y)) \Big] = \int_0^1 \gamma(u) \breve{F}_Y(u) \mathrm{d}u$

We work with a class of 2-Wasserstein-robust distortion risk measures

$$\rho^{\gamma,\epsilon}(Y) = \sup_{Y^\phi \in \varphi_Y^\epsilon} \mathbb{E}\Big[Y^\phi\, \gamma\Big(F_{Y^\phi}(Y^\phi) \Big) \Big], \quad \text{where} \quad \varphi_Y^\epsilon = \Big\{ Y^\phi \,:\, d_2[Y^\phi, Y] \le \epsilon \Big\}$$

- takes into account the uncertainty
- allows risk-averse and risk-seeking behaviors
- are (conditionally) elicitable

Robustifying Static Risk Measures

Let $\breve{F}_Y$ be the quantile function of Y

- 2-Wasserstein distance: $d_2[Y, Z] = \left(\int_0^1 \left| \breve{F}_Y(u) - \breve{F}_Z(u) \right|^2 \mathrm{d}u \right)^{1/2}$
- Distortion risk measure: $\rho^\gamma(Y) = \mathbb{E}\Big[Y\, \gamma(F_Y(Y)) \Big] = \int_0^1 \gamma(u) \breve{F}_Y(u) \mathrm{d}u$

We work with a class of 2-Wasserstein-robust distortion risk measures

$$\rho^{\gamma,\epsilon}(Y) = \sup_{Y^\phi \in \varphi_Y^\epsilon} \mathbb{E}\Big[Y^\phi\, \gamma\Big(F_{Y^\phi}(Y^\phi) \Big) \Big], \quad \text{where} \quad \varphi_Y^\epsilon = \Big\{ Y^\phi : d_2[Y^\phi, Y] \le \epsilon \Big\}$$

- takes into account the uncertainty
- allows risk-averse and risk-seeking behaviors
- are (conditionally) elicitable

Robustifying the Dynamic RL Setup

Problems of the form

$$\min_\theta \rho_{0,T}^{\gamma,\epsilon}\left(\{c_t^\theta\}_{t\in\mathcal{T}}\right) = \min_\theta \rho_0^{\gamma_0,\epsilon_0}\left(c_0^\theta + \rho_1^{\gamma_1,\epsilon_1}\left(c_1^\theta + \cdots + \rho_{T-1}^{\gamma_{T-1},\epsilon_{T-1}}\left(c_{T-1}^\theta + \rho_T^{\gamma_T,\epsilon_T}(c_T^\theta)\right)\cdots\right)\right)$$

where c_t^θ are $\mathcal{F}_{t+1}$-measurable random costs and $\rho_t^{\gamma_t,\epsilon_t}$ are robust distortion risk measures

DP equations for the *value function* for $s \in \mathcal{S}$:

$$V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi_{Y_t^\theta}^{\epsilon_t}} \mathbb{E}\left[Y_t^\phi\, \gamma_t\left(F_{Y_t^\phi}(Y_t^\phi)\right) \,\middle|\, s_t = s\right],$$

with $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

Robustifying the Dynamic RL Setup

Problems of the form

$$\min_{\theta}\ \rho_{0,T}^{\gamma,\epsilon}\Big(\{c_t^\theta\}_{t\in\mathcal{T}}\Big) = \min_{\theta}\ \rho_0^{\gamma_0,\epsilon_0}\left(c_0^\theta + \rho_1^{\gamma_1,\epsilon_1}\left(c_1^\theta + \cdots + \rho_{T-1}^{\gamma_{T-1},\epsilon_{T-1}}\left(c_{T-1}^\theta + \rho_T^{\gamma_T,\epsilon_T}\left(c_T^\theta\right)\right)\cdots\right)\right)$$

where c_t^θ are $\mathcal{F}_{t+1}$-measurable random costs and $\rho_t^{\gamma_t,\epsilon_t}$ are robust distortion risk measures

DP equations for the *value function* for $s \in \mathcal{S}$:

$$V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi_{Y_t^\theta}^{\epsilon_t}} \mathbb{E}\left[Y_t^\phi\ \gamma_t\Big(F_{Y_t^\phi}(Y_t^\phi)\Big)\ \Big|\ s_t = s\right],$$

with $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

Robust Dynamic RL Algorithm

Step 1: Estimate the distribution $F_{Y_t^\theta}$, where $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

- Continuous ranked probability score:

$$F_Y = \arg\min_{F\in\mathbb{F}} \mathbb{E}_{Y\sim F_Y}\Big[S(F,Y)\Big] \quad \text{with} \quad S(F,z) = \int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y\geq z}\Big)^2 \mathrm{d}y$$

Step 2: Estimate $V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi^{\epsilon_t}_{Y_t^\theta}} \mathbb{E}\Big[Y_t^\phi\,\gamma\Big(F_{Y_t^\phi}(Y_t^\phi)\Big)\,\Big|\, s\Big] = \sup_{\breve{F}_\phi \in \varphi^{\epsilon_t}_{\breve{F}_{Y_t^\theta}(\cdot|s)}} \int_0^1 \gamma_t(u)\breve{F}_\phi(u|s)\mathrm{d}u$

- Optimal quantile function is known:

$$\breve{F}_\phi^*(\cdot|s) = \left(\breve{F}_{Y_t^\theta}(\cdot|s) + \frac{\gamma_t(\cdot)}{2\lambda^*}\right)^{\uparrow}, \quad \text{with } \lambda^* > 0 \text{ such that} \quad \int_0^1 \Big|\breve{F}_\phi^*(u|s) - \breve{F}_{Y_t^\theta}(u|s)\Big|^2 \mathrm{d}u = \epsilon_t^2$$

Step 3: Update π^θ via a policy gradient method

- Optimization problem is convex over the space of quantile functions:

$$\nabla_\theta V_t(s;\theta) = -2\,\mathbb{E}\left[\lambda^*\Big(Y_t^{\phi,c} - Y_t^\theta\Big)\frac{\nabla_\theta F_{Y_t^\theta}(x|s)}{f_{Y_t^\theta}(x|s)}\bigg|_{x=Y_t^\theta}\right]$$

Robust Dynamic RL Algorithm

Step 1: Estimate the distribution $F_{Y_t^\theta}$, where $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

- Continuous ranked probability score:

$$F_Y = \arg\min_{F\in\mathbb{F}} \mathbb{E}_{Y\sim F_Y}\Big[S(F,Y)\Big] \quad \text{with} \quad S(F,z) = \int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y\geq z}\Big)^2 \mathrm{d}y$$

Step 2: Estimate $V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi^{\epsilon_t}_{Y_t^\theta}} \mathbb{E}\Big[Y_t^\phi\, \gamma\Big(F_{Y_t^\phi}(Y_t^\phi)\Big)\,\Big|\, s\Big] = \sup_{\breve{F}_\phi \in \varphi^{\epsilon_t}_{\breve{F}_{Y_t^\theta}(\cdot|s)}} \int_0^1 \gamma_t(u)\breve{F}_\phi(u|s)\mathrm{d}u$

- Optimal quantile function is known:

$$\breve{F}^*_\phi(\cdot|s) = \left(\breve{F}_{Y_t^\theta}(\cdot|s) + \frac{\gamma_t(\cdot)}{2\lambda^*}\right)^{\uparrow}, \quad \text{with } \lambda^* > 0 \text{ such that} \quad \int_0^1 \left|\breve{F}^*_\phi(u|s) - \breve{F}_{Y_t^\theta}(u|s)\right|^2 \mathrm{d}u = \epsilon_t^2$$

Step 3: Update π^θ via a policy gradient method

- Optimization problem is convex over the space of quantile functions:

$$\nabla_\theta V_t(s;\theta) = -2\,\mathbb{E}\left[\lambda^*\Big(Y_t^{\phi,c} - Y_t^\theta\Big) \frac{\nabla_\theta F_{Y_t^\theta}(x|s)}{f_{Y_t^\theta}(x|s)}\bigg|_{x=Y_t^\theta}\right]$$

Robust Dynamic RL Algorithm

Step 1: Estimate the distribution $F_{Y_t^\theta}$, where $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta; \theta)$

- Continuous ranked probability score:

$$F_Y = \arg\min_{F \in \mathbb{F}} \mathbb{E}_{Y \sim F_Y}\Big[S(F, Y)\Big] \quad \text{with} \quad S(F, z) = \int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y \geq z}\Big)^2 \mathrm{d}y$$

Step 2: Estimate $V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi^{\epsilon_t}_{Y_t^\theta}} \mathbb{E}\Big[Y_t^\phi \, \gamma\Big(F_{Y_t^\phi}(Y_t^\phi)\Big) \,\Big|\, s\Big] = \sup_{\breve{F}_\phi \in \varphi^{\epsilon_t}_{\breve{F}_{Y_t^\theta}(\cdot|s)}} \int_0^1 \gamma_t(u) \breve{F}_\phi(u|s) \mathrm{d}u$

- Optimal quantile function is known:

$$\breve{F}_\phi^*(\cdot|s) = \left(\breve{F}_{Y_t^\theta}(\cdot|s) + \frac{\gamma_t(\cdot)}{2\lambda^*}\right)^{\uparrow}, \quad \text{with } \lambda^* > 0 \text{ such that } \int_0^1 \left|\breve{F}_\phi^*(u|s) - \breve{F}_{Y_t^\theta}(u|s)\right|^2 \mathrm{d}u = \epsilon_t^2$$

Step 3: Update π^θ via a policy gradient method

- Optimization problem is convex over the space of quantile functions:

$$\nabla_\theta V_t(s;\theta) = -2\,\mathbb{E}\left[\lambda^* \left(Y_t^{\phi,c} - Y_t^\theta\right) \frac{\nabla_\theta F_{Y_t^\theta}(x|s)}{f_{Y_t^\theta}(x|s)}\bigg|_{x = Y_t^\theta}\right]$$

Robust Dynamic RL Algorithm

Step 1: Estimate the distribution $F_{Y_t^\theta}$, where $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

- Continuous ranked probability score:

$$F_Y = \arg\min_{F\in\mathbb{F}} \mathbb{E}_{Y\sim F_Y}\Big[S(F,Y)\Big] \quad \text{with} \quad S(F,z) = \int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y\geq z}\Big)^2 \mathrm{d}y$$

Step 2: Estimate $V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi^{\epsilon_t}_{Y_t^\theta}} \mathbb{E}\Big[Y_t^\phi\,\gamma\Big(F_{Y_t^\phi}(Y_t^\phi)\Big)\,\Big|\,s\Big] = \sup_{\breve{F}_\phi \in \varphi^{\epsilon_t}_{\breve{F}_{Y_t^\theta}(\cdot|s)}} \int_0^1 \gamma_t(u)\breve{F}_\phi(u|s)\mathrm{d}u$

- Optimal quantile function is known:

$$\breve{F}_\phi^*(\cdot|s) = \left(\breve{F}_{Y_t^\theta}(\cdot|s) + \frac{\gamma_t(\cdot)}{2\lambda^*}\right)^{\uparrow}, \quad \text{with } \lambda^* > 0 \text{ such that} \quad \int_0^1 \left|\breve{F}_\phi^*(u|s) - \breve{F}_{Y_t^\theta}(u|s)\right|^2 \mathrm{d}u = \epsilon_t^2$$

Step 3: Update π^θ via a policy gradient method

- Optimization problem is convex over the space of quantile functions:

$$\nabla_\theta V_t(s;\theta) = -2\,\mathbb{E}\left[\lambda^*\Big(Y_t^{\phi,c} - Y_t^\theta\Big)\frac{\nabla_\theta F_{Y_t^\theta}(x|s)}{f_{Y_t^\theta}(x|s)}\Bigg|_{x=Y_t^\theta}\right]$$

Robust Dynamic RL Algorithm

Step 1: Estimate the distribution $F_{Y_t^\theta}$, where $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

- Continuous ranked probability score:

$$F_Y = \underset{F\in\mathbb{F}}{\arg\min}\,\mathbb{E}_{Y\sim F_Y}\Big[S(F,Y)\Big] \quad \text{with} \quad S(F,z) = \int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y\geq z}\Big)^2 \mathrm{d}y$$

Step 2: Estimate $V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi^{\epsilon_t}_{Y_t^\theta}} \mathbb{E}\Big[Y_t^\phi\,\gamma\Big(F_{Y_t^\phi}(Y_t^\phi)\Big)\;\Big|\; s\Big] = \sup_{\breve{F}_\phi \in \varphi^{\epsilon_t}_{\breve{F}_{Y_t^\theta}(\cdot|s)}} \int_0^1 \gamma_t(u)\breve{F}_\phi(u|s)\mathrm{d}u$

- Optimal quantile function is known:

$$\breve{F}_\phi^*(\cdot|s) = \left(\breve{F}_{Y_t^\theta}(\cdot|s) + \frac{\gamma_t(\cdot)}{2\lambda^*}\right)^{\uparrow}, \quad \text{with } \lambda^* > 0 \text{ such that} \quad \int_0^1 \left|\breve{F}_\phi^*(u|s) - \breve{F}_{Y_t^\theta}(u|s)\right|^2 \mathrm{d}u = \epsilon_t^2$$

Step 3: Update π^θ via a policy gradient method

- Optimization problem is convex over the space of quantile functions:

$$\nabla_\theta V_t(s;\theta) = -2\,\mathbb{E}\left[\lambda^*\Big(Y_t^{\phi,c} - Y_t^\theta\Big)\frac{\nabla_\theta F_{Y_t^\theta}(x|s)}{f_{Y_t^\theta}(x|s)}\bigg|_{x=Y_t^\theta}\right]$$

Robust Dynamic RL Algorithm

Step 1: Estimate the distribution $F_{Y_t^\theta}$, where $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta; \theta)$

- Continuous ranked probability score:

$$F_Y = \underset{F \in \mathbb{F}}{\arg\min}\, \mathbb{E}_{Y \sim F_Y}\Big[S(F, Y)\Big] \quad \text{with} \quad S(F, z) = \int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y \geq z}\Big)^2 \mathrm{d}y$$

Step 2: Estimate $V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi_{Y_t^\theta}^{\epsilon_t}} \mathbb{E}\Big[Y_t^\phi \, \gamma\Big(F_{Y_t^\phi}(Y_t^\phi)\Big) \,\Big|\, s\Big] = \sup_{\breve{F}_\phi \in \varphi_{\breve{F}_{Y_t^\theta}(\cdot|s)}^{\epsilon_t}} \int_0^1 \gamma_t(u) \breve{F}_\phi(u|s) \mathrm{d}u$

- Optimal quantile function is known:

$$\breve{F}_\phi^*(\cdot|s) = \left(\breve{F}_{Y_t^\theta}(\cdot|s) + \frac{\gamma_t(\cdot)}{2\lambda^*}\right)^{\uparrow}, \quad \text{with } \lambda^* > 0 \text{ such that} \quad \int_0^1 \left|\breve{F}_\phi^*(u|s) - \breve{F}_{Y_t^\theta}(u|s)\right|^2 \mathrm{d}u = \epsilon_t^2$$

Step 3: Update π^θ via a policy gradient method

- Optimization problem is convex over the space of quantile functions:

$$\nabla_\theta V_t(s;\theta) = -2\,\mathbb{E}\left[\lambda^* \Big(Y_t^{\phi,c} - Y_t^\theta\Big) \frac{\nabla_\theta F_{Y_t^\theta}(x|s)}{f_{Y_t^\theta}(x|s)}\bigg|_{x = Y_t^\theta}\right]$$

Contributions & Future Directions

A unifying, practical framework for risk-aware RL with dynamic risk measures

- Generalization to the broad class of *dynamic convex risk measures*
- Novel setting utilizing *elicitable mappings* to avoid nested simulations
- *Robustification* to protect against model uncertainty

Future directions

- Risk-aware dynamic RL for multi-agent systems
- Implied volatility surfaces for dynamical hedging of exotic options
  (ongoing work with Vedant Choudhary & Sebastian Jaimungal)
- Inverse RL for dynamic risk measures
  (ongoing work with Ziteng Cheng & Sebastian Jaimungal)

Contributions & Future Directions

A unifying, practical framework for risk-aware RL with dynamic risk measures

- Generalization to the broad class of *dynamic convex risk measures*
- Novel setting utilizing *elicitable mappings* to avoid nested simulations
- *Robustification* to protect against model uncertainty

Future directions

- Risk-aware dynamic RL for multi-agent systems
- Implied volatility surfaces for dynamical hedging of exotic options
  (ongoing work with Vedant Choudhary & Sebastian Jaimungal)
- Inverse RL for dynamic risk measures
  (ongoing work with Ziteng Cheng & Sebastian Jaimungal)

Thank you!

More info and slides: anthonycoache.ca

References I

Bäuerle, N. and Glauner, A. (2022). Markov decision processes with recursive risk measures. *European Journal of Operational Research*, 296(3):953–966.

Bielecki, T. R., Cialenco, I., and Ruszczyński, A. (2022). Risk filtering and risk-averse control of Markovian systems subject to model uncertainty. *arXiv preprint arXiv:2206.09235*.

Cheng, Z. and Jaimungal, S. (2022). Markov decision processes with Kusuoka-type conditional risk mappings. *arXiv preprint arXiv:2203.09612*.

Coache, A. and Jaimungal, S. (2021). Reinforcement learning with dynamic convex risk measures. *arXiv preprint arXiv:2112.13414*.

Coache, A., Jaimungal, S., and Cartea, Á. (2022). Conditionally elicitable dynamic risk measures for deep reinforcement learning. *arXiv preprint arXiv:2206.14666*.

Di Castro, D., Oren, J., and Mannor, S. (2019). Practical risk measures in reinforcement learning. *arXiv preprint arXiv:1908.08379*.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707.

Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and stochastics*, 6(4):429–447.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Jaimungal, S., Pesenti, S. M., Wang, Y. S., and Tatsat, H. (2022). Robust risk-aware reinforcement learning. *SIAM Journal on Financial Mathematics*, 13(1):213–226.

References II

Kusuoka, S. (2001). On law invariant coherent risk measures. In *Advances in Mathematical Economics*, pages 83–95. Springer.

Marzban, S., Delage, E., and Li, J. Y. (2021). Deep reinforcement learning for equal risk pricing and hedging under dynamic expectile risk measures. *arXiv preprint arXiv:2109.04001*.

Nass, D., Belousov, B., and Peters, J. (2019). Entropic risk measure in policy search. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1101–1106. IEEE.

Ruszczyński, A. (2010). Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125(2):235–261.

Smirnova, E., Dohmatob, E., and Mary, J. (2019). Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. (2016). Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338.

Algorithms

Algorithm 1: Actor-critic algorithm – Nested simulation approach

Input: ANNs π^θ, V^ϕ, numbers of epochs K's, mini-batch sizes B's

Set initial learning rates for ϕ, θ;
for *each iteration* $k = 1, \dots, K$ **do**
  for *each epoch* $k^\phi = 1, \dots, K^\phi$ **do**
    Simulate a mini-batch of B^ϕ episodes induced by π^θ;
    Generate M^ϕ additional (inner) transitions induced by π^θ;
    Compute the loss $\mathcal{L}(\phi)$: expected square loss between predicted and target values;
    Update ϕ by performing an Adam optimisation step, tune the learning rate for ϕ;
  for *each epoch* $k^\theta = 1, \dots, K^\theta$ **do**
    Simulate a mini-batch of B^θ episodes induced by π^θ;
    Generate M^θ additional (inner) transitions induced by π^θ;
    Compute the loss $\mathcal{L}(\theta)$: policy gradient;
    Update θ by performing an Adam optimisation step, tune the learning rate for θ;

Output: Optimal policy π^θ and its value function V^ϕ

Algorithms

Algorithm 2: Actor-critic algorithm – Elicitable approach

Input: ANNs π^θ, V^ϕ, numbers of epochs K's, mini-batch sizes B's

```
Set initial learning rates for φ, θ;
for each iteration k = 1, ..., K do
    for each epoch k^φ = 1, ..., K^φ do
        Simulate a mini-batch of B^φ episodes induced by π^θ;
        Compute the loss L(φ): minimization of the expected consistent score;
        Update φ by performing an Adam optimisation step, tune the learning rate for φ;
        if k^φ mod K* = 0 then
            Update the target networks φ̃;
    for each epoch k^θ = 1, ..., K^θ do
        Simulate a mini-batch of ⌈B^θ/(1 − α)⌉ episodes induced by π^θ;
        Compute the loss L(θ): policy gradient;
        Update θ by performing an Adam optimisation step, tune the learning rate for θ;
```

Output: Optimal policy π^θ and its value function V^ϕ