Robust Reinforcement Learning with Dynamic Distortion Risk Measures

Anthony Coache (Imperial College London)
anthonycoache.ca

Joint work with
Sebastian Jaimungal (U. Toronto)

Control and Optimization Seminar
University of Connecticut
November 4, 2024

IMPERIAL

Statistical Sciences
UNIVERSITY OF TORONTO

NSERC
CRSNG

Agenda

Motivations

Risk Assessment

Problem Setup

Algorithm

Experiments

Discussion

Agenda

Reinforcement Learning (RL)

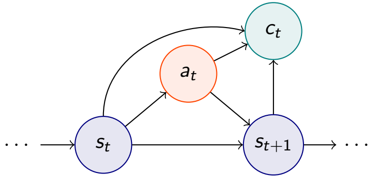
Principled model-agnostic framework for learning-based control

During a training phase, the agent:

↳ interacts with a virtual environment

↳ observes feedback in the form of costs

↳ updates its behaviour; finds best course of action



Applications of interest:

- Portfolio allocation
- Pricing and hedging
- Robot control
- Route optimisation
- Resource allocation
- Healthcare treatments
- Self-driving vehicles
- Control in agriculture
- etc.

Reinforcement Learning (RL)

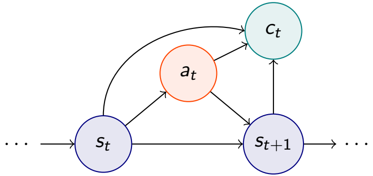
Principled model-agnostic framework for learning-based control

During a training phase, the agent:

↳ interacts with a virtual environment

↳ observes feedback in the form of costs

↳ updates its behaviour; finds best course of action



Applications of interest:

- Portfolio allocation
- Pricing and hedging
- Robot control
- Route optimisation
- Resource allocation
- Healthcare treatments
- Self-driving vehicles
- Control in agriculture
- etc.

Robust Risk-Aware RL

Standard RL: aim at optimising problems of the form $\min_\theta \mathbb{E}[Y^\theta]$, where $Y^\theta = \sum_t \gamma^t c_t^\theta$

- ✗ Ignores the risk of the costs!

Risk-aware RL: e.g. expected utility [Nass et al., 2019], risk-constrained $\mathbb{E}$ [Di Castro et al., 2019], coherent risk [Tamar et al., 2016], etc.

- ✗ Optimising static risk measures leads to optimal precommitment policies!

Robust risk-aware RL: e.g. distributional RL and KL divergence [Smirnova et al., 2019], risk-neutral RL and Wasserstein ball [Abdullah et al., 2019], distributional RL and ϕ-divergence [Clavier et al., 2022], RDEU and Wasserstein ball [Jaimungal et al., 2022], etc.

- ✓ Accounts for model uncertainty
- ✗ Gives time-inconsistent optimal policies!

Robust Risk-Aware RL

Standard RL: aim at optimising problems of the form $\min_\theta \mathbb{E}[Y^\theta]$, where $Y^\theta = \sum_t \gamma^t c_t^\theta$

- × Ignores the risk of the costs!

Risk-aware RL: e.g. expected utility [Nass et al., 2019], risk-constrained $\mathbb{E}$ [Di Castro et al., 2019], coherent risk [Tamar et al., 2016], etc.

- × Optimising static risk measures leads to optimal precommitment policies!

Robust risk-aware RL: e.g. distributional RL and KL divergence [Smirnova et al., 2019], risk-neutral RL and Wasserstein ball [Abdullah et al., 2019], distributional RL and ϕ-divergence [Clavier et al., 2022], RDEU and Wasserstein ball [Jaimungal et al., 2022], etc.

- ✓ Accounts for model uncertainty
- × Gives time-inconsistent optimal policies!

Robust Risk-Aware RL

Standard RL: aim at optimising problems of the form $\min_\theta \mathbb{E}[Y^\theta]$, where $Y^\theta = \sum_t \gamma^t c_t^\theta$

- ✗ Ignores the risk of the costs!

Risk-aware RL: e.g. expected utility [Nass et al., 2019], risk-constrained $\mathbb{E}$ [Di Castro et al., 2019], coherent risk [Tamar et al., 2016], etc.

- ✗ Optimising static risk measures leads to optimal precommitment policies!

Robust risk-aware RL: e.g. distributional RL and KL divergence [Smirnova et al., 2019], risk-neutral RL and Wasserstein ball [Abdullah et al., 2019], distributional RL and ϕ-divergence [Clavier et al., 2022], RDEU and Wasserstein ball [Jaimungal et al., 2022], etc.

- ✓ Accounts for model uncertainty
- ✗ Gives time-inconsistent optimal policies!

Robust Risk-Aware RL (cont'd)

Time-consistent approaches: e.g. recursive risk measures [Chu and Zhang, 2014; Bäuerle and Glauner, 2022], dynamic risk measures [Tamar et al., 2016; Ahmadi et al., 2021; Cheng and Jaimungal, 2022], etc.

- × Applicable only in discrete spaces or tuned to a specific risk measure!

[Bielecki et al., 2023]: DP equations for risk-averse control with partially observable costs

- ✓ Accounts for model uncertainty via Bayesian perspective
- × Requires finite state and action spaces

[Marzban et al., 2023; Coache et al., 2023]: Algorithms for solving RL problems for various classes of dynamic risk measures

- ✓ Efficient estimation method avoiding nested simulations
- × Does not allow robustification

Robust Risk-Aware RL (cont'd)

Time-consistent approaches: e.g. recursive risk measures [Chu and Zhang, 2014; Bäuerle and Glauner, 2022], dynamic risk measures [Tamar et al., 2016; Ahmadi et al., 2021; Cheng and Jaimungal, 2022], etc.

- ✗ Applicable only in discrete spaces or tuned to a specific risk measure!

[Bielecki et al., 2023]: DP equations for risk-averse control with partially observable costs

- ✓ Accounts for model uncertainty via Bayesian perspective
- ✗ Requires finite state and action spaces

[Marzban et al., 2023; Coache et al., 2023]: Algorithms for solving RL problems for various classes of dynamic risk measures

- ✓ Efficient estimation method avoiding nested simulations
- ✗ Does not allow robustification

Robust Risk-Aware RL (cont'd)

Time-consistent approaches: e.g. recursive risk measures [Chu and Zhang, 2014; Bäuerle and Glauner, 2022], dynamic risk measures [Tamar et al., 2016; Ahmadi et al., 2021; Cheng and Jaimungal, 2022], etc.

- ✗ Applicable only in discrete spaces or tuned to a specific risk measure!

[Bielecki et al., 2023]: DP equations for risk-averse control with partially observable costs

- ✓ Accounts for model uncertainty via Bayesian perspective
- ✗ Requires finite state and action spaces

[Marzban et al., 2023; Coache et al., 2023]: Algorithms for solving RL problems for various classes of dynamic risk measures

- ✓ Efficient estimation method avoiding nested simulations
- ✗ Does not allow robustification

Contributions

Goal: develop deep RL algorithms to solve robust risk-aware problems with dynamic risk

- ✓ Actor-critic algorithm optimising dynamic robust risk measures
- ✓ Accounts for model uncertainty and risk in a time-consistent manner
- ✓ Analysis with uncertainty sets induced by the conditional Wasserstein distance
- ✓ Derivation of deterministic policy gradient formulas
- ✓ Universal approximation theorem of the value function
- ✓ Performance evaluation on a portfolio allocation example

Agenda

Dynamic Risk Measures

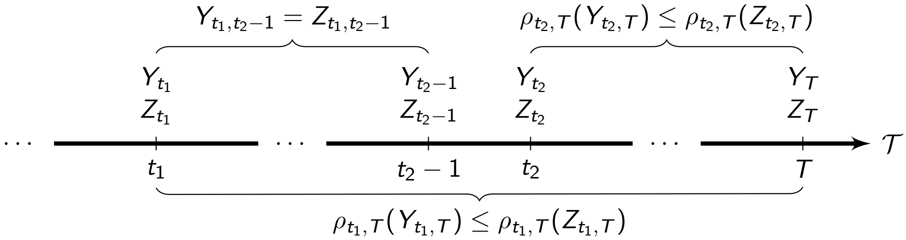
- Let $\mathcal{T} := \{0, 1, \dots, T\}$
- We work on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \in \mathcal{T}}, \mathbb{P})$
- $\mathcal{F}_t$-measurable bounded random costs: $\mathcal{Y}_t := \mathcal{L}^\infty(\Omega, \mathcal{F}_t, \mathbb{P})$
- $\mathcal{Y}_{t_1,t_2} := \mathcal{Y}_{t_1} \times \cdots \times \mathcal{Y}_{t_2}$

Dynamic risk measure: A sequence of maps $\{\rho_{t,T}\}_{t \in \mathcal{T}}$ such that $\rho_{t,T} : \mathcal{Y}_{t,T} \to \mathcal{Y}_t$

Time-Consistency

Strong time-consistency: For any $Y_{t_1,T}, Z_{t_1,T} \in \mathcal{Y}_{t_1,T}$ and $0 \le t_1 < t_2 \le T$, we have

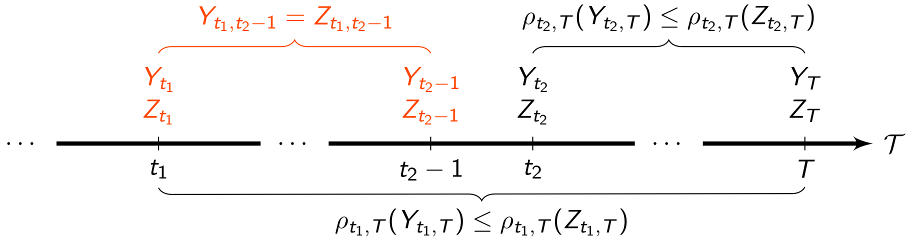
$$\begin{matrix} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \le \rho_{t_2,T}(Z_{t_2,T}) \end{matrix} \implies \rho_{t_1,T}(Y_{t_1,T}) \le \rho_{t_1,T}(Z_{t_1,T})$$



Time-Consistency

Strong time-consistency: For any $Y_{t_1,T}, Z_{t_1,T} \in \mathcal{Y}_{t_1,T}$ and $0 \le t_1 < t_2 \le T$, we have

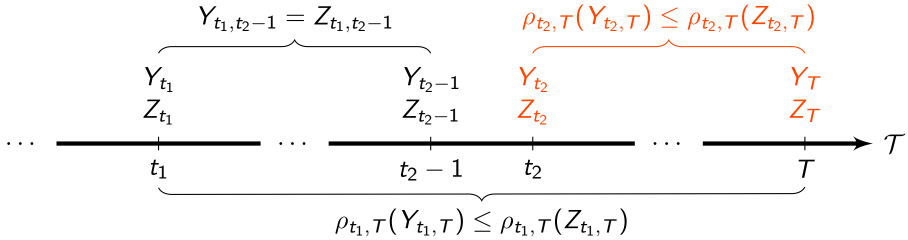
$$\begin{matrix} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \le \rho_{t_2,T}(Z_{t_2,T}) \end{matrix} \implies \rho_{t_1,T}(Y_{t_1,T}) \le \rho_{t_1,T}(Z_{t_1,T})$$



Time-Consistency

Strong time-consistency: For any $Y_{t_1,T}, Z_{t_1,T} \in \mathcal{Y}_{t_1,T}$ and $0 \le t_1 < t_2 \le T$, we have

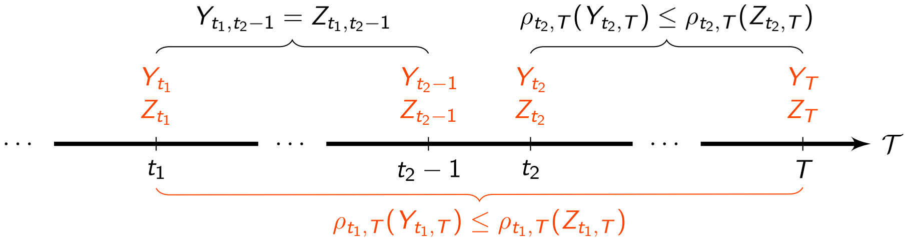
$$\begin{matrix} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \le \rho_{t_2,T}(Z_{t_2,T}) \end{matrix} \implies \rho_{t_1,T}(Y_{t_1,T}) \le \rho_{t_1,T}(Z_{t_1,T})$$



Time-Consistency

Strong time-consistency: For any $Y_{t_1,T}, Z_{t_1,T} \in \mathcal{Y}_{t_1,T}$ and $0 \le t_1 < t_2 \le T$, we have

$$\begin{matrix} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \le \rho_{t_2,T}(Z_{t_2,T}) \end{matrix} \implies \rho_{t_1,T}(Y_{t_1,T}) \le \rho_{t_1,T}(Z_{t_1,T})$$



Time-Consistent Dynamic Risk

Theorem 1 of Ruszczyński [2010]

Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a time-consistent, dynamic risk measure. Suppose that it satisfies

- $\rho_{t,T}(Y_t, Y_{t+1}, \dots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \dots, Y_T)$
- $\rho_{t,T}(0, \dots, 0) = 0$
- $Y \leq Z$ a.s. $\implies \rho_{t,T}(Y) \leq \rho_{t,T}(Z)$

Then $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ may be expressed as

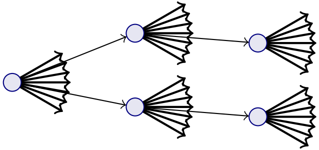
$$\rho_{t,T}(Y_{t,T}) = Y_t + \rho_t\Bigg(Y_{t+1} + \rho_{t+1}\bigg(Y_{t+2} + \cdots + \rho_{T-2}\Big(Y_{T-1} + \rho_{T-1}\big(Y_T\big)\Big)\cdots\bigg)\Bigg),$$

where each one-step conditional risk measure $\rho_t : \mathcal{Y}_{t+1} \to \mathcal{Y}_t$ satisfies $\rho_t(Y) = \rho_{t,t+1}(0, Y)$ for any $Y \in \mathcal{Y}_{t+1}$.

Elicitability

Nested simulations are computationally expensive...

- Simulation of N episodes with T periods
- Additional M inner transitions for each state



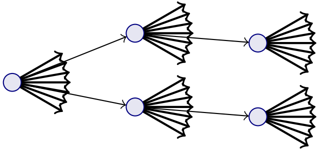
ρ_t is k-elicitable [Gneiting, 2011] iff there exists a scoring function $S : \mathbb{R}^k \times \mathbb{Y} \to \mathbb{R}$ s.t.

$$\rho_t(Y) = \arg\min_{\mathfrak{a} \in \mathbb{R}^k} \mathbb{E}_{Y \sim F_{Y|\mathcal{F}_t}} \Big[S(\mathfrak{a}, Y) \Big].$$

Elicitability

Nested simulations are computationally expensive...

- Simulation of N episodes with T periods
- Additional M inner transitions for each state



ρ_t is k-elicitable [Gneiting, 2011] iff there exists a scoring function $S : \mathbb{R}^k \times \mathbb{Y} \to \mathbb{R}$ s.t.

$$\rho_t(Y) = \arg\min_{\mathfrak{a} \in \mathbb{R}^k} \mathbb{E}_{Y \sim F_{Y|\mathcal{F}_t}} \Big[S(\mathfrak{a}, Y) \Big].$$

Elicitable Mappings

Expectation: $\mathbb{E}[Y] = \underset{\mathfrak{a}\in\mathbb{R}}{\arg\min}\, \mathbb{E}_{Y\sim F_Y}\left[(\mathfrak{a} - Y)^2\right]$

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

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

Conditional elicitable maps:

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

Elicitable Mappings

Expectation: $\mathbb{E}[Y] = \arg\min_{\mathfrak{a}\in\mathbb{R}} \mathbb{E}_{Y\sim F_Y}\left[(\mathfrak{a} - Y)^2\right]$

Pair VaR_α-CVaR_α: $\left(\text{VaR}_\alpha(Y), \text{CVaR}_\alpha(Y)\right) = \arg\min_{(\mathfrak{a}_1,\mathfrak{a}_2)\in\mathbb{R}^2} \mathbb{E}_{Y\sim F_Y}\left[S(\mathfrak{a}_1, \mathfrak{a}_2, Y)\right]$, with

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

Conditional elicitable maps:

$$\rho_t(Y \mid s_t = s) = \arg\min_{h:\mathcal{S}\to\mathbb{R}} \mathbb{E}_{Y\sim F_Y}\left[S(h(s), Y)\right]$$

Elicitable Mappings

Expectation: $\mathbb{E}[Y] = \underset{\mathfrak{a}\in\mathbb{R}}{\arg\min}\, \mathbb{E}_{Y\sim F_Y}\left[(\mathfrak{a} - Y)^2\right]$

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

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

Conditional elicitable maps:

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

Agenda

Problem Setup

Problems of the form

$$\min_{\pi} \rho_{0,T}\Big(\{c_t^\pi\}_t\Big) = \min_{\pi} \rho_0\Bigg(c_0^\pi + \rho_1\Big(c_1^\pi + \cdots + \rho_{T-1}\Big(c_{T-1}^\pi + \rho_T\big(c_T^\pi\big)\Big)\cdots\Big)\Bigg)$$

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

Running risk-to-go satisfies dynamic programming equations:

$$V_t(s;\pi) = \rho_t\Big(c_t^\pi + V_{t+1}(s_{t+1}^\pi;\pi) \;\Big|\; s_t = s\Big)$$

$$Q_t(s,a;\pi) = \rho_t\Big(c_t + Q_{t+1}(s_{t+1},\pi(s_{t+1});\pi) \;\Big|\; s_t = s,\ a_t = a\Big)$$

Problem Setup

Problems of the form

$$\min_{\pi} \rho_{0,T}\left(\{c_t^\pi\}_t\right) = \min_{\pi} \rho_0\left(c_0^\pi + \rho_1\left(c_1^\pi + \cdots + \rho_{T-1}\left(c_{T-1}^\pi + \rho_T\left(c_T^\pi\right)\right)\cdots\right)\right)$$

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

Running risk-to-go satisfies dynamic programming equations:

$$V_t(s;\pi) = \rho_t\left(c_t^\pi + V_{t+1}(s_{t+1}^\pi;\pi) \;\middle|\; s_t = s\right)$$

$$Q_t(s,a;\pi) = \rho_t\left(c_t + Q_{t+1}(s_{t+1}, \pi(s_{t+1});\pi) \;\middle|\; s_t = s,\ a_t = a\right)$$

Account for Model Uncertainty

Training experience should reflect events similar to those likely to occur during testing

↳ What if there is model uncertainty?

We include uncertainty sets within dynamic risk measures [Moresco et al., 2024]

↳ Leads to time-consistent optimal policies

↳ Provides a general equivalence between time-consistency and robust dynamic risk

↳ Shows equivalence between uncertainty on the entire stochastic process and one-step uncertainty sets

Robust one-step conditional risk: For an uncertainty set $\varphi^\epsilon : \mathcal{Y}_{t+1} \to 2^{\mathcal{Y}_{t+1}}$, define

$$\varrho_t^\epsilon(Y) = \operatorname{ess\,sup} \left\{ \rho_t(Y^\phi) \ : \ Y^\phi \in \varphi_Y^\epsilon \right\}$$

Account for Model Uncertainty

Training experience should reflect events similar to those likely to occur during testing

↳ What if there is model uncertainty?

We include uncertainty sets within dynamic risk measures [Moresco et al., 2024]

↳ Leads to time-consistent optimal policies

↳ Provides a general equivalence between time-consistency and robust dynamic risk

↳ Shows equivalence between uncertainty on the entire stochastic process and one-step uncertainty sets

Robust one-step conditional risk: For an uncertainty set $\varphi^{\epsilon} : \mathcal{Y}_{t+1} \to 2^{\mathcal{Y}_{t+1}}$, define

$$\varrho_t^{\epsilon}(Y) = \operatorname{ess\,sup} \left\{ \rho_t(Y^{\phi}) \ : \ Y^{\phi} \in \varphi_{Y}^{\epsilon} \right\}$$

Dynamic Robust Distortion Risk Measures

We aim to optimise a class of dynamic robust distortion risk measures with piecewise linear γ_s and uncertainty sets induced by the conditional 2-Wasserstein distance

$$\varrho_t^{\epsilon_s,\gamma_s}(Y_t^\pi) = \operatorname*{ess\,sup}_{Y^\phi \in \varphi^{\epsilon_s}_{Y_t^\pi}} \left\langle \gamma_s, \breve{F}_\phi(\cdot|s,a) \right\rangle \quad \text{with} \quad Y_t^\pi := c_t(s,a,s') + V_{t+1}(s';\pi).$$

- **w/o moment constraints**:

$$\vartheta_Y^\epsilon = \left\{ Y^\phi \in \mathcal{Y}_{t+1} : \ \left\| \breve{F}_{Y|\mathcal{F}_t} - \breve{F}_{Y^\phi|\mathcal{F}_t} \right\| \le \epsilon \right\}$$

- **w/ moment constraints**:

$$\varsigma_Y^\epsilon = \left\{ Y^\phi \in \mathcal{Y}_{t+1} : \ \begin{array}{c} \left\| \breve{F}_{Y|\mathcal{F}_t} - \breve{F}_{Y^\phi|\mathcal{F}_t} \right\| \le \epsilon, \\ \langle \breve{F}_{Y^\phi|\mathcal{F}_t}, 1 \rangle = \langle \breve{F}_{Y|\mathcal{F}_t}, 1 \rangle, \\ \left\| \breve{F}_{Y^\phi|\mathcal{F}_t} \right\|^2 = \left\| \breve{F}_{Y|\mathcal{F}_t} \right\|^2 \end{array} \right\}$$

Dynamic Robust Distortion Risk Measures

We aim to optimise a class of dynamic robust distortion risk measures with piecewise linear γ_s and uncertainty sets induced by the conditional 2-Wasserstein distance

$$\varrho_t^{\epsilon_s,\gamma_s}(Y_t^\pi) = \operatorname*{ess\,sup}_{Y^\phi \in \varphi_{Y_t^\pi}^{\epsilon_s}} \left\langle \gamma_s, \breve{F}_\phi(\cdot|s,a) \right\rangle \quad \text{with} \quad Y_t^\pi := c_t(s,a,s') + V_{t+1}(s';\pi).$$

- **w/o moment constraints**:

$$\vartheta_Y^\epsilon = \left\{ Y^\phi \in \mathcal{Y}_{t+1} \ : \ \left\| \breve{F}_{Y|_{\mathcal{F}_t}} - \breve{F}_{Y^\phi|_{\mathcal{F}_t}} \right\| \leq \epsilon \ \right\}$$

- **w/ moment constraints**:

$$\varsigma_Y^\epsilon = \left\{ Y^\phi \in \mathcal{Y}_{t+1} \ : \ \begin{array}{c} \left\| \breve{F}_{Y|_{\mathcal{F}_t}} - \breve{F}_{Y^\phi|_{\mathcal{F}_t}} \right\| \leq \epsilon, \\ \langle \breve{F}_{Y^\phi|_{\mathcal{F}_t}}, 1 \rangle = \langle \breve{F}_{Y|_{\mathcal{F}_t}}, 1 \rangle, \\ \left\| \breve{F}_{Y^\phi|_{\mathcal{F}_t}} \right\|^2 = \left\| \breve{F}_{Y|_{\mathcal{F}_t}} \right\|^2 \end{array} \right\}$$

Agenda

Optimal Quantile Function w/o Moments

[Thm. 3.9, Pesenti and Jaimungal, 2023]

Consider dynamic robust distortion risk measures, where $\varphi^{\epsilon_s}_{Y^\theta_t} = \vartheta^{\epsilon_s}_{Y^\theta_t}$. The optimal quantile function is given by

$$\breve{F}^*_\phi(\cdot|s,a) = \left(\breve{F}_{Y^\theta_t}(\cdot|s,a) + \frac{\gamma_s(\cdot)}{2\lambda^*}\right)^{\uparrow},$$

where $\lambda^* > 0$ is such that $\left\|\breve{F}^*_\phi(\cdot|s,a) - \breve{F}_{Y^\theta_t}(\cdot|s,a)\right\|^2 = \epsilon_s^2$ and $F^\uparrow := \arg\min_{G\in\mathbb{F}}\{\|G-F\|^2\}$ denotes the isotonic projection of a function F, where

$$\mathbb{F} = \{F \in \mathbb{L}^2([0,1]) \,:\, F \text{ is nondecreasing and left-continuous}\}.$$

Optimal Quantile Function w/o Moments (cont'd)

If γ_s is nondecreasing, then $\breve{F}^*_\phi(\cdot|s,a) = \breve{F}_{Y_t^\theta}(\cdot|s,a) + \frac{\epsilon_s \gamma_s}{\|\gamma_s\|}$. In addition, we obtain

$$
\begin{aligned}
Q_t(s,a;\theta) &= \operatorname*{ess\,sup}_{Y_t^\phi \in \vartheta^{\epsilon_s}_{Y_t^\theta}} \left\langle \gamma_s, \breve{F}_{Y_t^\phi}(\cdot|s,a) \right\rangle \\
&= \left\langle \gamma_s, \breve{F}_{Y_t^\theta}(\cdot|s,a) \right\rangle + \epsilon_s \|\gamma_s\| \\
&= \left\langle \gamma_s, \breve{F}_{\underbrace{(c_t + \epsilon_s \|\gamma_s\|)}_{c'_t} + Q_{t+1}(s_{t+1}, \pi^\theta(s_{t+1}); \theta)}(\cdot|s,a) \right\rangle
\end{aligned}
$$

↳ The $\mathcal{F}_t$-measurable shift $\epsilon_s \|\gamma_s\|$ may be included as part of the cost function

↳ State-independent ϵ, γ lead to identical robust and non-robust optimal policies

Optimal Quantile Function w/o Moments (cont'd)

If γ_s is nondecreasing, then $\breve{F}^*_\phi(\cdot|s,a) = \breve{F}_{Y_t^\theta}(\cdot|s,a) + \frac{\epsilon_s \gamma_s}{\|\gamma_s\|}$. In addition, we obtain

$$
\begin{aligned}
Q_t(s,a;\theta) &= \operatorname*{ess\,sup}_{Y_t^\phi \in \vartheta^{\epsilon_s}_{Y_t^\theta}} \left\langle \gamma_s, \breve{F}_{Y_t^\phi}(\cdot|s,a) \right\rangle \\
&= \left\langle \gamma_s, \breve{F}_{Y_t^\theta}(\cdot|s,a) \right\rangle + \epsilon_s \|\gamma_s\| \\
&= \left\langle \gamma_s, \breve{F}_{\underbrace{(c_t + \epsilon_s \|\gamma_s\|)}_{c'_t} + Q_{t+1}(s_{t+1}, \pi^\theta(s_{t+1}); \theta)}(\cdot|s,a) \right\rangle
\end{aligned}
$$

↳ The $\mathcal{F}_t$-measurable shift $\epsilon_s \|\gamma_s\|$ may be included as part of the cost function

↳ State-independent ϵ, γ lead to identical robust and non-robust optimal policies

Optimal Quantile Function w/ Moments

We cast in a dynamic setting [Thm. 3.1, Bernard et al., 2023]:

Theorem [C., Jaimungal, 2024]

Consider dynamic robust distortion risk measures, where γ_s is nondecreasing and

$$\varsigma_{Y_t^\theta}^{\epsilon_s} = \left\{ Y^\phi \in \mathcal{Y}_{t+1} : \|\breve{F}_{Y_t^\theta|_{\mathcal{F}_t}} - \breve{F}_{Y^\phi|_{\mathcal{F}_t}}\| \le \epsilon_s, \quad \mu = \langle \breve{F}_{Y^\phi|_{\mathcal{F}_t}}, 1 \rangle, \quad \mu^2 + \sigma^2 = \|\breve{F}_{Y^\phi|_{\mathcal{F}_t}}\|^2 \right\}.$$

The optimal quantile function is then given by

$$\breve{F}_\phi^*(u|s,a) = \mu + \frac{\lambda^* \left(\breve{F}_{Y_t^\theta}(u|s,a) - \mu \right) + \gamma_s(u) - 1}{b_{\lambda^*}},$$

where λ^* and b_{λ^*} depend non-trivially on ϵ_s, γ_s, and $\breve{F}_{Y_t^\theta}$.

Additionally, the optimal solution remains valid with $\lambda^* = 0$ if the tolerance ϵ_s is sufficiently large.

Deterministic Gradient

Theorem [C., Jaimungal, 2024]

Consider dynamic robust distortion risk measures, where γ_s is non-decreasing and

$$\varsigma_{Y_t^\theta}^{\epsilon_s} = \left\{ Y^\phi \in \mathcal{Y}_{t+1} \,:\, \|\breve{F}_{Y_t^\theta|\mathcal{F}_t} - \breve{F}_{Y^\phi|\mathcal{F}_t}\| \leq \epsilon_s, \quad \mu = \langle \breve{F}_{Y^\phi|\mathcal{F}_t}, 1 \rangle, \quad \mu^2 + \sigma^2 = \|\breve{F}_{Y^\phi|\mathcal{F}_t}\|^2 \right\}.$$

The gradient of the value function is given by

$$\begin{aligned}\nabla_\theta V_t(s;\theta) = \nabla_a Q_t(s,a;\theta)\Big|_{a=\pi^\theta(s)} \nabla_\theta \pi^\theta(s) \\ - \frac{b_{\lambda^*} - \lambda^*}{b_{\lambda^*}} \mathbb{E}_{t,s}\left[\Big((b_{\lambda^*} - \lambda^*)(Y_t^\theta - \mu) + 1 \Big) \frac{\nabla_a F_{Y_t^\theta}(x|s,a)}{\nabla_x F_{Y_t^\theta}(x|s,a)} \Bigg|_{(x,a)=(Y_t^\theta, \pi^\theta(s))} \right] \nabla_\theta \pi^\theta(s).\end{aligned}$$

↳ Reduces to deterministic policy gradient [Silver et al., 2014] when $\epsilon_s \downarrow 0$

Algorithm

We parameterise the functionals by neural networks, and wish to optimise the value function $V_t(s, \theta) = Q_t(s, \pi^\theta(s); \theta)$ over policies θ via policy gradient approach:

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

Actor-critic style algorithm composed of interleaved procedures:

- ✓ estimate the distribution of costs-to-go
- ✓ approximate the running risk-to-go
- ✓ update the policy via deterministic policy gradient

Algorithm (cont'd)

Step 1: Estimate the distribution $F_{Y_t^\theta|_{(s,a)}}$ where $Y_t^\theta := c_t(s,a,s') + Q_{t+1}^\theta(s', \pi^\theta(s'))$

↳ Continuous ranked probability score as strictly proper scoring rule

↳ Requires an estimation of the Q-function...

Step 2: Approximate the running risk-to-go $Q_t^\theta(s,a) = \operatorname*{ess\,sup}_{\breve{F}_\phi \in \varphi^{\epsilon_s}_{\breve{F}_{Y_t^\theta|_{(s,a)}}}} \left\langle \gamma_s, \breve{F}_\phi(\cdot|s,a) \right\rangle$

↳ Known optimal quantile function $\breve{F}_\phi^*$, and class of elicitable one-step risk measures

↳ Changes the distribution of Y_t^θ...

Step 3: Update π^θ with the analytical deterministic gradient formula

↳ Convex optimisation over the space of quantile functions

Algorithm (cont'd)

Step 1: Estimate the distribution $F_{Y_t^\theta|_{(s,a)}}$ where $Y_t^\theta := c_t(s,a,s') + Q_{t+1}^\theta(s', \pi^\theta(s'))$

↳ Continuous ranked probability score as strictly proper scoring rule

↳ Requires an estimation of the Q-function...

Step 2: Approximate the running risk-to-go $Q_t^\theta(s,a) = \operatorname*{ess\,sup}_{\breve{F}_\phi \in \varphi^{\epsilon_s}_{\breve{F}_{Y_t^\theta|_{(s,a)}}}} \left\langle \gamma_s, \breve{F}_\phi(\cdot|s,a) \right\rangle$

↳ Known optimal quantile function $\breve{F}_\phi^*$, and class of elicitable one-step risk measures

↳ Changes the distribution of Y_t^θ...

Step 3: Update π^θ with the analytical deterministic gradient formula

↳ Convex optimisation over the space of quantile functions

Algorithm (cont'd)

Step 1: Estimate the distribution $F_{Y_t^\theta|_{(s,a)}}$ where $Y_t^\theta := c_t(s,a,s') + Q_{t+1}^\theta(s', \pi^\theta(s'))$

↳ Continuous ranked probability score as strictly proper scoring rule

↳ Requires an estimation of the Q-function...

Step 2: Approximate the running risk-to-go $Q_t^\theta(s,a) = \operatorname*{ess\,sup}_{\breve{F}_\phi \in \varphi^{\epsilon_s}_{\breve{F}_{Y_t^\theta|_{(s,a)}}}} \left\langle \gamma_s, \breve{F}_\phi(\cdot|s,a) \right\rangle$

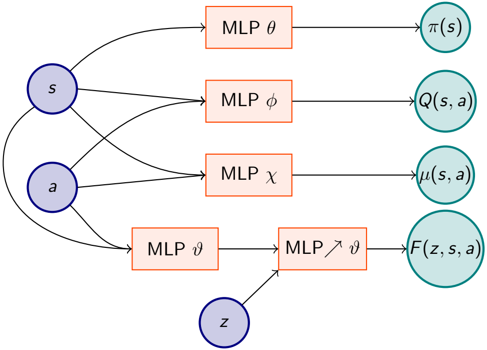
↳ Known optimal quantile function $\breve{F}_\phi^*$, and class of elicitable one-step risk measures

↳ Changes the distribution of Y_t^θ...

Step 3: Update π^θ with the analytical deterministic gradient formula

↳ Convex optimisation over the space of quantile functions

Neural Network Structure



- For layers that are descendant of z, we constrain the weights to non-negative values and use monotonic activation function to ensure a nondecreasing mapping [Sill, 1997]
- There exists a sufficiently large ANN approximating Q to any arbitrary accuracy

Agenda

Experimental Setup

Consider a market with multiple assets, where an agent

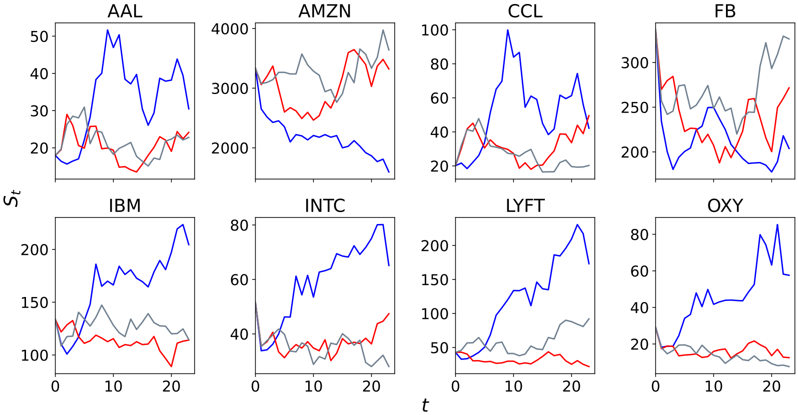
- ↳ observes the time and asset prices
- ↳ decides on the proportion of wealth to invest in each asset
- ↳ receives feedback from P&L differences
- ↳ assumes a null interest rate, no leveraging nor short-selling

We estimate a co-integration model with daily data from different stocks and use the resulting estimated model as a simulation engine to generate price paths

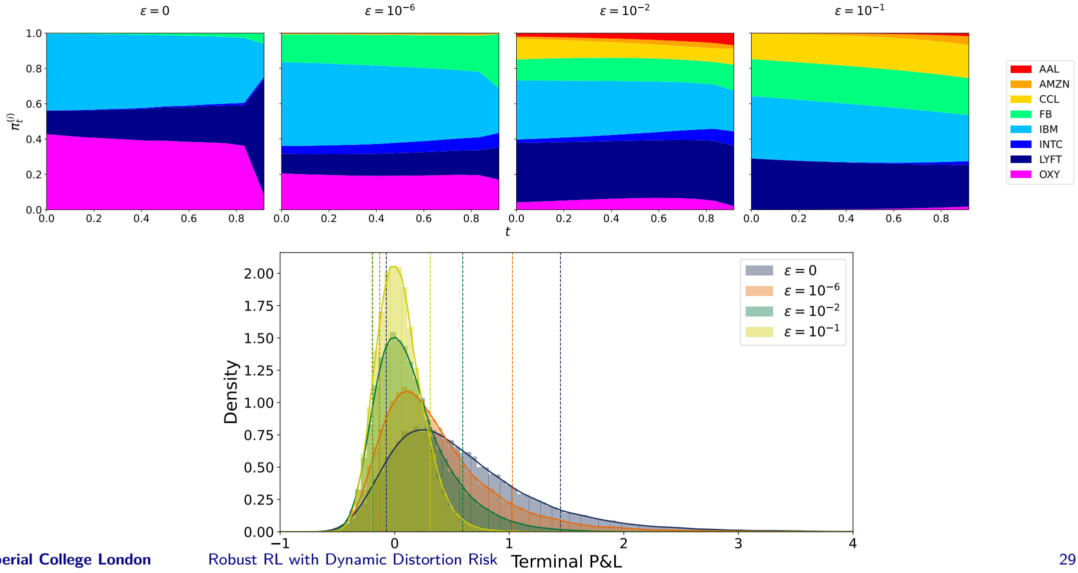
$$\Delta S_\tau = \alpha \beta^\intercal S_{\tau-1} + \Gamma_1 \Delta S_{\tau-1} + \cdots + \Gamma_{k_{ar}-1} \Delta S_{\tau-k_{ar}+1} + C D_\tau + u_\tau$$

Simulation Engine

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



Robust Portfolio Allocation



Agenda

Motivations

Risk Assessment

Problem Setup

Algorithm

Experiments

Discussion

Future Directions

Practical algorithm for risk-sensitive RL with dynamic robust risk measures

↳ Accounts simultaneously for **risk** and **model uncertainty**

↳ Utilises **elicitable mappings** to avoid nested simulations

↳ Proves that classical deterministic policy gradient is a limiting case

Future directions:

- Other classes of dynamic robust risk measures
- Multi-agent RL with dynamic risk measures
- Identification of risk-aversion using inverse RL
- Model-based methods for partially observable MDPs

Thank you!

More info and slides:

References I

Abdullah, M. A., Ren, H., Ammar, H. B., Milenkovic, V., Luo, R., Zhang, M., and Wang, J. (2019). Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*.

Ahmadi, M., Rosolia, U., Ingham, M. D., Murray, R. M., and Ames, A. D. (2021). Constrained risk-averse Markov decision processes. In *The 35th AAAI Conference on Artificial Intelligence (AAAI-21)*.

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

Bernard, C., Pesenti, S. M., and Vanduffel, S. (2023). Robust distortion risk measures. *Mathematical Finance*.

Bielecki, T. R., Cialenco, I., and Ruszczyński, A. (2023). Risk filtering and risk-averse control of markovian systems subject to model uncertainty. *Mathematical Methods of Operations Research*, 98(2):231–268.

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

Chu, S. and Zhang, Y. (2014). Markov decision processes with iterated coherent risk measures. *International Journal of Control*, 87(11):2286–2293.

Clavier, P., Allassonière, S., and Pennec, E. L. (2022). Robust reinforcement learning with distributional risk-averse formulation. *arXiv preprint arXiv:2206.06841*.

Coache, A., Jaimungal, S., and Cartea, Á. (2023). Conditionally elicitable dynamic risk measures for deep reinforcement learning. *SIAM Journal on Financial Mathematics*, 14(4):1249–1289.

References II

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

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.

Marzban, S., Delage, E., and Li, J. Y.-M. (2023). Deep reinforcement learning for option pricing and hedging under dynamic expectile risk measures. *Quantitative Finance*, 23(10):1411–1430.

Moresco, M. R., Mailhot, M., and Pesenti, S. M. (2024). Uncertainty propagation and dynamic robust risk measures. *Mathematics of Operations Research*.

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.

Pesenti, S. M. and Jaimungal, S. (2023). Portfolio optimization within a wasserstein ball. *SIAM Journal on Financial Mathematics*, 14(4):1175–1214.

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

Sill, J. (1997). Monotonic networks. *Advances in Neural Information Processing Systems*, 10.

References III

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, pages 387–395. PMLR.

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.