Robust Reinforcement Learning with Dynamic Distortion Risk Measures

Anthony Coache (Imperial College London)
anthonycoache.ca

Joint work with
Sebastian Jaimungal (U. Toronto)

Workshop on Mathematical Insights from Markets, Control, and Learning ⋆ Sept. 26, 2024 ⋆ Aussois

IMPERIAL

Statistical Sciences UNIVERSITY OF TORONTO

NSERC CRSNG

Agenda

Reinforcement Learning (RL)

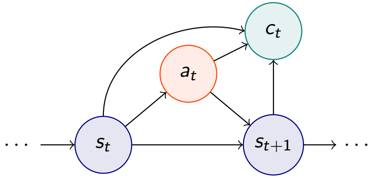
Principled model-agnostic framework for learning-based control

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.

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!

(Robust) risk-sensitive RL: e.g. expected utility [Nass et al., 2019], risk-constrained $\mathbb{E}$ [Di Castro et al., 2019], coherent risk [Tamar et al., 2016], distributional RL and ϕ-divergence [Smirnova et al., 2019; Clavier et al., 2022], RDEU and Wasserstein ball [Jaimungal et al., 2022], etc.

× Optimising static risk measures leads to optimal precommitment policies!

Time-consistent approaches: e.g. recursive risk measures [Chu and Zhang, 2014; Bäuerle and Glauner, 2022; Bielecki et al., 2023], dynamic risk measures [Tamar et al., 2016; Ahmadi et al., 2021; Cheng and Jaimungal, 2022; Marzban et al., 2023], etc.

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

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!

(Robust) risk-sensitive RL: e.g. expected utility [Nass et al., 2019], risk-constrained $\mathbb{E}$ [Di Castro et al., 2019], coherent risk [Tamar et al., 2016], distributional RL and ϕ-divergence [Smirnova et al., 2019; Clavier et al., 2022], RDEU and Wasserstein ball [Jaimungal et al., 2022], etc.

× Optimising static risk measures leads to optimal precommitment policies!

Time-consistent approaches: e.g. recursive risk measures [Chu and Zhang, 2014; Bäuerle and Glauner, 2022; Bielecki et al., 2023], dynamic risk measures [Tamar et al., 2016; Ahmadi et al., 2021; Cheng and Jaimungal, 2022; Marzban et al., 2023], etc.

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

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!

(Robust) risk-sensitive RL: e.g. expected utility [Nass et al., 2019], risk-constrained $\mathbb{E}$ [Di Castro et al., 2019], coherent risk [Tamar et al., 2016], distributional RL and ϕ-divergence [Smirnova et al., 2019; Clavier et al., 2022], RDEU and Wasserstein ball [Jaimungal et al., 2022], etc.

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

Time-consistent approaches: e.g. recursive risk measures [Chu and Zhang, 2014; Bäuerle and Glauner, 2022; Bielecki et al., 2023], dynamic risk measures [Tamar et al., 2016; Ahmadi et al., 2021; Cheng and Jaimungal, 2022; Marzban et al., 2023], etc.

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

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 Q-function
- ✓ Performance evaluation on a portfolio allocation example

Dynamic Risk Measures

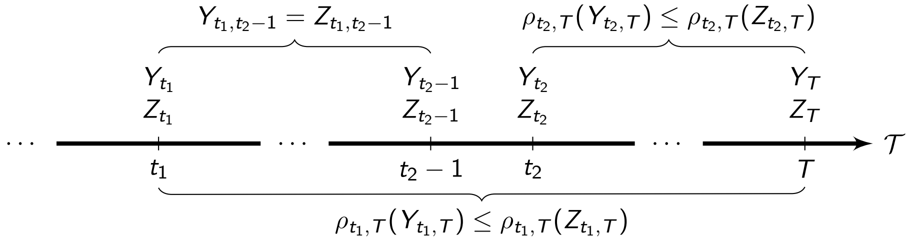
- Let $\mathcal{T} := \{0, 1, \dots, T\}$
- We work on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathcal{T}}, \mathbb{P})$
- $\mathcal{F}_t$-measurable bounded random costs: $\mathcal{Y}_t := \mathcal{L}^\infty(\Omega, \mathcal{F}_t, \mathbb{P})$
- $\mathcal{Y}_{t_1,t_2} := \mathcal{Y}_{t_1} \times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

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 $0 \leq t_1 < t_2 \leq T$, we have

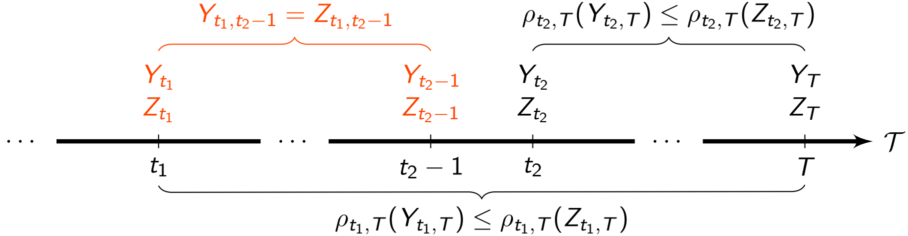
$$\begin{matrix} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \leq \rho_{t_2,T}(Z_{t_2,T}) \end{matrix} \implies \rho_{t_1,T}(Y_{t_1,T}) \leq \rho_{t_1,T}(Z_{t_1,T})$$



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

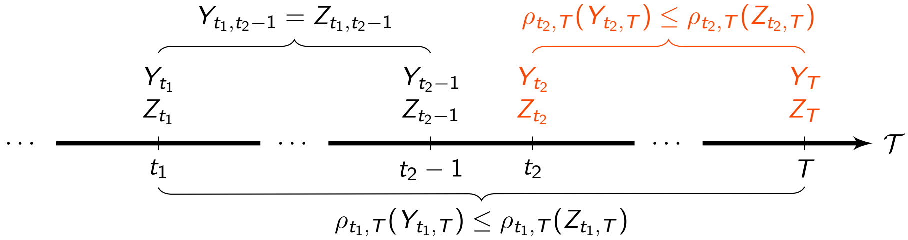
$$\left.\begin{array}{c} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \le \rho_{t_2,T}(Z_{t_2,T}) \end{array}\right. \implies \rho_{t_1,T}(Y_{t_1,T}) \le \rho_{t_1,T}(Z_{t_1,T})$$



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

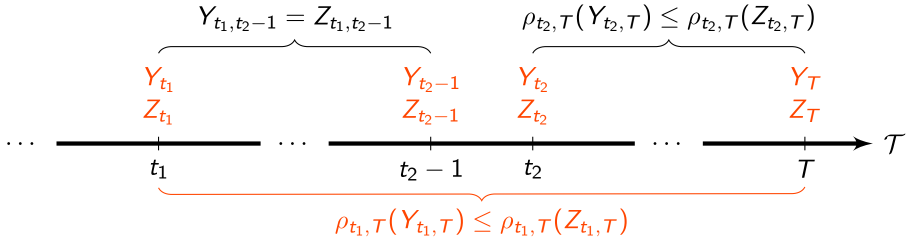
$$\begin{matrix} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \le \rho_{t_2,T}(Z_{t_2,T}) \end{matrix} \implies \rho_{t_1,T}(Y_{t_1,T}) \le \rho_{t_1,T}(Z_{t_1,T})$$



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

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



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

- $\rho_{t,T}(Y_t, Y_{t+1}, \ldots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \ldots, Y_T)$
- $\rho_{t,T}(0, \ldots, 0) = 0$
- $Y \leq Z$ 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 $Y \in \mathcal{Y}_{t+1}$.

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^π are $\mathcal{F}_{t+1}$-measurable random costs and ρ_t are one-step conditional risk measures.

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

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

where c_t^π are $\mathcal{F}_{t+1}$-measurable random costs and ρ_t are one-step conditional risk measures.

Running risk-to-go satisfies dynamic programming equations:

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

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

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]

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

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

We aim to optimise a class of dynamic robust distortion risk measure with uncertainty sets induced by the conditional 2-Wasserstein distance

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

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]

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

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

We aim to optimise a class of dynamic robust distortion risk measure with uncertainty sets induced by the conditional 2-Wasserstein distance

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

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]

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

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

We aim to optimise a class of dynamic robust distortion risk measure with uncertainty sets induced by the conditional 2-Wasserstein distance

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

Analytical Worst-Case Distribution

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

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

The optimal quantile function is then given by

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

where λ^* and b_{λ^*} depend non-trivially on the quantile function $\breve{F}_{Y^\theta_t}$.

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

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

The gradient of the value function is given by

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

↪ 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 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:

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

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

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

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:

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

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

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

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:

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

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

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

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

↳ Convex optimisation over the space of quantile functions

Experimental Setup

Consider a market with multiple assets, where an agent

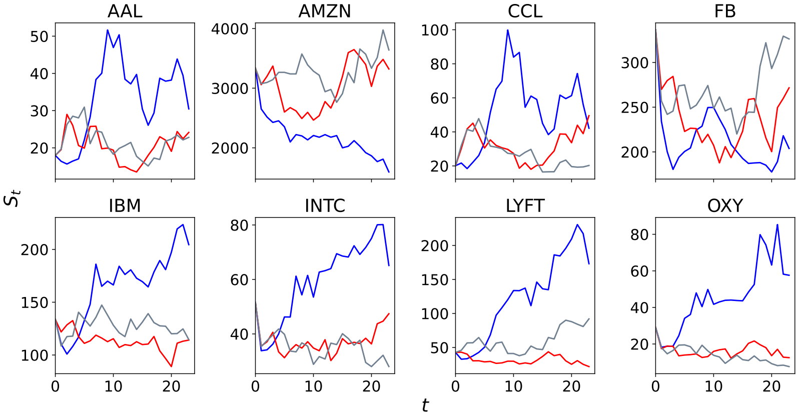
- ↳ observes the time and asset prices
- ↳ decides on the proportion of wealth to invest in each asset
- ↳ receives feedback from P&L differences
- ↳ assumes a null interest rate, no leveraging nor short-selling

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

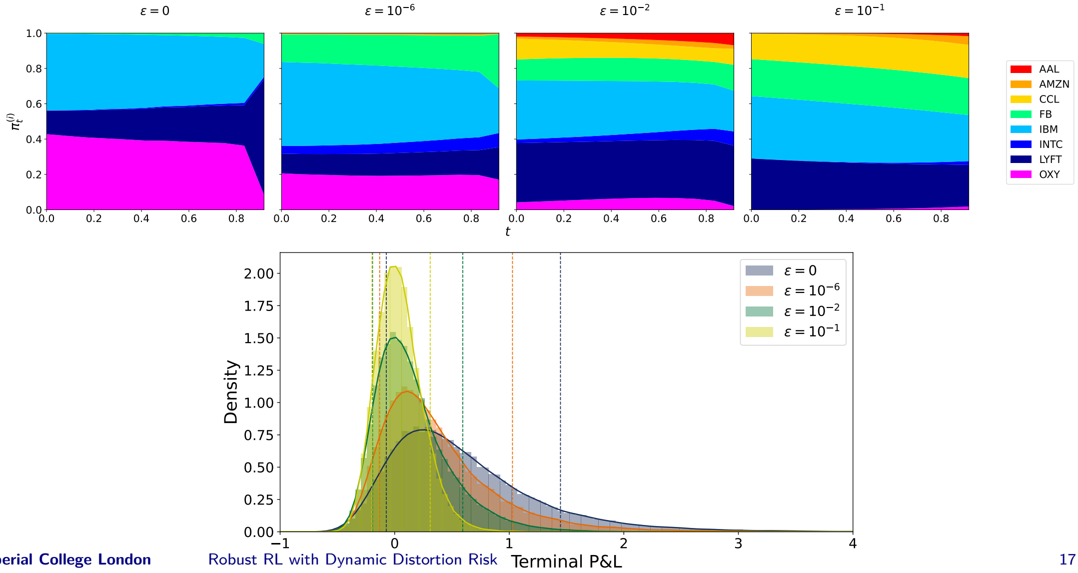
$$\Delta S_\tau = \alpha\beta^\intercal S_{\tau-1} + \Gamma_1 \Delta S_{\tau-1} + \cdots + \Gamma_{k_{ar}-1} \Delta S_{\tau-k_{ar}+1} + C D_\tau + u_\tau$$

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



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

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*.

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

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

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.

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

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.