Robust Reinforcement Learning for Dynamic Risk Measures

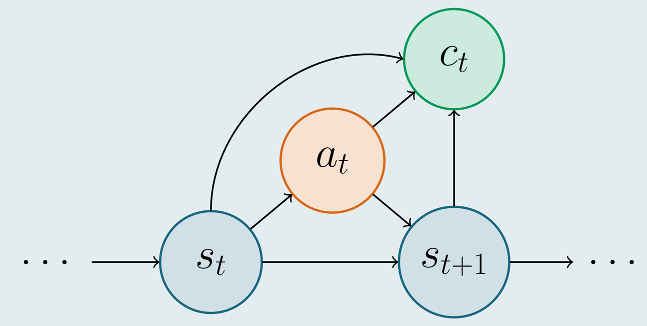
Anthony Coache

Statistical Sciences
UNIVERSITY OF TORONTO

The Problem

Reinforcement learning (RL) is a model-agnostic framework for learning-based control. The agent aims to discover the best possible actions based on a certain criterion by updating its behavior according to its experience. At each period, the agent:

- begins in a state $s_t \in \mathcal{S}$
- takes an action $a_t \in \mathcal{A}$ according to a deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$
- moves to a next state s_{t+1} and receives a cost $c_t = c(s_t, a_t, s_{t+1})$



Main issues:

- Real-world uncertainty may result in algorithms optimized on training models to perform poorly during testing.
- Optimizing static risk measures in sequential decision making problems leads to optimal precommitment strategies, i.e. they are time-inconsistent.

How can we simultaneously (i) robustify the actions against the uncertainty of the environment and (ii) account for risk in a time-consistent manner in RL problems?

Previous work

Robust RL: via KL divergence (Smirnova et al., 2019), Wasserstein distance (Abdullah et al., 2019; Jaimungal et al., 2022), Bayesian perspective (Bielecki et al., 2022)

Time-consistent RL: with dynamic spectral (Coache et al., 2022), expectile (Marzban et al., 2021), distortion (Jaimungal et al., 2023) risk measures

Goal: To the best of our knowledge, no RL methodology bridges the gaps between these works, that is:

- a deep RL algorithm
- optimization of time-consistent dynamic risk measures
- robustification against environmental uncertainty

Elicitability

Elicitable mappings admit the existence of a loss function that can be used as a penalizer when updating their point estimate:

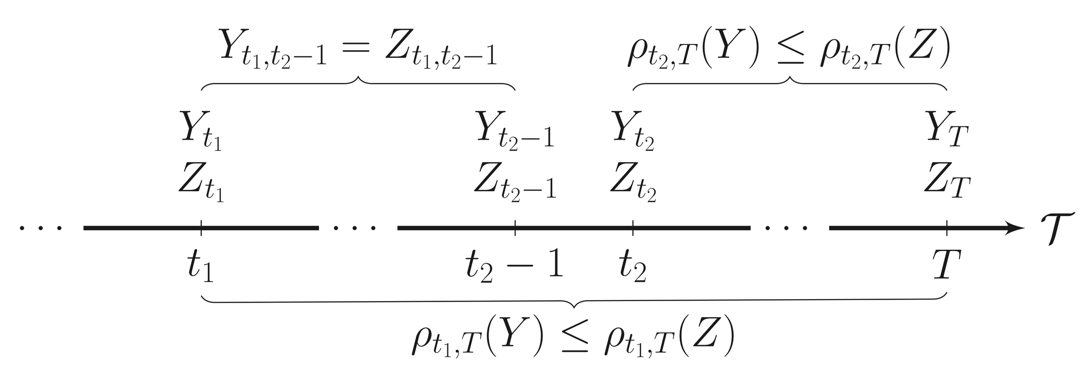
- $\mathbb{E}[Y] = \underset{\mathfrak{a} \in \mathbb{R}}{\arg\min}\, \mathbb{E}\Big[(\mathfrak{a} - Y)^2\Big]$
- $\Big(\text{VaR}_\alpha(Y), \text{CVaR}_\alpha(Y)\Big) = \underset{(\mathfrak{a}_1, \mathfrak{a}_2) \in \mathbb{R}^2}{\arg\min}\, \mathbb{E}\Big[S(\mathfrak{a}_1, \mathfrak{a}_2, Y)\Big]$
- Conditional maps
$$\rho(Y \mid s_t = s) = \underset{h : \mathcal{S} \to \mathbb{R}}{\arg\min}\, \mathbb{E}\Big[S(h(s), Y)\Big]$$
- Cumulative distribution functions
$$F_Y = \underset{F \in \mathbb{F}}{\arg\min}\, \mathbb{E}\left[\int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y \geq Y}\Big)^2 \mathrm{d}y\right]$$

Dynamic risk measures

Let $\mathcal{T} := \{0, \ldots, T\}$ denote a sequence of periods, and define $\mathcal{Y}_{t_1,t_2} := \mathcal{Y}_{t_1} \times \cdots \times \mathcal{Y}_{t_2}$ as the space of sequences of $\mathcal{F}_t$-measurable random costs.

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

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



Objective function: We optimize dynamic risk measures

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

with a class of robust distortion one-step risk measures

$$\rho_t(Y) = \sup_{Y^\phi \in \varphi_Y^{\epsilon_t}} \mathbb{E}\left[Y^\phi\, \gamma_t\Big(F_{Y^\phi|_{\mathcal{F}_t}}(Y^\phi)\Big) \,\Big|\, \mathcal{F}_t\right]$$
$$\varphi_Y^\epsilon = \{Y^\phi \in \mathcal{Y}_{t+1} \,:\, \text{2-Wass}(Y^\phi|_{\mathcal{F}_t}, Y|_{\mathcal{F}_t}) \leq \epsilon_t\}$$

This class of risk measures:

- takes into account the uncertainty
- allows risk-averse and risk-seeking behaviors
- is elicitable
- is time-consistent

Setup

Time-consistency leads to a dynamic programming principle. We want to minimize the running risk-to-go $Q_t(s, \pi^\theta(s); \theta)$ over policies θ:

$$Q_t(s, a; \theta) = \sup_{Y_t^\phi \in \varphi_{Y_t^\theta}^{\epsilon_t}} \mathbb{E}\left[Y_t^\phi\, \gamma_{t,s}\Big(F_{Y_t^\phi|_{s_t=s,a_t=a}}(Y_t^\phi)\Big) \,\middle|\, \begin{matrix} s_t = s \\ a_t = a \end{matrix}\right]$$

with costs-to-go $Y_t^\theta := c(s_t, a_t, s_{t+1}) + Q_{t+1}(s_{t+1}, \pi^\theta(s); \theta)$.

Following the work from Pesenti and Jaimungal (2020), the quantile reformulation

$$\sup_{\breve{F}_\phi \in \varphi^{\epsilon_t}_{\breve{F}_{Y_t^\theta}(\cdot|s,a)}} \int_0^1 \gamma_{t,s}(u) \breve{F}_\phi(u|s,a) \mathrm{d}u$$

leads to an equivalent convex optimization problem. It aids in obtaining a closed-form formula of the optimal $\breve{F}_\phi$.

Main results

Proposition 1: The quantile function of the optimal random variable in $Q_t(s, a; \theta)$ is given by

$$\breve{F}_\phi^*(\cdot|s,a) = \left(\breve{F}_{Y_t^\theta}(\cdot|s,a) + \frac{\gamma_{t,s}(\cdot)}{2\lambda^*}\right)^{\uparrow},$$

where $\lambda^* > 0$ is such that

$$\int_0^1 \left|\breve{F}_\phi^*(u|s,a) - \breve{F}_{Y_t^\theta}(u|s,a)\right|^2 \mathrm{d}u = \epsilon_{t,s}^2.$$

Proposition 2: Using the deterministic policy gradient (Silver et al., 2014),

$$\nabla_\theta \mathbb{E}\Big[Q_t(s, \pi^\theta(s); \theta) \,\Big|\, s_t = s\Big]$$
$$= \mathbb{E}\left[\nabla_\theta \pi^\theta(s) \nabla_a Q_t(s, a; \theta)\Big|_{a=\pi^\theta(s)} \,\middle|\, s_t = s\right].$$

Algorithm

We propose an actor-critic-adversary style algorithm, analogous to the DDPG algorithm, with feed-forward neural nets for the 3 components:

Actor: update the current policy π^θ

- Off-policy deterministic policy gradient

Critic: estimate $Q_t(s, a; \theta)$ of the current policy

- Optimal distribution from Proposition 1
- Strictly consistent score for distortion risk

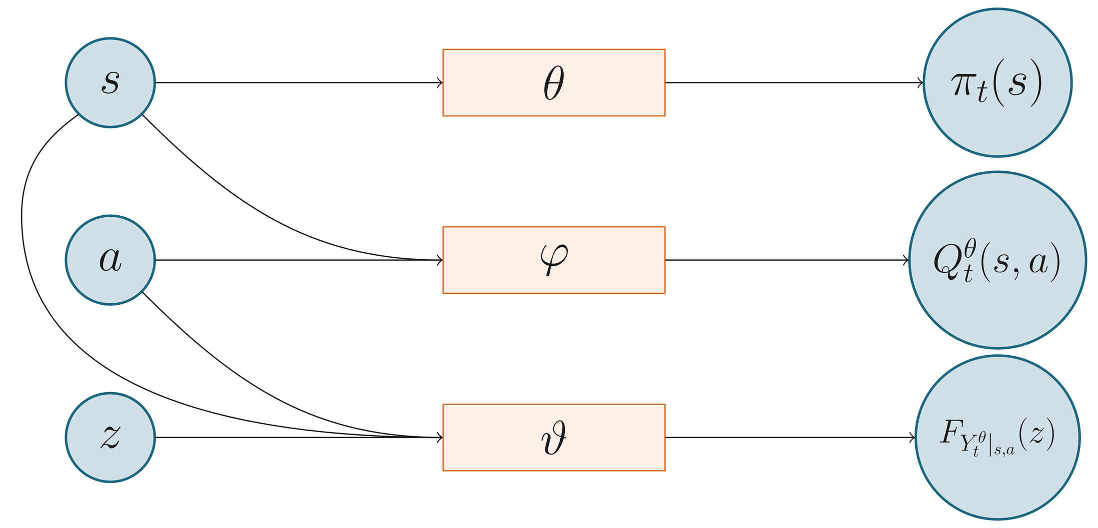
Adversary: estimate the CDF F of Y_t^θ

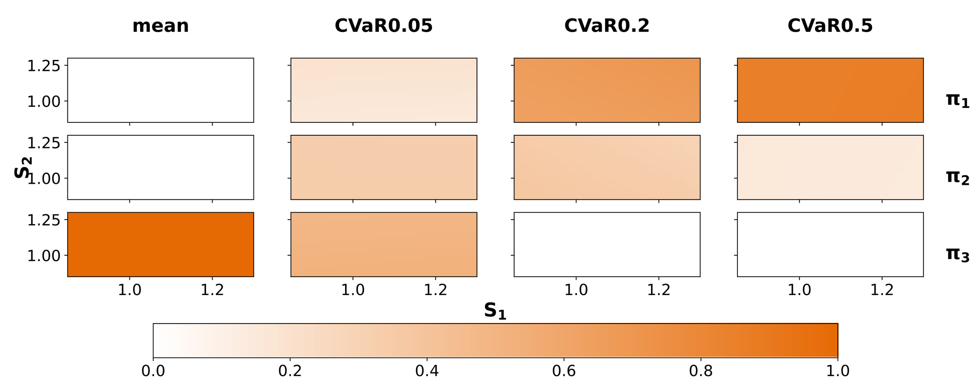
- Expected scoring rule for CDF
- Monotonicity penalty

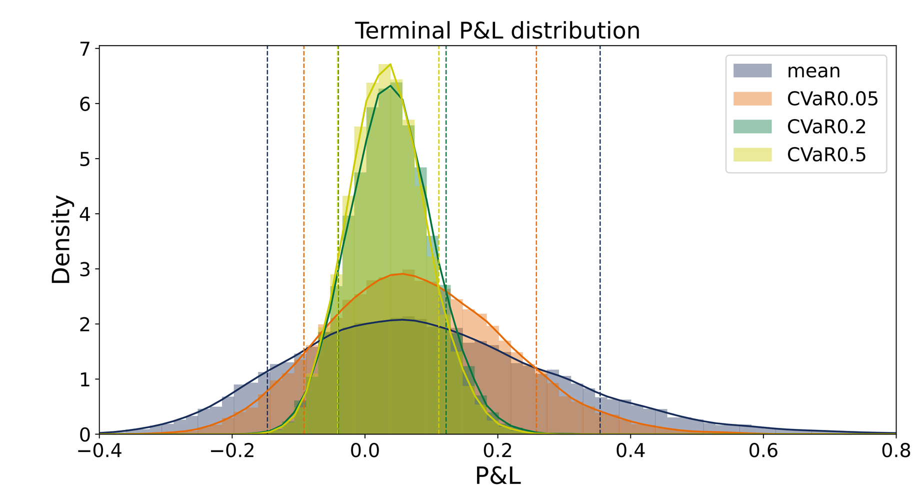
Preliminary results

Portfolio allocation problem on a market of correlated GBMs with respectively

- drifts of $\boldsymbol{\mu} = [0.03; 0.06; 0.09]$
- volatilities of $\boldsymbol{\sigma} = [0.06; 0.12; 0.18]$







References

Thank you for your attention! Preprint available soon.
For more information, please visit: `https://anthonycoache.ca/`

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*.

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*.

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

Jaimungal, S., Pesenti, S. M., Saporito, Y. F., and Targino, R. S. (2023). Risk budgeting allocation for dynamic risk measures. *arXiv preprint arXiv:2305.11319*.

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. (2021). Deep reinforcement learning for equal risk pricing and hedging under dynamic expectile risk measures. *arXiv preprint arXiv:2109.04001*.

Pesenti, S. and Jaimungal, S. (2020). Portfolio optimisation within a Wasserstein ball. *arXiv preprint arXiv:2012.04500*.

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*.

Fields-CFI Conference on Recent Advances in Mathematical Finance and Insurance ⋆ September 25–27, 2023 ⋆ Fields Institute, Toronto, Canada