Conditionally Elicitable Dynamic Risk Measures for Deep Reinforcement Learning

Anthony Coache (University of Toronto)

Joint work with

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

and

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

SIAG/FME Conference Paper Prize Session ⋆ June 9, 2023 ⋆ Philadelphia, USA

Statistical Sciences UNIVERSITY OF TORONTO

UNIVERSITY OF OXFORD

OXFORD-MAN INSTITUTE

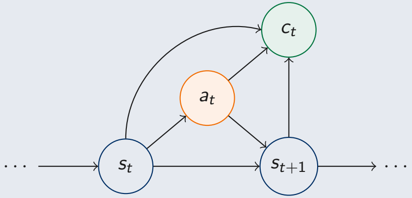
siam 2023 Conference on Financial Mathematics and Engineering

Reinforcement Learning (RL)

- Model-agnostic framework for learning-based control
- Learning optimal behaviours from interactions to minimise a cost signal

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

- $\mathcal{S}$ – State space
- $\mathcal{A}$ – Action space
- $\pi^\theta(a_t|s_t)$ – Randomised policy characterised by θ
- $\mathbb{P}(s_0), \mathbb{P}(s_{t+1}|s_t, a_t)$ – Transition probabilities
- $c_t(s_t, a_t, s_{t+1})$ – Cost function



Risk-Aware RL

Standard RL aims at minimising problems of the form: $\min_\theta \mathbb{E}\left[Y^\theta\right]$, where $Y^\theta = \sum_t c_t^\theta$

- Ignores the risk of the costs!

Risk-aware RL with static risk measures, 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!

Recent approaches to overcome the time-consistency issue, e.g.:

- *Dynamic risk measures* [Marzban et al., 2021; Coache and Jaimungal, 2021], *conditional risk mappings* [Cheng and Jaimungal, 2022], *recursive risk filters* [Bielecki et al., 2022]...

In this paper, we:

- develop a computational approach to solve RL problems with dynamic risk
- devise an efficient deep estimation method for elicitable dynamic risk measures
- prove that these dynamic risk measures may be approximated to an arbitrary accuracy using NNs

Risk-Aware RL

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

- Ignores the risk of the costs!

Risk-aware RL with static risk measures, 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!

Recent approaches to overcome the time-consistency issue, e.g.:

- *Dynamic risk measures* [Marzban et al., 2021; Coache and Jaimungal, 2021], *conditional risk mappings* [Cheng and Jaimungal, 2022], *recursive risk filters* [Bielecki et al., 2022]...

In this paper, we:

- develop a computational approach to solve RL problems with dynamic risk
- devise an efficient deep estimation method for elicitable dynamic risk measures
- prove that these dynamic risk measures may be approximated to an arbitrary accuracy using NNs

Time-Consistent Dynamic Risk

- $\mathcal{Y}_{t_1,t_2} := \mathcal{Y}_{t_1} \times \cdots \times \mathcal{Y}_{t_2}$ – Sequence of $\mathcal{F}_t$-measurable random costs on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_t, \mathbb{P})$
- $\rho_{t,T} : \mathcal{Y}_{t,T} \to \mathcal{Y}_t$ – Dynamic risk measure $\{\rho_{t,T}\}_t$
- *Strong time-consistency* – 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 $\rho_{t_1,T}(Y_{t_1}, \dots, Y_T) \leq \rho_{t_1,T}(Z_{t_1}, \dots, Z_T)$.

[Thm. 1, Ruszczyński, 2010]

Let $\{\rho_{t,T}\}_t$ be a monotone, cash-additive, and normalised dynamic risk measure.
Then $\{\rho_{t,T}\}_t$ is strongly time-consistent iff it may be expressed with one-step conditional risk measures $\rho_t : \mathcal{Y}_{t+1} \to \mathcal{Y}_t$ as

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

Time-Consistent Dynamic Risk

- $\mathcal{Y}_{t_1,t_2} := \mathcal{Y}_{t_1} \times \cdots \times \mathcal{Y}_{t_2}$ – Sequence of $\mathcal{F}_t$-measurable random costs on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_t, \mathbb{P})$
- $\rho_{t,T} : \mathcal{Y}_{t,T} \to \mathcal{Y}_t$ – Dynamic risk measure $\{\rho_{t,T}\}_t$
- *Strong time-consistency* – 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 $\rho_{t_1,T}(Y_{t_1}, \ldots, Y_T) \leq \rho_{t_1,T}(Z_{t_1}, \ldots, Z_T)$.

[Thm. 1, Ruszczyński, 2010]

Let $\{\rho_{t,T}\}_t$ be a monotone, cash-additive, and normalised dynamic risk measure.
Then $\{\rho_{t,T}\}_t$ is strongly time-consistent iff it may be expressed with one-step conditional risk measures $\rho_t : \mathcal{Y}_{t+1} \to \mathcal{Y}_t$ as

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

Problem Setup

Problems of the form

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

where c_t^θ are $\mathcal{F}_{t+1}$-measurable random costs and ρ_t's are static risk measures.

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

Problem Setup

Problems of the form

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

where c_t^θ are $\mathcal{F}_{t+1}$-measurable random costs and ρ_t's are static risk measures.

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

Policy Gradient

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

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

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

Under some regularity assumptions, the gradient of the value function at any period $t \in \mathcal{T}$ and any state $s \in \mathcal{S}$ for dynamic spectral risk measures with finite support spectrum is

$$\nabla_\theta V_t(s;\theta) = \sum_{m=1}^{k-1} \frac{p_m}{1-\alpha_m} \mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t=s)} \left[\left(c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta) - \lambda_m^* \right)_+ \left(\nabla_\theta \log \pi^\theta(a|s_t) \Big|_{a=a_t^\theta} \right) \right]$$
$$+ \mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t=s)} \left[\left(\nabla_\theta V_{t+1}(s';\theta) \Big|_{s'=s_{t+1}^\theta} \right) \xi_m^*(a_t^\theta, s_{t+1}^\theta) \right],$$

Actor-critic style algorithm composed of two interleaved procedures:

- Value function estimation given a policy
- Policy update given a value function

Policy Gradient

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

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

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

Under some regularity assumptions, the gradient of the value function at any period $t \in \mathcal{T}$ and any state $s \in \mathcal{S}$ for dynamic spectral risk measures with finite support spectrum is

$$\nabla_\theta V_t(s;\theta) = \sum_{m=1}^{k-1} \frac{p_m}{1-\alpha_m} \mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t=s)}\left[\left(c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta) - \lambda_m^* \right)_+ \left(\nabla_\theta \log \pi^\theta(a|s_t) \Big|_{a=a_t^\theta} \right) \right]$$
$$+ \mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t=s)}\left[\left(\nabla_\theta V_{t+1}(s';\theta) \Big|_{s'=s_{t+1}^\theta} \right) \xi_m^*(a_t^\theta, s_{t+1}^\theta) \right],$$

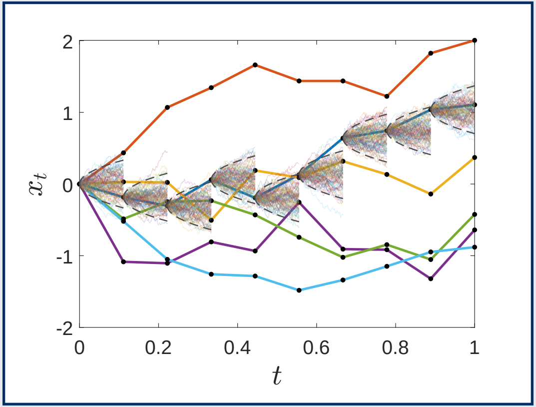
Actor-critic style algorithm composed of two interleaved procedures:

- Value function estimation given a policy
- Policy update given a value function

Estimation of V

Previous approaches: nested simulations [Tamar et al., 2016; Coache and Jaimungal, 2021]

- Generate (outer) episodes and (inner) transitions for every visited state
- Computationally expensive...



Elicitability

We leverage the elicitability to efficiently estimate dynamic risk measures

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
$S(\mathfrak{a}, y)$	$(\mathfrak{a} - y)^2$	$\lvert\mathfrak{a} - y\rvert$	$\mathbb{1}_{\mathfrak{a} \leq y} - \alpha$	$\emptyset$

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

- Elicitability of (static) spectral risk measures [Fissler and Ziegel, 2016]
- Characterization of their scoring function S

Elicitability

We leverage the elicitability to efficiently estimate dynamic risk measures

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
$S(\mathfrak{a}, y)$	$(\mathfrak{a} - y)^2$	$\lvert\mathfrak{a} - y\rvert$	$\mathbb{1}_{\mathfrak{a} \leq y} - \alpha$	$\emptyset$

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

- Elicitability of (static) spectral risk measures [Fissler and Ziegel, 2016]
- Characterization of their scoring function S

Conditional Elicitability

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

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

In our RL problem, the costs are supported by observed features, i.e. the states $s \in \mathcal{S}$

$$\Big(\mathrm{VaR}_\alpha(Y|s_t = s), \mathrm{CVaR}_\alpha(Y|s_t = s)\Big) = \underset{h_1,h_2\,:\,\mathcal{S}\to\mathbb{R}}{\arg\min}\; \mathbb{E}_{Y\sim F_Y}\Big[S(h_1(s), h_2(s), Y)\Big]$$

- Model $V_t(s;\theta)$ with NNs $H_t^\psi(s), V_t^\phi(s)$
- Use empirical estimates based on observed data

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

Similar results for a class of spectral risk measures

Conditional Elicitability

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

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

In our RL problem, the costs are supported by observed features, i.e. the states $s \in \mathcal{S}$

$$\Big(\mathrm{VaR}_\alpha(Y|s_t = s), \mathrm{CVaR}_\alpha(Y|s_t = s)\Big) = \underset{h_1, h_2 : \mathcal{S} \to \mathbb{R}}{\arg\min}\ \mathbb{E}_{Y \sim F_Y}\Big[S(h_1(s), h_2(s), Y)\Big]$$

- Model $V_t(s; \theta)$ with NNs $H_t^\psi(s), V_t^\phi(s)$
- Use empirical estimates based on observed data

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

Similar results for a class of spectral risk measures

Conditional Elicitability

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

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

In our RL problem, the costs are supported by observed features, i.e. the states $s \in \mathcal{S}$

$$\Big(\mathrm{VaR}_\alpha(Y|s_t = s), \mathrm{CVaR}_\alpha(Y|s_t = s)\Big) = \underset{h_1,h_2\,:\,\mathcal{S}\to\mathbb{R}}{\arg\min}\ \mathbb{E}_{Y\sim F_Y}\Big[S(h_1(s), h_2(s), Y)\Big]$$

- Model $V_t(s;\theta)$ with NNs $H_t^\psi(s), V_t^\phi(s)$
- Use empirical estimates based on observed data

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

Similar results for a class of spectral risk measures

Accuracy of the Elicitable Approach

- We can approximate the value function to an arbitrary accuracy using this framework

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

Suppose π^θ is a fixed policy, with its corresponding value function $V_t(s;\theta)$. Then for any $\varepsilon_1^*, \dots, \varepsilon_k^* > 0$, there exist NNs denoted $H_{1,t}^{\psi_1}, \dots, H_{k,t}^{\psi_k}$ such that for any $t \in \mathcal{T}$, we have

$$\operatorname*{ess\,sup}_{s \in \mathcal{S}} \left\| V_t(s;\theta) - \left(H_{k,t}^{\psi_k}(s;\theta) + \sum_{m=1}^{k-1} p_m \sum_{l=1}^{m} H_{l,t}^{\psi_l}(s;\theta) \right) \right\| < \varepsilon^*.$$

Portfolio Allocation

Consider a market with d assets. An agent

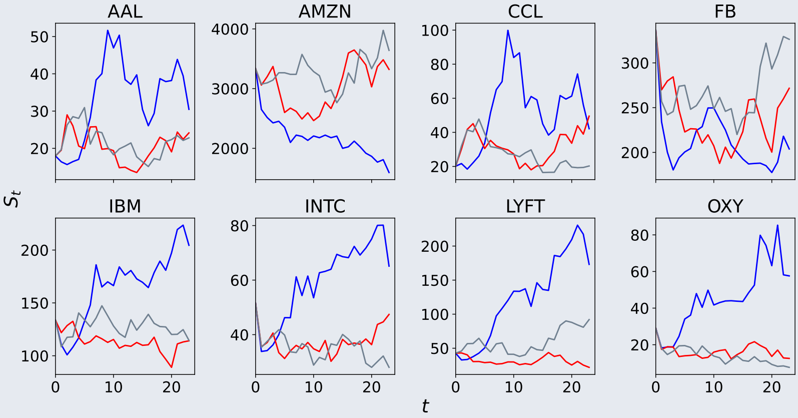
- observes the time t and asset prices $\{S_t^{(i)}\}_{i=1,\dots,d}$
- decides on the proportion of its wealth $\pi_t^{(i)}$ to invest in asset i
- receives feedback from P&L differences $y_t - y_{t+1}$, where its wealth y_t varies according to

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

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

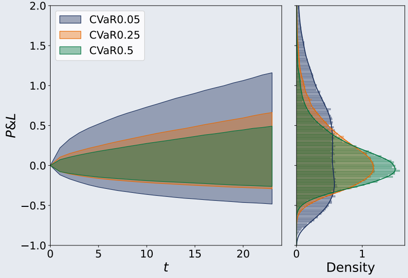
Portfolio Allocation

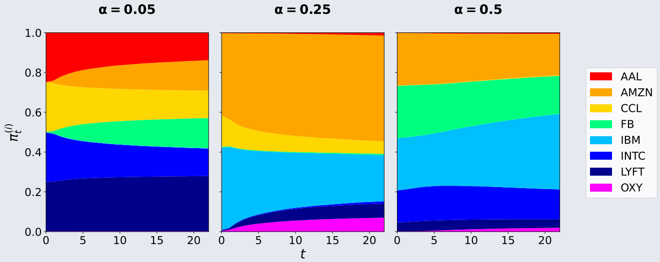
Co-integration model using daily data from eight different stocks listed on the NASDAQ exchange between September 31, 2020 and December 31, 2021.



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.





Contributions

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

- Novel setting utilising *elicitability* for efficient & accurate estimation
- Performance validation on several benchmark optimisation problems

Future directions

- Robustification to protect against model uncertainty
- DDPG approach for dynamic risk measures
- Risk-aware dynamic RL for multi-agent systems

Thank you!

Paper, code and slides available at: `anthonycoache.ca`

References

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.

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.

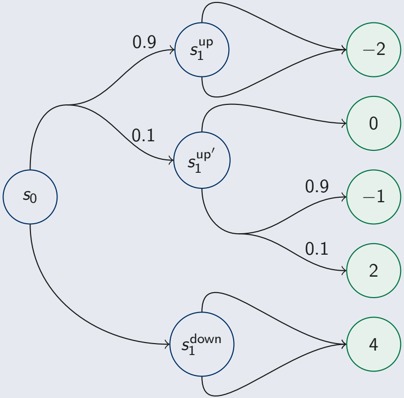
Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. (2016). Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338.

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 initial optimal strategy...



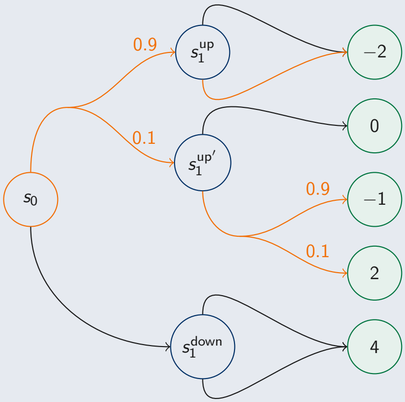
Optimizing static risk measures leads to optimal precommitment policies!

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 initial optimal strategy...



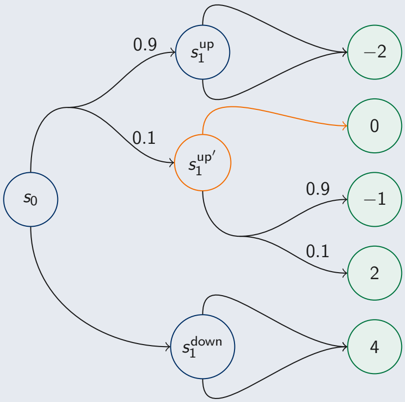
Optimizing static risk measures leads to optimal precommitment policies!

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 initial optimal strategy...



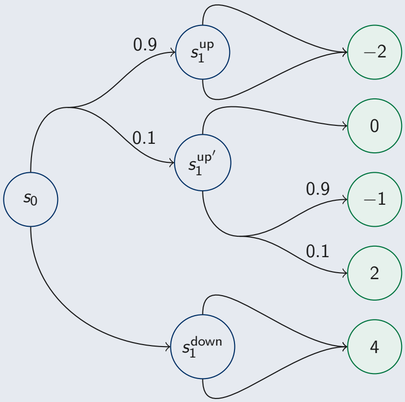
Optimizing static risk measures leads to optimal precommitment policies!

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 initial optimal strategy...



Optimizing static risk measures leads to optimal precommitment policies!

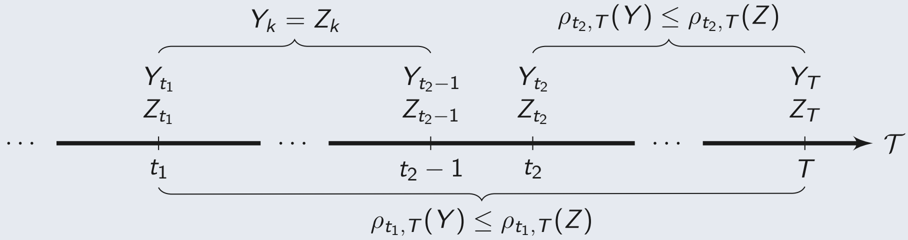
Time-Consistency

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



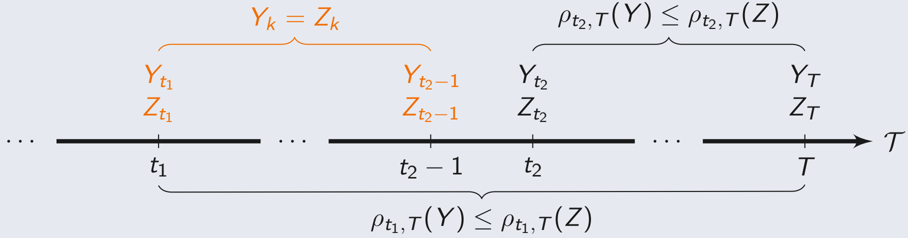
Time-Consistency

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

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

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



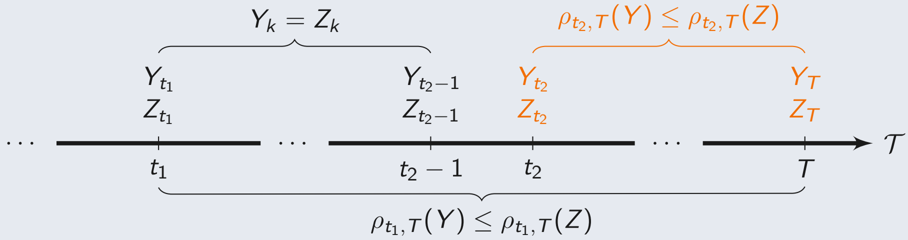
Time-Consistency

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



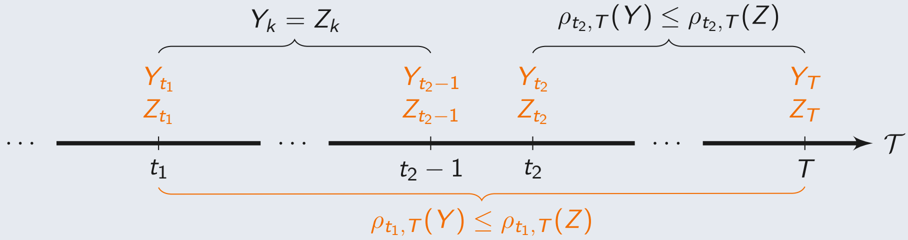
Time-Consistency

Strong time-consistency

$\{\rho_{t,T}\}_t$ is *strongly time-consistent* iff for any $Y, Z \in \mathcal{Y}_{t_1,T}$ and $0 \leq t_1 < t_2 \leq T$, 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 $\rho_{t_1,T}(Y_{t_1}, \ldots, Y_T) \leq \rho_{t_1,T}(Z_{t_1}, \ldots, Z_T)$.



Algorithms

Algorithm 1: Actor-critic algorithm – Elicitable approach

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

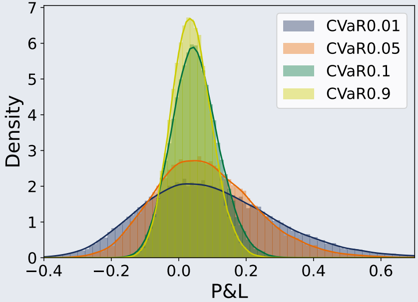
```
Set initial learning rates for φ, θ;
for each iteration k = 1, ..., K do
    for each epoch k^φ = 1, ..., K^φ do
        Simulate a mini-batch of B^φ episodes induced by π^θ;
        Compute the loss L(φ): minimization of the expected consistent score;
        Update φ by performing an Adam optimisation 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;
```

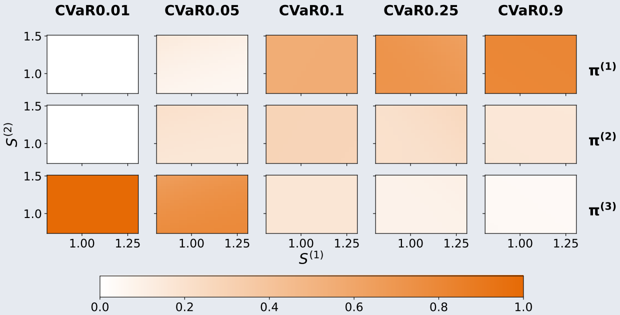
Output: Optimal policy π^θ and its value function V^ϕ

Portfolio Allocation

$$\mathrm{d}S_t^{(i)} = \mu^{(i)} S_t^{(i)} \mathrm{d}t + \sigma^{(i)} S_t^{(i)} \mathrm{d}W_t^{(i)}$$

Drifts and volatilities are $\mu = [0.03; 0.06; 0.09]$ and $\sigma = [0.06; 0.12; 0.18]$





Portfolio Allocation

$$\mathrm{d}X_t^{(i)} = -\kappa X_t^{(i)}\mathrm{d}t + \sigma^{(i)}\mathrm{d}W_t^{(i)} \quad \text{with} \quad S_t^{(i)} = e^{X_t^{(i)}+\mu^{(i)}t-(\sigma^{(i)})^2\frac{1-e^{-2\kappa t}}{4\kappa}}$$

Drifts and volatilities are $\mu = [0.03; 0.06; 0.09]$ and $\sigma = [0.06; 0.12; 0.18]$