Reinforcement Learning with Dynamic Risk Measures

Anthony Coache

anthonycoache.ca

Joint work with

Sebastian Jaimungal

and

Álvaro Cartea

2022 INFORMS Annual Meeting ⋆ Indianapolis, US ⋆ October 18, 2022

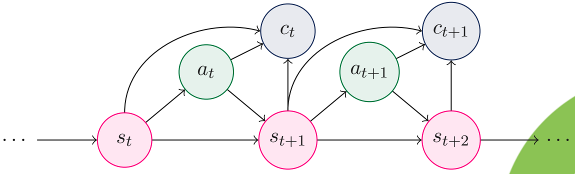
Reinforcement Learning (RL)

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

Standard RL: $\min_\theta \mathbb{E}\left[\{c_t^\theta\}_t\right]$

Risk-aware RL: $\min_\theta \rho\left(\{c_t^\theta\}_t\right)$



Risk-Sensitive RL

Risk-aware RL: applying risk measures *recursively* [e.g. Rus10]

- Offers a *remedy to environment uncertainty*
- Provides *time-consistent* optimal strategies
- Tuned to *agent's risk preference*

Several policy search algorithms in the dynamic framework

- [TCGM16] studies *stationary policies*, restricted to *coherent risk* measures
- [MDL21] proposes ad hoc actor-critic algorithm for *dynamic expectile risk*

We develop a generalized, practical setting to solve a wider class of RL problems

- Considers *non-stationary policies*
- Extended to dynamic *convex* risk measures
- Improved algorithm for *elicitable* dynamic risk measures

Risk-Sensitive RL

Risk-aware RL: applying risk measures *recursively* [e.g. Rus10]

- Offers a *remedy to environment uncertainty*
- Provides *time-consistent* optimal strategies
- Tuned to *agent's risk preference*

Several policy search algorithms in the dynamic framework

- [TCGM16] studies *stationary policies*, restricted to *coherent risk* measures
- [MDL21] proposes ad hoc actor-critic algorithm for *dynamic expectile risk*

We develop a generalized, practical setting to solve a wider class of RL problems

- Considers *non-stationary policies*
- Extended to dynamic *convex* risk measures
- Improved algorithm for *elicitable* dynamic risk measures

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
- $\mathcal{Y}_{t,T} := \mathcal{Y}_t \times \cdots \times \mathcal{Y}_T$ – Sequence of random variables

Dynamic risk measure $\{\rho_{t,T}\}_t$

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

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

Time-Consistency

Time-consistency

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

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

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, Rus10]

Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic risk measure satisfying for any $Y \in \mathcal{Y}_{t,T}$, $t \in \mathcal{T}$

$$\rho_{t,T}(Y_t, Y_{t+1}, \ldots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \ldots, Y_T) \quad \text{and} \quad \rho_{t,T}(0, \ldots, 0) = 0.$$

Then $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ is time-consistent iff for any $0 \leq t_1 \leq t_2 \leq T$ 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

Time-consistency

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

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

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

[Thm. 1, Rus10]

Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic risk measure satisfying for any $Y \in \mathcal{Y}_{t,T},\ t \in \mathcal{T}$

$$\rho_{t,T}(Y_t, Y_{t+1}, \dots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \dots, Y_T) \quad \text{and} \quad \rho_{t,T}(0, \dots, 0) = 0.$$

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

Time-consistency

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

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

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

[Thm. 1, Rus10]
Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic risk measure satisfying for any $Y \in \mathcal{Y}_{t,T}$, $t \in \mathcal{T}$

$$\rho_{t,T}(Y_t, Y_{t+1}, \dots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \dots, Y_T) \text{ and } \rho_{t,T}(0, \dots, 0) = 0.$$

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

Time-consistency

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

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

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

[Thm. 1, Rus10]
Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic risk measure satisfying for any $Y \in \mathcal{Y}_{t,T},\ t \in \mathcal{T}$

$$\rho_{t,T}(Y_t, Y_{t+1},\dots,Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1},\dots,Y_T) \ \text{ and } \ \rho_{t,T}(0,\dots,0) = 0.$$

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

Time-consistency

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

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

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

[Thm. 1, Rus10]
Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic risk measure satisfying for any $Y \in \mathcal{Y}_{t,T},\ t \in \mathcal{T}$

$$\rho_{t,T}(Y_t, Y_{t+1}, \dots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \dots, Y_T) \text{ and } \rho_{t,T}(0, \dots, 0) = 0.$$

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

Time-consistency

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

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

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

[Thm. 1, Rus10]
Let $\{\rho_{t,T}\}_{t\in\mathcal{T}}$ be a dynamic risk measure satisfying for any $Y \in \mathcal{Y}_{t,T}, \ t \in \mathcal{T}$

$$\rho_{t,T}(Y_t, Y_{t+1}, \ldots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \ldots, Y_T) \text{ and } \rho_{t,T}(0, \ldots, 0) = 0.$$

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

Recursive relationship for time-consistent dynamic risk

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

- Axioms of convex risk measures [FS02]: monotone, translation invariant and convex
- Markovian: not allowed to depend on the whole past

Problem Setup

Problems of the form

$$\min_{\theta} \rho_{0,T}\Big(\{c_t^\theta\}_{t\in\mathcal{T}}\Big) = \min_{\theta} \rho_0\bigg(c_0^\theta + \rho_1\Big(c_1^\theta + \cdots + \rho_{T-2}\Big(c_{T-2}^\theta + \rho_{T-1}\big(c_{T-1}^\theta\big)\Big)\cdots\Big)\bigg)$$

where $c_t^\theta := c(s_t, a_t^\theta, s_{t+1}^\theta)$ 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 $\mathbb{P}^\theta(a, s'|s_t = s) = \mathbb{P}(s'|s,a)\pi^\theta(a|s_t = s)$

Problem Setup

Problems of the form

$$\min_{\theta} \rho_{0,T}\Big(\{c_t^\theta\}_{t\in\mathcal{T}}\Big) = \min_{\theta} \rho_0\Bigg(c_0^\theta + \rho_1\bigg(c_1^\theta + \cdots + \rho_{T-2}\Big(c_{T-2}^\theta + \rho_{T-1}\big(c_{T-1}^\theta\big)\Big)\cdots\bigg)\Bigg)$$

where $c_t^\theta := c(s_t, a_t^\theta, s_{t+1}^\theta)$ 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 [CJ21]

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 [KT00] 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 [CJ21]

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 [KT00] 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 [CJ21]

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

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

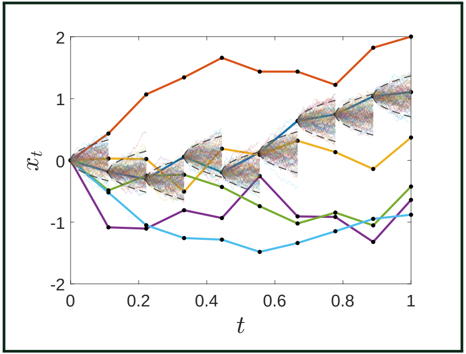
Actor-critic style algorithm [KT00] composed of two interleaved procedures:

- *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 [CJ21]

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

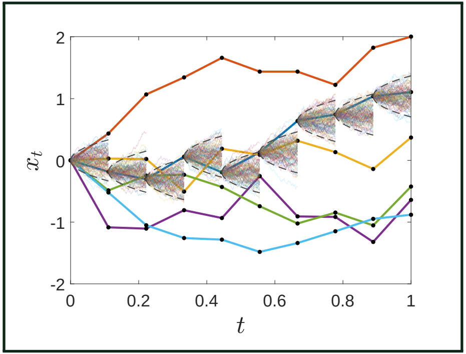


Elicitable approach [CJC22]

- *Conditional elicitability* of dynamic spectral risk measures [FZ16]
- Avoids nested simulations, *memory efficient*

We derive universal approximation theorems for $V_t(s;\theta)$ in both cases

Estimation of V



Nested simulation approach [CJ21]

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

Elicitable approach [CJC22]

- *Conditional elicitability* of dynamic spectral risk measures [FZ16]
- Avoids nested simulations, *memory efficient*

We derive universal approximation theorems for $V_t(s;\theta)$ in both cases

Elicitability

Elicitable mapping [Gne11]

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

$$M(Y) = \arg\min_{\mathfrak{a} \in \mathbb{A}} \mathbb{E}_{Y \sim F}\Big[S(\mathfrak{a}, Y)\Big].$$

Conditional elicitability from [Osb85]. Recently, [FZ16]:

- showed that $M(Y) = (\mathrm{VaR}_\alpha(Y), \mathrm{CVaR}_\alpha(Y))$ is elicitable
- characterized the scoring function S

Modeling of $V_t(s;\theta)$ with ANNs $H_t^\psi(s), V_t^\phi(s)$; empirical estimates based on observed data

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

Similar results for a class of spectral risk measures

Elicitability

Elicitable mapping [Gne11]

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

$$M(Y) = \arg\min_{\mathfrak{a} \in \mathbb{A}} \mathbb{E}_{Y \sim F}\Big[S(\mathfrak{a}, Y)\Big].$$

Conditional elicitability from [Osb85]. Recently, [FZ16]:

- showed that $M(Y) = \big(\mathrm{VaR}_\alpha(Y),\, \mathrm{CVaR}_\alpha(Y)\big)$ is elicitable
- characterized the scoring function S

Modeling of $V_t(s;\theta)$ with ANNs $H_t^\psi(s), V_t^\phi(s)$; empirical estimates based on observed data

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

Similar results for a class of spectral risk measures

Dynamic Risk Measures

We consider the following one-step conditional risk measures:

- Expectation: $\rho_{\mathbb{E}}(Y) = \mathbb{E}[Y]$
- Conditional value-at-risk: $\rho_{\text{CVaR}}(Y;\alpha) = \sup_{\xi \in \mathcal{U}(\mathbb{P})} \left\{ \mathbb{E}^{\xi}[Y] \right\}$
- Penalized CVaR: $\rho_{\text{CVaR-p}}(Y;\alpha,\kappa) = \sup_{\xi \in \mathcal{U}(\mathbb{P})} \left\{ \mathbb{E}^{\xi}[Y] - \kappa \mathbb{E}^{\xi}[\log \xi] \right\}$

where

$$\mathcal{U}(\mathbb{P}) = \left\{ \xi : \sum_{\omega} \xi(\omega)\mathbb{P}(\omega) = 1,\ \xi \in \left[0, \frac{1}{\alpha}\right] \right\}.$$

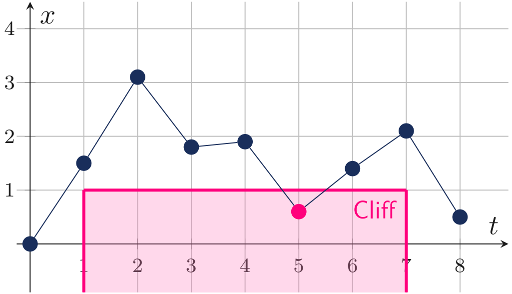
Special cases

- $\rho_{\text{CVaR-p}}(Y;\alpha,\kappa) \to \rho_{\text{CVaR}}(Y;\alpha)$ as $\kappa \to 0$
- $\rho_{\text{CVaR-p}}(Y;\alpha,\kappa) \to \rho_{\mathbb{E}}(Y)$ as $\kappa \to \infty$

Cliff Walking

Consider an autonomous rover that:

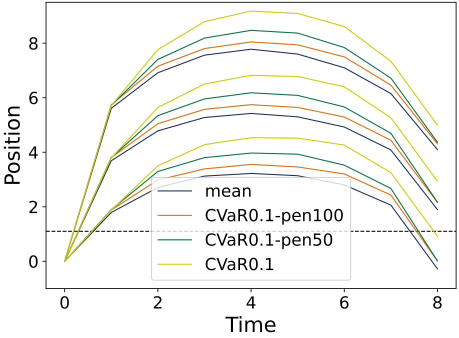
- starts at $(0, 0)$, wants to go at $(T, 0)$
- takes actions $a_t^\theta \sim \pi^\theta = \mathcal{N}(\mu^\theta, \sigma)$
- moves from (t, x_t) to $(t + 1, x_t + a_t)$
- receives penalties when stepping into the cliff and landing away from (T, x)

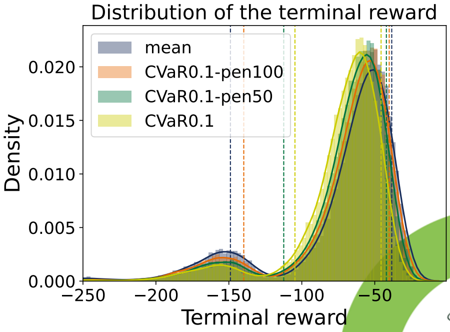


Cliff Walking

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)



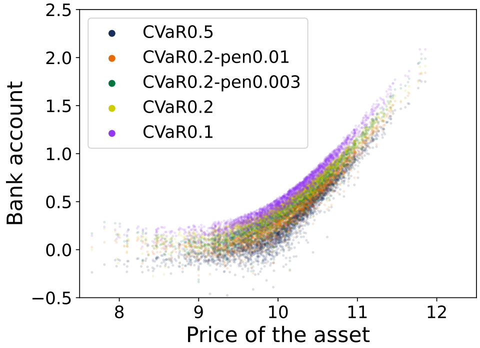


informs.

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

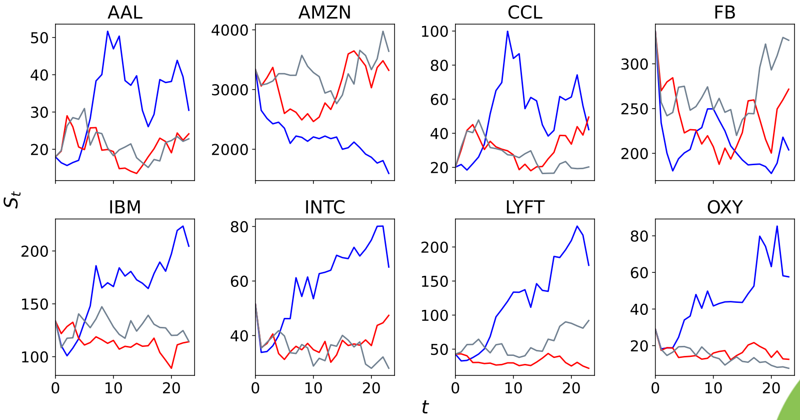
- observes the time t and asset prices $\{S_t^{(i)}\}_{i=1,\ldots,d}$
- decides on the proportion of its wealth $\pi_t^{(i)}$ to invest in asset i
- receives feedback from P&L differences $y_t - y_{t+1}$, where its wealth y_t varies according to

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

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

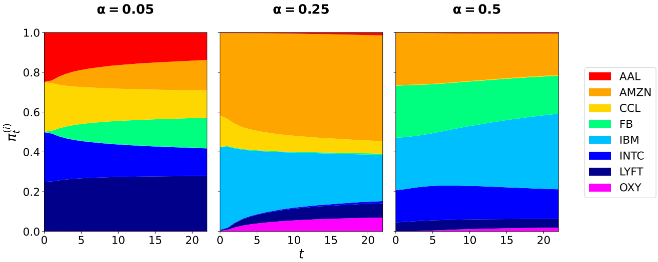
Portfolio Allocation

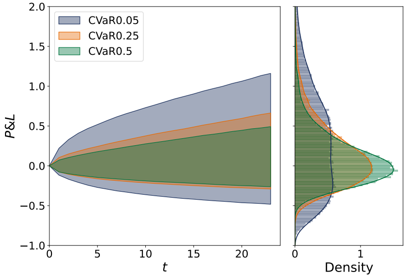
Co-integration model using daily data from eight different stocks listed on the NASDAQ exchange between September 31, 2020 and December 31, 2021.



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 & Future Directions

A unifying, practical framework for policy gradient with dynamic risk measures

- *Risk-sensitive* optimization with *non-stationary policies*
- Generalization to the broad class of *dynamic convex risk measures*
- Novel setting utilizing *elicitable mappings* to avoid nested simulations

Future directions

- Multi-agent RL with dynamic risk measures
- Robust time-consistent RL

Code: `https://github.com/acoache/RL-DynamicConvexRisk`
`https://github.com/acoache/RL-ElicitableDynamicRisk`
Papers: `https://arxiv.org/pdf/2112.13414.pdf`
`https://www.ssrn.com/abstract=4149461`
More info: `anthonycoache.ca`

References

[CJ21] Anthony Coache and Sebastian Jaimungal. Reinforcement learning with dynamic convex risk measures. *arXiv preprint arXiv:2112.13414*, 2021.

[CJC22] Anthony Coache, Sebastian Jaimungal, and Álvaro Cartea. Conditionally elicitable dynamic risk measures for deep reinforcement learning. *arXiv preprint arXiv:2206.14666*, 2022.

[FS02] Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002.

[FZ16] Tobias Fissler and Johanna F Ziegel. Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016.

[Gne11] Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[KT00] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014. Citeseer, 2000.

[MDL21] Saeed Marzban, Erick Delage, and Jonathan Yumeng Li. Deep reinforcement learning for equal risk pricing and hedging under dynamic expectile risk measures. *arXiv preprint arXiv:2109.04001*, 2021.

[Osb85] Kent Osband. *Providing incentives for better cost forecasting*. PhD thesis, University of California, Berkeley, 1985.

[Rus10] Andrzej Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125(2):235–261, 2010.

[TCGM16] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338, 2016.

Algorithms

Algorithm 1: Actor-critic algorithm – Nested simulation approach

Input: ANNs π^θ, V^ϕ, numbers of epochs K, K_1, K_2, mini-batch sizes B_1, B_2, M transitions

Set initial learning rates for ϕ, θ;

for *each iteration* $k = 1, \ldots, K$ **do**

- **for** *each epoch* $k_1 = 1, \ldots, K_1$ **do**
  - Zero out the gradients of V^ϕ;
  - Simulate a mini-batch of B_1 episodes induced by π^θ;
  - Generate M additional (inner) transitions induced by π^θ;
  - Compute the target values of the value function;
  - 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 rates for ϕ with a scheduler;
- **for** *each epoch* $k_2 = 1, \ldots, K_2$ **do**
  - Zero out the gradient of π^θ;
  - Simulate a mini-batch of B_2 episodes induced by π^θ;
  - Generate M 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 with a scheduler;

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

Algorithms

Algorithm 2: Actor-critic algorithm – Conditional elicitability approach

Input: ANNs π^θ, V^ϕ, numbers of epochs K, K_1, K_2, mini-batch sizes B_1, B_2
Set initial learning rates for ϕ, θ;
for *each iteration* $k = 1, \ldots, K$ **do**
  for *each epoch* $k_1 = 1, \ldots, K_1$ **do**
    Zero out the gradients of V^ϕ;
    Simulate a mini-batch of B_1 episodes induced by π^θ;
    Compute the loss $\mathcal{L}(\phi)$: minimization of the expected score;
    Update ϕ by performing an Adam optimisation step;
    if $k_1 \mod K^* = 0$ **then**
      Update the target networks $\tilde{\phi}$;
    Tune the learning rates for ϕ with a scheduler;
  for *each epoch* $k_2 = 1, \ldots, K_2$ **do**
    Zero out the gradient of π^θ;
    Simulate a mini-batch of $\lceil B_2/(1-\alpha) \rceil$ episodes 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 with a scheduler;
Output: Optimal policy π^θ and its value function V^ϕ