Reinforcement Learning for Dynamic Risk Measures

Anthony Coache

Joint work with
Sebastian Jaimungal
and
Álvaro Cartea

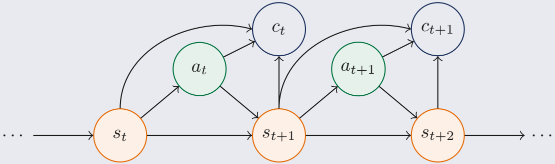
anthonycoache.ca
sebastian.statistics.utoronto.ca
sites.google.com/site/alvarocartea/home

Bachelier Finance Society, 11th World Congress ⋆ June 13-17, 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



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: *risk-neutral objective* function of a cost

$$\min_\theta \mathbb{E}[Y^\theta].$$

Risk-aware RL: *risk measure* ρ of a cost

$$\min_\theta \rho(Y^\theta) \qquad \text{or} \qquad \min_\theta \mathbb{E}[Y^\theta] \ \text{subj. to} \ \rho(Y^\theta) \leq Y^*.$$

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: *risk-neutral objective* function of a cost

$$\min_\theta \mathbb{E}[Y^\theta].$$

Risk-aware RL: *risk measure ρ* of a cost

$$\min_\theta \rho(Y^\theta) \qquad \text{or} \qquad \min_\theta \mathbb{E}[Y^\theta] \text{ subj. to } \rho(Y^\theta) \leq Y^*.$$

Risk-Sensitive RL

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

- Offers a *remedy to environment uncertainty*
- Provides strategies that are more *robust*
- Tuned to *agent's risk preference*

[TCGM16] provide policy search algorithms in the dynamic framework:

- Studies *stationary policies*
- Restricted to *coherent* risk measures

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

- Considers finite-horizon problems and *non-stationary policies*
- Extended to dynamic *convex* risk measures
- Leads to *time-consistent* solutions

Risk-Sensitive RL

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

- Offers a *remedy to environment uncertainty*
- Provides strategies that are more *robust*
- Tuned to *agent's risk preference*

[TCGM16] provide policy search algorithms in the dynamic framework:

- Studies *stationary policies*
- Restricted to *coherent* risk measures

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

- Considers finite-horizon problems and *non-stationary policies*
- Extended to dynamic *convex* risk measures
- Leads to *time-consistent* solutions

Risk-Sensitive RL

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

- Offers a *remedy to environment uncertainty*
- Provides strategies that are more *robust*
- Tuned to *agent's risk preference*

[TCGM16] provide policy search algorithms in the dynamic framework:

- Studies *stationary policies*
- Restricted to *coherent* risk measures

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

- Considers finite-horizon problems and *non-stationary policies*
- Extended to dynamic *convex* risk measures
- Leads to *time-consistent* solutions

Convex Risk Measures

Convex $\rho : \mathcal{Y} \to \mathbb{R}$ [FS02]

- *monotone:* $Y_1 \leq Y_2$ implies $\rho(Y_1) \leq \rho(Y_2)$
- *translation invariant:* $\rho(Y + m) = \rho(Y) + m, \ \forall m \in \mathbb{R}$
- *convex:* $\rho(\lambda Y_1 + (1-\lambda)Y_2) \leq \lambda\rho(Y_1) + (1-\lambda)\rho(Y_2)$

Representation Theorem [SDR14]

Let $\mathbb{E}^\xi[Y] = \sum_\omega Y(\omega)\xi(\omega)d\mathbb{P}(\omega)$ and ρ^* be a convex penalty.

A risk measure ρ is convex, proper and lower semicontinuous iff there exists $\mathcal{U} \subset \{\xi : \sum_\omega \xi(\omega)\mathbb{P}(\omega) = 1, \ \xi \geq 0\}$ such that

$$\rho(Y) = \sup_{\xi \in \mathcal{U}(\mathbb{P})} \left\{ \mathbb{E}^\xi [Y] - \rho^*(\xi) \right\}.$$

We assume an explicit form of the *risk envelope* $\mathcal{U}$ is known

Convex Risk Measures

Convex $\rho : \mathcal{Y} \to \mathbb{R}$ [FS02]

- *monotone:* $Y_1 \leq Y_2$ implies $\rho(Y_1) \leq \rho(Y_2)$
- *translation invariant:* $\rho(Y + m) = \rho(Y) + m, \ \forall m \in \mathbb{R}$
- *convex:* $\rho(\lambda Y_1 + (1-\lambda)Y_2) \leq \lambda\rho(Y_1) + (1-\lambda)\rho(Y_2)$

Representation Theorem [SDR14]

Let $\mathbb{E}^{\xi}[Y] = \sum_{\omega} Y(\omega)\xi(\omega)d\mathbb{P}(\omega)$ and ρ^* be a convex penalty.

A risk measure ρ is convex, proper and lower semicontinuous iff there exists $\mathcal{U} \subset \{\xi : \sum_{\omega}\xi(\omega)\mathbb{P}(\omega) = 1, \ \xi \geq 0\}$ such that

$$\rho(Y) = \sup_{\xi \in \mathcal{U}(\mathbb{P})} \left\{\mathbb{E}^{\xi}\left[Y\right] - \rho^*(\xi)\right\}.$$

We assume an explicit form of the *risk envelope* $\mathcal{U}$ is known

Dynamic Risk Measures

Consider

- $(\Omega, \mathcal{F}, \mathbb{P})$ – Probability space
- $\mathcal{T} := \{0, \dots, T\}$
- $\mathcal{F}_0 \subseteq \cdots \subseteq \mathcal{F}_T$ – Filtration
- $\mathcal{Y}_t := \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ – p-integrable, $\mathcal{F}_t$-measurable random variables
- $\mathcal{Y}_{t,T} := \mathcal{Y}_t \times \cdots \mathcal{Y}_T$ – 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 $0 \le t_1 < t_2 \le T$, 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 $\rho_{t_1,T}(Y_{t_1}, \ldots, Y_T) \le \rho_{t_1,T}(Z_{t_1}, \ldots, Z_T)$.

Theorem [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

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

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

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

Theorem [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 $0 \leq t_1 \leq t_2 \leq T$ and $Y \in \mathcal{Y}_{0,T}$, we have

$$\rho_{t_1,T}(Y_{t_1}, \ldots, Y_T) = \rho_{t_1,t_2}\Big(Y_{t_1}, \ldots, Y_{t_2-1}, \rho_{t_2,T}\big(Y_{t_2}, \ldots, Y_T\big)\Big).$$

Time-Consistency

Time-consistency

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

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

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

Theorem [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 $0 \leq t_1 \leq t_2 \leq T$ and $Y \in \mathcal{Y}_{0,T}$, we have

$$\rho_{t_1,T}(Y_{t_1},\ldots,Y_T) = \rho_{t_1,t_2}\Big(Y_{t_1},\ldots,Y_{t_2-1}, \rho_{t_2,T}\big(Y_{t_2},\ldots,Y_T\big)\Big).$$

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

Theorem [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

$$\rho_{t_1,T}(Y_{t_1}, \ldots, Y_T) = \rho_{t_1,t_2}\Big(Y_{t_1}, \ldots, Y_{t_2-1}, \rho_{t_2,T}\big(Y_{t_2}, \ldots, Y_T\big)\Big).$$

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
- Markovian, i.e. not allowed to depend on the whole past

Problem Setup

Problems of the form $\min_\theta \rho_{0,T}(Y^\theta)$ induced by a policy π^θ, i.e.

$$\min_\theta \rho_0\left(c_0^\theta + \rho_1\left(c_1^\theta + \cdots + \rho_{T-2}\left(c_{T-2}^\theta + \rho_{T-1}\left(c_{T-1}^\theta\right)\right)\cdots\right)\right)$$

Note, here $c_t^\theta := c(s_t, a_t^\theta, s_{t+1}^\theta)$ is a $\mathcal{F}_{t+1}$-measurable random cost

DP equations for the *value function*, i.e. running risk-to-go:

$$V_{T-1}(s;\theta) = \max_{\xi \in \mathcal{U}(\mathbb{P}^\theta(\cdot,\cdot|s_{T-1}=s))} \left\{ \mathbb{E}_{T-1}^\xi\Big[\underbrace{c_{T-1}^\theta}_{\text{final cost}}\Big] - \rho_{T-1}^*(\xi) \right\},$$

$$V_t(s;\theta) = \max_{\xi \in \mathcal{U}(\mathbb{P}^\theta(\cdot,\cdot|s_t=s))} \left\{ \mathbb{E}_t^\xi\Big[\underbrace{c_t^\theta}_{\text{current cost}} + \underbrace{V_{t+1}(s_{t+1}^\theta;\theta)}_{\text{one-step ahead risk-to-go}}\Big] - \rho_t^*(\xi) \right\},$$

for $s \in \mathcal{S}$ and $t = T-2, \ldots, 1$, where $\mathbb{P}^\theta(a, s'|s_t = s) = \mathbb{P}(s'|s,a)\pi^\theta(a|s_t = s)$

Problem Setup

Problems of the form $\min_\theta \rho_{0,T}(Y^\theta)$ induced by a policy π^θ, i.e.

$$\min_\theta \rho_0\left(c_0^\theta + \rho_1\left(c_1^\theta + \cdots + \rho_{T-2}\left(c_{T-2}^\theta + \rho_{T-1}\left(c_{T-1}^\theta\right)\right)\cdots\right)\right)$$

Note, here $c_t^\theta := c(s_t, a_t^\theta, s_{t+1}^\theta)$ is a $\mathcal{F}_{t+1}$-measurable random cost

DP equations for the *value function*, i.e. running risk-to-go:

$$V_{T-1}(s;\theta) = \max_{\xi \in \mathcal{U}(\mathbb{P}^\theta(\cdot,\cdot|s_{T-1}=s))} \left\{ \mathbb{E}_{T-1}^\xi\left[\underbrace{c_{T-1}^\theta}_{\text{final cost}} \right] - \rho_{T-1}^*(\xi) \right\},$$

$$V_t(s;\theta) = \max_{\xi \in \mathcal{U}(\mathbb{P}^\theta(\cdot,\cdot|s_t=s))} \left\{ \mathbb{E}_t^\xi\left[\underbrace{c_t^\theta}_{\text{current cost}} + \underbrace{V_{t+1}(s_{t+1}^\theta;\theta)}_{\text{one-step ahead risk-to-go}} \right] - \rho_t^*(\xi) \right\},$$

for $s \in \mathcal{S}$ and $t = T-2, \ldots, 1$, where $\mathbb{P}^\theta(a, s'|s_t = s) = \mathbb{P}(s'|s,a)\pi^\theta(a|s_t = s)$

Problem Setup

Problems of the form $\min_\theta \rho_{0,T}(Y^\theta)$ induced by a policy π^θ, i.e.

$$\min_\theta \rho_0\left(c_0^\theta + \rho_1\left(c_1^\theta + \cdots + \rho_{T-2}\left(c_{T-2}^\theta + \rho_{T-1}\left(c_{T-1}^\theta\right)\right)\cdots\right)\right)$$

Note, here $c_t^\theta := c(s_t, a_t^\theta, s_{t+1}^\theta)$ is a $\mathcal{F}_{t+1}$-measurable random cost

DP equations for the *value function*, i.e. running risk-to-go:

$$V_{T-1}(s;\theta) = \max_{\xi \in \mathcal{U}(\mathbb{P}^\theta(\cdot,\cdot|s_{T-1}=s))} \left\{ \mathbb{E}_{T-1}^\xi \Big[\underbrace{c_{T-1}^\theta}_{\text{final cost}} \Big] - \rho_{T-1}^*(\xi) \right\},$$

$$V_t(s;\theta) = \max_{\xi \in \mathcal{U}(\mathbb{P}^\theta(\cdot,\cdot|s_t=s))} \left\{ \mathbb{E}_t^\xi \Big[\underbrace{c_t^\theta}_{\text{current cost}} + \underbrace{V_{t+1}(s_{t+1}^\theta;\theta)}_{\text{one-step ahead risk-to-go}} \Big] - \rho_t^*(\xi) \right\},$$

for $s \in \mathcal{S}$ and $t = T-2, \ldots, 1$, where $\mathbb{P}^\theta(a, s'|s_t = s) = \mathbb{P}(s'|s,a)\pi^\theta(a|s_t = s)$

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]

The gradient of the value function at period $T-1$ is

$$\nabla_\theta V_{T-1}(s;\theta) = \mathbb{E}^{\xi^*}_{T-1}\left[\left(c(s, a^\theta_{T-1}, s^\theta_T) - \lambda^*\right)\nabla_\theta \log \pi^\theta(a^\theta_{T-1}|s)\right] - \nabla_\theta \rho^*_{T-1}(\xi^*),$$

and the gradient of the value function at periods $t = T-2, \ldots, 0$ is

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

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]

The gradient of the value function at period $T-1$ is

$$\nabla_\theta V_{T-1}(s;\theta) = \mathbb{E}^{\xi^*}_{T-1}\left[\left(c(s, a^\theta_{T-1}, s^\theta_T) - \lambda^*\right)\nabla_\theta \log \pi^\theta(a^\theta_{T-1}|s)\right] - \nabla_\theta \rho^*_{T-1}(\xi^*),$$

and the gradient of the value function at periods $t = T-2, \dots, 0$ is

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

Algorithm

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

Algorithm 1: Main algorithm

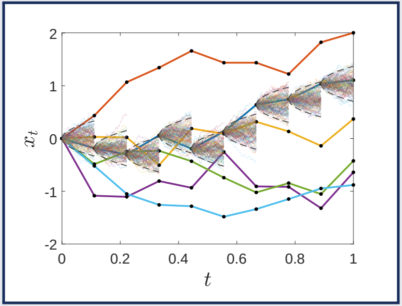
Input: Value function V^ϕ, policy π^θ
Initialize environment and optimizers;
for *each epoch* $k = 1, \dots, K$ **do**
    Generate trajectories;
    Estimate V^ϕ using π^θ;
    Update π^θ using V^ϕ;
Output: Optimal policy $\pi^\theta \approx \pi^*$

- We parametrize policy and value function by ANNs, denoted θ and ϕ

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 (working paper: Coache, Jaimungal & Cartea (2022))

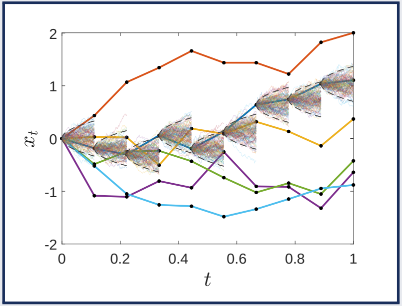
- *Conditional elicitability* of dynamic spectral risk measures [FZ16]
- Avoids nested simulations, *memory efficient*

- 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 (working paper: Coache, Jaimungal & Cartea (2022))

- *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

Background on elicitability [see e.g. Gne11].
Let $\mathfrak{a} \in \mathbb{A}$ be a point estimate of the mapping of interest $M(Y)$, $Y \sim \mathbb{F}$

Elicitable mapping

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

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

Modeling $M(Y|X = x)$ with an ANN $H^{\psi}(x) : \mathbb{X} \to \mathbb{A}$, and empirical estimates based on observed data

$$\hat{\psi} = \arg\min_{\psi} \frac{1}{n} \sum_{i=1}^{n} \left[S\Big(H^{\psi}(x^{(i)}), Y^{(i)}\Big)\right]$$

Elicitability

Background on elicitability [see e.g. Gne11].
Let $\mathfrak{a} \in \mathbb{A}$ be a point estimate of the mapping of interest $M(Y)$, $Y \sim \mathbb{F}$

Elicitable mapping

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

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

Modeling $M(Y|X = x)$ with an ANN $H^{\psi}(x) : \mathbb{X} \to \mathbb{A}$, and empirical estimates based on observed data

$$\hat{\psi} = \arg\min_{\psi} \frac{1}{n} \sum_{i=1}^{n} \left[S\Big(H^{\psi}(x^{(i)}), Y^{(i)} \Big) \right]$$

Conditional Elicitability

Originating from the work of [Osb85], where components of a k-elicitable vector-valued mapping can fail to be 1-elicitable

Conditional elicitability of the CVaR [FZ16]

Let distribution functions of Y, denoted $\mathbb{F}$, have finite first moments, unique α-quantiles, and be supported on $\mathbb{Y} \subseteq \mathbb{R}$. Define the mapping

$$M(Y) = \Big(\mathsf{VaR}_\alpha(Y),\ \mathsf{CVaR}_\alpha(Y)\Big) \quad \text{and} \quad \mathbb{A} = \Big\{\mathfrak{a} \in \mathbb{Y}^2 \mid \mathfrak{a}_1 \leq \mathfrak{a}_2\Big\}.$$

Then

- the mapping M is 2-elicitable wrt $\mathbb{F}$;
- a scoring function $S : \mathbb{A} \times \mathbb{Y} \to \mathbb{R}$ of this form is strictly $\mathbb{F}$-consistent for M

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

- Similar result for classes 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 (CVaR): $\rho_{\text{CVaR}}(Y;\alpha) = \sup_{\xi \in \mathcal{U}(\mathbb{P})} \left\{ \mathbb{E}^{\xi}\left[Y\right] \right\}$
- Penalized CVaR: $\rho_{\text{CVaR-p}}(Y;\alpha,\kappa) = \sup_{\xi \in \mathcal{U}(\mathbb{P})} \left\{ \mathbb{E}^{\xi}\left[Y\right] - \kappa \mathbb{E}^{\xi}\left[\log \xi\right] \right\}$

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

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

Statistical Arbitrage

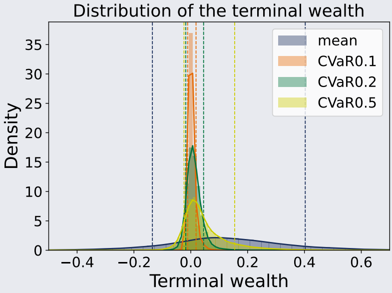
Consider a market with a single asset. An agent:

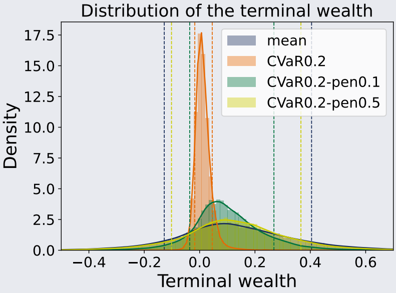
- invests during T periods
- observes its inventory $q_t \in (-q_{\max}, q_{\max})$ and the asset price S_t
- trades quantities $a_t \in (-a_{\max}, a_{\max})$ of the asset
- faces cost transactions and a terminal penalty imposed by the market
- receives a cost that affects its wealth $y_t \in \mathbb{R}$

Statistical Arbitrage

Consider a market with a single asset. An agent:

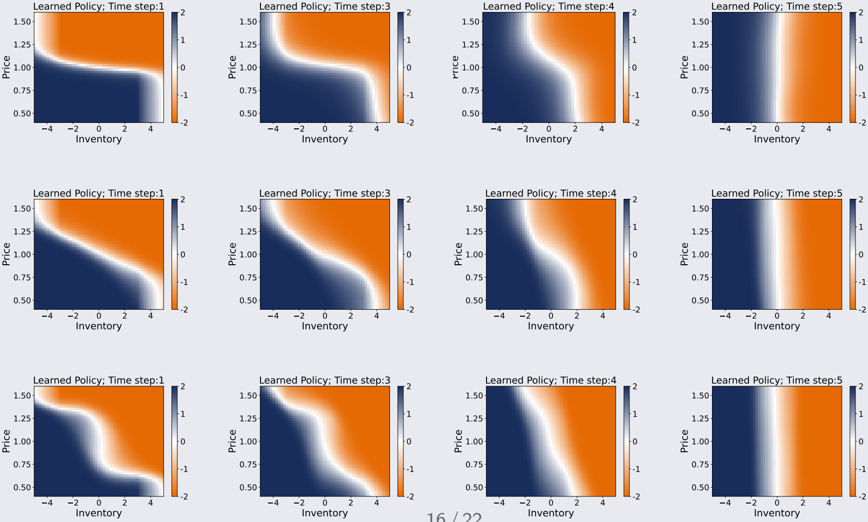
- invests during T periods
- observes its inventory $q_t \in (-q_{\max}, q_{\max})$ and the asset price S_t
- trades quantities $a_t \in (-a_{\max}, a_{\max})$ of the asset
- faces cost transactions and a terminal penalty imposed by the market
- receives a cost that affects its wealth $y_t \in \mathbb{R}$





Statistical Arbitrage

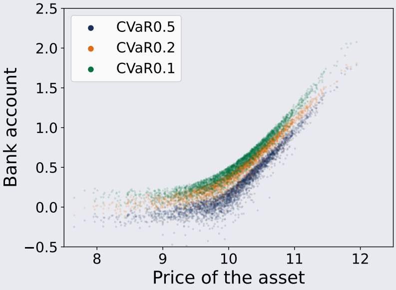
- Asset price: Ornstein-Uhlenbeck process with mean-reversion level at 1
- $\rho_{\mathbb{E}}$ (top), $\rho_{\mathsf{CVaR}_{0.2}}$ with $\kappa = 0.1$ (middle), $\rho_{\mathsf{CVaR}_{0.2}}$ (bottom)



Option Hedging

Consider a call option where the underlying asset dynamics follow the Heston model. An agent:

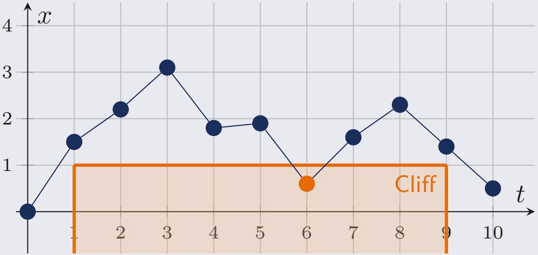
- sells the call option, and aims to hedge it trading solely the asset
- observes its previous position a_t, its bank account B_t, the price S_t
- trades in a market with transaction costs (per share) and an interest rate
- receives a cost that affect its wealth y_t



Cliff Walking

Consider an autonomous rover that:

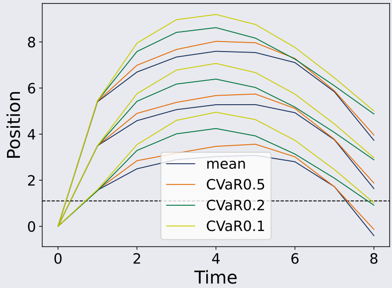
- starts at $(0, 0)$, wants to go at $(T, 0)$
- moves from (t, x_t) to $(t+1, x_t + a_t)$
- takes actions $a_t^\theta \sim \pi^\theta = \mathcal{N}(\mu^\theta, \sigma)$
- receives a big penalty when stepping into the cliff
- gets a penalty when landing further from the goal at (T, x)

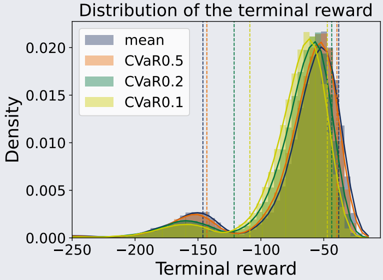


Cliff Walking

Consider an autonomous rover that:

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





Portfolio Allocation

Consider a market with 3 assets. An agent

- changes its portfolio allocation during T periods
- observes the time t and asset prices $\{S_t^{(i)}\}_{i=1,2,3}$
- decides on the proportion of its wealth $\pi_t^{(i)}$ to invest in asset i
- sees its wealth y_t vary according to

$$\mathrm{d}y_t = y_t \left(\sum_{i=1}^{3} \pi_t^{(i)} \frac{\mathrm{d}S_t^{(i)}}{S_t^{(i)}} \right)$$

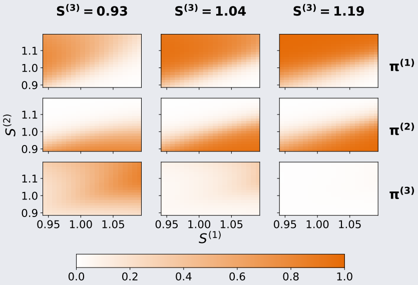
- receives feedback from P&L differences $y_{t+1} - y_t$

We assume a null interest rate, correlated financial instruments, no leveraging nor short-selling

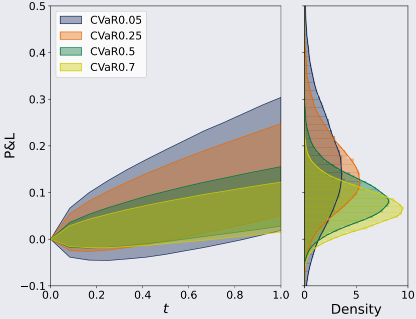
Portfolio Allocation

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

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



$\alpha = 0.4$



P&L

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

- Deep deterministic policy gradient with dynamic risk measures
- Robust time-consistent reinforcement learning

Code: `https://github.com/acoache/RL-DynamicConvexRisk`
Paper: `https://arxiv.org/pdf/2112.13414.pdf`
More info: `anthonycoache.ca`

References

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

[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.

[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.

[SDR14] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[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.

Algorithm – Estimation of Value Function

Algorithm 2: Estimation of the value function V (Nested approach)

Input: V^ϕ, π^θ, N trajectories, M transitions, K epochs, batch size B

for *each epoch* $k = 1, \dots, K$ **do**

Set the gradients to zero;

Sample B states $s_t^{(b)}$, $b = 1, \dots, B$, $t \in \mathcal{T}$;

Obtain from π^θ the transitions $(a_t^{(b,m)}, s_{t+1}^{(b,m)}, c_t^{(b,m)})$, $m = 1, \dots, M$;

for *each state* $b = 1, \dots, B$, $t \in \mathcal{T}$ **do**

Compute the *predicted values* $\hat{v}_t^b = V_t^\phi(s_t^{(b)}; \theta)$;

Set the *target value* as

$$v_t^b = \max_{\xi \in \mathcal{U}(\mathbb{P}^\theta(\cdot,\cdot | s_t = s_t^{(b)}))} \left\{ \mathbb{E}^\xi_{t, s_t^{(b)}} \left[c_t^{(b,m)} + V_{t+1}^\phi(s_{t+1}^{(b,m)}; \theta) \right] + \rho_t^*(\xi) \right\};$$

Compute the expected square loss between v_t^b and $\hat{v}_t^b$;

Update ϕ by performing an Adam optimizer step;

Output: An estimate of the value function $V_t^\phi(s; \theta) \approx V_t(s; \theta)$

Algorithm – Estimation of Value Function

Algorithm 3: Estimation of the value function V (Elicitable approach)

Input: $H_1^{\psi_1}$, $H_2^{\psi_2}$, $V^\phi = H_1^{\psi_1} + H_2^{\psi_2}$, π^θ, N trajectories, K epochs, batch size B

for *each epoch* $k = 1, \dots, K$ **do**

Set the gradients to zero;

Simulate B episodes induced by π^θ;

Compute the loss

$$\mathcal{L}^\phi = \sum_{t \in \mathcal{T}} \sum_{b=1}^{B} \left[S\left(H_1^{\psi_1}\left(s_t^{(b)}; \theta\right);\ V^\phi\left(s_t^{(b)}; \theta\right);\ c_t^{(b)} + V^{\tilde{\phi}}\left(s_{t+1}^{(b)}; \theta\right)\right)\right];$$

Update $\phi = \{\psi_1, \psi_2\}$ by performing an Adam optimizer step;

Output: An estimate of the value function $V^\phi(s_t; \theta) \approx V_t(s; \theta)$

Algorithm – Update of Policy

Algorithm 4: Update of the policy π

Input: π^θ, V^ϕ, N trajectories, M transitions, K epochs, batch size B

for *each epoch* $k = 1, \dots, K$ **do**

Set the gradients to zero;

Sample B states $s_t^{(b)}$, $b = 1, \dots, B$, $t \in \mathcal{T}$;

Obtain from π^θ the transitions $(a_t^{(b,m)}, s_{t+1}^{(b,m)}, c_t^{(b,m)})$, $m = 1, \dots, M$;

for *each state* $b = 1, \dots, B$, $t \in \mathcal{T}$ **do**

Obtain $\hat{z}_t^{(b,m)} = \nabla_\theta \log \pi^\theta(a_t^{(b,m)} | s_t^{(b)})$ from reparametrization trick;

Obtain $\hat{v}_{t+1}^{(b,m)} = V_{t+1}^\phi(s_{t+1}^{(b,m)}; \theta)$;

Obtain $\hat{\rho}_t^{(b)} = \nabla_\theta \rho_t^*(\xi^*)$;

Calculate the *gradient* $\nabla_\theta V_t(s_t^{(b)}; \theta)$ using empirical estimates

$$\ell_t^{(b)} = \frac{1}{M} \sum_{m=1}^{M} \left(\left(c_t^{(b,m)} + \hat{v}_{t+1}^{(b,m)} - \lambda^* \right) \hat{z}_t^{(b,m)} - \hat{\rho}_t^{(b)} \right);$$

Take the average $\ell = \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=0}^{T-1} \ell_t^{(b)}$;

Update θ by performing an Adam optimizer step ;

Output: An updated policy π^θ