Optimising a Dynamic CVaR over Policies using Conditional Elicitability

Anthony Coache

Joint work with
Sebastian Jaimungal   Álvaro Cartea

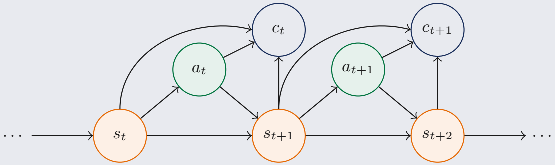
anthonycoache.ca
sebastian.statistics.utoronto.ca
sites.google.com/site/alvarocartea/home

Oxford-Man Institute Presentations ⋆ May 6, 2022

Reinforcement Learning

Markov decision process (MDP)

- $\mathcal{S}$ – State space
- $\mathcal{A}$ – Action space
- $\pi^\theta(a_t|s_t)$ – Randomised policy parametrised by θ
- $\mathbb{P}(s_0), \mathbb{P}(s_{t+1}|s_t, a_t)$ – Transition probability distribution
- $c(s_t, a_t, s_{t+1}) \in \mathcal{C}$ – Cost function
- $(s_0, a_0, c_0, \ldots, s_{T-1}, a_{T-1}, c_{T-1}, s_T)$ – Episode



Dynamic Risk

We consider dynamic risk measures [see e.g. Rus10]

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

where Y_t is a $\mathcal{F}_t$-measurable random cost

- Constitutes a class of time-consistent risk measures
- Leads to time-consistent solutions, i.e. an optimal behaviour planned for a future state of the environment is still optimal once the agent visits the state

- ρ_t is a static CVaR

$$\text{CVaR}_\alpha(Y) = \frac{1}{1-\alpha}\int_\alpha^1 \text{VaR}_u(Y)\text{d}u, \quad \alpha \in (0,1)$$

- Can be generalised to spectral risk measures with finite support spectrum

Dynamic Risk

We consider dynamic risk measures [see e.g. Rus10]

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

where Y_t is a $\mathcal{F}_t$-measurable random cost

- Constitutes a class of time-consistent risk measures
- Leads to time-consistent solutions, i.e. an optimal behaviour planned for a future state of the environment is still optimal once the agent visits the state
- ρ_t is a static CVaR

$$\mathsf{CVaR}_\alpha(Y) = \frac{1}{1-\alpha}\int_\alpha^1 \mathsf{VaR}_u(Y)\mathrm{d}u, \quad \alpha \in (0,1)$$

- Can be generalised to spectral risk measures with finite support spectrum

Problem Setup

We aim to solve risk-sensitive RL problems minimising the dynamic CVaR:

$$\min_{\theta} \mathrm{CVaR}_\alpha\left(c_0^\theta + \mathrm{CVaR}_\alpha\left(c_1^\theta + \cdots + \mathrm{CVaR}_\alpha\left(c_{T-2}^\theta + \mathrm{CVaR}_\alpha\left(c_{T-1}^\theta\right)\right)\cdots\right)\right)$$

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

Dynamic programming equations for the value function – running risk-to-go:

$$V(s_{T-1};\theta) = \mathrm{CVaR}_\alpha\left(\underbrace{c_{T-1}^\theta}_{\text{final cost}} \middle| s_{T-1}\right), \quad \text{and}$$

$$V(s_t;\theta) = \mathrm{CVaR}_\alpha\left(\underbrace{c_t^\theta}_{\text{current cost}} + \underbrace{V(s_{t+1}^\theta;\theta)}_{\text{one-step ahead risk-to-go}} \middle| s_t\right).$$

- How to efficiently estimate the value function?
- How to optimise over policies under the same framework?

Problem Setup

We aim to solve risk-sensitive RL problems minimising the dynamic CVaR:

$$\min_{\theta} \mathsf{CVaR}_{\alpha}\left(c_0^{\theta} + \mathsf{CVaR}_{\alpha}\left(c_1^{\theta} + \cdots + \mathsf{CVaR}_{\alpha}\left(c_{T-2}^{\theta} + \mathsf{CVaR}_{\alpha}\left(c_{T-1}^{\theta} \right) \right) \cdots \right) \right)$$

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

Dynamic programming equations for the value function – running risk-to-go:

$$V(s_{T-1};\theta) = \mathsf{CVaR}_{\alpha}\left(\underbrace{c_{T-1}^{\theta}}_{\text{final cost}} \,\middle|\, s_{T-1} \right), \quad \text{and}$$

$$V(s_t;\theta) = \mathsf{CVaR}_{\alpha}\left(\underbrace{c_t^{\theta}}_{\text{current cost}} + \underbrace{V(s_{t+1}^{\theta};\theta)}_{\text{one-step ahead risk-to-go}} \,\middle|\, s_t \right).$$

- How to efficiently estimate the value function?
- How to optimise over policies under the same framework?

Problem Setup

We aim to solve risk-sensitive RL problems minimising the dynamic CVaR:

$$\min_{\theta} \mathrm{CVaR}_\alpha\left(c_0^\theta + \mathrm{CVaR}_\alpha\left(c_1^\theta + \cdots + \mathrm{CVaR}_\alpha\left(c_{T-2}^\theta + \mathrm{CVaR}_\alpha\left(c_{T-1}^\theta\right)\right)\cdots\right)\right)$$

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

Dynamic programming equations for the value function – running risk-to-go:

$$V(s_{T-1};\theta) = \mathrm{CVaR}_\alpha\left(\underbrace{c_{T-1}^\theta}_{\text{final cost}} \middle| s_{T-1}\right), \quad \text{and}$$

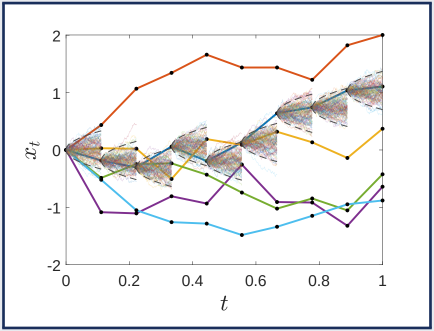
$$V(s_t;\theta) = \mathrm{CVaR}_\alpha\left(\underbrace{c_t^\theta}_{\text{current cost}} + \underbrace{V(s_{t+1}^\theta;\theta)}_{\text{one-step ahead risk-to-go}} \middle| s_t\right).$$

- How to efficiently estimate the value function?
- How to optimise over policies under the same framework?

Nested Simulation Approach

Previous approach: nested simulation framework [see e.g. CJ21]

- Computationally expensive in terms of memory
- Acquisition of new observations may not be possible
- Highly ineffective or even impracticable



Goal: develop a computationally efficient framework for this class of problems

Elicitability

Background on elicitability [see e.g. Gne11]

- $Y \sim \mathbb{F}$ – d-dimensional random variable with support on $\mathbb{Y} \subseteq \mathbb{R}^d$
- $a \in \mathbb{A} \subseteq \mathbb{R}^k$ – k-dimensional point approximation
- $M : \mathbb{Y} \to \mathbb{A}$ – statistical mapping of interest
- $S : \mathbb{A} \times \mathbb{Y} \to \mathbb{R}$ – scoring function

Objective:
Find a scoring function S such that when observing a realisation $y \in \mathbb{Y}$, our current point forecast $a \in \mathbb{A}$ is penalized by $S(a, y)$.

Examples:

- $S(a, y) = (a - y)^2$ for the mean
- $S(a, y) = |a - y|$ for the median

Elicitability

$\mathbb{F}$-consistent scoring function

$S : \mathbb{A} \times \mathbb{Y} \to \mathbb{R}$ such that for any $F \in \mathbb{F}$ and $a \in \mathbb{A}$,

$$\mathbb{E}_{Y \sim F}\Big[S\big(M(Y), Y\big)\Big] \leq \mathbb{E}_{Y \sim F}\Big[S\big(a, Y\big)\Big].$$

Furthermore, S is *strictly* $\mathbb{F}$-consistent for M if the equality implies $a = M(Y)$.

k-elicitable mapping

$M : \mathbb{Y} \to \mathbb{A}$ such that there exists a strictly $\mathbb{F}$-consistent scoring function S.

A mapping M is k-elicitable iff there exists a scoring function such that the correct estimate of M is the unique minimiser of the expected score, i.e.

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

Elicitability

$\mathbb{F}$-consistent scoring function

$S : \mathbb{A} \times \mathbb{Y} \to \mathbb{R}$ such that for any $F \in \mathbb{F}$ and $a \in \mathbb{A}$,

$$\mathbb{E}_{Y\sim F}\Big[S\big(M(Y), Y\big)\Big] \leq \mathbb{E}_{Y\sim F}\Big[S\big(a, Y\big)\Big].$$

Furthermore, S is *strictly* $\mathbb{F}$-consistent for M if the equality implies $a = M(Y)$.

k-elicitable mapping

$M : \mathbb{Y} \to \mathbb{A}$ such that there exists a strictly $\mathbb{F}$-consistent scoring function S.

A mapping M is k-elicitable iff there exists a scoring function such that the correct estimate of M is the unique minimiser of the expected score, i.e.

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

Deep Composite Modelling

Original problem:

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

- Suppose the random variable Y is supported by observed features $x \in \mathbb{R}^q$
- Modelling $M(Y)$ with an ANN $H^\psi : \mathbb{R}^q \to \mathbb{A}$

$$\psi^* = \arg\min_{\psi} \mathbb{E}_{Y \sim F}\left[S\Big(H^\psi(x), Y\Big)\right].$$

Replace the expectation by the empirical mean based on some observed data

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

Valid for any strictly consistent scoring function S

Deep Composite Modelling

Original problem:

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

- Suppose the random variable Y is supported by observed features $x \in \mathbb{R}^q$
- Modelling $M(Y)$ with an ANN $H^\psi : \mathbb{R}^q \to \mathbb{A}$

$$\psi^* = \arg\min_{\psi} \mathbb{E}_{Y \sim F}\Big[S\Big(H^\psi(x), Y\Big)\Big].$$

Replace the expectation by the empirical mean based on some observed data

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

Valid for any strictly consistent scoring function S

Deep Composite Modelling

Original problem:

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

- Suppose the random variable Y is supported by observed features $x \in \mathbb{R}^q$
- Modelling $M(Y)$ with an ANN $H^{\psi} : \mathbb{R}^q \to \mathbb{A}$

$$\psi^* = \arg\min_{\psi} \mathbb{E}_{Y \sim F}\Big[S\Big(H^{\psi}(x), Y\Big)\Big].$$

Replace the expectation by the empirical mean based on some observed data

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

Valid for any strictly consistent scoring function S

Conditional Elicitability

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

- Variance is 2-elicitable (conditionally on the mean), but not 1-elicitable
- CVaR is 2-elicitable (conditionally on the VaR), but not 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

$$M(Y) = \Big(\mathrm{VaR}_\alpha(Y),\, \mathrm{CVaR}_\alpha(Y)\Big) \quad \text{and} \quad \mathbb{A} = \Big\{a \in \mathbb{Y}^2 \mid a_1 \leq a_2\Big\}.$$

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

$$S(a_1, a_2, y) = \Big(\mathbb{1}(y \leq a_1) - \alpha\Big)\Big(G_1(a_1) - G_1(y)\Big) - G_2(a_2) + G_2(y)$$
$$+ \nabla G_2(a_2)\left[a_2 + \frac{1}{1-\alpha}\Big(\Big(\mathbb{1}(y > a_1) - (1-\alpha)\Big)a_1 - \mathbb{1}(y > a_1)y\Big)\right]$$

Conditional Elicitability

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

- Variance is 2-elicitable (conditionally on the mean), but not 1-elicitable
- CVaR is 2-elicitable (conditionally on the VaR), but not 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

$$M(Y) = \Big(\mathrm{VaR}_\alpha(Y),\, \mathrm{CVaR}_\alpha(Y)\Big) \quad \text{and} \quad \mathbb{A} = \Big\{a \in \mathbb{Y}^2 \mid a_1 \leq a_2\Big\}.$$

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

$$\begin{aligned} S(a_1, a_2, y) = {} & \Big(\mathbb{1}(y \leq a_1) - \alpha\Big)\Big(G_1(a_1) - G_1(y)\Big) - G_2(a_2) + G_2(y) \\ & + \nabla G_2(a_2)\left[a_2 + \frac{1}{1-\alpha}\left(\Big(\mathbb{1}(y > a_1) - (1-\alpha)\Big)a_1 - \mathbb{1}(y > a_1)y\right)\right] \end{aligned}$$

Algorithm

Actor-critic style algorithm

- Critic: Estimate V (for a fixed π) with deep composite modelling
- Actor: Update π (for a fixed V) with a policy gradient method

Suppose we simulate B full episodes composed of T transitions:

$$\left(s_t^{(b)}, a_t^{(b)}, s_{t+1}^{(b)}, c_t^{(b)}\right), \quad t \in \mathcal{T},\ b \in B$$

Define the following ANN structures:

- $\pi^\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ – Policy
- $V^{\psi_1,\psi_2}(s_t;\theta) = H_1^{\psi_1}(s_t;\theta) + H_2^{\psi_2}(s_t;\theta)$ – Value function
- $H_1^{\psi_1}(s_t;\theta) \in \mathbb{R}$ – Estimate of $\mathrm{VaR}_\alpha\left(c_t^\theta + V(s_{t+1}^\theta;\theta) \middle| s_t\right)$
- $H_2^{\psi_2}(s_t;\theta) \in \mathbb{R}_+$ – Estimate of $\left(V(s_t;\theta) - \mathrm{VaR}_\alpha\left(c_t^\theta + V(s_{t+1}^\theta;\theta) \middle| s_t\right)\right)$

$$V(s_t;\theta) = \mathrm{CVaR}_\alpha\left(c_t^\theta + V(s_{t+1}^\theta;\theta) \middle| s_t\right)$$

Algorithm

Actor-critic style algorithm

- Critic: Estimate V (for a fixed π) with deep composite modelling
- Actor: Update π (for a fixed V) with a policy gradient method

Suppose we simulate B full episodes composed of T transitions:

$$\left(s_t^{(b)}, a_t^{(b)}, s_{t+1}^{(b)}, c_t^{(b)}\right), \quad t \in \mathcal{T},\ b \in B$$

Define the following ANN structures:

- $\pi^\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ – Policy
- $V^{\psi_1,\psi_2}(s_t;\theta) = H_1^{\psi_1}(s_t;\theta) + H_2^{\psi_2}(s_t;\theta)$ – Value function
- $H_1^{\psi_1}(s_t;\theta) \in \mathbb{R}$ – Estimate of $\mathrm{VaR}_\alpha\left(c_t^\theta + V(s_{t+1}^\theta;\theta) \middle| s_t\right)$
- $H_2^{\psi_2}(s_t;\theta) \in \mathbb{R}_+$ – Estimate of $\left(V(s_t;\theta) - \mathrm{VaR}_\alpha\left(c_t^\theta + V(s_{t+1}^\theta;\theta) \middle| s_t\right)\right)$

$$V(s_t;\theta) = \mathrm{CVaR}_\alpha\left(c_t^\theta + V(s_{t+1}^\theta;\theta) \middle| s_t\right)$$

Critic: Update of Value Function

$$\min_{\psi_1,\psi_2} \mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t)}\left[S\left(\underbrace{H_1^{\psi_1}(s_t;\theta)}_{\text{VaR}_\alpha};\ \underbrace{H_1^{\psi_1}(s_t;\theta)+H_2^{\psi_2}(s_t;\theta)}_{\text{CVaR}_\alpha};\ \underbrace{c_t^\theta + V^{\psi_1,\psi_2}(s_{t+1}^\theta;\theta)}_{\text{running risk-to-go}}\right)\right]$$

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

- Conditional elicitability of the CVaR with scoring function $S(a_1,a_2,y)$
- Empirical mean over observed transitions
- Update ψ_1,ψ_2 using an optimisation rule, e.g. Adam

Loss for the update of ψ_1,ψ_2:

$$\mathcal{L}^{\psi_1,\psi_2} = \sum_{t\in\mathcal{T}}\sum_{b=1}^{B}\left[S\left(\underbrace{H_1^{\psi_1}\left(s_t^{(b)};\theta\right)}_{\text{VaR}_\alpha(\cdot|s_t)};\ \underbrace{V^{\psi_1,\psi_2}\left(s_t^{(b)};\theta\right)}_{\text{CVaR}_\alpha(\cdot|s_t)};\ \underbrace{c_t^{(b)} + V^{\tilde\psi_1,\tilde\psi_2}\left(s_{t+1}^{(b)};\theta\right)}_{\text{running risk-to-go}}\right)\right]$$

Critic: Update of Value Function

$$\min_{\psi_1,\psi_2} \mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t)}\left[S\left(\underbrace{H_1^{\psi_1}(s_t;\theta)}_{\text{VaR}_\alpha};\ \underbrace{H_1^{\psi_1}(s_t;\theta)+H_2^{\psi_2}(s_t;\theta)}_{\text{CVaR}_\alpha};\ \underbrace{c_t^\theta + V^{\psi_1,\psi_2}(s_{t+1}^\theta;\theta)}_{\text{running risk-to-go}}\right)\right]$$

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

- Conditional elicitability of the CVaR with scoring function $S(a_1,a_2,y)$
- Empirical mean over observed transitions
- Update ψ_1,ψ_2 using an optimisation rule, e.g. Adam

Loss for the update of ψ_1,ψ_2:

$$\mathcal{L}^{\psi_1,\psi_2} = \sum_{t\in\mathcal{T}}\sum_{b=1}^{B}\left[S\left(\underbrace{H_1^{\psi_1}\left(s_t^{(b)};\theta\right)}_{\text{VaR}_\alpha(\cdot|s_t)};\ \underbrace{V^{\psi_1,\psi_2}\left(s_t^{(b)};\theta\right)}_{\text{CVaR}_\alpha(\cdot|s_t)};\ \underbrace{c_t^{(b)} + V^{\tilde\psi_1,\tilde\psi_2}\left(s_{t+1}^{(b)};\theta\right)}_{\text{running risk-to-go}}\right)\right]$$

Critic: Update of Value Function

$$\min_{\psi_1,\psi_2} \; \mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t)}\left[S\Big(\underbrace{H_1^{\psi_1}(s_t;\theta)}_{\text{VaR}_\alpha};\; \underbrace{H_1^{\psi_1}(s_t;\theta) + H_2^{\psi_2}(s_t;\theta)}_{\text{CVaR}_\alpha};\; \underbrace{c_t^\theta + V^{\psi_1,\psi_2}(s_{t+1}^\theta;\theta)}_{\text{running risk-to-go}} \Big) \right]$$

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

- Conditional elicitability of the CVaR with scoring function $S(a_1,a_2,y)$
- Empirical mean over observed transitions
- Update ψ_1,ψ_2 using an optimisation rule, e.g. Adam

Loss for the update of ψ_1,ψ_2:

$$\mathcal{L}^{\psi_1,\psi_2} = \sum_{t\in\mathcal{T}} \sum_{b=1}^{B} \left[S\Big(\underbrace{H_1^{\psi_1}\left(s_t^{(b)};\theta\right)}_{\text{VaR}_\alpha(\cdot|s_t)};\; \underbrace{V^{\psi_1,\psi_2}\left(s_t^{(b)};\theta\right)}_{\text{CVaR}_\alpha(\cdot|s_t)};\; \underbrace{c_t^{(b)} + V^{\tilde{\psi}_1,\tilde{\psi}_2}\left(s_{t+1}^{(b)};\theta\right)}_{\text{running risk-to-go}} \Big) \right]$$

Actor: Update of Policy

Using both the representation theorem [SDR14] and Envelope theorem [MS02]:

$$\nabla_\theta V(s_t;\theta) = \frac{1}{1-\alpha}\mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t)}\left[\left(c_t^\theta + V(s_{t+1}^\theta;\theta) - \lambda^*\right)_+ \left(\nabla_\theta \log \pi^\theta(a_t^\theta|s_t)\right)\right]$$

where λ^* is any α-quantile of $c_t^\theta + V(s_{t+1}^\theta;\theta)$

- Policy gradient method with the gradient of the value function V
- Empirical mean over observed transitions
- Estimation of the VaR_α with $H_1^{\psi_1}$
- Samples from π^θ using the reparameterization trick
- Update θ using an optimisation rule

Loss for the update of θ:

$$\mathcal{L}^\theta = \frac{1}{1-\alpha}\sum_{t\in\mathcal{T}}\sum_{b=1}^{B}\left[\left(c_t^{(b)} + V^{\psi_1,\psi_2}(s_{t+1}^{(b)};\theta) - H_1^{\psi_1}(s_t^{(b)};\theta)\right)_+ \left(\nabla_\theta \log \pi^\theta(a_t^{(b)}|s_t^{(b)})\right)\right]$$

Actor: Update of Policy

Using both the representation theorem [SDR14] and Envelope theorem [MS02]:

$$\nabla_\theta V(s_t;\theta) = \frac{1}{1-\alpha}\mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t)}\left[\left(c_t^\theta + V(s_{t+1}^\theta;\theta) - \lambda^*\right)_+ \left(\nabla_\theta \log \pi^\theta(a_t^\theta|s_t)\right)\right]$$

where λ^* is any α-quantile of $c_t^\theta + V(s_{t+1}^\theta;\theta)$

- Policy gradient method with the gradient of the value function V
- Empirical mean over observed transitions
- Estimation of the VaR$_\alpha$ with $H_1^{\psi_1}$
- Samples from π^θ using the reparameterization trick
- Update θ using an optimisation rule

Loss for the update of θ:

$$\mathcal{L}^\theta = \frac{1}{1-\alpha}\sum_{t\in\mathcal{T}}\sum_{b=1}^{B}\left[\left(c_t^{(b)} + V^{\psi_1,\psi_2}(s_{t+1}^{(b)};\theta) - H_1^{\psi_1}(s_t^{(b)};\theta)\right)_+ \left(\nabla_\theta \log \pi^\theta(a_t^{(b)}|s_t^{(b)})\right)\right]$$

Actor: Update of Policy

Using both the representation theorem [SDR14] and Envelope theorem [MS02]:

$$\nabla_\theta V(s_t;\theta) = \frac{1}{1-\alpha}\mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t)}\left[\left(c_t^\theta + V(s_{t+1}^\theta;\theta) - \lambda^*\right)_+\left(\nabla_\theta \log \pi^\theta(a_t^\theta|s_t)\right)\right]$$

where λ^* is any α-quantile of $c_t^\theta + V(s_{t+1}^\theta;\theta)$

- Policy gradient method with the gradient of the value function V
- Empirical mean over observed transitions
- Estimation of the VaR_α with $H_1^{\psi_1}$
- Samples from π^θ using the reparameterization trick
- Update θ using an optimisation rule

Loss for the update of θ:

$$\mathcal{L}^\theta = \frac{1}{1-\alpha}\sum_{t\in\mathcal{T}}\sum_{b=1}^{B}\left[\left(c_t^{(b)} + V^{\psi_1,\psi_2}(s_{t+1}^{(b)};\theta) - H_1^{\psi_1}(s_t^{(b)};\theta)\right)_+\left(\nabla_\theta \log \pi^\theta(a_t^{(b)}|s_t^{(b)})\right)\right]$$

Actor: Update of Policy

Using both the representation theorem [SDR14] and Envelope theorem [MS02]:

$$\nabla_\theta V(s_t;\theta) = \frac{1}{1-\alpha}\mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t)}\left[\left(c_t^\theta + V(s_{t+1}^\theta;\theta) - \lambda^*\right)_+ \left(\nabla_\theta \log \pi^\theta(a_t^\theta|s_t)\right)\right]$$

where λ^* is any α-quantile of $c_t^\theta + V(s_{t+1}^\theta;\theta)$

- Policy gradient method with the gradient of the value function V
- Empirical mean over observed transitions
- Estimation of the VaR_α with $H_1^{\psi_1}$
- Samples from π^θ using the reparameterization trick
- Update θ using an optimisation rule

Loss for the update of θ:

$$\mathcal{L}^\theta = \frac{1}{1-\alpha}\sum_{t\in\mathcal{T}}\sum_{b=1}^{B}\left[\left(c_t^{(b)} + V^{\psi_1,\psi_2}(s_{t+1}^{(b)};\theta) - H_1^{\psi_1}(s_t^{(b)};\theta)\right)_+ \left(\nabla_\theta \log \pi^\theta(a_t^{(b)}|s_t^{(b)})\right)\right]$$

Actor: Update of Policy

Using both the representation theorem [SDR14] and Envelope theorem [MS02]:

$$\nabla_\theta V(s_t;\theta) = \frac{1}{1-\alpha}\mathbb{E}_{\mathbb{P}^\theta(\cdot,\cdot|s_t)}\left[\left(c_t^\theta + V(s_{t+1}^\theta;\theta) - \lambda^*\right)_+ \left(\nabla_\theta \log \pi^\theta(a_t^\theta|s_t)\right)\right]$$

where λ^* is any α-quantile of $c_t^\theta + V(s_{t+1}^\theta;\theta)$

- Policy gradient method with the gradient of the value function V
- Empirical mean over observed transitions
- Estimation of the VaR_α with $H_1^{\psi_1}$
- Samples from π^θ using the reparameterization trick
- Update θ using an optimisation rule

Loss for the update of θ:

$$\mathcal{L}^\theta = \frac{1}{1-\alpha}\sum_{t\in\mathcal{T}}\sum_{b=1}^{B}\left[\left(c_t^{(b)} + V^{\psi_1,\psi_2}(s_{t+1}^{(b)};\theta) - H_1^{\psi_1}(s_t^{(b)};\theta)\right)_+ \left(\nabla_\theta \log \pi^\theta(a_t^{(b)}|s_t^{(b)})\right)\right]$$

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

$$\mathrm{d}y_t = y_t \left(\sum_{i=0}^{I} \pi_t^{(i)} \frac{\mathrm{d}S_t^{(i)}}{S_t^{(i)}} \right)$$

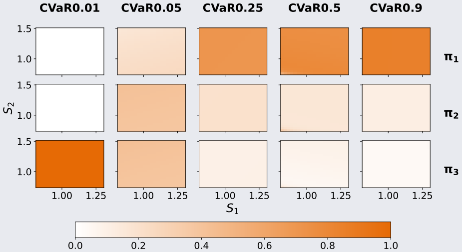
- receives feedback from P&L differences $y_{t+1} - y_t$

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

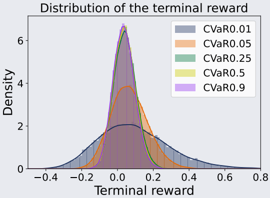
Results - GBM

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

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



Optimal policy

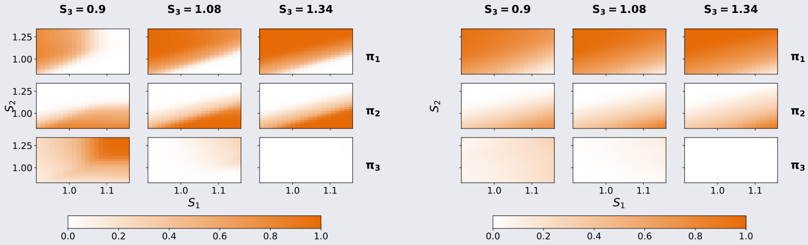


P&L

Results - Mean-Reversion

$$\mathrm{d}X_t^{(i)} = 2\left(\vartheta^{(i)} - X_t^{(i)}\right)\mathrm{d}t + \sigma^{(i)}\mathrm{d}W_t^{(i)}, \qquad S_t^{(i)} = \exp(X_t^{(i)})$$

Drifts and volatilities are $\vartheta = [0.028; 0.053; 0.074]$ and $\sigma = [0.06; 0.12; 0.18]$



$\alpha = 0.5$ $\qquad$ $\alpha = 0.9$

Contributions & Future Directions

Novel setting to solve RL problems in a time-consistent manner using dynamic spectral risk measures

- Efficient approach for estimating dynamic spectral risk with ANNs
- Risk-aware actor-critic algorithm that uses only full episodes

Future directions

- Consider deterministic policies
- Apply the proposed methodology on a real dataset
- Find baselines to reduce the variance of the gradient estimator
- Explore optimisation performances with different characterisations of scoring functions

References

Code: Available soon on https://github.com/acoache
Paper: Available soon
anthonycoache.ca

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

[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.

[MS02] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[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.