Introduction à l'apprentissage par renforcement sensible au risque avec des mesures de risque dynamiques

Anthony Coache (U. Toronto)
anthonycoache.ca

Travaux conjoints avec
Sebastian Jaimungal (U. Toronto)
Álvaro Cartea (Oxford)

Statistical Sciences
UNIVERSITY OF TORONTO

Séminaire STATQAM ⋆ 15 février 2024

Table des matières

Table des matières

Apprentissage par renforcement / *Reinforcement Learning* (RL)

Sous-domaine de l'apprentissage automatique

- Cadre agnostique de modèles pour contrôle fondé sur l'apprentissage
- Apprendre des stratégies optimales à partir d'interactions pour minimiser un signal

Pendant la phase d'apprentissage, l'agent:

↳ interagit avec un environnement virtuel

↳ observe un rendement sous forme de coûts

↳ met à jour son comportement pour découvrir la meilleure stratégie

Applications d'intérêt:

- Allocation de portefeuille, couverture, tarification, trading algorithmique
- Contrôle de robots, véhicules autonomes, contrôle en l'agriculture, systèmes biologiques

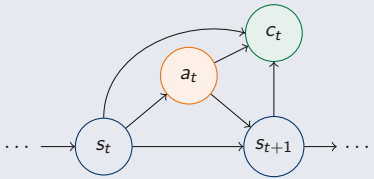
Notation RL

Processus de décision markovien

- $\mathcal{S}$ – Ensemble d'états (*states*)
- $\mathcal{A}$ – Ensemble d'actions (*actions*)
- $\pi^\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ – Politique (*policy*)
- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ – Fonction de transition (*transition probabilities*)
- $c : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ – Fonction de coût (*cost*)

RL vise à minimiser des problèmes de la forme:

$$\min_\theta \; J\Big(\{c_t^\theta\}_t \;\Big|\; s_0 = s\Big)$$



RL sensible au risque

RL standard: $\min_\theta \mathbb{E}[Y^\theta]$, où $Y^\theta = \sum_t \gamma^t c_t^\theta$

- ✕ L'espérance ignore le risque associé aux coûts!

RL sensible au risque utilise des mesures de risque, p. ex.

- ↪ Utilité espérée [Nass et al., 2019]: $\mathbb{E}[U(Y^\theta)]$
- ↪ Contrainte sur le risque [Di Castro et al., 2019]: $\mathbb{E}[Y^\theta]$ tel que $\rho(Y^\theta) \leq c^*$
- ↪ Mesure de risque cohérente [Tamar et al., 2016]: $\rho(Y^\theta)$
- ↪ *Conditional value-at-risk* ou *expected-shortfall* (CVaR_α):

$$\frac{1}{1-\alpha} \int_{[\alpha,1]} \text{VaR}_u(Y^\theta) \mathrm{d}u$$

RL sensible au risque

RL standard: $\min_\theta \mathbb{E}[Y^\theta]$, où $Y^\theta = \sum_t \gamma^t c_t^\theta$

- ✕ L'espérance ignore le risque associé aux coûts!

RL sensible au risque utilise des mesures de risque, p. ex.

- ↳ Utilité espérée [Nass et al., 2019]: $\mathbb{E}[U(Y^\theta)]$
- ↳ Contrainte sur le risque [Di Castro et al., 2019]: $\mathbb{E}[Y^\theta]$ tel que $\rho(Y^\theta) \leq c^*$
- ↳ Mesure de risque cohérente [Tamar et al., 2016]: $\rho(Y^\theta)$
- ↳ *Conditional value-at-risk* ou *expected-shortfall* (CVaR_α):

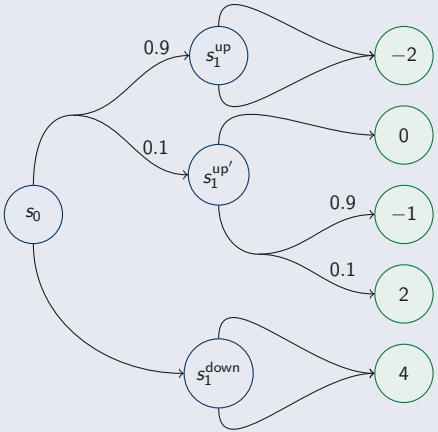
$$\frac{1}{1-\alpha} \int_{[\alpha,1]} \text{VaR}_u(Y^\theta) \mathrm{d}u$$

Problème de cohérence temporelle

Minimisons $\text{CVaR}_{0.9}$ du coût terminal.

- **Actions optimales à s_0**:
  Mouvement vers le haut, puis vers le bas
- **Actions optimales à $s_1^{\text{up}'}$**:
  Mouvement vers le haut

Contradiction avec la stratégie optimale initiale...

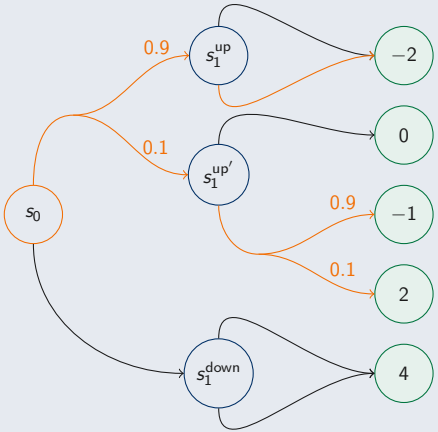


Problème de cohérence temporelle

Minimisons $\text{CVaR}_{0.9}$ du coût terminal.

- **Actions optimales à s_0**:
  Mouvement vers le haut, puis vers le bas
- **Actions optimales à $s_1^{\text{up}'}$**:
  Mouvement vers le haut

Contradiction avec la stratégie optimale initiale...

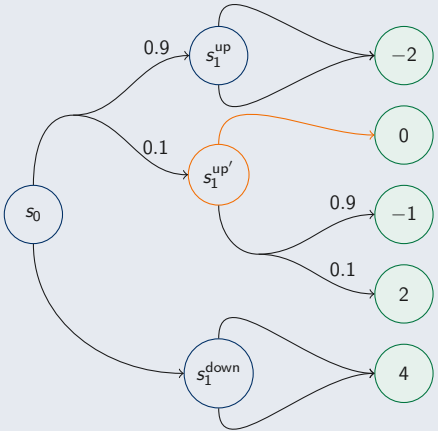


Problème de cohérence temporelle

Minimisons $\text{CVaR}_{0.9}$ du coût terminal.

- **Actions optimales à s_0**:
  Mouvement vers le haut, puis vers le bas
- **Actions optimales à $s_1^{up'}$**:
  Mouvement vers le haut

Contradiction avec la stratégie optimale initiale...



RL sensible au risque avec mesures de risque dynamiques

Optimiser une mesure de risque statique mène à une politique de pré-engagement optimale

Approches récentes pour remédier à ce problème:

- ↳ Programmation dynamique avec coûts latents et actions aléatoires pour des **fonctions de risque conditionnelles** de type Kusuoka [Cheng and Jaimungal, 2022]
- ↳ Approche bayésienne pour tenir compte de l'incertitude du modèle avec coûts latents et des **filtres de risque récursifs** [Bielecki et al., 2023]
- ↳ Itération de la politique pour des **mesures de risque cohérentes récursives** [Bäuerle and Glauner, 2022]
- ↳ Algorithme de type deep Q-learning pour des **mesures de risque expectile dynamiques** [Marzban et al., 2023]
- ↳ etc.

Table des matières

Mesures de risque dynamiques

Considérons

- $\mathcal{T} := \{0, \ldots, T\}$
- $\mathcal{F}_0 \subseteq \cdots \subseteq \mathcal{F}_T$ – Filtration sur $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathcal{T}}, \mathbb{P})$
- $\mathcal{Y}_t := \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ – Variables aléatoires p-intégrables, $\mathcal{F}_t$-mesurables
- $\mathcal{Y}_{t_1,t_2} := \mathcal{Y}_{t_1} \times \cdots \times \mathcal{Y}_{t_2}$ – Séquence de variables aléatoires

Mesure de risque dynamique $\{\rho_{t,T}\}_{t\in\mathcal{T}}$

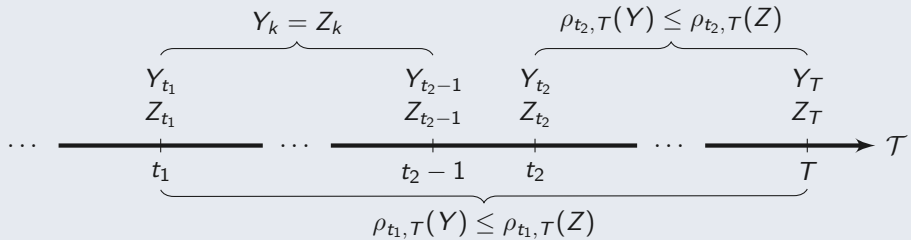
Séquence de mesures de risque conditionnelles $\rho_{t,T} : \mathcal{Y}_{t,T} \to \mathcal{Y}_t$ où

$$\rho_{t,T}(Y) \leq \rho_{t,T}(Z), \text{ pour tout } Y, Z \in \mathcal{Y}_{t,T} \text{ tel que } Y \leq Z$$

Forte cohérence dans le temps

Pour tout $Y_{t_1,T}, Z_{t_1,T} \in \mathcal{Y}_{t_1,T}$ et $0 \leq t_1 < t_2 \leq T$, nous avons

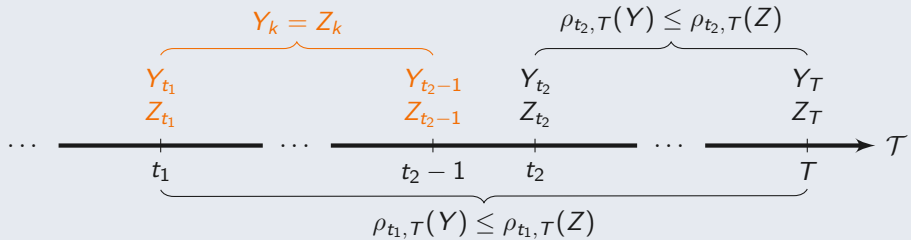
$$\left.\begin{array}{c} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \leq \rho_{t_2,T}(Z_{t_2,T}) \end{array}\right. \Longrightarrow \rho_{t_1,T}(Y_{t_1,T}) \leq \rho_{t_1,T}(Z_{t_1,T}).$$



Forte cohérence dans le temps

Pour tout $Y_{t_1,T}, Z_{t_1,T} \in \mathcal{Y}_{t_1,T}$ et $0 \leq t_1 < t_2 \leq T$, nous avons

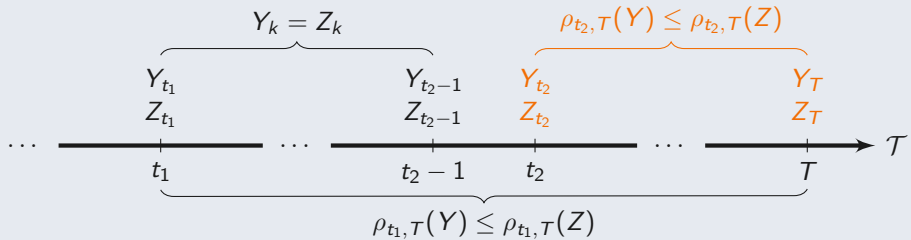
$$\begin{matrix} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \leq \rho_{t_2,T}(Z_{t_2,T}) \end{matrix} \implies \rho_{t_1,T}(Y_{t_1,T}) \leq \rho_{t_1,T}(Z_{t_1,T}).$$



Forte cohérence dans le temps

Pour tout $Y_{t_1,T}, Z_{t_1,T} \in \mathcal{Y}_{t_1,T}$ et $0 \leq t_1 < t_2 \leq T$, nous avons

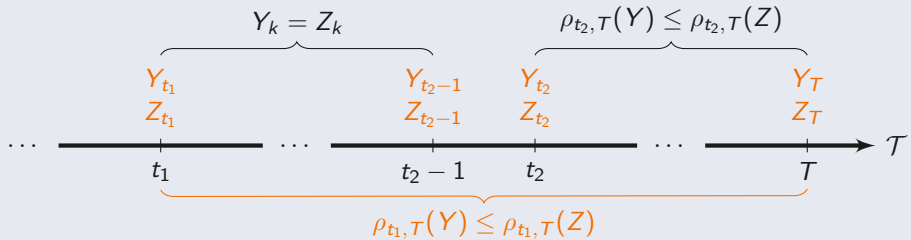
$$\left.\begin{array}{c} Y_{t_1,t_2-1} = Z_{t_1,t_2-1} \\ \rho_{t_2,T}(Y_{t_2,T}) \leq \rho_{t_2,T}(Z_{t_2,T}) \end{array}\right. \Longrightarrow \rho_{t_1,T}(Y_{t_1,T}) \leq \rho_{t_1,T}(Z_{t_1,T}).$$



Forte cohérence dans le temps

Pour tout $Y_{t_1,T}, Z_{t_1,T} \in \mathcal{Y}_{t_1,T}$ et $0 \leq t_1 < t_2 \leq T$, nous avons

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



Cohérence temporelle / *Time-Consistency*

[Thm. 1, Ruszczyński, 2010]

Soit $\{\rho_{t,T}\}_t$ une mesure de risque dynamique monotone, normalisée et invariante par translation.
$\{\rho_{t,T}\}_t$ est fortement cohérente dans le temps si et seulement si elle peut être exprimée à l'aide de mesures de risque conditionnelles à une étape (*one-step conditional risk measures*) $\rho_t : \mathcal{Y}_{t+1} \to \mathcal{Y}_t$:

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

Nous pouvons supposer des propriétés supplémentaires pour ρ_t, p. ex. les axiomes pour mesures de risque convexes, cohérentes, spectrales, etc.

Table des matières

Problème d'optimisation

Problèmes de la forme suivante:

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

où c_t^θ sont des coûts aléatoires $\mathcal{F}_{t+1}$-mesurables
et ρ_t des mesures de risque conditionnelles à une étape.

Programmation dynamique pour la *value function*, c.-à-d. risque futur mobile, pour $s \in \mathcal{S}$:

$$V_t(s;\theta) = \rho_t\Big(\underbrace{c_t^\theta}_{\text{coût courant}} + \underbrace{V_{t+1}(s_{t+1}^\theta;\theta)}_{\text{risque futur pour le prochaines périodes}} \Big| s_t = s\Big)$$

Problème d'optimisation

Problèmes de la forme suivante:

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

où c_t^θ sont des coûts aléatoires $\mathcal{F}_{t+1}$-mesurables
et ρ_t des mesures de risque conditionnelles à une étape.

Programmation dynamique pour la *value function*, c.-à-d. risque futur mobile, pour $s \in \mathcal{S}$:

$$V_t(s;\theta) = \rho_t\Bigg(\underbrace{c_t^\theta}_{\text{coût courant}} + \underbrace{V_{t+1}(s_{t+1}^\theta;\theta)}_{\text{risque futur pour le prochaines périodes}} \Bigg| s_t = s\Bigg)$$

Méthodes policy-gradient

- Nous désirons optimiser V sur les politiques θ via une méthode policy-gradient:

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

Gradient de V [C., Jaimungal, 2023]

Sous certaines hypothèses concernant la forme de l'enveloppe de risque, le gradient de la *value function* à une période $t \in \mathcal{T}$ et l'état $s \in \mathcal{S}$ pour une mesure de risque dynamique convexe est

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

Algorithme **acteur-critique** composé de deux procédures entrelacées:

- Le **critique** calcule la *value function* d'une politique
- L'**acteur** met à jour la politique en gardant V fixe
- Nous modélisons la politique et la *value function* par des réseaux de neurones

Méthodes policy-gradient

- Nous désirons optimiser V sur les politiques θ via une méthode policy-gradient:

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

Gradient de V [C., Jaimungal, 2023]

Sous certaines hypothèses concernant la forme de l'enveloppe de risque, le gradient de la *value function* à une période $t \in \mathcal{T}$ et l'état $s \in \mathcal{S}$ pour une mesure de risque dynamique convexe est

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

Algorithme **acteur-critique** composé de deux procédures entrelacées:

- Le **critique** calcule la *value function* d'une politique
- L'**acteur** met à jour la politique en gardant V fixe
- Nous modélisons la politique et la *value function* par des réseaux de neurones

Résultats principaux

Coache & Jaimungal (2023) Reinforcement Learning with Dynamic Convex Risk Measures. *Mathematical Finance*. DOI: 10.1111/mafi.12388

- Classe de **mesures de risque dynamiques convexes**:

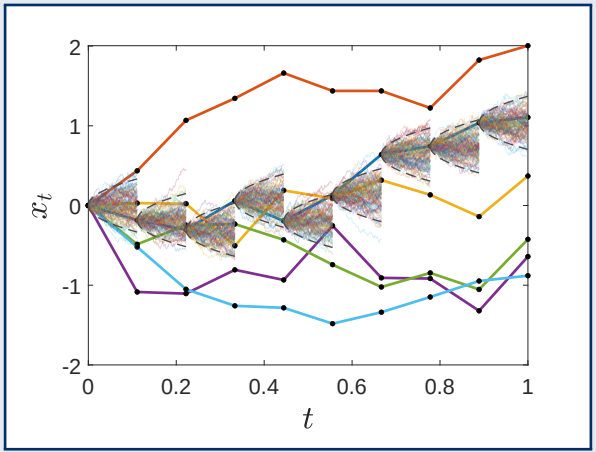
$$\rho_t(Y) = \sup_{\xi_t \in \mathcal{U}(P)} \left\{ \mathbb{E}_t^{\xi_t}[Y] - \rho_t^*(\xi_t) \right\}$$

$$\mathcal{U}(P) = \left\{ \xi : \sum_{\omega} \xi(\omega) P(\omega) = 1,\ \xi \geq 0,\ \underbrace{g_e(\xi, P) = 0, \forall e \in \mathcal{E}}_{\text{fcts affine w.r.t. } \xi},\ \underbrace{f_i(\xi, P) \leq 0, \forall i \in \mathcal{I}}_{\text{fcts convexes w.r.t. } \xi} \right\}$$

- Algorithme acteur-critique avec un cadre de simulations imbriquées
- Dérivation du gradient de V en utilisant le *Envelope theorem for saddle-point problems*
- Théorème d'approximation universelle pour $V_t(s; \theta)$

Simulations imbriquées

- ✕ Les simulations imbriquées sont coûteuses...



Élicitabilité

Amélioration de l'estimation en limitant les fonctions objectifs

ρ est élicitable [Gneiting, 2011] si et seulement si il existe une fonction de score $S : \mathbb{R} \times \mathbb{Y} \to \mathbb{R}$ telle que

$$\rho(Y) = \underset{\mathfrak{a} \in \mathbb{R}}{\arg\min}\, \mathbb{E}_{Y \sim F_Y}\Big[S(\mathfrak{a}, Y)\Big].$$

$\rho(Y)$	$\mathbb{E}$	Médiane	VaR_α	$\mathbf{CVaR}_\alpha$
$S(\mathfrak{a}, y)$	$(\mathfrak{a} - y)^2$	$\lvert\mathfrak{a} - y\rvert$	$\mathbb{1}_{\mathfrak{a} \leq y} - \alpha$	$\emptyset$

Les fonctions non-élicitables peuvent faire partie d'une fonction vectorielle élicitable:

- Élicitabilité de mesures de risque (statiques) spectrales [Fissler and Ziegel, 2016]
- Caractérisation de leur fonction de score S

Élicitabilité

Amélioration de l'estimation en limitant les fonctions objectifs

ρ est élicitable [Gneiting, 2011] si et seulement si il existe une fonction de score $S : \mathbb{R} \times \mathbb{Y} \to \mathbb{R}$ telle que

$$\rho(Y) = \underset{\mathfrak{a} \in \mathbb{R}}{\arg\min}\, \mathbb{E}_{Y \sim F_Y}\Big[S(\mathfrak{a}, Y)\Big].$$

$\rho(Y)$	$\mathbb{E}$	Médiane	VaR_α	$\mathbf{CVaR}_\alpha$
$S(\mathfrak{a}, y)$	$(\mathfrak{a} - y)^2$	$\lvert\mathfrak{a} - y\rvert$	$\mathbb{1}_{\mathfrak{a} \leq y} - \alpha$	$\emptyset$

Les fonctions non-élicitables peuvent faire partie d'une fonction vectorielle élicitable:

- Élicitabilité de mesures de risque (statiques) spectrales [Fissler and Ziegel, 2016]
- Caractérisation de leur fonction de score S

Élicitabilité conditionnelle

Exemple: $(\mathrm{VaR}_\alpha(Y), \mathrm{CVaR}_\alpha(Y))$ est élicitable, c.-à-d.

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

Dans notre problème RL, les coûts dépendent des états $s \in \mathcal{S}$

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

Nous modélisons V avec des réseaux de neurones et minimisons l'espérance empirique

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

Élicitabilité conditionnelle

Exemple: $(\mathrm{VaR}_\alpha(Y), \mathrm{CVaR}_\alpha(Y))$ est élicitable, c.-à-d.

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

Dans notre problème RL, les coûts dépendent des états $s \in \mathcal{S}$

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

Nous modélisons V avec des réseaux de neurones et minimisons l'espérance empirique

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

Élicitabilité conditionnelle

Exemple: $(\mathrm{VaR}_\alpha(Y), \mathrm{CVaR}_\alpha(Y))$ est élicitable, c.-à-d.

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

Dans notre problème RL, les coûts dépendent des états $s \in \mathcal{S}$

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

Nous modélisons V avec des réseaux de neurones et minimisons l'espérance empirique

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

Résultats principaux

Coache, Jaimungal & Cartea (2023) Conditionally Elicitable Dynamic Risk Measures for Deep Reinforcement Learning. *SIAM J. Financial Mathematics*. DOI: 10.1137/22M1527209

- Classe de **mesures de risque dynamiques spectrales**:

$$\rho_t^{\varphi}(Y) = \int_{[0,1]} \mathrm{CVaR}_{\alpha}(Y)\, \varphi_t(\mathrm{d}\alpha)$$

$$\varphi = \sum_{m=1}^{k-1} p_m \delta_{\alpha_m}, \quad \text{où} \quad p_m \in (0,1], \ \sum_{m=1}^{k-1} p_m = 1$$

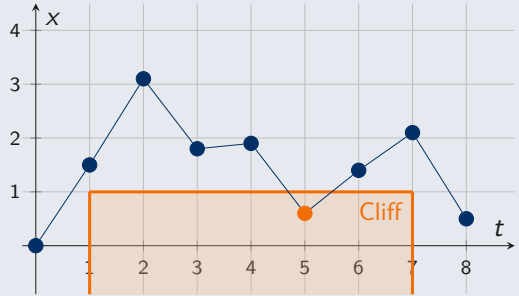
- Algorithme acteur-critique sans simulations imbriquées
- Théorème d'approximation universelle pour $V_t(s;\theta)$

Table des matières

Contrôle de robot

Considérons un robot autonome qui:

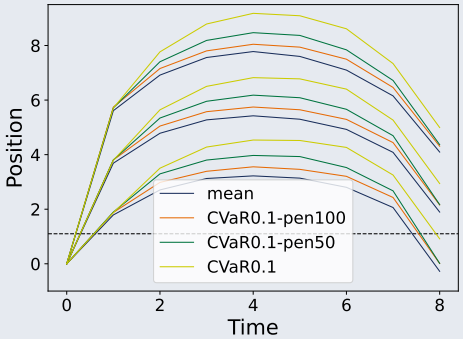
- débute à $(0, 0)$, désire terminer à $(T, 0)$
- effectue des actions $a_t^\theta \sim \pi^\theta = \mathcal{N}(\mu^\theta, \sigma)$
- bouge de (t, x_t) à $(t + 1, x_t + a_t)$
- reçoit des pénalités s'il tombe dans la falaise et s'éloigne de (T, x)

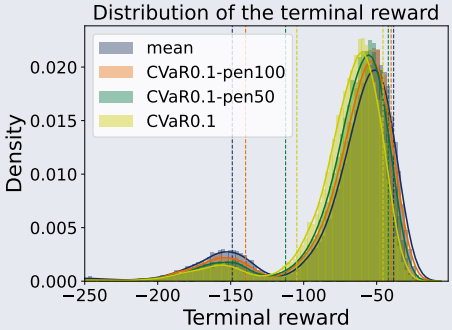


Contrôle de robot

Considérons un robot autonome qui:

- débute à $(0, 0)$, désire terminer à $(T, 0)$
- effectue des actions $a_t^\theta \sim \pi^\theta = \mathcal{N}(\mu^\theta, \sigma)$
- bouge de (t, x_t) à $(t + 1, x_t + a_t)$
- reçoit des pénalités s'il tombe dans la falaise et s'éloigne de (T, x)

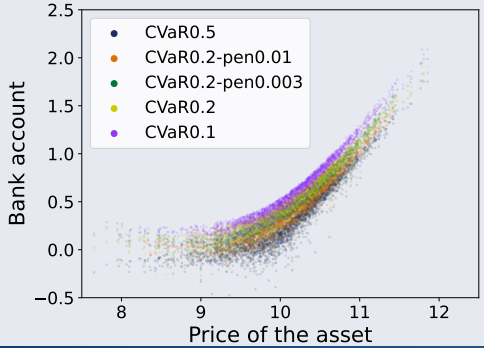




Couverture d'options

Posons une option d'achat dont le prix de l'actif sous-jacent suit un modèle Heston. L'agent:

- vend l'option d'achat, désire couvrir son risque en transigeant seulement l'actif
- observe sa dernière position, sa richesse totale, le prix de l'actif
- effectue des transactions dans un marché avec frictions
- reçoit un coût qui affecte sa richesse



Allocation de portefeuille

Considérons un marché avec d actifs. L'agent:

- observe la période t et le prix des actifs $\{S_t^{(i)}\}_{i=1,\dots,d}$
- décide la proportion de sa richesse $\pi_t^{(i)}$ à investir dans l'actif i
- reçoit un coût via la différence de P&L $y_t - y_{t+1}$, où sa richesse y_t varie selon

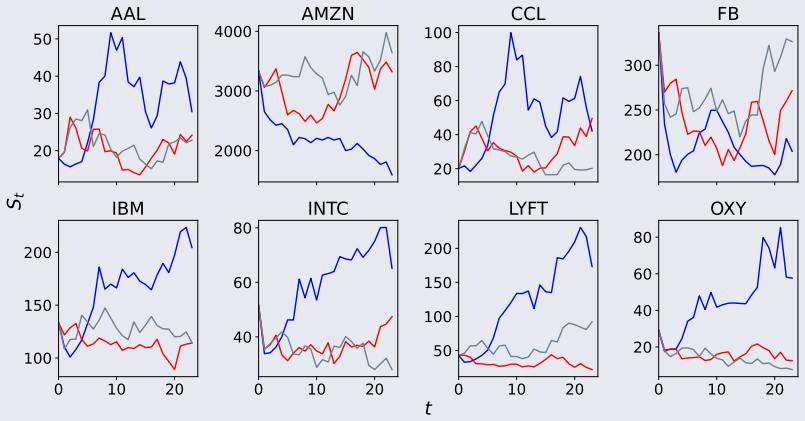
$$\mathrm{d}y_t = y_t \left(\sum_{i=1}^{d} \pi_t^{(i)} \frac{\mathrm{d}S_t^{(i)}}{S_t^{(i)}} \right), \quad y_0 = 1.$$

Nous estimons un modèle de co-intégration avec des données quotidiennes de différents actifs et utilisons le modèle estimé résultant comme générateur de trajectoires de prix

$$\Delta S_\tau = \alpha\beta^\mathsf{T} S_{\tau-1} + \Gamma_1 \Delta S_{\tau-1} + \cdots + \Gamma_{k_{ar}-1} \Delta S_{\tau-k_{ar}+1} + CD_\tau + u_\tau,$$

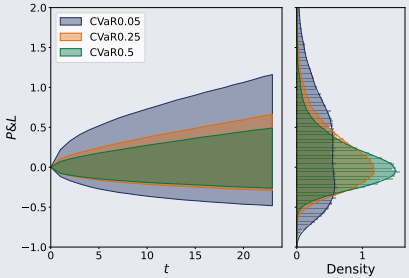
Allocation de portefeuille

Modèle de co-intégration utilisant des données quotidiennes de 8 actifs cotées sur le NASDAQ entre le 31 septembre 2020 et le 31 décembre 2021.



Allocation de portefeuille

Modèle de co-intégration utilisant des données quotidiennes de 8 actifs cotées sur le NASDAQ entre le 31 septembre 2020 et le 31 décembre 2021.



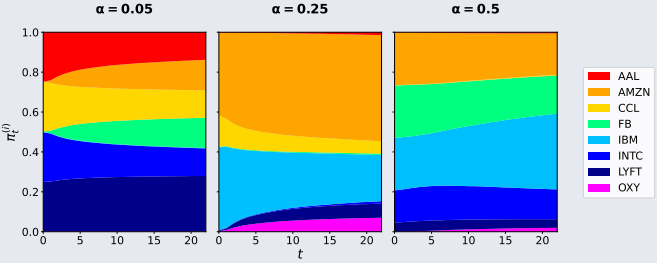


Table des matières

Incertitude sur le modèle

- La phase d'apprentissage doit refléter des évènements similaires à la phase de test
- Qu'arrive-t-il en présence d'incertitude sur le modèle?

Robustification des approches RL:

↳ Algorithme deep RL pour résoudre des problèmes où l'agent minimise une mesure de risque (statique) *RDEU* de variables aléatoires situées dans une **sphère Wasserstein** [Jaimungal et al., 2022]

↳ Algorithme RL robuste, se limitant aux politiques ayant une **divergence KL** à un ϵ près d'une distribution de probabilité d'action de référence [Smirnova et al., 2019]

↳ Approche bayésienne pour **tenir compte de l'incertitude du modèle** avec coûts latents et des filtres de risque récursifs [Bielecki et al., 2023]

↳ etc.

Robustification

Nous incluons des ensembles d'incertitude à l'intérieur des mesures de risque dynamiques [Moresco et al., 2023]

Mesure de risque conditionnelle à une étape robuste

La mesure de risque robuste de $\rho_t(Y)$ sous l'ensemble d'incertitude $\varphi : \mathcal{Y}_{t+1} \to 2^{\mathcal{Y}_{t+1}}$ est

$$\varrho_t^{\varphi}(Y) := \operatorname{ess\,sup} \left\{ \rho_t(Y^{\phi}) \in \mathcal{Y}_t \ : \ Y^{\phi} \in \varphi_Y \right\}.$$

- Fonction quantile $\breve{F}_{Y|_{\mathcal{F}_t}}$ de Y conditionnel à $\mathcal{F}_t$
- $\varphi_Y^{\epsilon} = \left\{ Y^{\phi} \in \mathcal{Y}_{t+1} \ : \ \left\| \breve{F}_{Y|_{\mathcal{F}_t}} - \breve{F}_{Y^{\phi}|_{\mathcal{F}_t}} \right\|_2 \leq \epsilon \right\}$

Robustification

Programmation dynamique: $V_t(s;\theta) = \varrho_t\Big(Y_t^\theta \;\Big|\; s_t = s\Big)$, où $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

Nous supposons que ϱ_t est une mesure de distorsion de risque robuste

$$\varrho_t(Y) = \operatorname*{ess\,sup}_{Y^\phi \in \varphi_Y^{\epsilon_t}} \left\langle \gamma_t, \breve{F}_{Y^\phi|_{\mathcal{F}_t}} \right\rangle, \quad \text{où } \gamma_t \text{ est linéaire par morceaux}$$

- prend en compte l'incertitude du modèle
- permet des stratégies prudentes (*risk-averse*) et hostiles (*risk-seeking*)
- est élicitable
- devient un problème d'optimisation convexe avec les fonctions quantiles [Pesenti and Jaimungal, 2023]

Résultats principaux

Proposition

Avec $\varphi^{\epsilon_t}_{Y^\theta_t} = \{Y^\phi \in \mathcal{Y}_{t+1} : \|\breve{F}_{Y^\theta_t|\mathcal{F}_t} - \breve{F}_{Y^\phi|\mathcal{F}_t}\|_2 \leq \epsilon_t\}$ et une distorsion γ_t non décroissante, la fonction quantile optimale est donnée par $\breve{F}^*(\cdot|s) = \breve{F}_{Y^\theta_t}(\cdot|s) + \frac{\epsilon_t \gamma_t}{\|\gamma_t\|_2}$.

- Cette robustification est équivalente à moduler la fonction de coût

Proposition

Avec $\varphi^{\epsilon_t}_{Y^\theta_t} = \{Y^\phi \in \mathcal{Y}_{t+1} : \|\breve{F}_{Y^\theta_t|\mathcal{F}_t} - \breve{F}_{Y^\phi|\mathcal{F}_t}\|_2 \leq \epsilon_t,\ \|\breve{F}_{Y^\phi|\mathcal{F}_t}\|_2 = \|\breve{F}_{Y^\theta_t|\mathcal{F}_t}\|_2\}$ et une distorsion γ_t non décroissante, la fonction quantile optimale est donnée par

$$\breve{F}^*(\cdot|s) = \frac{\lambda^\theta_t \breve{F}_{Y^\theta_t}(\cdot|s) + \gamma_t}{b_{\lambda^\theta_t}}.$$

Résultats principaux

Proposition

Avec $\varphi^{\epsilon_t}_{Y^\theta_t} = \{Y^\phi \in \mathcal{Y}_{t+1} : \|\breve{F}_{Y^\theta_t|\mathcal{F}_t} - \breve{F}_{Y^\phi|\mathcal{F}_t}\|_2 \leq \epsilon_t\}$ et une distorsion γ_t non décroissante, la fonction quantile optimale est donnée par $\breve{F}^*(\cdot|s) = \breve{F}_{Y^\theta_t}(\cdot|s) + \frac{\epsilon_t \gamma_t}{\|\gamma_t\|_2}$.

- Cette robustification est équivalente à moduler la fonction de coût

Proposition

Avec $\varphi^{\epsilon_t}_{Y^\theta_t} = \{Y^\phi \in \mathcal{Y}_{t+1} : \|\breve{F}_{Y^\theta_t|\mathcal{F}_t} - \breve{F}_{Y^\phi|\mathcal{F}_t}\|_2 \leq \epsilon_t,\ \|\breve{F}_{Y^\phi|\mathcal{F}_t}\|_2 = \|\breve{F}_{Y^\theta_t|\mathcal{F}_t}\|_2\}$ et une distorsion γ_t non décroissante, la fonction quantile optimale est donnée par

$$\breve{F}^*(\cdot|s) = \frac{\lambda^\theta_t \breve{F}_{Y^\theta_t}(\cdot|s) + \gamma_t}{b_{\lambda^\theta_t}}.$$

Algorithme

Étape 1: Estimation de la distribution $F_{Y_t^\theta}$, où $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta; \theta)$

↳ *Continuous ranked probability score*:

$$F_Y = \underset{F \in \mathbb{F}}{\arg\min}\, \mathbb{E}_{Y \sim F_Y}\Big[S(F, Y)\Big] \quad \text{with} \quad S(F, z) = \int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y \geq z}\Big)^2 \mathrm{d}y$$

Étape 2: Estimation de $V_t(s; \theta) = \underset{Y^\phi \in \varphi_Y^{\epsilon_t}}{\operatorname{ess\,sup}} \left\langle \gamma_t, \breve{F}_{Y^\phi|_{s_t}} \right\rangle$ par élicitabilité

↳ La fonction quantile optimale $\breve{F}^*$ est connue

Étape 3: Mise à jour de π^θ via une méthode policy-gradient

↳ Problème d'optimisation convexe sur l'espace des fonctions quantile

Algorithme

Étape 1: Estimation de la distribution $F_{Y_t^\theta}$, où $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta; \theta)$

↳ *Continuous ranked probability score*:

$$F_Y = \underset{F \in \mathbb{F}}{\arg\min}\, \mathbb{E}_{Y \sim F_Y}\Big[S(F, Y)\Big] \quad \text{with} \quad S(F, z) = \int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y \geq z}\Big)^2 \mathrm{d}y$$

Étape 2: Estimation de $V_t(s; \theta) = \underset{Y^\phi \in \varphi_Y^{\epsilon_t}}{\operatorname{ess\,sup}} \left\langle \gamma_t, \check{F}_{Y^\phi|_{s_t}} \right\rangle$ par élicitabilité

↳ La fonction quantile optimale $\check{F}^*$ est connue

Étape 3: Mise à jour de π^θ via une méthode policy-gradient

↳ Problème d'optimisation convexe sur l'espace des fonctions quantile

Algorithme

Étape 1: Estimation de la distribution $F_{Y_t^\theta}$, où $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta; \theta)$

↳ *Continuous ranked probability score*:

$$F_Y = \underset{F \in \mathbb{F}}{\arg\min}\, \mathbb{E}_{Y \sim F_Y}\Big[S(F, Y)\Big] \quad \text{with} \quad S(F, z) = \int_{\mathbb{R}} \Big(F(y) - \mathbb{1}_{y \geq z}\Big)^2 \mathrm{d}y$$

Étape 2: Estimation de $V_t(s; \theta) = \underset{Y^\phi \in \varphi_Y^{\epsilon_t}}{\operatorname{ess\,sup}} \left\langle \gamma_t, \breve{F}_{Y^\phi|_{s_t}} \right\rangle$ par élicitabilité

↳ La fonction quantile optimale $\breve{F}^*$ est connue

Étape 3: Mise à jour de π^θ via une méthode policy-gradient

↳ Problème d'optimisation convexe sur l'espace des fonctions quantile

Résultats principaux

Coache & Jaimungal (travaux en cours) Robust Reinforcement Learning with Dynamic Risk Measures.

- Classe de **mesures de risque dynamiques robustes de distortion**
- Algorithme RL prenant en compte simultanément le risque et l'incertitude du modèle
- Dérivation du gradient d'une politique déterministe (*deterministic policy gradient*)
- Résultats pour des fonctions de distorsion générales γ_t avec des projections isotoniques

Table des matières

Contributions

Développement de méthodologies profondes pour l'apprentissage par renforcement sensible au risque avec des mesures de risque dynamiques

- Algorithmes RL pour plusieurs classes de **mesures de risque dynamiques convexes**
- Cadre utilisant des **fonctions élicitables** pour éliminer les simulations imbriquées
- **Robustification** pour se protéger contre l'incertitude sur le modèle

Pistes de recherche

↳ Robustification sous d'autres ensembles d'incertitude, p. ex. divergence KL

↳ Modification des réseaux de neurones, p. ex. RNNs, méchanisme d'attention

↳ Cadre avec une multitude d'agents

- Étude des situations d'équilibre avec des agent cohérents dans le temps
- Idées en théorie des jeux à champ moyen, c.-à-d. MFGs

↳ RL inverse avec mesures de risque dynamiques

- Peut-on apprendre la fonction objectif d'un agent en observant leur stratégie?
- Travaux en cours avec Ziteng Cheng et Sebastian Jaimungal

Cheng, Coache & Jaimungal (2023) Eliciting Risk Aversion with Inverse Reinforcement Learning via Interactive Questioning. *arXiv*. DOI: 10.48550/arXiv.2308.08427

Merci!

Pour plus d'informations: `anthonycoache.ca`

Table des matières

Convex Risk Measures

Consider $\mathcal{Y} := \mathcal{L}_p(\Omega, \mathcal{F}, P)$ – p-integrable, $\mathcal{F}$-measurable random variables
A convex risk measure $\rho : \mathcal{Y} \to \mathbb{R}$ [Föllmer and Schied, 2002] is

- **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 [Shapiro et al., 2014]

A convex risk measure is proper and lower semicontinuous iff there exists a set

$$\mathcal{U}(P) \subset \left\{ \xi : \sum_{\omega} \xi(\omega) P(\omega) = 1, \ \xi \geq 0 \right\},$$

often referred to as risk envelope, such that

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

Gradient of V for Dynamic Convex Risk

Derivation of the gradient of $V_t(s;\theta)$ using the Envelope theorem for saddle-point problems

Gradient of V [C., Jaimungal, 2023]

Under regularity assumptions on the gradient of $\mathbb{P}^\theta(a, s'|s)$, for any state $s \in \mathcal{S}$, the gradient of the value function is given by

$$\nabla_\theta V_t(s;\theta) = \mathbb{E}_t^{\xi^*}\left[\left(c(s, a_t^\theta, s_{t+1}^\theta) + V_{t+1}(s_{t+1}^\theta;\theta) - \lambda^*\right)\nabla_\theta \log \pi^\theta(a_t^\theta|s_t = s) + \nabla_\theta V_{t+1}(s_{t+1}^\theta;\theta)\right]$$
$$- \nabla_\theta \rho_t^*(\xi^*) - \sum_{e\in\mathcal{E}}\left(\lambda^{*,\mathcal{E}}(e)\nabla_\theta g_e(\xi^*, \mathbb{P}^\theta)\right) - \sum_{i\in\mathcal{I}}\left(\lambda^{*,\mathcal{I}}(i)\nabla_\theta f_i(\xi^*, \mathbb{P}^\theta)\right),$$

where $(\xi^*, \lambda^*, \lambda^{*,\mathcal{E}}, \lambda^{*,\mathcal{I}})$ is any saddle-point of the Lagrangian function.

Universal Approximation Theorem for Dynamic Convex Risk

Approximation of the dynamic risk to any arbitrary accuracy using NNs

Approximation of V [C., Jaimungal, 2023]

Let π^θ denote a fixed policy, with corresponding value function $V_t(s;\theta)$. Then for any $\epsilon^* > 0$, there exists a NN V_t^φ such that for any $t \in \mathcal{T}$, we have

$$\operatorname*{ess\,sup}_{s\in\mathcal{S}} \left\| V_t(s;\theta) - V_t^\varphi(s;\theta) \right\| < \epsilon^*.$$

Elicitable Mappings

Expectation is elicitable: $\mathbb{E}[Y] = \underset{\mathfrak{a}\in\mathbb{R}}{\arg\min}\, \mathbb{E}_{Y\sim F_Y}\Big[(\mathfrak{a} - Y)^2\Big]$

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

$$S(\mathfrak{a}_1, \mathfrak{a}_2, y) = \Big(\mathbb{1}_{y\le\mathfrak{a}_1} - \alpha\Big)\Big(G_1(\mathfrak{a}_1) - G_1(y)\Big) - G_2(\mathfrak{a}_2) + G_2(y) + G_2'(\mathfrak{a}_2)\left[\mathfrak{a}_2 + \frac{1}{1-\alpha}\Big(\mathfrak{a}_1\Big(\mathbb{1}_{y>\mathfrak{a}_1} - (1-\alpha)\Big) - y\,\mathbb{1}_{y>\mathfrak{a}_1}\Big)\right]$$

Conditional maps are elicitable:

$$\rho(Y \mid s_t = s) = \underset{h\,:\,\mathcal{S}\to\mathbb{R}}{\arg\min}\, \mathbb{E}_{Y\sim F_Y}\Big[S(h(s), Y)\Big]$$

Any CDF is elicitable:

$$F_Y = \underset{F\in\mathbb{F}}{\arg\min}\, \mathbb{E}_{Y\sim F_Y}\left[\int_{\mathbb{R}}\Big(F(y) - \mathbb{1}_{y\ge Y}\Big)^2 \mathrm{d}y\right]$$

Universal Approximation Theorem of Dynamic Spectral Risk

Approximation of the dynamic risk to any arbitrary accuracy using NNs

Theorem [C., Jaimungal, Cartea, 2023]

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

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

Hedging with Friction Example

The asset price $(S_t)_{t\in\mathcal{T}}$ is simulated using the Milstein discretization scheme:

$$\begin{aligned} \mathrm{d}S_t &= \mu\, S_t \mathrm{d}t + \sqrt{\nu_t}\, S_t\, \mathrm{d}W_t^S, \\ \mathrm{d}\nu_t &= \kappa\, (\vartheta - \nu_t)\, \mathrm{d}t + \varsigma \sqrt{\nu_t}\, \mathrm{d}W_t^\nu \end{aligned}$$

The agent:

- sells a call option, aims to hedge it trading solely in the underlying asset
- observes the asset price and its previous hedge position
- takes an action a_t^θ, i.e. the number of shares to hold over the next time interval

Bank account B

$$\begin{cases} B_{t^+} = B_t - \left(a_t^\theta - a_{t-1}^\theta\right) S_t - \left|a_t^\theta - a_{t-1}^\theta\right| \epsilon \\ B_{t+1} = e^{r\Delta t} B_{t^+} \\ B_T = e^{r\Delta t} B_{(T-1)^+} + a_{T-1}^\theta S_T - \left|a_{T-1}^\theta\right| \epsilon - (S_T - K)_+ \end{cases}$$

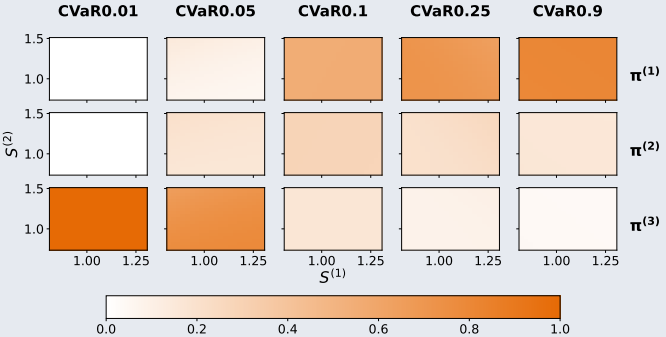
Wealth y

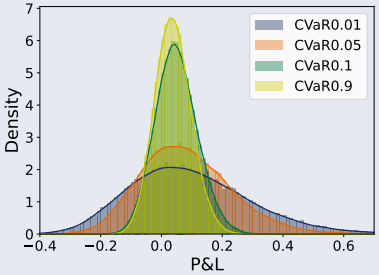
$$\begin{cases} y_{t^+} = B_{t^+} + a_t^\theta S_t \\ y_{t+1} = B_{t+1} + a_t^\theta S_{t+1} \\ y_T = B_T \end{cases}$$

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 $\sigma = [0.06; 0.12; 0.18]$

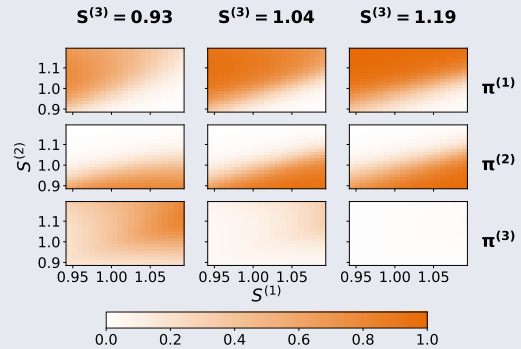


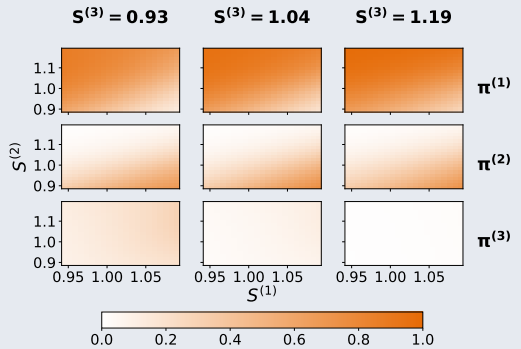


Portfolio Allocation

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

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





Optimal Quantile Function

Consider the value function for dynamic robust risk measures, where γ_t is nondecreasing and

$$\varphi^{\epsilon_t}_{Y^\theta_t} = \left\{ Y^\phi \in \mathcal{Y}_{t+1} : \|\breve{F}_{Y^\theta_t|\mathcal{F}_t} - \breve{F}_{Y^\phi|\mathcal{F}_t}\|_2 \le \epsilon_t,\ \|\breve{F}_{Y^\phi|\mathcal{F}_t}\|_2 = \|\breve{F}_{Y^\theta_t|\mathcal{F}_t}\|_2 \right\}.$$

The optimal quantile function is then given by $\breve{F}^*_\phi(u|s) = \frac{\lambda^* \breve{F}_{Y^\theta_t}(u|s) + \gamma_t(u)}{b_{\lambda^*}}$, where

$$b_{\lambda^*} = \frac{\left\| \lambda^* \breve{F}_{Y^\theta_t}(\cdot|s) + \gamma_t \right\|}{\left\| \breve{F}_{Y^\theta_t}(\cdot|s) \right\|}, \quad \lambda^* = \frac{-2\left\langle \breve{F}_{Y^\theta_t}(\cdot|s), \gamma_t \right\rangle + \sqrt{\Delta}}{2\left\| \breve{F}_{Y^\theta_t}(\cdot|s) \right\|^2}, \quad K = \left\| \breve{F}_{Y^\theta_t}(\cdot|s) \right\|^2 - \frac{\epsilon_t^2}{2}$$

$$\Delta = 4\left(\left\langle \breve{F}_{Y^\theta_t}(\cdot|s), \gamma_t \right\rangle^2 + \left\| \breve{F}_{Y^\theta_t}(\cdot|s) \right\|^2 \frac{\left\| \breve{F}_{Y^\theta_t}(\cdot|s) \right\|^2 \left\langle \breve{F}_{Y^\theta_t}(\cdot|s), \gamma_t \right\rangle^2 - K^2 \|\gamma_t\|^2}{K^2 - \left\| \breve{F}_{Y^\theta_t}(\cdot|s) \right\|^4} \right).$$

The optimal solution remains valid with $\lambda^* = 0$ if the tolerance ϵ_t satisfies

$$\epsilon_t^2 > 2\left\| \breve{F}_{Y^\theta_t}(\cdot|s) \right\|^2 \left(1 - \frac{\left\langle \gamma_t, \breve{F}_{Y^\theta_t}(\cdot|s) \right\rangle}{\left\| \breve{F}_{Y^\theta_t}(\cdot|s) \right\| \|\gamma_t\|} \right).$$

Deterministic Gradient for Robust Dynamic Risk

Consider the value function for dynamic robust risk measures, where γ_t is nondecreasing and

$$\varphi^{\epsilon_t}_{Y^\theta_t} = \left\{ Y^\phi \in \mathcal{Y}_{t+1} \,:\, \|\breve{F}_{Y^\theta_t|_{\mathcal{F}_t}} - \breve{F}_{Y^\phi|_{\mathcal{F}_t}}\|_2 \le \epsilon_t,\ \|\breve{F}_{Y^\phi|_{\mathcal{F}_t}}\|_2 = \|\breve{F}_{Y^\theta_t|_{\mathcal{F}_t}}\|_2 \right\}.$$

The gradient of the value function is given by

$$\begin{aligned}
\nabla_\theta V_t(s;\theta) &= \nabla_\theta Q_t(s, \pi^\theta(s);\theta) \\
&= \left(\nabla_a Q_t(s,a;\theta)\Big|_{a=\pi^\theta(s)} - \frac{(b_{\lambda^*} - \lambda^*)^2}{b_{\lambda^*}} \mathbb{E}_{t,s}\left[Y^\theta_t \left. \frac{\nabla_a F_{Y^\theta_t}(x|s,a)}{\nabla_x F_{Y^\theta_t}(x|s,a)} \right|_{(x,a)=(Y^\theta_t, \pi^\theta(s))} \right] \right) \nabla_\theta \pi^\theta(s).
\end{aligned}$$

Algorithm – Dynamic Convex Risk

Algorithm 1: Actor-critic algorithm – Nested simulation approach

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

Set initial learning rates for ϕ, θ;
for each iteration $k = 1, \ldots, K$ **do**
    for each epoch $k^\phi = 1, \ldots, K^\phi$ **do**
        Simulate a mini-batch of B^ϕ episodes induced by π^θ;
        Generate M^ϕ additional (inner) transitions induced by π^θ;
        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 rate for ϕ;
    for each epoch $k^\theta = 1, \ldots, K^\theta$ **do**
        Simulate a mini-batch of B^θ episodes induced by π^θ;
        Generate M^θ 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;

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

Algorithm – Dynamic Elicitable Risk

Algorithm 2: Actor-critic algorithm – Elicitable approach

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

Set initial learning rates for ϕ, θ;
for each iteration $k = 1, \ldots, K$ **do**
    for each epoch $k^\phi = 1, \ldots, K^\phi$ **do**
        Simulate a mini-batch of B^ϕ episodes induced by π^θ;
        Compute the loss $\mathcal{L}(\phi)$: minimization of the expected consistent score;
        Update ϕ by performing an Adam optimisation step, tune the learning rate for ϕ;
        if $k^\phi \bmod K^* = 0$ **then**
            Update the target networks $\tilde{\phi}$;
    for each epoch $k^\theta = 1, \ldots, K^\theta$ **do**
        Simulate a mini-batch of $\lceil B^\theta/(1-\alpha) \rceil$ 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;

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

Algorithm – Dynamic Robust Risk

- **Adversary** estimates the distribution of the costs-to-go
- **Critic** calculates the value function of the given policy
- **Actor** updates the current policy

Input: ANNs $\pi^{\theta}, V^{\phi}, F^{\vartheta}$, numbers of epochs K's

for each iteration $k = 1, \ldots, K$ **do**
  while convergence is not achieved do
    for $k^{\vartheta} = 1, \ldots, K^{\vartheta}$ **do**
      Adversary: minimization of the expected scoring rule for F to update ϑ;
    for $k^{\phi} = 1, \ldots, K^{\phi}$ **do**
      Critic: minimization of the expected consistent score to update ϕ;
  for $k^{\theta} = 1, \ldots, K^{\theta}$ **do**
    Actor: policy gradient to update θ ;

Output: Optimal policy π^{θ}, its value function $V_t(s;\theta)$, and the CDF $F_{Y_t^{\theta}}$

Inverse RL Setup (Infinite-Horizon Setting)

We assume finite state $\mathcal{S}$ and action $\mathcal{A}$ spaces, and a risk-aversion modeled by a dynamic spectral risk measure (defined through φ).

At each iteration,

- the **learner designs a controlled transition matrix** $G \in \mathcal{T}(\mathcal{S})^{\mathcal{A}}$
- the **agent demonstrates its optimal policy** satisfying

$$\pi^*(s) \in \arg\min_{a \in \mathcal{A}} \rho_\varphi \left(V^* \left(X_G^{s,a} \right) \right), \quad s \in \mathcal{S},$$

where

$$V^*(s) = C(s) + r \min_{a \in \text{Å}} \left\{ \rho_\varphi \left(V^* \left(X_G^{s,a} \right) \right) \right\}, \quad s \in \mathcal{S}.$$

The learner aims to identify (C, r, φ) from a set of candidates by only observing $(G_n, \pi_n^*)_{n=1}^N$

Environment Design Approaches

We define:

- **Regret**: $\Phi^s(a; G, \ell) = \sum_{s \in \mathcal{S}} C_\ell(s) + r_\ell \rho_{\varphi_\ell} \left(V^*_{G,\ell}\left(X_G^{s,a}\right)\right) - V^*_{G,\ell}(s)$
  - Difference between optimal and actual reward
- **Distinguishing power**: $\Psi(G, i, j) = -\sum_{s \in \mathcal{S}} \Phi^s\left(\pi^{*,i}; G, j\right) \Phi^s\left(\pi^{*,j}; G, i\right)$
  - $\Psi \leq 0$, and large $|\Psi|$ indicates that G easily distinguishes i and j
- **Gibbs measure on the risk candidates**: $\mathbb{Q}_N(\{\ell\}) \propto \exp\left(-k \sum_{n=1}^{N} \Phi(\pi^*_n; G_n, \ell)\right)$
  - Current belief of the risk candidates being the true risk-aversion

Main Contributions

Cheng, Coache & Jaimungal (2023) Eliciting Risk Aversion with Inverse Reinforcement Learning via Interactive Questioning. *arXiv*. DOI: 10.48550/arXiv.2308.08427

- Existence of a distinguishing environment
- Gibbs measure $\mathbb{Q}_n$ is consistent, i.e. if $G_n \sim \text{Unif}(\mathcal{T}(\mathcal{S})^{\mathcal{A}})$ iid and (C_0, r_0, φ_0) be the true risk-aversion, then

$$\lim_{N\to\infty} \mathbb{Q}_N(\{0\}) = 1 \text{ a.s.}$$

- Environment design approaches to improve the learning speed:
  - $G_{N+1} \in \arg\min_G \Psi(G, \eta^*, \zeta^*)$, where $\eta = \arg\max \mathbb{Q}_N(i)$, $\zeta = \arg\max \mathbb{Q}_N(i) \setminus \{\eta\}$
  - $G_{N+1} \in \arg\min_G \mathbb{E}[\Psi(G, \eta, \zeta)]$, where $\eta \sim \mathbb{Q}_N$, $\zeta \sim \mathbb{Q}_N | \zeta \neq \eta$