Robust Reinforcement Learning for Dynamic Risk Measures

Anthony Coache (University of Toronto)

anthonycoache.ca

Joint work with

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

SIAM Conference on Financial Mathematics and Engineering ⋆ June 6–9, 2023 ⋆ Philadelphia, USA

Statistical Sciences
UNIVERSITY OF TORONTO

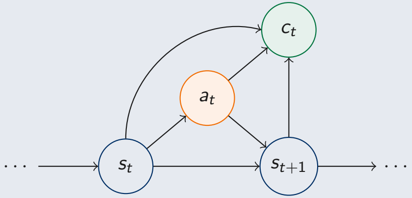
siam 2023 Conference on Financial Mathematics and Engineering

Reinforcement Learning (RL)

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

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

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



Robust Risk-Aware RL

Risk-aware RL with static risk measures as objectives instead of a (risk-neutral) expectation:

- Expectation ignores the risk of the costs!
- Optimizing static risk measures leads to optimal precommitment policies!

Risk-aware RL with dynamic risk, e.g.:

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

Robust RL to account for uncertainties, e.g.:

- KL divergence [Smirnova et al., 2019], Wasserstein ball [Jaimungal et al., 2022], Bayesian approach [Bielecki et al., 2022]

Our goal is to develop a practical, computational RL framework that simultaneously

- accounts for model uncertainty
- accounts for risk in a time-consistent manner

Robust Risk-Aware RL

Risk-aware RL with static risk measures as objectives instead of a (risk-neutral) expectation:

- Expectation ignores the risk of the costs!
- Optimizing static risk measures leads to optimal precommitment policies!

Risk-aware RL with dynamic risk, e.g.:

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

Robust RL to account for uncertainties, e.g.:

- KL divergence [Smirnova et al., 2019], Wasserstein ball [Jaimungal et al., 2022], Bayesian approach [Bielecki et al., 2022]

Our goal is to develop a practical, computational RL framework that simultaneously

- accounts for model uncertainty
- accounts for risk in a time-consistent manner

Robust Risk-Aware RL

Risk-aware RL with static risk measures as objectives instead of a (risk-neutral) expectation:

- Expectation ignores the risk of the costs!
- Optimizing static risk measures leads to optimal precommitment policies!

Risk-aware RL with dynamic risk, e.g.:

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

Robust RL to account for uncertainties, e.g.:

- KL divergence [Smirnova et al., 2019], Wasserstein ball [Jaimungal et al., 2022], Bayesian approach [Bielecki et al., 2022]

Our goal is to develop a practical, computational RL framework that simultaneously

- accounts for model uncertainty
- accounts for risk in a time-consistent manner

Robust Risk-Aware RL

Risk-aware RL with static risk measures as objectives instead of a (risk-neutral) expectation:

- Expectation ignores the risk of the costs!
- Optimizing static risk measures leads to optimal precommitment policies!

Risk-aware RL with dynamic risk, e.g.:

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

Robust RL to account for uncertainties, e.g.:

- KL divergence [Smirnova et al., 2019], Wasserstein ball [Jaimungal et al., 2022], Bayesian approach [Bielecki et al., 2022]

Our goal is to develop a practical, computational RL framework that simultaneously

- accounts for model uncertainty
- accounts for risk in a time-consistent manner

Dynamic Risk Measures

Consider

- $\mathcal{F}_0 \subseteq \cdots \subseteq \mathcal{F}_T$ – Filtration on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_t, \mathbb{P})$
- $\mathcal{Y}_t := \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ – p-integrable, $\mathcal{F}_t$-measurable random variables
- $\mathcal{Y}_{t_1,t_2} := \mathcal{Y}_{t_1} \times \cdots \times \mathcal{Y}_{t_2}$ – Sequence of random variables

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

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

- $\mathcal{F}_t$-measurable charge one would be willing to incur instead of a sequence of future costs

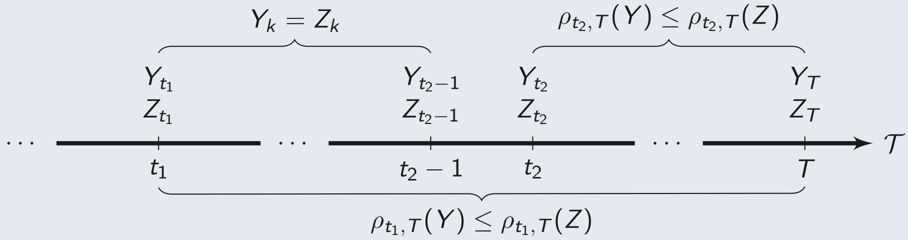
Time-Consistency

Strong time-consistency

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

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

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



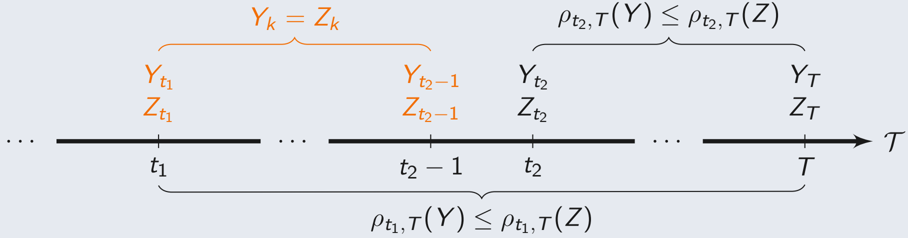
Time-Consistency

Strong time-consistency

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

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

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



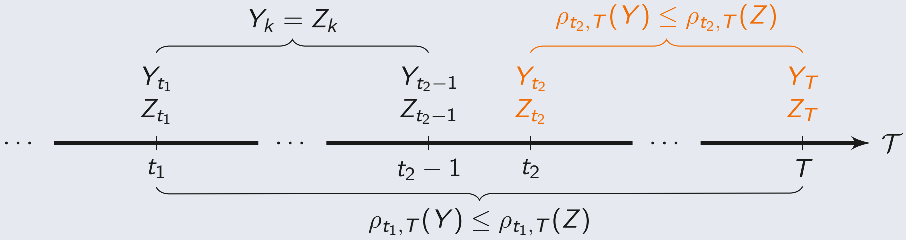
Time-Consistency

Strong time-consistency

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

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

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



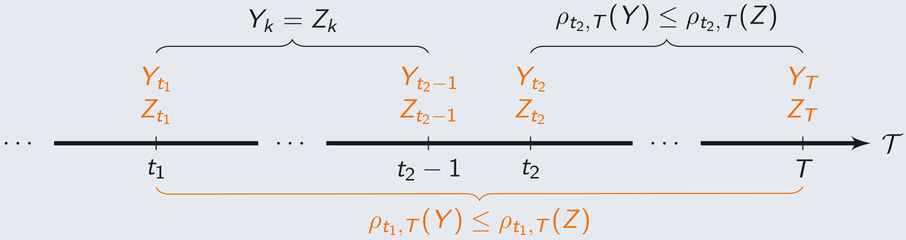
Time-Consistency

Strong time-consistency

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

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

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



Time-Consistency

[Thm. 1, Ruszczyński, 2010]
Let $\{\rho_{t,T}\}_t$ be a dynamic risk measure satisfying for any $Y, Z \in \mathcal{Y}_{t,T}$

- $\rho_{t,T}(Y) \leq \rho_{t,T}(Z)$ for all $Y \leq Z$;
- $\rho_{t,T}(Y_t, Y_{t+1}, \ldots, Y_T) = Y_t + \rho_{t,T}(0, Y_{t+1}, \ldots, Y_T)$;
- $\rho_{t,T}(0, \ldots, 0) = 0$;

Then $\{\rho_{t,T}\}_t$ is strongly time-consistent iff it may be expressed as

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

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

Robustifying the Dynamic Risk

Let $\breve{F}_Y$ be the quantile function of Y

- 2-Wasserstein distance: $d_2[Y, Z] = \left(\int_0^1 \left| \breve{F}_Y(u) - \breve{F}_Z(u) \right|^2 \mathrm{d}u \right)^{1/2}$
- Distortion risk measure: $\rho^{\gamma}(Y) = \mathbb{E}\left[Y \, \gamma(F_Y(Y)) \right] = \int_0^1 \gamma(u) \breve{F}_Y(u) \mathrm{d}u$

We work with 2-Wasserstein-robust distortion risk measures (with piecewise constant γ)

$$\rho^{\gamma,\epsilon}(Y) = \sup_{Y^{\phi} \in \varphi_Y^{\epsilon}} \mathbb{E}\left[Y^{\phi} \, \gamma\left(F_{Y^{\phi}}(Y^{\phi}) \right) \right], \quad \text{where} \quad \varphi_Y^{\epsilon} = \left\{ Y^{\phi} \, : \, d_2[Y^{\phi}, Y] \leq \epsilon \right\}$$

- take into account the uncertainty
- allow risk-averse and risk-seeking behaviors
- are elicitable

Robustifying the Dynamic Risk

Let $\breve{F}_Y$ be the quantile function of Y

- 2-Wasserstein distance: $d_2[Y, Z] = \left(\int_0^1 \left| \breve{F}_Y(u) - \breve{F}_Z(u) \right|^2 \mathrm{d}u \right)^{1/2}$
- Distortion risk measure: $\rho^\gamma(Y) = \mathbb{E}\Big[Y \, \gamma(F_Y(Y)) \Big] = \int_0^1 \gamma(u) \breve{F}_Y(u) \mathrm{d}u$

We work with 2-Wasserstein-robust distortion risk measures (with piecewise constant γ)

$$\rho^{\gamma,\epsilon}(Y) = \sup_{Y^\phi \in \varphi_Y^\epsilon} \mathbb{E}\Big[Y^\phi \, \gamma\Big(F_{Y^\phi}(Y^\phi) \Big) \Big], \quad \text{where} \quad \varphi_Y^\epsilon = \Big\{ Y^\phi \, : \, d_2[Y^\phi, Y] \leq \epsilon \Big\}$$

- take into account the uncertainty
- allow risk-averse and risk-seeking behaviors
- are elicitable

Robustifying the Dynamic Risk

Let $\breve{F}_Y$ be the quantile function of Y

- 2-Wasserstein distance: $d_2[Y, Z] = \left(\int_0^1 \left| \breve{F}_Y(u) - \breve{F}_Z(u) \right|^2 \mathrm{d}u \right)^{1/2}$
- Distortion risk measure: $\rho^\gamma(Y) = \mathbb{E}\Big[Y \, \gamma(F_Y(Y)) \Big] = \int_0^1 \gamma(u) \breve{F}_Y(u) \mathrm{d}u$

We work with 2-Wasserstein-robust distortion risk measures (with piecewise constant γ)

$$\rho^{\gamma,\epsilon}(Y) = \sup_{Y^\phi \in \varphi_Y^\epsilon} \mathbb{E}\Big[Y^\phi \, \gamma\Big(F_{Y^\phi}(Y^\phi) \Big) \Big], \quad \text{where} \quad \varphi_Y^\epsilon = \Big\{ Y^\phi \, : \, d_2[Y^\phi, Y] \le \epsilon \Big\}$$

- take into account the uncertainty
- allow risk-averse and risk-seeking behaviors
- are elicitable

Elicitability

We leverage the elicitability to efficiently estimate dynamic risk measures

Elicitable risk measure

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

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

Elicitability of (static) spectral risk measures, e.g. CVaR_α [Fissler and Ziegel, 2016]

- Proof of elicitability, and characterization of their scoring function

Extension to the class of (dynamic) spectral risk measures [Coache et al., 2022]

- May be approximated to any arbitrary accuracy with NNs
- *(SIAG/FME Conference Paper Prize Session, 11:45 AM – 1:15 PM)*

Elicitability

We leverage the elicitability to efficiently estimate dynamic risk measures

Elicitable risk measure

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

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

Elicitability of (static) spectral risk measures, e.g. CVaR_α [Fissler and Ziegel, 2016]

- Proof of elicitability, and characterization of their scoring function

Extension to the class of (dynamic) spectral risk measures [Coache et al., 2022]

- May be approximated to any arbitrary accuracy with NNs
- *(SIAG/FME Conference Paper Prize Session, 11:45 AM – 1:15 PM)*

Elicitable Mappings

Expectation is elicitable: $\mathbb{E}[Y] = \arg\min_{\mathfrak{a}\in\mathbb{R}} \mathbb{E}_{Y\sim F_Y}\left[(\mathfrak{a} - Y)^2\right]$

$(\mathrm{VaR}_\alpha, \mathrm{CVaR}_\alpha)$ is elicitable:

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

Distortion risk measures (with piecewise constant γ) are elicitable

Conditional maps are elicitable:

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

Any CDF is elicitable:

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

Elicitable Mappings

Expectation is elicitable: $\mathbb{E}[Y] = \arg\min_{\mathfrak{a}\in\mathbb{R}} \mathbb{E}_{Y\sim F_Y}\left[(\mathfrak{a} - Y)^2\right]$

$(\mathrm{VaR}_\alpha, \mathrm{CVaR}_\alpha)$ is elicitable:

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

Distortion risk measures (with piecewise constant γ) are elicitable

Conditional maps are elicitable:

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

Any CDF is elicitable:

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

Elicitable Mappings

Expectation is elicitable: $\mathbb{E}[Y] = \arg\min_{\mathfrak{a}\in\mathbb{R}} \mathbb{E}_{Y\sim F_Y}\left[(\mathfrak{a} - Y)^2\right]$

$(\mathrm{VaR}_\alpha,\ \mathrm{CVaR}_\alpha)$ is elicitable:

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

Distortion risk measures (with piecewise constant γ) are elicitable

Conditional maps are elicitable:

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

Any CDF is elicitable:

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

Elicitable Mappings

Expectation is elicitable: $\mathbb{E}[Y] = \arg\min_{\mathfrak{a}\in\mathbb{R}} \mathbb{E}_{Y\sim F_Y}\left[(\mathfrak{a} - Y)^2\right]$

$(\mathrm{VaR}_\alpha,\ \mathrm{CVaR}_\alpha)$ is elicitable:

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

Distortion risk measures (with piecewise constant γ) are elicitable

Conditional maps are elicitable:

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

Any CDF is elicitable:

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

Problem Setup

Problems of the form

$$\min_{\theta} \rho_{0,T}^{\gamma,\epsilon}\Big(\{c_t^\theta\}_t\Big) = \min_{\theta} \rho_0^{\gamma_0,\epsilon_0}\left(c_0^\theta + \rho_1^{\gamma_1,\epsilon_1}\left(c_1^\theta + \cdots + \rho_{T-1}^{\gamma_{T-1},\epsilon_{T-1}}\Big(c_{T-1}^\theta + \rho_T^{\gamma_T,\epsilon_T}(c_T^\theta)\Big)\cdots\right)\right)$$

where c_t^θ are $\mathcal{F}_{t+1}$-measurable random costs and $\rho_t^{\gamma_t,\epsilon_t}$ are robust distortion risk measures

DP equations for the *value function*, i.e. running risk-to-go, for $s \in \mathcal{S}$:

$$V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi_{Y_t^\theta}^{\epsilon_t}} \mathbb{E}\left[Y_t^\phi \, \gamma_t\Big(F_{Y_t^\phi|_{s_t=s}}(Y_t^\phi)\Big) \;\Big|\; s_t = s\right],$$

with $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

Problem Setup

Problems of the form

$$\min_{\theta} \rho_{0,T}^{\gamma,\epsilon}\left(\{c_t^{\theta}\}_t\right) = \min_{\theta} \rho_0^{\gamma_0,\epsilon_0}\left(c_0^{\theta} + \rho_1^{\gamma_1,\epsilon_1}\left(c_1^{\theta} + \cdots + \rho_{T-1}^{\gamma_{T-1},\epsilon_{T-1}}\left(c_{T-1}^{\theta} + \rho_T^{\gamma_T,\epsilon_T}\left(c_T^{\theta}\right)\right)\cdots\right)\right)$$

where c_t^{θ} are $\mathcal{F}_{t+1}$-measurable random costs and $\rho_t^{\gamma_t,\epsilon_t}$ are robust distortion risk measures

DP equations for the *value function*, i.e. running risk-to-go, for $s \in \mathcal{S}$:

$$V_t(s;\theta) = \sup_{Y_t^{\phi} \in \varphi_{Y_t^{\theta}}^{\epsilon_t}} \mathbb{E}\left[Y_t^{\phi}\, \gamma_t\left(F_{Y_t^{\phi}|_{s_t=s}}(Y_t^{\phi})\right) \,\middle|\, s_t = s\right],$$

with $Y_t^{\theta} := c_t^{\theta} + V_{t+1}(s_{t+1}^{\theta};\theta)$

Policy Gradient

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

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

- requires maximizing the worst case risk over ϕ
- $\nabla_\theta V(\cdot;\theta)$ depends on $V(\cdot;\theta)$ itself due to the DP equations

Adversary-actor-critic style algorithm composed of 3 interleaved procedures:

- *Adversary* estimates the distribution of the costs-to-go
- *Critic* calculates the value function of the given policy
- *Actor* updates the current policy
- We parametrize the components we optimize by NNs

Policy Gradient

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

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

- requires maximizing the worst case risk over ϕ
- $\nabla_\theta V(\cdot;\theta)$ depends on $V(\cdot;\theta)$ itself due to the DP equations

Adversary-actor-critic style algorithm composed of 3 interleaved procedures:

- *Adversary* estimates the distribution of the costs-to-go
- *Critic* calculates the value function of the given policy
- *Actor* updates the current policy
- We parametrize the components we optimize by NNs

Step 1: Distribution of Y_t^θ

We aim to estimate the distribution $F_{Y_t^\theta}(\cdot|s_t)$, where $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

- Estimation of $F_{Y_t^\theta}$ with the elicitable framework:

$$F_{Y_t^\theta}(\cdot|s_t) = \arg\min_{F\in\mathbb{F}} \mathbb{E}_{\substack{a_t^\theta\sim\pi^\theta \\ s_{t+1}^\theta\sim\mathbb{P}}}\left[\int_{\mathbb{R}}\left(F(y|s_t) - \mathbb{1}_{y\geq Y_t^\theta}\right)^2 \mathrm{d}y\right]$$

- Additional monotonicity penalty to discourage increasing NN functions
- Mini-batch of realizations Y_t^θ induced by π^θ

This requires an estimation of $V_t(s;\theta)$...

Step 1: Distribution of Y_t^θ

We aim to estimate the distribution $F_{Y_t^\theta}(\cdot|s_t)$, where $Y_t^\theta := c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta)$

- Estimation of $F_{Y_t^\theta}$ with the elicitable framework:

$$F_{Y_t^\theta}(\cdot|s_t) = \operatorname*{arg\,min}_{F\in\mathbb{F}} \mathbb{E}_{\substack{a_t^\theta\sim\pi^\theta \\ s_{t+1}^\theta\sim\mathbb{P}}}\left[\int_{\mathbb{R}}\left(F(y|s_t) - \mathbb{1}_{y\geq Y_t^\theta}\right)^2 \mathrm{d}y\right]$$

- Additional monotonicity penalty to discourage increasing NN functions
- Mini-batch of realizations Y_t^θ induced by π^θ

This requires an estimation of $V_t(s;\theta)$...

Step 2: Value function $V_t(s;\theta)$

We aim to estimate $V_t(s;\theta) = \sup\limits_{Y_t^\phi \in \varphi^{\epsilon_t}_{Y_t^\theta}} \rho_t\left(Y_t^\phi \,\middle|\, s_t = s\right) = \sup\limits_{\breve{F}_\phi \in \varphi^{\epsilon_t}_{\breve{F}_{Y_t^\theta}(\cdot|s)}} \int_0^1 \gamma_t(u) \breve{F}_\phi(u|s) \mathrm{d}u$

- Reformulation of the problem with quantile functions

Proposition

The optimal quantile function in $V_t(s;\theta)$ is given by

$$\breve{F}^*_\phi(\cdot|s) = \left(\breve{F}_{Y_t^\theta}(\cdot|s) + \frac{\gamma_t(\cdot)}{2\lambda^*}\right)^\uparrow, \quad \text{with } \lambda^* > 0 \text{ such that } \int_0^1 \left|\breve{F}^*_\phi(u|s) - \breve{F}_{Y_t^\theta}(u|s)\right|^2 \mathrm{d}u = \epsilon_t^2.$$

- Estimation of $V_t(s;\theta)$ with the elicitable framework:

$$\min_{h:\mathcal{S}\to\mathbb{R}} \mathbb{E}_{\substack{a_t^\theta \sim \pi^\theta \\ s_{t+1}^\theta \sim \mathbb{P}}}\left[S\left(h(s_t);\, Y_t^\phi\right)\right], \quad Y_t^\phi \sim \breve{F}^*_\phi(\cdot|s_t)$$

This requires an estimation of $F_{Y_t^\theta}(\cdot|s_t)$...

Step 2: Value function $V_t(s;\theta)$

We aim to estimate $V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi^{\epsilon_t}_{Y_t^\theta}} \rho_t\left(Y_t^\phi \,\middle|\, s_t = s\right) = \sup_{\breve{F}_\phi \in \varphi^{\epsilon_t}_{\breve{F}_{Y_t^\theta}(\cdot|s)}} \int_0^1 \gamma_t(u) \breve{F}_\phi(u|s) \mathrm{d}u$

- Reformulation of the problem with quantile functions

Proposition

The optimal quantile function in $V_t(s;\theta)$ is given by

$$\breve{F}_\phi^*(\cdot|s) = \left(\breve{F}_{Y_t^\theta}(\cdot|s) + \frac{\gamma_t(\cdot)}{2\lambda^*}\right)^{\uparrow}, \quad \text{with } \lambda^* > 0 \text{ such that } \int_0^1 \left|\breve{F}_\phi^*(u|s) - \breve{F}_{Y_t^\theta}(u|s)\right|^2 \mathrm{d}u = \epsilon_t^2.$$

- Estimation of $V_t(s;\theta)$ with the elicitable framework:

$$\min_{h:\mathcal{S}\to\mathbb{R}} \mathbb{E}_{\substack{a_t^\theta \sim \pi^\theta \\ s_{t+1}^\theta \sim \mathbb{P}}} \left[S\left(h(s_t);\, Y_t^\phi\right)\right], \quad Y_t^\phi \sim \breve{F}_\phi^*(\cdot|s_t)$$

This requires an estimation of $F_{Y_t^\theta}(\cdot|s_t)$...

Step 2: Value function $V_t(s;\theta)$

We aim to estimate $V_t(s;\theta) = \sup_{Y_t^\phi \in \varphi^{\epsilon_t}_{Y_t^\theta}} \rho_t\left(Y_t^\phi \;\middle|\; s_t = s\right) = \sup_{\breve{F}_\phi \in \varphi^{\epsilon_t}_{\breve{F}_{Y_t^\theta}(\cdot|s)}} \int_0^1 \gamma_t(u)\breve{F}_\phi(u|s)\mathrm{d}u$

- Reformulation of the problem with quantile functions

Proposition

The optimal quantile function in $V_t(s;\theta)$ is given by

$$\breve{F}^*_\phi(\cdot|s) = \left(\breve{F}_{Y_t^\theta}(\cdot|s) + \frac{\gamma_t(\cdot)}{2\lambda^*}\right)^{\uparrow}, \quad \text{with } \lambda^* > 0 \text{ such that } \int_0^1 \left|\breve{F}^*_\phi(u|s) - \breve{F}_{Y_t^\theta}(u|s)\right|^2 \mathrm{d}u = \epsilon_t^2.$$

- Estimation of $V_t(s;\theta)$ with the elicitable framework:

$$\min_{h:\mathcal{S}\to\mathbb{R}} \mathbb{E}_{\substack{a_t^\theta \sim \pi^\theta \\ s_{t+1}^\theta \sim \mathbb{P}}}\left[S\left(h(s_t);\; Y_t^\phi\right)\right], \quad Y_t^\phi \sim \breve{F}^*_\phi(\cdot|s_t)$$

This requires an estimation of $F_{Y_t^\theta}(\cdot|s_t)$...

Step 3: Gradient $\nabla_\theta V_t(s;\theta)$

We aim to update the policy π^θ via a policy gradient method

- Optimization problem is convex over the space of quantile functions

Proposition

The gradient of $V_t(s;\theta)$ wrt policy parameters θ is given by

$$\nabla_\theta V_t(s;\theta) = -2\,\mathbb{E}\left[\lambda^*\left(Y_t^{\phi,c} - Y_t^\theta\right)\frac{\nabla_\theta F_{Y_t^\theta}(x|s)}{f_{Y_t^\theta}(x|s)}\bigg|_{x=Y_t^\theta}\right],$$

where $(Y_t^{\phi,c}, Y_t^\theta)$ is comonotonic with marginals $\breve{F}_\phi^*(\cdot|s)$ and $\breve{F}_{Y_t^\theta}(\cdot|s)$.

- Mini-batches of observed costs Y_t^θ (from π^θ) and distorted costs $Y_t^{\phi,c}$ (from $\breve{F}_\phi^*$)
- Optimization of the policy parameters θ wrt $\nabla_\theta V_t(s;\theta)$

Step 3: Gradient $\nabla_\theta V_t(s;\theta)$

We aim to update the policy π^θ via a policy gradient method

- Optimization problem is convex over the space of quantile functions

Proposition

The gradient of $V_t(s;\theta)$ wrt policy parameters θ is given by

$$\nabla_\theta V_t(s;\theta) = -2\,\mathbb{E}\left[\lambda^*\left(Y_t^{\phi,c} - Y_t^\theta\right)\frac{\nabla_\theta F_{Y_t^\theta}(x|s)}{f_{Y_t^\theta}(x|s)}\bigg|_{x=Y_t^\theta}\right],$$

where $(Y_t^{\phi,c}, Y_t^\theta)$ is comonotonic with marginals $\breve{F}_\phi^*(\cdot|s)$ and $\breve{F}_{Y_t^\theta}(\cdot|s)$.

- Mini-batches of observed costs Y_t^θ (from π^θ) and distorted costs $Y_t^{\phi,c}$ (from $\breve{F}_\phi^*$)
- Optimization of the policy parameters θ wrt $\nabla_\theta V_t(s;\theta)$

Step 3: Gradient $\nabla_\theta V_t(s;\theta)$

We aim to update the policy π^θ via a policy gradient method

- Optimization problem is convex over the space of quantile functions

Proposition

The gradient of $V_t(s;\theta)$ wrt policy parameters θ is given by

$$\nabla_\theta V_t(s;\theta) = -2\,\mathbb{E}\left[\lambda^*\left(Y_t^{\phi,c} - Y_t^\theta\right)\frac{\nabla_\theta F_{Y_t^\theta}(x|s)}{f_{Y_t^\theta}(x|s)}\bigg|_{x=Y_t^\theta}\right],$$

where $(Y_t^{\phi,c}, Y_t^\theta)$ is comonotonic with marginals $\breve{F}_\phi^*(\cdot|s)$ and $\breve{F}_{Y_t^\theta}(\cdot|s)$.

- Mini-batches of observed costs Y_t^θ (from π^θ) and distorted costs $Y_t^{\phi,c}$ (from $\breve{F}_\phi^*$)
- Optimization of the policy parameters θ wrt $\nabla_\theta V_t(s;\theta)$

Full Algorithm

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;

            Update ϑ via Adam optimization;

        for $k^\phi = 1, \ldots, K^\phi$ **do**

            Critic: minimization of the expected consistent score;

            Update ϕ via Adam optimization;

    for $k^\theta = 1, \ldots, K^\theta$ **do**

        Actor: policy gradient;

        Update θ via Adam optimization;

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

Full Algorithm

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

for *each iteration* $k = 1, \dots, K$ **do**
- **while** *convergence is not achieved* **do**
  - **for** $k^\vartheta = 1, \dots, K^\vartheta$ **do**
    - **Adversary:** minimization of the expected scoring rule for F;
    - Update ϑ via Adam optimization;
  - **for** $k^\phi = 1, \dots, K^\phi$ **do**
    - **Critic:** minimization of the expected consistent score;
    - Update ϕ via Adam optimization;
- **for** $k^\theta = 1, \dots, K^\theta$ **do**
  - **Actor:** policy gradient;
  - Update θ via Adam optimization;

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

Full Algorithm

Input: ANNs π^θ, V^ϕ, F^ϑ, 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;
            Update ϑ via Adam optimization;
        for $k^\phi = 1, \ldots, K^\phi$ **do**
            Critic: minimization of the expected consistent score;
            Update ϕ via Adam optimization;
    for $k^\theta = 1, \ldots, K^\theta$ **do**
        Actor: policy gradient;
        Update θ via Adam optimization;

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

Full Algorithm

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;
    - Update ϑ via Adam optimization;
  - **for** $k^\phi = 1, \ldots, K^\phi$ **do**
    - **Critic:** minimization of the expected consistent score;
    - Update ϕ via Adam optimization;
- **for** $k^\theta = 1, \ldots, K^\theta$ **do**
  - **Actor:** policy gradient;
  - Update θ via Adam optimization;

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

Contributions & Future Directions

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

- Efficient estimation method utilizing *elicitable mappings*
- *Robustification* to protect against model uncertainty

Future directions

- Alternative uncertainty sets, e.g. KL divergence
- Multi-agent settings, e.g. MFGs

Thank you!

More info and slides: `anthonycoache.ca`

Registration and travel support for this presentation was provided by the Society for Industrial and Applied Mathematics.

References I

Bielecki, T. R., Cialenco, I., and Ruszczyński, A. (2022). Risk filtering and risk-averse control of Markovian systems subject to model uncertainty. *arXiv preprint arXiv:2206.09235*.

Cheng, Z. and Jaimungal, S. (2022). Markov decision processes with Kusuoka-type conditional risk mappings. *arXiv preprint arXiv:2203.09612*.

Coache, A. and Jaimungal, S. (2021). Reinforcement learning with dynamic convex risk measures. *arXiv preprint arXiv:2112.13414*.

Coache, A., Jaimungal, S., and Cartea, Á. (2022). Conditionally elicitable dynamic risk measures for deep reinforcement learning. *arXiv preprint arXiv:2206.14666*.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707.

Jaimungal, S., Pesenti, S. M., Wang, Y. S., and Tatsat, H. (2022). Robust risk-aware reinforcement learning. *SIAM Journal on Financial Mathematics*, 13(1):213–226.

Marzban, S., Delage, E., and Li, J. Y. (2021). Deep reinforcement learning for equal risk pricing and hedging under dynamic expectile risk measures. *arXiv preprint arXiv:2109.04001*.

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

Smirnova, E., Dohmatob, E., and Mary, J. (2019). Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*.