Reinforcement Learning with Dynamic Convex Risk Measures

Anthony Coache Sebastian Jaimungal

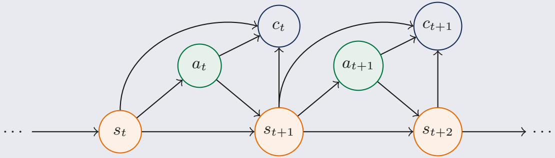
anthonycoache.ca
sebastian.statistics.utoronto.ca

Department of Statistical Sciences
University of Toronto

Reinforcement Learning (RL)

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

- $\mathcal{S}$ – State space
- $\mathcal{A}$ – Action space
- $\pi^\theta(a_t|s_t)$ – Randomized policy characterized by θ
- $\mathbb{P}(s_0), \mathbb{P}(s_{t+1}|s_t, a_t)$ – Transition probability distribution
- $c(s, a, s') \in \mathcal{C}$ – Cost function



Reinforcement Learning (RL)

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

- $\mathcal{S}$ – State space
- $\mathcal{A}$ – Action space
- $\pi^\theta(a_t|s_t)$ – Randomized policy characterized by θ
- $\mathbb{P}(s_0), \mathbb{P}(s_{t+1}|s_t, a_t)$ – Transition probability distribution
- $c(s, a, s') \in \mathcal{C}$ – Cost function

Standard RL: *risk-neutral objective* function of a cost

$$\min_\theta \mathbb{E}[Z].$$

Risk-aware RL: *risk measure* ρ of the cost Z

$$\min_\theta \rho(Z) \quad \text{or} \quad \min_\theta \mathbb{E}[Z] \text{ subj. to } \rho(Z) \le Z^*.$$

Reinforcement Learning (RL)

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

- $\mathcal{S}$ – State space
- $\mathcal{A}$ – Action space
- $\pi^\theta(a_t|s_t)$ – Randomized policy characterized by θ
- $\mathbb{P}(s_0), \mathbb{P}(s_{t+1}|s_t, a_t)$ – Transition probability distribution
- $c(s, a, s') \in \mathcal{C}$ – Cost function

Standard RL: *risk-neutral objective* function of a cost

$$\min_\theta \mathbb{E}[Z].$$

Risk-aware RL: *risk measure ρ* of the cost Z

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

Motivations

Risk-aware RL: applying risk measures *recursively* [e.g. Rus10; CZ14], or applying a *static* risk measure [e.g. NBP19; BG20]

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

[TCGM15] provide policy search algorithms in both the static and dynamic framework, but some potential shortcomings remain:

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

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

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

Motivations

Risk-aware RL: applying risk measures *recursively* [e.g. Rus10; CZ14], or applying a *static* risk measure [e.g. NBP19; BG20]

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

[TCGM15] provide policy search algorithms in both the static and dynamic framework, but some potential shortcomings remain:

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

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

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

Motivations

Risk-aware RL: applying risk measures *recursively* [e.g. Rus10; CZ14], or applying a *static* risk measure [e.g. NBP19; BG20]

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

[TCGM15] provide policy search algorithms in both the static and dynamic framework, but some potential shortcomings remain:

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

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

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

Risk Measures

$\rho : \mathcal{Z} \to \mathbb{R}$ is

- *monotone:* $Z_1 \leq Z_2$ implies $\rho(Z_1) \leq \rho(Z_2)$
- *translation invariant:* $\rho(Z + m) = \rho(Z) + m, \ \forall m \in \mathbb{R}$
- *positive homogeneous:* $\rho(\beta Z) = \beta\rho(Z), \ \forall \beta > 0$
- *subadditive:* $\rho(Z_1 + Z_2) \leq \rho(Z_1) + \rho(Z_2)$
- *convex:* $\rho(\lambda Z_1 + (1 - \lambda)Z_2) \leq \lambda\rho(Z_1) + (1 - \lambda)\rho(Z_2)$

Coherent ρ [ADEH99]

Monotone, translation invariant, positive homogeneous and subadditive

Convex ρ [FS02]

Monotone, translation invariant and convex

Risk Measures

$\rho : \mathcal{Z} \to \mathbb{R}$ is

- *monotone:* $Z_1 \leq Z_2$ implies $\rho(Z_1) \leq \rho(Z_2)$
- *translation invariant:* $\rho(Z + m) = \rho(Z) + m, \ \forall m \in \mathbb{R}$
- *positive homogeneous:* $\rho(\beta Z) = \beta\rho(Z), \ \forall \beta > 0$
- *subadditive:* $\rho(Z_1 + Z_2) \leq \rho(Z_1) + \rho(Z_2)$
- *convex:* $\rho(\lambda Z_1 + (1 - \lambda)Z_2) \leq \lambda\rho(Z_1) + (1 - \lambda)\rho(Z_2)$

Coherent ρ [ADEH99]

Monotone, translation invariant, positive homogeneous and subadditive

Convex ρ [FS02]

Monotone, translation invariant and convex

Risk Measures

$\rho : \mathcal{Z} \to \mathbb{R}$ is

- *monotone:* $Z_1 \leq Z_2$ implies $\rho(Z_1) \leq \rho(Z_2)$
- *translation invariant:* $\rho(Z + m) = \rho(Z) + m, \ \forall m \in \mathbb{R}$
- *positive homogeneous:* $\rho(\beta Z) = \beta \rho(Z), \ \forall \beta > 0$
- *subadditive:* $\rho(Z_1 + Z_2) \leq \rho(Z_1) + \rho(Z_2)$
- *convex:* $\rho(\lambda Z_1 + (1 - \lambda) Z_2) \leq \lambda \rho(Z_1) + (1 - \lambda) \rho(Z_2)$

Coherent ρ [ADEH99]

Monotone, translation invariant, positive homogeneous and subadditive

Convex ρ [FS02]

Monotone, translation invariant and convex

Dual Representation

Representation Theorem [SDR14]

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

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

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

Moreover, ρ coherent iff $\rho(Z) = \sup_{\xi \in \mathcal{U}(P)} \left\{\mathbb{E}^{\xi}\left[Z\right]\right\}$

We assume the *risk envelope* $\mathcal{U}$ is of the form [TCGM15]

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

Dual Representation

Representation Theorem [SDR14]

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

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

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

Moreover, ρ coherent iff $\rho(Z) = \sup_{\xi \in \mathcal{U}(P)} \left\{\mathbb{E}^{\xi}[Z]\right\}$

We assume the *risk envelope $\mathcal{U}$* is of the form [TCGM15]

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

Dynamic Risk Measures

Consider

- $(\Omega, \mathcal{F}, P)$ – Probability space
- $\mathcal{F}_0 \subseteq \ldots \subseteq \mathcal{F}_T$ – Filtration
- $\mathcal{Z}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ – p-integrable random variables
- $\mathcal{Z}_{t,T} = \mathcal{Z}_t \times \cdots \mathcal{Z}_T$

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

Sequence of $\rho_{t,T} : \mathcal{Z}_{t,T} \to \mathcal{Z}_t$ where $\rho_{t,T}(Z) \leq \rho_{t,T}(W), \ \forall Z \leq W$

Time-consistency [Rus10]

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

$$\rho_{t_2,T}(Z_{t_2}, \ldots, Z_T) \leq \rho_{t_2,T}(W_{t_2}, \ldots, W_T) \ \text{ and } \ Z_k = W_k, \ \forall k = t_1, \ldots, t_2$$

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

Dynamic Risk Measures

Consider

- $(\Omega, \mathcal{F}, P)$ – Probability space
- $\mathcal{F}_0 \subseteq \ldots \subseteq \mathcal{F}_T$ – Filtration
- $\mathcal{Z}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ – p-integrable random variables
- $\mathcal{Z}_{t,T} = \mathcal{Z}_t \times \cdots \mathcal{Z}_T$

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

Sequence of $\rho_{t,T} : \mathcal{Z}_{t,T} \to \mathcal{Z}_t$ where $\rho_{t,T}(Z) \leq \rho_{t,T}(W), \ \forall Z \leq W$

Time-consistency [Rus10]

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

$$\rho_{t_2,T}(Z_{t_2}, \ldots, Z_T) \leq \rho_{t_2,T}(W_{t_2}, \ldots, W_T) \ \text{ and } \ Z_k = W_k, \ \forall k = t_1, \ldots, t_2$$

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

Dynamic Risk Measures

One-step conditional risk measure ρ_t

Risk measure $\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$ such that $\rho_t(Z_{t+1}) = \rho_{t,t+1}(0, Z_{t+1})$.

Suppose a time-consistent $\{\rho_{t,T}\}_t$ satisfies

- $\rho_{t,T}(Z_t, Z_{t+1}, \ldots, Z_T) = Z_t + \rho_{t,T}(0, Z_{t+1}, \ldots, Z_T)$
- $\rho_{t,T}(0, \ldots, 0) = 0$
- $\rho_{t_1,t_2}(\mathbf{1}_A Z) = \mathbf{1}_A \rho_{t_1,t_2}(Z)$, $\forall A \in \mathcal{F}_{t_1}$

Then [Rus10] we have

$$\rho_{t,T}(Z_t, \ldots, Z_T) = Z_t + \rho_t\left(Z_{t+1} + \rho_{t+1}\left(Z_{t+2} + \cdots + \rho_{T-1}\left(Z_T\right)\cdots\right)\right)$$

Additional assumed properties for ρ_t:

- Axioms of convex risk measures
- Markovian, i.e. not allowed to depend on the whole past

Dynamic Risk Measures

One-step conditional risk measure ρ_t

Risk measure $\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$ such that $\rho_t(Z_{t+1}) = \rho_{t,t+1}(0, Z_{t+1})$.

Suppose a time-consistent $\{\rho_{t,T}\}_t$ satisfies

- $\rho_{t,T}(Z_t, Z_{t+1}, \ldots, Z_T) = Z_t + \rho_{t,T}(0, Z_{t+1}, \ldots, Z_T)$
- $\rho_{t,T}(0, \ldots, 0) = 0$
- $\rho_{t_1,t_2}(\mathbf{1}_A Z) = \mathbf{1}_A \rho_{t_1,t_2}(Z)$, $\forall A \in \mathcal{F}_{t_1}$

Then [Rus10] we have

$$\rho_{t,T}(Z_t, \ldots, Z_T) = Z_t + \rho_t\left(Z_{t+1} + \rho_{t+1}\left(Z_{t+2} + \cdots + \rho_{T-1}\left(Z_T\right) \cdots\right)\right)$$

Additional assumed properties for ρ_t:

- Axioms of convex risk measures
- Markovian, i.e. not allowed to depend on the whole past

Problem Setup

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

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

Using the dual representation and recursive equations, we have

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

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

for $s \in \mathcal{S}$ and $t = T-2, \ldots, 1$, where

- $c_t^\theta = c(s_t, a_t^\theta, s_{t+1}^\theta)$ – Cost of transitions at t induced by π^θ
- $\mathbb{P}^\theta(a, s'|s_t = s) = \mathbb{P}(s'|s,a)\pi^\theta(a|s_t = s)$ – Transition probability induced by π^θ

Problem Setup

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

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

Using the dual representation and recursive equations, we have

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

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

for $s \in \mathcal{S}$ and $t = T-2, \ldots, 1$, where

- $c_t^\theta = c(s_t, a_t^\theta, s_{t+1}^\theta)$ – Cost of transitions at t induced by π^θ
- $\mathbb{P}^\theta(a, s'|s_t = s) = \mathbb{P}(s'|s,a)\pi^\theta(a|s_t = s)$ – Transition probability induced by π^θ

Problem Setup

- We wish to optimize the value function over policies θ
- We parametrize both policy and value function by ANNs, denoted θ and ϕ
- The Lagrangian of the *maximization problem* is

$$L^{\theta}(\xi, \lambda) = \sum_{(a,s')} \xi(a,s')\mathbb{P}^{\theta}(a,s'|s_t = s)\left(c_t(s,a,s') + V_{t+1}(s';\theta)\right) - \rho_t^*(\xi)$$
$$- \lambda\left(\sum_{(a,s')} \xi(a,s')\mathbb{P}^{\theta}(a,s'|s_t = s) - 1\right)$$
$$- \underbrace{\sum_{e\in\mathcal{E}}\left(\lambda^{\mathcal{E}}(e)g_e(\xi,\mathbb{P}^{\theta})\right)}_{\text{equality constraints}} - \underbrace{\sum_{i\in\mathcal{I}}\left(\lambda^{\mathcal{I}}(i)f_i(\xi,\mathbb{P}^{\theta})\right)}_{\text{inequality constraints}}.$$

Problem Setup

- We wish to optimize the value function over policies θ
- We parametrize both policy and value function by ANNs, denoted θ and ϕ
- The Lagrangian of the *maximization problem* is

$$L^\theta(\xi,\lambda) = \sum_{(a,s')} \xi(a,s')\mathbb{P}^\theta(a,s'|s_t=s)\left(c_t(s,a,s') + V_{t+1}(s';\theta)\right) - \rho_t^*(\xi)$$
$$- \lambda\left(\sum_{(a,s')} \xi(a,s')\mathbb{P}^\theta(a,s'|s_t=s) - 1\right)$$
$$- \underbrace{\sum_{e\in\mathcal{E}} \left(\lambda^{\mathcal{E}}(e) g_e(\xi,\mathbb{P}^\theta)\right)}_{\text{equality constraints}} - \underbrace{\sum_{i\in\mathcal{I}} \left(\lambda^{\mathcal{I}}(i) f_i(\xi,\mathbb{P}^\theta)\right)}_{\text{inequality constraints}} . \; .$$

Problem Setup

- We wish to optimize the value function over policies θ
- We parametrize both policy and value function by ANNs, denoted θ and ϕ
- The Lagrangian of the *maximization problem* is

$$\begin{aligned} L^{\theta}(\xi,\lambda) = & \sum_{(a,s')} \xi(a,s')\mathbb{P}^{\theta}(a,s'|s_t=s)\left(c_t(s,a,s') + V_{t+1}(s';\theta)\right) - \rho_t^*(\xi) \\ & - \lambda\left(\sum_{(a,s')} \xi(a,s')\mathbb{P}^{\theta}(a,s'|s_t=s) - 1\right) \\ & \underbrace{-\sum_{e\in\mathcal{E}}\left(\lambda^{\mathcal{E}}(e)g_e(\xi,\mathbb{P}^{\theta})\right)}_{\text{equality constraints}} \underbrace{-\sum_{i\in\mathcal{I}}\left(\lambda^{\mathcal{I}}(i)f_i(\xi,\mathbb{P}^{\theta})\right).}_{\text{inequality constraints}} \end{aligned}$$

Problem Setup

The Envelope Theorem [MS02] says

$$\nabla_\theta \left(\max_{\xi \in \mathcal{U}(\mathbb{P}^\theta(\cdot,\cdot|s_t = s))} \left\{ \mathbb{E}^{\xi}_{t,s} \left[c^\theta_t + V_{t+1}(s^\theta_{t+1};\theta) \right] - \rho^*_t(\xi) \right\} \right) = \nabla_\theta L^\theta(\xi,\lambda) \Big|_{\xi^*,\lambda^*}$$

Gradient of V [CJ21]

$$\nabla_\theta V_t(s;\theta) = \mathbb{E}^{\xi^*}_t \Bigg[\overbrace{\left(c^\theta_t + V_{t+1}(s^\theta_{t+1};\theta) - \lambda^*\right) \nabla_\theta \log \pi^\theta(a^\theta_t | s_t = s)}^{\text{transition}} + \overbrace{\nabla_\theta V_{t+1}(s^\theta_{t+1};\theta)}^{\text{risk-to-go } V_{t+1}} \Bigg]$$
$$- \underbrace{\nabla_\theta \rho^*_t(\xi^*)}_{\text{convex penalty}} - \underbrace{\sum_{e\in\mathcal{E}} \left(\lambda^{*,\mathcal{E}}(e) \nabla_\theta g_e(\xi^*,\mathbb{P}^\theta)\right)}_{\text{equality constraints}} - \underbrace{\sum_{i\in\mathcal{I}} \left(\lambda^{*,\mathcal{I}}(i) \nabla_\theta f_i(\xi^*,\mathbb{P}^\theta)\right)}_{\text{inequality constraints}}$$

Problem Setup

The Envelope Theorem [MS02] says

$$\nabla_\theta \left(\max_{\xi \in \mathcal{U}(\mathbb{P}^\theta(\cdot,\cdot|s_t=s))} \left\{ \mathbb{E}^{\xi}_{t,s} \left[c^\theta_t + V_{t+1}(s^\theta_{t+1};\theta) \right] - \rho^*_t(\xi) \right\} \right) = \nabla_\theta L^\theta(\xi,\lambda) \Big|_{\xi^*,\lambda^*}$$

Gradient of V [CJ21]

$$\nabla_\theta V_t(s;\theta) = \mathbb{E}^{\xi^*}_t \Bigg[\overbrace{\left(c^\theta_t + V_{t+1}(s^\theta_{t+1};\theta) - \lambda^* \right) \nabla_\theta \log \pi^\theta(a^\theta_t | s_t = s)}^{\text{transition}} + \overbrace{\nabla_\theta V_{t+1}(s^\theta_{t+1};\theta)}^{\text{risk-to-go } V_{t+1}} \Bigg]$$
$$- \underbrace{\nabla_\theta \rho^*_t(\xi^*)}_{\text{convex penalty}} - \underbrace{\sum_{e\in\mathcal{E}} \left(\lambda^{*,\mathcal{E}}(e) \nabla_\theta g_e(\xi^*,\mathbb{P}^\theta) \right)}_{\text{equality constraints}} - \underbrace{\sum_{i\in\mathcal{I}} \left(\lambda^{*,\mathcal{I}}(i) \nabla_\theta f_i(\xi^*,\mathbb{P}^\theta) \right)}_{\text{inequality constraints}}$$

Algorithm

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

- *Critic* calculates the value function given a policy
- *Actor* updates the policy given a value function

Algorithm 1: Main algorithm

Input: Environment, risk measure, π^θ, V^ϕ

for *each epoch* $\kappa = 1, \ldots, K$ **do**

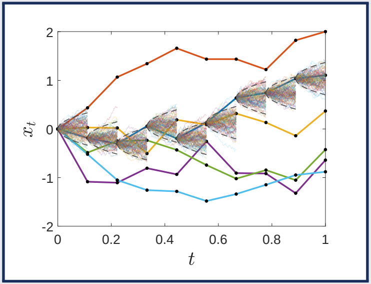
    Generate (outer) trajectories ;

    Generate (inner) transitions ;

    Estimate the value function (*critic*) ;

    Update the policy (*actor*) ;

Output: Optimal policy $\pi^\theta \approx \pi^*$



- Function approximation for estimating the policy and value function

Estimation of the Value Function

Recall that for $s \in \mathcal{S}$ and $t = 1, \ldots, T-2$,

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

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

Estimate the risk measure using (inner) transitions

$$(s_t, a^{(m)}_t, s^{(m)}_{t+1}, c^{(m)}_t), \; m = 1, \ldots, M$$

- ANN $V^\phi : s_t \mapsto \mathbb{R}$
- Expected square loss between predicted and target values
- Mini-batches of states from the (outer) trajectories
- Adam optimization step to update ϕ

Estimation of the Value Function

Recall that for $s \in \mathcal{S}$ and $t = 1, \dots, T-2$,

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

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

Estimate the risk measure using (inner) transitions

$$(s_t, a_t^{(m)}, s_{t+1}^{(m)}, c_t^{(m)}), \;\; m = 1, \dots, M$$

- ANN $V^{\phi} : s_t \mapsto \mathbb{R}$
- Expected square loss between predicted and target values
- Mini-batches of states from the (outer) trajectories
- Adam optimization step to update ϕ

Update of the Policy

Recall that for $s \in \mathcal{S}$ and $t = 1, \ldots, T-1$,

$$\nabla_\theta V_t(s;\theta) = \mathbb{E}_t^{\xi^*}\Bigg[\overbrace{\Big(c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta) - \lambda^*\Big)\nabla_\theta \log \pi^\theta(a_t^\theta | s_t = s)}^{\text{transition}} + \overbrace{\nabla_\theta V_{t+1}(s_{t+1}^\theta;\theta)}^{\text{risk-to-go } V_{t+1}} \Bigg]$$
$$- \underbrace{\nabla_\theta \rho_t^*(\xi^*)}_{\text{convex penalty}} - \underbrace{\sum_{e\in\mathcal{E}}\Big(\lambda^{*,\mathcal{E}}(e)\nabla_\theta g_e(\xi^*,\mathbb{P}^\theta)\Big)}_{\text{equality constraints}} - \underbrace{\sum_{i\in\mathcal{I}}\Big(\lambda^{*,\mathcal{I}}(i)\nabla_\theta f_i(\xi^*,\mathbb{P}^\theta)\Big)}_{\text{inequality constraints}}$$

V: obtained using the critic V^ϕ

$\pi^\theta(a_t^\theta | s_t = s)$: reparametrization trick

- ANN $\pi^\theta : s_t \mapsto \mathcal{P}(\mathcal{A})$
- Computation of $\nabla_\theta V_t$
- Mini-batches of states from the (outer) trajectories
- Stochastic Gradient Descent optimization step to update θ

Update of the Policy

Recall that for $s \in \mathcal{S}$ and $t = 1, \ldots, T-1$,

$$\nabla_\theta V_t(s;\theta) = \mathbb{E}_t^{\xi^*}\Bigg[\overbrace{\left(c_t^\theta + V_{t+1}(s_{t+1}^\theta;\theta) - \lambda^*\right)\nabla_\theta \log \pi^\theta(a_t^\theta | s_t = s)}^{\text{transition}} + \overbrace{\nabla_\theta V_{t+1}(s_{t+1}^\theta;\theta)}^{\text{risk-to-go } V_{t+1}}\Bigg]$$
$$- \underbrace{\nabla_\theta \rho_t^*(\xi^*)}_{\text{convex penalty}} - \underbrace{\sum_{e\in\mathcal{E}}\left(\lambda^{*,\mathcal{E}}(e)\nabla_\theta g_e(\xi^*,\mathbb{P}^\theta)\right)}_{\text{equality constraints}} - \underbrace{\sum_{i\in\mathcal{I}}\left(\lambda^{*,\mathcal{I}}(i)\nabla_\theta f_i(\xi^*,\mathbb{P}^\theta)\right)}_{\text{inequality constraints}}$$

V: obtained using the critic V^ϕ

$\pi^\theta(a_t^\theta | s_t = s)$: reparametrization trick

- ANN $\pi^\theta : s_t \mapsto \mathcal{P}(\mathcal{A})$
- Computation of $\nabla_\theta V_t$
- Mini-batches of states from the (outer) trajectories
- Stochastic Gradient Descent optimization step to update θ

Risk Measures

Different risk measures

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

where

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

Special cases

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

Statistical Arbitrage Example

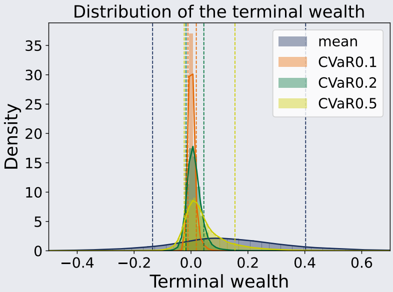
Consider a market with a single asset. An agent:

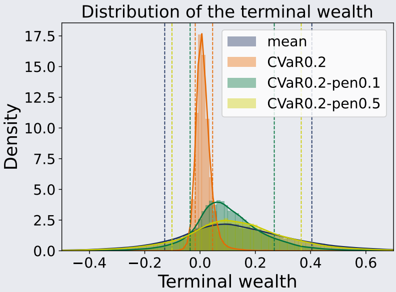
- invests during T periods, denoted $t = 0, \ldots, T-1$
- observes its inventory $q_t \in (-q_{\max}, q_{\max})$ and the price $S_t \in \mathbb{R}_+$
- trades quantities $a_t \in (-a_{\max}, a_{\max})$ of the asset
- faces cost transactions and a terminal penalty imposed by the market
- receives a cost that affects its wealth $y_t \in \mathbb{R}$, $y_0 = 0$

Statistical Arbitrage Example

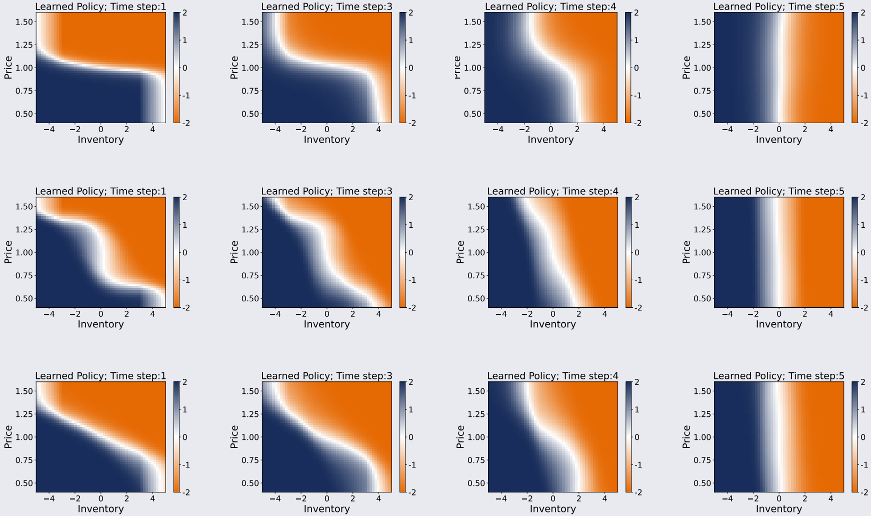
Consider a market with a single asset. An agent:

- invests during T periods, denoted $t = 0, \ldots, T-1$
- observes its inventory $q_t \in (-q_{\max}, q_{\max})$ and the price $S_t \in \mathbb{R}_+$
- trades quantities $a_t \in (-a_{\max}, a_{\max})$ of the asset
- faces cost transactions and a terminal penalty imposed by the market
- receives a cost that affects its wealth $y_t \in \mathbb{R}$, $y_0 = 0$





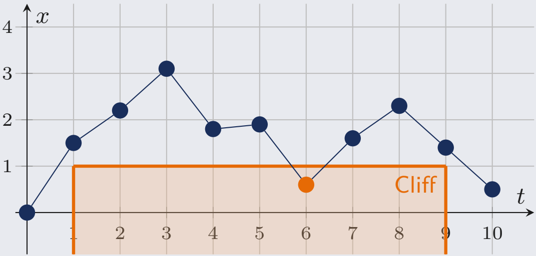
Statistical Arbitrage Example



Cliff Walking Example

Consider an autonomous rover that:

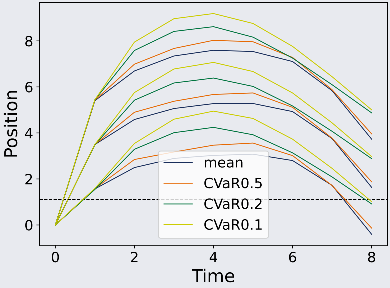
- starts at $(0,0)$ and wants to go at $(T,0)$
- moves from (t,x_1) to $(t+1,x_2)$, which incurs a cost
- receives a big penalty when stepping into the cliff
- takes actions $a_t^\theta \sim \pi^\theta = \mathcal{N}(\mu^\theta, \sigma)$, with $\mu^\theta \in (-a_{\max}, a_{\max})$
- gets a penalty when landing further from the goal at (T,x)

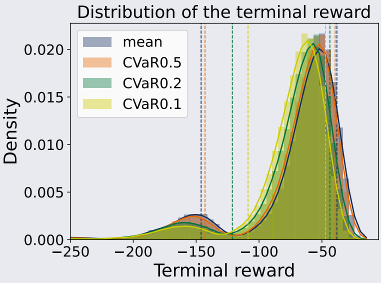


Cliff Walking Example

Consider an autonomous rover that:

- starts at $(0, 0)$ and wants to go at $(T, 0)$
- moves from (t, x_1) to $(t + 1, x_2)$, which incurs a cost
- receives a big penalty when stepping into the cliff
- takes actions $a_t^\theta \sim \pi^\theta = \mathcal{N}(\mu^\theta, \sigma)$, with $\mu^\theta \in (-a_{\max}, a_{\max})$
- gets a penalty when landing further from the goal at (T, x)





Hedging with Friction Example

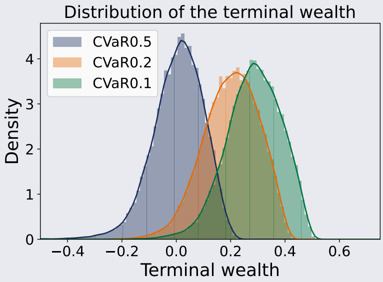
Consider a call option where the underlying asset dynamics follow the Heston model. An agent:

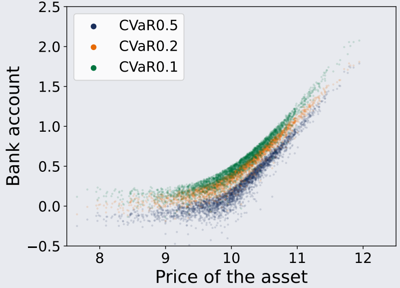
- sells the call option and aims to hedge it trading solely the asset
- observes its previous position a_t, its bank account B_t, and the price S_t
- trades in a market with transaction costs (per share) and an interest rate r
- receives a cost that affect its wealth y_t

Hedging with Friction Example

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

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





Contributions

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

- *Risk-sensitive* optimization with *non-stationary policies*
- Generalization to the broad class of *dynamic convex risk measures*

Future directions

- *Applications* on various problems (e.g. financial maths, grid worlds)
- Applications on data sets *with an offline setting*
- *Robust optimization* over Wasserstein balls
- *Computationally efficient* approach for large-scale problems

References

[ADEH99] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[BG20] Nicole Bäuerle and Alexander Glauner. Minimizing spectral risk measures applied to markov decision processes. *arXiv preprint arXiv:2012.04521*, 2020.

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

[CZ14] Shanyun Chu and Yi Zhang. Markov decision processes with iterated coherent risk measures. *International Journal of Control*, 87(11):2286–2293, 2014.

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

[KT00] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer, 2000.

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

[NBP19] David Nass, Boris Belousov, and Jan Peters. Entropic risk measure in policy search. *arXiv preprint arXiv:1906.09090*, 2019.

[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.

[TCGM15] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. *Advances in Neural Information Processing Systems*, 28:1468–1476, 2015.

Statistical Arbitrage Example

The agent:

- begins each episode with zero inventory
- observes the asset's price $S_t \in \mathbb{R}_+$ and their inventory $q_t \in (-q_{\max}, q_{\max})$
- performs a trade $a_t^\theta \in (-a_{\max}, a_{\max})$, resulting in wealth $y_t \in \mathbb{R}$ according to

$$\begin{cases} y_0 = 0, \\ y_t = y_{t-1} - a_{t-1}^\theta S_{t-1} - \varphi (a_{t-1}^\theta)^2, \qquad t = 1, \dots, T-1 \\ y_T = y_{T-1} - a_{T-1}^\theta S_{T-1} - \varphi (a_{T-1}^\theta)^2 + q_T S_T - \psi q_T^2. \end{cases}$$

The asset price follows an Ornstein-Uhlenbeck process:

$$\mathrm{d}S_t = \kappa(\mu - S_t)\mathrm{d}t + \sigma \mathrm{d}W_t$$

We suppose that $T = 5$, $q_{\max} = 5$, $a_{\max} = 2$, $\varphi = 0.005$ (transaction costs), $\psi = 0.5$ (terminal penalty), $\kappa = 2$, $\mu = 1$, $\sigma = 0.2$ and W_t is a standard $\mathbb{P}$-Brownian motion

Cliff Walking Example

Consider an autonomous rover that:

- starts at $(0, 0)$ and wants to go at $(T, 0)$
- moves from (t, x_1) to $(t + 1, x_2)$, which incurs a cost of $1 + (x_2 - x_1)^2$
- receives a penalty of 100 when stepping into the cliff $x \leq C$
- takes actions $a_t^\theta \sim \pi^\theta = \mathcal{N}(\mu^\theta, \sigma)$, with $\mu^\theta \in (-a_{\max}, a_{\max})$
- gets a penalty of size x^2 when landing further from the goal at (T, x)

We suppose that $T = 9$, $C = 1$, $a_{\max} = 4$, $\sigma = 1.5$

Hedging with Friction Example

The asset price $(S_t)_{t\in\mathcal{T}}$:

- is simulated using the Milstein discretization scheme
- evolves according to the Heston model

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

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

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

We suppose that $T = 10$ (over a month), $K = 10$, $\mu = 0.1$, $\kappa = 9$, $\vartheta = (0.25)^2$, $\varsigma = 1$, $(W_t^S)_{t\in\mathcal{T}}, (W_t^\nu)_{t\in\mathcal{T}}$ are two $\mathbb{P}$-Brownian motions with correlation $\rho = -0.5$, $S_0 = 10$, $\nu_0 = (0.2)^2$