Reinforcement Learning for Dynamic Convex Risk Measures

Anthony Coache Sebastian Jaimungal

anthonycoache.ca
sebastian.statistics.utoronto.ca

Department of Statistical Sciences
University of Toronto

SIAM Conference on Financial Mathematics and Engineering ⋆ June 1–4, 2021

Reinforcement Learning (RL)

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

- $\mathcal{S}$ – State space
- $\mathcal{A}$ – Action space
- $\pi^\theta(a|s)$ – Policy characterized by θ
- $P(s_1), P(s'|s,a)$ – Transition probability distribution
- $c(s,a) \in \mathcal{C}$ – State-action dependent cost function
- $\gamma \in (0,1)$ – Discount factor

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

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

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

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

Reinforcement Learning (RL)

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

- $\mathcal{S}$ – State space
- $\mathcal{A}$ – Action space
- $\pi^\theta(a|s)$ – Policy characterized by θ
- $P(s_1), P(s'|s,a)$ – Transition probability distribution
- $c(s,a) \in \mathcal{C}$ – State-action dependent cost function
- $\gamma \in (0,1)$ – Discount factor

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

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

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

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

Reinforcement Learning (RL)

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

- $\mathcal{S}$ – State space
- $\mathcal{A}$ – Action space
- $\pi^\theta(a|s)$ – Policy characterized by θ
- $P(s_1), P(s'|s,a)$ – Transition probability distribution
- $c(s,a) \in \mathcal{C}$ – State-action dependent cost function
- $\gamma \in (0,1)$ – Discount factor

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

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

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

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

Motivation

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 *convex* risk measures

Motivation

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 *convex* risk measures

Motivation

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 *convex* risk measures

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

If a risk measure ρ is convex, proper and lower semicontinuous, then 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]

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

Dual Representation

Representation Theorem [SDR14]

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

If a risk measure ρ is convex, proper and lower semicontinuous, then 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
- $\mathcal{F}_1 \subset \ldots \subset \mathcal{F}_T$ – 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 $1 \leq t_1 < t_2 \leq T$, and any $Z, W \in \mathcal{Z}_{t_1,T}$, we have

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

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
- $\mathcal{F}_1 \subset \ldots \subset \mathcal{F}_T$ – 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 $1 \leq t_1 < t_2 \leq T$, and any $Z, W \in \mathcal{Z}_{t_1,T}$, 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

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

Then [Rus10] we have

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

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

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

Then [Rus10] we have

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

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 $\min_\theta \rho_{1,T+1}(Z)$ induced by π^θ, i.e.

$$\min_\theta c(s_1, a_1) + \gamma\rho_1\left(c(s_2, a_2) + \cdots + \gamma\rho_{T-1}\left(c(s_T, a_T) + \gamma\rho_T\left(c(s_{T+1})\right)\cdots\right)\right)$$

Using the dual representation and recursive equations, we have

$$\begin{aligned} V_{T+1}(s) &= c_{T+1}(s), \\ V_t(s) &= \underbrace{c_t^\theta(s)}_{\text{cost for present state}} + \underbrace{\max_{\xi \in \mathcal{U}(s, P_\theta(\cdot|s_t=s))} \mathbb{E}^\xi\left[V_{t+1}(s_{t+1}^\theta) - \rho^*(\xi)\right]}_{\text{risk for next state}}, \end{aligned}$$

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

- $c_t^\theta(s) = \sum_a c_t(s,a)\pi^\theta(a|s_t = s)$ – Cost of π^θ
- $P_\theta(s'|s_t = s) = \sum_a P(s'|s,a)\pi^\theta(a|s_t = s)$ – Transition probability induced by π^θ

Problem Setup

Problems of the form $\min_\theta \rho_{1,T+1}(Z)$ induced by π^θ, i.e.

$$\min_\theta c(s_1, a_1) + \gamma\rho_1\left(c(s_2, a_2) + \cdots + \gamma\rho_{T-1}\left(c(s_T, a_T) + \gamma\rho_T\left(c(s_{T+1})\right)\cdots\right)\right)$$

Using the dual representation and recursive equations, we have

$$\begin{aligned} V_{T+1}(s) &= c_{T+1}(s), \\ V_t(s) &= \underbrace{c_t^\theta(s)}_{\text{cost for present state}} + \underbrace{\max_{\xi\in\mathcal{U}(s, P_\theta(\cdot|s_t=s))} \mathbb{E}^\xi\left[V_{t+1}(s_{t+1}^\theta) - \rho^*(\xi)\right]}_{\text{risk for next state}}, \end{aligned}$$

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

- $c_t^\theta(s) = \sum_a c_t(s, a)\pi^\theta(a|s_t = s)$ – Cost of π^θ
- $P_\theta(s'|s_t = s) = \sum_a P(s'|s, a)\pi^\theta(a|s_t = s)$ – Transition probability induced by π^θ

Problem Setup

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

$$L_t^{\theta,\phi}(\xi,\lambda) = \sum_{s'\in\mathcal{S}} \xi(s') P_\theta(s'|s) \left(V_{t+1}^{\phi}(s') - \rho^*(\xi(s'))\right) - \lambda \left(\sum_{s'\in\mathcal{S}} \xi(s') P_\theta(s'|s) - 1\right).$$

- The Envelope Theorem [MS02], says

$$\nabla_\theta \left(\max_{\xi\in\mathcal{U}(s,P_\theta(\cdot|s_t=s))} \mathbb{E}^{\xi}\left[V_{t+1}^{\phi}(s_{t+1}^{\theta}) - \rho^*(\xi)\right]\right) = \nabla_\theta L_t^{\theta,\phi}(\xi,\lambda)\Big|_{\xi^*,\lambda^*}$$

Problem Setup

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

$$L_t^{\theta,\phi}(\xi,\lambda) = \sum_{s'\in\mathcal{S}} \xi(s') P_\theta(s'|s) \left(V_{t+1}^{\phi}(s') - \rho^*(\xi(s'))\right) - \lambda \left(\sum_{s'\in\mathcal{S}} \xi(s') P_\theta(s'|s) - 1\right).$$

- The Envelope Theorem [MS02], says

$$\nabla_\theta \left(\max_{\xi \in \mathcal{U}(s, P_\theta(\cdot|s_t=s))} \mathbb{E}^{\xi}\left[V_{t+1}^{\phi}(s_{t+1}^{\theta}) - \rho^*(\xi)\right]\right) = \nabla_\theta L_t^{\theta,\phi}(\xi,\lambda)\Big|_{\xi^*,\lambda^*}$$

Problem Setup

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

$$L_t^{\theta,\phi}(\xi,\lambda) = \sum_{s'\in\mathcal{S}} \xi(s') P_\theta(s'|s) \left(V_{t+1}^{\phi}(s') - \rho^*(\xi(s'))\right) \\ - \lambda \left(\sum_{s'\in\mathcal{S}} \xi(s') P_\theta(s'|s) - 1\right).$$

- The Envelope Theorem [MS02], says

$$\nabla_\theta \left(\max_{\xi\in\mathcal{U}(s, P_\theta(\cdot|s_t=s))} \mathbb{E}^{\xi}\left[V_{t+1}^{\phi}(s_{t+1}^{\theta}) - \rho^*(\xi)\right] \right) = \nabla_\theta L_t^{\theta,\phi}(\xi,\lambda)\Big|_{\xi^*,\lambda^*}$$

Problem Setup

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

$$L_t^{\theta,\phi}(\xi,\lambda) = \sum_{s'\in\mathcal{S}} \xi(s') P_\theta(s'|s) \left(V_{t+1}^{\phi}(s') \;-\; \rho^*(\xi(s'))\right) \\ - \lambda \left(\sum_{s'\in\mathcal{S}} \xi(s') P_\theta(s'|s) - 1\right).$$

- The Envelope Theorem [MS02], says

$$\nabla_\theta \left(\max_{\xi\in\mathcal{U}(s,P_\theta(\cdot|s_t=s))} \mathbb{E}^{\xi}\left[V_{t+1}^{\phi}(s_{t+1}^{\theta}) - \rho^*(\xi)\right]\right) = \nabla_\theta L_t^{\theta,\phi}(\xi,\lambda)\Big|_{\xi^*,\lambda^*}$$

Problem Setup

Using an *ensemble of ANNs* $\{\pi^{\theta_t}\}_t$: $V_t^\phi(s) = V_t^\phi(s;\theta_t,\theta_{t+1},\dots)$

$$\nabla_{\theta_t} V_t^\phi(s) = \overbrace{\mathbb{E}\left[c_t(s,a_t^{\theta_t})\nabla_{\theta_t}\log\pi^{\theta_t}(a_t^{\theta_t}|s_t)\;\middle|\; s_t = s\right]}^{\text{actions}}$$
$$+\underbrace{\mathbb{E}^{\xi^*}\left[\left(V_t^\phi(s_{t+1}^{\theta_t}) - \rho^*(\xi^*) - \lambda^*\right)\nabla_{\theta_t}\log\pi^{\theta_t}(a_t^{\theta_t}|s_t)\;\middle|\; s_t = s\right]}_{\text{next states}}.$$

Using a *single ANN* π^θ: $V_t^\phi(s) = V_t^\phi(s;\theta)$

$$\nabla_\theta V_t^\phi(s) = \overbrace{\mathbb{E}\left[c_t(s,a_t^\theta)\,\nabla_\theta\log\pi^\theta(a_t^\theta|s_t)\;\middle|\; s_t = s\right]}^{\text{actions}} + \overbrace{\mathbb{E}^{\xi^*}\left[\nabla_\theta V_{t+1}^\phi(s_{t+1}^\theta)\;\middle|\; s_t = s\right]}^{\text{gradient of future } V\text{'s}}$$
$$+\underbrace{\mathbb{E}^{\xi^*}\left[\left(V_{t+1}^\phi(s_{t+1}^\theta) - \rho^*(\xi^*) - \lambda^*\right)\nabla_\theta\log\pi^\theta(a_t^\theta|s_t)\;\middle|\; s_t = s\right]}_{\text{next states}}.$$

Problem Setup

Using an *ensemble of ANNs* $\{\pi^{\theta_t}\}_t$: $V_t^\phi(s) = V_t^\phi(s;\theta_t,\theta_{t+1},\dots)$

$$\nabla_{\theta_t} V_t^\phi(s) = \overbrace{\mathbb{E}\left[c_t(s,a_t^{\theta_t})\nabla_{\theta_t}\log\pi^{\theta_t}(a_t^{\theta_t}|s_t)\;\middle|\; s_t = s\right]}^{\text{actions}}$$
$$+\underbrace{\mathbb{E}^{\xi^*}\left[\left(V_t^\phi(s_{t+1}^{\theta_t}) - \rho^*(\xi^*) - \lambda^*\right)\nabla_{\theta_t}\log\pi^{\theta_t}(a_t^{\theta_t}|s_t)\;\middle|\; s_t = s\right]}_{\text{next states}}.$$

Using a *single ANN* π^θ: $V_t^\phi(s) = V_t^\phi(s;\theta)$

$$\nabla_\theta V_t^\phi(s) = \overbrace{\mathbb{E}\left[c_t(s,a_t^\theta)\,\nabla_\theta\log\pi^\theta(a_t^\theta|s_t)\;\middle|\; s_t = s\right]}^{\text{actions}} + \overbrace{\mathbb{E}^{\xi^*}\left[\nabla_\theta V_{t+1}^\phi(s_{t+1}^\theta)\;\middle|\; s_t = s\right]}^{\text{gradient of future } V\text{'s}}$$
$$+\underbrace{\mathbb{E}^{\xi^*}\left[\left(V_{t+1}^\phi(s_{t+1}^\theta) - \rho^*(\xi^*) - \lambda^*\right)\nabla_\theta\log\pi^\theta(a_t^\theta|s_t)\;\middle|\; s_t = s\right]}_{\text{next states}}.$$

Algorithm

Actor-critic style algorithm 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 - Ensemble Approach

Input: Environment, risk measure, $\{\pi^{\theta_t}\}_t$, V^ϕ

```
for each period t = 1, ..., T do
    for each epoch κ = 1, ..., K do
        Generate trajectories with additional transitions for each state ;
        Estimate the value function (critic) ;
        Update the policy (actor) ;
```

Output: An optimal policy $\pi^\theta \approx \pi^*$

- *Simulation upon simulation* (or *nested simulation*) approach
- Function approximation for estimating the policy and value function

Estimation of the Value Function

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

$$V^{\phi}_{T+1}(s) = c_{T+1}(s)$$
$$V^{\phi}_{t}(s) = \underbrace{c^{\theta}_{t}(s)}_{\text{cost for present state}} + \underbrace{\max_{\xi \in \mathcal{U}(s, P_{\theta}(\cdot | s_t = s))} \left\{ \mathbb{E}^{\xi} \left[V^{\phi}_{t+1}(s^{\theta}_{t+1}) - \rho^{*}(\xi) \right] \right\}}_{\text{risk for the next state}}$$

$c^{\theta}_{t}(s)$: mean of $c_t(s, a)$ over transitions from π^{θ}

Risk measure: risk of V^{ϕ}_{t+1} for the next states of transitions from π^{θ}

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

Update of the Policy

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

$$\nabla_{\theta_t} V_t^{\phi}(s) = \overbrace{\mathbb{E}\left[c_t(s, a_t^{\theta_t})\, \nabla_{\theta_t} \log \pi^{\theta_t}(a_t^{\theta_t} | s_t) \;\middle|\; s_t = s \right]}^{\text{actions}}$$
$$\underbrace{+\, \mathbb{E}^{\xi^*}\left[\left(V_{t+1}^{\phi}(s_{t+1}^{\theta_t}) - \rho^*(\xi^*) - \lambda^* \right) \nabla_{\theta_t} \log \pi^{\theta_t}(a_t^{\theta_t} | s_t) \;\middle|\; s_t = s \right]}_{\text{next states}}.$$

$\pi^{\theta_t}(a_t^{\theta_t} | s_t = s)$: estimated by the reparameterization trick

V^{ϕ}: obtained using the critic

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

Trading Problem

Consider a market with a single asset. An agent:

- invests during T periods, denoted $t = 1, \ldots, T$
- observes its inventory $q_t \in (-q_{\max}, q_{\max})$ and the price $x_t \in \mathbb{R}_+$
- trades quantities $u_t \in (-u_{\max}, u_{\max})$ of the asset
- receives a cost that affects its wealth $y_t \in \mathbb{R}$

$$\begin{cases} y_1 = 0 \\ y_{t+1} = y_t - x_t u_t - \phi u_t^2, \quad t = 1, \ldots, T-1 \\ y_{T+1} = y_T - x_T u_T - \phi u_T^2 + q_{T+1} x_{T+1} - \psi q_{T+1}^2 \end{cases} .$$

Different risk measures

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

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

Trading Problem

Consider a market with a single asset. An agent:

- invests during T periods, denoted $t = 1, \dots, T$
- observes its inventory $q_t \in (-q_{\max}, q_{\max})$ and the price $x_t \in \mathbb{R}_+$
- trades quantities $u_t \in (-u_{\max}, u_{\max})$ of the asset
- receives a cost that affects its wealth $y_t \in \mathbb{R}$

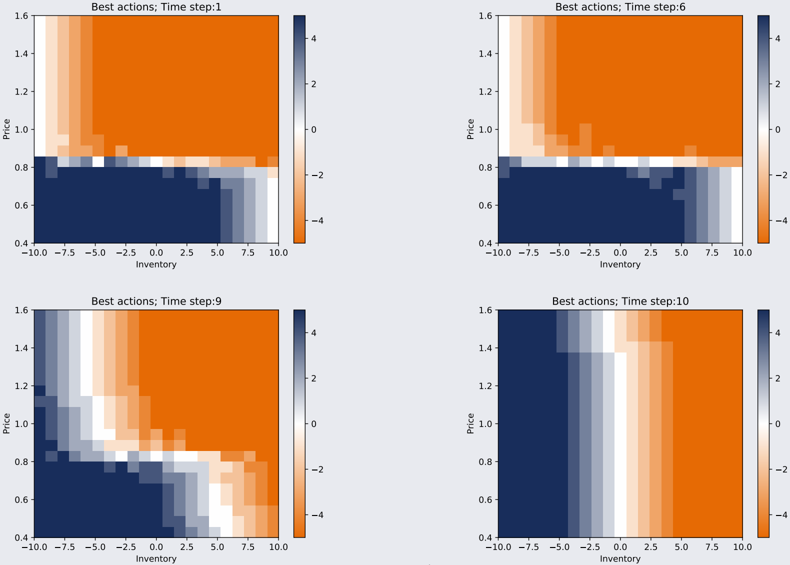
$$\begin{cases} y_1 = 0 \\ y_{t+1} = y_t - x_t u_t - \phi u_t^2, \quad t = 1, \dots, T-1 \\ y_{T+1} = y_T - x_T u_T - \phi u_T^2 + q_{T+1} x_{T+1} - \psi q_{T+1}^2 \end{cases} .$$

Different risk measures

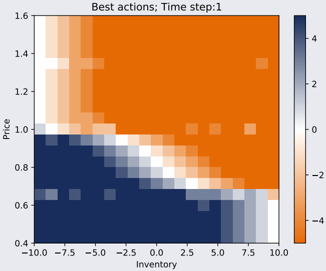
- Expectation: $\rho_{\mathbb{E}}(Z) = \mathbb{E}[Z]$
- Conditional value-at-risk (CVaR): $\rho_{\mathsf{CVaR}}(Z; \alpha) = \sup_{\xi \in \mathcal{U}(P)} \left\{ \mathbb{E}^{\xi}[Z] \right\}$
- Penalized CVaR: $\rho_{\mathsf{CVaR\text{-}p}}(Z; \alpha, \kappa) = \sup_{\xi \in \mathcal{U}(P)} \left\{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

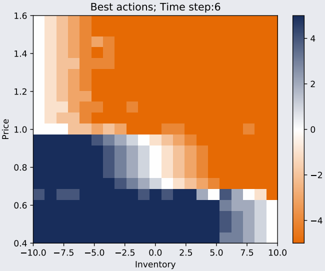
where $\mathcal{U}(P) = \left\{ \xi : \sum_{\omega} \xi(\omega) P(\omega) = 1,\ \xi \in [0, 1/\alpha] \right\}$

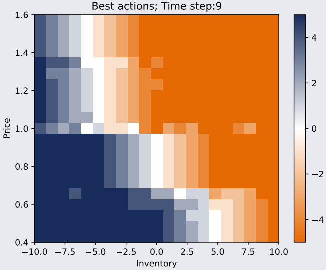
Optimal policy - Expectation

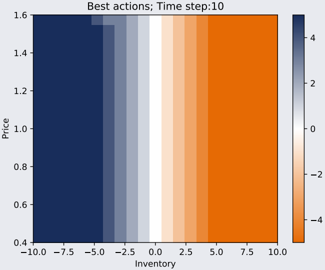


Optimal policy - CVaR

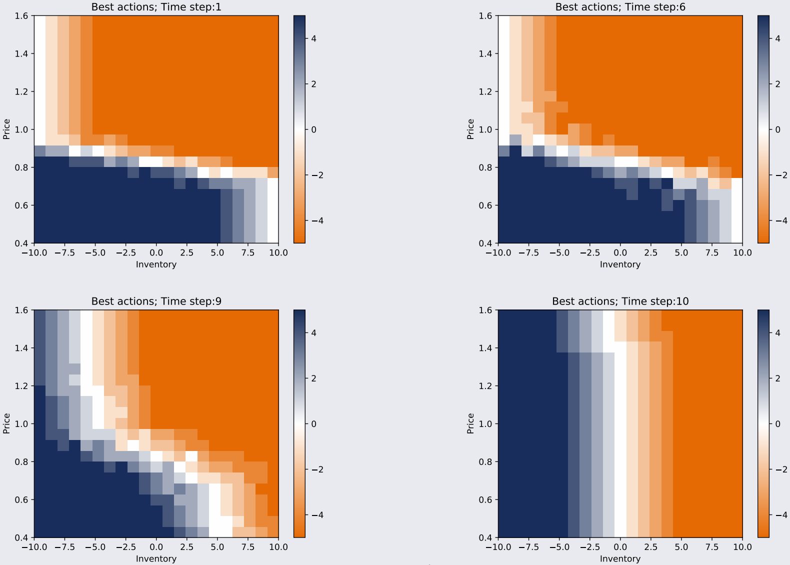




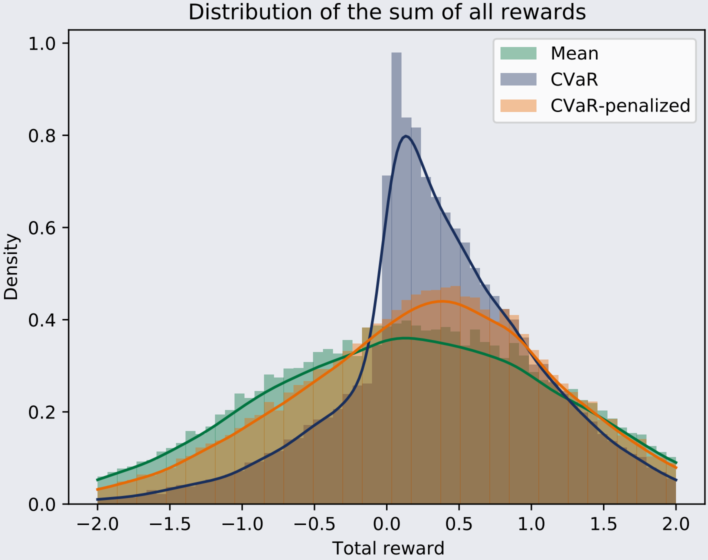




Optimal policy - Penalized CVaR



Terminal Reward Under Optimal Policy



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

- *Computationally efficient* approach for large-scale problems
- *Multi-agent system framework* to solve these problems
- *Deep Deterministic Policy Gradient* with dynamic risk measures
- *Applications* on various financial maths problems

References

Thank you!

[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.

[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.

[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.