Risk-Sensitive Reinforcement Learning with Dynamic Risk Measures

Anthony Coache   Sebastian Jaimungal

anthonycoache.ca
sebastian.statistics.utoronto.ca

Department of Statistical Sciences
University of Toronto

Summer Research Retreat ⋆ August 11–13, 2021

UNIVERSITY OF TORONTO

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

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

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

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

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

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

Risk-Sensitive RL

Risk-aware RL: applying risk measures *recursively* at each period [e.g. Rus10]

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

[TCGM15] provide policy search algorithms in the dynamic framework

- 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-Sensitive RL

Risk-aware RL: applying risk measures *recursively* at each period [e.g. Rus10]

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

[TCGM15] provide policy search algorithms in the dynamic framework

- 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-Sensitive RL

Risk-aware RL: applying risk measures *recursively* at each period [e.g. Rus10]

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

[TCGM15] provide policy search algorithms in the dynamic framework

- 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

Convex Risk Measures

Convex $\rho : \mathcal{Z} \to \mathbb{R}$ [FS02]

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

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

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

Convex Risk Measures

Convex $\rho : \mathcal{Z} \to \mathbb{R}$ [FS02]

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

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

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

Dynamic Convex Risk Measures

Consider

- $\mathcal{F}_1 \subset \ldots \subset \mathcal{F}_T$ – Filtration on $(\Omega, \mathcal{F}, P)$
- $\mathcal{Z}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ – p-integrable random variables

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

Sequence of $\rho_{t,T} : \mathcal{Z}_t \times \cdots \times \mathcal{Z}_T \to \mathcal{Z}_t$ 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 $Z, W \in \mathcal{Z}_{t_1,T}$,

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

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

Then for a time-consistent $\{\rho_{t,T}\}_t$, we have [Rus10]

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

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

Dynamic Convex Risk Measures

Consider

- $\mathcal{F}_1 \subset \ldots \subset \mathcal{F}_T$ – Filtration on $(\Omega, \mathcal{F}, P)$
- $\mathcal{Z}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ – p-integrable random variables

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

Sequence of $\rho_{t,T} : \mathcal{Z}_t \times \cdots \times \mathcal{Z}_T \to \mathcal{Z}_t$ 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 $Z, W \in \mathcal{Z}_{t_1,T}$,

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

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

Then for a time-consistent $\{\rho_{t,T}\}_t$, we have [Rus10]

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

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

Dynamic Convex Risk Measures

Consider

- $\mathcal{F}_1 \subset \ldots \subset \mathcal{F}_T$ – Filtration on $(\Omega, \mathcal{F}, P)$
- $\mathcal{Z}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ – p-integrable random variables

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

Sequence of $\rho_{t,T} : \mathcal{Z}_t \times \cdots \times \mathcal{Z}_T \to \mathcal{Z}_t$ 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 $Z, W \in \mathcal{Z}_{t_1,T}$,

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

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

Then for a time-consistent $\{\rho_{t,T}\}_t$, we have [Rus10]

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

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

Problem Setup

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

$$\min_\theta \, \rho_0 \left(c(s_0, a_0^\theta) + \rho_1 \left(c(s_1^\theta, a_1^\theta) + \cdots + \rho_{T-1} \left(c(s_{T-1}^\theta, a_{T-1}^\theta) + c(s_T^\theta) \right) \cdots \right) \right)$$

Using the dual representation and recursive equations, we have

$$V_T(s;\theta) = c_T(s),$$

$$V_t(s;\theta) = \max_{\xi \in \mathcal{U}(s, P^\theta(\cdot,\cdot|s_t=s))} \left\{ \mathbb{E}^\xi \Big[\underbrace{c_t(s, a_t)}_{\text{cost for present state}} + \underbrace{V_{t+1}(s_{t+1};\theta)}_{\text{risk for next state}} \Big] - \rho_t^*(\xi) \right\},$$

for $s \in \mathcal{S}$ and $t = T-1, \ldots, 0$, where

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

Problem Setup

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

$$\min_\theta \rho_0 \left(c(s_0, a_0^\theta) + \rho_1 \left(c(s_1^\theta, a_1^\theta) + \cdots + \rho_{T-1} \left(c(s_{T-1}^\theta, a_{T-1}^\theta) + c(s_T^\theta) \right) \cdots \right) \right)$$

Using the dual representation and recursive equations, we have

$$V_T(s;\theta) = c_T(s),$$

$$V_t(s;\theta) = \max_{\xi \in \mathcal{U}(s, P^\theta(\cdot,\cdot|s_t = s))} \left\{ \mathbb{E}^\xi \Big[\underbrace{c_t(s, a_t)}_{\text{cost for present state}} + \underbrace{V_{t+1}(s_{t+1};\theta)}_{\text{risk for next state}} \Big] - \rho_t^*(\xi) \right\},$$

for $s \in \mathcal{S}$ and $t = T-1, \ldots, 0$, where

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

Problem Setup

We wish to optimize the value function ϕ over policies θ

The Envelope Theorem [MS02] states

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

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

$$\nabla_{\theta_t} V_t^\phi(s) = \underbrace{\mathbb{E}^{\xi^*} \left[\left(c_t(s, a_t^{\theta_t}) + V_t^\phi(s_{t+1}^{\theta_t}) - \lambda^* \right) \nabla_{\theta_t} \log \pi^{\theta_t}(a_t^{\theta_t} | s_t) \,\middle|\, s_t = s \right]}_{\text{transition}} - \underbrace{\nabla_\theta \rho_t^*(\xi^*)}_{\text{convex penalty}}$$

Problem Setup

We wish to optimize the value function ϕ over policies θ

The Envelope Theorem [MS02] states

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

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

$$\nabla_{\theta_t} V_t^\phi(s) = \underbrace{\mathbb{E}^{\xi^*}\left[\left(c_t(s, a_t^{\theta_t}) + V_t^\phi(s_{t+1}^{\theta_t}) - \lambda^* \right) \nabla_{\theta_t} \log \pi^{\theta_t}(a_t^{\theta_t}|s_t) \,\middle|\, s_t = s \right]}_{\text{transition}} - \underbrace{\nabla_\theta \rho_t^*(\xi^*)}_{\text{convex penalty}}$$

Problem Setup

We wish to optimize the value function ϕ over policies θ

The Envelope Theorem [MS02] states

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

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

$$\nabla_{\theta_t} V_t^\phi(s) = \underbrace{\mathbb{E}^{\xi^*} \left[\left(c_t(s, a_t^{\theta_t}) + V_t^\phi(s_{t+1}^{\theta_t}) - \lambda^* \right) \nabla_{\theta_t} \log \pi^{\theta_t}(a_t^{\theta_t} | s_t) \;\middle|\; s_t = s \right]}_{\text{transition}} - \underbrace{\nabla_\theta \rho_t^*(\xi^*)}_{\text{convex penalty}}$$

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

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

```
for each period t = T, ..., 1 do
    for each epoch κ = 1, ..., K do
        Generate transitions for a batch of states ;
        Estimate the value function (critic) ;
        Generate transitions for a batch of states ;
        Update the policy (actor) ;
```

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

- Function approximation for estimating the policy and value function

Algorithm

Estimation of the value function V^{ϕ}:

$$V_t^{\phi}(s) = \max_{\xi \in \mathcal{U}(s, P^{\theta}(\cdot,\cdot|s_t=s))} \left\{ \mathbb{E}^{\xi} \left[\underbrace{c_t(s, a_t)}_{\text{cost for present state}} + \underbrace{V_{t+1}^{\phi}(s_{t+1})}_{\text{risk for next state}} \right] - \rho_t^*(\xi) \right\}$$

- ANN $V_t^{\phi} : s_t \mapsto \mathbb{R}$
- Expected square loss between predicted and target values

Update of the policy π^{θ}:

$$\nabla_{\theta_t} V_t^{\phi}(s) = \underbrace{\mathbb{E}^{\xi^*} \left[\left(c_t(s, a_t^{\theta_t}) + V_t^{\phi}(s_{t+1}^{\theta_t}) - \lambda^* \right) \nabla_{\theta_t} \log \pi^{\theta_t}(a_t^{\theta_t} | s_t) \,\middle|\, s_t = s \right]}_{\text{transition}} - \underbrace{\nabla_{\theta} \rho_t^*(\xi^*)}_{\text{convex penalty}}$$

- ANN $\pi^{\theta_t} : s_t \mapsto \mathcal{P}(\mathcal{A})$
- Gradient descent step with $\nabla_{\theta_t} V_t^{\phi}$

Algorithm

Estimation of the value function V^{ϕ}:

$$V_t^{\phi}(s) = \max_{\xi \in \mathcal{U}(s, P^{\theta}(\cdot,\cdot|s_t = s))} \left\{ \mathbb{E}^{\xi}\left[\underbrace{c_t(s, a_t)}_{\text{cost for present state}} + \underbrace{V_{t+1}^{\phi}(s_{t+1})}_{\text{risk for next state}} \right] - \rho_t^*(\xi) \right\}$$

- ANN $V_t^{\phi} : s_t \mapsto \mathbb{R}$
- Expected square loss between predicted and target values

Update of the policy π^{θ}:

$$\nabla_{\theta_t} V_t^{\phi}(s) = \underbrace{\mathbb{E}^{\xi^*}\left[\left(c_t(s, a_t^{\theta_t}) + V_t^{\phi}(s_{t+1}^{\theta_t}) - \lambda^* \right) \nabla_{\theta_t} \log \pi^{\theta_t}(a_t^{\theta_t} | s_t) \,\middle|\, s_t = s \right]}_{\text{transition}} - \underbrace{\nabla_{\theta} \rho_t^*(\xi^*)}_{\text{convex penalty}}$$

- ANN $\pi^{\theta_t} : s_t \mapsto \mathcal{P}(\mathcal{A})$
- Gradient descent step with $\nabla_{\theta_t} V_t^{\phi}$

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 $y_t \in \mathbb{R}$, $y_1 = 0$

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

Different risk measures

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

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 $t = 1, \ldots, T$
- 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 $y_t \in \mathbb{R}$, $y_1 = 0$

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

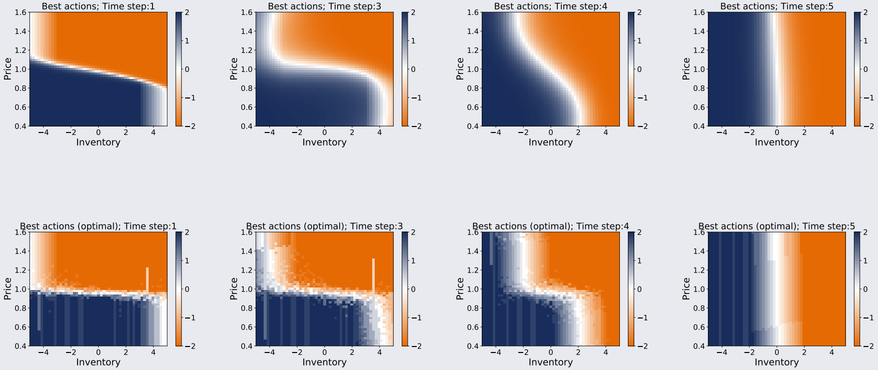
Different risk measures

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

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

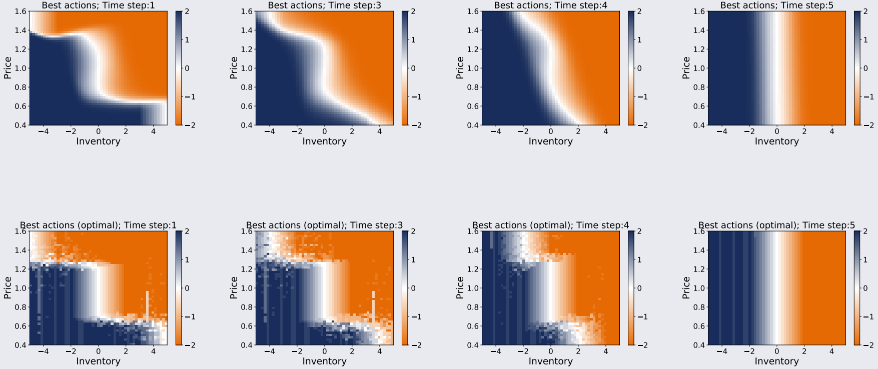
Optimal policy – Expectation

- $\rho_{\mathbb{E}}$



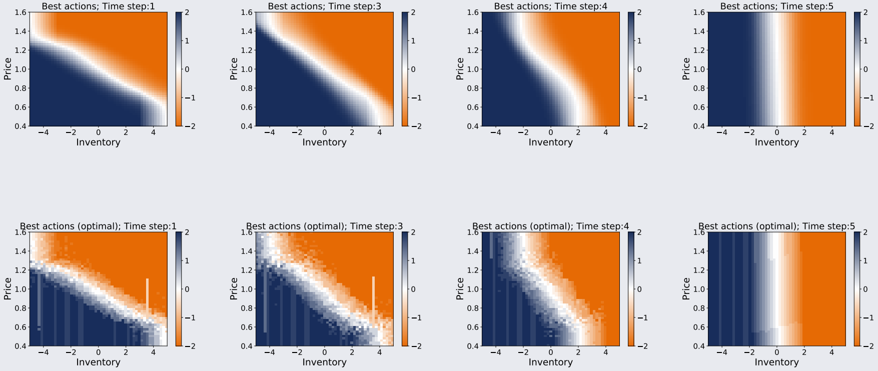
Optimal policy – CVaR

- ρ_{CVaR} with $\alpha = 0.2$

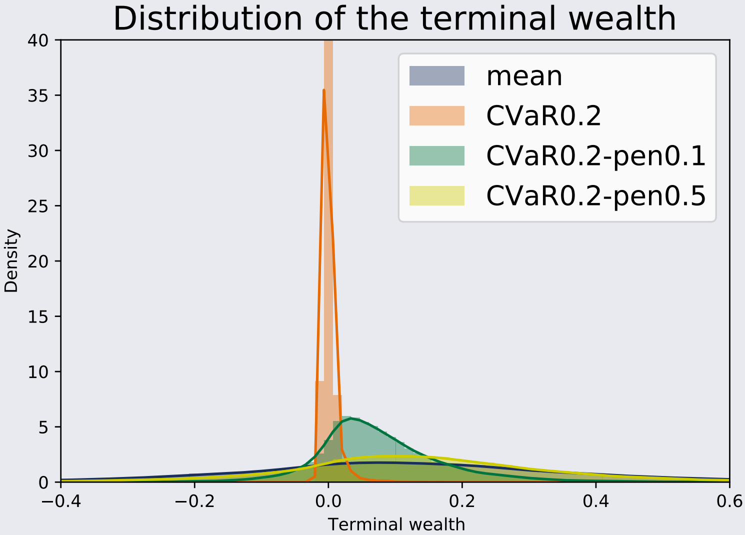


Optimal policy – Penalized CVaR

- $\rho_{\text{CVaR-p}}$ with $\alpha = 0.2$, $\kappa = 0.1$



Terminal Reward Under Learned Policies



Contributions & References

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

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

Future directions

- Implementation with *a single ANN*
- Various *applications* (e.g. financial maths, grid worlds, offline setting)
- *Deep Deterministic Policy Gradient* with dynamic risk measures

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

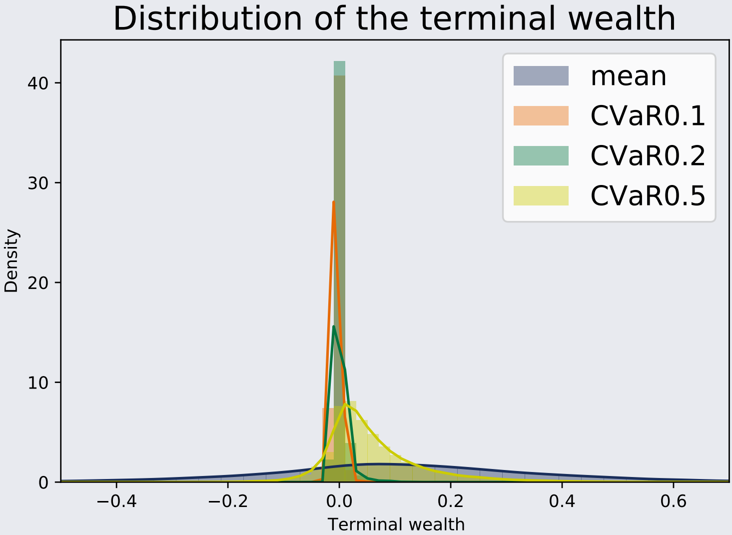
[MS02] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[Rus10] Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

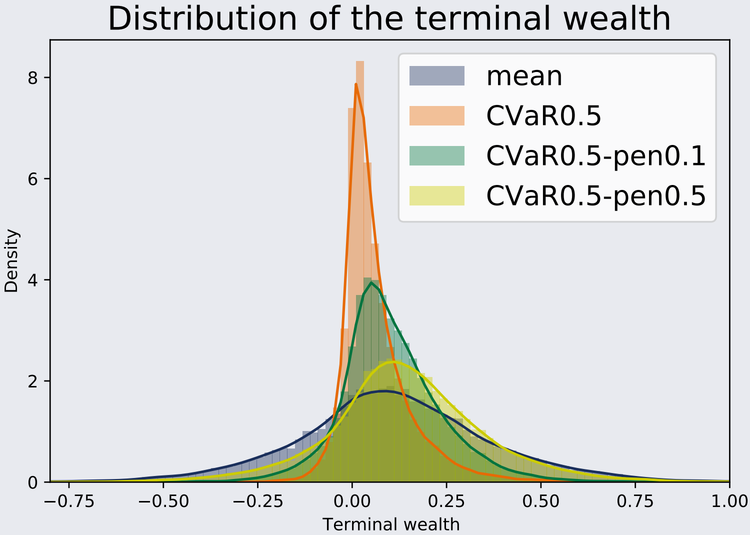
[SDR14] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[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.

Terminal Reward Under Learned Policies



Terminal Reward Under Learned Policies



Risk Envelope & Gradient Formula

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

$$\mathcal{U}(s, P^\theta(\cdot,\cdot|s)) = \Bigg\{ \xi P^\theta : \sum_{(a,s')} \xi(a,s') P^\theta(a,s'|s) = 1,\ \xi \geq 0,$$

$$\underbrace{g_e(\xi, P^\theta) = 0, \forall e \in \mathcal{E},}_{\text{equality constraints}} \quad \underbrace{f_i(\xi, P^\theta) \leq 0, \forall i \in \mathcal{I}}_{\text{inequality constraints}} \Bigg\}.$$

The full *gradient formula* is

$$\nabla_\theta V_t(s;\theta) = \overbrace{\mathbb{E}^{\xi^*}\left[\left(c_t(s,a) + V_{t+1}(s^\theta_{t+1};\theta) - \lambda^* \right) \nabla_\theta \log \pi^\theta(a|s_t = s) \right]}^{\text{transition}}$$

$$\underbrace{- \sum_{e \in \mathcal{E}} \left(\lambda^{*,\mathcal{E}}(e) \nabla_\theta g_e(\xi^*, P^\theta) \right)}_{\text{equality constraints}} \underbrace{- \sum_{i \in \mathcal{I}} \left(\lambda^{*,\mathcal{I}}(i) \nabla_\theta f_i(\xi^*, P^\theta) \right)}_{\text{inequality constraints}}$$

$$- \underbrace{\nabla_\theta \rho^*_t(\xi^*)}_{\text{conjugate}}$$

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