Risk-Sensitive Optimization in Reinforcement Learning

Anthony Coache

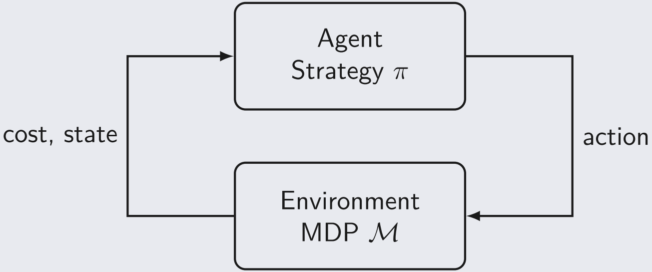
Department of Statistical Sciences
University of Toronto

ACTSCI / MAFI Research Meeting ⋆ January 28, 2021

Ideas Behind Reinforcement Learning

RL

- Idea: Collect data via an interactive process over many time steps
- Goal: Find a behavior which minimizes a cost



- The environment is often represented as a Markov decision process

Markov Decision Process

MDP

A **Markov decision process** is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, C, P, \pi, \gamma)$, where

- $\mathcal{S}$ – State space
  - Information available from the environment
- $\mathcal{A}$ – Action space
  - Action taken at a certain time
- $C(s, a) \in [-C_{\max}, C_{\max}]$ – State-action dependent cost function
  - Cost when being in state s and action a is taken
- $P(s_0), P(s' \mid s, a)$ – Transition probability distribution
  - Probability of being in state s' if in state s and action a is taken
- $\pi(a \mid s)$ – Policy
  - Probability of taking action a when being in state s
- $\gamma \in (0, 1)$ – Discount factor

Risk-sensitive RL

One trajectory of length T from $\mathcal{M}$ is denoted by

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T).$$

Let Z be the cumulative discounted cost of a trajectory induced by $\mathcal{M}$ with a policy π

$$Z(\tau) = C(a_0, s_0) + \gamma C(a_1, s_1) + \cdots + \gamma^T C(s_T).$$

Standard RL deals with a risk-neutral objective function of the cost Z

$$\min_\pi \mathbb{E}\left[Z(\tau)\right].$$

Optimization problem

Risk-sensitive RL considers problems in which the objective involves a risk measure ρ:

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

Risk-sensitive RL

- Risk-awareness provides strategies that are more robust to the environment
  - Autonomous car that accounts for environmental uncertainties, investing strategy that avoids losses of large amount of money, etc.

- Assumption of risk-aversion (as opposed to risk-neutrality) raises the complexity

- Risk sensitive criteria often lead to non-standard MDPs
  - Extend the state space to recover an ordinary MDP for CVaR optimization (Chow et al., 2015)

- Problem cannot be solved in a straightforward way by using Bellman equation, time-inconsistency issue
  - Adapt theory of risk measures to dynamic programming models with Markov risk measures (Ruszczyński, 2010)

Risk-sensitive RL

- Risk-awareness provides strategies that are more robust to the environment
  - Autonomous car that accounts for environmental uncertainties, investing strategy that avoids losses of large amount of money, etc.
- Assumption of risk-aversion (as opposed to risk-neutrality) raises the complexity
- Risk sensitive criteria often lead to non-standard MDPs
  - Extend the state space to recover an ordinary MDP for CVaR optimization (Chow et al., 2015)
- Problem cannot be solved in a straightforward way by using Bellman equation, time-inconsistency issue
  - Adapt theory of risk measures to dynamic programming models with Markov risk measures (Ruszczyński, 2010)

Optimization with Coherent Risk Measures

Policy Gradient for Coherent Risk Measures

A. Tamar, Y. Chow, M. Ghavamzadeh, S. Mannor, NeurIPS 2015.

- A gradient estimation algorithm for general coherent risk measures
  - Using sampling and convex programming
  - Consistency result provided
- A policy gradient theorem for Markov coherent risk measures
  - Dynamic programming approach to obtain a Bellman equation
  - Actor-critic algorithm for learning optimal policies

Coherent Risk Measures

Coherence

Consider two random variables X and Y. A risk measure ρ is said to be **coherent** (Artzner et al., 1999) if

- (Convexity) $\rho(\lambda X + (1-\lambda)Y) = \lambda\rho(X) + (1-\lambda)\rho(Y)$, $\forall\lambda \in [0,1]$
  - Diversification is favored by the risk measure.
- (Monotonicity) If $X \leq Y$, then $\rho(X) \leq \rho(Y)$
  - A portfolio with a higher cost for every scenario is indeed riskier.
- (Translation invariance) For all $a \in \mathbb{R}$, $\rho(X + a) = \rho(X) + a$
  - The deterministic part of a portfolio does not contribute to its risk.
- (Positive homogeneity) If $\lambda \geq 0$, then $\rho(\lambda X) = \lambda\rho(X)$
  - The risk is proportional to the size of the portfolio.

Duality Result

Representation Theorem

(Shapiro et al., 2014) A risk measure ρ is coherent iff. there exists a convex, bounded and closed set $\mathcal{U} \in \{P : \int_{\omega \in \Omega} P(\omega) = 1, P \geq 0\}$ called **risk envelope** such that

$$\rho(X) = \max_{\xi : \xi P \in \mathcal{U}(P)} \mathbb{E}^{\xi}[X] = \max_{\xi : \xi P \in \mathcal{U}(P)} \sum_{\omega \in \Omega} \xi(\omega) P(\omega) X(\omega).$$

In (Tamar et al., 2015), they assume that

$$\mathcal{U}(P) = \Big\{ \xi P : \sum_{\omega \in \Omega} \xi(\omega) P(\omega) = 1,\ \xi \geq 0,$$
$$g_e(\xi, P) = 0, \forall e \in \mathcal{E},\ f_i(\xi, P) \leq 0, \forall i \in \mathcal{I} \Big\},$$

where $g_e(\xi, P)$ are affine functions w.r.t. ξ, $f_i(\xi, P)$ are convex functions w.r.t. ξ, and $\mathcal{E}$ (resp. $\mathcal{I}$) denotes the set of equality (resp. inequality) constraints.

Duality Result

Representation Theorem

(Shapiro et al., 2014) A risk measure ρ is coherent iff. there exists a convex, bounded and closed set $\mathcal{U} \in \{P : \int_{\omega \in \Omega} P(\omega) = 1, P \geq 0\}$ called **risk envelope** such that

$$\rho(X) = \max_{\xi : \xi P \in \mathcal{U}(P)} \mathbb{E}^{\xi}[X] = \max_{\xi : \xi P \in \mathcal{U}(P)} \sum_{\omega \in \Omega} \xi(\omega) P(\omega) X(\omega).$$

In (Tamar et al., 2015), they assume that

$$\mathcal{U}(P) = \Big\{ \xi P : \sum_{\omega \in \Omega} \xi(\omega) P(\omega) = 1, \ \xi \geq 0, \quad g_e(\xi, P) = 0, \forall e \in \mathcal{E}, \ f_i(\xi, P) \leq 0, \forall i \in \mathcal{I} \Big\},$$

where $g_e(\xi, P)$ are affine functions w.r.t. ξ, $f_i(\xi, P)$ are convex functions w.r.t. ξ, and $\mathcal{E}$ (resp. $\mathcal{I}$) denotes the set of equality (resp. inequality) constraints.

Static Risk Problem

All actions are chosen according to a policy $\pi_\theta(\cdot|s)$, parameterized by θ. For a coherent risk measure ρ, the problem to solve is

$$\min_\theta \rho(Z) = \min_\theta \max_{\xi : \xi P_\theta \in \mathcal{U}(P_\theta)} \sum_{\omega \in \Omega} \xi(\omega) P_\theta(\omega) Z(\omega)$$

Z could be the cumulative discounted cost of a trajectory induced by $\mathcal{M}$ with a policy π_θ

- Use the assumption on $\mathcal{U}$ to write the Lagrangian function of $\rho(Z)$

$$L_\theta(\xi, \lambda^P, \lambda^\mathcal{E}, \lambda^\mathcal{I}) = \overbrace{\sum_{\omega \in \Omega} \xi(\omega) P_\theta(\omega) Z(\omega)}^{\text{risk measure}} - \overbrace{\lambda^P \left(\sum_{\omega \in \Omega} \xi(\omega) P_\theta(\omega) - 1 \right)}^{\text{density constr. on } \xi P_\theta}$$
$$- \underbrace{\sum_{e \in \mathcal{E}} \left(\lambda^\mathcal{E}(e) g_e(\xi, P_\theta) \right)}_{\text{equality constr. } \mathcal{E}} - \underbrace{\sum_{i \in \mathcal{I}} \left(\lambda^\mathcal{I}(i) f_i(\xi, P_\theta) \right)}_{\text{inequality constr. } \mathcal{I}}.$$

Static Risk Problem

All actions are chosen according to a policy $\pi_\theta(\cdot|s)$, parameterized by θ. For a coherent risk measure ρ, the problem to solve is

$$\min_\theta \rho(Z) = \min_\theta \max_{\xi : \xi P_\theta \in \mathcal{U}(P_\theta)} \sum_{\omega \in \Omega} \xi(\omega) P_\theta(\omega) Z(\omega)$$

Z could be the cumulative discounted cost of a trajectory induced by $\mathcal{M}$ with a policy π_θ

- Use the assumption on $\mathcal{U}$ to write the Lagrangian function of $\rho(Z)$

$$L_\theta(\xi, \lambda^P, \lambda^\mathcal{E}, \lambda^\mathcal{I}) = \overbrace{\sum_{\omega \in \Omega} \xi(\omega) P_\theta(\omega) Z(\omega)}^{\text{risk measure}} - \overbrace{\lambda^P \left(\sum_{\omega \in \Omega} \xi(\omega) P_\theta(\omega) - 1 \right)}^{\text{density constr. on } \xi P_\theta}$$
$$- \underbrace{\sum_{e \in \mathcal{E}} \left(\lambda^\mathcal{E}(e) g_e(\xi, P_\theta) \right)}_{\text{equality constr. } \mathcal{E}} - \underbrace{\sum_{i \in \mathcal{I}} \left(\lambda^\mathcal{I}(i) f_i(\xi, P_\theta) \right)}_{\text{inequality constr. } \mathcal{I}} .$$

Static Risk Problem

Gradient formula (static)

For any saddle point $(\xi^*, \lambda^{*,P}, \lambda^{*,\mathcal{E}}, \lambda^{*,\mathcal{I}})$ of L_θ, we have

$$\nabla_\theta \rho(Z) = \mathbb{E}^{\xi^*}\left[\nabla_\theta \log P_\theta(\omega)\left(Z - \lambda^{*,P}\right)\right] - \underbrace{\sum_{e \in \mathcal{E}} \left(\lambda^{*,\mathcal{E}}(e) \nabla_\theta g_e(\xi^*, P_\theta)\right)}_{\text{equality constr. } \mathcal{E}} - \underbrace{\sum_{i \in \mathcal{I}} \left(\lambda^{*,\mathcal{I}}(i) \nabla_\theta f_i(\xi^*, P_\theta)\right)}_{\text{inequality constr. } \mathcal{I}}.$$

- Saddle-point known analytically: Sampling-based estimator
- Saddle-point not known analytically: Convex optimization step, sampling step

Examples

Expectation

The expectation is a coherent risk measure, since

$$\rho_E(Z) = \mathbb{E}[Z] = \max_{\xi\,:\,\xi \in \mathcal{U}} \mathbb{E}^{\xi}[Z],$$

where its risk envelope is

$$\mathcal{U} = \{\xi \mid \xi \equiv 1\}.$$

Any saddle point $(\xi^*, \lambda^{*,P})$ satisfies $\xi^* = 1$ and $\lambda^{*,P} = 0$. Therefore,

$$\nabla_\theta \rho_E(Z) = \mathbb{E}\left[Z \, \nabla_\theta \log P_\theta(\omega)\right].$$

We recover the result for a risk-neutral objective (Sutton and Barto, 2018)

Examples

Conditional value-at-risk

The conditional value-at-risk (Rockafellar et al., 2000) is

$$\rho_{\mathrm{CVaR}}(Z,\alpha) = \inf_{t\in\mathbb{R}} \left\{ t + \alpha^{-1}\,\mathbb{E}\left[(Z-t)_+\right]\right\}$$
$$= \max_{\xi\,:\,\xi P_\theta \in \mathcal{U}(P_\theta)} \mathbb{E}^{\xi}\left[Z\right]$$

where

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

Any saddle point $(\xi^*, \lambda^{*,P})$ satisfies $\xi^*(\omega) = \frac{1}{\alpha}$ if $Z(\omega) > \lambda^{*,P}$ and $\xi^*(\omega) = 0$ otherwise, where $\lambda^{*,P}$ is any $(1-\alpha)$-quantile of Z. Therefore we obtain

$$\nabla_\theta \rho_{\mathrm{CVaR}}(Z,\alpha) = \mathbb{E}\left[(Z - q_\alpha)\,\nabla_\theta \log P_\theta(\omega) \mid Z > q_\alpha\right].$$

All spectral risk measures (Acerbi, 2002) are also coherent risk measures.

Policy Gradient Algorithm

How to compute $\nabla_\theta \log P_\theta(\omega)$ in the gradient formula?

$$\nabla_\theta \rho_E(Z) = \mathbb{E}\left[Z\,\nabla_\theta \log P_\theta(\omega)\right]$$
$$\nabla_\theta \rho_{\text{CVaR}}(Z,\alpha) = \mathbb{E}\left[(Z - q_\alpha)\,\nabla_\theta \log P_\theta(\omega) \mid Z > q_\alpha\right].$$

The gradient of the log-probability of a trajectory is

$$\nabla_\theta \log\left(P(\tau|\pi_\theta)\right) = \nabla_\theta \log\left(p(s_0)\prod_{t=0}^{T-1}\pi_\theta(a_t|s_t)P(s_{t+1}|a_t,s_t)\right)$$
$$= \sum_{t=0}^{T-1}\nabla_\theta \log\left(\pi_\theta(a_t|s_t)\right).$$

- $\nabla_\theta \log P_\theta$ depends only on $\nabla_\theta \log \pi_\theta$.

Policy Gradient Algorithm

$$\theta^* = \arg\min_{\theta} J(\theta) = \arg\min_{\theta} \; \mathbb{E}\left[Z\right]$$

Input: Policy to improve π_θ ;

Initialize number of samples N and learning rates $\{\nu_m\}_m$;

foreach *iteration* $m = 1, \ldots, M$ **do**

  Generate $\tau_1, \ldots, \tau_N$ trajectories from the MDP $\mathcal{M}$ under π_θ ;

  foreach *trajectory* $n = 1, \ldots, N$ **do**

    Compute $\nabla_\theta \log \pi_\theta(a_t | x_t)$ for each transition of τ_n ;

    Set $J_n \leftarrow Z(\tau_n) \sum_{t=0}^{T_n} \nabla_\theta \log \pi_\theta(a_t | x_t)$; `(Policy gradient thm)`

  Calculate $\widehat{\nabla J} \leftarrow \frac{1}{N} \sum_{n=1}^{N} J_n$; `(Sampling-based estimator)`

  Update $\theta \leftarrow \theta - \nu_m \widehat{\nabla J}$; `(Gradient descent)`

Output: $\theta \approx \theta^*$

Policy Gradient Algorithm

$$\theta^* = \arg\min_\theta J(\theta) = \arg\min_\theta \rho_{\text{CVaR}}(Z, \alpha)$$

Input: Policy to improve π_θ ;

Initialize number of samples N and learning rates $\{\nu_m\}_m$;

foreach *iteration* $m = 1, \ldots, M$ **do**

  Generate $\tau_1, \ldots, \tau_N$ trajectories from the MDP $\mathcal{M}$ under π_θ ;

  Estimate $\hat{q}_\alpha$, the quantile of $Z(\tau_1), \ldots, Z(\tau_N)$;

  foreach *trajectory* $n = 1, \ldots, N$ **do**

    Compute $\nabla_\theta \log \pi_\theta(a_t|x_t)$ for each transition of τ_n ;

    Set $J_n \leftarrow (Z(\tau_n) - \hat{q}_\alpha) \sum_{t=0}^{T_n} \nabla_\theta \log \pi_\theta(a_t|x_t)$; (Gradient formula)

  Calculate $\widehat{\nabla} J \leftarrow$ average of J_n s.t. $Z(\tau_n) > \hat{q}_\alpha$; (Estimator)

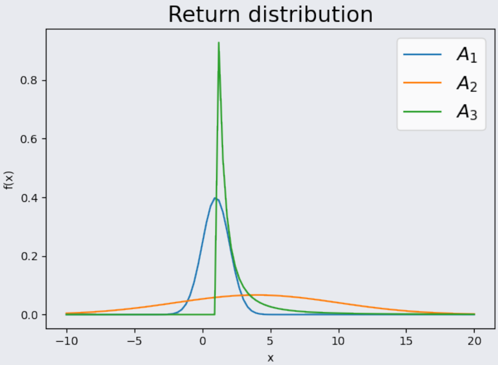
  Update $\theta \leftarrow \theta - \nu_m \widehat{\nabla} J$; (Gradient descent)

Output: $\theta \approx \theta^*$

Illustration

Three risky assets with the same initial price but different returns Z:

- A_1 – $Z \sim \mathcal{N}(\mu = 1, \sigma = 1)$
- A_2 – $Z \sim \mathcal{N}(\mu = 4, \sigma = 6)$
- A_3 – $Z \sim$ Pareto$(\alpha = 1.5)$ (i.e. $\mathbb{E}[Z] = 3$ and Var$[Z] = \infty$)



Discrete Action Space - Policies

One time step per trajectory, and the agent's policy is characterized by

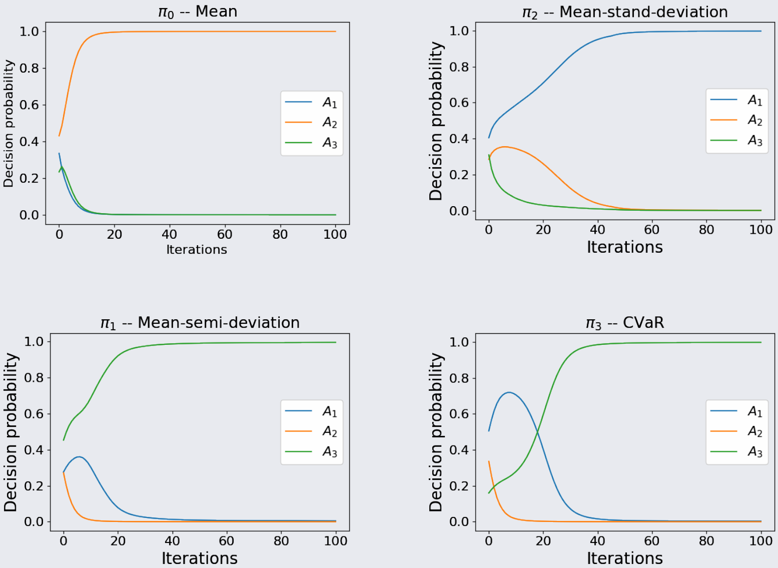
$$\pi_\theta(A_i) = \mathbb{P}\left[\text{Agent invests in } A_i\right] = \frac{e^{\theta_i}}{\sum_k e^{\theta_k}}, \quad i = 1, 2, 3.$$

Policies are trained with different objective functions, for agent's risk preferences

- Policy π^0: $\theta^* = \arg\min_\theta \mathbb{E}[Z]$
- Policy π^1: $\theta^* = \arg\min_\theta \mathbb{E}[Z] + \mathbb{SD}[Z]$
- Policy π^2: $\theta^* = \arg\min_\theta \mathbb{E}[Z] + \sqrt{\text{Var}[Z]}$
- Policy π^3: $\theta^* = \arg\min_\theta \text{CVaR}_{0.1}[Z]$

Policy π^2 was trained using the algorithm from (Tamar et al., 2012) for policy gradient with variance related risk criteria.

Discrete Action Space - Results



Discrete Action Space - Results

- π^0 favors the asset A_2
  - Expected behavior since A_2 has the highest mean return and the policy is risk-neutral, i.e. $\max_\theta \mathbb{E}[Z]$

- π^1 and π^3, which optimize coherent risk measures, favor A_3
  - Risk-averse policies choose the Pareto distributed returns, because it has a lower downside
  - Lower mean return, but less risky

- π^2 favors the asset A_1
  - Risk-averse policy that controls for the variance, **not coherent**
  - It does not choose A_3 because of the heavy upper-tail
  - Counter-intuitive since we avert high returns

Continuous Action Space - Policies

Now suppose the agent can invest a portion of its wealth in each asset, and the agent's policy is characterized by

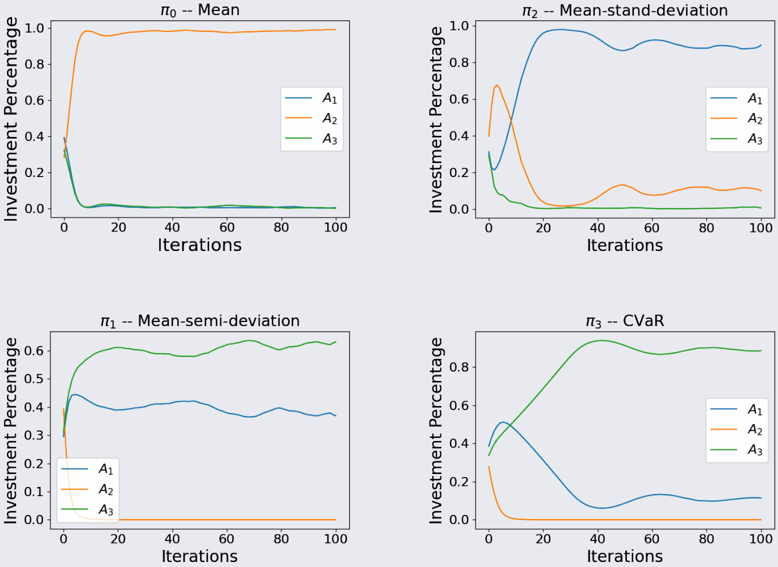
$$\pi_\theta(x) = \mathbb{P}\left[\text{Agent invests } x_i \text{ in } A_i,\ i = 1, 2, 3\right] \sim \text{Dirichlet}\left(\theta_1, \theta_2, \theta_3\right).$$

Policies are trained with different objective functions, for agent's risk preferences

- Policy π^0: $\theta^* = \arg\min_\theta \mathbb{E}[Z]$
- Policy π^1: $\theta^* = \arg\min_\theta \mathbb{E}[Z] + \mathbb{SD}[Z]$
- Policy π^2: $\theta^* = \arg\min_\theta \mathbb{E}[Z] + \sqrt{\text{Var}[Z]}$
- Policy π^3: $\theta^* = \arg\min_\theta \text{CVaR}_{0.1}[Z]$

Policy π^2 was trained using the algorithm from (Tamar et al., 2012) for policy gradient with variance related risk criteria.

Continuous Action Space - Results



Dynamic Risk Problem

Markov coherent risk measure

A **Markov coherent risk measure** (Ruszczyński, 2010) is a dynamic risk measure

$$\rho_\infty(\mathcal{M}) = C(s_0) + \gamma\rho\big(C(s_1) + \gamma\rho\big(C(s_2) + \cdots + \gamma\rho\big(C(s_T) + \cdots\big)\cdots\big)\big)$$

with a (static) coherent risk measure ρ, and a trajectory drawn from $\mathcal{M}$ under the policy π_θ.

- Dynamic risk measure: how to evaluate risk of future costs from today's perspective
- Markov risk measure: ρ is not allowed to depend on the whole past
- Time-consistent: if Z will be at least as good as W at time t_2, and they are identical between t_1 and t_2, then Z should not be worse than W at time t_1

Dynamic Risk Problem

The dynamic problem to solve is $\min_\theta \rho_\infty(\mathcal{M})$. Define the value function, the risk when starting in state s, as

$$V_\theta(s) = \rho_\infty(\mathcal{M} \mid s_0 = s).$$

Risk-sensitive Bellman equation

(Ruszczyński, 2010) With a dynamic programming decomposition, it can be shown that the value function is the unique solution to

$$V_\theta(s) = C(s) + \gamma \max_{\xi P_\theta(\cdot|s) \in \mathcal{U}(s, P_\theta(\cdot|s))} \mathbb{E}^\xi \left[V_\theta(s') \right].$$

- They extended the policy gradient theorem by developing a formula for $\nabla_\theta V_\theta(s)$
- Used to develop an actor-critic sampling-based algorithm
- Used to construct a Q-learning style algorithm for risk-aware MDPs (Huang and Haskell, 2017)

Conclusion

- Sequential decision making modeled as MDPs in order to optimize a policy that achieves good risk performance
  - Results generalized to the whole class of coherent risk measures
  - Appropriate risk measure that suits agent's risk preference

- Two policy gradient formulas and algorithms
  - Static risk problem: Sampling-based estimator
  - Dynamic risk problem: Actor-critic style algorithm

- Future directions
  - Dynamic risk problem for a finite-time horizon?
  - Multi-agent system framework for Markov coherent risk measures?
  - Extend it to a broader class of risk measures, e.g. distortion risk measures?

Acknowledgments and References

I am grateful to NSERC, University of Toronto and FRQNT for their financial support during my PhD.

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, 9(3):203–228.

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. (2015). Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28:1522–1530.

Huang, W. and Haskell, W. B. (2017). Risk-aware q-learning for markov decision processes. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 4928–4933. IEEE.

Rockafellar, R. T., Uryasev, S., et al. (2000). Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42.

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

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on stochastic programming: modeling and theory*. SIAM.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. (2015). Policy gradient for coherent risk measures. *Advances in Neural Information Processing Systems*, 28:1468–1476.

Examples

Mean-semi-deviation

Denote the semi-deviation by

$$\mathbb{SD}[Z] = \left(\mathbb{E}\left[(Z - \mathbb{E}[Z])_+^2\right]\right)^{1/2}.$$

The risk of The mean-semi-deviation is a coherent risk measure

$$\rho_{\mathbb{SD}}(Z, \alpha) = \mathbb{E}[Z] + \alpha \mathbb{SD}[Z],$$

and its gradient is given by

$$\nabla_\theta \mathbb{SD}[Z] = \frac{\mathbb{E}\left[(Z - \mathbb{E}[Z])_+ \times (\nabla_\theta \log P(\omega)(Z - \mathbb{E}[Z]) - \nabla_\theta \mathbb{E}[Z])\right]}{\mathbb{SD}(Z)}.$$

Dynamic Risk Measures

Consider a filtration $\{\mathcal{F}_t\}_t$, and denote the spaces $\mathcal{L}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ and $\mathcal{L}_{t,T} = \mathcal{L}_t \times \ldots \times \mathcal{L}_T$.

Dynamic risk measure

A **dynamic risk measure** (Ruszczyński, 2010) is a sequence $\{\rho_{t,T}\}_{t=1,\ldots,T}$, $\rho_{t,T} : \mathcal{L}_{t,T} \to \mathcal{L}_t$ where $\rho_{t,T}(Z) \leq \rho_{t,T}(W)$, $\forall\, Z \leq W$.

- How to evaluate the risk of future costs $Z_t, \ldots, Z_T$ at time t

Time-consistency

$\{\rho_{t,T}\}_t$ is said to be **time-consistent** iff. for any $1 \leq t_1 < t_2 \leq T$ and any sequence $Z, W \in \mathcal{L}_{t_1,T}$, we have

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

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

- If Z will be at least as good as W at time t_2, and they are identical between t_1 and t_2, then Z should not be worse than W at time t_1

Dynamic Risk Measures

Consider a filtration $\{\mathcal{F}_t\}_t$, and denote the spaces $\mathcal{L}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ and $\mathcal{L}_{t,T} = \mathcal{L}_t \times \ldots \times \mathcal{L}_T$.

Dynamic risk measure

A **dynamic risk measure** (Ruszczyński, 2010) is a sequence $\{\rho_{t,T}\}_{t=1,\ldots,T}$, $\rho_{t,T} : \mathcal{L}_{t,T} \to \mathcal{L}_t$ where $\rho_{t,T}(Z) \leq \rho_{t,T}(W)$, $\forall\, Z \leq W$.

- How to evaluate the risk of future costs $Z_t, \ldots, Z_T$ at time t

Time-consistency

$\{\rho_{t,T}\}_t$ is said to be **time-consistent** iff. for any $1 \leq t_1 < t_2 \leq T$ and any sequence $Z, W \in \mathcal{L}_{t_1,T}$, we have

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

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

- If Z will be at least as good as W at time t_2, and they are identical between t_1 and t_2, then Z should not be worse than W at time t_1

Dynamic Risk Measures

Recursive relationship

If $\{\rho_{t,T}\}_t$ satisfies $\rho_{t,T}(0,\ldots,0) = 0$ and

$$\rho_{t,T}(Z_t, Z_{t+1}, \ldots, Z_T) = Z_t + \rho_{t,T}(0, Z_{t+1}, \ldots, Z_T),$$

then time-consistency is equivalent to

$$\rho_{t_1,T}(Z_{t_1}, \ldots, Z_{t_2}, \ldots, Z_T) = \rho_{t_1,t_2}(Z_{t_1}, \ldots, Z_{t_2-1}, \rho_{t_2,T}(Z_{t_2}, \ldots, Z_T)).$$

We obtain the following relation

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

where $\rho_t : \mathcal{L}_{t+1} \to \mathcal{L}_t$ are one-step conditional risk measures such that

$$\rho_t(Z_{t+1}) = \rho_{t,t+1}(0, Z_{t+1}).$$

Dynamic Risk Measures

Recursive relationship

If $\{\rho_{t,T}\}_t$ satisfies $\rho_{t,T}(0,\dots,0) = 0$ and

$$\rho_{t,T}(Z_t, Z_{t+1}, \dots, Z_T) = Z_t + \rho_{t,T}(0, Z_{t+1}, \dots, Z_T),$$

then time-consistency is equivalent to

$$\rho_{t_1,T}(Z_{t_1}, \dots, Z_{t_2}, \dots, Z_T) = \rho_{t_1,t_2}(Z_{t_1}, \dots, Z_{t_2-1}, \rho_{t_2,T}(Z_{t_2}, \dots, Z_T)).$$

We obtain the following relation

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

where $\rho_t : \mathcal{L}_{t+1} \to \mathcal{L}_t$ are one-step conditional risk measures such that

$$\rho_t(Z_{t+1}) = \rho_{t,t+1}(0, Z_{t+1}).$$

Dynamic Risk Problem

Using Markov coherent risk measures, define

$$\rho_\infty(\mathcal{M}) = C(s_0) + \gamma\rho\left(C(s_1) + \gamma\rho\left(C(s_2) + \cdots + \gamma\rho\left(C(s_T) + \cdots\right)\cdots\right)\right),$$

with a (static) coherent risk measure ρ, and a trajectory drawn from $\mathcal{M}$ under the policy π_θ. The dynamic problem to solve is $\min_\theta \rho_\infty(\mathcal{M})$.

Risk-sensitive Bellman equation

With a dynamic programming decomposition, it can be shown that the value function is the unique solution to

$$V_\theta(s) = C(s) + \gamma \max_{\xi P_\theta(\cdot|s) \in \mathcal{U}(s, P_\theta(\cdot|s))} \mathbb{E}^\xi \left[V_\theta(s')\right],$$

where $V_\theta(s) = \rho_\infty(\mathcal{M} \mid s_0 = s)$.

- Used to develop an actor-critic sampling-based algorithm
- Used to construct a Q-learning style algorithm for risk-aware MDPs (Huang and Haskell, 2017)

Dynamic Risk Problem

Gradient formula (dynamic)

Let

$$L_\theta(\xi, \lambda^P, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) = \overbrace{\sum_{s' \in \mathcal{S}} \xi(s') P_\theta(s'|s) V_\theta(s')}^{\text{risk measure}} - \overbrace{\lambda^P \sum_{s' \in \mathcal{S}} \xi(s') P_\theta(s'|s) - 1}^{\text{density constr.}}$$
$$- \underbrace{\sum_{e \in \mathcal{E}} \left(\lambda^{\mathcal{E}}(e) g_e(\xi, P_\theta) \right)}_{\text{equality constr. } \mathcal{E}} - \underbrace{\sum_{i \in \mathcal{I}} \left(\lambda^{\mathcal{I}}(i) f_i(\xi, P_\theta) \right)}_{\text{inequality constr. } \mathcal{I}} .$$

For each $s \in \mathcal{S}$, we have saddle points $(\xi_s^*, \lambda_s^{*,P}, \lambda_s^{*,\mathcal{E}}, \lambda_s^{*,\mathcal{I}})$ of L_θ, and

$$\nabla_\theta V_\theta(s) = \mathbb{E}^{\xi_s^*} \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log(\pi_\theta(a_t|s_t)) h_\theta(s_t, a_t) \;\middle|\; s_0 = s \right]$$

$$h_\theta(s, a) = C(s) + \sum_{s' \in \mathcal{S}} P(s'|s, a) \xi_s^*(s') \Bigg[\gamma V_\theta(s') - \lambda_s^{*,P}$$
$$- \underbrace{\sum_{e \in \mathcal{E}} \left(\lambda_s^{*,\mathcal{E}}(e) \frac{dg_e(\xi_s^*, p)}{dp(s')} \right)}_{\text{equality constr. } \mathcal{E}} - \underbrace{\sum_{i \in \mathcal{I}} \left(\lambda_s^{*,\mathcal{I}}(i) \frac{df_i(\xi_s^*, p)}{dp(s')} \right)}_{\text{inequality constr. } \mathcal{I}} \Bigg].$$

Dynamic Risk Problem

Gradient formula (dynamic)

Let

$$L_\theta(\xi, \lambda^P, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) = \overbrace{\sum_{s' \in \mathcal{S}} \xi(s') P_\theta(s'|s) V_\theta(s')}^{\text{risk measure}} - \overbrace{\lambda^P \sum_{s' \in \mathcal{S}} \xi(s') P_\theta(s'|s) - 1}^{\text{density constr.}}$$

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

For each $s \in \mathcal{S}$, we have saddle points $(\xi_s^*, \lambda_s^{*,P}, \lambda_s^{*,\mathcal{E}}, \lambda_s^{*,\mathcal{I}})$ of L_θ, and

$$\nabla_\theta V_\theta(s) = \mathbb{E}^{\xi_s^*} \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log(\pi_\theta(a_t|s_t)) h_\theta(s_t, a_t) \,\middle|\, s_0 = s \right]$$

$$h_\theta(s, a) = C(s) + \sum_{s' \in \mathcal{S}} P(s'|s, a) \xi_s^*(s') \Bigg[\gamma V_\theta(s') - \lambda_s^{*,P}$$

$$- \underbrace{\sum_{e \in \mathcal{E}} \left(\lambda_s^{*,\mathcal{E}}(e) \frac{d g_e(\xi_s^*, p)}{d p(s')} \right)}_{\text{equality constr. } \mathcal{E}} - \underbrace{\sum_{i \in \mathcal{I}} \left(\lambda_s^{*,\mathcal{I}}(i) \frac{d f_i(\xi_s^*, p)}{d p(s')} \right)}_{\text{inequality constr. } \mathcal{I}} \Bigg].$$