Risk-averse Decision-making & Control

Mohammad Ghavamzadeh

Adobe Research

February 4, 2017

Outline

Sequential Decision-Making

Risk-Sensitive Sequential Decision-Making

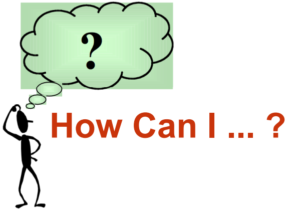
Mean-Variance Optimization

Mean-CVaR Optimization

Expected Exponential Utility

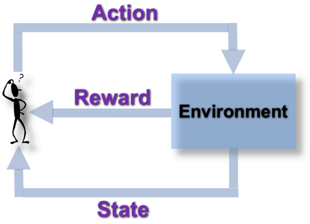
Outline

Sequential Decision-Making under Uncertainty



- Move around in the physical world ***(navigation)***
- Play and win a game
- Control the throughput of a power plant ***(process control)***
- Manage a portfolio ***(finance)***
- Medical diagnosis and treatment

Reinforcement Learning (RL)



- **RL:** A class of learning problems in which an agent interacts with a dynamic, stochastic, and incompletely known environment
- **Goal:** Learn an action-selection strategy, or ***policy***, to optimize some measure of its long-term performance
- **Interaction:** Modeled as a MDP

Markov Decision Process

MDP

- An MDP $\mathcal{M}$ is a tuple $\langle \mathcal{X}, \mathcal{A}, R, P, P_0 \rangle$.
- $\mathcal{X}$: set of states
- $\mathcal{A}$: set of actions
- $R(x,a)$: reward random variable, $r(x,a) = \mathbb{E}\big[R(x,a)\big]$
- $P(\cdot|x,a)$: transition probability distribution
- $P_0(\cdot)$: initial state distribution

- **Stationary Policy:** a distribution over actions, conditioned on the current state $\mu(\cdot|x)$

Discounted Reward MDPs

For a given policy μ

Return

$$D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu$$

Discounted Reward MDPs

For a given policy μ

Return

$$D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu$$

Risk-Neutral Objective

$$\mu^* = \arg\max_\mu \sum_{x \in \mathcal{X}} P_0(x) V^\mu(x)$$

where $V^\mu(x) = \mathbb{E}\big[D^\mu(x)\big]$.

Discounted Reward MDPs

For a given policy μ

Return

$$D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^{t} R(x_t, a_t) \mid x_0 = x,\ \mu$$

Risk-Neutral Objective ***(for simplicity)***

$$\mu^{*} = \arg\max_{\mu} V^{\mu}(x^{0})$$

x^{0} is the initial state, i.e., $P_0(x) = \delta(x - x^{0})$.

Average Reward MDPs

For a given policy μ

Average Reward

$$\rho(\mu) \;=\; \lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right]$$

Average Reward MDPs

For a given policy μ

Average Reward

$$\rho(\mu) \quad = \quad \lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] \quad = \quad \sum_{x,a}\pi^{\mu}(x,a)\,r(x,a)$$

Average Reward MDPs

For a given policy μ

Average Reward

$$\rho(\mu) \;=\; \lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] \;=\; \sum_{x,a}\pi^{\mu}(x,a)\,r(x,a)$$

$\pi^{\mu}(x,a)$: stationary dist. of state-action pair (x,a) under policy μ.

Average Reward MDPs

For a given policy μ

Average Reward

$$\rho(\mu) \;=\; \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] \;=\; \sum_{x,a} \pi^{\mu}(x,a)\, r(x,a)$$

$\pi^{\mu}(x,a)$: stationary dist. of state-action pair (x,a) under policy μ.

Risk-Neutral Objective

$$\mu^* = \arg\max_{\mu} \rho(\mu)$$

Return Random Variable

$$\overbrace{D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

Return Random Variable

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^{t} R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

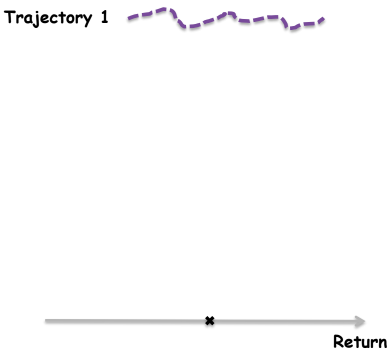
Policy μ

Return

Return Random Variable

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return}\quad\textit{random variable}}$$

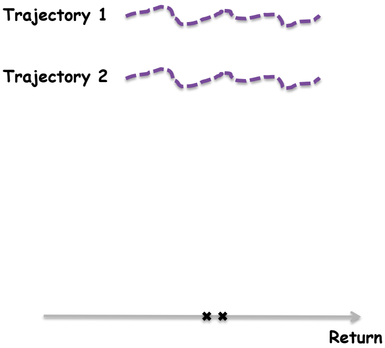
Policy μ



Return Random Variable

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return} \quad \textit{random variable}}$$

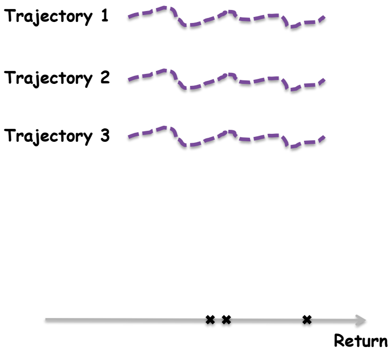
Policy μ



Return Random Variable

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

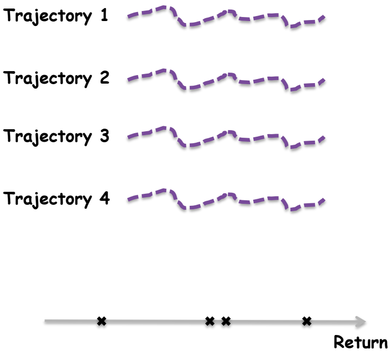
Policy μ



Return Random Variable

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return} \quad \textit{random variable}}$$

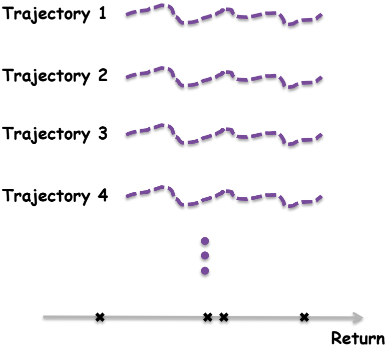
Policy μ



Return Random Variable

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^{t} R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

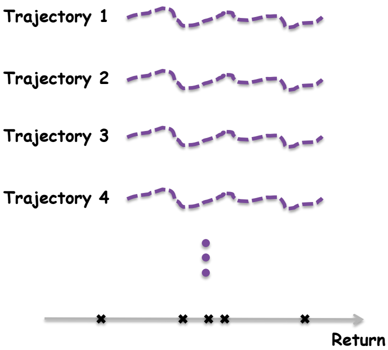
Policy μ



Return Random Variable

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return} \quad \textit{random variable}}$$

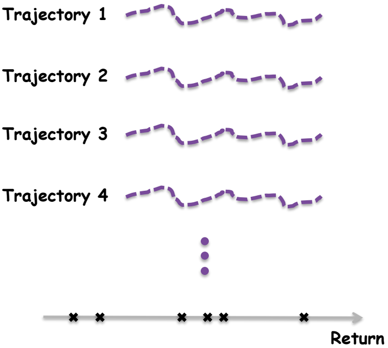
Policy μ



Return Random Variable

$$\overbrace{D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return}\quad \textit{random variable}}$$

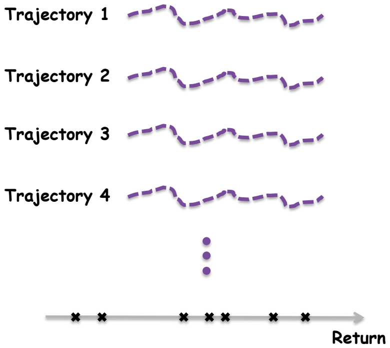
Policy μ



Return Random Variable

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

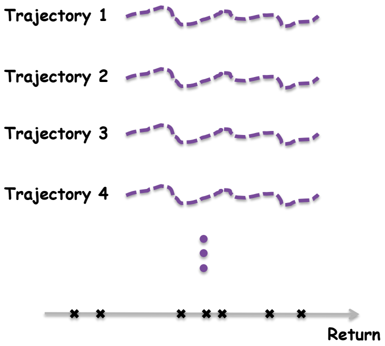
Policy μ

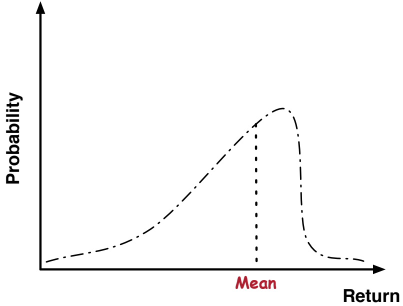


Return Random Variable

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^{t} R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

Policy μ





Outline

Risk-Sensitive Sequential Decision-Making

$$\overbrace{D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

Risk-Sensitive Sequential Decision-Making

$$\overbrace{D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return}\quad \textit{random variable}}$$

- a criterion that penalizes the ***variability*** induced by a given policy

Risk-Sensitive Sequential Decision-Making

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textbf{\textit{return random variable}}}$$

- a criterion that penalizes the ***variability*** induced by a given policy
- minimize some measure of ***risk*** as well as maximizing the usual optimization criterion

Risk-Sensitive Sequential Decision-Making

Objective: to optimize a risk-sensitive criterion such as

- expected exponential utility ***(Howard & Matheson 1972)***
- variance-related measures ***(Sobel 1982; Filar et al. 1989)***
- percentile performance ***(Filar et al. 1995)***

Risk-Sensitive Sequential Decision-Making

Objective: to optimize a risk-sensitive criterion such as

- expected exponential utility ***(Howard & Matheson 1972)***
- variance-related measures ***(Sobel 1982; Filar et al. 1989)***
- percentile performance ***(Filar et al. 1995)***

Open Question ???

construct conceptually meaningful and computationally tractable criteria

Risk-Sensitive Sequential Decision-Making

Objective: to optimize a risk-sensitive criterion such as

- expected exponential utility ***(Howard & Matheson 1972)***
- variance-related measures ***(Sobel 1982; Filar et al. 1989)***
- percentile performance ***(Filar et al. 1995)***

Open Question ???

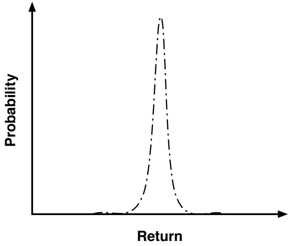
construct conceptually meaningful and computationally tractable criteria

mainly negative results

(e.g., Sobel 1982; Filar et al., 1989; Mannor & Tsitsiklis, 2011)

Risk-Sensitive Sequential Decision-Making

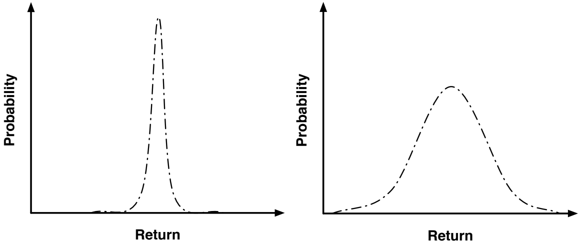
$$\overbrace{D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return}\quad \textit{random variable}}$$



$$\max_{\mu}\ \mathrm{Mean}(D^\mu)$$

Risk-Sensitive Sequential Decision-Making

$$\overbrace{D^{\mu}(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textbf{\textit{return random variable}}}$$

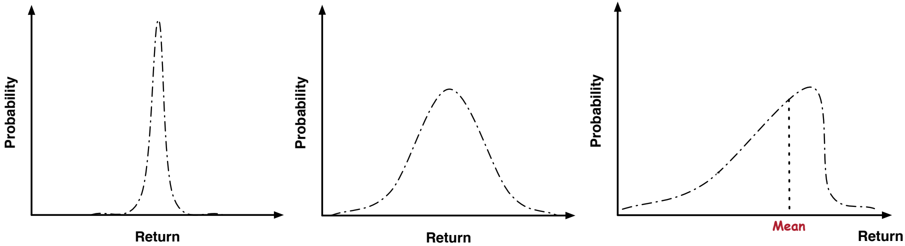


$$\max_{\mu}\ \mathrm{Mean}(D^{\mu})$$

$$\max_{\mu}\ \mathrm{Mean}(D^{\mu}) \quad \text{s.t.}\ \ \mathrm{Var}_{\alpha}(D^{\mu}) \leq \beta$$

Risk-Sensitive Sequential Decision-Making

$$\overbrace{D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

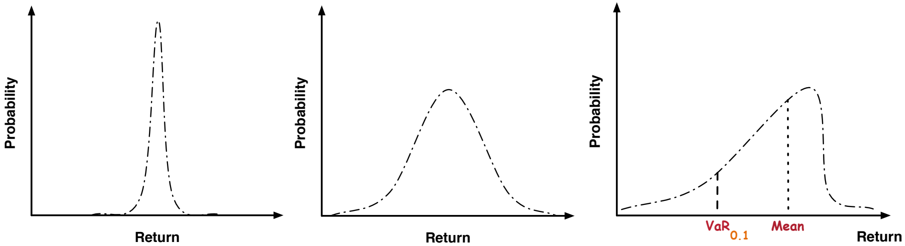


$$\max_\mu \ \mathrm{Mean}(D^\mu)$$

$$\max_\mu \ \mathrm{Mean}(D^\mu)$$
$$\text{s.t.} \ \ \mathrm{Var}_\alpha(D^\mu) \le \beta$$

Risk-Sensitive Sequential Decision-Making

$$\overbrace{D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

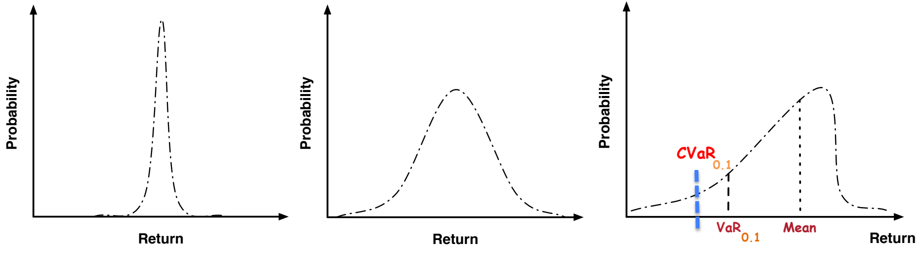


$$\max_\mu \ \text{Mean}(D^\mu)$$

$$\begin{aligned} &\max_\mu \ \text{Mean}(D^\mu) \\ &\text{s.t.} \ \text{Var}_\alpha(D^\mu) \le \beta \end{aligned}$$

Risk-Sensitive Sequential Decision-Making

$$\overbrace{D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$

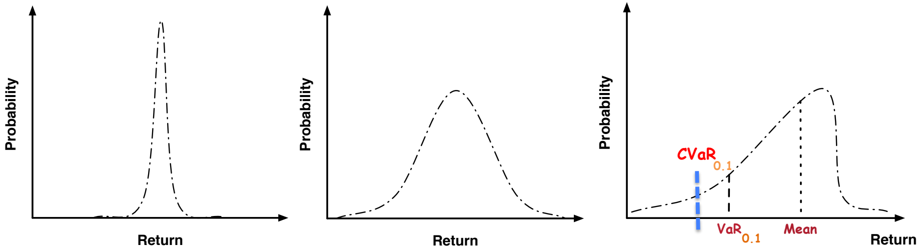


$$\max_\mu \ \mathrm{Mean}(D^\mu)$$

$$\max_\mu \ \mathrm{Mean}(D^\mu)$$
$$\text{s.t.} \ \ \mathrm{Var}_\alpha(D^\mu) \le \beta$$

Risk-Sensitive Sequential Decision-Making

$$\overbrace{D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{return random variable}}$$



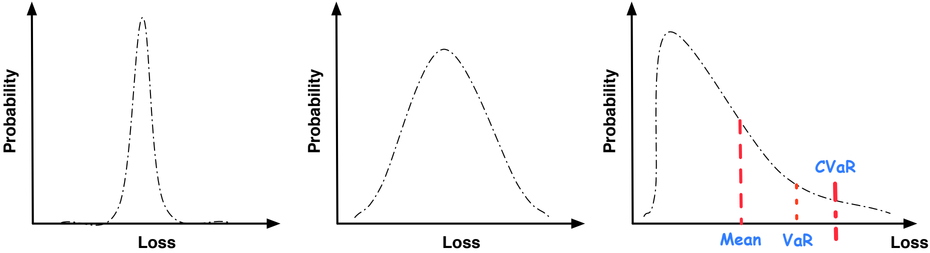
$$\max_\mu \ \mathrm{Mean}(D^\mu)$$

$$\max_\mu \ \mathrm{Mean}(D^\mu) \quad \text{s.t.} \ \mathrm{Var}_\alpha(D^\mu) \le \beta$$

$$\max_\mu \ \mathrm{Mean}(D^\mu) \quad \text{s.t.} \ \mathrm{CVaR}_\alpha(D^\mu) \ge \beta$$

Risk-Sensitive Sequential Decision-Making

$$D^\mu(x) = \overbrace{\sum_{t=0}^{\infty} \gamma^t C(x_t, a_t) \mid x_0 = x,\ \mu}^{\textit{loss}\quad \textit{random variable}}$$



$$\min_\mu \ \mathrm{Mean}(D^\mu)$$

$$\begin{aligned} \min_\mu \quad & \mathrm{Mean}(D^\mu) \\ \text{s.t.} \quad & \mathrm{Var}(D^\mu) \le \beta \end{aligned}$$

$$\begin{aligned} \min_\mu \quad & \mathrm{Mean}(D^\mu) \\ \text{s.t.} \quad & \mathrm{CVaR}_\alpha(D^\mu) \le \beta \end{aligned}$$

Risk-Sensitive Sequential Decision-Making

long history in operations research

- most work has been in the context of MDPs ***(model is known)***
- much less work in reinforcement learning (RL) framework

Risk-Sensitive RL

- expected exponential utility ***(Borkar 2001, 2002)***
- variance-related measures ***(Tamar et al., 2012, 2013; Prashanth & MGH, 2013, 2016)***
- CVaR optimization ***(Chow & MGH, 2014; Tamar et al., 2015)***
- coherent risk measures ***(Tamar, Chow, MGH, Mannor, 2015, 2017)***

Outline

Mean-Variance Optimization

$$\min_{\mu} \; \mathrm{Mean}(D^{\mu})$$
$$\text{s.t.} \; \mathrm{Var}(D^{\mu}) \leq \beta$$

Outline

Discounted Reward Setting

Discounted Reward MDPs

Return

$$D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu$$

Mean of Return ***(value function)***

$$V^\mu(x) = \mathbb{E}\big[D^\mu(x)\big]$$

Variance of Return ***(measure of variability)***

$$\Lambda^\mu(x) = \mathbb{E}\big[D^\mu(x)^2\big] - V^\mu(x)^2 = U^\mu(x) - V^\mu(x)^2$$

Discounted Reward MDPs

Risk-Sensitive Criteria

1. Maximize $V^\mu(x^0)$ s.t. $\Lambda^\mu(x^0) \leq \alpha$
2. Minimize $\Lambda^\mu(x^0)$ s.t. $V^\mu(x^0) \geq \alpha$
3. Maximize the **Sharpe Ratio**: $V^\mu(x^0)/\sqrt{\Lambda^\mu(x^0)}$
4. Maximize $V^\mu(x^0) - \alpha\Lambda^\mu(x^0)$

Policy Evaluation (Estimating Mean and Variance)

1. A. Tamar, D. Di Castro, and S. Mannor. *"Temporal Difference Methods for the Variance of the Reward To Go"*. ***ICML-2013***.

2. A. Tamar, D. Di Castro, and S. Mannor. *"Learning the Variance of the Reward-To-Go"*. ***JMLR-2016***.

Value Function

Return

$$D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu$$

Value Function ***(mean of return)*** $V^\mu : \mathcal{X} \to \mathbb{R}$

$$V^\mu(x) = \mathbb{E}\big[D^\mu(x)\big]$$

Action-value Function

Return

$$D^{\mu}(x,a) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ a_0 = a,\ \mu$$

Action-value Function ***(mean of return)*** $Q^{\mu} : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$

$$Q^{\mu}(x,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \mid X_0 = x,\ A_0 = a,\ \mu\right]$$

Bellman Equation

For a policy μ

- **Bellman Equation for Value Function**

$$V^\mu(x) = r\big(x, \mu(x)\big) + \gamma \sum_{x' \in \mathcal{X}} P\big(x'|x, \mu(x)\big) V^\mu(x')$$

- **Bellman Equation for Action-value Function**

$$\begin{aligned} Q^\mu(x,a) &= r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x,a) V^\mu(x') \\ &= r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x,a) Q^\mu\big(x', \mu(x')\big) \end{aligned}$$

Variance of Return

Variance of Return *(measure of variability)*

$$\Lambda^\mu(x) = \overbrace{\mathbb{E}\big[D^\mu(x)^2\big]}^{U^\mu(x)} - V^\mu(x)^2$$

Square Reward Value Function

$$U^\mu(x) = \mathbb{E}\big[D^\mu(x)^2\big]$$

Square Reward Action-value Function

$$W^\mu(x,a) = \mathbb{E}\big[D^\mu(x,a)^2\big]$$

Bellman Equation for Variance *(Sobel, 1982)*

For a policy μ

- **Bellman Equation for Square Reward Value Function**

$$U^\mu(x) = r\big(x,\mu(x)\big)^2 + \gamma^2 \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)U^\mu(x') + 2\gamma r\big(x,\mu(x)\big) \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)V^\mu(x')$$

- **Bellman Equation for Square Reward Action-value Function**

$$W^\mu(x,a) = r(x,a)^2 + \gamma^2 \sum_{x'\in\mathcal{X}} P(x'|x,a)U^\mu(x') + 2\gamma r(x,a) \sum_{x'\in\mathcal{X}} P(x'|x,a)V^\mu(x')$$

Dynamic Programming for Optimizing Variance *(Sobel, 1982)*

V is amenable to optimization with ***policy iteration***

$$V^{\mu_1}(x) \geq V^{\mu_2}(x),\ \forall x \in \mathcal{X} \quad \Longrightarrow \quad Q^{\mu_1}(x,a) \geq Q^{\mu_2}(x,a),\ \forall x \in \mathcal{X},\ \forall a \in \mathcal{A}$$

Λ is not amenable to optimization with ***policy iteration***

$$\Lambda^{\mu_1}(x) \geq \Lambda^{\mu_2}(x),\ \forall x \in \mathcal{X} \quad \not\Longrightarrow \quad \Lambda^{\mu_1}(x,a) \geq \Lambda^{\mu_2}(x,a),\ \forall x \in \mathcal{X},\ \forall a \in \mathcal{A}$$

Dynamic Programming for Optimizing Variance

U alone does **not** satisfy the implication

$$U^{\mu_1}(x) \geq U^{\mu_2}(x),\ \forall x \in \mathcal{X} \quad \not\Longrightarrow \quad W^{\mu_1}(x,a) \geq W^{\mu_2}(x,a),\ \forall x \in \mathcal{X},\ \forall a \in \mathcal{A}$$

but U and V together **do**

$$\left.\begin{array}{l} V^{\mu_1}(x) \geq V^{\mu_2}(x),\ \forall x \in \mathcal{X} \\ \\ U^{\mu_1}(x) \geq U^{\mu_2}(x),\ \forall x \in \mathcal{X} \end{array}\right\} \quad \Longrightarrow \quad W^{\mu_1}(x,a) \geq W^{\mu_2}(x,a),\ \forall x \in \mathcal{X},\ \forall a \in \mathcal{A}$$

Bellman Equation for Variance

Bellman equation for U^μ is linear in V^μ and U^μ

$$U^\mu(x) = r\big(x,\mu(x)\big)^2 + \gamma^2 \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)U^\mu(x') + 2\gamma r\big(x,\mu(x)\big) \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)V^\mu(x')$$

Bellman equation for Λ^μ is **not** linear in V^μ and Λ^μ

$$\Lambda^\mu(x) = U^\mu(x) - V^\mu(x)^2$$

TD Methods for Variance

$$\begin{cases} V^\mu(x) = r\big(x,\mu(x)\big) + \gamma \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)V^\mu(x') \\ \\ U^\mu(x) = r\big(x,\mu(x)\big)^2 + \gamma^2 \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)U^\mu(x') \\ \qquad + 2\gamma r\big(x,\mu(x)\big) \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)V^\mu(x') \end{cases} \tag{1}$$

- solution to **(1)** may be expressed as the fixed point of a linear mapping in the joint space V and U

TD Methods for Variance

$$
\begin{cases}
V^\mu(x) &= \overbrace{r\big(x,\mu(x)\big) + \gamma \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)V^\mu(x')}^{[\mathcal{T}^\mu Z]_V(x)} \\
U^\mu(x) &= \overbrace{r\big(x,\mu(x)\big)^2 + \gamma^2 \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)U^\mu(x')} ^{[\mathcal{T}^\mu Z]_U(x)} \\
&+ \underbrace{2\gamma r\big(x,\mu(x)\big) \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)V^\mu(x')}_{[\mathcal{T}^\mu Z]_U(x)}
\end{cases}
\tag{1}
$$

- solution to **(1)** may be expressed as the fixed point of a linear mapping in the joint space V and U

$$\mathcal{T}^\mu : \mathbb{R}^{2|\mathcal{X}|} \rightarrow \mathbb{R}^{2|\mathcal{X}|} \quad , \quad Z = (Z_V \in \mathbb{R}^{|\mathcal{X}|}, Z_U \in \mathbb{R}^{|\mathcal{X}|}) \quad , \quad \mathcal{T}^\mu Z = Z$$

TD Methods for Variance

$$
\begin{cases}
V^\mu(x) &= \overbrace{r\big(x,\mu(x)\big) + \gamma \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)V^\mu(x')}^{[\mathcal{T}^\mu Z]_V(x)} \\
U^\mu(x) &= \overbrace{r\big(x,\mu(x)\big)^2 + \gamma^2 \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)U^\mu(x')} ^{[\mathcal{T}^\mu Z]_U(x)} \\
&+ \underbrace{2\gamma r\big(x,\mu(x)\big) \sum_{x'\in\mathcal{X}} P\big(x'|x,\mu(x)\big)V^\mu(x')}_{[\mathcal{T}^\mu Z]_U(x)}
\end{cases}
\qquad (1)
$$

- projection of this mapping onto a linear feature space is contracting
  (allowing us to use TD methods)

$$
S_V = \{v^\top \phi_v(x) \mid v \in \mathbb{R}^{\kappa_2}, x\in\mathcal{X}\} \quad , \quad S_U = \{u^\top \phi_u(x) \mid u \in \mathbb{R}^{\kappa_3}, x\in\mathcal{X}\}
$$

$$
\Pi_V : \mathbb{R}^{|\mathcal{X}|} \to S_V \ , \ \Pi_U : \mathbb{R}^{|\mathcal{X}|} \to S_U \quad , \quad \Pi = \begin{pmatrix} \Pi_V & 0 \\ 0 & \Pi_U \end{pmatrix} \ , \ Z = \Pi \mathcal{T}^\mu Z
$$

TD(0) Algorithm for Variance

TD(0) for Variance *(Tamar et al., 2013)*

$$v_{t+1} = v_t + \zeta(t)\delta_t\phi_v(x_t) \qquad\qquad u_{t+1} = u_t + \zeta(t)\epsilon_t\phi_u(x_t)$$

where the TD-errors δ_t and ϵ_t are computed as

$$\delta_t = r(x_t, a_t) + \gamma v_t^\top \phi_v(x_{t+1}) - v_t^\top \phi_v(x_t)$$

$$\epsilon_t = r(x_t, a_t)^2 + 2\gamma r(x_t, a_t) v_t^\top \phi_v(x_{t+1}) + \gamma^2 u_t^\top \phi_u(x_{t+1}) - u_t^\top \phi_u(x_t)$$

Relevant Publications

1. T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. *"Parametric return density estimation for reinforcement learning"*. arXiv, 2012
2. M. Sato, H. Kimura, and S. Kobayashi. *"TD algorithm for the variance of return and mean-variance reinforcement learning"*. Transactions of the Japanese Society for Artificial Intelligence, 2001.
3. M. Sobel, *"The variance of discounted Markov decision processes"*. Applied Probability, 1982.
4. A. Tamar, D. Di Castro, and S. Mannor. *"Temporal Difference Methods for the Variance of the Reward To Go"*. ICML, 2013.
5. A. Tamar, D. Di Castro, and S. Mannor. *"Learning the Variance of the Reward-To-Go"*. JMLR, 2016.

Policy Gradient Algorithms

1. A. Tamar, D. Di Castro, and S. Mannor. *"Policy Gradients with Variance Related Risk Criteria"*. ***ICML-2012***.

Mean-Variance Optimization for Discounted MDPs

Optimization Problem

$$\max_{\mu} V^{\mu}(x^0) \quad \text{s.t.} \quad \Lambda^{\mu}(x^0) \leq \alpha$$

$$\Updownarrow$$

$$\max_{\theta} L_{\lambda}(\theta) \triangleq V^{\theta}(x^0) - \lambda \overbrace{\Gamma\big(\Lambda^{\theta}(x^0) - \alpha\big)}^{\text{penalty function}}$$

A class of parameterized stochastic policies

$$\big\{\mu(\cdot|x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R}^{\kappa_1}\big\}$$

Mean-Variance Optimization for Discounted MDPs

Optimization Problem

$$\max_{\mu} V^{\mu}(x^0) \quad \text{s.t.} \quad \Lambda^{\mu}(x^0) \leq \alpha$$

$$\Updownarrow$$

$$\max_{\theta} L_{\lambda}(\theta) \stackrel{\triangle}{=} V^{\theta}(x^0) - \lambda \overbrace{\Gamma\big(\Lambda^{\theta}(x^0) - \alpha\big)}^{\text{penalty function}}$$

A class of parameterized stochastic policies

$$\big\{\mu(\cdot|x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R}^{\kappa_1}\big\}$$

To tune θ, one needs to evaluate

$$\nabla_{\theta} L_{\lambda}(\theta) = \nabla_{\theta} V^{\theta}(x^0) - \lambda \Gamma'\big(\Lambda^{\theta}(x^0) - \alpha\big) \nabla_{\theta} \Lambda^{\theta}(x^0)$$

Computing the Gradient

Computing the Gradient $\nabla_\theta L_\lambda(\theta)$

$$\nabla_\theta V^\theta(x^0) = \mathbb{E}_\xi \left[D^\theta(\xi) \; \nabla_\theta \log \mathbb{P}(\xi|\theta) \right]$$

$$\nabla_\theta \Lambda^\theta(x^0) = \mathbb{E}_\xi \left[D^\theta(\xi)^2 \; \nabla_\theta \log \mathbb{P}(\xi|\theta) \right] - 2V^\theta(x^0) \; \nabla_\theta V^\theta(x^0)$$

A **System Trajectory** of length τ generated by policy θ:

$$\xi = \big(x_0 = x^0, a_0 \sim \mu(\cdot|x_0), x_1, a_1 \sim \mu(\cdot|x_1), \ldots, x_{\tau-1}, a_{\tau-1} \sim \mu(\cdot|x_{\tau-1})\big)$$

Computing the Gradient

Computing the Gradient $\nabla_\theta L_\lambda(\theta)$

$$\nabla_\theta V^\theta(x^0) = \mathbb{E}_\xi\left[D^\theta(\xi)\ \nabla_\theta \log \mathbb{P}(\xi|\theta)\right]$$

$$\nabla_\theta \Lambda^\theta(x^0) = \mathbb{E}_\xi\left[D^\theta(\xi)^2\ \nabla_\theta \log \mathbb{P}(\xi|\theta)\right] - 2V^\theta(x^0)\ \nabla_\theta V^\theta(x^0)$$

$$\nabla_\theta \log \mathbb{P}(\xi|\theta) = \sum_{t=0}^{\tau-1} \nabla_\theta \log \mu(a_t|x_t;\theta)$$

A **System Trajectory** of length τ generated by policy θ:

$$\xi = \big(x_0 = x^0, a_0 \sim \mu(\cdot|x_0), x_1, a_1 \sim \mu(\cdot|x_1), \dots, x_{\tau-1}, a_{\tau-1} \sim \mu(\cdot|x_{\tau-1})\big)$$

Risk-Sensitive Policy Gradient Algorithms

At each iteration k, the algorithm

- Generates a trajectory ξ_k by following the policy θ_k and
- Update the parameters as

$$\theta_{k+1} = \theta_k + \zeta_2(k)\Big(D(\xi_k) - \lambda\Gamma(\widehat{\Lambda}_k - \alpha)\big(D(\xi_k)^2 - 2\widehat{V}_k D(\xi_k)\big)\Big)\nabla_\theta \log \mathbb{P}(\xi_k|\theta_k)$$

$$\widehat{V}_{k+1} = \widehat{V}_k + \zeta_1(k)\big(D(\xi_k) - \widehat{V}_k\big)$$

$$\widehat{\Lambda}_{k+1} = \widehat{\Lambda}_k + \zeta_1(k)\big(D(\xi_k)^2 - \widehat{V}_k^2 - \widehat{\Lambda}_k\big)$$

Risk-Sensitive Policy Gradient Algorithms

At each iteration k, the algorithm

- Generates a trajectory ξ_k by following the policy θ_k and
- Update the parameters as

$$\theta_{k+1} = \theta_k + \zeta_2(k)\Big(D(\xi_k) - \lambda\Gamma(\widehat{\Lambda}_k - \alpha)\big(D(\xi_k)^2 - 2\widehat{V}_k D(\xi_k)\big)\Big)\nabla_\theta \log \mathbb{P}(\xi_k|\theta_k)$$

$$\widehat{V}_{k+1} = \widehat{V}_k + \zeta_1(k)\big(D(\xi_k) - \widehat{V}_k\big)$$

$$\widehat{\Lambda}_{k+1} = \widehat{\Lambda}_k + \zeta_1(k)\big(D(\xi_k)^2 - \widehat{V}_k^2 - \widehat{\Lambda}_k\big)$$

step-sizes $\{\zeta_2(k)\}$ and $\{\zeta_1(k)\}$ are chosen such that the policy parameter update is on the faster time-scale than the mean and variance parameters.

$$\zeta_1(k) = o\big(\zeta_2(k)\big) \qquad \text{or equivalently} \qquad \lim_{k\to\infty} \frac{\zeta_2(k)}{\zeta_1(k)} = 0$$

Risk-Sensitive Policy Gradient Algorithms *(Optimizing Sharpe Ratio)*

At each iteration k, the algorithm

- Generates a trajectory ξ_k by following the policy θ_k and
- Update the parameters as

$$\theta_{k+1} = \theta_k + \frac{\zeta_2(k)}{\sqrt{\widehat{\Lambda}_k}} \left(D(\xi_k) - \frac{\widehat{V}_k D(\xi_k)^2 - 2D(\xi_k)\widehat{V}_k^2}{2\widehat{\Lambda}_k} \right) \nabla_\theta \log \mathbb{P}(\xi_k|\theta_k)$$

$$\widehat{V}_{k+1} = \widehat{V}_k + \zeta_1(k)\big(D(\xi_k) - \widehat{V}_k\big)$$

$$\widehat{\Lambda}_{k+1} = \widehat{\Lambda}_k + \zeta_1(k)\big(D(\xi_k)^2 - \widehat{V}_k^2 - \widehat{\Lambda}_k\big)$$

Risk-Sensitive Policy Gradient Algorithms *(Optimizing Sharpe Ratio)*

At each iteration k, the algorithm

- Generates a trajectory ξ_k by following the policy θ_k and
- Update the parameters as

$$\theta_{k+1} = \theta_k + \frac{\zeta_2(k)}{\sqrt{\widehat{\Lambda}_k}} \left(D(\xi_k) - \frac{\widehat{V}_k D(\xi_k)^2 - 2D(\xi_k)\widehat{V}_k^2}{2\widehat{\Lambda}_k} \right) \nabla_\theta \log \mathbb{P}(\xi_k|\theta_k)$$

$$\widehat{V}_{k+1} = \widehat{V}_k + \zeta_1(k)\big(D(\xi_k) - \widehat{V}_k\big)$$

$$\widehat{\Lambda}_{k+1} = \widehat{\Lambda}_k + \zeta_1(k)\big(D(\xi_k)^2 - \widehat{V}_k^2 - \widehat{\Lambda}_k\big)$$

two time-scale stochastic approximation algorithm

Experimental Results

Simple Portfolio Management Problem *(Tamar et al., 2012)*

Problem Description

State: $x_t \in \mathbb{R}^{N+2}$

$x_t^{(1)} \in [0,1]$ fraction of investment in liquid assets

$x_t^{(2)}, \dots, x_t^{(N+1)} \in [0,1]$ fraction of investment in non-liquid assets with time to maturity $1, \dots, N$ time steps

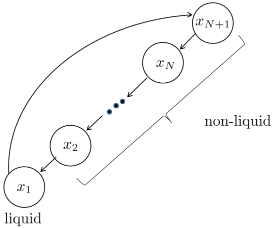
$x_t^{(N+2)}$ deviation of interest rate of non-liquid assets from its mean

Action: investing a fraction α of the total available cash in a non-liquid asset

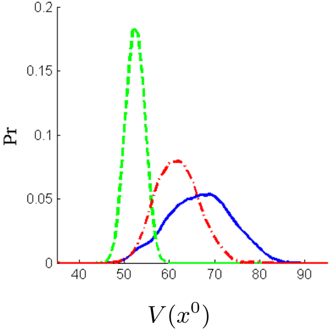
Cost: logarithm of the return from the investment

Aim: find a risk-sensitive investment strategy to mix liquid assets with fixed interest rate & risky non-liquid assets with time-variant interest rate

Results - Simple Portfolio Management Problem



Dynamics of the investment



risk neutral - *mean-var* - *Sharpe Ratio*

Summary - Risk-Sensitive Policy Gradient Algorithms

- Algorithms can be implemented as single time-scale
  (generating several trajectories from each policy & then update)

- λ is assumed to be ***fixed*** ***(selecting λ from a list)***
  (learning λ adds another time-scale to the algorithm)

- The unit of observation is a system trajectory ***(not state-action pair)***
  - algorithms are ***simple*** ***(+)***
  - better-suited to ***un-discounted*** problems ***(episodic)***
  - ***unbiased*** estimates of the gradient ***(+)***
  - ***high variance*** estimates of the gradient
    (variance grows with the length of the trajectories) ***(−)***

Actor-Critic Algorithms

1. Prashanth L. A. and **MGH**. *"Actor-Critic Algorithms for Risk-Sensitive MDPs"*. ***NIPS-2013***.

2. Prashanth L. A. and **MGH**. *"Variance-constrained Actor-Critic Algorithms for Discounted and Average Reward MDPs"*. ***MLJ-2016***.

Mean-Variance Optimization for Discounted MDPs

Optimization Problem

$$\max_{\mu} V^{\mu}(x^0) \quad \text{s.t.} \quad \Lambda^{\mu}(x^0) \leq \alpha$$

$$\Updownarrow$$

$$\max_{\lambda} \min_{\theta} L(\theta,\lambda) \overset{\triangle}{=} -V^{\theta}(x^0) + \lambda\big(\Lambda^{\theta}(x^0) - \alpha\big)$$

A class of ***parameterized stochastic policies***

$$\big\{\mu(\cdot|x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R}^{\kappa_1}\big\}$$

Mean-Variance Optimization for Discounted MDPs

Optimization Problem

$$\max_{\mu} V^{\mu}(x^0) \quad \text{s.t.} \quad \Lambda^{\mu}(x^0) \le \alpha$$

$$\Big\Updownarrow$$

$$\max_{\lambda} \min_{\theta} L(\theta, \lambda) \stackrel{\triangle}{=} -V^{\theta}(x^0) + \lambda\big(\Lambda^{\theta}(x^0) - \alpha\big)$$

A class of ***parameterized stochastic policies***

$$\big\{\mu(\cdot|x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R}^{\kappa_1}\big\}$$

One needs to evaluate $\nabla_\theta L(\theta, \lambda)$ and $\nabla_\lambda L(\theta, \lambda)$ to tune θ and λ

Mean-Variance Optimization for Discounted MDPs

Optimization Problem

$$\max_{\mu} V^{\mu}(x^0) \quad \text{s.t.} \quad \Lambda^{\mu}(x^0) \leq \alpha$$

$$\Updownarrow$$

$$\max_{\lambda} \min_{\theta} L(\theta, \lambda) \stackrel{\triangle}{=} -V^{\theta}(x^0) + \lambda\big(\Lambda^{\theta}(x^0) - \alpha\big)$$

A class of ***parameterized stochastic policies***

$$\big\{\mu(\cdot|x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R}^{\kappa_1}\big\}$$

One needs to evaluate $\nabla_\theta L(\theta, \lambda)$ and $\nabla_\lambda L(\theta, \lambda)$ to tune θ and λ

The goal is to find the ***saddle point*** of $L(\theta, \lambda)$

$$(\theta^*, \lambda^*) \quad \text{s.t} \quad L(\theta, \lambda^*) \geq L(\theta^*, \lambda^*) \geq L(\theta^*, \lambda) \quad \forall\theta, \forall\lambda > 0$$

Computing the Gradients

Computing the Gradient $\nabla_\theta L(\theta,\lambda)$

$$(1-\gamma)\nabla_\theta V^\theta(x^0) = \sum_{x,a} \pi_\gamma^\theta(x,a|x^0)\ \nabla_\theta \log\mu(a|x;\theta)\ Q^\theta(x,a)$$

$$\begin{aligned}(1-\gamma^2)\nabla_\theta U^\theta(x^0) &= \sum_{x,a} \widetilde{\pi}_\gamma^\theta(x,a|x^0)\ \nabla_\theta \log\mu(a|x;\theta)\ W^\theta(x,a) \\ &\quad + 2\gamma \sum_{x,a,x'} \widetilde{\pi}_\gamma^\theta(x,a|x^0)\ P(x'|x,a)\ r(x,a)\ \nabla_\theta V^\theta(x')\end{aligned}$$

$\pi_\gamma^\theta(x,a|x^0)$ and $\widetilde{\pi}_\gamma^\theta(x,a|x^0)$ are γ and γ^2 discounted visiting state distributions of the Markov chain under policy θ

Why Estimating the Gradient is Challenging?

Computing the Gradient $\nabla_\theta L(\theta, \lambda)$

$$(1-\gamma)\nabla_\theta V^\theta(x^0) = \sum_{x,a} \pi_\gamma^\theta(x,a|x^0)\ \nabla_\theta \log\mu(a|x;\theta)\ Q^\theta(x,a)$$

$$(1-\gamma^2)\nabla_\theta U^\theta(x^0) = \sum_{x,a} \widetilde{\pi}_\gamma^\theta(x,a|x^0)\ \nabla_\theta \log\mu(a|x;\theta)\ W^\theta(x,a) + 2\gamma \sum_{x,a,x'} \widetilde{\pi}_\gamma^\theta(x,a|x^0)\ P(x'|x,a)\ r(x,a)\ \nabla_\theta V^\theta(x')$$

$\pi_\gamma^\theta(x,a|x^0)$ and $\widetilde{\pi}_\gamma^\theta(x,a|x^0)$ are γ and γ^2 discounted visiting state distributions of the Markov chain under policy θ

Why Estimating the Gradient is Challenging?

Computing the Gradient $\nabla_\theta L(\theta, \lambda)$

$$(1-\gamma)\nabla_\theta V^\theta(x^0) = \sum_{x,a} \pi_\gamma^\theta(x,a|x^0)\ \nabla_\theta \log \mu(a|x;\theta)\ Q^\theta(x,a)$$

$$\begin{aligned}(1-\gamma^2)\nabla_\theta U^\theta(x^0) = &\sum_{x,a} \widetilde{\pi}_\gamma^\theta(x,a|x^0)\ \nabla_\theta \log \mu(a|x;\theta)\ W^\theta(x,a) \\ &+ 2\gamma \sum_{x,a,x'} \widetilde{\pi}_\gamma^\theta(x,a|x^0)\ P(x'|x,a)\ r(x,a)\ \nabla_\theta V^\theta(x')\end{aligned}$$

$\pi_\gamma^\theta(x,a|x^0)$ and $\widetilde{\pi}_\gamma^\theta(x,a|x^0)$ are γ and γ^2 discounted visiting state distributions of the Markov chain under policy θ

Simultaneous Perturbation (SP) Methods

Idea: Estimate the gradients $\nabla_\theta V^\theta(x^0)$ and $\nabla_\theta U^\theta(x^0)$ using two simulated trajectories of the system corresponding to policies with parameters θ and $\theta^+ = \theta + \beta\Delta,\ \beta > 0$.

Our actor-critic algorithms are based on two SP methods

1. Simultaneous Perturbation Stochastic Approximation (SPSA)
2. Smoothed Functional (SF)

Simultaneous Perturbation Methods

SPSA Gradient Estimate

$$\partial_{\theta^{(i)}} \widehat{V}^{\theta}(x^0) \quad \approx \quad \frac{\widehat{V}^{\theta+\beta\Delta}(x^0) - \widehat{V}^{\theta}(x^0)}{\beta\Delta^{(i)}}, \qquad i = 1, \dots, \kappa_1$$

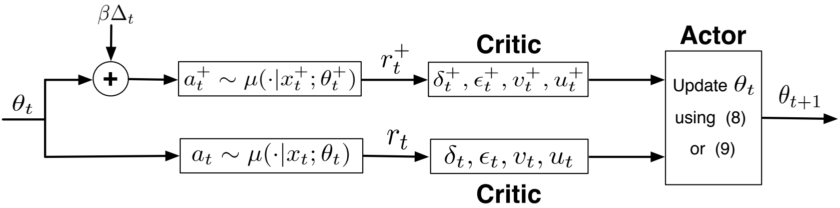
Δ is a vector of independent Rademacher random variables

SF Gradient Estimate

$$\partial_{\theta^{(i)}} \widehat{V}^{\theta}(x^0) \quad \approx \quad \frac{\Delta^{(i)}}{\beta} \left(\widehat{V}^{\theta+\beta\Delta}(x^0) - \widehat{V}^{\theta}(x^0) \right), \qquad i = 1, \dots, \kappa_1$$

Δ is a vector of independent Gaussian $\mathcal{N}(0,1)$ random variables

Mean-Variance Actor-Critic Algorithm



Trajectory 1 take action $a_t \sim \mu(\cdot|x_t;\theta_t)$, observe reward $r(x_t, a_t)$ and next state x_{t+1}

Trajectory 2 take action $a_t^+ \sim \mu(\cdot|x_t^+;\theta_t^+)$, observe reward $r(x_t^+, a_t^+)$ and next state x_{t+1}^+

Critic update the critic parameters v_t, v_t^+ for value and u_t, u_t^+ for square value functions in a TD-like fashion

Actor estimate $\nabla V^\theta(x^0)$ and $\nabla U^\theta(x^0)$ using SPSA or SF and update the policy parameter θ and the Lagrange multiplier λ

Mean-Variance Actor-Critic Algorithm

Critic Updates *(Tamar et al., 2013)*

$$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t) \qquad v^+_{t+1} = v^+_t + \zeta_3(t)\delta^+_t\phi_v(x^+_t)$$
$$u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t) \qquad u^+_{t+1} = u^+_t + \zeta_3(t)\epsilon^+_t\phi_u(x^+_t)$$

where the TD-errors $\delta_t, \delta^+_t, \epsilon_t, \epsilon^+_t$ are computed as

$$\delta_t = r(x_t, a_t) + \gamma v_t^\top\phi_v(x_{t+1}) - v_t^\top\phi_v(x_t)$$
$$\delta^+_t = r(x^+_t, a^+_t) + \gamma v_t^{+\top}\phi_v(x^+_{t+1}) - v_t^{+\top}\phi_v(x^+_t)$$
$$\epsilon_t = r(x_t, a_t)^2 + 2\gamma r(x_t, a_t)v_t^\top\phi_v(x_{t+1}) + \gamma^2 u_t^\top\phi_u(x_{t+1}) - u_t^\top\phi_u(x_t)$$
$$\epsilon^+_t = r(x^+_t, a^+_t)^2 + 2\gamma r(x^+_t, a^+_t)v_t^{+\top}\phi_v(x^+_{t+1}) + \gamma^2 u_t^{+\top}\phi_u(x^+_{t+1}) - u_t^{+\top}\phi_u(x^+_t)$$

Mean-Variance Actor-Critic Algorithm

Actor Updates

$$\theta_{t+1}^{(i)} = \Gamma_i \left[\theta_t^{(i)} + \frac{\zeta_2(t)}{\beta \Delta_t^{(i)}} \Big(\big(1 + 2\lambda_t v_t^\top \phi_v(x^0)\big)(v_t^+ - v_t)^\top \phi_v(x^0) - \lambda_t (u_t^+ - u_t)^\top \phi_u(x^0) \Big) \right]$$

$$\lambda_{t+1} = \Gamma_\lambda \left[\lambda_t + \zeta_1(t) \Big(u_t^\top \phi_u(x^0) - \big(v_t^\top \phi_v(x^0)\big)^2 - \alpha \Big) \right]$$

Mean-Variance Actor-Critic Algorithm

Actor Updates

$$\theta_{t+1}^{(i)} = \Gamma_i\left[\theta_t^{(i)} + \frac{\zeta_2(t)}{\beta\Delta_t^{(i)}}\Big(\big(1 + 2\lambda_t v_t^\top \phi_v(x^0)\big)(v_t^+ - v_t)^\top \phi_v(x^0) - \lambda_t(u_t^+ - u_t)^\top \phi_u(x^0)\Big)\right]$$

$$\lambda_{t+1} = \Gamma_\lambda\left[\lambda_t + \zeta_1(t)\Big(u_t^\top \phi_u(x^0) - \big(v_t^\top \phi_v(x^0)\big)^2 - \alpha\Big)\right]$$

step-sizes $\{\zeta_3(t)\}$, $\{\zeta_2(t)\}$, and $\{\zeta_1(t)\}$ are chosen such that the critic, policy parameter, and Lagrange multiplier updates are on the fastest, intermediate, and slowest time-scales, respectively.

Mean-Variance Actor-Critic Algorithm

Actor Updates

$$\theta_{t+1}^{(i)} = \Gamma_i \left[\theta_t^{(i)} + \frac{\zeta_2(t)}{\beta \Delta_t^{(i)}} \Big(\big(1 + 2\lambda_t v_t^\top \phi_v(x^0)\big)(v_t^+ - v_t)^\top \phi_v(x^0) - \lambda_t (u_t^+ - u_t)^\top \phi_u(x^0) \Big) \right]$$

$$\lambda_{t+1} = \Gamma_\lambda \left[\lambda_t + \zeta_1(t) \Big(u_t^\top \phi_u(x^0) - \big(v_t^\top \phi_v(x^0)\big)^2 - \alpha \Big) \right]$$

step-sizes $\{\zeta_3(t)\}$, $\{\zeta_2(t)\}$, and $\{\zeta_1(t)\}$ are chosen such that the critic, policy parameter, and Lagrange multiplier updates are on the fastest, intermediate, and slowest time-scales, respectively.

$\boxed{\textbf{three time-scale stochastic approximation algorithm}}$

Outline

Average Reward Setting

Average Reward MDPs

Average Reward

$$\rho(\mu) \;=\; \lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] \;=\; \sum_{x,a}\pi^{\mu}(x,a)\,r(x,a)$$

Long-Run Variance *(measure of variability)*

$$\Lambda(\mu) \;=\; \sum_{x,a}\pi^{\mu}(x,a)\big[r(x,a)-\rho(\mu)\big]^2 \;=\; \lim_{T\to\infty}\frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1}\big(R_t-\rho(\mu)\big)^2 \mid \mu\right]$$

The frequency of visiting state-action pairs, $\pi^{\mu}(x,a)$, determines the variability in the average reward.

Average Reward MDPs

Average Reward

$$\rho(\mu) \;=\; \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] \;=\; \sum_{x,a} \pi^{\mu}(x,a)\, r(x,a)$$

Long-Run Variance ***(measure of variability)***

$$\Lambda(\mu) \;=\; \sum_{x,a} \pi^{\mu}(x,a)\big[r(x,a) - \rho(\mu)\big]^2 \;=\; \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left[\sum_{t=0}^{T-1} \big(R_t - \rho(\mu)\big)^2 \mid \mu\right]$$

$$=\; \eta(\mu) - \rho(\mu)^2, \qquad \textbf{where} \qquad \eta(\mu) \;=\; \sum_{x,a} \pi^{\mu}(x,a)\, r(x,a)^2$$

Mean-Variance Optimization for Average Reward MDPs

Optimization Problem

$$\max_{\mu} \rho(\mu) \quad \text{s.t.} \quad \Lambda(\mu) \leq \alpha$$

$$\Updownarrow$$

$$\max_{\lambda} \; \min_{\theta} \; L(\theta, \lambda) \; \stackrel{\triangle}{=} \; -\rho(\theta) + \lambda\big(\Lambda(\theta) - \alpha\big)$$

One needs to evaluate $\nabla_\theta L(\theta, \lambda)$ and $\nabla_\lambda L(\theta, \lambda)$ to tune θ and λ

Computing the Gradients

Computing the Gradient $\nabla_\theta L(\theta, \lambda)$

$$\nabla \rho(\theta) = \sum_{x,a} \pi(x,a;\theta) \nabla \log \mu(a|x;\theta) Q(x,a;\theta)$$

$$\nabla \eta(\theta) = \sum_{x,a} \pi(x,a;\theta) \nabla \log \mu(a|x;\theta) W(x,a;\theta)$$

U^μ and W^μ are the differential value and action-value functions associated with the square reward, satisfying the following Poisson equations:

$$\eta(\mu) + U^\mu(x) = \sum_a \mu(a|x) \left[r(x,a)^2 + \sum_{x'} P(x'|x,a) U^\mu(x') \right]$$

$$\eta(\mu) + W^\mu(x,a) = r(x,a)^2 + \sum_{x'} P(x'|x,a) U^\mu(x')$$

Mean-Variance Actor-Critic Algorithm

Input: policy $\mu(\cdot|\cdot;\theta)$ and value function feature vectors $\phi_v(\cdot)$ and $\phi_u(\cdot)$
Initialization: policy parameters $\theta = \theta_0$; value function weight vectors $v = v_0$ and $u = u_0$; initial state $x_0 \sim P_0(x)$
for $t = 0, 1, 2, \dots$ **do**
Draw action $a_t \sim \mu(\cdot|x_t;\theta_t)$ and observe reward $R(x_t, a_t)$ and next state x_{t+1}

Average Updates:
$$\widehat{\rho}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\rho}_t + \zeta_4(t)R(x_t, a_t)$$
$$\widehat{\eta}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\eta}_t + \zeta_4(t)R(x_t, a_t)^2$$

TD Errors:
$$\delta_t = R(x_t, a_t) - \widehat{\rho}_{t+1} + v_t^\top \phi_v(x_{t+1}) - v_t^\top \phi_v(x_t)$$
$$\epsilon_t = R(x_t, a_t)^2 - \widehat{\eta}_{t+1} + u_t^\top \phi_u(x_{t+1}) - u_t^\top \phi_u(x_t)$$

Critic Update:
$$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t), \qquad u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t)$$

Actor Update:
$$\theta_{t+1} = \Gamma\Big(\theta_t - \zeta_2(t)\big(-\delta_t\psi_t + \lambda_t(\epsilon_t\psi_t - 2\widehat{\rho}_{t+1}\delta_t\psi_t)\big)\Big)$$
$$\lambda_{t+1} = \Gamma_\lambda\Big(\lambda_t + \zeta_1(t)(\widehat{\eta}_{t+1} - \widehat{\rho}^2_{t+1} - \alpha)\Big)$$

end for
return policy and value function parameters θ, λ, v, u

Mean-Variance Actor-Critic Algorithm

Input: policy $\mu(\cdot|\cdot;\theta)$ and value function feature vectors $\phi_v(\cdot)$ and $\phi_u(\cdot)$
Initialization: policy parameters $\theta = \theta_0$; value function weight vectors $v = v_0$ and $u = u_0$; initial state $x_0 \sim P_0(x)$
for $t = 0, 1, 2, \dots$ **do**
Draw action $a_t \sim \mu(\cdot|x_t;\theta_t)$ and observe reward $R(x_t, a_t)$ and next state x_{t+1}

Average Updates:
$$\widehat{\rho}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\rho}_t + \zeta_4(t)R(x_t, a_t)$$
$$\widehat{\eta}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\eta}_t + \zeta_4(t)R(x_t, a_t)^2$$

TD Errors:
$$\delta_t = R(x_t, a_t) - \widehat{\rho}_{t+1} + v_t^\top \phi_v(x_{t+1}) - v_t^\top \phi_v(x_t)$$
$$\epsilon_t = R(x_t, a_t)^2 - \widehat{\eta}_{t+1} + u_t^\top \phi_u(x_{t+1}) - u_t^\top \phi_u(x_t)$$

Critic Update:
$$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t), \qquad u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t)$$

Actor Update:
$$\theta_{t+1} = \Gamma\Big(\theta_t - \zeta_2(t)\big(-\delta_t\psi_t + \lambda_t(\epsilon_t\psi_t - 2\widehat{\rho}_{t+1}\delta_t\psi_t)\big)\Big)$$
$$\lambda_{t+1} = \Gamma_\lambda\Big(\lambda_t + \zeta_1(t)(\widehat{\eta}_{t+1} - \widehat{\rho}_{t+1}^2 - \alpha)\Big)$$

end for
return policy and value function parameters θ, λ, v, u

Mean-Variance Actor-Critic Algorithm

Input: policy $\mu(\cdot|\cdot;\theta)$ and value function feature vectors $\phi_v(\cdot)$ and $\phi_u(\cdot)$
Initialization: policy parameters $\theta = \theta_0$; value function weight vectors $v = v_0$ and $u = u_0$; initial state $x_0 \sim P_0(x)$
for $t = 0, 1, 2, \dots$ **do**
Draw action $a_t \sim \mu(\cdot|x_t;\theta_t)$ and observe reward $R(x_t, a_t)$ and next state x_{t+1}

Average Updates:
$$\widehat{\rho}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\rho}_t + \zeta_4(t)R(x_t, a_t)$$
$$\widehat{\eta}_{t+1} = \big(1 - \zeta_4(t)\big)\widehat{\eta}_t + \zeta_4(t)R(x_t, a_t)^2$$

TD Errors:
$$\delta_t = R(x_t, a_t) - \widehat{\rho}_{t+1} + v_t^\top \phi_v(x_{t+1}) - v_t^\top \phi_v(x_t)$$
$$\epsilon_t = R(x_t, a_t)^2 - \widehat{\eta}_{t+1} + u_t^\top \phi_u(x_{t+1}) - u_t^\top \phi_u(x_t)$$

Critic Update:
$$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t), \qquad u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t)$$

Actor Update:
$$\theta_{t+1} = \Gamma\Big(\theta_t - \zeta_2(t)\big(-\delta_t\psi_t + \lambda_t(\epsilon_t\psi_t - 2\widehat{\rho}_{t+1}\delta_t\psi_t)\big)\Big)$$
$$\lambda_{t+1} = \Gamma_\lambda\Big(\lambda_t + \zeta_1(t)(\widehat{\eta}_{t+1} - \widehat{\rho}^2_{t+1} - \alpha)\Big)$$

end for
return policy and value function parameters θ, λ, v, u

three time-scale stochastic approximation algorithm

Experimental Results

Traffic Signal Control Problem *(Prashanth & MGH, 2016)*

Problem Description

State: vector of queue lengths and elapsed times
$x_t = (q_1, \ldots, q_N, t_1, \ldots, t_N)$

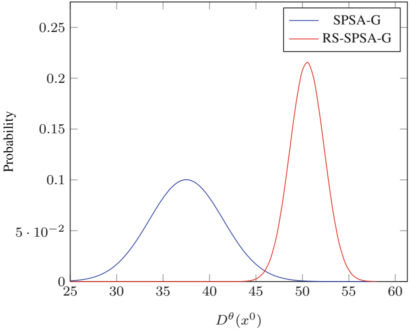
Action: feasible sign configurations

Cost:

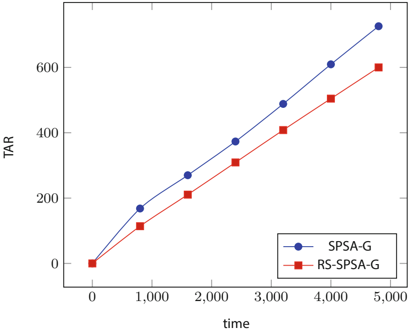
$$h(x_t) = r_1 * \Big[\sum_{i \in I_p} r_2 * q_i(t) + \sum_{i \notin I_p} s_2 * q_i(t) \Big] + s_1 * \Big[\sum_{i \in I_p} r_2 * t_i(t) + \sum_{i \notin I_p} s_2 * t_i(t) \Big]$$

Aim: find a risk-sensitive control strategy that minimizes the total delay experienced by road users, while also reducing the variations

Results - Discounted Reward Setting



Distribution of $D^\theta(x^0)$

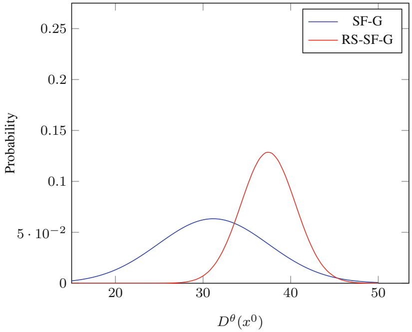


Total arrived drivers

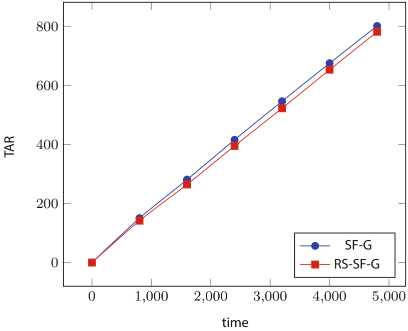
Total Arrived Drivers

Algorithm	Risk-Neutral	Risk-Sensitive
SPSA-G	754.84 ± 317.06	622.38 ± 28.36

Results - Discounted Reward Setting



Distribution of $D^\theta(x^0)$

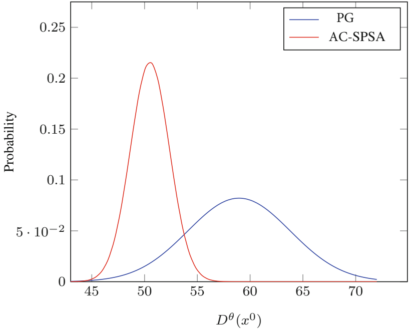


Total arrived drivers

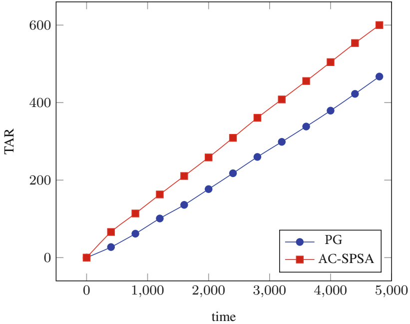
Total Arrived Drivers

Algorithm	Risk-Neutral	Risk-Sensitive
SF-G	832.34 ± 82.24	810.82 ± 36.56

Results - Actor-Critic vs. Policy Gradient

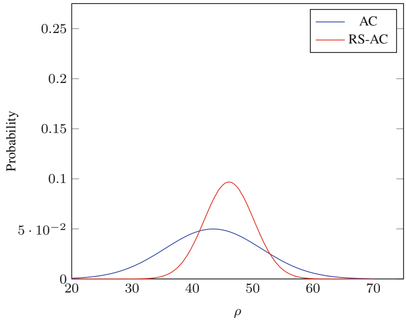


Distribution of $D^\theta(x^0)$

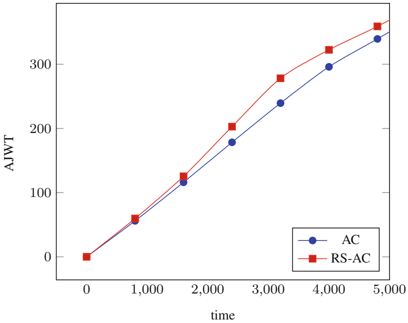


Total arrived drivers

Results - Average Reward Setting



Distribution of ρ



Average junction waiting time

Conclusions

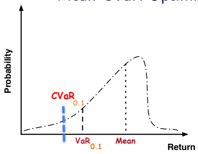
For ***discounted*** and ***average*** reward MDPs, we

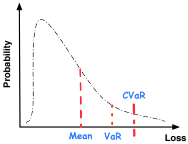
- define a set of ***(variance-related)*** ***risk-sensitive criteria***
- show how to ***estimate the gradient*** of these risk-sensitive criteria
- propose ***actor-critic algorithms*** to optimize these risk-sensitive criteria
- establish the ***asymptotic convergence*** of the algorithms
- demonstrate their usefulness in a ***traffic signal control*** problem

Relevant Publications

1. J. Filar, L. Kallenberg, and H. Lee. *"Variance-penalized Markov decision processes"*. Mathematics of OR, 1989.
2. P. Geibel and F. Wysotzki. *"Risk-sensitive reinforcement learning applied to control under constraints"*. JAIR, 2005.
3. R. Howard and J. Matheson. *"Risk-sensitive Markov decision processes"*. Management Science, 1972.
4. Prashanth L. A. and **MGH**. *"Actor-Critic Algorithms for Risk-Sensitive MDPs"*. NIPS, 2013.
5. Prashanth L. A. and **MGH**. *"Variance-constrained Actor-Critic Algorithms for Discounted and Average Reward MDPs"*. MLJ, 2016.
6. M. Sobel. *"The variance of discounted Markov decision processes"*. Applied Probability, 1982.
7. A. Tamar, D. Di Castro, and S. Mannor. *"Policy Gradients with Variance Related Risk Criteria"*. ICML, 2012.
8. A. Tamar, D. Di Castro, and S. Mannor. *"Temporal difference methods for the variance of the reward to go"*. ICML, 2013.

Outline





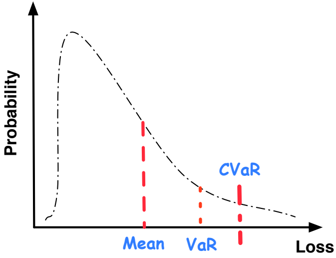
$$\max_{\mu} \ \mathrm{Mean}(D^\mu) \quad \text{s.t.} \ \mathrm{CVaR}_\alpha(D^\mu) \geq \beta$$

$$\min_{\mu} \ \mathrm{Mean}(D^\mu) \quad \text{s.t.} \ \mathrm{CVaR}_\alpha(D^\mu) \leq \beta$$

Mean-CVaR Optimization

1. Y. Chow and **MGH**. *"Algorithms for CVaR Optimization in MDPs"*. ***NIPS-2014***.
2. Y. Chow, **MGH**, L. Janson, and M. Pavone. *"Risk-Constrained Reinforcement Learning with Percentile Risk Criteria"*. ***JMLR-2017***.
3. A. Tamar, Y. Glassner, and S. Mannor. *"Optimizing the CVaR via Sampling"*. ***AAAI-2015***.

Value-at-Risk (VaR)



Cumulative Distribution $F(z) = \mathbb{P}(Z \leq z)$

Value-at-Risk at the Confidence Level $\alpha \in (0, 1)$

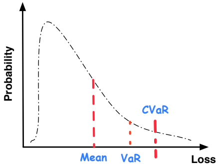
$$\mathsf{VaR}_\alpha(Z) = \min\{z \mid F(z) \geq \alpha\}$$

Properties of VaR

$$\text{VaR}_\alpha(Z) = \min\{z \mid F(z) \geq \alpha\}$$

- when F is ***continuous*** and ***strictly increasing***, $\text{VaR}_\alpha(Z)$ is the unique z satisfying $F(z) = \alpha$
- otherwise, $\text{VaR}_\alpha(Z)$ can have ***no solution*** or ***a whole range of solutions***
- often numerically unstable and difficult to work with
- is **not** a ***coherent*** risk measure
- does not quantify the losses that might be suffered beyond its value at the $(1-\alpha)$-tail of the distribution ***(Rockafellar & Uryasev, 2000)***

Conditional Value-at-Risk (CVaR)

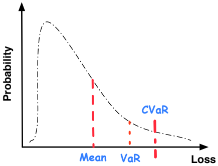


Conditional Value-at-Risk at the Confidence Level $\alpha \in (0,1)$

$$\mathsf{CVaR}_\alpha(Z) = \mathbb{E}\big[Z \mid Z \geq \mathsf{VaR}_\alpha(Z)\big]$$

coherent risk measure

Conditional Value-at-Risk (CVaR)



Conditional Value-at-Risk at the Confidence Level $\alpha \in (0,1)$

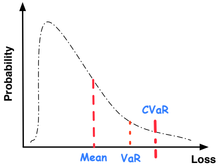
$$\mathsf{CVaR}_\alpha(Z) = \mathbb{E}\big[Z \mid Z \geq \mathsf{VaR}_\alpha(Z)\big]$$

coherent risk measure

A Different Formula for CVaR ***(Rockafellar & Uryasev, 2002)***

$$\mathsf{CVaR}_\alpha(Z) = \min_{\nu\in\mathbb{R}} \; H_\alpha(Z,\nu) \triangleq \min_{\nu\in\mathbb{R}} \; \left\{\nu + \frac{1}{1-\alpha}\mathbb{E}\big[\; \overbrace{(Z-\nu)^+}^{\max(Z-\nu,0)} \;\big]\right\}$$

Conditional Value-at-Risk (CVaR)



Conditional Value-at-Risk at the Confidence Level $\alpha \in (0,1)$

$$\text{CVaR}_\alpha(Z) = \mathbb{E}\big[Z \mid Z \geq \text{VaR}_\alpha(Z)\big]$$

coherent risk measure

A Different Formula for CVaR ***(Rockafellar & Uryasev, 2002)***

$$\text{CVaR}_\alpha(Z) = \min_{\nu \in \mathbb{R}} \; H_\alpha(Z,\nu) \triangleq \min_{\nu \in \mathbb{R}} \; \Big\{ \nu + \frac{1}{1-\alpha}\mathbb{E}\big[\, \overbrace{(Z-\nu)^+}^{\max(Z-\nu,0)} \,\big] \Big\}$$

$H_\alpha(Z,\nu)$ is finite and convex, hence continuous, as a function of ν

Mean-CVaR Optimization

Optimization Problem *(Rockafellar & Uryasev, 2000, 2002)*

$$\min_{\mu} V^{\mu}(x^0) \qquad \text{s.t.} \qquad \text{CVaR}_{\alpha}\big(D^{\mu}(x^0)\big) \leq \beta$$

Mean-CVaR Optimization

Optimization Problem *(Rockafellar & Uryasev, 2000, 2002)*

$$\min_{\mu} V^{\mu}(x^0) \qquad \text{s.t.} \qquad \text{CVaR}_{\alpha}\big(D^{\mu}(x^0)\big) \leq \beta$$

Nice Property of CVaR Optimization *(Bäuerle & Ott, 2011)*

- there exists a ***deterministic history-dependent*** optimal policy for CVaR optimization

- does not depend on the complete history, just the ***accumulated discounted cost***

at time t, only depends on x_t and $\sum_{k=0}^{t-1} \gamma^k C(x_k, a_k)$

Mean-CVaR Optimization

Optimization Problem

$$\min_{\mu} V^{\mu}(x^0) \qquad \text{s.t.} \qquad \mathrm{CVaR}_{\alpha}\big(D^{\mu}(x^0)\big) \leq \beta$$

$$\Updownarrow$$

$$\min_{\theta,\nu} V^{\theta}(x^0) \qquad \text{s.t.} \qquad H_{\alpha}\big(D^{\theta}(x^0),\nu\big) \leq \beta$$

$$\Updownarrow$$

$$\max_{\lambda\geq 0}\min_{\theta,\nu}\left(L(\theta,\nu,\lambda) \triangleq V^{\theta}(x^0) + \lambda\Big(H_{\alpha}\big(D^{\theta}(x^0),\nu\big) - \beta\Big)\right)$$

Mean-CVaR Optimization

Optimization Problem

$$\min_{\mu} V^{\mu}(x^0) \qquad \text{s.t.} \qquad \mathrm{CVaR}_{\alpha}\big(D^{\mu}(x^0)\big) \leq \beta$$

$$\Updownarrow$$

$$\min_{\theta,\nu} V^{\theta}(x^0) \qquad \text{s.t.} \qquad H_{\alpha}\big(D^{\theta}(x^0),\nu\big) \leq \beta$$

$$\Updownarrow$$

$$\max_{\lambda\geq 0}\min_{\theta,\nu}\left(L(\theta,\nu,\lambda) \stackrel{\triangle}{=} V^{\theta}(x^0) + \lambda\Big(H_{\alpha}\big(D^{\theta}(x^0),\nu\big) - \beta\Big)\right)$$

The goal is to find the ***saddle point*** of $L(\theta,\nu,\lambda)$

$(\theta^*,\nu^*,\lambda^*)$ s.t $L(\theta,\nu,\lambda^*) \geq L(\theta^*,\nu^*,\lambda^*) \geq L(\theta^*,\nu^*,\lambda) \quad \forall\theta,\nu,\forall\lambda > 0$

Computing the Gradients

Computing the Gradients $\nabla_\theta L(\theta,\nu,\lambda)$, $\partial_\nu L(\theta,\nu,\lambda)$, $\nabla_\lambda L(\theta,\nu,\lambda)$

$$\nabla_\theta L(\theta,\nu,\lambda) = \nabla_\theta V^\theta(x^0) + \frac{\lambda}{(1-\alpha)}\nabla_\theta \mathbb{E}\Big[\big(D^\theta(x^0)-\nu\big)^+\Big]$$

$$\begin{aligned}\partial_\nu L(\theta,\nu,\lambda) &= \lambda\left(1 + \frac{1}{(1-\alpha)}\partial_\nu \mathbb{E}\Big[\big(D^\theta(x^0)-\nu\big)^+\Big]\right)\\ &\ni \lambda\left(1 - \frac{1}{(1-\alpha)}\mathbb{P}\big(D^\theta(x^0)\geq \nu\big)\right)\end{aligned}$$

$$\nabla_\lambda L(\theta,\nu,\lambda) = \nu + \frac{1}{(1-\alpha)}\mathbb{E}\Big[\big(D^\theta(x^0)-\nu\big)^+\Big] - \beta$$

$\ni$ means that the term is a member of the sub-gradient set $\partial_\nu L(\theta,\nu,\lambda)$

Policy Gradient Algorithm for Mean-CVaR Optimization

Input: parameterized policy $\mu(\cdot|\cdot;\theta)$, confidence level α, loss tolerance β
Init: Policy parameter $\theta = \theta_0$, VaR parameter $\nu = \nu_0$, Lagrangian parameter $\lambda = \lambda_0$
for $i = 0, 1, 2, \ldots$ **do**
  for $j = 1, 2, \ldots$ **do**
    Generate N trajectories $\{\xi_{j,i}\}_{j=1}^N$, starting at $x_0 = x^0$ & following the policy θ_i
  end for

ν **Update:** $$\nu_{i+1} = \Gamma_\nu\Big[\nu_i - \zeta_3(i)\Big(\lambda_i - \frac{\lambda_i}{(1-\alpha)N}\sum_{j=1}^N \mathbf{1}\big\{D(\xi_{j,i}) \geq \nu_i\big\}\Big)\Big]$$

θ **Update:** $$\theta_{i+1} = \Gamma_\theta\Big[\theta_i - \zeta_2(i)\Big(\frac{1}{N}\sum_{j=1}^N \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})|_{\theta=\theta_i} D(\xi_{j,i}) + \frac{\lambda_i}{(1-\alpha)N}\sum_{j=1}^N \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})|_{\theta=\theta_i}\big(D(\xi_{j,i}) - \nu_i\big)\mathbf{1}\big\{D(\xi_{j,i}) \geq \nu_i\big\}\Big)\Big]$$

λ **Update:** $$\lambda_{i+1} = \Gamma_\lambda\Big[\lambda_i + \zeta_1(i)\Big(\nu_i - \beta + \frac{1}{(1-\alpha)N}\sum_{j=1}^N \big(D(\xi_{j,i}) - \nu_i\big)\mathbf{1}\big\{D(\xi_{j,i}) \geq \nu_i\big\}\Big)\Big]$$

end for
return parameters ν, θ, λ

$\boxed{\textbf{three time-scale stochastic approximation algorithm}}$

Main Problem of VaR and CVaR Optimization

- sampling-based approaches to quantile estimation ***(including VaR and CVaR)*** suffer from ***high variance***
- only αN among N samples are effective
  (more variance for α close to 1)
- using ***importance sampling*** for variance reduction
  (Bardou et al., 2009; Tamar et al., 2015)

ν ***Update:***

$$\nu_{i+1} = \Gamma_\nu\left[\nu_i - \zeta_3(i)\left(\lambda_i - \frac{\lambda_i}{(1-\alpha)N}\sum_{j=1}^{N}\mathbf{1}\left\{D(\xi_{j,i}) \geq \nu_i\right\}\right)\right]$$

Other Notes on Mean-CVaR Optimization Algorithm

- estimating ν is in fact estimating VaR_α
- we can also estimate ν using the empirical α-quantile

$$\widehat{\nu} = \min_z \widehat{F}(z) \geq \alpha$$

$$\widehat{F}(z) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big\{D(\xi_i) \leq z\big\}$$ ***(empirical C.D.F.)***

Actor-Critic Algorithms for Mean-CVaR Optimization

Original MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, C, P, P_0)$

Augmented MDP $\bar{\mathcal{M}} = (\bar{\mathcal{X}}, \bar{\mathcal{A}}, \bar{C}, \bar{P}, \bar{P}_0)$

$$\bar{\mathcal{X}} = \mathcal{X} \times \mathbb{R}, \qquad \bar{\mathcal{A}} = \mathcal{A}, \qquad \bar{P}_0(x, s) = P_0(x)\mathbf{1}\{s_0 = s\}$$

$$\bar{C}(x, s, a) = \begin{cases} \lambda(-s)^+/(1-\alpha) & \text{if } x = x_T, \\ C(x, a) & \text{otherwise.} \end{cases}$$

$$\bar{P}(x', s'|x, s, a) = \begin{cases} P(x'|x, a) & \text{if } s' = \big(s - C(x, a)\big)/\gamma, \\ 0 & \text{otherwise.} \end{cases}$$

x_T: a terminal state of $\mathcal{M}$

s_T: value of the s-part of the state at a terminal state x_T after T steps

$$s_T = \frac{1}{\gamma^T}\Big[\nu - \sum_{t=0}^{T-1} \gamma^t C(x_t, a_t)\Big]$$

Actor-Critic Algorithms for Mean-CVaR Optimization

$$\nabla_\theta L(\theta,\nu,\lambda) = \nabla_\theta \left(\overbrace{\mathbb{E}\big[D^\theta(x^0)\big] + \frac{\lambda}{(1-\alpha)}\mathbb{E}\Big[\big(D^\theta(x^0)-\nu\big)^+\Big]}^{V^\theta(x^0,\nu)} \right)$$

$$\nabla_\lambda L(\theta,\nu,\lambda) = \nu - \beta + \nabla_\lambda \left(\underbrace{\mathbb{E}\big[D^\theta(x^0)\big] + \frac{\lambda}{(1-\alpha)}\mathbb{E}\Big[\big(D^\theta(x^0)-\nu\big)^+\Big]}_{V^\theta(x^0,\nu)} \right)$$

$V^\theta(x^0,\nu)$: value function of policy θ at state (x^0,ν) in augmented MDP $\bar{\mathcal{M}}$

Experimental Results

American Option Pricing Problem *(Chow & MGH, 2014)*

Problem Description

State: vector of cost and time $x_t = (c_t, t)$

Action: accept the present cost ***or*** wait ***(2 actions)***

Cost:

$$c(x_t) = \begin{cases} c_t & \text{if price is accepted } \textbf{\textit{or}} \;\; t = T, \\ p_h & \text{otherwise.} \end{cases}$$

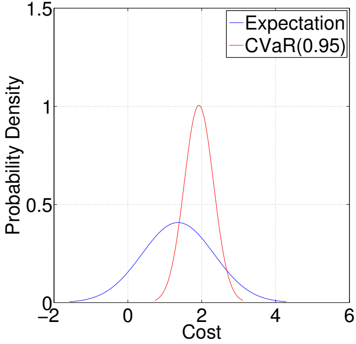
Dynamics: $x_{t+1} = (c_{t+1}, t+1)$, and

$$c_{t+1} = \begin{cases} f_u c_t & \text{w.p. } p, \\ f_d c_t & \text{w.p. } 1-p. \end{cases}$$

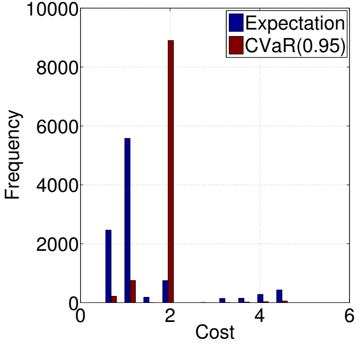
Aim: find a risk-sensitive control strategy that minimizes the total cost, while also avoiding large values of total cost

Results - American Option Pricing Problem

Policy Gradient ***mean-CVaR optimization*** $\alpha = 0.95, \quad \beta = 3$



Distribution of $D^\theta(x^0)$

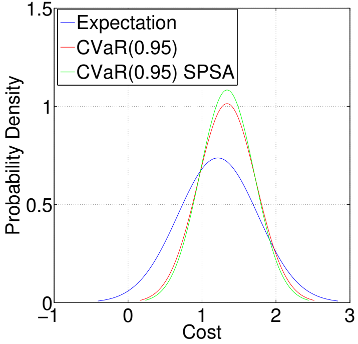


Histogram of $D^\theta(x^0)$

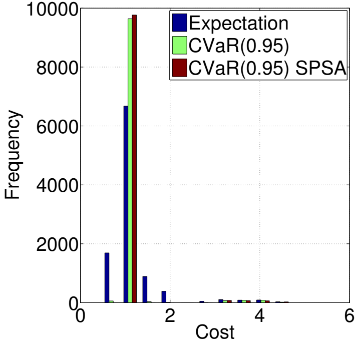
RS-PG vs. Risk-Neutral PG: slightly higher cost – significantly lower variance

Results - American Option Pricing Problem

Actor-Critic ***mean-CVaR optimization*** $\alpha = 0.95, \quad \beta = 3$



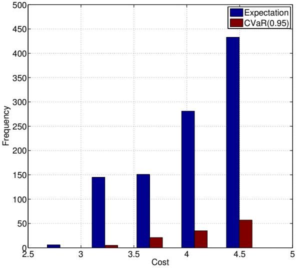
Distribution of $D^\theta(x^0)$



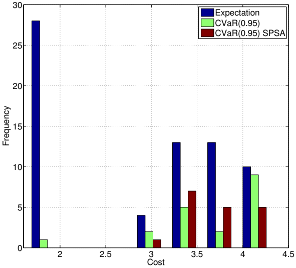
Histogram of $D^\theta(x^0)$

RS-AC vs. Risk-Neutral AC: slightly higher cost – lower variance

Results - American Option Pricing Problem



Tail of $D^\theta(x^0)$



Tail of $D^\theta(x^0)$

	$\mathbb{E}[D^\theta(x^0)]$	$\sigma[D^\theta(x^0)]$	CVaR$[D^\theta(x^0)]$
PG	1.177	1.065	4.464
PG-CVaR	1.997	0.060	2.000
AC	1.113	0.607	3.331
AC-CVaR-SPSA	1.326	0.322	2.145
AC-CVaR	1.343	0.346	2.208

Risk-Neutral PG and AC have much heavier tail than RS-PG and RS-AC

Relevant Publications

1. N. Bäuerle and J. Ott. *"Markov decision processes with average-value-at-risk criteria"*. Mathematical Methods of Operations Research, 2011.
2. K. Boda and J. Filar. *"Time consistent dynamic risk measures"*. Mathematical Methods of Operations Research, 2006.
3. V. Borkar and R. Jain. *"Risk-constrained Markov decision processes"*. IEEE Transaction on Automatic Control, 2014.
4. Y. Chow and **MGH**. *"Algorithms for CVaR Optimization in MDPs"*. NIPS, 2014.
5. Y. Chow, **MGH**, L. Janson, and M. Pavone. *"Risk-Constrained Reinforcement Learning with Percentile Risk Criteria"*. JMLR, 2017.
6. T. Morimura, M. Sugiyama, M. Kashima, H. Hachiya, and T. Tanaka. *"Non-parametric return distribution approximation for reinforcement learning"*. ICML, 2010.
7. J. Ott. *"A Markov Decision Model for a Surveillance Application and Risk-Sensitive Markov Decision Processes"*. PhD thesis, 2010.
8. M. Petrik and D. Subramanian. *"An approximate solution method for large risk-averse Markov decision processes"*. UAI, 2012.
9. R. Rockafellar and S. Uryasev. *"Conditional value-at-risk for general loss distributions"*. Journal of Banking and Finance, 2000.

Relevant Publications

10. R. Rockafellar and S. Uryasev. *"Optimization of conditional value-at-risk"*. Journal of Risk, 2002.
11. A. Tamar, Y. Glassner, and S. Mannor. *"Optimizing the CVaR via Sampling"*. AAAI, 2015.
12. A. Tamar, Y. Chow, **MGH**, and S. Mannor. *"Policy Gradient for Coherent Risk Measures"*. NIPS, 2015.
13. A. Tamar, Y. Chow, **MGH**, and S. Mannor. *"Sequential Decision Making with Coherent Risk"*. IEEE-TAC, 2017.

Outline

- Sequential Decision-Making
- Risk-Sensitive Sequential Decision-Making
- Mean-Variance Optimization
  - Discounted Reward Setting
    - Policy Evaluation (Estimating Mean and Variance)
    - Policy Gradient Algorithms
    - Actor-Critic Algorithms
  - Average Reward Setting
- Mean-CVaR Optimization
- Expected Exponential Utility

Expected Exponential Utility

Expected Exponential Loss

Objective: to find a policy μ^* such that

$$\mu^* = \arg\min_{\mu} \left(\lambda^{\mu} \triangleq \limsup_{n \to \infty} \frac{1}{\beta T} \log \mathbb{E}\left[e^{\beta \sum_{t=0}^{T-1} \gamma^t C\left(X_t, \mu(X_t)\right)} \right] \right)$$

Expected Exponential Loss

Objective: to find a policy μ^* such that

$$\mu^* = \arg\min_{\mu} \left(\lambda^{\mu} \triangleq \limsup_{n\to\infty} \frac{1}{\beta T} \log \mathbb{E}\left[e^{\beta \sum_{t=0}^{T-1} \gamma^t C\left(X_t, \mu(X_t)\right)} \right] \right)$$

Similarity to ***Mean-Variance*** Optimization

$$\frac{1}{\beta T} \log \mathbb{E}\left[e^{\beta \sum_{t=0}^{T-1} \gamma^t C\left(X_t, \mu(X_t)\right)} \right] \approx \mathbb{E}\big[D^{\mu}(x^0)\big] + \frac{\beta}{2}\mathbf{Var}\big[D^{\mu}(x^0)\big] + O(\beta^2)$$

Expected Exponential Loss

Objective: to find a policy μ^* such that

$$\mu^* = \arg\min_{\mu} \left(\lambda^{\mu} \triangleq \limsup_{n\to\infty} \frac{1}{\beta T} \log \mathbb{E}\left[e^{\beta \sum_{t=0}^{T-1} \gamma^t C\big(X_t, \mu(X_t)\big)} \right] \right)$$

Similarity to ***Mean-Variance*** Optimization

$$\frac{1}{\beta T} \log \mathbb{E}\left[e^{\beta \sum_{t=0}^{T-1} \gamma^t C\big(X_t, \mu(X_t)\big)} \right] \approx \mathbb{E}\big[D^{\mu}(x^0)\big] + \frac{\beta}{2} \mathbf{Var}\big[D^{\mu}(x^0)\big] + O(\beta^2)$$

How to choose the mean-variance tradeoff β ???

Expected Exponential Loss

Objective: to find a policy μ^* such that

$$\mu^* = \arg\min_{\mu} \left(\lambda^{\mu} \triangleq \limsup_{n\to\infty} \frac{1}{T} \log \mathbb{E}\left[e^{\sum_{t=0}^{T-1} C\left(X_t,\mu(X_t)\right)} \right] \right)$$

DP Equation: is ***non-linear eigenvalue*** problem

$$\lambda^* V^*(x) = \min_{a\in\mathcal{A}} \left(e^{C(x,a)} \sum_{x'\in\mathcal{X}} P(x'|x,a) V^*(x') \right), \quad \forall x \in \mathcal{X} \qquad \textit{\textbf{(deterministic)}}$$

$$V^*(x) = \min_{\mu} \left(\sum_{a\in\mathcal{A}} \mu(a|x) \frac{e^{C(x,a)}}{\lambda^*} \sum_{x'\in\mathcal{X}} P(x'|x,a) V^*(x') \right), \quad \forall x \in \mathcal{X} \qquad \textit{\textbf{(stochastic)}}$$

Value Iteration for Expected Exponential Loss

- Fix $x^0 \in \mathcal{X}$ and pick an arbitrary initial guess V_0

- At each iteration k, for all $x \in \mathcal{X}$, do

$$\widetilde{V}_{k+1}(x) = \min_{a \in \mathcal{A}} \left(e^{C(x,a)} \sum_{x' \in \mathcal{X}} P(x'|x,a) V_k(x') \right)$$

$$V_{k+1}(x) = \frac{\widetilde{V}_{k+1}(x)}{\widetilde{V}_{k+1}(x^0)}$$

- converges to V^* with $\lambda^* = V^*(x^0)$

Policy Iteration for Expected Exponential Loss

- Pick an arbitrary initial guess μ_0

- At each iteration k, solve the ***principle eigenvalue*** problem ***(policy evaluation)***

$$\lambda_k V_k(x) = e^{C\big(x,\mu_k(x)\big)} \sum_{x' \in \mathcal{X}} P\big(x'|x,\mu_k(x)\big) V_k(x'), \quad \forall x \in \mathcal{X}, \quad \text{with } V_k(x^0) = 1$$

- For all $x \in \mathcal{X}$, set ***(policy improvement - greedification)***

$$\mu_{k+1}(x) \in \arg\min_{a \in \mathcal{A}} \left(e^{C(x,a)} \sum_{x' \in \mathcal{X}} P(x'|x,a) V_k(x') \right)$$

- (V_k, λ_k) converges to (V^*, λ^*) with $V^*(x^0) = 1$

Q-Learning for Expected Exponential Loss

Action-value Function

$$Q^\mu(x,a) = \frac{e^{C(x,a)}}{\lambda^\mu} \sum_{x' \in \mathcal{X}} P(x'|x,a) V^\mu(x')$$

DP Equation

$$Q^*(x,a) = \frac{e^{C(x,a)}}{\lambda^*} \sum_{x' \in \mathcal{X}} P(x'|x,a) \min_{a' \in \mathcal{A}} Q^*(x',a')$$

Q-value Iteration

$(\forall x \in \mathcal{X},\ \forall a \in \mathcal{A}$, fix $x^0 \in \mathcal{X},\ a^0 \in \mathcal{A})$

$$\widetilde{Q}_{k+1}(x,a) = e^{C(x,a)} \sum_{x' \in \mathcal{X}} P(x'|x,a) \min_{a' \in \mathcal{A}} Q_k(x',a'), \quad Q_{k+1}(x,a) = \frac{\widetilde{Q}_{k+1}(x,a)}{\widetilde{Q}_{k+1}(x^0,a^0)}$$

Q-Learning

$$Q_{k+1}(x,a) = Q_k(x,a) + \zeta(k) \left(\frac{e^{C(x,a)}}{Q_k(x^0,a^0)} \min_{a' \in \mathcal{A}} Q_k(x',a') - Q_k(x,a) \right)$$

Actor-Critic for Expected Exponential Loss

DP Eq. for Policy θ

$$V^\theta(x) = \sum_{a\in\mathcal{A}} \mu(a|x;\theta)\frac{e^{C(x,a)}}{\lambda^\theta}\sum_{x'\in\mathcal{X}} P(x'|x,a)V^\theta(x')$$

Markov Chain Induced by Policy θ

$$P^\theta(x'|x) = \frac{\sum_{a\in\mathcal{A}} \mu(a|x;\theta)e^{C(x,a)}P(x'|x,a)V^\theta(x')}{\lambda^\theta V^\theta(x)}$$

with stationary distributions $d^\theta(x)$ and $\pi^\theta(x,a) = d^\theta(x)\mu(a|x;\theta)$

Actor-Critic for Expected Exponential Loss

Gradient of the Performance Measure

$$\nabla_\theta \log(\lambda^\theta) = \frac{\nabla_\theta \lambda^\theta}{\lambda^\theta} = \sum_{x,a} \pi^\theta(x,a) \nabla_\theta \mu(a|x;\theta) q^\theta(x,a)$$

$$= \sum_{x,a \neq a^0} \pi^\theta(x,a) \nabla_\theta \mu(a|x;\theta) \big[q^\theta(x,a) - q^\theta(x^0,a^0) \big]$$

where

$$q^\theta(x,a) = \frac{e^{C(x,a)}}{V^\theta(x)\lambda^\theta} \sum_{x' \in \mathcal{X}} P(x'|x,a) V^\theta(x')$$

Actor-Critic for Expected Exponential Loss

Critic Update

$$q(x_t, a_t) = q(x_t, a_t) + \zeta_2(t) \left(\frac{e^{C(x_t, a_t)} q(x_{t+1}, a_{t+1})}{q(x^0, a^0)} - q(x_t, a_t) \right)$$

Actor Update

$$\theta_{t+1} = \theta_t - \zeta_1(t) \nabla_\theta \mu(a_t | x_t; \theta) \big[q^\theta(x_t, a_t) - q^\theta(x^0, a^0) \big]$$

Two Time-Scale Stochastic Approximation

$$\zeta_1(t) = o\big(\zeta_2(t)\big) \qquad , \qquad \lim_{t \to \infty} \frac{\zeta_1(t)}{\zeta_2(t)} = 0$$

Relevant Publications

1. V. Borkar. *"A sensitivity formula for the risk-sensitive cost and the actor-critic algorithm"*. Systems & Control Letters, 2001.
2. V. Borkar. *"Q-learning for risk-sensitive control"*. Mathematics of Operations Research, 2002.
3. V. Borkar and S. Meyn. *"Risk-sensitive optimal control for Markov decision processes with monotone cost"*. Mathematics of Operations Research, 2002.

Thank you!!

Mohammad Ghavamzadeh

ghavamza@adobe.com OR
mohammad.ghavamzadeh@inria.fr