Algorithms for Risk-Sensitive Reinforcement Learning

Prashanth L.A.

INRIA Lille – Team SequeL

joint work with **Mohammad Ghavamzadeh**

Motivation

Risk is like fire: If controlled it will help you; if uncontrolled it will rise up and destroy you.

Theodore Roosevelt

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

Douglas Adams

Motivation

Risk is like fire: If controlled it will help you; if uncontrolled it will rise up and destroy you.

Theodore Roosevelt

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

Douglas Adams

Risk-Sensitive Sequential Decision-Making

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

Return r.v. Reward Policy

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

Risk-Sensitive Sequential Decision-Making

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

Return r.v.   Reward   Policy

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

Risk-Sensitive Sequential Decision-Making

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

Return r.v.

Reward

Policy

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

- minimize some measure of ***risk*** as well as maximizing a usual optimization criterion

Risk-Sensitive Sequential Decision-Making

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

Return r.v. Reward Policy

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

Risk-Sensitive Sequential Decision-Making

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

Return r.v. Reward Policy

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

- minimize some measure of ***risk*** as well as maximizing a usual optimization criterion

Risk-Sensitive Sequential Decision-Making

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

Return r.v. Reward Policy

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

- minimize some measure of ***risk*** as well as maximizing a usual optimization criterion

Risk-Sensitive Sequential Decision-Making

$$\underbrace{D^{\mu}(x)}_{\text{Return r.v.}} = \sum_{t=0}^{\infty} \gamma^t \underbrace{R(x_t, a_t)}_{\text{Reward}} \mid x_0 = x, \underbrace{\mu}_{\text{Policy}}$$

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

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

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

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)

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)*

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

Variance of Return *(measure of variability)*

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

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)***

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

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

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

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 $V^\mu(x^0) - \alpha\Lambda^\mu(x^0)$

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 $V^{\mu}(x^0) - \alpha\Lambda^{\mu}(x^0)$

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 $V^{\mu}(x^0) - \alpha\Lambda^{\mu}(x^0)$

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 $V^\mu(x^0) - \alpha\Lambda^\mu(x^0)$

Risk-Sensitive Discounted MDPs

A class of parameterized stochastic policies

$$\left\{\mu(\cdot|x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \Re^{\kappa_1}\right\}$$

Optimization Problem

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

$$\Updownarrow$$

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

Risk-Sensitive Discounted MDPs

A class of parameterized stochastic policies

$$\left\{\mu(\cdot|x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta \subseteq \Re^{\kappa_1}\right\}$$

Optimization Problem

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

$$\Updownarrow$$

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

Solving the risk-sensitive MDP

Three-Stage Solution:

inner-most stage Simulate the MDP and estimate $V^\mu(x^0)$ and $\Lambda^\mu(x^0)$ using a TD-critic;

next outer stage Estimate $\nabla_\theta L(\theta, \lambda)$ using TD critic and then update θ along descent direction; and

outer-most stage update the Lagrange multipliers λ using the variance constraint ($\nabla_\lambda L(\theta, \lambda) = \Lambda^\theta(x^0) - \alpha$).

Using multi-timescale stochastic approximation all three stages happen simultaneously with varying step-sizes

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

Solving the risk-sensitive MDP

Three-Stage Solution:

inner-most stage Simulate the MDP and estimate $V^{\mu}(x^0)$ and $\Lambda^{\mu}(x^0)$ using a TD-critic;

next outer stage Estimate $\nabla_{\theta} L(\theta, \lambda)$ using TD critic and then update θ along descent direction; and

outer-most stage update the Lagrange multipliers λ using the variance constraint $(\nabla_{\lambda} L(\theta, \lambda) = \Lambda^{\theta}(x^0) - \alpha)$.

Using multi-timescale stochastic approximation all three stages happen simultaneously with varying step-sizes

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

Solving the risk-sensitive MDP

Three-Stage Solution:

inner-most stage Simulate the MDP and estimate $V^{\mu}(x^0)$ and $\Lambda^{\mu}(x^0)$ using a TD-critic;

next outer stage Estimate $\nabla_{\theta} L(\theta, \lambda)$ using TD critic and then update θ along descent direction; and

outer-most stage update the Lagrange multipliers λ using the variance constraint $(\nabla_{\lambda} L(\theta, \lambda) = \Lambda^{\theta}(x^0) - \alpha)$.

Using multi-timescale stochastic approximation all three stages happen simultaneously with varying step-sizes

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

Solving the risk-sensitive MDP

Three-Stage Solution:

inner-most stage Simulate the MDP and estimate $V^{\mu}(x^0)$ and $\Lambda^{\mu}(x^0)$ using a TD-critic;

next outer stage Estimate $\nabla_{\theta} L(\theta, \lambda)$ using TD critic and then update θ along descent direction; and

outer-most stage update the Lagrange multipliers λ using the variance constraint $(\nabla_{\lambda} L(\theta, \lambda) = \Lambda^{\theta}(x^0) - \alpha)$.

Using multi-timescale stochastic approximation all three stages happen simultaneously with varying step-sizes

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

Computing the Gradients

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

Why Estimating the Gradient is Challenging?

The Gradient $\nabla_\theta L(\theta, \lambda)$

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

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

$\pi_\gamma^\theta(x, a|x^0)$ 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 Simultaneous Perturbation?

Challenge: estimating $\nabla_\theta L(\theta, \lambda)$

- two different sampling distributions (π^θ_γ and $\widetilde{\pi}^\theta_\gamma$) used for $\nabla V^\theta(x^0)$ and $\nabla U^\theta(x^0)$
- $\nabla V^\theta(x')$ appears in the second sum of $\nabla U^\theta(x^0)$ equation

Solution: use SPSA

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

Δ is a vector of independent Rademacher random variables

Why Simultaneous Perturbation?

Challenge: estimating $\nabla_\theta L(\theta, \lambda)$

- two different sampling distributions (π^θ_γ and $\widetilde{\pi}^\theta_\gamma$) used for $\nabla V^\theta(x^0)$ and $\nabla U^\theta(x^0)$
- $\nabla V^\theta(x')$ appears in the second sum of $\nabla U^\theta(x^0)$ equation

Solution: use SPSA

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

Δ is a vector of independent Rademacher random variables

SPSA idea

Scalar θ:

$$\frac{dV(\theta)}{d\theta} = \lim_{\beta \to 0} \left(\frac{V(\theta+\beta) - V(\theta)}{\beta} \right).$$

Using a Taylor expansion of $V(\theta)$ around θ, we obtain:

$$V(\theta+\beta) = V(\theta) + \beta \frac{dV(\theta)}{d\theta} + \frac{\beta^2}{2} \frac{d^2 V(\theta)}{d\theta^2} + o(\beta^2),$$

$$\text{Thus, } \frac{V(\theta+\beta) - V(\theta)}{\beta} = \frac{dV(\theta)}{d\theta} + o(\beta).$$

Vector $\theta \in \mathbb{R}^{\kappa_1}$:

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

where Δ is a vector of independent Rademacher random variables.

SPSA idea

Scalar θ:

$$\frac{dV(\theta)}{d\theta} = \lim_{\beta\to 0}\left(\frac{V(\theta+\beta)-V(\theta)}{\beta}\right).$$

Using a Taylor expansion of $V(\theta)$ around θ, we obtain:

$$V(\theta+\beta) = V(\theta) + \beta\frac{dV(\theta)}{d\theta} + \frac{\beta^2}{2}\frac{d^2V(\theta)}{d\theta^2} + o(\beta^2),$$

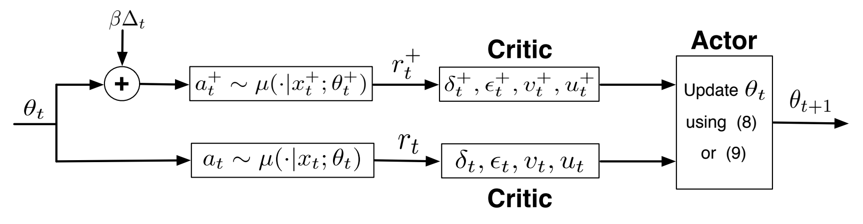
$$\text{Thus, } \frac{V(\theta+\beta)-V(\theta)}{\beta} = \frac{dV(\theta)}{d\theta} + o(\beta).$$

Vector $\theta \in \mathbb{R}^{\kappa_1}$:

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

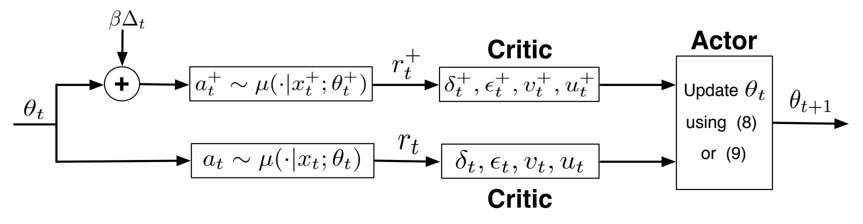
where Δ is a vector of independent Rademacher random variables.

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 corresponding to policies with parameters θ and $\theta^+ = \theta + \beta\Delta,\ \beta > 0$.

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 corresponding to policies with parameters θ and $\theta^+ = \theta + \beta\Delta,\ \beta > 0$.

Critic Update
Approximation

$$\widehat{V}(x) \approx v^\top \phi_v(x) \text{ and } \widehat{U}(x) \approx u^\top \phi_u(x)$$

Update rule

	Trajectory 1	Trajectory 2
Value	$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t)$	$v^+_{t+1} = v^+_t + \zeta_3(t)\delta^+_t\phi_v(x^+_t)$
Square-Value	$u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t)$	$u^+_{t+1} = u^+_t + \zeta_3(t)\epsilon^+_t\phi_u(x^+_t)$

$\delta_t, \delta^+_t, \epsilon_t, \epsilon^+_t$ denote the TD-errors.

Tamar et al (2013) Temporal difference methods for the variance of the reward to go. In: ICML

Critic Update

Approximation

$\widehat{V}(x) \approx v^\top \phi_v(x)$ and $\widehat{U}(x) \approx u^\top \phi_u(x)$

Update rule

	Trajectory 1	Trajectory 2
Value	$v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t)$	$v^+_{t+1} = v^+_t + \zeta_3(t)\delta^+_t\phi_v(x^+_t)$
Square-Value	$u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t)$	$u^+_{t+1} = u^+_t + \zeta_3(t)\epsilon^+_t\phi_u(x^+_t)$

$\delta_t, \delta^+_t, \epsilon_t, \epsilon^+_t$ denote the TD-errors.

Tamar et al (2013) Temporal difference methods for the variance of the reward to go. In: ICML

Critic Update (contd)

TD-errors δ_t, ϵ_t in Trajectory 1 (policy θ)

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

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

TD-errors δ_t^+, ϵ_t^+ in Trajectory 2 (perturbed policy $\theta + \beta\Delta$)

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

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

Critic Update (contd)

TD-errors δ_t, ϵ_t in Trajectory 1 (policy θ)

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

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

TD-errors δ_t^+, ϵ_t^+ in Trajectory 2 (perturbed policy $\theta + \beta\Delta$)

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

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

Actor Update

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

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

Step-sizes $\{\zeta_3(t)\}$, $\{\zeta_2(t)\}$, and $\{\zeta_1(t)\}$ chosen s.t.

- Critic is on the fastest time-scale,
- Policy parameter update is on the intermediate, and
- Lagrange multiplier update is on the slowest time-scale

Actor Update

$$\theta_{t+1}^{(i)} = \Gamma_i\left[\theta_t^{(i)} + \zeta_2(t)\Bigg(\underbrace{\frac{(1+2\lambda_t v_t^\top\phi_v(x^0))(v_t^+ - v_t)^\top\phi_v(x^0) - \lambda_t(u_t^+ - u_t)^\top\phi_u(x^0)}{\beta\Delta_t^{(i)}}}_{\nabla_\theta L(\theta,\lambda)}\Bigg)\right]$$

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

Step-sizes $\{\zeta_3(t)\}$, $\{\zeta_2(t)\}$, and $\{\zeta_1(t)\}$ chosen s.t.

- Critic is on the fastest time-scale,
- Policy parameter update is on the intermediate, and
- Lagrange multiplier update is on the slowest time-scale

Average Reward Setting

Notation

Average Reward

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

Variance

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

Stream of rewards: (0,0,0,0,. . .) or (100,-100,100,-100,. . .)

The long-term frequency of occurrence of state-action pairs determines the variability in the average reward

Filar et al(1989) Variance-penalized Markov decision processes. Mathematics of Operations Research 14(1):147–161

Notation

Average Reward

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

Variance

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

Stream of rewards: (0,0,0,0,. . .) or (100,-100,100,-100,. . .)

The long-term frequency of occurrence of state-action pairs determines the variability in the average reward

Filar et al(1989) Variance-penalized Markov decision processes. Mathematics of Operations Research 14(1):147–161

Notation

Average Reward

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

Variance

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

Stream of rewards: (0,0,0,0,...) or (100,-100,100,-100,...)

The long-term frequency of occurrence of state-action pairs determines the variability in the average reward

Filar et al(1989) Variance-penalized Markov decision processes. Mathematics of Operations Research 14(1):147–161

Notation

Average Reward

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

Variance

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

Stream of rewards: (0,0,0,0,. . .) or (100,-100,100,-100,. . .)

The long-term frequency of occurrence of state-action pairs determines the variability in the average reward

Filar et al(1989) Variance-penalized Markov decision processes. Mathematics of Operations Research 14(1):147–161

Risk-sensitive MDP

Objective

$$\max_{\theta} \rho(\theta) \qquad \text{subject to} \qquad \Lambda(\theta) \leq \alpha$$

$$\Updownarrow$$

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

As before, one needs $\nabla_\theta L(\theta, \lambda)$ to tune policy parameter θ

Risk-sensitive MDP

Objective

$$\max_{\theta} \rho(\theta) \qquad \text{subject to} \qquad \Lambda(\theta) \le \alpha$$

$$\Updownarrow$$

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

As before, one needs $\nabla_\theta L(\theta, \lambda)$ to tune policy parameter θ

Notation (again)

Average Reward

$$\rho(\mu) \;=\; \sum_{x,a} d^{\mu}(x)\mu(a|x)r(x,a) \;=\; \sum_{x,a} \pi^{\mu}(x,a)r(x,a),$$

Variance

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

Computing the gradients

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

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

U^{μ} and W^{μ} are the differential value and action-value functions that satisfy

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

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

Bhatnagar et al (2009) Natural actor-critic algorithms. In: Automatica

RS-AC algorithm

Initialization: policy parameters θ_0; value function weights v_0, u_0; initial state x_0
for $t = 0, 1, 2, \ldots$ **do**
Draw action $a_t \sim \mu(\cdot|x_t;\theta_t)$ and observe next state x_{t+1}, reward $R(x_t, a_t)$

$$\begin{aligned}
\textbf{Average Updates:}\quad & \widehat{\rho}_{t+1} = \big(1-\zeta_4(t)\big)\widehat{\rho}_t + \zeta_4(t)R(x_t,a_t)\\
& \widehat{\eta}_{t+1} = \big(1-\zeta_4(t)\big)\widehat{\eta}_t + \zeta_4(t)R(x_t,a_t)^2\\
\textbf{TD Errors:}\quad & \delta_t = R(x_t,a_t) - \widehat{\rho}_{t+1} + v_t^\top\phi_v(x_{t+1}) - v_t^\top\phi_v(x_t)\\
& \epsilon_t = R(x_t,a_t)^2 - \widehat{\eta}_{t+1} + u_t^\top\phi_u(x_{t+1}) - u_t^\top\phi_u(x_t)\\
\textbf{Critic Update:}\quad & v_{t+1} = v_t + \zeta_3(t)\delta_t\phi_v(x_t), \quad u_{t+1} = u_t + \zeta_3(t)\epsilon_t\phi_u(x_t)\\
\textbf{Actor Update:}\quad & \theta_{t+1} = \Gamma\Big(\theta_t - \zeta_2(t)\big(-\delta_t\psi_t + \lambda_t(\epsilon_t\psi_t - 2\widehat{\rho}_{t+1}\delta_t\psi_t)\big)\Big)\\
& \lambda_{t+1} = \Gamma_\lambda\Big(\lambda_t + \zeta_1(t)(\widehat{\eta}_{t+1} - \widehat{\rho}_{t+1}^2 - \alpha)\Big)
\end{aligned}$$

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

Experimental Results

Traffic Signal Control MDP

Problem Description

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

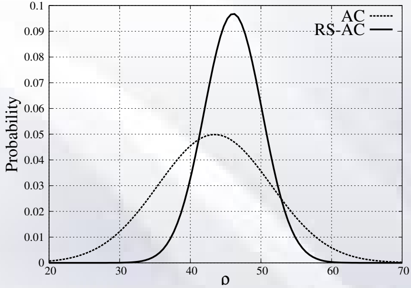
Action: feasible sign configurations

Cost:

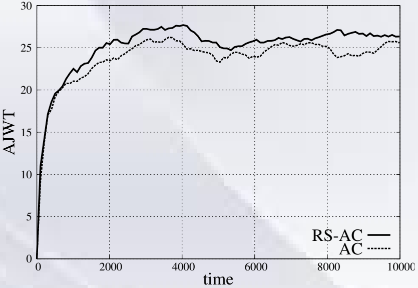
$$h(x_t) = r_1 * \Big[\sum_{i \in I_p} r_2 * q_i(t) + \sum_{i \notin I_p} s_2 * q_i(t)\Big] + s_1 * \Big[\sum_{i \in I_p} r_2 * t_i(t) + \sum_{i \notin I_p} s_2 * t_i(t)\Big]$$

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

Results - Average Reward Setting



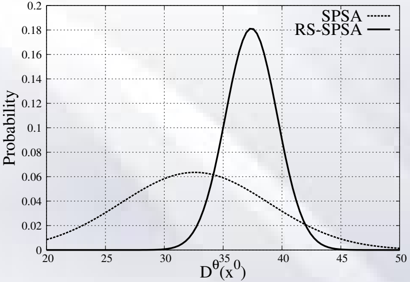
(a) Distribution of ρ



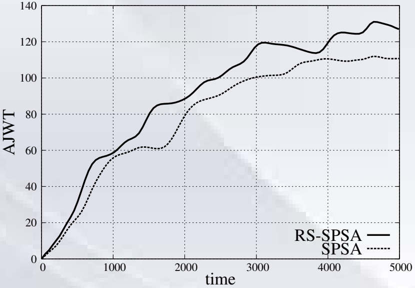
(b) Average junction waiting time

RS-AC vs. Risk-Nutral AC: higher return with lower variance

Results - Discounted Reward Setting



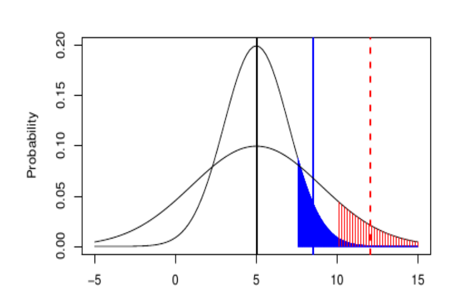
(c) Distribution of $D^{\theta}(x^0)$



(d) Average junction waiting time

CVaR as Risk Measure

Conditional Value-at-Risk (CVaR)



$$\mathrm{VaR}_\alpha(X) := \inf\left\{\xi \mid \mathbb{P}\left(X \le \xi\right) \ge \alpha\right\}$$
$$\mathrm{CVaR}_\alpha(X) := \mathbb{E}\left[X \mid X \ge \mathrm{VaR}_\alpha(X)\right].$$

Unlike VaR, CVaR is a coherent risk measure [1]

[1] convex, monotone, positive homogeneous and translation equi-variant

Practical Motivation

Portfolio Re-allocation

- **Portfolio** composed of assets (e.g. stocks)
- **Stochastic** gains for buying/selling assets
- **Aim** find an investment strategy that achieves a targeted asset allocation

A *risk-averse* investor would prefer a strategy that

1. quickly achieves the target asset allocation;
2. minimizes the worst-case losses incurred

Practical Motivation

Portfolio Re-allocation

Portfolio composed of assets (e.g. stocks)

Stochastic gains for buying/selling assets

Aim find an investment strategy that achieves a targeted asset allocation

A *risk-averse* investor would prefer a strategy that

1. quickly achieves the target asset allocation;
2. minimizes the worst-case losses incurred

CVaR-Constrained SSP

Stochastic Shortest Path

State. $\mathcal{S} = \{0, 1, \ldots, r\}$

Actions. $\mathcal{A}(s) = \{\text{feasible actions in state } s\}$

Costs. $g(s, a)$ and $c(s, a)$

used in the objective used in the constraint

Stochastic Shortest Path

State. $\mathcal{S} = \{0, 1, \ldots, r\}$

Actions. $\mathcal{A}(s) = \{\text{feasible actions in state } s\}$

Costs. $g(s, a)$ and $c(s, a)$

used in the objective

used in the constraint

CVaR-Constrained SSP

minimize the total cost:

$$\underbrace{\mathbb{E}\left[\sum_{m=0}^{\tau-1} g(s_m, a_m) \,\middle|\, s_0 = s^0\right]}_{G^\theta(s^0)}$$

subject to (CVaR constraint):

$$\underbrace{\mathrm{CVaR}_\alpha\left[\sum_{m=0}^{\tau-1} c(s_m, a_m) \,\middle|\, s_0 = s^0\right]}_{C^\theta(s^0)}$$

CVaR-Constrained SSP

minimize the total cost:

$$\underbrace{\mathbb{E}\left[\sum_{m=0}^{\tau-1} g(s_m, a_m) \,\Big|\, s_0 = s^0\right]}_{G^\theta(s^0)}$$

subject to (CVaR constraint):

$$\mathrm{CVaR}_\alpha \underbrace{\left[\sum_{m=0}^{\tau-1} c(s_m, a_m) \,\Big|\, s_0 = s^0\right]}_{C^\theta(s^0)}$$

Lagrangian Relaxation

$$\min_\theta \; G^\theta(s^0) \quad \text{s.t.} \quad \mathrm{CVaR}_\alpha(C^\theta(s^0) \le K_\alpha$$

$$\Updownarrow$$

$$\max_\lambda \min_\theta \left[\mathcal{L}^{\theta,\lambda}(s^0) \; := \; G^\theta(s^0) + \lambda\big(\mathrm{CVaR}_\alpha(C^\theta(s^0)) - K_\alpha\big)\right]$$

Solving the CVaR-constrained SSP

$$\max_{\lambda} \min_{\theta} \left[\mathcal{L}^{\theta,\lambda}(s^0) \ := \ G^{\theta}(s^0) + \lambda\big(\mathrm{CVaR}_{\alpha}(C^{\theta}(s^0)) - K_{\alpha}\big) \right]$$

Three-Stage Solution:

inner-most stage Simulate the SSP for several episodes and aggregate the costs;

next outer stage Estimate $\nabla_{\theta}\mathcal{L}^{\theta,\lambda}(s^0)$ using simulated values and update θ along descent direction[1]; and

outer-most stage update the Lagrange multipliers λ using the variance constraint

[1] Note: $\nabla_{\theta}\mathcal{L}^{\theta,\lambda}(s^0) = \nabla_{\theta}G^{\theta}(s^0) + \lambda\nabla_{\theta}\mathrm{CVaR}_{\alpha}(C^{\theta}(s^0)), \qquad \nabla_{\lambda}\mathcal{L}^{\theta,\lambda}(s^0) = \mathrm{CVaR}_{\alpha}(C^{\theta}(s^0)) - K_{\alpha}$

Solving the CVaR-constrained SSP

$$\max_{\lambda} \min_{\theta} \left[\mathcal{L}^{\theta,\lambda}(s^0) \; := \; G^{\theta}(s^0) + \lambda \big(\text{CVaR}_{\alpha}(C^{\theta}(s^0)) - K_{\alpha} \big) \right]$$

Three-Stage Solution:

inner-most stage Simulate the SSP for several episodes and aggregate the costs;

next outer stage Estimate $\nabla_{\theta} \mathcal{L}^{\theta,\lambda}(s^0)$ using simulated values and update θ along descent direction[1]; and

outer-most stage update the Lagrange multipliers λ using the variance constraint

[1] Note: $\nabla_{\theta} \mathcal{L}^{\theta,\lambda}(s^0) = \nabla_{\theta} G^{\theta}(s^0) + \lambda \nabla_{\theta} \text{CVaR}_{\alpha}(C^{\theta}(s^0))$, $\quad \nabla_{\lambda} \mathcal{L}^{\theta,\lambda}(s^0) = \text{CVaR}_{\alpha}(C^{\theta}(s^0)) - K_{\alpha}$

Solving the CVaR-constrained SSP

$$\max_{\lambda} \min_{\theta} \left[\mathcal{L}^{\theta,\lambda}(s^0) \; := \; G^{\theta}(s^0) + \lambda\big(\mathrm{CVaR}_{\alpha}(C^{\theta}(s^0)) - K_{\alpha}\big)\right]$$

Three-Stage Solution:

inner-most stage: Simulate the SSP for several episodes and aggregate the costs;

next outer stage: Estimate $\nabla_{\theta}\mathcal{L}^{\theta,\lambda}(s^0)$ using simulated values and update θ along descent direction[1]; and

outer-most stage: update the Lagrange multipliers λ using the variance constraint

[1] Note: $\nabla_{\theta}\mathcal{L}^{\theta,\lambda}(s^0) = \nabla_{\theta}G^{\theta}(s^0) + \lambda\nabla_{\theta}\mathrm{CVaR}_{\alpha}(C^{\theta}(s^0))$, $\quad \nabla_{\lambda}\mathcal{L}^{\theta,\lambda}(s^0) = \mathrm{CVaR}_{\alpha}(C^{\theta}(s^0)) - K_{\alpha}$

Solving the CVaR-constrained SSP

Three-Stage Solution:

inner-most stage Simulate the SSP for several episodes and aggregate the costs;

next outer stage Estimate $\nabla_\theta \mathcal{L}^{\theta,\lambda}(s^0)$ using simulated values and update θ along descent direction[1]; and

outer-most stage update the Lagrange multipliers λ using the variance constraint

$$\theta_{n+1} = \Gamma\left(\theta_n - \gamma_n \nabla_\theta \mathcal{L}^{\theta,\lambda}(s^0)\right) \quad \text{and} \quad \lambda_{n+1} = \Gamma_\lambda\left(\lambda_n + \gamma_n \nabla_\lambda \mathcal{L}^{\theta,\lambda}(s^0)\right),$$

[1] converge to a (local) saddle point of $^{\theta,\lambda}(s^0)$, i.e., to a tuple (θ^*, λ^*) that are a local minimum w.r.t. θ and a local maximum w.r.t. λ of $\mathcal{L}^{\theta,\lambda}(s^0)$

Solving the CVaR-constrained SSP

Three-Stage Solution:

inner-most stage Simulate the SSP for several episodes and aggregate the costs;

next outer stage Estimate $\nabla_\theta \mathcal{L}^{\theta,\lambda}(s^0)$ using simulated values and update θ along descent direction[1]; and

outer-most stage update the Lagrange multipliers λ using the variance constraint

$$\theta_{n+1} = \Gamma\left(\theta_n - \gamma_n \nabla_\theta \mathcal{L}^{\theta,\lambda}(s^0)\right) \quad \text{and} \quad \lambda_{n+1} = \Gamma_\lambda\left(\lambda_n + \gamma_n \nabla_\lambda \mathcal{L}^{\theta,\lambda}(s^0)\right),$$

[1] converge to a (local) saddle point of $^{\theta,\lambda}(s^0)$, i.e., to a tuple (θ^*, λ^*) that are a local minimum w.r.t. θ and a local maximum w.r.t. λ of $\mathcal{L}^{\theta,\lambda}(s^0)$

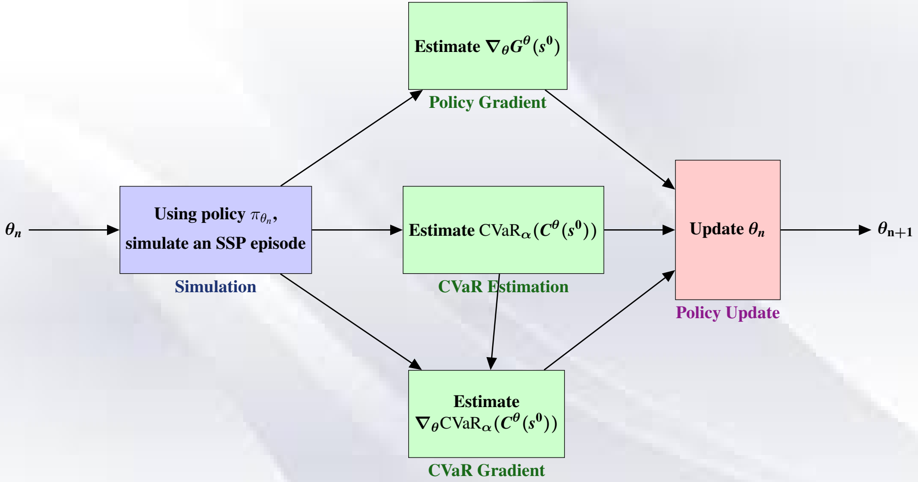


Figure : Overall flow of our algorithms.

Estimating CVaR: A convex optimization problem [2]

For any random variable X, let

$$v(\xi, X) := \xi + \frac{1}{1-\alpha}(X-\xi)_+ \text{ and}$$

$$V(\xi) = \mathbb{E}\left[v(\xi, X)\right]$$

Then,

$$\mathrm{VaR}_\alpha(X) = (\arg\min V := \{\xi \in \mathbb{R} \mid V'(\xi) = 0\})$$

$$\mathrm{CVaR}_\alpha(X) = V(\mathrm{VaR}_\alpha(X))$$

[2] Rockafellar, R.T., Uryasev, S. (2000), "Optimization of conditional value-at-risk". In: *Journal of risk*

Estimating CVaR: A convex optimization problem [2]

For any random variable X, let

$$v(\xi, X) := \xi + \frac{1}{1-\alpha}(X - \xi)_+ \text{ and}$$

$$V(\xi) = \mathbb{E}\left[v(\xi, X)\right]$$

Then,

$$\mathrm{VaR}_\alpha(X) = (\arg\min V := \{\xi \in \mathbb{R} \mid V'(\xi) = 0\})$$

$$\mathrm{CVaR}_\alpha(X) = V(\mathrm{VaR}_\alpha(X))$$

[2] Rockafellar, R.T., Uryasev, S. (2000), "Optimization of conditional value-at-risk". In: *Journal of risk*

Estimating $\mathrm{VaR}_\alpha(C^\theta(s^0))$

Observation: to estimate VaR, one needs to find ξ^* that satisfies $V'(\xi^*) = 0$

ξ_{n-1} → **Observe a new sample C_n of $C^\theta(s^0)$** (SSP simulation) → **Update ξ_n using $\frac{\partial v}{\partial \xi}(\xi, C_n)$** (GD Update) → ξ_n

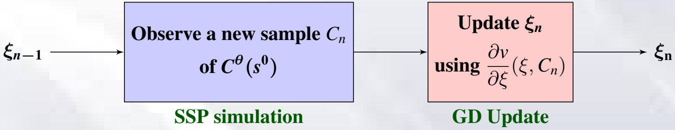
- Step-sizes

$$\xi_n = \xi_{n-1} - \zeta_{n,1} \left(1 - \frac{1}{1-\alpha} \mathbf{1}_{\{C_n \geq \xi\}}\right)$$

- Sample gradient

Estimating $\mathrm{VaR}_\alpha(C^\theta(s^0))$

Observation: to estimate VaR, one needs to find ξ^* that satisfies $V'(\xi^*) = 0$



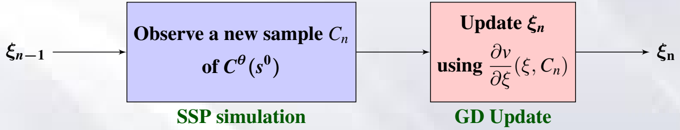
- Step-sizes

$$\xi_n = \xi_{n-1} - \zeta_{n,1}\left(1 - \frac{1}{1-\alpha}\mathbf{1}_{\{C_n \geq \xi\}}\right)$$

- Sample gradient

Estimating $\mathrm{VaR}_\alpha(C^\theta(s^0))$

Observation: to estimate VaR, one needs to find ξ^* that satisfies $V'(\xi^*) = 0$



- Step-sizes

$$\xi_n = \xi_{n-1} - \zeta_{n,1} \left(1 - \frac{1}{1-\alpha}\mathbf{1}_{\{C_n \geq \xi\}}\right)$$

- Sample gradient

Estimating $\mathrm{VaR}_\alpha(C^\theta(s^0))$

Observation: to estimate VaR, one needs to find ξ^* that satisfies $V'(\xi^*) = 0$

ξ_{n-1} → **Observe a new sample C_n of $C^\theta(s^0)$** (**SSP simulation**) → **Update ξ_n using $\frac{\partial v}{\partial \xi}(\xi, C_n)$** (**GD Update**) → ξ_n

- Step-sizes

$$\xi_n = \xi_{n-1} - \zeta_{n,1} \left(1 - \frac{1}{1-\alpha} \mathbf{1}_{\{C_n \geq \xi\}}\right)$$

- Sample gradient

Estimating $\text{CVaR}_\alpha(C^\theta(s^0))$[3]

Recall $\text{CVaR}_\alpha(C^\theta(s^0)) = \mathbb{E}\left[v(\text{VaR}_\alpha(C^\theta(s^0)), C^\theta(s^0))\right]$

To estimate CVaR, one can

Monte-Carlo Average

$$\frac{1}{m}\sum_{n=1}^{m} v(\xi_{n-1}, C_n)$$

Use Stochastic Approximation

$$\psi_n = \psi_{n-1} - \zeta_{n,2}\left(\psi_{n-1} - v(\xi_{n-1}, C_n)\right)$$

[3] O. Bardou et al. (2009) "Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling." In: Monte Carlo Methods and Applications

Estimating $\text{CVaR}_\alpha(C^\theta(s^0))$[3]

Recall $\text{CVaR}_\alpha(C^\theta(s^0)) = \mathbb{E}\left[v(\text{VaR}_\alpha(C^\theta(s^0)), C^\theta(s^0))\right]$

To estimate CVaR, one can

Monte-Carlo Average

$$\frac{1}{m}\sum_{n=1}^{m} v(\xi_{n-1}, C_n)$$

Use Stochastic Approximation

$$\psi_n = \psi_{n-1} - \zeta_{n,2}\left(\psi_{n-1} - v(\xi_{n-1}, C_n)\right)$$

[3] O. Bardou et al. (2009) "Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling." In: Monte Carlo Methods and Applications

Estimating $\mathrm{CVaR}_\alpha(C^\theta(s^0))$[3]

Recall $\mathrm{CVaR}_\alpha(C^\theta(s^0)) = \mathbb{E}\left[v(\mathrm{VaR}_\alpha(C^\theta(s^0)), C^\theta(s^0))\right]$

To estimate CVaR, one can

Monte-Carlo Average

$$\frac{1}{m}\sum_{n=1}^{m} v(\xi_{n-1}, C_n)$$

Use Stochastic Approximation

$$\psi_n = \psi_{n-1} - \zeta_{n,2}\left(\psi_{n-1} - v(\xi_{n-1}, C_n)\right)$$

[3] O. Bardou et al. (2009) "Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling." In: Monte Carlo Methods and Applications

Likelihood ratios for gradient estimation[4]

Markov chain. $\{X_n\}$

States. 0 recurrent and $1, \ldots, r$ transient

Transition probability matrix. $P(\theta) := [[p_{X_i X_j}(\theta)]]_{i,j=0}^{r}$

Performance measure. $F(\theta) = \mathbb{E}[f(X)]$

Simulate (using $P(\theta)$) and obtain $X := (X_0, \ldots, X_{\tau-1})^T$

$$\nabla_\theta F(\theta) = \mathbb{E}\left[f(X) \sum_{m=0}^{\tau-1} \frac{\nabla_\theta p_{X_m X_{m+1}}(\theta)}{p_{X_m X_{m+1}}(\theta)} \right]$$

[4] Glynn, P.W. (1987) |"Likelilood ratio gradient estimation: an overview." In: Winter simulation conference

Likelihood ratios for gradient estimation[4]

Markov chain. $\{X_n\}$

States. 0 recurrent and $1, \ldots, r$ transient

Transition probability matrix. $P(\theta) := [[p_{X_i X_j}(\theta)]]_{i,j=0}^{r}$

Performance measure. $F(\theta) = \mathbb{E}[f(X)]$

Simulate (using $P(\theta)$) and obtain $X := (X_0, \ldots, X_{\tau-1})^T$

$$\nabla_\theta F(\theta) = \mathbb{E}\left[f(X) \sum_{m=0}^{\tau-1} \frac{\nabla_\theta p_{X_m X_{m+1}}(\theta)}{p_{X_m X_{m+1}}(\theta)} \right]$$

[4] Glynn, P.W. (1987) |"Likelilood ratio gradient estimation: an overview." In: Winter simulation conference

Policy gradient for the objective [5]

Policy gradient:

$$\nabla_\theta G^\theta(s^0) = \mathbb{E}\left[\left(\sum_{n=0}^{\tau-1} g(s_n, a_n)\right) \nabla \log P(s_0, \ldots, s_{\tau-1}) \mid s_0 = s^0\right],$$

Likelihood derivative:

$$\nabla \log P(s_0, \ldots, s_{\tau-1}) = \sum_{m=0}^{\tau-1} \nabla \log \pi_\theta(a_m | s_m)$$

[5] Bartlett, P.L., Baxter, J. (2011) "Infinite-horizon policy-gradient estimation."

Policy gradient for the CVaR constraint [6]

$$\nabla_\theta \mathrm{CVaR}_\alpha(C^\theta(s^0))$$
$$= \mathbb{E}\left[\left(C^\theta(s^0) - \mathrm{VaR}_\alpha(C^\theta(s^0))\right) \nabla \log P(s_0, \ldots, s_{\tau-1}) \mid C^\theta(s^0) \geq \mathrm{VaR}_\alpha(C^\theta(s^0)) \right],$$

where $\nabla \log P(s_0, \ldots, s_\tau)$ is the likelihood derivative

[6] Tamar, A. et al. (2014) "Policy Gradients Beyond Expectations: Conditional Value-at-Risk." In: arxiv:1404.3862

Putting it all together. . .

Input: parameterized policy $\pi_\theta(\cdot|\cdot)$, step-sizes $\{\zeta_{n,1}, \zeta_{n,2}, \gamma_n\}_{n\geq 1}$
For each $n = 1, 2, \ldots$ **do**

Simulate the SSP using $\pi_{\theta_{n-1}}$ and obtain:

$$G_n := \sum_{j=0}^{\tau_n - 1} g(s_{n,j}, a_{n,j}),\, C_n := \sum_{j=0}^{\tau_n - 1} c(s_{n,j}, a_{n,j}) \text{ and } z_n := \sum_{j=0}^{\tau_n - 1} \nabla \log \pi_\theta(s_{n,j}, a_{n,j})$$

VaR/CVaR estimation:

$$\text{VaR: } \xi_n = \xi_{n-1} - \zeta_{n,1}\left(1 - \frac{1}{1-\alpha}\mathbf{1}_{\{C_n \geq \xi_{n-1}\}}\right), \qquad \text{CVaR: } \psi_n = \psi_{n-1} - \zeta_{n,2}\left(\psi_{n-1} - v(\xi_{n-1}, C_n)\right)$$

Policy Gradient:

$$\text{Total Cost: } \bar{G}_n = \bar{G}_{n-1} - \zeta_{n,2}(G_n - \bar{G}_n), \quad \text{Gradient: } \partial G_n = \bar{G}_n z_n$$

CVaR Gradient:

$$\text{Total Cost: } \tilde{C}_n = \tilde{C}_{n-1} - \zeta_{n,2}(C_n - \tilde{C}_n), \quad \text{Gradient: } \partial C_n = (\tilde{C}_n - \xi_n) z_n \mathbf{1}_{\{C_n \geq \xi_n\}}$$

Policy and Lagrange Multiplier Update:

$$\theta_n = \theta_{n-1} - \gamma_n(\partial G_n + \lambda_{n-1}(\partial C_n)), \qquad \lambda_n = \Gamma_\lambda\left(\lambda_{n-1} + \gamma_n(\psi_n - K_\alpha)\right).$$

mini-Batches

θ_{n-1} → **Using policy $\pi_{\theta_{n-1}}$, simulate m_n episodes** → **Obtain $\{G_{n,j}, C_{n,j}, z_{n,j}\}_{j=1}^{m_n}$** → **Compute $\mathrm{CVaR}_\alpha(C^\theta(s^0))$ and $\nabla_\theta \mathrm{CVaR}_\alpha(C^\theta(s^0)), \nabla_\theta G^\theta(s^0)$** → θ_n

Simulation | **Cost/Likelihood Estimates** | **Averaging**

Figure : mini-batch idea

$$\text{VaR: } \xi_n = \frac{1}{m_n}\sum_{j=1}^{m_n}\left(1 - \frac{\mathbf{1}_{\{C_{n,j} \geq \xi_{n-1}\}}}{1-\alpha}\right), \quad \text{CVaR: } \psi_n = \frac{1}{m_n}\sum_{j=1}^{m_n} v(\xi_{n-1}, C_{n,j})$$

$$\text{Total Cost: } \bar{G}_n = \frac{1}{m_n}\sum_{j=1}^{m_n} G_{n,j}, \qquad \text{Policy Gradient: } \partial G_n = \bar{G}_n z_n.$$

$$\text{Total Cost: } \bar{C}_n = \frac{1}{m_n}\sum_{j=1}^{m_n} C_{n,j}, \qquad \text{CVaR Gradient: } \partial C_n = (\bar{C}_n - \xi_n) z_n \mathbf{1}_{\{\bar{C}_n \geq \xi_n\}}.$$

Comparison to Previous Work

Borkar V et al. (2010) propose an algorithm for a (finite horizon) CVaR constrained MDP, under a separability condition.

Tamar et al. (2014) do not consider a risk-constrained SSP and instead optimize only CVaR.

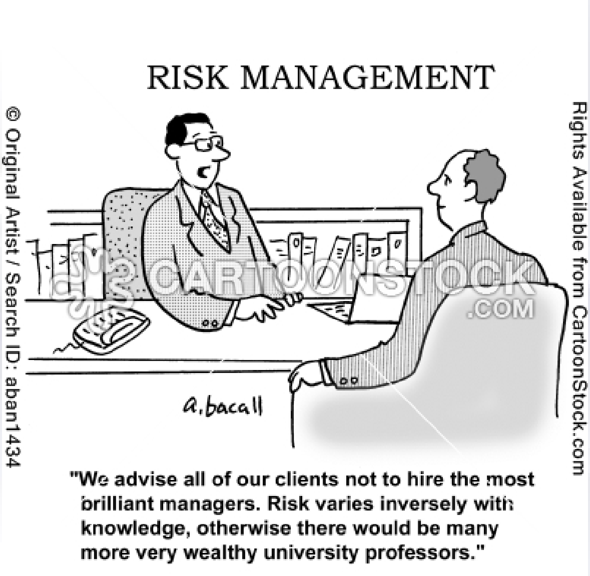
[1] Borkar V (2010) “Risk-constrained Markov decision processes” In: CDC

[2] Tamar et al (2014) “Policy Gradients Beyond Expectations: Conditional Value-at-Risk” In: arxiv:1404.3862

References I

- Prashanth L.A. and Mohammad Ghavamzadeh.
  Actor-Critic Algorithms for Risk-Sensitive MDPs.
  NIPS (Full oral presentation), 2013.
- S.Bhatnagar, H.L.Prasad and Prashanth.L.A.,
  Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods,
  Lecture Notes in Control and Information Sciences Series, Springer, 2012.
- Prashanth L.A.
  Policy Gradients for CVaR-Constrained MDPs.
  ALT, 2013.

What next?


RISK MANAGEMENT
© Original Artist / Search ID: aban1434
Rights Available from CartoonStock.com
CARTOONSTOCK .COM
a.bacall
"We advise all of our clients not to hire the most brilliant managers. Risk varies inversely with knowledge, otherwise there would be many more very wealthy university professors."