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
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

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

- a criterion that penalizes the \textit{variability} induced by a given policy
- minimize some measure of \textit{risk} as well as maximizing a usual optimization criterion
Risk-Sensitive Sequential Decision-Making

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

- 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

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

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

Return r.v. Reward Policy

Risk-Sensitive Sequential Decision-Making

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

- a criterion that penalizes the \textit{variability} induced by a given policy
- minimize some measure of \textit{risk} as well as maximizing a usual optimization criterion
Risk-Sensitive Sequential Decision-Making

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

- a criterion that penalizes the \textit{variability} induced by a given policy
- minimize some measure of \textit{risk} as well as maximizing a usual optimization criterion
Risk-Sensitive Sequential Decision-Making

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

- 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

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

- a criterion that penalizes the \textit{variability} induced by a given policy

- minimize some measure of \textit{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)

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

Variance of Return (measure of variability)

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

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

Variance of Return (measure of variability)

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

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

Variance of Return (measure of variability)

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

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

Optimization Problem

\[
\begin{align*}
\max_{\theta} & \quad V^\theta(x^0) \\
\text{s.t.} & \quad \Lambda^\theta(x^0) \leq \alpha
\end{align*}
\]

\[
\begin{align*}
\max_{\lambda} \min_{\theta} & \quad L(\theta, \lambda) \\
\triangleq & \quad -V^\theta(x^0) + \lambda (\Lambda^\theta(x^0) - \alpha)
\end{align*}
\]
Risk-Sensitive Discounted MDPs

A class of parameterized stochastic policies

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

Optimization Problem

\[ \max_{\theta} \ V^\theta(x^0) \ \text{s.t.} \ \Lambda^\theta(x^0) \leq \alpha \]

\[ \triangleq \]

\[ \max_{\lambda} \ \min_{\theta} \ L(\theta, \lambda) \triangleq -V^\theta(x^0) + \lambda (\Lambda^\theta(x^0) - \alpha) \]
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 $\theta$ along descent direction; and

outer-most stage update the Lagrange multipliers $\lambda$ using the variance constraint

\[ \nabla_{\chi}L(\theta, \lambda) = \Lambda^{\theta}(x^0) - \alpha. \]

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

One needs to evaluate $\nabla_{\theta}L(\theta, \lambda)$ and $\nabla_{\chi}L(\theta, \lambda)$ to tune $\theta$ and $\lambda$. 
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 $\theta$ along descent direction; and

outer-most stage  update the Lagrange multipliers $\lambda$ 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 $\theta$ and $\lambda$. 
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 $\theta$ along descent direction; and

outer-most stage  update the Lagrange multipliers $\lambda$ 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 $\theta$ and $\lambda$
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 $\theta$ along descent direction; and

outer-most stage  update the Lagrange multipliers $\lambda$ 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 $\theta$ and $\lambda$
Computing the Gradients

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

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

$$(1 - \gamma^2) \nabla_\theta U^\theta(x^0) = \sum_{x,a} \tilde{\pi}_\gamma^\theta(x, a|x^0) \nabla_\theta \log \mu(a|x; \theta) \ W^\theta(x, a)$$

$$+ 2\gamma \sum_{x,a,x'} \tilde{\pi}_\gamma^\theta(x, a|x^0) \ P(x'|x, a) \ r(x, a) \ \nabla_\theta V^\theta(x')$$

$\pi_\gamma^\theta(x, a|x^0)$ and $\tilde{\pi}_\gamma^\theta(x, a|x^0)$ are $\gamma$ and $\gamma^2$ discounted visiting state distributions of the Markov chain under policy $\theta$
Why Estimating the Gradient is Challenging?

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

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

$$(1 - \gamma^2) \nabla_{\theta} U^\theta(x^0) = \sum_{x,a} \tilde{\pi}_\gamma^\theta(x,a|x^0) \nabla_{\theta} \log \mu(a|x; \theta) W^\theta(x,a)$$

$$+ 2\gamma \sum_{x,a,x'} \tilde{\pi}_\gamma^\theta(x,a|x^0) P(x'|x,a) r(x,a) \nabla_{\theta} V^\theta(x')$$

$\pi_\gamma^\theta(x,a|x^0)$ and $\tilde{\pi}_\gamma^\theta(x,a|x^0)$ are $\gamma$ and $\gamma^2$ discounted visiting state distributions of the Markov chain under policy $\theta$
Why Simultaneous Perturbation?

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

- two different sampling distributions ($\pi^\theta_\gamma$ and $\tilde{\pi}^\theta_\gamma$) 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**

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

$\Delta$ is a vector of independent Rademacher random variables
Why Simultaneous Perturbation?

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

- two different sampling distributions ($\pi_\gamma^\theta$ and $\tilde{\pi}_\gamma^\theta$) 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**

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

$\Delta$ is a vector of independent Rademacher random variables
SPSA idea

**Scalar \( \theta \):**

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

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

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

Thus,

\[
\frac{V(\theta + \beta) - V(\theta)}{\beta} = \frac{dV(\theta)}{d\theta} + o(\beta).
\]

**Vector \( \theta \in \mathbb{R}^{\kappa_1} \):**

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

where \( \Delta \) is a vector of independent Rademacher random variables.
SPSA idea

**Scalar $\theta$:**

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

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

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

Thus,

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

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

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

where $\Delta$ 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 $\theta$ and $\theta^+ = \theta + \beta \Delta$, $\beta > 0$. 

$$\theta_t + \beta \Delta_t \rightarrow a_t^+ \sim \mu(\cdot|x_t^+; \theta_t^+)$$

$$r_t^+ \rightarrow \delta_t^+, \epsilon_t^+, v_t^+, u_t^+$$

$$\theta_t \rightarrow a_t \sim \mu(\cdot|x_t; \theta_t)$$

$$r_t \rightarrow \delta_t, \epsilon_t, v_t, u_t$$

Critic

Actor

Update $\theta_t$ using (8) or (9)
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 $\theta$ and $\theta^+ = \theta + \beta \Delta$, $\beta > 0$. 

Prashanth L.A. (INRIA) Algorithms for Risk-Sensitive Reinforcement Lear

14 / 48
\[ \hat{V}(x) \approx v^\top \phi_v(x) \text{ and } \hat{U}(x) \approx u^\top \phi_u(x) \]

<table>
<thead>
<tr>
<th>Update rule</th>
<th>Trajectory 1</th>
<th>Trajectory 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>( v_{t+1} = v_t + \zeta_3(t) \delta_t \phi_v(x_t) )</td>
<td>( v_{t+1}^+ = v_t^+ + \zeta_3(t) \delta_t^+ \phi_v(x_t^+) )</td>
</tr>
<tr>
<td>Square-Value</td>
<td>( u_{t+1} = u_t + \zeta_3(t) \epsilon_t \phi_u(x_t) )</td>
<td>( u_{t+1}^+ = u_t^+ + \zeta_3(t) \epsilon_t^+ \phi_u(x_t^+) )</td>
</tr>
</tbody>
</table>

\( \delta_t, \delta_t^+, \epsilon_t, \epsilon_t^+ \) denote the TD-errors.

Critic Update
Approximation

\[ \hat{V}(x) \approx v^\top \phi_v(x) \] and \[ \hat{U}(x) \approx u^\top \phi_u(x) \]

Update rule

<table>
<thead>
<tr>
<th></th>
<th>Trajectory 1</th>
<th>Trajectory 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Value</strong></td>
<td>( v_{t+1} = v_t + \zeta_3(t) \delta_t \phi_v(x_t) )</td>
<td>( v_{t+1}^+ = v_t^+ + \zeta_3(t) \delta_t^+ \phi_v(x_t^+) )</td>
</tr>
<tr>
<td><strong>Square-Value</strong></td>
<td>( u_{t+1} = u_t + \zeta_3(t) \epsilon_t \phi_u(x_t) )</td>
<td>( u_{t+1}^+ = u_t^+ + \zeta_3(t) \epsilon_t^+ \phi_u(x_t^+) )</td>
</tr>
</tbody>
</table>

\( \delta_t, \delta_t^+, \epsilon_t, \epsilon_t^+ \) denote the TD-errors.

Critic Update (contd)

**TD-errors $\delta_t$, $\epsilon_t$ in Trajectory 1 (policy $\theta$)**

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

**TD-errors $\delta_t^+$, $\epsilon_t^+$ in Trajectory 2 (perturbed policy $\theta + \beta \Delta$)**

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

TD-errors $\delta_t, \epsilon_t$ in Trajectory 1 (policy $\theta$)

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

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

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

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

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

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

\[
\theta_{t+1}^{(i)} = \Gamma_i \left[ \theta_t^{(i)} + \zeta_2(t) \left( \frac{(1 + 2\lambda_t v_t^\top \phi_v(x^0))(v_t^{+} - v_t)^\top \phi_v(x^0) - \lambda_t (u_t^{+} - u_t)^\top \phi_u(x^0)}{\beta \Delta_t^{(i)}} \right) \right] \\
\lambda_{t+1} = \Gamma_\lambda \left[ \lambda_t + \zeta_1(t) \left( u_t^\top \phi_u(x^0) - (v_t^\top \phi_v(x^0))^2 - \alpha \right) \right]
\]

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

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

Variance

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

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.


19 / 48
Notation

Average Reward

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

Variance

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

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

Notation

Average Reward

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

Variance

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

Stream of rewards: (0,0,0,0,\ldots) or (100,-100,100,-100,\ldots)

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

Notation

Average Reward

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

Variance

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

Stream of rewards: \((0,0,0,0,\ldots)\) or \((100,-100,100,-100,\ldots)\)

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

Risk-sensitive MDP

Objective

\[
\max_{\theta} \rho(\theta) \quad \text{subject to} \quad \Lambda(\theta) \leq \alpha
\]

\[
\iff
\]

\[
\max_{\lambda} \min_{\theta} \left( L(\theta, \lambda) \triangleq -\rho(\theta) + \lambda(\Lambda(\theta) - \alpha) \right)
\]

As before, one needs \(\nabla_{\theta} L(\theta, \lambda)\) to tune policy parameter \(\theta\).
Risk-sensitive MDP

Objective

\[
\max_{\theta} \rho(\theta) \quad \text{subject to} \quad \Lambda(\theta) \leq \alpha
\]

\[
\max_{\lambda} \min_{\theta} \left( L(\theta, \lambda) \triangleq -\rho(\theta) + \lambda(\Lambda(\theta) - \alpha) \right)
\]

As before, one needs \( \nabla_{\theta} L(\theta, \lambda) \) to tune policy parameter \( \theta \)
Notation (again)

Average Reward

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

Variance

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

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

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

\( U^\mu \) and \( W^\mu \) are the differential value and action-value functions that satisfy

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

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

RS-AC algorithm

**Initialization:** policy parameters $\theta_0$; 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 $R(x_t, a_t)$

**Average Updates:**

$\hat{\rho}_{t+1} = (1 - \zeta_4(t)) \hat{\rho}_t + \zeta_4(t) R(x_t, a_t)$

$\hat{\eta}_{t+1} = (1 - \zeta_4(t)) \hat{\eta}_t + \zeta_4(t) R(x_t, a_t)^2$

**TD Errors:**

$\delta_t = R(x_t, a_t) - \hat{\rho}_{t+1} + v_t^\top \phi_v(x_{t+1}) - v_t^\top \phi_v(x_t)$

$\epsilon_t = R(x_t, a_t)^2 - \hat{\eta}_{t+1} + u_t^\top \phi_u(x_{t+1}) - u_t^\top \phi_u(x_t)$

**Critic Update:**

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

**Actor Update:**

$\theta_{t+1} = \Gamma \left( \theta_t - \zeta_2(t) \left( - \delta_t \psi_t + \lambda_t (\epsilon_t \psi_t - 2 \hat{\rho}_{t+1} \delta_t \psi_t) \right) \right)$

$\lambda_{t+1} = \Gamma \lambda \left( \lambda_t + \zeta_1(t) (\hat{\eta}_{t+1} - \hat{\rho}_{t+1}^2 - \alpha) \right)$

end for

return policy and value function parameters $\theta, \lambda, v, u$
Experimental Results
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:

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

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 $\rho$

(b) Average junction waiting time

RS-AC vs. Risk-Neutral 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)

\[ \text{VaR}_\alpha(X) := \inf \{ \xi \mid \mathbb{P}(X \leq \xi) \geq \alpha \} \]

\[ \text{CVaR}_\alpha(X) := \mathbb{E}[X \mid X \geq \text{VaR}_\alpha(X)] \, . \]

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: \( S = \{0, 1, \ldots, r\} \)

Actions: \( A(s) = \{\text{feasible actions in state } s\} \)

Costs: \( g(s, a) \) and \( c(s, a) \) used in the objective and constraint.
Stochastic Shortest Path

State. \( S = \{0, 1, \ldots, r\} \)

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

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

\( g(s, a) \) used in the objective
\( c(s, a) \) used in the constraint
CVaR-Constrained SSP

minimize the total cost:

\[ \min \mathbb{E} \left[ \sum_{m=0}^{\tau-1} g(s_m, a_m) \left| s_0 = s^0 \right. \right] \]

subject to (CVaR constraint):

\[ \text{CVaR}_\alpha \left[ \sum_{m=0}^{\tau-1} c(s_m, a_m) \left| s_0 = s^0 \right. \right] \leq G^\theta(s^0) \]
CVaR-Constrained SSP

minimize the total cost:

\[
\mathbb{E} \left[ \sum_{m=0}^{\tau-1} g(s_m, a_m) \mid s_0 = s^0 \right]
\]

subject to (CVaR constraint):

\[
\text{CVaR}_\alpha \left[ \sum_{m=0}^{\tau-1} c(s_m, a_m) \mid s_0 = s^0 \right]
\]
Lagrangian Relaxation

\[
\min_\theta \quad G^\theta (s^0) \quad \text{s.t.} \quad \text{CVaR}_\alpha (C^\theta (s^0)) \leq K_\alpha
\]

\[
\overset{\Leftrightarrow}{\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \Quad
Solving the CVaR-constrained SSP

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

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 \( \theta \) along descent direction\(^1\); and

outer-most stage  update the Lagrange multipliers \( \lambda \) using the variance constraint

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

\(^1\) 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

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

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 \( \theta \) along descent direction \(^1\); and

outer-most stage: update the Lagrange multipliers \( \lambda \) 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

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

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 \( \theta \) along descent direction\(^1\); and

outer-most stage  update the Lagrange multipliers \( \lambda \) 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

Three-Stage Solution:
inner-most stage  Simulate the SSP for several episodes and aggregate the costs;
next outer stage  Estimate $\nabla_\theta L^\theta,\lambda(s^0)$ using simulated values and update $\theta$ along descent direction\(^1\); and
outer-most stage  update the Lagrange multipliers $\lambda$ using the variance constraint

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

\(^1\) converge to a (local) saddle point of $\theta,\lambda(s^0)$, i.e., to a tuple $(\theta^*, \lambda^*)$ that are a local minimum w.r.t. $\theta$ and a local maximum w.r.t. $\lambda$ of $L^\theta,\lambda(s^0)$.
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 $\theta$ along descent direction$^1$; and

outer-most stage  
update the Lagrange multipliers $\lambda$ 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 $\theta^*, \lambda^*$, i.e., to a tuple $(\theta^*, \lambda^*)$ that are a local minimum w.r.t. $\theta$ and a local maximum w.r.t. $\lambda$ of $\mathcal{L}^{\theta,\lambda}(s^0)$.
Using policy $\pi_{\theta_n}$, simulate an SSP episode

$\theta_n$ \rightarrow Estimate $\nabla_{\theta} G^\theta(s^0)$ \rightarrow $\theta_{n+1}$

Simulation

Estimate $\text{CVaR}_\alpha(C^\theta(s^0))$ \rightarrow Update $\theta_n$

Policy Gradient

Estimate $\nabla_{\theta} \text{CVaR}_\alpha(C^\theta(s^0))$ \rightarrow $\theta_{n+1}$

CVaR Estimation

CVaR Gradient

Figure: Overall flow of our algorithms.
Estimating CVaR: A convex optimization problem

For any random variable $X$, let

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

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

Then,

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

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

---

Estimating CVaR: A convex optimization problem

For any random variable \( X \), let

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

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

Then,

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

\[
\text{CVaR}_\alpha(X) = V(\text{VaR}_\alpha(X))
\]

---

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

**Observation:** to estimate VaR, one needs to find $\xi^*$ 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**

- **SSP simulation**

- **GD Update**

Observe a new sample $C_n$ of $C^\theta(s^0)$

Update $\xi_n$ using $\frac{\partial v}{\partial \xi}(\xi, C_n)$
Estimating $\text{VaR}_\alpha(C^\theta(s^0))$

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

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

$$\xi_n = \xi_{n-1} - \zeta_{n,1}$$

$$\left(1 - \frac{1}{1 - \alpha} \mathbf{1}_{\{C_n \geq \xi\}}\right)$$
Estimating \( \text{VaR}_\alpha(C^\theta(s^0)) \)

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

- Observe a new sample \( C_n \) of \( C^\theta(s^0) \)
- **SSP simulation**
- **Update** \( \xi_n \) using
  \[
  \frac{\partial v}{\partial \xi}(\xi, C_n)
  \]
- **GD Update**

- **Step-sizes**
  \[
  \xi_n = \xi_{n-1} - \zeta_{n,1}
  \]

- **Sample gradient**
  \[
  \left( 1 - \frac{1}{1 - \alpha} \mathbf{1}_{\{C_n \geq \xi\}} \right)
  \]
Estimating $\text{VaR}_\alpha(C^\theta(s^0))$

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

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

- **Sample gradient**
Estimating $\text{CVaR}_\alpha(C^\theta(s^0))^3$

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

To estimate CVaR, one can

Monte-Carlo Average

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

Use Stochastic Approximation

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

---

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**
  \[
  \psi_n = \psi_{n-1} - \zeta_{n,2} (\psi_{n-1} - v(\xi_{n-1}, C_n))
  \]

---

\(^3\) 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 $\text{CVaR}_\alpha(C^\theta(s^0)) = \mathbb{E}\left[\nu(\text{VaR}_\alpha(C^\theta(s^0)), C^\theta(s^0))\right]$ 

To estimate CVaR, one can

**Monte-Carlo Average**

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

**Use Stochastic Approximation**

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

---

\(^3\) 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

Markov chain. \( \{X_n\} \)

States. \( 0 \) recurrent and \( 1, \ldots, r \) transient

Transition probability matrix. \( P(\theta) := \begin{bmatrix} [p_{X_iX_j}(\theta)]_{i,j=0}^r \end{bmatrix} \)

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

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

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

---

Likelihood ratios for gradient estimation

Markov chain. \( \{X_n\} \)

States. 0 recurrent and 1, \ldots, \( r \) transient

Transition probability matrix. \( P(\theta) := \left[ [p_{X_iX_j}(\theta)] \right]_{i,j=0}^r \)

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

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

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

---

Policy gradient for the objective

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:

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

Policy gradient for the CVaR constraint

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

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

where \( \nabla \log P(s_0, \ldots, s_{\tau}) \) is the likelihood derivative.
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:

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

**VaR/CVaR estimation:**

\[
\text{VaR: } \xi_n = \xi_{n-1} - \zeta_{n,1} \left( 1 - \frac{1}{1-\alpha} \right) 1\{C_n \geq \xi_{n-1}\}, \quad \text{CVaR: } \psi_n = \psi_{n-1} - \zeta_{n,2} (\psi_{n-1} - \nu(\xi_{n-1}, C_n))
\]

**Policy Gradient:**

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

**CVaR Gradient:**

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

**Policy and Lagrange Multiplier Update:**

\[
\theta_n = \theta_{n-1} - \gamma_n (\partial G_n + \lambda_{n-1} (\partial C_n)), \quad \lambda_n = \Gamma \lambda \left( \lambda_{n-1} + \gamma_n (\psi_n - K\alpha) \right).
\]
mini-Batches

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 $\text{CVaR}_\alpha(C^\theta(s^0))$ and $\nabla_\theta \text{CVaR}_\alpha(C^\theta(s^0)), \nabla_\theta G^\theta(s^0)$

Averaging

Prashanth L.A. (INRIA) Algorithms for Risk-Sensitive Reinforcement Lear

Figure: mini-batch idea

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

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

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

Prashanth L.A.
Policy Gradients for CVaR-Constrained MDPs. 
*ALT, 2013.*
What next?

RISK MANAGEMENT

© Original Artist/ Search ID: aban1434

"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."

Rights Available from CartoonStock.com