Policy Gradients for CVaR-Constrained MDPs

Prashanth L.A.

INRIA Lille – Team SequeL

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

Risk-neutral Objective:

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

Total Cost Cost 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

Risk-neutral Objective:

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

Total Cost   Cost   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-neutral Objective:

$$\min_{\theta\in\Theta} G^{\theta}(s^0) = \mathbb{E}\left[\sum_{m=0}^{\tau-1} g(s_m, a_m) \mid s_0 = s^0, \quad \theta\right]$$

Total Cost   Cost   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

Risk-neutral Objective:

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

Total Cost Cost 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

Risk-neutral Objective:

$$\min_{\theta \in \Theta} \underbrace{G^{\theta}(s^0)}_{\text{Total Cost}} = \mathbb{E}\left[\sum_{m=0}^{\tau-1} \underbrace{g(s_m, a_m)}_{\text{Cost}} \mid s_0 = s^0, \underbrace{\theta}_{\text{Policy}}\right]$$

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

A brief history of risk measures

Risk measures considered in the literature:

- 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)

A brief history of risk measures

Risk measures considered in the literature:

- 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)

A brief history of risk measures

Risk measures considered in the literature:

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

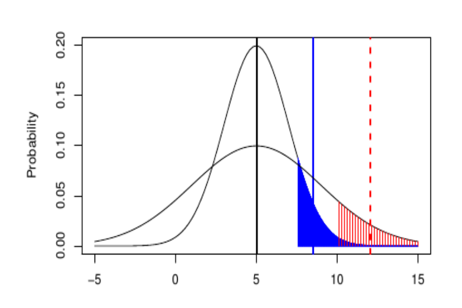
Open Question ???

construct conceptually meaningful and computationally tractable criteria

mainly negative results:

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

Conditional Value-at-Risk (CVaR)



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

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

Our Contributions

define a CVaR-constrained ***stochastic shortest path*** problem

derive CVaR estimation procedures using ***stochastic approximation***

propose ***policy gradient algorithms*** to optimize CVaR-constrained SSP

establish the ***asymptotic convergence*** of the algorithms

adapt our proposed algorithms to incorporate importance sampling (IS)

Our Contributions

define a CVaR-constrained ***stochastic shortest path*** problem

derive CVaR estimation procedures using ***stochastic approximation***

propose ***policy gradient algorithms*** to optimize CVaR-constrained SSP

establish the ***asymptotic convergence*** of the algorithms

adapt our proposed algorithms to incorporate importance sampling (IS)

Our Contributions

define a CVaR-constrained ***stochastic shortest path*** problem

derive CVaR estimation procedures using ***stochastic approximation***

propose ***policy gradient algorithms*** to optimize CVaR-constrained SSP

establish the ***asymptotic convergence*** of the algorithms

adapt our proposed algorithms to incorporate importance sampling (IS)

Our Contributions

define a CVaR-constrained ***stochastic shortest path*** problem

derive CVaR estimation procedures using ***stochastic approximation***

propose ***policy gradient algorithms*** to optimize CVaR-constrained SSP

establish the ***asymptotic convergence*** of the algorithms

adapt our proposed algorithms to incorporate importance sampling (IS)

Our Contributions

define a CVaR-constrained ***stochastic shortest path*** problem

derive CVaR estimation procedures using ***stochastic approximation***

propose ***policy gradient algorithms*** to optimize CVaR-constrained SSP

establish the ***asymptotic convergence*** of the algorithms

adapt our proposed algorithms to incorporate importance sampling (IS)

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. $\mathcal{A}(s) =$ {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:

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

subject to (CVaR constraint):

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

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):

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

Lagrangian Relaxation

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

$$\Updownarrow$$

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

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

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

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

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

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 $\mathcal{L}^{\theta,\lambda}(s^0)$

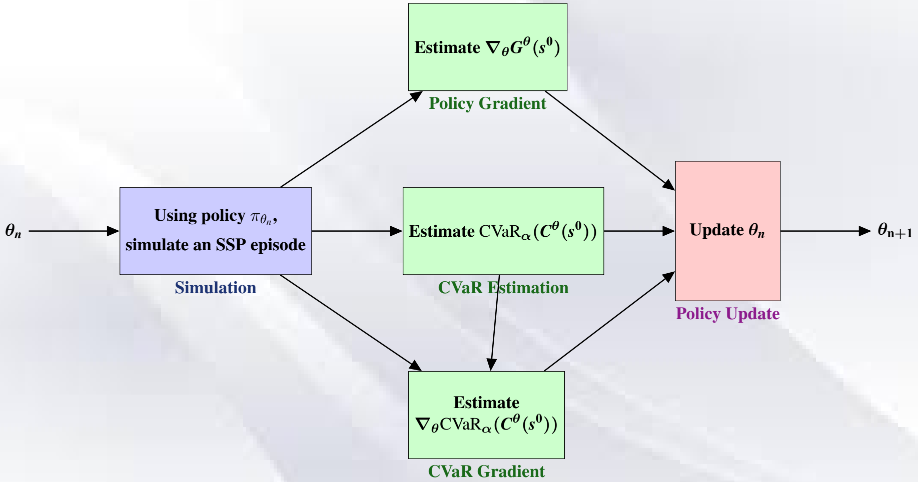


Figure: Overall flow of our algorithms.

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 $C^\theta(s^0)$** → **Update ξ_n using $\frac{\partial v}{\partial \xi}(\xi, C_n)$** → ξ_n

SSP simulation

GD Update

- 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 $V'(\xi^*) = 0$

$$\xi_{n-1} \longrightarrow \boxed{\text{Observe a new sample } C_n \text{ of } C^\theta(s^0)} \longrightarrow \boxed{\text{Update } \xi_n \text{ using } \frac{\partial v}{\partial \xi}(\xi, C_n)} \longrightarrow \xi_n$$

SSP simulation **GD Update**

- 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 $V'(\xi^*) = 0$

$$\xi_{n-1} \longrightarrow \boxed{\begin{array}{c}\textbf{Observe a new sample } C_n\\ \textbf{of } C^\theta(s^0)\end{array}} \longrightarrow \boxed{\begin{array}{c}\textbf{Update } \xi_n\\ \textbf{using } \frac{\partial v}{\partial \xi}(\xi, C_n)\end{array}} \longrightarrow \xi_n$$

SSP simulation **GD Update**

- 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 $C^\theta(s^0)$** → **Update ξ_n using $\frac{\partial v}{\partial \xi}(\xi, C_n)$** → ξ_n

SSP simulation | **GD Update**

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

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

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

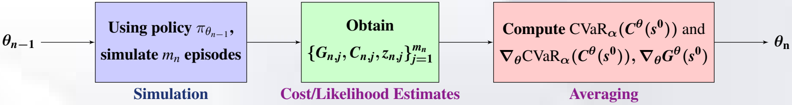


Figure: mini-batch idea

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

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

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

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

Conclusions

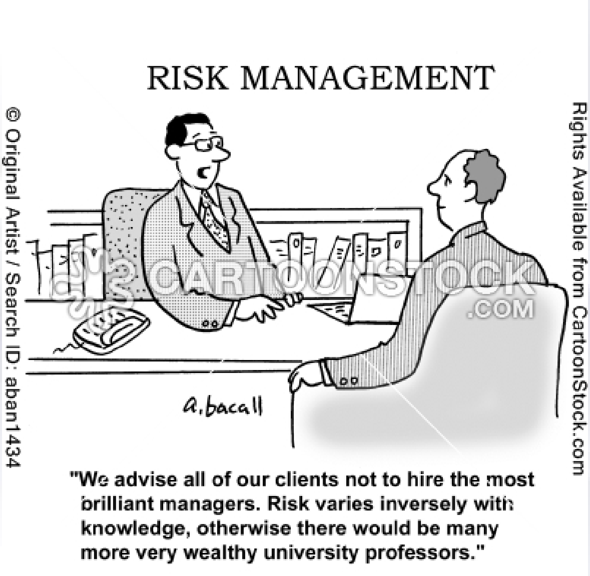
For ***stochastic shortest path*** problem, we

- defined ***CVaR*** as a ***risk measure***
- showed how to ***estimate*** both CVaR and its gradient
- proposed ***policy gradient algorithms*** to optimize the CVaR-constrained SSP
- established the ***asymptotic convergence*** of the algorithms
- adapted our algorithms to incorporate ***importance sampling*** for CVaR estimation

Future Work

- demonstrate the usefulness of our algorithms in a ***portfolio optimization*** application
- obtain finite-time bounds on the quality of solution of the policy gradient algorithms (esp. mini-batch - useful even for risk-neutral setting)

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."