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:

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

- 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

Risk-neutral Objective:

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

- 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

Total Cost

Cost

Policy
Risk-Sensitive Sequential Decision-Making

Risk-neutral Objective:

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

- 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

Risk-neutral Objective:

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

- 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:**

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

- 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)

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

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

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

Costs. \( g(s, a) \) and \( c(s, a) \)
Stochastic Shortest Path

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

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:

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

subject to (CVaR constraint):

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

minimize the total cost:

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

subject to (CVaR constraint):

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

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

\[ \Longleftrightarrow \]

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

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

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

**Three-Stage Solution:**

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

**next outer stage** Estimate \( \nabla_{\theta} L^{\theta, \lambda}(s^0) \) 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: \( \nabla_{\theta} L^{\theta, \lambda}(s^0) = \nabla_{\theta} G^{\theta}(s^0) + \lambda \nabla_{\theta} \text{CVaR}_\alpha(C^{\theta}(s^0)), \quad \nabla_{\lambda} L^{\theta, \lambda}(s^0) = \text{CVaR}_\alpha(C^{\theta}(s^0)) - K_\alpha \)
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 $\theta, \lambda (s^0)$, 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 $\mathcal{L}^{\theta,\lambda}(s^0)$.
Using policy $\pi_{\theta_n}$, simulate an SSP episode

Simulation

Estimate $\nabla_{\theta} G^{\theta}(s^0)$

Policy Gradient

Estimate CVaR$_{\alpha}(C^{\theta}(s^0))$

CVaR Estimation

Update $\theta_n$

Policy Update

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

CVaR Gradient

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

For any random variable $X$, let

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

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

Then,

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

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

---

Estimating CVaR: A convex optimization problem \(^2\)

For any random variable \(X\), let

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

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

Then,

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

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

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

- **Step-sizes**

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

  $$\xi_{n,1} = \left(1 - \frac{1}{1 - \alpha} \mathbb{1}_{\{c_n \geq \xi\}} \right)$$

- **Sample gradient**

  SSP simulation

  Update $\xi_n$

  Using $\frac{\partial v}{\partial \xi}(\xi, C_n)$

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

\[
\xi_n = \xi_{n-1} - \zeta_{n,1} 
\left( \frac{1}{1 - \alpha} \mathbb{1}_{\{c_n \geq \xi\}} \right)
\]

- **Step-sizes**

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

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

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

- **Sample gradient**
  \[
  \frac{1}{1 - \alpha} \mathbf{1}_{\{C_n \geq \xi\}}
  \]
**Observation:** to estimate VaR, one needs to find $\xi^*$ that satisfies $V'(\xi^*) = 0$

\[
\xi_{n-1} \quad \xrightarrow{\text{Observe a new sample } C_n \text{ of } C^\theta(s^0)} \quad \xi_n
\]

**SSP simulation**

**Update $\xi_n$ using**

\[
\frac{\partial v}{\partial \xi}(\xi, C_n)
\]

**GD Update**

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

- **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 $\text{CVaR}_\alpha(C^\theta(s^0))$\textsuperscript{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

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

\textsuperscript{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
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

$$\psi_n = \psi_{n-1} - \zeta_{n,2} (\psi_{n-1} - v(\xi_{n-1}, C_n))$$
Likelihood ratios for gradient estimation$^4$

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 \( X = (X_0, \ldots, X_{T-1})^T \)

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

---

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 \( X := (X_0, \ldots, X_{\tau-1})^T \)

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

---

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:

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

Policy gradient for the CVaR constraint \(^6\)

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

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:**

**VaR:** $\xi_n = \xi_{n-1} - \zeta_n,1 \left(1 - \frac{1}{1-\alpha} \mathbf{1}\{C_n \geq \xi_{n-1}\}\right)$, \quad **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: $\bar{C}_n = \bar{C}_{n-1} - \zeta_n,2 (C_n - \bar{C}_n)$, \quad Gradient: $\partial C_n = (\bar{C}_n - \xi_n) z_n \mathbf{1}\{C_n \geq \xi_n\}$

**Policy and Lagrange Multiplier Update:**

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

$\theta_{n-1} \rightarrow \theta_n$

Figure: mini-batch idea

VaR: $\xi_n = \frac{1}{m_n} \sum_{j=1}^{m_n} \left(1 - \frac{1}{C_{n,j}} \right)$, 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: $\partial C_n = (\bar{C}_n - \xi_n) z_n 1\{\bar{C}_n \geq \xi_n\}$. 
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
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

"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."