The Cross-Entropy Method

6.454 Seminar

17 September 2003

Guy Weichenberg

Outline

- Introduction.
- CE method for rare-event simulation (RES).
  - Example.
- CE method for combinatorial optimization problems (COPs).
  - Example.
- Enhancements.
- Convergence.
- Conclusions.

Introduction

- CE method broadened from a rare-event simulation technique to generic tool for solving different NP-hard problems.
- CE method a global iterative search algorithm comprising following two steps:
  1. Generate random data samples using set of dynamic parameters.
  2. Update parameters governing random data generation using data samples themselves with objective of improving future data samples.

CE Method for Rare-Event Simulation

Preliminaries

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ be random vector taking values in some space $\mathcal{X}$.
- Let S be real-valued function on $\mathcal{X}$.
- Let $f(\cdot; \mathbf{u})$ be probability density function on $\mathcal{X}$ parameterized by $\mathbf{u}$.

Suppose we are interested in probability of occurrence of a rare event. Specifically, we are interested in very small probability l that $S(\mathbf{x})$ is greater than or equal to real number γ under $f(\cdot; \mathbf{u})$:

$$l = \mathbb{P}_{\mathbf{u}}(S(\mathbf{X}) \geq \gamma) = \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geq \gamma\}}.$$

Crude Monte Carlo Simulation

- In CMC approach to estimating l, we simply draw random sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from $f(\cdot; \mathbf{u})$ and compute:

$$\frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \gamma\}}$$

  to arrive at unbiased estimate of l.

- For rare events, most terms in above summation will be zero, thus requiring very large value of N to obtain meaningful estimate of l.

Importance Sampling

- Using *importance sampling density* g, we represent l as:

$$l = \int I_{\{S(\mathbf{x}) \geq \gamma\}} \frac{f(\mathbf{x}; \mathbf{u})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = \mathbb{E}_g I_{\{S(\mathbf{X}) \geq \gamma\}} \frac{f(\mathbf{X}; \mathbf{u})}{g(\mathbf{X})}.$$

- Hence, unbiased estimator of l is:

$$\widehat{l} = \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \gamma\}} W(\mathbf{X_i}),$$

where $W(\mathbf{x}) = f(\mathbf{x}; \mathbf{u})/g(\mathbf{x})$ is *likelihood ratio* (LR), and $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are i.i.d. vectors drawn from g.

Importance Sampling (continued)

- Optimal importance sampling density is:

$$g^*(\mathbf{x}) = \frac{I_{\{S(\mathbf{x})\geq\gamma\}} f(\mathbf{x};\mathbf{u})}{l}$$

  $\rightarrow$ estimate has zero variance and only one sample is required.

- Problem with IS approach is obtaining g^*, as it depends on l which we are attempting to estimate. In addition, it is often convenient to choose importance sampling density of same form as $f(\cdot;\mathbf{u})$.

Importance Sampling (continued)

- When choosing importance sampling density of form $f(\cdot;\mathbf{u})$, we must find density $f(\cdot;\mathbf{v})$ which is "closest" to g^*. Convenient measure of "closeness" is *Kullback-Leibler (K-L) distance*:

$$\mathcal{D}(g,h) = \mathbb{E}_g \ln \frac{g(\mathbf{X})}{h(\mathbf{X})} = \int g(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x} - \int g(\mathbf{x}) \ln h(\mathbf{x}) d\mathbf{x}.$$

- Minimizing K-L distance between g^* and $f(\cdot;\mathbf{v})$ is equivalent to following maximization problem:

$$\max_{\mathbf{v}} \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X})\geq\gamma\}} \ln f(\mathbf{X};\mathbf{v}) = \max_{\mathbf{v}} \mathbb{E}_{\mathbf{w}} I_{\{S(\mathbf{X})\geq\gamma\}} W(\mathbf{X};\mathbf{u},\mathbf{w}) \ln f(\mathbf{X};\mathbf{v}),$$

where $W(\mathbf{x};\mathbf{u},\mathbf{w}) = f(\mathbf{x};\mathbf{u})/f(\mathbf{x};\mathbf{w})$ is the LR between $f(\cdot;\mathbf{u})$ and $f(\cdot;\mathbf{w})$ at $\mathbf{x}$.

Importance Sampling (continued)

- Maximizing $\mathbf{v}^*$ can be estimated by solving stochastic counterpart:

$$\mathbf{v}^* = \text{argmax}_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \gamma\}} W(\mathbf{X}_i; \mathbf{u}, \mathbf{w}) \ln f(\mathbf{X}_i; \mathbf{v}), \quad (1)$$

  where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are iid vectors drawn from $f(\cdot; \mathbf{w})$.

- When function to be maximized in (1) is differentiable w.r.t. $\mathbf{v}$, solution may be analytically obtained by solving following system of equations:

$$\frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \gamma\}} W(\mathbf{X}_i; \mathbf{u}, \mathbf{w}) \nabla \ln f(\mathbf{X}_i; \mathbf{v}) = \mathbf{0}. \quad (2)$$

Importance Sampling (continued)

- Approaches implied by (1) and (2) to estimating $\mathbf{v}^*$ only yield meaningful estimates when not too many of $I_{\{S(\mathbf{X}_i) \geq \gamma\}}$ terms are zero (i.e. when the event of interest is not rare).

CE Algorithm for Rare-Event Simulation

For rare events, we estimate $\mathbf{v}^*$ using following iterative algorithm in which $\mathbf{v}_t$ and γ_t associated with each iteration progressively approach $\mathbf{v}^*$ and γ, respectively:

1. Define $\widehat{\mathbf{v}}_o = \mathbf{u}$. Set $t = 1$.
2. Generate samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from density $f(\cdot; \mathbf{v}_{t-1})$ and compute sample $(1-\rho)$-quantile of $\widehat{\gamma}_t$ of performances (i.e. $\widehat{\gamma}_t = S_{(\lceil (1-\rho) \rceil)}$) provided $\widehat{\gamma}_t$ is less than γ. Otherwise, set $\widehat{\gamma}_t = \gamma$.
3. Use same sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to solve stochastic program:

$$\widehat{\mathbf{v}}_t = \text{argmax}_{\mathbf{v}} \; \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \widehat{\gamma}_t\}} W(\mathbf{X}_i; \mathbf{u}, \widehat{\mathbf{v}}_{t-1}) \ln f(\mathbf{X}_i; \mathbf{v}).$$

CE Algorithm for Rare-Event Simulation (continued)

4. If $\widehat{\gamma}_t < \gamma$, set $t = t + 1$ and reiterate from Step 2. Otherwise, proceed with Step 5.

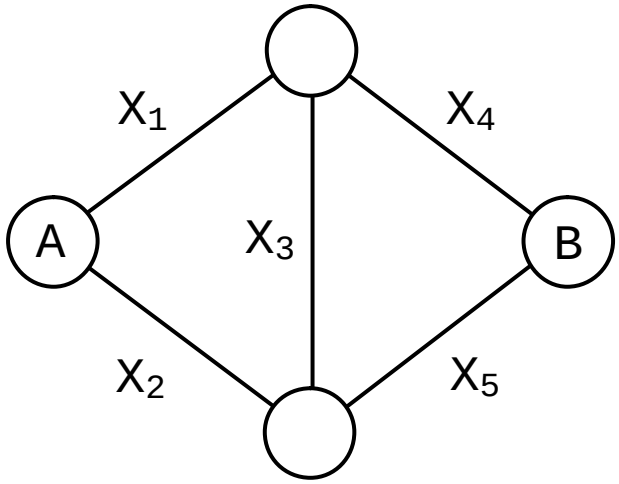
5. Estimate rare-event probability l using LR estimate:

$$\widehat{l} = \frac{1}{N_1} \sum_{i=1}^{N_1} I_{\{S(\mathbf{X}_i) \geq \gamma\}} W(\mathbf{X}_i; \mathbf{u}, \widehat{\mathbf{v}}_T),$$

where T denotes final number of iterations.

Shortest Path Example

Suppose the edge weights $\mathbf{X} = (X_1, \ldots, X_5)$ are independent and exponentially distributed with means $\mathbf{u} = (u_1, \ldots, u_5)$.



Let $S(\mathbf{X})$ be the shortest path length from A to B. We wish to estimate:

$$l = \mathbb{P}_{\mathbf{u}}(S(\mathbf{X}) \geq \gamma) = \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geq \gamma\}}.$$

Example (continued)

Step 1: Set $t = 1$, and the initial density to:

$$f(\mathbf{x}; \mathbf{v}_0) = \exp\left(-\sum_{j=1}^{5} \frac{x_j}{u_j}\right) \prod_{j=1}^{5} \frac{1}{u_j}$$

Step 2: Generate samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from the density:

$$f(\mathbf{x}; \mathbf{v}_{t-1}) = \exp\left(-\sum_{j=1}^{5} \frac{x_j}{v_{t-1,j}}\right) \prod_{j=1}^{5} \frac{1}{v_{t-1,j}}$$

Step 3: Solve the stochastic program:

$$\widehat{\mathbf{v}}_t = \text{argmax}_{\mathbf{v}} \; \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \widehat{\gamma}_t\}} W(\mathbf{X}_i; \mathbf{u}, \widehat{\mathbf{v}}_{t-1}) \ln f(\mathbf{X}_i; \mathbf{v}).$$

Example (continued)

where:

$$W(\mathbf{X}_i; \mathbf{u}, \widehat{\mathbf{v}}_{t-1}) = \exp\left(-\sum_{j=1}^{5} x_j \left(\frac{1}{u_j} - \frac{1}{\widehat{v}_{t-1,j}}\right)\right) \prod_{j=1}^{5} \frac{\widehat{v}_{t-1,j}}{u_j}$$

by solving the following set of equations (obtained from (2)):

$$\sum_{i=1}^{N} I_{\{S(\mathbf{x}_i) \geq \widehat{\gamma}_t\}} W(\mathbf{X}_i; \mathbf{u}, \widehat{\mathbf{v}}_{t-1}) \left(\frac{X_{ij}}{v_j^2} - \frac{1}{v_j}\right) = 0, \;\; j{=}1,\ldots,5.$$

Thus,

$$\widehat{v}_{t,j} = \frac{\sum_{i=1}^{N} I_{\{S(\mathbf{x}_i) \geq \widehat{\gamma}_t\}} W(\mathbf{X}_i; \mathbf{u}, \widehat{\mathbf{v}}_{t-1}) X_{ij}}{\sum_{i=1}^{N} I_{\{S(\mathbf{x}_i) \geq \widehat{\gamma}_t\}} W(\mathbf{X}_i; \mathbf{u}, \widehat{\mathbf{v}}_{t-1})}.$$

Example (continued)

Performance comparison:

- CMC: Using 10^8 samples, we obtain $\widehat{l} = 1.30 \times 10^{-5}$ with relative error $(\mathrm{Var}(\widehat{l})^{1/2}/\widehat{l})$ of 0.03 in 6350 seconds.
- CE: Using $N = 1000$, $N_1 = 10^5$ and $\rho = 0.1$, we obtain $\widehat{l} = 1.34 \times 10^{-5}$ with relative error of 0.03 in 3 seconds (5 iterations).

CE Method for Combinatorial Optimization

Preliminaries

- Let $\mathcal{X}$ be a finite set of states.
- Let S be a real-valued performance function we wish to maximize over $\mathcal{X}$.
- Let γ^* be maximum value of S.
- Let x^* be state at which this maximum occurs .
- Let $f(\cdot; \mathbf{v})$ be probability mass function on $\mathcal{X}$, parameterized by vector $\mathbf{v}$.

Recasting the COP

Main idea: Recast deterministic COP into probabilistic framework, where RES technique can be used.

- For certain parameter vector $\mathbf{u}$ and threshold γ, define $l(\gamma)$ as probability that performance function S exceeds γ:

$$l(\gamma) = \mathbb{P}_{\mathbf{u}}(S(\mathbf{X}) \geq \gamma) = \sum_{\mathbf{x}} I_{\{S(\mathbf{x}) \geq \gamma\}} f(\mathbf{x}; \mathbf{u}) = \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geq \gamma\}}.$$

- In typical COPs, $|\mathcal{X}|$ is very large and $l(\gamma^*) = f(\mathbf{x}^*; \mathbf{u}) \approx 1/|\mathcal{X}|$ is consequently very small $\rightarrow$ $l(\gamma^*)$ a *rare event*.

CE Algorithm for Combinatorial Optimization

1. Define $\widehat{\mathbf{v}}_o = \mathbf{u}$. Set $t = 1$.
2. Generate samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from density $f(\cdot; \mathbf{v}_{t-1})$ and compute sample $(1-\rho)$-quantile of $\widehat{\gamma}_t$ of performances (i.e. $\widehat{\gamma}_t = S_{(\lceil (1-\rho) \rceil)}$) provided $\widehat{\gamma}_t$ is less than γ. Otherwise, set $\widehat{\gamma}_t = \gamma$.
3. Use same sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to solve stochastic program:

$$\widehat{\mathbf{v}}_t = \text{argmax}_{\mathbf{v}} \; \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \widehat{\gamma}_t\}} \ln f(\mathbf{X}_i; \mathbf{v}). \tag{3}$$

4. If for some $t \geq d$:

$$\widehat{\gamma}_t = \widehat{\gamma}_{t-1} = \cdots = \widehat{\gamma}_{t-d},$$

then stop; otherwise, set $t = t + 1$ and reiterate from Step 2.

Remarks

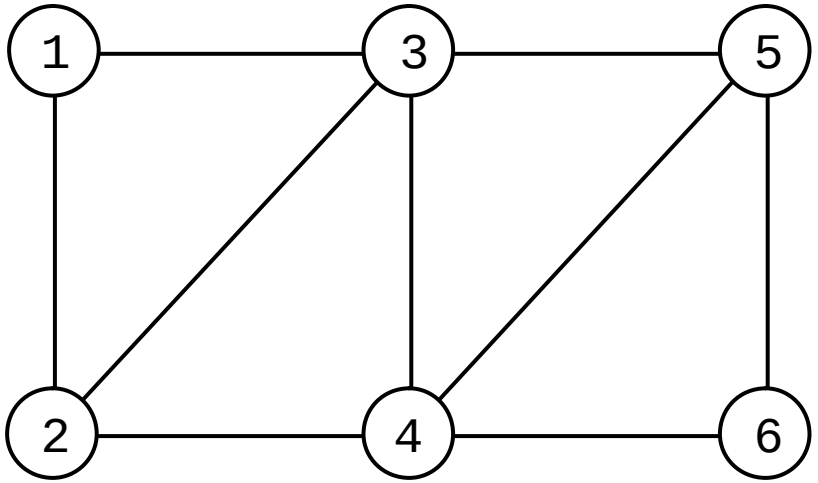
- No Importance Sampling step in algorithm.
- Significant difference between RES and COP Algorithms is role of initial parameter vector $\mathbf{u}$. In RES, $\mathbf{u}$ uniquely characterizes system under consideration, whereas in COP, $\mathbf{u}$ is arbitrarily chosen to initialize algorithm. LR term W is thus implicitly set to unity in COP because it is meaningless.
- Rather than updating $\widehat{\mathbf{v}}_t$ through (3), it is often beneficial to use following smoothing function to prevent harmful recurrences of zeros and ones in parameter vectors:

$$\widehat{\mathbf{v}}_t = \alpha \widehat{\mathbf{w}}_t + (1 - \alpha)\widehat{\mathbf{v}}_{t-1},$$

where $\widehat{\mathbf{w}}_t$ is the vector derived via (3).

Max-Cut Example

We wish to partition nodes in weighted graph into two subsets such that sum of weights from one subset to other is maximized.



Let $\mathbf{X} = (X_1, \ldots, X_6)$ be cut vector, where $X_i = 1$ if node i belongs to same partition as node 1, and 0 otherwise. Let $S(\mathbf{X})$ be the cost of cut $\mathbf{X}$. We wish to maximize S.

Example (continued)

Step 1: Arbitrarily set initial distribution of cut vectors to product of independent Bernoulli random variables $\mathbf{q} = (1, q_2, \ldots, q_6)$:

$$f(\mathbf{x}; \mathbf{p}_0) = \prod_{i=2}^{6} q_i^{x_i} \left(1 - q_i\right)^{1-x_i}$$

Step 2: Generate samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from the density:

$$f(\mathbf{x}; \mathbf{p}_{t-1}) = \prod_{i=2}^{6} p_{t-1,i}^{x_i} \left(1 - p_{t-1,i}\right)^{1-x_i}$$

Step 3: Solve the stochastic program:

$$\widehat{\mathbf{p}}_t = \text{argmax}_{\mathbf{p}} \; \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{x}_i) \geq \widehat{\gamma}_t\}} \ln f(\mathbf{X}_i; \mathbf{p}).$$

Example (continued)

by solving the following set of equations (obtained from (2)):

$$\frac{1}{(1-p_j)p_j}\sum_{i=1}^{N} I_{\{S(\mathbf{x}_i)\geq\widehat{\gamma}_t\}}(X_{ij}-p_j)=0,\ \ j{=}2,\ldots,6.$$

Thus,

$$\widehat{p}_{t,j}=\frac{\sum_{i=1}^{N} I_{\{S(\mathbf{x}_i)\geq\widehat{\gamma}_t\}}X_{ij}}{\sum_{i=1}^{N} I_{\{S(\mathbf{x}_i)\geq\widehat{\gamma}_t\}}}.$$

Enhancements

Alternative Performance Functions

- The CE method can be sped up by using $l(\gamma) = \mathbb{E}_{\mathbf{u}} \varphi(s;\gamma)$ instead of $l(\gamma) = \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geq \gamma\}}$ for updating $\mathbf{v}_t$, for some $\varphi(s;\gamma)$ which, for a maximization problem, is increasing in s.
- Empirical evidence shows that $\varphi = s I_{\{S(\mathbf{X}) \geq \gamma\}}$ speeds up algorithm, while high-power polynomials should be avoided as they lead more easily to local minima.

Fully Adaptive CE Method

- In FACE, parameters are updated using best performing constant N^e samples.
- Number of samples at each iteration N_t is obtained by attempting to satisfy:

$$\begin{aligned} S_{t,(N_t)} &\geq S_{t-1,(N_{t-1})} \\ \widehat{\gamma}_t &> \widehat{\gamma}_{t-1}. \end{aligned}$$

  i.e. Improvement in both the best and worst of the N^e samples in each iteration.

- Empirically, it was found that FACE converges up to two times faster than original CE algorithm.

Convergence

- Convergence of CE method to an estimate of (a possibly local) optimal CE parameter in a finite number of iterations with a finite sample size was shown in [Homem de Mello, 2002] under following assumptions:
  - Probability being estimated in CE method does not vanish in a neighborhood of optimal parameter $\mathbf{v}^*$.
  - ρ is adaptively decreased, and the sample size N is adaptively increased.
- In same work, authors showed that convergence to optimal CE parameter is exponential in sample size N.

Conclusions

- Summarized basic theory of CE method, and specialized method to RES and COPs.
- More work needs to be done. For example,:
  - Relationship of CE parameters to instance of problem.
  - Relationship of CE parameters to convergence (rates).