Outline

1. The Cross-Entropy Method
2. An Application to Risk Analysis
Importance Sampling (IS) Estimator for $\mu = E(h(X))$

- **Plain Monte Carlo:** Draw $n$ i.i.d. samples $X_1, \ldots, X_n$ from the target distribution $F(x)$, and use the estimator

$$\hat{\mu}_{\text{plain}} = \frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

- **Importance Sampling:** Instead of sampling from $F(x)$, draw $n$ i.i.d. samples $Y_1, \ldots, Y_n$ from an alternative distribution $G(x)$ that emphasizes “important values”, and use the estimator

$$\hat{\mu}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} h(Y_i) \frac{f(Y_i)}{g(Y_i)} \left( \text{or} \frac{1}{n} \sum_{i=1}^{n} h(Y_i) \frac{p(Y_i)}{q(Y_i)} \right),$$

where Radon-Nikodym derivative $w_i = \frac{f(Y_i)}{g(Y_i)}$ is used to correct the biasness.
How to Choose an Alternative Distribution $g(x)$?

The variance of the importance sampling estimator is

$$\text{var}(\hat{\mu}_{\text{IS}}) = \frac{1}{n} \text{var}_G \left( h(Y_i) \frac{f(Y_i)}{g(Y_i)} \right).$$

Ideally, if we could select an alternative distribution:

$$g^*(x) = \frac{1}{\mu} h(x) f(x), \quad x \in \mathbb{R},$$

then $\text{var}(\hat{\mu}_{\text{IS}}) = \frac{1}{n} \text{var}_{G^*}(\mu^{-1}) = 0$. But this requires the knowledge on the unknown parameter $\mu$. 
How to Choose an Alternative Distribution $g(x)$?

The variance of the importance sampling estimator is

$$\text{var}(\hat{\mu}_{IS}) = \frac{1}{n} \text{var}_G \left( h(Y_i) \frac{f(Y_i)}{g(Y_i)} \right).$$

Ideally, if we could select an alternative distribution:

$$g^*(x) = \frac{1}{\mu} h(x) f(x), \quad x \in \mathbb{R},$$

then $\text{var}(\hat{\mu}_{IS}) = \frac{1}{n} \text{var}_{G^*}(\mu^{-1}) = 0$. But this requires the knowledge on the unknown parameter $\mu$.

A Useful Idea

- Practically, we can find an alternative distribution $f_\theta(x)$ that is closest to $g^*(x)$ in some sense.
- The closeness is measured by some kind of “distance”.
The Cross-Entropy

Definition (Solomon Kullback and Richard Leibler, NSA, 1951)

Let $P$ and $Q$ denote two probability distributions with density functions $p(x)$ and $q(x)$ respectively. Kullback-Leibler Divergence or the cross entropy from $Q$ to $P$ is defined as

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} \left[ \log \frac{p(x)}{q(x)} \right] p(x) dx = E_P \left[ \log \frac{p(X)}{q(X)} \right].$$

- $P$ = “true” distribution; $Q$ = model distribution.
- $D_{KL}(P \parallel Q)$ is a measure of the information lost when $Q$ is used to approximate $P$.
- $D_{KL}(P \parallel Q)$ is not symmetric; that is,

$$D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P).$$

- If $P = Q$, then $D_{KL}(P \parallel Q) = 0.$
When $Q$ becomes closer to $P$, $D_{KL}(P \parallel Q)$ is smaller.
The Cross-Entropy Method (R. Rubinstein, 1997)

- \( p(x) = g^*(x) = \frac{1}{\mu} h(x)f(x) \) is the ideal distribution.
- \( q(x) = f_\theta(x) \) represents an alternative distribution to approximate \( g^*(x) \).
- The parameter \( \theta \) is chosen to minimize the information loss:
  \[
  \min_{\theta} D_{KL}(g^* \| f_\theta).
  \]
- Rewrite the KL divergence as follows
  \[
  D_{KL}(g^* \| f_\theta) = \int_{-\infty}^{\infty} g^*(x) \log g^*(x) \, dx - \frac{1}{\mu} \int_{-\infty}^{\infty} h(x)f(x) \log f_\theta(x) \, dx.
  \]

**Optimization Problem**

Select an alternative sampling distribution \( f_{\theta^*} \), where \( \theta^* \) is a maximizer

\[
\theta^* = \arg\max_\theta \int_{-\infty}^{\infty} h(x)f(x) \log f_\theta(x) \, dx = \arg\max_\theta E_f \left[ h(X) \log f_\theta(X) \right].
\]
Implementation

- Let $H(\theta) = E_f[h(X) \log f_\theta(X)]$. Solve
  
  $$H'(\theta) = E_f\left[h(X) \frac{\partial}{\partial \theta} \log f_\theta(X)\right] = 0$$

  for $\theta$. But it is in general difficult to solve this analytically. We solve the discretized version of this equation.

- Draw i.i.d. copies $X_1, \ldots, X_N$ from $f(x)$ (called pilot samples), and solve
  
  $$\frac{1}{N} \sum_{i=1}^{N} h(X_i) \frac{\partial}{\partial \theta} \log f_\theta(X_i) = 0$$

  for the optimal $\theta^*$. 

- If $f(x)$ is the density of the standard normal distribution $N(0, 1)$ and $f_\theta(x)$ is the density of $N(\theta, 1)$, then
  
  $$\theta^* = \frac{\sum_{i=1}^{N} h(X_i) X_i}{\sum_{i=1}^{N} h(X_i)}.$$
The Basic Cross-Entropy Algorithm

Goal: Estimate $\mu = E(h(X))$, where $X \sim f(x)$.

1. Generate $N$ independent pilot samples $x_1, \ldots, x_N$ from $f(x)$.
2. Obtain $\theta^*$ by solving
   \[
   \sum_{i=1}^{N} h(x_i) \frac{\partial}{\partial \theta} \log f_\theta(x_i) = 0.
   \]
3. Generate a sample $y$ from the alternative distribution $f_{\theta^*}(x)$.
4. Set
   \[
   H = h(y) \times \frac{f(y)}{f_{\theta^*}(y)}.
   \]
5. Repeat (3)-(4) $n$ times to obtain values $H_1, \ldots, H_n$, and set
   \[
   \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} H_i.
   \]
Remarks

- The pilot sample size $N$ is usually smaller than the IS simulation sample size $n$.

- If $h(t) = I\{X > t\}$ where $t$ is the threshold, then $\theta^*$ satisfies that

$$
\sum_{i=1}^{N} \frac{\partial}{\partial \theta} \log f_\theta(x_i) = 0, \text{ for } x_i > t, \forall i = 1, \ldots, N.
$$

That is, the cross-entropy method is closely related to the maximum likelihood estimation.

- Since the pilot samples are drawn from the original distribution $f(x)$, there could be very few samples or no sample in tail events $\{X > t\}$ for higher threshold $t$. As such, cross-entropy estimates for tail probabilities could be misleading.
Example: European Call Option (BLS_Call)

- Assume that the stock price follows a geometric Brownian motion:

\[ S_t = S_0 e^{(r - \sigma^2/2)t + \sigma B_t}, \quad t \geq 0, \]

where the risk-free interest rate is given by \( r \) and volatility \( \sigma \) is the (physical) parameter of the stock price process.

- \( X = (S_T - K)^+ = \max\{0, S_T - K\} \).

- Calculate \( v = E[e^{-rT}(S_T - K)^+] \), where

\[ S_T \sim \text{LogN} \left( \log S_0 + (r - \sigma^2/2)T, \sigma^2 T \right). \]

- The true values are calculated using Black-Scholes Formula:

\[ v = S_0 \Phi(\sigma \sqrt{T} - \theta) - Ke^{-rT} \Phi(-\theta). \]

\[ \theta = \frac{1}{\sigma \sqrt{T}} \log \frac{K}{S_0} + \left(\frac{\sigma}{2} - r\right) \sqrt{T}. \]
Example: European Call Option (cont’d)

- Estimate $v = E[h(Z)]$, where $Z \sim N(0, 1)$ and

$$h(u) = e^{-rT} (S_0 e^{(r-\sigma^2/2)T + \sigma \sqrt{T} z} - K)^+ \to 0, \text{ as } K \to \infty.$$ 

- To overcome the problem of no-sample when $K$ is large, we can use the importance sampling and the alternative distribution $g(x) \sim N(\theta^*, 1)$, and

$$\theta^* = \arg\max_x h(x)f(x),$$

where $f(x) = (2\pi)^{-1/2} e^{-x^2/2}, x \in \mathbb{R}$.

<table>
<thead>
<tr>
<th>K</th>
<th>Theoretical value</th>
<th>Estimate</th>
<th>S.E.</th>
<th>R.E.</th>
<th>$x^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>5.2253</td>
<td>5.2349</td>
<td>0.0247</td>
<td>0.47%</td>
<td>0.9849</td>
</tr>
<tr>
<td>60</td>
<td>1.6237</td>
<td>1.6310</td>
<td>0.0112</td>
<td>0.68%</td>
<td>1.4849</td>
</tr>
<tr>
<td>80</td>
<td>0.0795</td>
<td>0.0802</td>
<td>0.0008</td>
<td>1.00%</td>
<td>2.6002</td>
</tr>
<tr>
<td>100</td>
<td>0.0024</td>
<td>0.0024</td>
<td>3.0 x 10^{-5}</td>
<td>1.26%</td>
<td>3.6014</td>
</tr>
<tr>
<td>120</td>
<td>6.066 x 10^{-5}</td>
<td>5.957 x 10^{-5}</td>
<td>8.631 x 10^{-7}</td>
<td>1.45%</td>
<td>4.4569</td>
</tr>
</tbody>
</table>

**Figure:** Exponential Tilting: $S_0 = 50, r = 0.05, \sigma = 0.2, T = 1, n = 10000$
Example: European Call Option (cont’d)

Instead of obtaining $\theta$ numerically, we can estimate $\theta$ from pilot samples via the cross-entropy method.

Algorithm:

1. Generate $N$ independent pilot samples $x_1, \ldots, x_N$ from $N(0, 1)$.
2. Set
   \[
   \hat{\theta} = \frac{\sum_{i=1}^{N} h(x_i) x_i}{\sum_{i=1}^{N} h(x_i)}.
   \]
3. Generate a sample $y$ from the alternative distribution $N(\hat{\theta}, 1)$.
4. Set
   \[
   H = h(y) e^{-\hat{\theta} y + \hat{\theta}^2 / 2}.
   \]
5. Repeat (3)-(4) $n$ times to obtain values $H_1, \ldots, H_n$, and set
   \[
   \hat{\nu} = \frac{1}{n} \sum_{i=1}^{n} H_i.
   \]
Example: European Call Option (cont’d)

Figure: Cross-Entropy: $S_0 = 50, r = 0.05, \sigma = 0.2, T = 1, n = 10000$, NaN = Not a Number

<table>
<thead>
<tr>
<th></th>
<th>$K = 50$</th>
<th>$K = 60$</th>
<th>$K = 80$</th>
<th>$K = 100$</th>
<th>$K = 120$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True value</strong></td>
<td>5.2253</td>
<td>1.6237</td>
<td>0.0795</td>
<td>0.0024</td>
<td>$6.066 \times 10^{-5}$</td>
</tr>
<tr>
<td><strong>Estimate</strong></td>
<td>5.2166</td>
<td>1.6273</td>
<td>0.0805</td>
<td>0.0024</td>
<td>NaN</td>
</tr>
<tr>
<td><strong>S.E.</strong></td>
<td>0.0243</td>
<td>0.0109</td>
<td>0.0008</td>
<td>$3.2 \times 10^{-5}$</td>
<td>NaN</td>
</tr>
<tr>
<td><strong>R.E.</strong></td>
<td>0.47%</td>
<td>0.67%</td>
<td>0.99%</td>
<td>1.32%</td>
<td>NaN</td>
</tr>
<tr>
<td>$\hat{\theta}$</td>
<td>1.2057</td>
<td>1.7525</td>
<td>2.8247</td>
<td>3.3781</td>
<td>NaN</td>
</tr>
</tbody>
</table>

- In exponential tilting, $\theta^* = \arg \max_x h(x)f(x)$, which can be determined numerically. This is efficient for exponential families.
- In cross-entropy, $\hat{\theta} = \frac{\sum_{i=1}^{N} h(X_i)X_i}{\sum_{i=1}^{N} h(X_i)}$, which is estimated from pilot samples. This is useful for general distributions.
Remark

To overcome the problem of no pilot sample for rare events, one can use the iterative cross-entropy method that is based, again, on Radon-Nikodym derivatives.

The difficult task in the cross-entropy method is to solve

\[ E_f \left[ h(X) \frac{\partial}{\partial \theta} \log f_\theta(X) \right] = \int_{-\infty}^{\infty} \left[ h(x) \frac{\partial}{\partial \theta} \log f_\theta(x) \right] f(x) dx = 0. \]

Define the Radon-Nikodym derivative

\[ w(x) = \frac{f(x)}{f_{\theta'}(x)}, \quad (\theta' \text{ can be different from } \theta). \]

The equation becomes

\[ \int_{-\infty}^{\infty} \left[ h(x) w(x) \frac{\partial}{\partial \theta} \log f_\theta(x) \right] f_{\theta'}(x) dx = 0. \]
Iterative Cross-Entropy Method

Iterative Idea:

1. Suppose that at stage $k$, we has estimated the parameter $\theta'$.
2. At stage $k + 1$, draw pilot samples $Y_1, \ldots, Y_N$ from $f_{\theta'}$, and obtain $\theta$ by solving
   \[
   \frac{1}{N} \sum_{i=1}^{N} h(y_i)w(y_i) \frac{\partial}{\partial \theta} \log f_{\theta}(y_i) = 0.
   \]
3. Repeat (1)-(2) to obtain a sequence of estimates $\theta_1, \ldots, \theta_k$ iteratively.
4. Under some regularity conditions, $\theta_k$ converges to $\theta^*$, as $k \to \infty$. Use the limit $\theta^*$ as the optimal parameter in the importance sampling.
Example (de Boer, Kroese, Mannor, Rubinstein, 2005)

- Let $\mathcal{X}$ denote a set of finite states, and $S : \mathcal{X} \to \mathbb{R}$, the performance function. The goal is to find a maximizer $x^*$ such that
  \[ S(x^*) = \gamma^* = \max_{x \in \mathcal{X}} S(x) \]

- Convert this deterministic problem into a stochastic problem. Let $f(x)$ be a PMF defined on $\mathcal{X}$, and $X \sim f(x)$. Estimate the tail probability using the cross-entropy method
  \[ \nu = P(S(X) > \gamma) = E[I\{S(X) > \gamma\}] . \]

  When $\gamma$ is close to $\gamma^*$, $P(S(X) > \gamma)$ becomes very small.

- Use the cross-entropy method, obtain
  \[ \theta^* = \arg\max_{\theta} E_f[I\{S(X) > \gamma\} \log f_{\theta}(X)] , \]

  where $f_{\theta}(x)$ is the alternative distribution.

- When $\gamma$ is close to $\gamma^*$, $f_{\theta^*}(x)$ assigns most of its probability mass close to the optimal solution $x^*$, and thus can be used to generate an approximate optimal solution.
Remark

The cross-entropy method is a generic approach to

- rare event simulation (e.g., estimation of default probability, or rare events in complex stochastic networks);
- combinatorial optimization (e.g., DNA sequence alignment);
- continuous multi-extremal optimization (e.g., protein structure prediction, molecular dynamics simulation).
Two Popular Risk Measures

- Value-at-Risk (VaR) at level $\alpha$, defined by
  \[ \text{VaR}_\alpha(X) := \min\{x \in \mathbb{R} : P(X > x) = \alpha\}. \]

- Conditional VaR or Expected Tail Loss at level $\alpha$ is defined by
  \[ \text{CVaR}_\alpha(X) := E(X \mid X > \text{VaR}_\alpha(X)). \]
Let $L$ denote total loss with loss distribution $f(x)$. Then
\[ E(L|L > x) = \frac{E(I\{L > x\}L)}{P(L > x)} \]

$L_1, \ldots, L_n$ are drawn from an alternative distribution $g(x)$ and $w_1, \ldots, w_n$ are corresponding likelihood ratios.

Two unbiased estimators:
\[ \frac{1}{n} \sum_{i=1}^{n} I\{L_i > x\}w_i \rightarrow P(L > x) \]
\[ \frac{1}{n} \sum_{i=1}^{n} I\{L_i > x\}L_iw_i \rightarrow E(I\{L > x\}L). \]

An IS estimator for $E(L|L > x)$ is given by
\[ \frac{\sum_{i=1}^{n} I\{L_i > x\}L_iw_i}{\sum_{i=1}^{n} I\{L_i > x\}w_i} \rightarrow \frac{E(I\{L > x\}L)}{P(L > x)} = E(L|L > x). \]