Efficient Rare Event Simulation: A Tutorial on Importance Sampling

Michele Pagano*, Werner Sandmann†

Dept. Information Systems and Applied Computer Science, University of Bamberg, Germany
Dept. of Information Engineering, University of Pisa, Italy

HET-NETs ’05
July, 2005

Key Topics in Rare Event Simulation

- Relevance of Rare Events on network performance
- The Basic Problem of Rare Event Simulation
- Variance Reduction Techniques
  - RESTART and Importance Sampling
- Importance Sampling: Theory
  - Basic Principle and Definitions
  - Efficiency Criteria
  - The core of IS: the Change of Measure
- Importance Sampling: Applications
  - A case study: M/M/1 queue
  - Tandem queues
  - Advanced approaches: The Cross Entropy Method
- Concluding Remarks
First Words on Rare Event Simulation

Rare Events

- influence many real-world systems,
- may cause serious consequences,
- should occur with extremely small probability (e.g. $< 10^{-9}$),
- occur rarely in stochastic simulations, too,
- are therefore difficult to simulate.

Reliable statistics require sufficiently large number of observations: for probability $10^{-9}$ on average $10^9$ trials for one observation.

⇒ Direct simulation lasts days, weeks, months, years, lifetimes!

In principle, that’s the whole story...

Thank you for your attention!

Application Areas

- Nuclear Physics, e.g. atomic accident
- Security systems, e.g. false alarms in radar
- Technical defects, e.g. aircraft, spacecraft
- Mathematical Finance and Insurance Risk, e.g. ruins
- Manufacturing Systems, e.g. breakdowns

- Computer and Communication Systems
  - bit errors in digital communications
  - component and system failures and breakdowns (fault-tolerant systems, reliability, availability)
  - Queueing Systems
  - excessive backlogs, waiting times, delays
  - buffer overflows → packet loss, cell loss
Rare Events informally

Informal Characterization

- Rare events occur with small probability.

Immediate Questions

- What is a small probability?
- Is there a probability threshold such that all less probable events are called rare?
- Does only an event's probability determine its relevance?
- Are all events with small probability of practical interest?
- Is there a formal mathematical characterization of rare events?

Rough Answers

- There is no fixed probability threshold for characterizing rarity.
- Rare events of practical interest often depend on system parameters.
- Probability becomes asymptotically small with parameter changes.
- Rare events are often defined on tails of probability distributions.

Efficient Rare Event Simulation: A Tutorial on Importance Sampling

Rare Events formally?

Simple Example: stable single-server queueing system

- The probability of more than \( n \in \mathbb{N} \) jobs in the system converges to zero for \( n \to \infty \).
- The probability of waiting time greater than \( x \in \mathbb{R}^+ \) converges to zero for \( x \to \infty \).
- Convergence speed depends on system utilization.

Waiting Time in a M/M/1 queue:

\[
P(W > x) = \frac{\lambda}{\mu} e^{-(\mu-\lambda)x}
\]

Rare events are characterized by large deviation from normality

The mathematical theory related to rare events is known as Large Deviations Theory

Large Deviation Principle for the Waiting Time Distribution

\[
\lim_{x \to \infty} \frac{1}{x} \log P(W > x) = -\delta \quad \text{where} \quad \delta = \sup_{\theta} \{ \theta : \Lambda(\theta) < 0 \} = \inf_{y} \frac{\Lambda^*(y)}{y}
\]

or, roughly speaking:

\[
P(W > x) \approx e^{-\Delta x}
\]
The Problem of Rare Event Simulation

Given a random variable $X$ with distribution $P$, estimate by direct simulation

$$\gamma := \mathbb{P}\{A\} = \mathbb{E}_P[I_A] \quad \text{for some event } A$$

Generate samples $X_1, X_2, \ldots, X_N$, iid as $X$ according to distribution $P$,

$$\hat{\gamma} = \frac{1}{N} \sum_{i=1}^{N} I_A(X_i) \quad \text{(unbiased sample mean)},$$
$$\sigma^2(\hat{\gamma}) = \text{Var}[\hat{\gamma}] = \frac{\gamma(1-\gamma)}{N} \quad \text{(variance of sample mean)}.$$

Confidence interval:

$$\hat{\gamma} \pm z_{1-\alpha/2} \sqrt{\frac{\gamma(1-\gamma)}{N}}$$

Relative error (Estimate is good, if absolute error is small compared to absolute value):

$$\delta_{rel}(\hat{\gamma}) = \frac{\sigma(\hat{\gamma})}{\mathbb{E}[\hat{\gamma}]} = \sqrt{\frac{1-\gamma}{\gamma N}} \quad \gamma \to 0 \implies \infty$$

Graphical Interpretation of the Confidence interval

Confidence interval for $\gamma$:

$$\hat{\gamma} \pm z_{1-\alpha/2} \sqrt{\frac{\gamma(1-\gamma)}{N}}$$

$z_{1-\alpha/2}$ is defined by the equation $\mathbb{P}\left(\mathcal{N}(0, 1) \geq z_{1-\alpha/2}\right) = \alpha/2$

$\mathcal{N}(0, 1)$ denotes a normally distributed random variable with zero mean and variance one.
Direct Rare Event Simulation – Sample Size

Requirements in evaluating the confidence interval for \( \gamma \):

- confidence level \( 1 - \alpha \)
- maximum relative half-width \( \beta \)

What sample size is needed?

\[
z_{1-\alpha/2} \sqrt{\frac{1 - \gamma}{\gamma N}} \leq \beta \quad \Rightarrow \quad N \geq \frac{z_{1-\alpha/2}^2}{\beta^2} \cdot \frac{1 - \gamma}{\gamma} \quad \gamma \to 0 \to \infty.
\]

Numerical Example

- 99% confidence interval \( \Rightarrow \alpha = 0.01 \) and \( z_{1-\alpha/2} = 2.576 \)
- maximum relative half-width of 10% \( \Rightarrow \beta = 0.1 \)

\[
\Rightarrow N \geq 100 \cdot 2.576^2 \cdot \frac{1 - \gamma}{\gamma}
\]

- for instance, if \( \gamma = 10^{-9} \Rightarrow N \geq 6.64 \cdot 10^{11} \)

Efficient Rare Event Simulation: A Tutorial on Importance Sampling

Samples in Network Models

- In practical situations (realistic models) one sample may include generation of millions or billions, or trillions, or ... of random numbers!
- Consider a single server queue simulation, simulation of \( n \) jobs

  - One single sample requires
    - generation of \( n \) (inter)arrival times and \( n \) service times
    - altogether \( 2n \) non-uniform random variates
  - Steady-state simulations require a very large number of jobs to simulate
  - Altogether in \( N \) runs \( 2 \cdot n \cdot N \) non-uniform random variates.
  - For example 99\% confidence interval for a probability of about \( 10^{-12} \) with 1 million jobs in each run: \( 1.38 \cdot 10^{21} \) random variates!

- Imagine what you would need for a network simulation!

  - \( k \) service times per job for \( k \) nodes, additionally routing probabilities

  \( \Rightarrow \) Direct rare event simulation is impracticable!
Simulation Speed-Up

Stochastic simulations are statistical estimations.

By simulation speed-up it is meant that the time to determine statistical estimates of desired accuracy (confidence interval half-width, relative error) is significantly reduced.

Basically two different types of approach

- make more experiments in same time
- need less experiments for desired accuracy

Speed-up techniques overview

- **Parallel and Distributed Simulation**: exploits a multiprocessor environment
  - process distribution (synchronization?)
  - simulation replica
- **Hybrid techniques**: combine analytic results with simulation
  - decomposition (into independent sub-models – in time or space)
  - conditional sampling
- **Variance reduction by use of correlation**: exploits a known correlation in input and output samples
  - Antithetic Variates
  - Common Random Numbers
  - Control Variates
  - Stratified Sampling
- **Rare Event Provoking Techniques**: increase the frequency of the rare event of interest
  - Importance Splitting (RESTART)
  - Importance Sampling
RESTART (REpetitive Simulation Trials After Reaching Thresholds) exploits the idea of sampling the rare event from a reduced state space, that includes the event of interest.

1991: REST ART was first introduced by Villén-Altamirano in its one threshold version.

1994: multiple thresholds version of REST ART.

- **Importance Sampling in a Nutshell**
  
  Importance Sampling (IS) is a general variance reduction technique, not limited to rare events.

  **Idea**
  
  *Simulation Speed–up by Rare Event Provoking*
  
  (generate more rare events during simulation, in same time)

  Simulation method based on IS:
  
  - modify underlying stochastics, e.g. interarrival or service time distributions, component failure rates, densities, transition probabilities etc. *(Change of Measure, Biasing)*
  - perform simulation under modified probability measure
  - unbias results by correcting factor, the Likelihood Ratio

  When applied *properly*, enormous variance reduction (several orders of magnitude) can be obtained.
  
  If not, variance may even grow infinitely.

  **The Art of Importance Sampling**
  
  How to perform the change of measure?
Objective: determine the area of region B

- The analytical solution would require a mathematical description of the boundary of B as well as a complex integration procedure.
- When this knowledge is not obtainable, computer simulation using the Monte Carlo (MC) method is one alternative:
  - Generate \( N \) statistically independent random samples, uniformly distributed over the entire space \( A \)
  - Estimate the area of \( B \) as \( \hat{B} = N_B / N \), where \( N_B \) is the number of hits within region \( B \)
  - The variance of this estimate is inversely proportional to \( N \), while the precision of the estimate is related to the number of hits in the important region.

Objective: determine the area of region C

- Using MC simulation, much larger number of samples, \( \bar{N} \), would have to be generated for an equivalent estimator variance.
- Using Importance Sampling, we would bias the sampling procedure to increase the fraction of samples that result in hits.
  - Double the probability that samples are generated in the quadrant \( D \), containing \( C \)
  - The average number of samples in region \( C \) is doubled, increasing the estimator precision
  - Each sample which results in a hit in region \( D \) must be weighted by a factor of 1/2 to yield the correct, statistically unbiased result.
The art of Importance Sampling

Ensuring that the regions in the space with increased sampling frequency include the important region. This can be a problem in case of insufficient prior knowledge of system behavior.

Objective: determine the area of region $E$

If the region of interest is actually $E$, the biased scheme used here reduces the number of hits and the corresponding estimator precision by a factor of two (the weight of each hit is 2!).

Given:

- a random variable (RV) $X$ with density $f$
- a real-valued function $g$
- another density $f^*$, such that

$$g(x)f(x) > 0 \Rightarrow f^*(x) > 0$$

Likelihood Ratio of $f$ and $f^*$

$$L(x) := \begin{cases} \frac{f(x)}{f^*(x)}, & \text{if } f^*(x) \neq 0, \\ 0, & \text{otherwise,} \end{cases}$$

Then

$$\mathbb{E}_f[g(X)] = \int g(x)f(x)dx = \int g(x)L(x)f^*(x)dx = \mathbb{E}_{f^*}[g(X)L(X)]$$
Importance Sampling Estimator

- Samples $X_1, \ldots, X_N$, iid, according to $f^*$, (e.g. generated by simulation)
- The Importance Sampling Estimator
  \[ \hat{\gamma}_{IS} := \frac{1}{N} \sum_{i=1}^{N} g(X_i) L(X_i) \]
  is an unbiased estimator for $\gamma := \mathbb{E}_f[g(X)]$
- For estimating probabilities of an event $A$
  - $g$ is the indicator function of the event $A$
  - $I_A(x) f(x) > 0 \Rightarrow f^*(x) > 0$ is required
  - The Importance Sampling Estimator
    \[ \hat{\gamma}_{IS} := \frac{1}{N} \sum_{i=1}^{N} I_A(X_i) L(X_i) \]
    is an unbiased estimator for $\gamma := \mathbb{P}(A) = \mathbb{E}_f[I_A(X)]$

General Mathematical Basis of Importance Sampling

- Importance Sampling is not limited to real-valued continuous random variables!
- Formally, in measure-theoretic terms, Importance Sampling in general is based on an application of the Radon-Nikodym theorem, and the likelihood ratio is what is known as the Radon-Nikodym derivative
- For some arbitrary RV $H$ and probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on a measurable space $(\Omega, \mathcal{A})$
  \[ \mathbb{E}_{\mathbb{P}}[H] = \int H(\omega) d\mathbb{P} = \int H(\omega) L(\omega) d\mathbb{Q} = \mathbb{E}_{\mathbb{Q}}[HL] \]
  where the likelihood ratio is
  \[ L(\omega) = \frac{d\mathbb{P}}{d\mathbb{Q}} \]
- In particular, defining $H(\omega) := I_A(\omega)$ yields for the probability of each $A \in \mathcal{A}$
  \[ \mathbb{P}(A) = \mathbb{E}_{\mathbb{P}}[I_A] = \int I_A(\omega) d\mathbb{P} = \int I_A(\omega) L(\omega) d\mathbb{Q} = \mathbb{E}_{\mathbb{Q}}[I_A L] \]
- In Importance Sampling the probability measure $\mathbb{Q}$ is called the Importance Sampling Measure, and the corresponding density is called the Importance Sampling Density
Remarks on absolute continuity

- The existence of the Radon-Nikodym derivative $L(\omega)$ requires that the measure $\mathbb{P}$ be absolutely continuous with respect to the measure $\mathbb{Q}$, i.e.
  \[ \forall A \in \mathcal{A} : \mathbb{Q}(A) = 0 \Rightarrow \mathbb{P}(A) = 0 \]
  which is equivalent to:
  \[ \forall A \in \mathcal{A} : \mathbb{P}(A) > 0 \Rightarrow \mathbb{Q}(A) > 0 \]

- The condition of absolute continuity allows that $\mathbb{Q}(A) > 0$ if $\mathbb{P}(A) = 0$

- The probability measure $\mathbb{Q}$ may assign positive probability to events that are impossible under probability measure $\mathbb{P}$

- Condition $g(x)f(x) > 0 \Rightarrow f^*(x) > 0$ means absolute continuity of probability measures

Informal Discussion of Special Cases

- The probabilistic setting to which IS applies is extremely general
  - real-valued one-dimensional random variables with densities $f$ and $f^*$
  - discrete random variables with probability distributions $\mathbb{P}$ and $\mathbb{P}^*$:
    - Set $f(x) = \mathbb{P}(X = x)$ and $f^*(x) = \mathbb{P}^*(X = x)$
  - random vectors:
    - densities $f$ and $f^*$ are multidimensional, real-valued or discrete similar to above. Arguments $x$ are vectors.
  - Markov chains:
    - density $f$ corresponds to probability distribution of Markov chain path probabilities
    - Importance Sampling density $f^*$ corresponds to probability distribution of path probabilities, not necessarily Markovian.
    - for instance, for a DTMC with initial distribution $p_0$ and transition probabilities $P(i, j)$:
      \[ L(X_0, X_1, \ldots, X_m) = \frac{p_0(X_0)}{p_0^*(X_0)} \prod_{i=1}^{m} \frac{P(X_{i-1}, X_i)}{P^*(X_{i-1}, X_i)} \]
The efficiency of an unbiased estimator is determined by its variance.

It is easy to show that

\[
\text{Var}[\hat{\gamma}_{IS}] = \text{Var}\left[\frac{1}{N} \sum_{i=1}^{N} g(X_i) L(X_i) \right] = \frac{1}{N} \left( \mathbb{E}_{f^*} \left[ (g(X)^2 L(X)^2) - \gamma^2 \right] \right)
\]

Variances are nonnegative, minimum possible variance is zero.

Optimal zero-variance Importance Sampling estimator always exist, since

\[
\frac{1}{N} \mathbb{E}_{f^*_{\text{opt}}} \left[ (g(X)L(X) - \gamma)^2 \right] = 0 \implies f^*_{\text{opt}}(x) = \frac{g(x)f(x)}{\gamma}
\]

Unfortunately

- \(f^*_{\text{opt}}\) depends explicitly on the unknown \(\gamma\) \(\implies\) generally not available.
- if available, requires sampling from conditional density.
- \(f^*_{\text{opt}}\) often belongs to different class of models/measures.

The Special Case of Estimating Probabilities

If \(\gamma = \mathbb{P}(A)\), then \(g = I_A\) and

\[
\text{Var}[\hat{\gamma}_{IS}] = \frac{1}{N} \left( \mathbb{E}_{f^*} \left[ I_A(X) L(X)^2 \right] - \gamma^2 \right)
\]

By direct substitution in the general expression,

\[
f^*_{\text{opt}}(x) = \frac{I_A(x)f(x)}{\gamma} = \begin{cases} 
\frac{f(x)}{\gamma} & \text{if } x \in A, \\
0 & \text{otherwise}.
\end{cases}
\]

The optimal change of measure is the ordinary distribution, conditioned that the rare event has occurred.

Aim:

Find a good change of measure, resulting in Importance Sampling estimators with small variance.
Efficiency Criteria

- Relative error of an estimator
  - defined as ratio of standard deviation and expectation
  - directly proportional to relative half-width of confidence intervals
  - therefore criterion for efficiency of estimators

- Relative error for direct simulation
  \[ \delta_{rel}(\hat{\gamma}) = \frac{\sigma(\hat{\gamma})}{E[\hat{\gamma}]} = \sqrt{\frac{1 - \gamma}{\gamma N}} \]  \( \gamma \to \infty \)

- Relative error for Importance Sampling
  \[ \delta_{rel}(\hat{\gamma}_{IS}) = \frac{\sqrt{\text{Var}[\hat{\gamma}_{IS}]} \sqrt{E[\hat{\gamma}_{IS}]} = \frac{\sqrt{\text{E}_{f^*}[g(X)^2L(X)^2] - \gamma^2}}{\gamma \sqrt{N}} }{E[\hat{\gamma}_{IS}]} \]
  - depends on \( E_{f^*}[g(X)^2L(X)^2] \), particularly on the likelihood ratio.
  - the likelihood ratio significantly influences efficiency
  - something can/may be done against convergence to infinity

Bounded Relative Error

- Let \( \gamma \) depend on a rarity parameter \( m > 0 \), such that the larger \( m \) the smaller \( \gamma \)
  \[ \lim_{m \to \infty} \gamma(m) = 0 \]
  - for instance rarity of buffer overflow grows with buffer capacity \( m \)
  - in principle, \( f^* \) may depend on \( m \), but usually does not

- The family of estimators \( \hat{\gamma}_{IS}(m) \) or, short, the estimator \( \hat{\gamma}_{IS} \) has bounded relative error (BRE), if there exists a constant \( c > 0 \), such that
  \[ \lim_{m \to \infty} \delta_{rel}(\hat{\gamma}_{IS}(m)) \leq c < \infty. \]

Interpretation of BRE
Relative Error remains bounded even if \( \gamma \) goes to zero.
Asymptotic Optimality

☞ As variances are nonnegative,
\[ \mathbb{E}_{f^*}[g(X)^2L(X)^2] \geq \gamma(m)^2 \Rightarrow \frac{\ln \mathbb{E}_{f^*}[g(X)^2L(X)^2]}{\ln \gamma(m)} \leq 2. \]

☞ The family of estimators \( \hat{\gamma}_{IS}(m) \) or, short, the estimator \( \hat{\gamma}_{IS} \) is called asymptotically optimal (AO), if
\[ \lim_{m \to \infty} \frac{\ln \mathbb{E}_{f^*}[g(X)^2L(X)^2]}{\ln \gamma(m)} = 2. \]

An Importance Sampling estimator for \( \gamma(m) \) is asymptotically optimal iff \( \gamma(m) \) converges faster to zero than \( \delta_{rel}(\hat{\gamma}_{IS}(m)) \) converges to infinity:
\[ \lim_{m \to \infty} \frac{\delta_{rel}(\hat{\gamma}_{IS}(m))}{\gamma(m)} = \lim_{m \to \infty} \delta_{rel}(\hat{\gamma}_{IS}(m)) \gamma(m) = 0. \]

☞ If \( \gamma(m) \) converges exponentially fast to zero for \( m \to \infty \) (as seen in Large Deviation Theory) and the Importance Sampling estimator has polynomial (i.e., polynomially increasing to infinity) relative error, then the estimator is asymptotically optimal.

Efficient Rare Event Simulation: A Tutorial on Importance Sampling

Relationship between AO and BRE

☞ Example for Asymptotic Optimality

⇒ \( \gamma(m) \) converges exponentially fast to zero, i.e.
\[ \gamma(m) = e^{-dm} \quad \text{for some } d > 1 \]

⇒ Asymptotic optimality means
\[ \lim_{m \to \infty} \frac{1}{m} \ln \mathbb{E}_{f^*}[g(X)^2L(X)^2] = -2d. \]

☞ Each IS estimator with bounded relative error is asymptotically optimal, i.e.

\textbf{BRE implies AO}

☞ There exist asymptotically optimal IS estimators not having bounded relative error, i.e.

\textbf{AO does not imply BRE}

Asymptotic optimality is a strictly weaker criterion than bounded relative error.
Historically Importance Sampling often used for tail probabilities

Typical application: bit error rates in digital communications

Scaling: more probability mass in the tails

\[ f^*(x) = \frac{1}{\alpha} f\left(\frac{x}{\alpha}\right) \quad \alpha \in \mathbb{R}^d \]

For complex systems of high dimensionality scaling in each dimension can be contraproductive since it does not generate more error events

Translation: Shift expectation to error region

\[ f^*(x) = f(x - T), \quad T \in \mathbb{R}^d \]

Translation is more system dependent and more difficult to apply than scaling

There is an obvious choice of \( T \), based on the most likely path to error

\( \Rightarrow \) a central idea in Importance Sampling

\( \Rightarrow \) for normal distribution equivalent to exponential change of measure

Exponential Change of Measure

Exponential Change of Measure (ECM), aka Exponential Twisting, Exponential Tilting

very common proof technique in LDT (e.g. lower bound in Cramér theorem)

Most popular change of measure technique in Importance Sampling for rare events in queueing systems over the last two decades

Basic idea:

(\( \Rightarrow \) restrict potential Importance Sampling densities/measures/distributions to a parametric family/class

(\( \Rightarrow \) determine optimal change of measure within this restricted class (Optimal Exponential Change of Measure, OECM)

Given a RV \( X \) with density \( f \) and moment generating function

\[ M(\vartheta) = \mathbb{E}_f [e^{\vartheta X}] = \int e^{\vartheta x} f(x) dx, \quad \vartheta \in \mathbb{R}^d \]

the exponentially twisted (or tilted) density \( f^*(x) \) with twisting (or tilting) parameter \( \vartheta \) is defined by

\[ f^*(x) := \frac{e^{\vartheta x} f(x)}{M(\vartheta)} \]
Likelihood Ratio and Moment Generating Function under ECM

\(L(x) = \frac{f(x)}{f^*(x)} = \frac{f(x)}{M(\vartheta) e^{\vartheta x}} = \frac{M(\vartheta)}{e^{\vartheta x}} = M(\vartheta) e^{-\vartheta x}\)

Moment Generating Function \(M^*\) of \(X\) according to the twisted density \(f^*\)

\[M^*(\eta) = \mathbb{E}_{f^*}[e^{\eta X}] = \frac{1}{M(\vartheta)} \mathbb{E}_f[e^{(\eta + \vartheta)X}] = \frac{M(\eta + \vartheta)}{M(\vartheta)}\]

ECM for some Distributions

\[\text{Exp}(\lambda) \rightarrow \text{Exp}(\lambda - \theta)\]
\[\Gamma(\lambda, \beta) \rightarrow \Gamma(\lambda - \vartheta, \beta)\]
\[N(\mu, \sigma^2) \rightarrow N(\mu + \vartheta, \sigma^2)\]
\[\text{Geo}(p) \rightarrow \text{Geo}(1 - (1 - p)e^\vartheta)\]

ECM for Sums of iid Random Variables

Many interesting properties of queueing models can be expressed in terms of sums of iid random variables

Let \(X_1, \ldots, X_n\) iid real-valued random variables and \(S_n := X_1 + \cdots + X_n\)

Moment generating function \(M_{S_n}\) of the sum \(S_n\):

\[M_{S_n}(\vartheta) = \mathbb{E}[e^{\vartheta S_n}] = \mathbb{E}[e^{\vartheta(X_1+\cdots+X_n)}] = \left(\mathbb{E}[e^{\vartheta X}]\right)^n = (M_X(\vartheta))^n.\]

Moment generating function according to exponentially twisted density of sum

\[M^*_{S_n}(\eta) = M^*_X(\eta)^n = \left(\frac{M_X(\eta + \vartheta)}{M_X(\vartheta)}\right)^n\]

Likelihood ratio of original and exponentially twisted density of sum

\[L(x_1, \ldots, x_n) = M_{S_n}(\vartheta) e^{-\vartheta s_n} = M_X(\vartheta)^n e^{-\vartheta s_n}\]
Queueing Models and Random Walks

Interpretation: The sequence of sums \((S_n)\) is a random walk with negative drift, i.e. with independent increments \(X_i\), where \(\mu = \mathbb{E}[X_i] < 0\) holds.

Relation: The waiting time in a stable G/G/1 queue has the same steady-state distribution as a random walk with negative drift.

Goal:

1. Probability \(\gamma(m)\) of steady-state waiting exceeds some (high) level \(m > 0\)
2. Probability \(\gamma(m)\) of random walk exceeds some (high) level \(m > 0\)
3. \(\gamma(m)\) corresponds to the probability that the First Passage Time
   \[\tau(m) = \inf_{n > 0} [S_n > m]\]
   is finite, i.e.
   \[\gamma(m) = P(\tau(m) < \infty)\]
4. Obviously for large \(m\), due to \(\mu < 0\), exceeding level \(m\) is a rare event

Optimal Exponential Change of Measure

An IS estimator for the probability that a simple random walk with negative drift exceeds some level \(m\), i.e. an IS estimator for the probability \(\gamma(m) = P\{\tau(m) < \infty\}\), is asymptotically optimal, if it is built according to the ECM, where the twisting parameter \(\vartheta^* > 0\) is chosen such that
\[M_X(\vartheta^*) = 1\] and thus \(\ln M_X(\vartheta^*) = 0\)
Twisting by \(\vartheta^*\) is called the optimal exponential change of measure (OECM).

Problem in determining optimal ECM: condition for asymptotically optimal exponential change of measure usually has no explicit solution

Empirical study by Asmussen and Rubinstein (1995) on the efficiency of ECM for single server queues showed that deviations up to 20% away from optimal parameter often yield good results in the sense of large amount of variance reduction
ECM for G/G/1 queue

\( A \) denotes interarrival times, \( B \) denotes service times and \( \mathbb{E}[B] < \mathbb{E}[A] \) is the stability condition.

From Lindley Recursion

\[
W_{n+1} = \max(0, W_n + B_n - A_{n+1}), \quad W_0 = 0
\]

where \( W_i, A_i, B_i \) denote waiting time, interarrival time and service time of the \( i \)-th customer.

Steady-steady waiting time has the same distribution as the maximum of random walk with negative drift \( X = B - A \).

OECM for G/G/1 queue

From \( M_X(\vartheta^*) = 1 \), we can get the equation for the asymptotically optimal change of measure:

\[
M_B(\vartheta^*) M_A(-\vartheta^*) = M_B(\vartheta^*) M_A(-\vartheta^*) = 1
\]

Unfortunately, this condition is explicitly solvable only for a few models:

- \( M/M/1 \)

\[
M_B(\vartheta^*) M_A(-\vartheta^*) = \frac{\mu}{\mu - \vartheta^*} \cdot \frac{\lambda}{\lambda + \vartheta^*} = 1 \quad \Rightarrow \quad \vartheta^* = \mu - \lambda
\]

the asymptotically optimal exponential change of measure corresponds to an interchange of arrival and service rate.

- \( M/Erlang(2)/1 \) and \( Erlang(2)/M/1 \): quadratic expression

- \( M/D/1 \) and \( D/M/1 \): transcendental equation, which has to solved numerically (if possible).

For instance, for the \( M/D/1 \) queue with \( B = 1 \) and \( \lambda < 1 \):

\[
M_B(\vartheta^*) M_A(-\vartheta^*) = e^{\vartheta^*} \cdot \frac{\lambda}{\lambda + \vartheta^*} = 1
\]
Applicability to Queueing Networks?

- For single server queues, ECM in conjunction with large deviations theory often yields quite good results.
- Application of LDT-based ECM to queueing networks turns out to be extremely difficult.
- Well known trial by Parekh and Walrand (1989)
  - Generalization of asymptotically optimal change of measure for M/M/1 queues to Markovian tandem networks.
  - Interchange interarrival rate and smallest service rate (service rate of bottleneck queue).
  - Other rates remain unchanged.
- Glasserman and Kou (1995) showed that, even in the case of only two queues in tandem, this generalization yields infinite variance in some parameter regions.
- Roughly speaking, this generalization for tandem queues is only efficient if there is one single bottleneck queues, i.e. contents/population of system is significantly dominated by one single queue.
- Obviously, for more complicated networks it is even more difficult to find an efficient change of measure.

Concluding Remarks

- Optimal zero variance Importance Sampling estimator typically unavailable.
- Bounded relative error or at least asymptotic optimality highly desirable.
- Change of measure in Importance sampling intimately related to large deviations theory.
- Scaling and translation not promising for queueing network models.
- Exponential change of measure
  - Asymptotically optimal estimators for some single server systems.
  - Relation to large deviations results for random walks.
  - Generalization difficult, not possible even for Markovian tandem queues.
- Further Problems with ECM
  - ECM restricts class of possible Importance Sampling measures.
  - Even best possible ECM may not be asymptotically optimal.
  - Today's state of the art: ECM not well-suited for complex networks.
- Try specialized methods for Markovian models and/or adaptive methods (e.g. Cross Entropy Method).