EFFICIENT RARE-EVENT SIMULATION FOR MANY-SERVER LOSS SYSTEMS

Jose Blanchet
IEOR Department
Columbia University
Rare-event simulation for stochastic system
Many-Server Loss System

- Loss system: GI/GI/s/0 no waiting room → customers are lost if all servers are busy...

- Assume $s = \#$ of servers large

Focus of this talk:
- Computing loss probabilities / overflow events
- Conditional distribution at the time of a loss
- Also discuss systems with time varying / Markov modulated arrivals
MANY-SERVER LOSS SYSTEM: THE MODEL

Customer

Server 1

Server 2

Server 3

Server 4

Server s
Many-Server Loss System

Customer

Server 1
Server 2
Server 3
Server 4

Server s
APPLICATION 1: LONG DISTANCE LINES

• A local company sets up long-distance call lines
• “Customers” are the employees (can be over 5000 in big companies)
• “Service times” are the call holding times

How many call lines should be set up to guarantee a loss probability of less than, say 0.1%?
APPLICATION 2: TELECOMMUNICATION SWITCHES

- Digital switches provide connections among phone calls, internet etc.
- Switch holds a buffer capacity; packets beyond the capacity are rejected
- What is the value of buffer capacity to achieve a loss probability typically in the order of $10^{-9}$?
APPLICATION 3: INSURANCE PORTFOLIO

- A life insurance company sells insurance contracts to policyholders.
- Policyholders pay regular (or lump-sum) premium to the company; in return, the company pays benefit to policyholders in the contingent event (e.g. death).
- “Customers” are the policyholders.
- “Service times” are the times to contingent event (or the tenor of contract, if shorter).
- Large insurance companies have millions of policyholders.
- The cash flow of insurance company is a functional of the statuses of customers in the system:

  \[ \text{net cash position} = \text{net discounted premium received} + \text{net discounted benefit paid} \]

- What is the probability that the insurance company suffers from insolvency?
IMPORTANT FEATURES OF MANY-SERVER LOSS SYSTEM
• Many servers! (order of $10^2$ to $10^3$ depending on context)
• Customers arrive frequently i.e. heavy traffic
• Stable system
• Loss event is rare (order of $10^{-3}$ to $10^{-9}$)
• Other features: time-varying, limited waiting room capacity etc.
The Basic Model

Customers arrive according to a renewal process with rate $\lambda s$ i.e. interarrival times $U_i$ are i.i.d. with mean $1/(\lambda s)$
THE BASIC MODEL

Notes:

- State-space of the process (if insisting on being Markov) at a time $t$ is high-dimensional (measure-valued). It consists of:
  - Number of customers
  - Residual service time for each present customer
  - Age of the process since last arrival
- One convenient way of encoding the state:

  $$ Y(t) = (Q(t, y), B(t)) \in \mathcal{D} \times \mathbb{R}_+ $$

  $Q(t, y)$: number of customers at time $t$ who have residual service times larger than $y$
  $B(t)$: age of process since last arrival

- Traffic intensity $\rho := \lambda E V < 1 \Rightarrow$ stable system
- Technical assumptions:
  - Interarrival times $U_t$ possess exponential moments i.e. $E e^{\theta U_t} < \infty$ for some $\theta$ in a neighborhood of 0
  - Service times $V_t$ have bounded support
MAIN GOAL OF THE TALK

- Provide an “optimal” importance sampling algorithm to estimate the steady-state loss probability

- Main Motivations:
  - Analytical solution not available except Poisson arrival
  - Typical order of magnitude $\approx 10^{-3}$ to $10^{-9}$ $\Rightarrow$ crude Monte Carlo is inefficient, if not infeasible

- More motivations:
  - Since our simulation is pathwise, other quantities of interest can be simulated e.g. conditional expectation of functional of the statuses of customers before loss happens
  - The algorithm can be generalized to a range of more complicated models
CRUDE MONTE CARLO SCHEME

Rare-event simulation for stochastic system

Required service time at arrival

$V = 0$

$U_0$  $U_1$

$M$

$s = 4$

$V_2$

$V_1$

Arrival time
**Crude Monte Carlo Scheme**

Required service time at arrival

$$V = 0$$

$$V_1, V_2$$

$$M$$

$$A_1, A_2$$

$$U_0, U_1$$

$$s = 4$$

Rare-event simulation for stochastic system

Arrival time
CRUDE MONTE CARLO SCHEME

Required service time at arrival

$V = 0$

$M$

$s = 4$

Arrival time

Rare-event simulation for stochastic system
CRUDE MONTE CARLO SCHEME

Required service time at arrival

$V = 0$

$A_1$

$A_2$

$M$

$V_1$

$V_2$

$U_0$

$U_1$

$s = 4$

$\tau_s$

Arrival time

Rare-event simulation for stochastic system
CRUDE MONTE CARLO SCHEME

- Run the process for a long time
  - $\hat{P}(\text{loss}) = \frac{\# \text{ loss}}{\lambda s \times \text{total time units simulated}}$
**A Numerical Example**

**Parameters/Assumptions:**

- \( s = 100, \lambda = 1, EV_i = 0.5 \)
- Poisson arrival with rate \( \lambda s = 100 \)
- Service time \( V_i \sim U[0,1] \)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss probability calculated from Erlang’s loss formula</td>
<td>( 1.630319 \times 10^{-10} )</td>
</tr>
<tr>
<td>Running time to simulate 1000 time units using crude Monte Carlo</td>
<td>( 5.16 \text{ seconds} )</td>
</tr>
<tr>
<td>Approximate number of customers that can be simulated in this time</td>
<td>( 1000 \times 100 = 10^5 )</td>
</tr>
<tr>
<td>Approximate time to simulate one loss event</td>
<td>( \frac{1}{1.630319 \times 10^{-10}} \times 5.16 = 3.66 \text{ days} )</td>
</tr>
<tr>
<td>Approximate time to simulate 100 loss events</td>
<td>( 366 \text{ days} )</td>
</tr>
</tbody>
</table>
Our Algorithm...
ORGANIZATION OF THE TALK

1. Introduce notions in rare-event simulation
2. Explain in detail our importance sampling scheme
3. Algorithmic efficiency
4. Generalizations e.g. renewal arrivals, Markov-modulation, time-varying system
LITERATURE REVIEW


- Rare event analysis / large deviations:
FUNDAMENTAL CHALLENGE OF RARE-EVENT SIMULATION

- Suppose one wants to estimate \( P(A_n) \downarrow 0 \) as \( n \uparrow \infty \)
- Crude Monte Carlo estimator i.e.
  \[
  \frac{1}{N} \sum_{i=1}^{N} 1(A_{i,n})
  \]
gives unbiased estimate with variance
  \[
  \frac{1}{N^2} P(A_n)(1 - P(A_n))
  \]
- Relative error (coefficient of variation) defined by the ratio of standard deviation to mean gives
  \[
  \sqrt{\frac{1 - P(A_n)}{NP(A_n)}}
  \]
- \( N \sim 1/P(A_n) \) number of samples is required to retain reasonable level of relative error
- If \( P(A_n) \) is exponentially small in \( n \), number of samples required also blows up exponentially in \( n \)
IMPORTANCE SAMPLING

- For illustration let \( A_n = \{S_n \in \mathcal{X}\} \) where \( S_n \) has density \( f(\cdot) \)
- Instead of sampling from density \( f(\cdot) \), we sample from \( \bar{f}(\cdot) \)
- Likelihood ratio \( L(S_n) = \frac{f(x_n)}{\bar{f}(x_n)} \)
**Importance Sampling**

**Definition 1:** An estimator is called strongly efficient if its relative error is bounded in $n$ i.e.

$$\frac{\bar{E}(L(S_n) \mathbb{1}(S_n \in \mathcal{X}))^2}{P(S_n \in \mathcal{X})^2} < C$$

for all $n$.

**Definition 2:** An estimator is called asymptotically optimal, or logarithmically efficient, if

$$\limsup_{n \to \infty} \frac{\log \bar{E}(L(S_n) \mathbb{1}(S_n \in \mathcal{X}))^2}{\log P(S_n \in \mathcal{X})} = 2$$

**Note:**
1. If $P(S_n \in \mathcal{X}) \to 0$ exponentially in $n$, then Definition 2 means that the second moment of the estimator decays in twice the exponential rate as $P(S_n \in \mathcal{X})$
2. A zero-variance sampler has a density

$$\frac{f(x) \mathbb{1}(x \in \mathcal{X})}{P(S_n \in \mathcal{X})}$$
A Simple (Simplest) Example...

- Consider $P(S_n > an)$ where $S_n = \sum_{i=1}^{n} X_i$, $X_i$ are i.i.d. r.v.'s with $EX_i = 0$ and $\psi(\theta) = \log E e^{\theta X_i} < \infty$ for all $\theta \in \mathbb{R}$, and $a > 0$
- By Law of Large Numbers, $P(S_n > an) \rightarrow 0$ as $n \uparrow 0$
- Consider the importance sampling scheme where the probability distribution of each $X_i$ is tilted along its exponential family so that $EX_i = a$ i.e. $d\tilde{P} = e^{\theta^* X_i - \psi(\theta^*)} dP$ where $\theta^*$ is the solution to $\psi'(\theta) = \alpha$
- Cramer’s Theorem:

$$P(S_n > an) = \tilde{E}[e^{-\theta^* S_n + n\psi(\theta^*)}; S_n > an]$$
$$= e^{-\theta^* an + n\psi(\theta^*)} \tilde{E}[e^{\theta^*(an - S_n)}; S_n > an] \approx e^{-nI(a)}$$

where $I(a) = \theta^* a - \psi(\theta^*)$ is called the rate function in large deviations theory
Notes from the Example

- The proof of large deviations suggests a natural importance sampling scheme
- This scheme can be shown to be asymptotically optimal:
  \[
  \tilde{E}(L1(S_n > an))^2 = E[L; S_n > an] \\
  = E[e^{-\theta^*S_n + n\psi(\theta^*)}; S_n > an] \\
  = e^{-nI(a)} E[e^{\theta^*(an-S_n)}; S_n > an] \approx e^{-2nI(a)}
  \]
- The importance sampling scheme mimics the zero-variance sampler in the sense that
  \[
P(X_1 \in B_1, \ldots, X_n \in B_n | S_n > an) \rightarrow \tilde{P}(X_1 \in B_1) \cdots \tilde{P}(X_n \in B_n)
  \]
  for all Borel sets \(B_1, \ldots, B_n\)
LARGE DEVIATIONS AND IMPORTANCE SAMPLING

- Contrary to central limit theorems where information on moments is enough, large deviations typically depend on the behavior of the moment generating function.
- Gartner-Ellis Theorem as a generalization of Cramer’s Theorem: Under regularity conditions, suppose a random object $S_n$ possesses logarithmic moment generating function $\psi_n(\theta) = \log E e^{\theta S_n}$ such that $\psi_\infty(\theta) := \lim_{n \to \infty} \frac{1}{n} \psi_n(\theta)$ (Gartner-Ellis limit) exists on a sufficiently large enough interval of $\theta$, then

$$\lim_{n \to \infty} \frac{1}{n} \log P(S_n \in X) = - \inf_{a \in X} I(a)$$

where $I(a) = \sup_{\theta \in \mathbb{R}} \langle \theta, a \rangle - \psi(\theta)$ is the rate function.
LARGE DEVIATIONS AND IMPORTANCE SAMPLING

To find an optimal importance sampler for large deviations event...

- Formulate Gartner-Ellis limit of the random object
- Decode from the limit the contributions of more “elementary” objects that lead to the rare event
- In many cases (but not all), the naturally suggested sampler is asymptotically optimal
BACK TO OUR PROBLEM...

- Do we know the Gartner-Ellis limit of the random object i.e. the steady-state loss distribution?
- Problem reformulation
CRUDE MONTE CARLO REVISITED

- Run the process for a long time

- \( \hat{P}(\text{loss}) = \frac{\# \text{ loss}}{\lambda s \times \text{total time units simulated}} \)
A Regenerative View

- Suppose $A \in \mathcal{D} \times \mathbb{R}_+$ is a regenerative set of the system
- Kac's formula:

$$P_\pi(\text{loss}) = \frac{E_A N_A}{\lambda \mathcal{E}_A \tau_A}$$

Notations:
- $N_A =$ number of loss before reaching $A$ again
- $\tau_A =$ time units to reach $A$ again
- $E_A[\cdot] =$ expectation with initial state in steady-state conditional on being in $A$

- If we choose $A$ to be a “good” set i.e. it does not take exponential amount of time to reach, then crude Monte Carlo is equivalent to using sample mean for both numerator and denominator
- Mixing guaranteed by finite support assumption on service time
A Regenerative View

Regeneration is hit. A sample of $N_A$ and $\tau_A$ is recorded.

Another regeneration is hit. Another sample of $N_A$ and $\tau_A$ is taken.

Cycle starts $D \times \mathbb{R}_+$. 

33
“Splitting” Algorithm

- Do importance sampling on the numerator
- Run crude Monte Carlo; every time $A$ is hit, “split” the path into two: one continue with original evolution; another is applied with importance sampling to get a sample of $N_A$. Then continue with the original path

$$\hat{P}(\text{loss}) = \frac{\text{weighted sum of } N_A}{\lambda s \times \text{total time units}}$$
“Splitting” Algorithm

Regeneration is hit. A sample of $\tau_A$ is recorded. “Split” the chain again.

Another $N_A$ is recorded.

An importance sampling sample of $N_A$ is recorded.

$D \times \mathbb{R}_+$

Regeneration is hit. “Split” into two chains $A$

Another regeneration is hit. Another sample of $\tau_A$ is taken. Continue the procedure.
What is a good choice of set $A$?

- $A$ should be occupied frequently (but not too frequently!) in steady-state.
- When $s$ is large, one can use Central Limit Theorem to approximate the “central limit” region of $(Q(t, y), B(t)) \in D \times \mathbb{R}_+$.
- (Pang and Whitt (2008)) Suppose there is no capacity constraint i.e. number of servers is infinite (but arrival rate is still $\lambda s$), then

$$
\frac{Q(\infty, y) - \lambda s \int_y^\infty \bar{F}(u) du}{\sqrt{s}} \approx Z(y)
$$

where $Z(y)$ is a Gaussian process with

$$
\text{Var } Z(y) = \lambda c^2 \int_y^\infty \bar{F}(u)^2 du + \lambda \int_y^\infty F(u) \bar{F}(u) du
$$

- We can choose

$$
A = \left\{ Q(t, y) \in \left( \lambda s \int_y^\infty \bar{F}(u) du - sd(Z(y)) \sqrt{s}, \lambda s \int_y^\infty \bar{F}(u) du + sd(Z(y)) \sqrt{s} \right), t \in \{\Delta, 2\Delta, \ldots\} \right\}
$$
**Key Questions**

- Our problem becomes estimating $E_r N_A$ for some $r \in A$
- Do we have information from large deviations theory (i.e. Gartner-Ellis limit)?
- How does the rare event i.e. loss happen?
- Does the intuition give an asymptotically optimal estimator (or more)?
A SIMPLER PROBLEM

- Consider a simpler problem in which Gartner-Ellis limit can be computed:
  - A “coupled” system that has no capacity constraint i.e. number of servers is infinite
  - Fix a time horizon $t$ and initial state, say 0
- What is the probability that there are more than $s$ customers in the system at time $t$?
SOLUTION TO THE SIMPLER PROBLEM

- This is mathematically

\[ P(Q(t) > s) \]

where

\[ Q(t) = \sum_{i=1}^{N(t)} 1(V_i > t - A_i) \]

is the number of customers in the system at time \( t \)

- Gartner-Ellis limit

\[ \psi_{\infty}(\theta) = \int_0^t \psi_N(\log(e^\theta F(t-u) + F(t-u)))du \]

where \( \psi_N(\cdot) \) is the infinitesimal logarithmic moment generating function of the arrival process
IMPORTANCE SAMPLING FOR THE SIMPLER PROBLEM

Required service time at arrival

\( M \)

\( V = 0 \)

Arrival time

\( s = 4 \)

Random-event simulation for stochastic system

40
**Importance Sampling for the Simpler Problem**

1. Arrival rate is accelerated towards $t$ by tilting the interarrival times $U_i$ by
   \[ \psi_N(\log(e^{\theta*F(t-A_i)} + F(t-A_i)) \]

Required service time at arrival

\[ M \]

\[ V = 0 \]

\[ s = 4 \]

Arrival time

Rare-event simulation for stochastic system
**Importance Sampling for the Simpler Problem**

1. Arrival rate is accelerated towards $t$ by tilting the interarrival times $U_i$ by
   $$\psi_N(\log(e^{\theta^*} F(t - A_i) + F(t - A_i)))$$
2. Service time density is increased by a factor of $e^{\theta^*}$ inside the triangle
INTUITION FOR OUR PROBLEM

Given an initial state \( r \in A \),

\[
E_r N_A = E_r [N_A; \tau_s < \tau_A] \approx P_r (\tau_s < \tau_A)
\]

\[
\approx \sum_t P_r (\tau_s = t < \tau_A) \approx \sum_t P_r (Q(t) > s) \approx P_r (Q(t^*) > s)
\]

\[
\approx e^{-sI_t^*}
\]

Idea 1: Before the first loss, the system acts the same as if there are infinite number of servers

Idea 2: Since time horizon for loss is not fixed, we shall randomize the time horizon (also preventing blowing up variance due to non-optimal sample paths)
Algorithm for $E_{\tau}N_A$

Step 1: Sample a random time $\tau$ according to $p(\tau = T + 1) = \frac{6}{(1 + 1)^2}$

Arrival time

$\tau$

$V = 0$

$M$

Required service time at arrival

$s = 4$
Algorithm for $E_r N_A$

Step 2: Run the sequential importance sampler as in the simpler problem, pretending infinite number of servers and time horizon of the realized $\tau$, until $\tau_s \wedge \tau_A$
Algorithm for $E_r N_A$

Step 2: Run the sequential importance sampler as in the simpler problem, pretending infinite number of servers and time horizon of the realized $\tau$, until $\tau_s \wedge \tau_A$. 

Required service time at arrival

$s = 4$

Arrival time

Rare-event simulation for stochastic system
Algorithm for $E_rN_A$

Step 2: Run the sequential importance sampler as in the simpler problem, pretending infinite number of servers and time horizon of the realized $\tau$, until $\tau_s \land \tau_A$
Algorithm for $E_r N_A$

Required service time at arrival

Step 2: Run the sequential importance sampler as in the simpler problem, pretending infinite number of servers and time horizon of the realized $\tau$, until $\tau_s \wedge \tau_A$

$M$

$V = 0$

$s = 4$

Arrival time

48
Algorithm for $E_r N_A$

Step 3: If $\tau_s$ is hit first, continue the process under the original measure until $\tau_A$
Algorithm for $E_{r}N_{A}$

Required service time at arrival

$V = 0$

$M$

Until $A$ is hit

$s = 4$

Step 3: If $\tau_{s}$ is hit first, continue the process under the original measure until $\tau_{A}$

Arrival time

50
ASYMPTOTIC OPTIMALITY

The likelihood ratio for the algorithm under $\tau_s < \tau_A$ is

$$L(Y_u, 0 \leq u \leq \tau_s) = \frac{1}{\sum_t P(\tau = t) L_t^{-1}(Y_u, 0 \leq u \leq \tau_s)}$$

where $L_t$ is the likelihood ratio conditional on $\tau = t$ given by

$$\exp\left\{ s \sum_{i=1}^{N(\tau_s)-1} \psi_N\left( \log\left( e^{\theta_t \bar{F}(t-A_i)} + F(t-A_i) \right) \right) U_i - \theta_t \sum_{i=1}^{N(\tau_s)-1} 1(V_i > t - A_i) \right\}$$

for $t \geq \tau_s$ and

$$\exp\left\{ s \sum_{i=1}^{N(t)} \psi_N\left( \log\left( e^{\theta_t \bar{F}(t-A_i)} + F(t-A_i) \right) \right) U_i - \theta_t \sum_{i=1}^{N(t)} 1(V_i > t - A_i) \right\}$$

for $t < \tau_s$.
ASYMPTOTIC OPTIMALITY

The likelihood ratio is bounded from above by

$$\frac{L[\tau_s]}{P(\tau = |\tau_s|)} \leq c[\tau_s]^{2e^{-sI_{\tau_s}}} + \theta[\tau_s]^R[\tau_s, |\tau_s|]$$

Required service time at arrival

$$s = 4$$
ASYMPTOTIC OPTIMALITY

- Second moment of likelihood ratio is
  \[ \tilde{E}[N_A^2L^2; \tau_s < \tau_A] = E[N_A^2L; \tau_s < \tau_A] \leq ce^{-sI_t^*}E[e^{\theta[\tau_s]}R[\tau_s, [\tau_s]]; \tau_s < \tau_A] \]

- When \( \tau \) is sampled at scale \( O\left(\frac{1}{s}\right) \):
  - For the case of Poisson arrival, given \( \tau_s, R[\tau_s, [\tau_s]] \sim Binomial(s, p) \) where \( p \) is the ratio of purple area to trapezoid
  - For general case, condition on the arrival times of the contributing customers

- One can get a logarithmic bound of \( e^{-2sI_t^*} \)
ASYMPTOTIC OPTIMALITY

**Theorem:** We have

\[
\lim_{s \to \infty} \frac{1}{s} \log P(\text{loss}) = -I_{t^*}
\]

where \( t^* = \text{argmin} I_t \) and

\[
\lim_{s \to \infty} \frac{1}{s} \log \tilde{E}[N_A^2 L^2; \tau_s < \tau_A] = -2I_{t^*}
\]

Hence the algorithm is asymptotically optimal.

**Sketch of Proof:**

- Lower bound \( \lim_{s \to \infty} \frac{1}{s} \log P(\text{loss}) \geq -I_{t^*} \) is established by explicitly identifying the optimal sample path
- For upper bound,

\[
-2I_{t^*} \leq \lim_{s \to \infty} \frac{1}{s} \log P(\text{loss})^2 \leq \lim_{s \to \infty} \frac{1}{s} \log \tilde{E}[N_A^2 L^2; \tau_s < \tau_A] \leq -2I_{t^*}
\]
SIMPLIFICATION AND EXTENSIONS

For Poisson arrival,

- A faster algorithm can be obtained by, after sampling $\tau$, generating $Q(t)$ using tilted measure and then sampling the customers exploiting the Poisson random measure description.

- It is interesting to note that the seemingly more powerful idea of conditionally sampling $Y_u, 0 \leq u \leq t | Q(t)$ (instead of exponential tilting) will blow up the second moment of the likelihood ratio at a neighborhood of the time of first loss, due to “discontinuity” of the likelihood ratio at $\tau_S$.

- However, this works for a discrete version of the process (Blanchet, Glynn and Lam (2009)).
SIMPLIFICATION AND EXTENSIONS

- The case of Markov-modulated arrivals can be simulated by restricting set $A$ to the optimal Markov state i.e. the state that gives the highest arrival rate.

- The case of both Markov-modulated arrivals and (possibly correlated) service times can be simulated by augmenting the state-space to identify the residual service times of customers who enter at each Markov state.

- Time-varying arrivals (in this case we are interested in loss during an interval instead of the steady-state) can be simulated using exactly the same methodology (with truncation of $\tau$).