RARE EVENT SIMULATION TECHNIQUES FOR STOCHASTIC MODELS: AN OVERVIEW

Perwez Shahabuddin

Columbia University
New York, U.S.A.
(perwez@ieor.columbia.edu)

Rare Events In Stochastic Models

- Current interest in rare events stems primarily from developments in computer and communication technology

- Example: Telecommunication: Standards for telecommunications call for extremely small probabilities of buffer overflow (or equivalently, packet loss) in telecom switches, e.g. 10^{-9}

- Example: Reliability: Many industrial and scientific applications require highly reliable computer systems (with correspondingly small failure probabilities)

- Exact analysis usually very difficult, due to
  - Large size (e.g., Markov chains)
  - Analytical Complexity (e.g., non-Markovian models)

Rare Events In Stochastic Models (Contd.)

- Two alternatives:

  1. Asymptotic Approximations (e.g. Large Deviation Theory).
     - Computationally inexpensive (+).
     - Little idea/control of error in the approach (-).
  2. Simulation
     - Usually computationally expensive (-).
     - Better idea of the error of the estimate using confidence interval (+).
     - Can control error by sampling more (+)

- We focus on simulation.

- Difficulty: Takes extremely long simulation runs for rare events to happen.

- General Idea: Use information provided by asymptotics, to speed-up simulation by orders of magnitude.

Reliability Model of a Simple Computer System

- System is considered up if the data is accessible: at least, 1 database and 1 processor is up.

- System state $X = (S_p, S_d)$ where
  - S_p = No. of processor up.
  - S_d = No. of database up.

- Failure rates: λ_p, λ_d, Repair rates: μ_p, μ_d
  (e.g. $\lambda_p = \lambda_d = 10^{-4}$/hr, $\mu_p = \mu_d = 1$/hr)

Reliability Model of a Simple Computer System (Contd.)

Performance Measures of Interest::

- Long-run fraction of time the system spends in failed states (steady state unavailability).
- Expected time to failure starting from "new" state.
- Chance of system failing before a given time (unreliability).

Rare Events in Queuing Models

- M/M/1/B Queue (B: Buffer Capacity)

- λ: Arrival Rate, μ: Service Rate ($\lambda < \mu$)
  (e.g. $\lambda = 0.5$/hr, $\mu = 1$/hr)

- When buffer is full arriving customers are lost.

- Problem: Estimate probability of customer lost when B is large (buffer overflow event is rare)

- System State, X= No. of people in system

- Transition Diagram:

General Rare Event Simulation Problem

- Internal probability dynamics $\mathbf{P}$ is given.
- Rare set F is given.
- Problem: find
  - $\pi_F :=$ Fraction of time in the long run, the system is in state F.
  - Mean time to failure, i.e., $E(T_F)$ where T_F is the time to reach F.
  - $P(T_F < t)$, where t is a fixed time

Importance Sampling

- Change probability dynamics $\mathbf{P}$ to make F occur more frequently. Call this new dynamics $\tilde{\mathbf{P}}$.

- Simulate using $\tilde{\mathbf{P}}$. Adjust estimate using knowledge of $\mathbf{P}$ and $\tilde{\mathbf{P}}$.

- Have to be very careful: Arbitrary $\tilde{\mathbf{P}}$ that makes F occur faster may decrease simulation efficiency.

- $\tilde{\mathbf{P}}$ which work, generally give orders of magnitude increase in efficiency.

- Lot of research in choosing good $\tilde{\mathbf{P}}$. (Anantharam, Asmussen, Anderson, Cottrell, Frater, Fort, Glasserman, Glynn, Hiedelberger, Iglehart, Malgouyres, Mandjes, Nicola, Parekh, Ridder, Rubinstein, Sadowsky, Seigmund, Walrand, etc.)

More Formally ...

- Let $\{X(t), t \geq 0\}$ be a stochastic process.
- Let $\mathcal{S}$ be the state space.
- Let $F \subset \mathcal{S}$ be the rare set.
- Let A be the *rare event*, e.g., chance of hitting F in $[0, T]$.
- Problem: Estimate $\gamma = P(A)$
- Note that: $\gamma = P(A) = E[I_A]$ where I_A is the "indicator" random variable:

$$I_A = \begin{cases} 1 & \text{if } A \text{ happens} \\ 0 & \text{otherwise} \end{cases}$$

Naive Simulation Algorithm

1. Generate "n" independent samples of the stochastic process

2. Thus we get n independent samples of I_A, say, $I_{A,1}, I_{A,2}, ..., I_{A,n}$.

3. Estimate $\gamma \equiv E[I_A]$ by $\hat{\gamma}_n = \frac{\sum_{i=1}^{n} I_{A,i}}{n}$

Note that $\hat{\gamma}_n$ is

(i) Unbiased: $E[\hat{\gamma}_n] = E[\frac{\sum_{i=1}^{n} I_{A,i}}{n}] = E[I_A] = \gamma$

(ii) Consistent: By Law of Large Numbers we have

$$\hat{\gamma}_n = \frac{\sum_{i=1}^{n} I_{A,i}}{n} \to E[I_A] = \gamma$$

as $n \to \infty$

Rare Event Simulation Problem

Question:
How fast does $\hat{\gamma}_n$ converge to γ?

Answer:

- Depends on $\text{Var}(\hat{\gamma}_n)$.
  $\text{Var}(\hat{\gamma}_n) = \frac{\text{Var}(I_A)}{n} = \frac{E[I_A^2]-E^2[I_A]}{n} = \frac{E[I_A]-\gamma^2}{n} = \frac{\gamma-\gamma^2}{n}$
  $\approx \frac{\gamma}{n}$ (since γ is small)

- Infact, the confidence interval (CI) is given by

$$\hat{\gamma}_n \pm C\sqrt{Var(\hat{\gamma}_n)} \approx \hat{\gamma}_n \pm C\sqrt{\frac{\gamma}{n}}$$

  where C is a constant obtained from Gaussian tables.

- Hence half width of CI, $HW = C\sqrt{\frac{\gamma}{n}}$

- Relative Error: $RE \equiv \frac{HW}{\gamma} = \frac{C}{\sqrt{n\gamma}}$

- $RE \to \infty$ as $\gamma \to 0$ for fixed effort n

Another Perspective on the Rare Event Simulation Problem

- Recall: $RE \equiv \frac{HW}{\gamma} = \frac{C}{\sqrt{n\gamma}}$
- Hence $n = \frac{C^2}{RE^2\gamma}$.

- Hence to achieve a given RE, the effort $n \to \infty$, as $\gamma \to 0$.

- In "light-tailed" stochastic systems, usually

$$\gamma = C_1 e^{-C_2 B},$$

  where C_1 and C_2 are constants, and B is some rarity parameter (e.g., buffer size of a switch)

- Hence $n = \frac{C^2 e^{C_2 B}}{RE^2 C_1}$.

- Hence the effort n increases *exponentially* w.r.t. B.

Importance Sampling Concept

Basic Equation:

- Let ω be a sample path (assume discrete for simplicity).
- Let Ω be the set of all sample paths.
- Let $\tilde{\mathbf{P}}$ be some other probability measure with corresponding density $\tilde{p}(w)$.
- $$\gamma = E[I_A] = \sum_{\omega \in \Omega} I_A(\omega) p(\omega)$$
$$= \sum_{\omega \in \Omega} I_A(\omega) \frac{p(\omega)}{\tilde{p}(\omega)} \tilde{p}(\omega) = \sum_{\omega \in \Omega} I_A(\omega) L(\omega) \tilde{p}(\omega)$$

  where $L(\omega) = \frac{p(\omega)}{\tilde{p}(\omega)}$ is the likelihood ratio.
- Hence $\gamma = E_{\tilde{p}}[I_A L]$

Importance Sampling Algorithm

1. Generate "n" sample paths of the stochastic process using $\tilde{\mathbf{P}}$ (instead of $\mathbf{P}$). Call them $\omega_1, \omega_2, ..., \omega_n$.

2. Set $I_{A,i} = 1$ if $\omega_i \in A$.

3. Compute $L_i = \frac{p(\omega_i)}{\tilde{p}(\omega_i)}$.

4. Estimate $\gamma = E_{\tilde{p}}[I_A.L]$ by $\tilde{\gamma}_n = \frac{\sum_{i=1}^{n} I_{A,i}.L_i}{n}$.

Note that

- $Var_{\tilde{p}}(\tilde{\gamma}_n) = \frac{Var_{\tilde{p}}(I_A L)}{n} = E_{\tilde{p}}(I_A^2 L^2) - \gamma^2$.

- Compare with $Var_p(\tilde{\gamma}_n) = E_p(I_A) - \gamma^2$.

- Question: Which $\tilde{\mathbf{P}}$ should we use so that $E_{\tilde{p}}(I_A^2 L^2) \ll E_p[I_A]$?

- Since $E_{\tilde{p}}(I_A^2 L^2) = E_p(I_A L)$ we want $E_p(I_A L) \ll E_p[I_A]$

Intuition in Selection of $\tilde{P}$

- Coarse Intuition:
  A should be made much more probable under $\tilde{\mathbf{P}}$, i.e., $\tilde{P}(A) \gg P(A)$

- Not sufficient; in fact, may make things much, much worse.

- Reasons:
  - Want $E_p[I_A L] \ll E_p[I_A]$.
  - Hence a sufficient condition is :
    $L = L(\omega) = \frac{p(\omega)}{\tilde{p}(\omega)} \ll 1$ for each $\omega \in A$

- Refined intuition(always works if implementable): Each $\omega \in A$ should be made much more probable under $\tilde{P}$.

- Very difficult problem as there may be uncountably infinite number of sample paths in A.

- Lot of research in finding $\tilde{\mathbf{P}}$ with this property.

Coarse Intuition

Refined Intuition

In Fact ...

- If
$$\tilde{p}(\omega) = p(\omega|\omega \in A) = \frac{p(\omega)}{p(A)}$$
then our importance sampling estimator has zero variance.

- Why? Because then $L(\omega) = p(\omega)/\tilde{p}(\omega) = P(A)$ (non-random) for all $\omega \in A$.

- Hence $I_A L$ is non-random under $\tilde{p}$
(Recall that $\gamma = E_{\tilde{p}}(I_A L)$)

- However, not implementable as for determining $\tilde{p}(\omega)$ one needs to know $P(A) \equiv \gamma$!

- Can we still use the basic intuition?

Use Large Deviations

- Large deviations theory gives an approximate idea of the "most likely path" (under some scaling) the rare event can happen for several stochastic processes.

- Also, the probability of deviations from this most likely path falls exponentially fast, i.e., only the most like path is important.

- Choose $\tilde{\mathbf{P}}$ to stear the stochastic process in that direction.

- Change of measure framework used: "exponential change of measure" or "exponentially twisting"
  (will cover in second talk)

A Canonical Example: The M/M/1 Queue

- Let λ be arrival rate and let μ be the service rate.
- Note:$\lambda < \mu$ for stability.
- Random walk associated with the M/M/1 queue has a negative drift with rate $(\lambda - \mu)/(\lambda\mu) < 0$.
- Large deviations theory says that the most likely path to buffer overfow (after scaling) is a straight line with slope $(\mu - \lambda)/(\lambda\mu) > 0$.
- This corresponds to the random walk for the M/M/1 queue with arrival and service rate interchanged.
- This is the new system we simulate.

Criteria for Evaluating the Efficiency of Rare Event Simulation

- Let B be a rarity parameter, i.e., B is such that $\gamma \equiv \gamma(B) \to 0$ as $B \to \infty$.

- Example, $B =$ buffer size in queue.

- $RE(B) = \frac{HW(B)}{\gamma(B)} = C\frac{\sqrt{\mathrm{Var}_{\tilde{p}}(I_A L^2)}}{n\gamma(B)}$

- First criterion: Bounded relative error (BRE)

$$\boxed{RE(B) \leq K \quad \text{as} \quad B \to \infty, \quad \text{where } K \text{ is a constant}}$$

- Recall, for naive simulation, $RE(B) \to \infty$ as $B \to \infty$

- BRE is very difficult to achieve in most situations.

- In that case we settle for something weaker called *asymptotic optimality*.

Asymptotic Optimality

- Motivation: Note that
  - $Var_{\tilde{p}}(I_A.L^2) = E_{\tilde{p}}[I_A.L^2] - \gamma^2 \geq 0$
  - In application satisfying large deviation properties, γ decreases at some exponential rate θ^*, i.e., $\lim_{B\to\infty} \frac{-\log(\gamma(B))}{B} = \theta^*$
  - Roughly speaking: $\gamma \approx C_1 e^{-\theta^* B}$, where C_1 is a constant.
  - Hence the fastest possible exponential rate of decreases of $E_{\tilde{p}}(I_A L^2)$ is $2\theta^*$, i.e.,

$$\boxed{\lim_{B\to\infty} \frac{-\log(E_{\tilde{p}}[I_A L^2])}{B} = 2\theta^*}$$

$\Uparrow$

Asymptotic optimality criteria

State of the Art for Reliability Models

- Exponential failure and repair times (i.e., Markov models).
- Systems with large (numerical methods face state space explosion) number of components, each of which fails & gets reparied.
- System considered failed if certain combination of components are failed.
- Interactions among components.
  - Sharing of repairmen.
  - Operational dependencies.
  - Repair dependencies.
  - Failure propagation.
- Performance Measures of Interest:
  - Steady state: steady-state unavailability.
  - Transient: unreliability, interval unavailability, mean time to failure
- Type of results: Bounded relative error.
- Heuristics exist for general failure & repair times (i.e., Non-Markovian models).

State of the Art for Queuing Models & Networks

- Continuous time queues:
  - GI/GI/K queue.
  - Tandem queues, e.g., M/GI/1 $\longrightarrow$ GI/1 queue.
  - Open problem: More general queueing network.

- Discrete time queues:
  - Acyclic communication networks

  - Open problem: cyclic communication networks.

- Fluid analog of above queues.

- Type of results: Asymptotic optimality.

- Heuristics for some of the open problems exist.

Multilevel Splitting

Basic Idea:

Issues in Multi-level Splitting

Question 1: Where should one place the levels?

Answer 1: Place such that the chance of hitting the next level from a given level is not rare. (for simplicity, place equidistant)

Question 2: How many splits at each level? (for simplicity, assume same number of splits at each level)

Answer 2:

- Not too low, as then then there may be no eventual hits to the rare set.
- Not too high, as then the computational effort will increase exponentially fast.

History & State of the Art

- Originally discussed in the context of estimating rare particle transmission probabilities in physics by Kahn & Harris (1952).

- Applied to fast simulation of rare events in computer and communication systems by Bayes (1972), Hopman and Kleijnen (1978) Villen-Altamirano and Villen Altamirano(1991, 1994), etc.,

- Little or no analysis of simulation efficiency or conditions under which the technique worked.

- Glasserman, Heidelberger, Shahabuddin, Zajic (1998 & 1999) identified a unifying set of models & implementation conditions under which this type of method is asymptotically optimal.

- Used Branching Process Theory for variance analysis.

- Main assumption:Process can be multidimensional, but rare event can occur only in one direction.

History & State of the Art (Contd.)

- Open problems: Selection of intermediate sets, numbers of splits & asymptotic optimality results for multidimensional rare event problem.

  – Needs large deviations theory as in importance sampling.
  – However, more robust as compared to importance sampling.

- Heuristics currently being developed for open problems.