An Introduction to Monte Carlo Methods and Rare Event Simulation

Gerardo Rubino and Bruno Tuffin

INRIA Rennes - Centre Bretagne Atlantique

QEST Tutorial, Budapest, September 2009
Outlook

1. Introduction to rare events
2. Monte Carlo: the basics
3. Inefficiency of crude Monte Carlo, and robustness issue
4. Importance Sampling
5. Splitting
6. Confidence interval issues
7. Some applications
Outline

1. Introduction to rare events
2. Monte Carlo: the basics
3. Inefficiency of crude Monte Carlo, and robustness issue
4. Importance Sampling
5. Splitting
6. Confidence interval issues
7. Some applications
Introduction: rare events

Rare events occur when dealing with performance evaluation in many different areas

- in *telecommunication networks*: loss probability of a small unit of information (a packet, or a cell in ATM networks), connectivity of a set of nodes,
- in *dependability analysis*: probability that a system is failed at a given time, availability, mean-time-to-failure,
- in *air control systems*: probability of collision of two aircrafts,
- in *particle transport*: probability of penetration of a nuclear shield,
- in *biology*: probability of some molecular reactions,
- in *insurance*: probability of ruin of a company,
- in *finance*: value at risk (maximal loss with a given probability in a predefined time),
- ...
What is a rare event? Why simulation?

- A rare event is an event occurring with a small probability.
- How small? Depends on the context.
- In many cases, these probabilities can be between $10^{-8}$ and $10^{-10}$, or even at lower values. Main example: critical systems, that is,  
  - systems where the rare event is a catastrophic failure with possible human losses,
  - or systems where the rare event is a catastrophic failure with possible monetary losses.
- In most of the above problems, the mathematical model is often too complicated to be solved by analytic or numeric methods because  
  - the assumptions are not stringent enough,
  - the mathematical dimension of the problem is too large,
  - the state space is too large to get a result in reasonable time,
  - ...
- Simulation is, most of the time, the only tool at hand.
Outline

1. Introduction to rare events
2. Monte Carlo: the basics
3. Inefficiency of crude Monte Carlo, and robustness issue
4. Importance Sampling
5. Splitting
6. Confidence interval issues
7. Some applications
In all the above problems, the goal is to compute $\mu = \mathbb{E}[X]$ for some random variable $X$ (that is, it can be written in this form).

Monte Carlo simulation (in its basic form) generates $n$ independent copies of $X$, $(X_i, 1 \leq i \leq n)$. Then,

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an approximation (an estimation) of $\mu$;

$\bar{X}_n \to \mu$ with probability 1, as $n \to \infty$ (Strong Law of Large Numbers).
**Accuracy**: how accurate is $\bar{X}_n$? We can evaluate the accuracy of $\bar{X}_n$ by means of the Central Limit Theorem, which allows us to build the following confidence interval:

$$CI = \left( \bar{X}_n - \frac{c_\alpha \sigma}{\sqrt{n}}, \bar{X}_n + \frac{c_\alpha \sigma}{\sqrt{n}} \right)$$

- meaning: $\mathbb{P}(\mu \in CI) \approx 1 - \alpha$; $\alpha$: *confidence level*
- (that is, on a large number $M$ of experiences (of estimations of $\mu$ using $\bar{X}_n$), we expect that in roughly a fraction $\alpha$ of the cases (in about $\alpha M$ cases), the confidence interval doesn’t contain $\mu$)
- $c_\alpha = \Phi^{-1}(1 - \alpha/2)$ where $\Phi$ is the cdf of $\mathcal{N}(0, 1)$
- $\sigma^2 = \text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X]$, usually unknown and estimated by

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \bar{X}_n^2.$$
Remarks on the confidence interval

- Size of the confidence interval: $2c_\alpha \sigma / \sqrt{n}$.
- The smaller $\alpha$, the more confident we are in the result:
  \[ P(\mu \text{ belongs to } CI) \approx 1 - \alpha. \]
- But, if we reduce $\alpha$ (without changing $n$), $c_\alpha$ increases:
  - $\alpha = 10\%$ gives $c_\alpha = 1.64$,
  - $\alpha = 5\%$ gives $c_\alpha = 1.96$,
  - $\alpha = 1\%$ gives $c_\alpha = 2.58$.
- The other way to have a better confidence interval is to increase $n$.
- The $1/\sqrt{n}$ factor says that to reduce the width of the confidence interval by 2, we need 4 times more replications.
A fundamental example: evaluating integrals

- Assume \( \mu = \int_{I} f(x) \, dx < \infty \), with \( I \) an interval in \( \mathbb{R}^d \).

- With an appropriate change of variable, we can assume that \( I = [0, 1]^d \).

- There are many numerical methods available for approximating \( \mu \). Their quality is captured by their convergence speed as a function of the number of calls to \( f \), which we denote by \( n \).

Some examples:

- Trapezoidal rule; convergence speed is in \( n^{-2/d} \),
- Simpson’s rule; convergence speed is in \( n^{-4/d} \),
- Gaussian quadrature method having \( m \) points; convergence speed is in \( n^{-(2m-1)/d} \).

For all these methods, the speed decreases when \( d \) increases (and \( \to 0 \) when \( d \to \infty \)).
The “independence of the dimension”

- Let now $X$ be an uniform r.v. on the cube $[0, 1]^d$.
- We immediately have $\mu = \mathbb{E}[X]$, which opens the path to the Monte Carlo technique for approximating $\mu$ statistically.
- We have that $\bar{X}_n$ is an estimator of our integral,
  - and that the convergence speed, as a function of $n$, is in $n^{-1/2}$, thus independent of the dimension $d$ of the problem.
- This independence of the dimension of the problem in the computational cost is the main advantage of the Monte Carlo approach over quadrature techniques.
- In many cases, it means that quadrature techniques can not be applied, and that Monte Carlo works in reasonable time with good accuracy.
Other examples

- **Reliability at** $t$:
  - $C(t)$ is the configuration of a multicomponent system at time $t$;
  - $s(c) = 1$ (when configuration is $c$, system is operational);
  - $X(t) = 1(s(C(u)) = 1$ for all $u \leq t$);
  - $\bar{X}_n(t) = n^{-1} \sum_{i=1}^{n} X_i(t)$ is an estimator of the reliability at $t$, with $X_1(t), \ldots, X_n(t)$ $n$ iid copies of $X(t)$.

- **Mean waiting time in equilibrium**:
  - $X_i$ is the waiting time of the $i$th customer arriving to a stationary queue,
  - $\bar{X}_n$ is an estimator of the mean waiting time in equilibrium.

- etc.
Improving Monte Carlo methods

- Given a problem (that is, given $X$), there are possibly many estimators for approximating $\mu = \mathbb{E}(X)$.
- For any such estimator $\tilde{X}$, we can usually write
  $$\tilde{X} = \phi(X_1, \cdots, X_n)$$

  where $X_1, \cdots, X_n$ are $n$ copies of $X$, not necessarily independent in the general case.
- How to compare $\tilde{X}$ with the standard $\bar{X}$? Or how to compare two possible estimators of $\mu$, $\tilde{X}_1$ and $\tilde{X}_2$?
- Which good property for a new estimator $\tilde{X}$ must we look for?
- A first example is unbiasedness: $\tilde{X}$ is unbiased if $\mathbb{E}(\tilde{X}) = \mu$, which obviously looks as a desirable property.
- Note that there are many useful estimators that are not unbiased.
From the accuracy point of view, the smaller the variability of an unbiased estimator (the smaller its variance), the better its accuracy.

For instance, in the case of the standard estimator $\bar{X}$, we have seen that its accuracy is captured by the size of the associated confidence interval, $2c_\alpha \sigma / \sqrt{n}$.

Now observe that this confidence interval size can be also written $2c_\alpha \sqrt{\mathbb{V}(X)}$.

A great amount of effort has been done in the research community looking for new estimators of the same target $\mu$ having smaller and smaller variances.

Another possibility (less explored so far) is to reduce the computational cost.

Let's look at this in some detail, focusing on the variance problem.
Before looking at some ideas developed to build estimators with “small” variances, let us look more formally at the accuracy concept.

The variability of an estimator \( \tilde{X}_n \) of \( \mu \) is formally captured by the Mean Squared Error

\[
\text{MSE}(\tilde{X}_n) = \mathbb{E}[(\tilde{X} - \mu)^2] = \mathbb{V}(\tilde{X}_n) + \mathbb{B}^2(\tilde{X}_n),
\]

where \( \mathbb{B}(\tilde{X}_n) \) is the Biais of \( \tilde{X}_n \),

\[
\mathbb{B}(\tilde{X}_n) = |\mathbb{E}(\tilde{X}_n) - \mu|.
\]

Recall that many estimators are *unbiased*, meaning that \( \mathbb{E}(\tilde{X}_n) = \mu \), that is, \( \mathbb{B}(\tilde{X}_n) = 0 \) (and then, that \( \text{MSE}(\tilde{X}_n) = \mathbb{V}(\tilde{X}_n) \)).

The dominant term is often the variance one.

In the following refresher, the goal is to estimate \( \mu = \mathbb{E}(X) \) where \( X \) has cdf \( F \) and variance \( \sigma^2 \). Recall that \( \mathbb{V}(\bar{X}_n) = \sigma^2/n \).
Variance reduction: antithetic variables

- Suppose \( n \) is even, that is, \( n = 2k \).
- Assume that the \( i \)th replication \( X_i \) is obtained using \( X_i = F^{-1}(U_i) \), with \( U_1, \cdots, U_n \) i.i.d. with the Uniform(0,1) distribution.
- Let us define a new estimator \( \tilde{X}_{2k} \) using half the previous number of uniform r.v.: \( \tilde{X}_{2k} \) is built from \( U_1, \cdots, U_k \) using
  \[
  \tilde{X}_{2k} = \frac{1}{2k} \sum_{j=1}^{k} \left[ F^{-1}(U_j) + F^{-1}(1 - U_j) \right].
  \]
- Observe that if \( U \) is Uniform(0,1), \( 1 - U \) has the same distribution and both variables are negatively correlated:
  \[
  \text{Cov}(U, 1 - U) = \mathbb{E}[U(1 - U)] - \mathbb{E}(U)\mathbb{E}(1 - U) = -1/12.
  \]
For the variance of $\tilde{X}_{2k}$,

$$\text{V}(\tilde{X}_{2k}) = \frac{1}{4k^2} \sum_{j=1}^{k} \text{V}(Y_j + Z_j),$$

with $Y_j = F^{-1}(U_j)$ and $Z_j = F^{-1}(1 - U_j)$.

After some algebra, writing back $2k = n$,

$$\text{V}(\tilde{X}_n) = \frac{1}{n} \left( \sigma^2 + \text{Cov}(Y, Z) \right),$$

with $(Y, Z)$ representing any generic pair $(Y_j, Z_j)$.

It can now be proven that $\text{Cov}(Y, Z) \leq 0$, due to the fact that $F^{-1}$ is not decreasing and that $U$ and $1 - U$ are negatively correlated, and thus

$$\text{V}(\tilde{X}_n) \leq \text{V}(\bar{X}_n).$$

This technique is called *antithetic variables* in Monte Carlo theory.
Suppose now that $X$ is naturally sampled as $X = Y - Z$, $Y$ and $Z$ being two r.v. defined on the same space, and dependent.

Let us denote $\mathbb{V}(Y) = \sigma_Y^2$, $\mathbb{V}(Z) = \sigma_Z^2$, $\mathbb{Cov}(Y, Z) = C_{Y,Z}$.

The standard estimator of $\mu$ is simply

$$\bar{X}_n = \bar{Y}_n - \bar{Z}_n.$$ 

Its variance is

$$\mathbb{V}(\bar{X}_n) = \frac{1}{n} \left( \sigma_Y^2 + \sigma_Z^2 - 2C_{Y,Z} \right).$$

To build $\bar{Y}_n$ and $\bar{Z}_n$ we typically use $Y_i = F_Y^{-1}(U_{1,i})$ and $Z_j = F_Z^{-1}(U_{2,j})$ where the $U_{m,h}$, $m = 1, 2$, $h = 1, \cdots, n$, are iid Uniform(0,1) r.v. and $F_Y$, $F_Z$ are the respective cdf of $Y$ and $Z$. 

G. Rubino and B. Tuffin (INRIA) 
Monte Carlo & Rare Events 
QEST, Sept. 2009 18 / 72
Suppose now that we sample each pair \((Y_k, Z_k)\) with the same uniform r.v. \(U_k\): \(Y_k = F_Y^{-1}(U_k)\) and \(Z_k = F_Z^{-1}(U_k)\).

Using the fact that \(F_Y^{-1}\) and \(F_Z^{-1}\) are non increasing, we can easily prove that \(\text{Cov}(Y_k, Z_k) \geq 0\).

This means that if we define a new estimator \(\tilde{X}_n\) as

\[
\tilde{X}_n = \frac{1}{n} \sum_{k=1}^{n} [F_Y^{-1}(U_k) - F_Z^{-1}(U_k)]
\]

we have

\[
\mathbb{E}(\tilde{X}_n) = \mathbb{E}(\tilde{X}_n) = \mu,
\]

and

\[
\mathbb{V}(\tilde{X}_n) \leq \mathbb{V}(\tilde{X}_n).
\]

This technique is called \textit{common variables} in Monte Carlo theory.
Variance reduction: control variables

- Here, we suppose that there is an auxiliary r.v. $C$ correlated with $X$, with known mean $\mathbb{E}(C)$ and easy to sample.
- Define $\tilde{X} = X + \gamma(C - \mathbb{E}(C))$ for an arbitrary coefficient $\gamma > 0$. See that $\mathbb{E}(\tilde{X}) = \mu$.
- We have $\mathbb{V}(\tilde{X}) = \sigma^2 - 2\gamma \text{Cov}(X, C) + \gamma^2 \mathbb{V}(C)$.
- If $\text{Cov}(X, C)$ and $\mathbb{V}(C)$ are known, we set $\gamma = \text{Cov}(X, C)/\mathbb{V}(C)$ and we get
  $$\tilde{X} = \left(1 - \rho_{X,C}^2\right)\sigma^2 \leq \sigma^2,$$
  $\rho_{X,C}$ being the coefficient of correlation between $X$ and $C$. 

G. Rubino and B. Tuffin (INRIA) 
Monte Carlo & Rare Events 
QEST, Sept. 2009 20 / 72
Variance reduction: conditional Monte Carlo

- Assume we have an auxiliary r.v. $C$, correlated with $X$, such that $E(X | C)$ is available analytically and $C$ is easy to sample.
- Since $E[E(X | C)] = \mu$, the r.v. $E(X | C)$ is an unbiased estimator of $\mu$.
- From
  \[ \sigma^2 = \nabla(X) = \nabla[E(X | C)] + E[\nabla(X | C)], \]
  we get
  \[ \nabla[E(X | C)] = \sigma^2 - E[\nabla(X | C)] \leq \sigma^2 \]
  because $\nabla(X | C)$ and thus $E[\nabla(X | C)]$ are non negative.
- The corresponding estimator is
  \[ \tilde{X}_n = \frac{1}{n} \sum_{i=1}^{n} E(X | C_i). \]
So, is there any problem with Monte Carlo approach?
Main one: the rare event problem
Another one: specification/validation of models
This tutorial focuses on the main one
There are many techniques for facing the rare event problem:
- for example, we have the variance reduction techniques described before (there are other similar methods available);
- we will focus on the most effective ones in case of performance or dependability (or performability) problems: importance sampling and splitting.
On accuracy

- Resuming: how to improve the accuracy? *Acceleration*
  - either by decreasing the simulation time to get a replication
  - or by reducing the variance of the estimator.

- For rare events, acceleration required! (see next slide).
Outline

1 Introduction to rare events
2 Monte Carlo: the basics
3 Inefficiency of crude Monte Carlo, and robustness issue
4 Importance Sampling
5 Splitting
6 Confidence interval issues
7 Some applications
What is crude simulation?

- Assume we want to estimate $\mu = \mathbb{P}(A)$ for some rare event $A$.
- **Crude Monte Carlo**: simulates the model directly.
- Estimation

  $$
  \mu \approx \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i
  $$

  where the $X_i$ are i.i.d. copies of Bernoulli r.v. $X = 1_A$.
- $\sigma[X_i] = \mu(1 - \mu)$ for a Bernoulli r.v.
Inefficiency of crude Monte Carlo: relative error

- Confidence interval

\[
\left( \hat{\mu}_n - c_\alpha \sqrt{\frac{\mu(1-\mu)}{n}}, \hat{\mu}_n + c_\alpha \sqrt{\frac{\mu(1-\mu)}{n}} \right)
\]

estimated by

\[
\left( \hat{\mu}_n - c_\alpha \sqrt{\frac{\hat{\mu}_n(1-\hat{\mu}_n)}{n}}, \hat{\mu}_n + c_\alpha \sqrt{\frac{\hat{\mu}_n(1-\hat{\mu}_n)}{n}} \right)
\]

where \( c_\alpha \) is the \( 1 - \alpha / 2 \) quantile of the normal distribution, for \( n \) large enough (Student law used otherwise).

- Relative half width \( c_\alpha \sigma / (\sqrt{n} \mu) = c_\alpha \sqrt{(1 - \mu) / \mu / n} \to \infty \) as \( \mu \to 0 \).

- For a given relative error \( RE \), the required value of

\[
n = (c_\alpha)^2 \frac{1 - \varepsilon}{RE^2 \varepsilon},
\]

inversely proportional to \( \mu \).
Inefficiency of crude Monte Carlo: occurrence of the event

- To get a single occurrence, we need in average $1/\mu$ replications ($10^9$ for $\mu = 10^{-9}$).
- If no observation the returned interval is $(0, 0)$
- Otherwise, if (unlikely) one observation when $n \ll 1/\mu$, over-estimation of mean and variance
- In general, bad coverage of the confidence interval unless $n \gg 1/\mu$.
- As we can see, something has to be done to accelerate the occurrence (and reduce variance).
- An estimator has to be “robust” to the rarity of the event.
In rare-event simulation models, we often parameterize with a rarity parameter $\epsilon > 0$ such that $\mu = \mathbb{E}[X(\epsilon)] \to 0$ as $\epsilon \to 0$.

Typical example

- For a direct Bernoulli r.v. $X = 1_A$, $\epsilon = \mu = \mathbb{E}[1_A]$.
- When simulating a system involving failures and repairs, $\epsilon$ can be the rate or probability of individual failures.
- For a queue or a network of queues, when estimating the overflow probability, $\epsilon = 1/C$ inverse of the capacity of the considered queue.

The question is then: how does an estimator behave as $\epsilon \to 0$, i.e., the event becomes rarer?
Robustness properties: Bounded relative error (BRE)

- An estimator $X(\epsilon)$ is said to have \textit{bounded relative variance} (or \textit{bounded relative error}) if $\sigma^2(X(\epsilon))/\mu^2(\epsilon)$ is bounded uniformly in $\epsilon$. Equivalent to saying that $\sigma(X(\epsilon))/\mu(\epsilon)$ is bounded uniformly in $\epsilon$.

- Interpretation: estimating $\mu(\epsilon)$ with a given relative accuracy can be achieved with a bounded number of replications even if $\epsilon \to 0$.

- When the confidence interval comes from the central limit theorem, it means that the relative half width

$$c_\alpha \frac{\sigma(X(\epsilon))}{\sqrt{n}}$$

remains bounded as $\epsilon \to 0$. 
Robustness properties: Asymptotic Optimality (AO)

- BRE has often been found difficult to verify in practice (ex: queueing systems).
- Weaker property: asymptotic optimality (or logarithmic efficiency) if
  \[
  \lim_{\epsilon \to 0} \frac{\ln(\mathbb{E}[X^2(\epsilon)])}{\ln(\mu(\epsilon))} = 2.
  \]

  - Equivalent to say that \(\lim_{\epsilon \to 0} \frac{\ln(\sigma^2[X(\epsilon)])}{\ln(\mu(\epsilon))} = 2\).
  - Property also called logarithmic efficiency or weak efficiency.
  - Quantity under limit is always positive and less than or equal to 2: \(\sigma^2[X(\epsilon)] \geq 0\), so \(\mathbb{E}[X^2(\epsilon)] \geq (\mu(\epsilon))^2\) and then \(\ln \mathbb{E}[X^2(\epsilon)] \geq 2 \ln \mu(\epsilon)\), i.e.,
    \[
    \frac{\ln \mathbb{E}[X^2(\epsilon)]}{\ln \mu(\epsilon)} \leq 2.
    \]
  - Interpretation: the second moment and the square of the mean go to zero at the same exponential rate.
Relation between BRE and AO

- AO weaker property: if we have BRE, $\exists \kappa > 0$ such that $\mathbb{E}[X^2(\epsilon)] \leq \kappa^2 \mu^2(\epsilon)$, i.e., $\ln \mathbb{E}[X^2(\epsilon)] \leq \ln \kappa^2 + 2 \ln \mu(\epsilon)$, leading to $\lim_{\epsilon \to 0} \ln \mathbb{E}[X^2(\epsilon)] / \ln \mu(\epsilon) \geq 2$. Since this ratio is always less than 2, we get the limit 2.

- Not an equivalence. Some counter-examples:
  - an estimator for which $\gamma = e^{-\eta/\epsilon}$ with $\eta > 0$, but for which the variance is $Q(1/\epsilon)e^{-2\eta/\epsilon}$ with $Q$ a polynomial;
  - exponential tilting in queueing networks.

- Other robustness measures exist (based on higher degree moments, on the Normal approximation, on simulation time...)

G. Rubino and B. Tuffin (INRIA)
Monte Carlo & Rare Events
QEST, Sept. 2009
Work-normalized properties

- Variance is not all, generation time is important (figure of merit).
- Let $\sigma_n^2(\epsilon)$ and $t_n(\epsilon)$ be the variance and generation time $t_n(\epsilon)$ for a sample of size $n$.
- When $t_n(\epsilon)$ is strongly dependent on $\epsilon$, any behavior is possible: increasing or decreasing to 0 as $\epsilon \to 0$.
- Work-normalized versions of the above properties:
  - The estimator verifies work-normalized relative variance if
    
    $$\frac{\sigma_n^2(\epsilon)t_n(\epsilon)}{\mu^2(\epsilon)}$$

    is upper-bounded whatever the rarity, and is therefore a work-normalized version of the bounded relative error property.
  - The estimator verifies work-normalized asymptotic optimality if
    
    $$\lim_{\epsilon \to 0} \frac{\ln t_n(\epsilon) + \ln \sigma_n^2(\epsilon)}{\ln \mu(\epsilon)} = 2.$$
Outline

1. Introduction to rare events
2. Monte Carlo: the basics
3. Inefficiency of crude Monte Carlo, and robustness issue
4. Importance Sampling
5. Splitting
6. Confidence interval issues
7. Some applications
Importance Sampling (IS)

- Let $X = h(Y)$ for some function $h$ where $Y$ obeys some probability law $\mathbb{P}$.
- IS replaces $\mathbb{P}$ by another probability measure $\tilde{\mathbb{P}}$, using

$$E[X] = \int h(y)d\mathbb{P}(y) = \int h(y)\frac{d\mathbb{P}(y)}{d\tilde{\mathbb{P}}(y)}d\tilde{\mathbb{P}}(y) = \tilde{\mathbb{E}}[h(Y)L(Y)]$$

- $L = \frac{d\mathbb{P}}{d\tilde{\mathbb{P}}}$ likelihood ratio,
- $\tilde{\mathbb{E}}$ is the expectation associated with probability law $\tilde{\mathbb{P}}$.

- Required condition: $d\tilde{\mathbb{P}}(y) \neq 0$ when $h(y)d\mathbb{P}(y) \neq 0$.
- If $\mathbb{P}$ and $\tilde{\mathbb{P}}$ continuous laws, $L$ ratio of density functions $f(y)/\tilde{f}(y)$.

$$E[X] = \int h(y)f(y)dy = \int h(y)\frac{f(y)}{\tilde{f}(y)}\tilde{f}(y)dy = \tilde{\mathbb{E}}[h(Y)L(Y)].$$

- If $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are discrete laws, $L$ ratio of indiv. prob $p(y_i)/\tilde{p}(y_i)$

$$E[X] = \sum_{i} h(y_i)p(y_i) = \sum_{i} h(y_i)\frac{p(y_i)}{\tilde{p}(y_i)}\tilde{p}(y_i) = \tilde{\mathbb{E}}[h(Y)L(Y)].$$
Estimator and goal of IS

- Take \((Y_i, 1 \leq i \leq n)\) i.i.d; copies of \(Y\), according to \(\tilde{P}\). The estimator is
  \[
  \frac{1}{n} \sum_{i=1}^{n} h(Y_i) L(Y_i).
  \]

- The estimator is unbiased:
  \[
  \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} h(Y_i) L(Y_i) \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} [h(Y_i) L(Y_i)] = \mu.
  \]

- Goal: select probability law \(\tilde{P}\) such that
  \[
  \tilde{\sigma}^2 [h(Y) L(Y)] = \tilde{\mathbb{E}}[(h(Y) L(Y))^2] - \mu^2 < \sigma^2 [h(Y)].
  \]

- It means changing the probability distribution such that the 2nd moment is smaller.
IS difficulty: system with exponential failure time

- $Y$: exponential r.v. with rate $\lambda$.
- $A = \text{“failure before } T\text{”} = [0, T]$.
- Goal: compute $\mu = \mathbb{E}[1_A(Y)] = 1 - e^{-\lambda T}$.
- Use for IS an exponential density with a different rate $\tilde{\lambda}$

$$\tilde{\mathbb{E}}[(1_A(Y)L(Y))^2] = \int_0^T \left( \frac{\lambda e^{-\lambda y}}{\tilde{\lambda} e^{-\tilde{\lambda} y}} \right)^2 \tilde{\lambda} e^{-\tilde{\lambda} y} dy = \frac{\lambda^2(1 - e^{-(2\lambda - \tilde{\lambda})T})}{\tilde{\lambda}(2\lambda - \tilde{\lambda})}.$$

- Variance ratio for $T = 1$ and $\lambda = 0.1$:
If $A = [T, \infty)$, i.e., $\mu = \mathbb{P}[Y \geq T]$, and IS with exponential with rate $\tilde{\lambda}$:

$$\tilde{\mathbb{E}}[(1_A(Y)L(Y))^2] = \int_T^\infty \left(\frac{\lambda e^{-\lambda y}}{\tilde{\lambda} e^{-\tilde{\lambda} y}}\right)^2 \tilde{\lambda} e^{-\tilde{\lambda} y} dy = \frac{\lambda^2 e^{-(2\lambda-\tilde{\lambda})T}}{\tilde{\lambda}(2\lambda - \tilde{\lambda})}.$$ 

Minimal value computable, but infinite variance wen $\tilde{\lambda} > 2\lambda$. If $\lambda = 1$: 

![Graph showing variance ratio vs. \tilde{\lambda} for different values of \lambda]
Optimal estimator for estimating $\mathbb{E}[h(Y)] = \int h(y)L(y)d\tilde{P}(y)$

- **Optimal change of measure:**
  \[
  \tilde{P} = \frac{|h(Y)|}{\mathbb{E}[|h(Y)|]}dP.
  \]

- **Proof:** for any alternative IS measure $P'$, leading to the likelihood ratio $L'$ and expectation $\mathbb{E}'$,
  \[
  \tilde{\mathbb{E}}[(h(Y)L(Y))^2] = (\mathbb{E}[|h(Y)|])^2 = (\mathbb{E}'[|h(Y)|L'(Y)])^2 \leq \mathbb{E}'[(h(Y)L'(Y))^2].
  \]

- If $h \geq 0$, $\tilde{\mathbb{E}}[(h(Y)L(Y))^2] = (\mathbb{E}[h(Y)])^2$, i.e., $\tilde{\sigma}^2(h(Y)L(Y)) = 0$. That is, IS provides a zero-variance estimator.

- Implementing it requires knowing $\mathbb{E}[|h(Y)|]$, i.e. what we want to compute; if so, no need to simulation!

- But provides a hint on the general form of a “good” IS measure.
IS for a discrete-time Markov chain (DTMC) \( \{ Y_j, j \geq 0 \} \)

- \( X = h(Y_0, \ldots, Y_\tau) \) function of the sample path with
  - \( P = (P(y, z)) \) transition matrix, \( \pi_0(y) = \mathbb{P}[Y_0 = y] \), initial probabilities
  - up to a stopping time \( \tau \), first time it hits a set \( \Delta \).
  - \( \mu(y) = \mathbb{E}_y[X] \).

- IS replaces the probabilities of paths \( (y_0, \ldots, y_n) \),
  \[
  \mathbb{P}[(Y_0, \ldots, Y_\tau) = (y_0, \ldots, y_n)] = \pi_0(y_0) \prod_{j=1}^{n-1} P(y_{j-1}, y_j),
  \]
  by \( \tilde{\mathbb{P}}[(Y_0, \ldots, Y_\tau) = (y_0, \ldots, y_n)] \) st \( \mathbb{E}[\tau] < \infty \).

- For convenience, the IS measure remains a DTMC, replacing \( P(y, z) \) by \( \tilde{P}(y, z) \) and \( \pi_0(y) \) by \( \tilde{\pi}_0(y) \).

- Then \( L(Y_0, \ldots, Y_\tau) = \frac{\pi_0(Y_0)}{\tilde{\pi}_0(Y_0)} \prod_{j=1}^{\tau-1} \frac{P(Y_{j-1}, Y_j)}{\tilde{P}(Y_{j-1}, Y_j)} \).
Illustration: a birth-death process

- Markov chain with state-space \{0, 1, \ldots, B\}, \( P(y, y + 1) = p_y \) and \( P(y, y - 1) = 1 - p_y \), for \( y = 1, \ldots, B - 1 \)
- \( \Delta = \{0, B\} \), and let \( \mu(y) = \mathbb{P}[Y_\tau = B \mid Y_0 = y] \).
- Rare event if \( B \) large or the \( p_y \)s are small.
- If \( p_y = p < 1 \) for \( y = 1, \ldots, B - 1 \), known as the gambler’s ruin problem.
- An \( M/M/1 \) queue with arrival rate \( \lambda \) and service rate \( \mu > \lambda \) fits the framework with \( p = \lambda / (\lambda + \mu) \).
- How to apply IS: increase the \( p_y \)s to \( \tilde{p}_y \) to accelerate the occurrence (but not too much again).
- Large deviation theory applies here, when \( B \) increases.
  - Strategy for an \( M/M/1 \) queue: exchange \( \lambda \) and \( \mu \)
  - Asymptotic optimality, but no bounded relative error.
Zero-variance IS estimator for Markov chains simulation

- Restrict to an additive (positive) cost

\[ X = \sum_{j=1}^{\tau} c(Y_{j-1}, Y_j) \]

- Is there a Markov chain change of measure yielding zero-variance?
- Yes we have zero variance with

\[ \tilde{P}(y, z) = \frac{P(y, z)(c(y, z) + \mu(z))}{\sum_{w} P(y, w)(c(y, w) + \mu(w))} = \frac{P(y, z)(c(y, z) + \mu(z))}{\mu(y)}. \]

- Without the additivity assumption the probabilities for the next state must depend in general of the entire history of the chain.
Zero-variance for Markov chains

- Proof by induction on the value taken by $\tau$, using the fact that $\mu(Y_\tau) = 0$. In that case, if $\tilde{X}$ denotes the IS estimator,

$$
\tilde{X} = \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{P(Y_{j-1}, Y_j)}{\tilde{P}(Y_{j-1}, Y_j)}
$$

$$
= \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{P(Y_{j-1}, Y_j)\mu(Y_{j-1})}{P(Y_{j-1}, Y_j)(c(Y_{j-1}, Y_j) + \mu(Y_j))}
$$

$$
= \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\mu(Y_{j-1})}{c(Y_{j-1}, Y_j) + \mu(Y_j)}
$$

$$
= \mu(Y_0)
$$

- Unique Markov chain implementation of the zero-variance estimator.
- Again, implementing it requires knowing $\mu(y)$ $\forall y$, the quantities we wish to compute.
- Approximation to be used.
Zero-variance approximation

- Use a heuristic approximation \( \hat{\mu}(\cdot) \) and plug it into the zero-variance change of measure instead of \( \mu(\cdot) \).
- More efficient but also more requiring technique: learn adaptively function \( \mu(\cdot) \), and still plug the approximation into the zero-variance change of measure formula instead of \( \mu(\cdot) \).
  - Adaptive Monte Carlo (AMC) proceeds iteratively.
    - Considers several steps and \( n_i \) independent simulation replications at step \( i \).
    - At step \( i \), replaces \( \mu(x) \) by a guess \( \mu^{(i)}(x) \).
    - Use probabilities
      \[
      \tilde{P}^{(i)}_{y,z} = \frac{P_{y,z}(c_{y,z} + \mu^{(i)}(z))}{\sum_w P_{y,w}(c_{y,w} + \mu^{(i)}(w))}.
      \]
    - Gives a new estimation \( \mu^{(i+1)}(y) \) of \( \mu(y) \), from which a new transition matrix \( \tilde{P}^{(i+1)} \) is defined.
Adaptive stochastic approximation (ASA)

- ASA just uses a single sample path \((y_0, \ldots, y_n)\).
- Initial distribution for \(y_0\), matrix \(\tilde{P}(0)\) and guess \(\mu(0)(\cdot)\).
- At step \(j\) of the path, if \(y_j \notin \Delta\),
  - matrix \(\tilde{P}(j)\) used to generate \(y_{j+1}\).
  - From \(y_{j+1}\), update the estimate of \(\mu(y_j)\) by
    \[
    \mu^{(j+1)}(y_j) = (1 - a_j(y_j))\mu^{(j)}(y_j) \\
    + a_j(y_j) \left[ c(y_j, y_{j+1}) + \mu^{(j)}(y_{j+1}) \right] \frac{P(y_j, y_{j+1})}{\tilde{P}(j)(y_j, y_{j+1})},
    \]
    where \(\{a_j(y), j \geq 0\}\), sequence of step sizes
  - For \(\delta > 0\) constant,
    \[
    \tilde{P}^{(j+1)}(y_j, y_{j+1}) = \max \left( P(y_j, y_{j+1}) \left[ c(y_j, y_{j+1}) + \mu^{(j+1)}(y_{j+1}) \right], \frac{\mu^{(j+1)}(y_j)}{\mu^{(j+1)}(y_j)} \right), \delta \right).
    \]
  - Otherwise \(\mu^{(j+1)}(y) = \mu^{(j)}(y), \tilde{P}^{(j+1)}(y, z) = P^{(j)}(y, z)\).
  - Normalize: \(P^{(j+1)}(y_j, y) = \frac{\tilde{P}^{(j+1)}(y_j, y)}{\sum_z \tilde{P}^{(j+1)}(y_j, z)}\).
- If \(y_j \in \Delta\), \(y_{j+1}\) generated from initial distribution, but estimations of \(P(\cdot, \cdot)\) and \(\mu(\cdot)\) kept.
- Batching techniques used to get a confidence interval.
Drawbacks of the learning techniques

- You have to store vectors $\mu^{(n)}(\cdot)$. State-space typically very large when we use simulation...
- This limits the practical effectiveness of the method.
- Other possibility:
  - Use $K$ basis functions $\mu^{(1)}(\cdot), \ldots, \mu^{(K)}(\cdot)$, and an approximation
    $$\mu(\cdot) \equiv \sum_{k=1}^{K} \alpha_k \mu^{(k)}(\cdot).$$
  - Learn coefficients $\alpha_k$ as in previous methods, instead of the function itself.
  - See also how best basis functions can be learnt, as done in dynamic programming.
Illustration of heuristics: birth-death process

- Let $P(i, i + 1) = p$ and $P(i, i - 1) = 1 - p$ for $1 \leq i \leq B - 1$, and $P(0, 1) = P(B, B - 1) = 1$.

- We want to compute $\mu(1)$, probability of reaching $B$ before coming back to 0.

- If $p$ small, to approach $\mu(\cdot)$, we can use

$$
\hat{\mu}(y) = p^{B-y} \quad \forall y \in \{1, \ldots , B-1\}
$$

with $\hat{\mu}(0) = 0$ and $\hat{\mu}(B) = 1$ based on the asymptotic estimate $\mu(i) = p^{B-i} + o(p^{B-i})$.

- We can verify that the variance of this estimator is going to 0 (for fixed sample size) as $p \to 0$. 
Other procedure: optimization within a parametric class

- No direct relation with the zero-variance change of measure.
- Parametric class of IS measures depending on vector $\theta$, $\{\tilde{P}_\theta, \theta \in \Theta\}$:
  - family of densities $\tilde{f}_\theta$, or of discrete probability vectors $\tilde{p}_\theta$.
- Find
  \[ \theta^* = \arg\max_\theta \mathbb{E}_\theta [(h(Y)L(Y))^2]. \]
- The optimization can sometimes be performed analytically
  - Ex: estimate $\mu = \mathbb{P}[X \geq na]$ for $X \text{ Binomial}(n, p)$
  - IS parametric family $\text{Binomial}(n, \theta)$.
  - Twisting the parameter $p$ to $\theta = a$ is optimal (from Large Deviations theory).
Adaptive learning of the best parameters

- The value of $\theta$ that minimize the variance can be learned adaptively in various ways.
- ASA method can be adapted to optimize $\theta$ by stochastic approximation.
- We may replace the variance in the optimization problem by some distance between $\tilde{P}_\theta$ and the optimal $d\tilde{P}^* = (|X|/E[|X|])dP$, simpler to optimize.
- Cross-entropy technique uses the Kullback-Leibler "distance"

$$D(\tilde{P}^*, \tilde{P}_\theta) = \tilde{E}^* \left[ \log \frac{d\tilde{P}^*}{d\tilde{P}_\theta} \right]$$

$$= E \left[ \frac{|X|}{E[|X|]} \log \left( \frac{|X|}{E[|X|]} dP \right) \right] - \frac{1}{E[|X|]} E \left[ |X| \log d\tilde{P}_\theta \right].$$

- Determine then

$$\max_{\theta \in \Theta} E \left[ |X| \log d\tilde{P}_\theta \right] = \max_{\theta \in \Theta} \tilde{E} \left[ \frac{dP}{d\tilde{P}} |X| \log d\tilde{P}_\theta \right].$$
Adaptive learning in Cross-Entropy (CE)

- CE method applied in an iterative manner, increasing the rarity at each step.
- Start with $\theta_0 \in \Theta$ and r.v. $X_0$ whose expectation is easier to estimate than $X$.
- At step $i \geq 0$, $n_i$ independent simulations are performed using IS with $\theta_i$, to approximate the previous maximization ($\tilde{P}$ replaced by $\tilde{P}_{\theta_i}$)
- Solution of the corresponding sample average problem

$$
\theta_{i+1} = \arg \max_{\theta \in \Theta} \frac{1}{n_i} \sum_{j=1}^{n_i} |X_i(\omega_{i,j})| \log(\frac{d\tilde{P}_\theta(\omega_{i,j})}{d\tilde{P}_{\theta_i}}(\omega_{i,j}))
$$

where $\omega_{i,j}$ represents the $j$th sample at step $i$.
- Kullback-Leibler distance is convenient for the case where $\tilde{P}_\theta$ is from an exponential family, because the log and the exponential cancel.
Outline

1. Introduction to rare events
2. Monte Carlo: the basics
3. Inefficiency of crude Monte Carlo, and robustness issue
4. Importance Sampling
5. Splitting
6. Confidence interval issues
7. Some applications
Splitting: general principle

- Splitting is the other main rare event simulation technique.
- Assume we want to compute the probability $P(D)$ of an event $D$.
- General idea:
  - Decompose $D_1 \supset \cdots \supset D_m = D$,
  - Use $P(D) = P(D_1)P(D_2 \mid D_1) \cdots P(D_m \mid D_{m-1})$, each conditional event being “not rare”,
  - Estimate each individual conditional probability by crude Monte Carlo, i.e., without changing the laws driving the model.
  - The final estimate is the product of individual estimates.
- Question: how to do it for a stochastic process? Difficult to sample conditionally to an intermediate event.
Graphical interpretation

\[ D_1 \supseteq D_2 \supseteq D_3 = D \]
Splitting and Markov chain \( \{ Y_j; j \geq 0 \} \in \mathcal{Y} \)

- **Goal:** compute \( \gamma_0 = \mathbb{P}[\tau_B < \tau_A] \) with
  - \( \tau_A = \inf\{ j > 0 : Y_{j-1} \notin A \text{ and } Y_j \in A \} \)
  - \( \tau_B = \inf\{ j > 0 : Y_j \in B \} \)

- **Intermediate levels from importance function** \( h : \mathcal{Y} \to \mathbb{R} \) with
  \( A = \{ x \in \mathcal{Y} : h(x) \leq 0 \} \) and \( B = \{ x \in \mathcal{Y} : h(x) \geq \ell \} \):
  - Partition \([0, \ell]\) in \( m \) subintervals with boundaries \( 0 = \ell_0 < \ell_1 < \cdots < \ell_m = \ell \).
  - Let \( T_k = \inf\{ j > 0 : h(Y_j) \geq \ell_k \} \) and \( D_k = \{ T_k < \tau_A \} \).

- **1st stage:**
  - simulate \( N_0 \) chains until \( \min(\tau_A, T_1) \).
  - If \( R_1 \) number of chains for which \( D_1 \) occurs, \( \hat{p}_1 = R_1/N_0 \) unbiased estimator of \( p_1 = \mathbb{P}(D_1) \).

- **Stage 1 < \( k \leq m \):**
  - If \( R_{k-1} = 0 \), \( \hat{p}_l = 0 \) for all \( l \geq k \) and the algorithm stops
  - Otherwise, start \( N_k \) chains from these \( R_k \) entrance states, by potentially cloning (splitting) some chains
  - simulate these chains up to \( \min(\tau_A, T_k) \).
  - \( \hat{p}_k = R_k/N_{k-1} \) unbiased estimator of \( p_k = \mathbb{P}(D_k|D_{k-1}) \).
Two-dimensional illustration
The different implementations

- **Fixed splitting:**
  - clone each of the $R_k$ chains reaching level $k$ in $c_k$ copies, for a fixed positive integer $c_k$.
  - $N_k = c_k R_k$ is random.

- **Fixed effort:**
  - $N_k$ fixed a priori
  - *random assignment* draws the $N_k$ starting states at random, with replacement, from the $R_k$ available states.
  - *fixed assignment*, on the other hand, we would split each of the $R_k$ states approximately the same number of times.
  - Fixed assignment gives a smaller variance than random assignment because it amounts to using stratified sampling over the empirical distribution $G_k$ at level $k$.

- Fixed splitting can be implemented in a depth-first way, recursively, while fixed effort cannot.

- On the other hand, you have no randomness (less variance) in the number of chains with fixed effort.
Diminishing the computational effort

- As $k$ increases, it is likely that the average time before reaching the next level or going back to $A$ increases significantly.
- We can kill (truncate) trajectories that go a given number $\beta$ of levels down (unlikely to come back), but biased.
- Unbiased solution: apply the Russian roulette principle
  - kill the trajectory going down with a probability $r_\beta$. If it survives, assign a multiplicative weight $1/(1 - r_\beta)$.
  - Several possible implementations to reduce the variance due to the introduction of weights.
Issues to be solved

- **How to define the importance function $h$?**
  - If the state space is one-dimensional and included in $\mathbb{R}$, the final time is an almost surely finite stopping time and the critical region is $B = [b, \infty)$, any strictly increasing function would be good (otherwise a mapping can be constructed, by just moving the levels), such as for instance $h(x) = x$.
  - If the state space is multidimensional: the importance function is a one-dimensional projection of the state space.
  - Desirable property: the probability to reach the next level should be the same, whatever the entrance state in the current level.
  - Ideally, $h(x) = \mathbb{P}[\tau_B \leq \tau_A \mid X(0) = x]$, but as in IS, they are a probabilities we are looking for.
  - This $h(\cdot)$ can also be learnt or estimated *a priori*, with a presimulation, by partitionning the state space and assuming it constant on each region.
Issues to be solved (2)

- **How many offsprings at each level?**
  - In fixed splitting:
    - if $c_k < 1/p_k$, we do not split enough, it will become unlikely to reach the next event;
    - if $c_k > 1/p_k$, the number of trajectories will exponentially explode with the number of levels.
    - The right amount is $c_k = 1/p_k$ ($c_k$ can be randomized to reach that value if not an integer).
  - In fixed effort, no explosion is possible.
  - In both cases, the right amount has to be found.

- **How many levels to define?**
  - i.e., what probability to reach the next level?
In a general setting, very few results exist:

- We only have a central limit theorem based on genetic type interacting particle systems, as the sample increases.
- Nothing exist on the definition of optimal number of levels...

Consider the simplified setting, with a single entrance state at each level.

Similar to coin-flipping to see if next level is reached or not.

In that case, asymptotically optimal results can be derived, providing hints of values to be used.
Simplified setting and fixed effort

\( N_0 = N_1 = \cdots = N_{m-1} = n \)

The \( \hat{p}_k \)'s binomial r.v. with parameters \( n \) and \( p_k = p = \mu_0^{1/m} \) assumed independent.

It can be shown that

\[
\text{Var}[\hat{p}_1 \cdots \hat{p}_m] = \prod_{k=1}^{m} \mathbb{E}[\hat{p}_k^2] - \gamma_0^2 = \left( p^2 + \frac{p(1-p)}{n} \right)^{m} - p^{2m}
\]

\[
= \frac{mp^{2m-1}(1-p)}{n} + \cdots + \frac{(p(1-p))^m}{n^m}.
\]

Assuming \( n \gg (m-1)(1-p)/p \),

\[
\text{Var}[\hat{p}_1 \cdots \hat{p}_m] \approx mp^{2m-1}(1-p)/n \approx m\gamma_0^{2-1/m}/n.
\]

The work normalized variance \( \approx \left[ \gamma_0^n m^2 \right]/n = \gamma_0^{2-1/m} m^2 \)

Minimized at \( m = -\ln(\gamma_0)/2 \)

This gives \( p^m = \gamma_0 = e^{-2m} \), so \( p = e^{-2} \).

But the relative error and its work-normalized version both increase toward infinity at a logarithmic rate.

There is no asymptotic optimality either.
Simplified setting: fixed splitting

- $N_0 = n$, $p_k = p = \gamma_0^{1/m}$ for all $k$, and $c = 1/p$; i.e., $N_k = R_k/p$.
- The process $\{N_k, k \geq 1\}$ is a branching process.
- From standard branching process theory

$$\text{Var}[\hat{p}_1 \cdots \hat{p}_m] = m(1 - p)p^{2m-1}/n.$$ 

- If $p$ fixed and $m \to \infty$, the squared relative error $m(1 - p)/(np)$ is unbounded,
- But it is asymptotically efficient:

$$\lim_{\gamma_0 \to 0^+} \frac{\log(E[\tilde{\gamma}_n^2])}{\log \gamma_0} = \lim_{\gamma_0 \to 0^+} \frac{\log(m(1 - p)\gamma_0^2/(np) + \gamma_0^2)}{\log \gamma_0} = 2.$$ 

- Fixed splitting is asymptotically better, but it is more sensitive to the values used.
Illustrative simple example: a tandem queue

- Illustrative of the impact of the importance function.
- Two queues in tandem
  - arrival rate at the first queue is \( \lambda = 1 \)
  - mean service time is \( \rho_1 = 1/4, \rho_2 = 1/2. \)
  - Embedded DTMC: \( Y = (Y_j, j \geq 0) \) with \( Y_j = (Y_{1,j}, Y_{2,j}) \) number of customers in each queue after the \( j \)th event
  - \( B = \{(x_1, x_2) : x_2 \geq L = 30\}, A = \{(0, 0)\}. \)

- Goal: impact of the choice of the importance function?
- Importance functions:
  
  \[
  \begin{align*}
  h_1(x_1, x_2) &= x_2, \\
  h_2(x_1, x_2) &= (x_2 + \min(0, x_2 + x_1 - L))/2, \\
  h_3(x_1, x_2) &= x_2 + \min(x_1, L - x_2 - 1) \times (1 - x_2/L).
  \end{align*}
  \]
Illustration, fixed effort: a tandem queue (2)

- $V_N$: variance per chain, ($N$ times the variance of the estimator) and the work-normalized variance per chain, $W_N = S_N V_N$, where $S_N$ is the expected total number of simulated steps of the $N$ Markov chains.
- With $h_1$, $\hat{V}_N$ and $\hat{W}_N$ were significantly higher than for $h_2$ and $h_3$.
- Estimators rescaled as $\tilde{V}_N = 10^{18} \times \hat{V}_N$ and $\tilde{W}_N = 10^{15} \times \hat{W}_N$.

<table>
<thead>
<tr>
<th></th>
<th>$N = 2^{10}$</th>
<th>$N = 2^{12}$</th>
<th>$N = 2^{14}$</th>
<th>$N = 2^{16}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tilde{V}_N$</td>
<td>$\tilde{W}_N$</td>
<td>$\tilde{V}_N$</td>
<td>$\tilde{W}_N$</td>
</tr>
<tr>
<td>$h_2$, Splitting</td>
<td>109</td>
<td>120</td>
<td>89</td>
<td>98</td>
</tr>
<tr>
<td>$h_2$, Rus. Roul.</td>
<td>178</td>
<td>67</td>
<td>99</td>
<td>37</td>
</tr>
<tr>
<td>$h_3$, Splitting</td>
<td>93</td>
<td>103</td>
<td>110</td>
<td>121</td>
</tr>
<tr>
<td>$h_3$, Rus. Roul.</td>
<td>90</td>
<td>34</td>
<td>93</td>
<td>35</td>
</tr>
</tbody>
</table>
Outline

1. Introduction to rare events
2. Monte Carlo: the basics
3. Inefficiency of crude Monte Carlo, and robustness issue
4. Importance Sampling
5. Splitting
6. Confidence interval issues
7. Some applications
Confidence interval issues

- Robustness is an issue, but what about the confidence interval validity?
- If the rare event has not occurred: empirical confidence interval is $(0, 0)$.
- The problem can even be more underhand: it may happen that the rare event happens due to some trajectories, but other important trajectories important for the variance estimation are still rare and not sampled: the empirical confidence confidence interval is not good then.
Illustrative example of the difficulty

- 4-component system with two classes of components, subject to failures and repairs. Discrete time Markov Chain
- $\mu$ probability starting from (2, 2) to we reach a down state before coming back to (2, 2).
IS probability used

- Failure Biasing scheme: for each up state $\neq (2, 2)$, we increase the probability of failure to the constant $q$ (ex: 0.8) and use individual probabilities proportional to the original ones.

\[
\begin{align*}
1 - q & \quad 1 \\
2, 1 & \quad (1 - q)/2 \\
1, 1 & \quad q\varepsilon \\
0 & \quad q/2 \\
2, 2 & \quad 1 \quad \varepsilon_q \\
1, 2 & \quad 1 - q/2 \\
\end{align*}
\]
Empirical evaluation as $\epsilon \to 0$

- Fix the number of samples, $n = 10^4$, using the same pseudo-random number generator, and varying $\epsilon$ from $10^{-2}$ down to 0.
- Remember that $\mu = 2\epsilon^2 + o(\epsilon^2)$ and $\sigma_{IS}^2 = \Theta(\epsilon^3)$.

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>$2\epsilon^2$</th>
<th>Est.</th>
<th>Confidence Interval</th>
<th>Est. RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e-02</td>
<td>2e-04</td>
<td>2.03e-04</td>
<td>( 1.811e-04 , 2.249e-04 )</td>
<td>1.08e-01</td>
</tr>
<tr>
<td>1e-03</td>
<td>2e-06</td>
<td>2.37e-06</td>
<td>( 1.561e-06 , 3.186e-06 )</td>
<td>3.42e-01</td>
</tr>
<tr>
<td>2e-04</td>
<td>8e-08</td>
<td>6.48e-08</td>
<td>( 1.579e-08 , 1.138e-07 )</td>
<td>7.56e-01</td>
</tr>
<tr>
<td>1e-04</td>
<td>2e-08</td>
<td>9.95e-09</td>
<td>( 9.801e-09 , 1.010e-08 )</td>
<td>1.48e-02</td>
</tr>
<tr>
<td>1e-06</td>
<td>2e-12</td>
<td>9.95e-13</td>
<td>( 9.798e-13 , 1.009e-12 )</td>
<td>1.48e-02</td>
</tr>
<tr>
<td>1e-08</td>
<td>2e-16</td>
<td>9.95e-17</td>
<td>( 9.798e-17 , 1.009e-16 )</td>
<td>1.48e-02</td>
</tr>
</tbody>
</table>

- The estimated value becomes bad as $\epsilon \to 0$.
- It seems that BRE is verified while it is not!
Asymptotic explanation

- When $\varepsilon$ small, transitions in $\Theta(\varepsilon)$ not sampled anymore.
- Asymptotic view of the Markov chain:
Asymptotic explanation (2)

- For this system:
  - the expectation is $\epsilon^2 + o(\epsilon^2)$
  - variance $\frac{1-q^2}{nq^2} \epsilon^4 + o(\epsilon^4)$.

- Results in accordance to the numerical values, and BRE is obtained.
- But does not correspond to the initial system, with different values.
- Reason: important paths are still rare under this IS scheme.
- Diagnostic procedures can be imagined.
Outline

1. Introduction to rare events
2. Monte Carlo: the basics
3. Inefficiency of crude Monte Carlo, and robustness issue
4. Importance Sampling
5. Splitting
6. Confidence interval issues
7. Some applications
Some applications

- HRMS (Highly Reliable Markovian Systems): IS examples
- STATIC MODELS (Network Reliability):
  - a recursive variance reduction technique
  - reducing time instead of variance