Rare-Event Simulation of Regenerative Systems: Estimation of the Mean and Distribution of Hitting Times

Bruno Tuffin
Based on joint works with P. L'Ecuyer, P. Glynn and M. Nakayama

The 12th International Conference on Monte Carlo Methods and Applications
July 8-12, 2019
Sydney, Australia

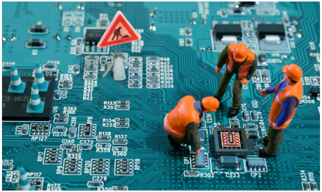
Outline

Introduction: rare events and dependability

- 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),
- ...

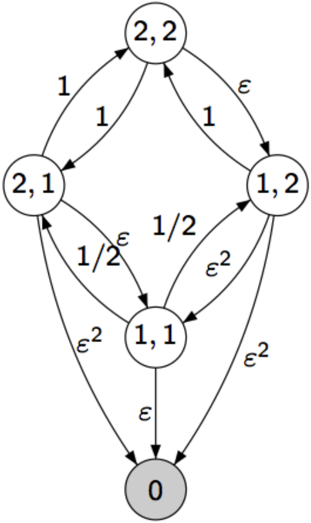
Context: Time To Failure (TTF) estimation

- Dependability analysis is of primary importance in many areas
  - nuclear power plants
  - telecommunications
  - manufacturing
  - transport systems
  - computer science
- Focus on the time to failure (TTF): random time to reach failure
- Even for Markov chains, models usually so large
  $\Rightarrow$ computation by simulation



Example: Highly Reliable Markovian Systems (HRMS)

- System with c types of components. $X = (S_1, \ldots, X_c)$ with X_i number of up components.
- Markov chain. Failure rates are $O(\varepsilon)$, but not repair rates. Failure propagations possible.
- System down when in grey state(s)
- Goal:
  - compute p probability from $(2, 2)$ to hit failure before being back $(2, 2)$: small if ε small.
  - compute TTF: long time if ε small.



$\mathcal{S}$-valued regenerative process $X = (X(t) : t \geq 0)$

- Goal: Compute $\alpha = \mathbb{E}[T]$, where

$$T = \inf\{t \geq 0 : X(t) \in \mathcal{A}\}$$

is the hitting time of subset $\mathcal{A}$

$\mathcal{S}$-valued regenerative process $X = (X(t) : t \geq 0)$

- Goal: Compute $\alpha = \mathbb{E}[T]$, where
$$T = \inf\{t \geq 0 : X(t) \in \mathcal{A}\}$$
is the hitting time of subset $\mathcal{A}$
- Regeneration times $0 = \Gamma(0) < \Gamma(1) < \cdots$,
with iid cycles $((\tau(k), (X(\Gamma(k-1)+s) : 0 \leq s < \tau(k)) : k \geq 1)$
- $\tau(k) = \Gamma(k) - \Gamma(k-1)$, length of the kth regenerative cycle

$\mathcal{S}$-valued regenerative process $X = (X(t) : t \geq 0)$

- Goal: Compute $\alpha = \mathbb{E}[T]$, where

$$T = \inf\{t \geq 0 : X(t) \in \mathcal{A}\}$$

  is the hitting time of subset $\mathcal{A}$
- Regeneration times $0 = \Gamma(0) < \Gamma(1) < \cdots$,
  with iid cycles $((\tau(k), (X(\Gamma(k-1)+s) : 0 \leq s < \tau(k)) : k \geq 1)$
- $\tau(k) = \Gamma(k) - \Gamma(k-1)$, length of the kth regenerative cycle

Ratio expression: $\alpha = \dfrac{\mathbb{E}[T \wedge \tau]}{\mathbb{P}(T < \tau)}$.

$\mathcal{S}$-valued regenerative process $X = (X(t) : t \geq 0)$

- Goal: Compute $\alpha = \mathbb{E}[T]$, where

$$T = \inf\{t \geq 0 : X(t) \in \mathcal{A}\}$$

  is the hitting time of subset $\mathcal{A}$
- Regeneration times $0 = \Gamma(0) < \Gamma(1) < \cdots$,
  with iid cycles $((\tau(k), (X(\Gamma(k-1)+s) : 0 \leq s < \tau(k)) : k \geq 1)$
- $\tau(k) = \Gamma(k) - \Gamma(k-1)$, length of the kth regenerative cycle

Ratio expression: $\alpha = \dfrac{\mathbb{E}[T \wedge \tau]}{\mathbb{P}(T < \tau)}$.

$$\begin{aligned}
\alpha &= \mathbb{E}[T; T < \tau] + \mathbb{E}[\tau + T - \tau; T > \tau] \\
&= \mathbb{E}[T; T < \tau] + \mathbb{E}[\tau; T > \tau] + \mathbb{E}[T - \tau; T > \tau] \\
&= \mathbb{E}[T \wedge \tau; T < \tau] + \mathbb{E}[T \wedge \tau; T > \tau] + \mathbb{E}[T - \tau \mid T > \tau]\,\mathbb{P}(T > \tau) \\
&= \mathbb{E}[T \wedge \tau] + \alpha(1 - \mathbb{P}(T < \tau))
\end{aligned}$$

Regenerative simulation

- $W(k) = \inf\{t \geq 0 : X(\Gamma(k-1)+t) \in \mathcal{A}\}$ first hitting to $\mathcal{A}$ after regeneration $\Gamma(k-1)$
- $I(k) = \mathcal{I}(W(k) < \tau(k))$ with $\mathcal{I}(\cdot)$ is the indicator function

Definition (Ratio estimator)

$$\hat{\alpha}(n) = \frac{(1/n)\sum_{k=1}^{n}[W(k) \wedge \tau(k)]}{(1/n)\sum_{k=1}^{n} I(k)}.$$

Proposition (Central Limit Theorem)

$$n^{1/2}[\hat{\alpha}(n) - \alpha] \Rightarrow \sigma_2 \mathcal{N}(0,1)$$

as $n \to \infty$, *where* $\sigma_2^2 = \frac{\mathbb{E}[(T \wedge \tau)^2]}{p^2} - 2\alpha \frac{\mathbb{E}[T\mathcal{I}(T<\tau)]}{p^2} + \frac{\alpha^2}{p}$.

Rare events: hitting $\mathcal{A}$ rarely occurs before τ

- Denominator p in $\alpha = \frac{\mathbb{E}[T \wedge \tau]}{\mathbb{P}(T < \tau)}$ a small probability
  $\Longrightarrow$ requires an acceleration technique
- Fraction β of cycles used to estimate the numerator with crude MC
- Fraction $1 - \beta$ to estimate the denominator with a variance reduction technique

Inefficiency of crude Monte Carlo for the denominator

- Compute the denominator/probability $p = \mathbb{E}[1_{[T<\tau]}] << 1$
- n iiid Y_i Bernoulli r.v.: 1 if the event is hit and 0 otherwise.
- To get a single occurence, we need in average $1/p$ replications (10^9 for $p = 10^{-9}$), and more to get a confidence interval.
- In most cases, you will get $(0, 0)$ as a confidence interval.
- $n\bar{Y}_n$ Binomial with parameters (n, p) and the confidence interval is

$$\left(\bar{Y}_n - \frac{c_\beta\sqrt{p(1-p)}}{\sqrt{n}}, \ \bar{Y}_n + \frac{c_\beta\sqrt{p(1-p)}}{\sqrt{n}}\right).$$

- *Relative half width* $c_\beta\sigma/(\sqrt{n}p) = c_\beta\sqrt{(1-p)/p/n} \to \infty$ as $p \to 0$.
- For a given relative error RE , the required value of

$$n = (c_\beta)^2 \frac{1-p}{RE^2 p},$$

inversely proportional to p.
- Two main families of techniques:
  - Splitting (also called *subset simulation*) and Importance Sampling.

Robustness properties

- In rare-event simulation models, we often parameterize with a rarity parameter $\epsilon > 0$ such that $\mu = \mathbb{E}[Y(\epsilon)] \to 0$ as $\epsilon \to 0$.
- An estimator $Y(\epsilon)$ is said to have *bounded relative variance* (or *bounded relative error*) if $\sigma^2(Y(\epsilon))/\mu^2(\epsilon)$ is bounded uniformly in ϵ.
  - Interpretation: estimating $\mu(\epsilon)$ with a given relative accuracy can be achieved with a bounded number of replications even if $\epsilon \to 0$.
- Weaker property: *asymptotic optimality* (or *logarithmic efficiency*) if $\lim_{\epsilon\to 0} \ln(\mathbb{E}[Y^2(\epsilon)])/\ln(\mu(\epsilon)) = 2$.
- Stronger property: *vanishing relative variance*: $\sigma^2(Y(\epsilon))/\mu^2(\epsilon) \to 0$ as $\epsilon \to 0$. Asymptotically, we get the zero-variance estimator.
- Other robustness measures exist (based on higher degree moments, on the Normal approximation, on simulation time...).

L'Ecuyer, Blanchet, T., Glynn, ACM ToMaCS 2010

Importance Sampling (IS)

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

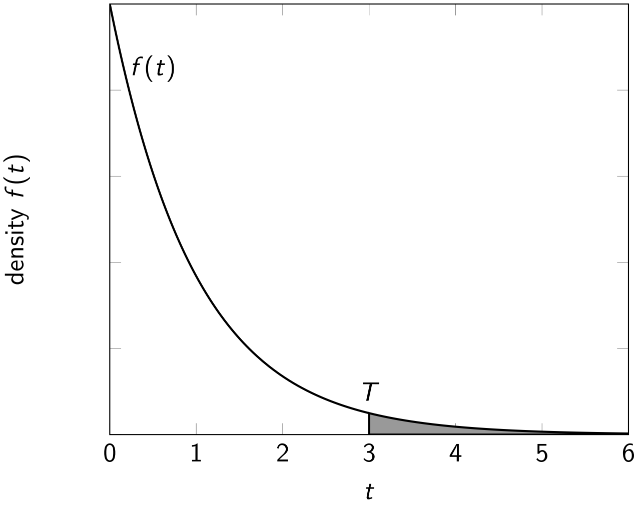
$$E[Y] = \int h(x) d\mathbb{P}(x) = \int h(x) \frac{d\mathbb{P}(x)}{d\tilde{\mathbb{P}}(x)} d\tilde{\mathbb{P}}(x) = \tilde{\mathbb{E}}\left[h(x)L(x)\right]$$

  - $L = d\mathbb{P}/d\tilde{\mathbb{P}}$ likelihood ratio,
  - $\tilde{\mathbb{E}}$ is the expectation associated to probability law $\mathbb{P}$.
- Required condition: $d\tilde{\mathbb{P}}(x) \neq 0$ when $h(x)d\mathbb{P}(x) \neq 0$.
- Unbiased estimator: $\displaystyle\frac{1}{n}\sum_{i=1}^{n} h(X_i)L(X_i)$ with $(X_i, 1 \leq i \leq n)$ i.i.d; copies of X, according to $\tilde{\mathbb{P}}$.
- Goal: select probability law $\tilde{\mathbb{P}}$ such that

$$\tilde{\sigma}^2[h(X)L(X)] = \tilde{\mathbb{E}}[(h(X)L(X))^2] - \mu^2 < \sigma^2[h(X)].$$

Example

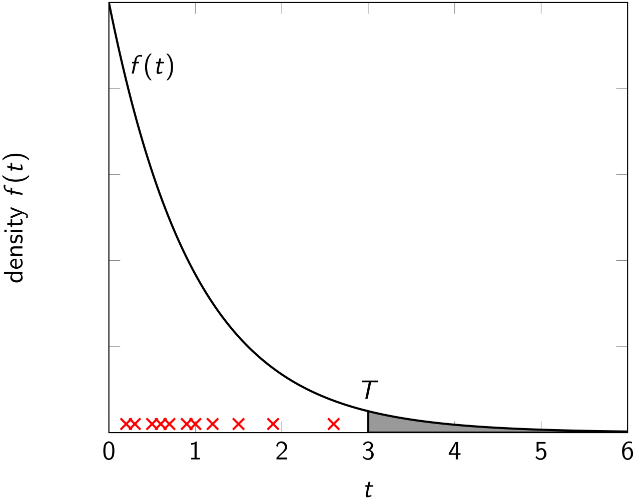
- We want to estimate the probability that a random variable exceeds T (area in grey under the density $f(t)$).



Reminder: the probability to be in an interval $[a, b]$ is the measure of the area

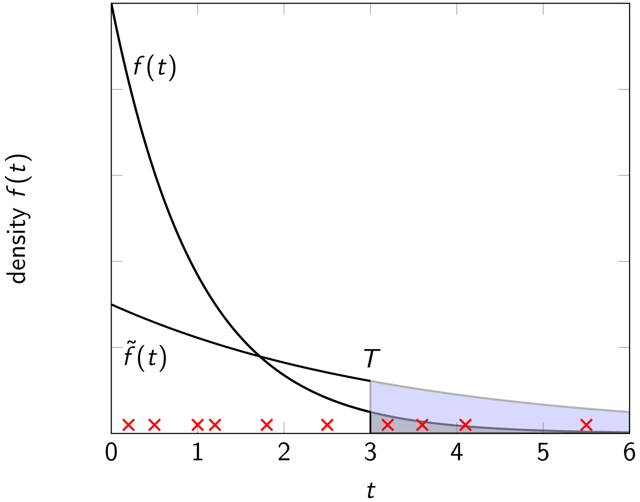
Rare event problem

- Draw values t_i (the red crosses X on the t-axis) according to density f
- Very few points (none) are $> T$.



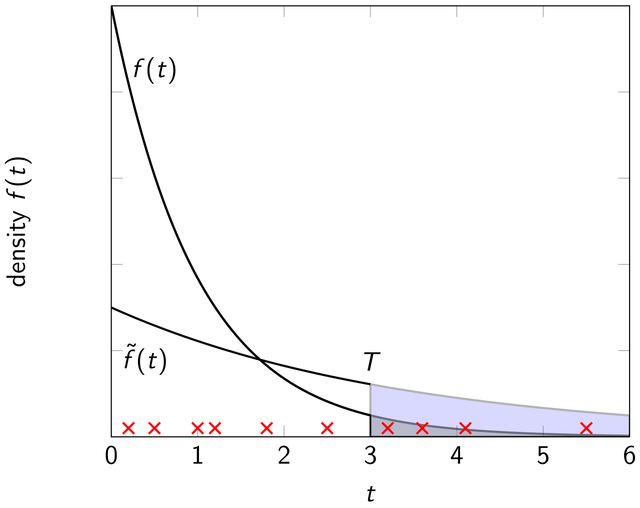
Importance sampling

- Sample according to another density $\tilde{f}$ increasing the probability to be $> T$.



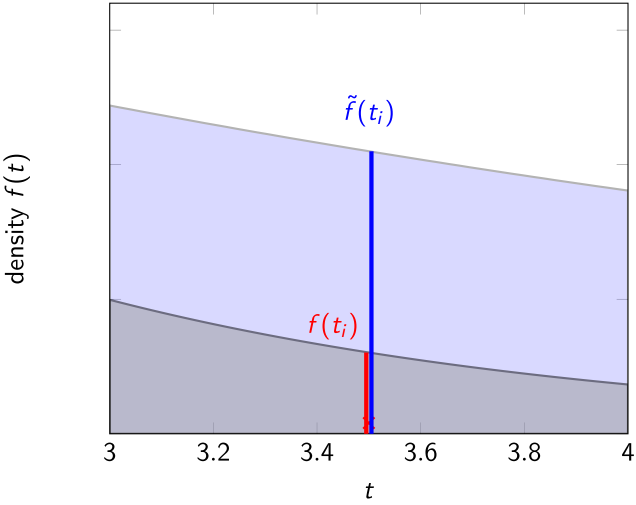
Importance sampling

- Sample according to another density $\tilde{f}$ increasing the probability to be $> T$.
- Rare set reached!



- Biased estimated probability then:
  - i.e., the proportion of points is the probability under the new density does not correspond to the grey area, but to the blue one.
- How to obtain a "valid" estimation?

- Instead of counting 1 each time we are $> T$ and look at the average value
- for each sample value t_i, we count $1(t_i > T)\frac{f(t_i)}{\tilde{f}(t_i)}$ (ratio of heights under densities at t_i) and look again at the average value
  $\Rightarrow$ unbiased estimation: the true probability is estimated.



IS for a discrete-time Markov chain (DTMC) $\{X_j, j \geq 0\}$

- $Y = h(X_0, \ldots, X_T)$ function of the sample path with
  - $P = (P(x,z))_{x,z\in\mathcal{S}}$ transition matrix, $\pi_0(x) = \mathbb{P}[X_0 = x]$, initial probabilities
  - up to a stopping time T
  - $\mu(x) = \mathbb{E}_x[Y]$.
- IS replaces the probabilities of paths $(x_0, \ldots, x_n)$,

$$\mathbb{P}[(X_0, \ldots, X_T) = (x_0, \ldots, x_n)] = \pi_0(x_0) \prod_{j=1}^{n-1} P(x_{j-1}, x_j),$$

  by $\tilde{\mathbb{P}}[(X_0, \ldots, X_T) = (x_0, \ldots, x_n)]$ st $\tilde{\mathbb{E}}[T] < \infty$.
- For convenience, the IS measure remains a DTMC, replacing $P(x,z)$ by $\tilde{P}(x,z)$ and $\pi_0(x)$ by $\tilde{\pi}_0(x)$.
- Then $L(X_0, \ldots, X_T) = \dfrac{\pi_0(X_0)}{\tilde{\pi}_0(X_0)} \prod_{j=1}^{T-1} \dfrac{P(X_{j-1}, X_j)}{\tilde{P}(X_{j-1}, X_j)}$.

Zero-variance IS estimator for Markov chains simulation

- Restrict to an additive (positive) cost

$$Y = \sum_{j=1}^{T} c(X_{j-1}, X_j)$$

  - For hitting proba: $c(x,z) = 1$ if $z \in \mathcal{A}$, 0 otherwise, $\mu(x) \equiv p(x)$
  - For hitting time: $c(x,z)$ avg time in x.
- Is there a Markov chain change of measure yielding zero-variance?
- We have zero variance with

$$\begin{aligned} \tilde{P}(x,z) &= \frac{P(x,z)(c(x,z)+\mu(z))}{\sum_w P(x,w)(c(x,w)+\mu(w))} \\ &= \frac{P(x,z)(c(x,z)+\mu(z))}{\mu(x)}. \end{aligned}$$

- Implementing it requires knowing $\mu(x)\ \forall x \in \mathcal{S}$, the quantities we wish to compute.

Zero-variance approximation

- Use a heuristic approximation $\hat{\mu}(\cdot)$ and plug it into the zero-variance change of measure instead of $\mu(\cdot)$

$$\tilde{P}(y,z) \quad = \quad \frac{P(y,z)(c(y,z)+\hat{\mu}(z))}{\sum_w P(y,w)(c(y,w)+\hat{\mu}(w))}$$

Outline

Highly Reliable Markovian Systems (HRMS)

- System with c types of components. $X = (X_1, \ldots, C_c)$ with C_i number of up components.
- $\mathbf{1}$: state with all components up.
- Failure rates are $O(\varepsilon)$, but not repair rates. Failure propagations possible.
- System down (in $\mathcal{A}$) when some combinations of components are down.
- Goal: compute $\mu(\mathbf{1}) \equiv p(\mathbf{1})$ with $p(y)$ probability to hit $\mathcal{A}$ before $\mathbf{1}$ starting from y (denominator of the ratio est. of MTTF)
- Simulation using the embedded DTMC. Failure probabilities are $O(\varepsilon)$ (except from $\mathbf{1}$). How to improve (accelerate) this?
- Existing method: $\forall y \neq \mathbf{1}$, increase the probability of the set of failures to constant $0.5 < q < 0.9$ and use individual probabilities proportional to the original ones (SFB), or uniformly (BFB).
- Failures not rare anymore. BRE property verified for BFB.

HRMS Example, and IS

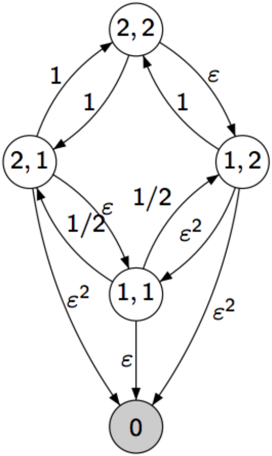


Figure: Original probabilities

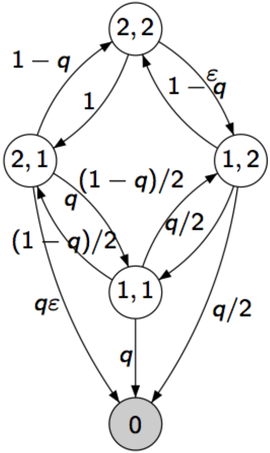


Figure: Probabilities under IS/BFB

HRMS, Zero-variance IS

L'Ecuyer & T., ANOR, 2011

- Recall the zero-variance approximation:

$$\tilde{P}(x,z) = \frac{P(x,z)(c(x,z)+\hat{p}(z))}{\sum_w P(y,w)(c(x,w)+\hat{p}(w))}$$

- The idea is to approach $p(y)$ by the probability $\hat{p}(y)$ of the path from y to $\mathcal{A}$ with the largest probability
- Intuition: as $\epsilon \to 0$, we get a good idea of the probability.

Proposition

Bounded Relative Error proved (as $\epsilon \to 0$) in general.
Even Vanishing Relative Error if $\hat{p}(y)$ contains all the paths with the smallest degree in ϵ.

- Other simple version: approach $p(y)$ by the (sum of) probability of paths from y with only failure components of a given type.
- Gain of several orders of magnitudes + stability of the results with respect to the literature.

HRMS: numerical illustrations

- Comparison of BFB and Zero-Variance Approximation (ZVA).
- $c = 3$ types of components, n_i of type i
- failure rates ε, 1.5ε, and $2\varepsilon^2$, repair rate 1
- System is down whenever fewer than two components of any one type are operational.

n_i	ε	μ_0	BFB est	ZVA est	BFB σ^2	ZVA σ^2
3	0.001	2.6×10^{-3}	2.7×10^{-3}	2.6×10^{-3}	6.2×10^{-5}	2.2×10^{-8}
6	0.01	1.8×10^{-7}	1.9×10^{-7}	1.8×10^{-7}	6.3×10^{-11}	2.0×10^{-14}
6	0.001	1.7×10^{-11}	1.8×10^{-11}	1.7×10^{-11}	8.8×10^{-19}	1.2×10^{-23}
12	0.1	6.0×10^{-8}	4.8×10^{-8}	6.0×10^{-8}	8.1×10^{-10}	1.6×10^{-10}
12	0.001	3.9×10^{-28}	(1.8×10^{-40})	3.9×10^{-28}	(3.2×10^{-74})	1.4×10^{-55}

Outline

Main ideas

Glynn, Nakayama & T., WSC, 2017

- Two potential estimators:
  - Direct estimator: repeat experiments up to failure of the system, and compute the average value
  - Literature, regenerative estimator: expresses the MTTF as a ratio of quantities over regenerative cycles

Main ideas

Glynn, Nakayama & T., WSC, 2017

- Two potential estimators:
  - Direct estimator: repeat experiments up to failure of the system, and compute the average value
  - Literature, regenerative estimator: expresses the MTTF as a ratio of quantities over regenerative cycles
- Question: Is there a reason why the regenerative estimator is used? Which one is "better"?

Main ideas

Glynn, Nakayama & T., WSC, 2017

- Two potential estimators:
  - Direct estimator: repeat experiments up to failure of the system, and compute the average value
  - Literature, regenerative estimator: expresses the MTTF as a ratio of quantities over regenerative cycles
- Question: Is there a reason why the regenerative estimator is used? Which one is "better"?
- Contributions
  - Crude (direct and regenerative) estimators are asymptotically similar in performance, in rare event settings
  - For Importance Sampling estimators, the regenerative one yield a efficient estimator when the crude can not.

Crude estimators of MTTF

- Notations for an S-valued regenerative process $X = (X(t) : t \geq 0)$
  - Compute $\alpha = \mathbb{E}[T]$, where $T = \inf\{t \geq 0 : X(t) \in \mathcal{A}\}$ is the hitting time of subset $\mathcal{A}$

Crude estimators of MTTF

- Notations for an S-valued regenerative process $X = (X(t) : t \geq 0)$
  - Compute $\alpha = \mathbb{E}[T]$, where $T = \inf\{t \geq 0 : X(t) \in \mathcal{A}\}$ is the hitting time of subset $\mathcal{A}$
  - Regeneration times $0 = \Gamma(0) < \Gamma(1) < \cdots$,
    with iid cycles $((\tau(k), (X(\Gamma(k-1)+s) : 0 \leq s < \tau(k)) : k \geq 1)$
  - $\tau(k) = \Gamma(k) - \Gamma(k-1)$, length of the kth regenerative cycle
  - $W(k) = \inf\{t \geq 0 : X(\Gamma(k-1)+t) \in \mathcal{A}\}$ first hitting to $\mathcal{A}$ after regeneration $\Gamma(k-1)$
  - $I(k) = \mathcal{I}(W(k) < \tau(k))$ with $\mathcal{I}(\cdot)$ is the indicator function

Crude estimators of MTTF

- Notations for an S-valued regenerative process $X = (X(t) : t \geq 0)$
  - Compute $\alpha = \mathbb{E}[T]$, where $T = \inf\{t \geq 0 : X(t) \in \mathcal{A}\}$ is the hitting time of subset $\mathcal{A}$
  - Regeneration times $0 = \Gamma(0) < \Gamma(1) < \cdots$,
    with iid cycles $((\tau(k), (X(\Gamma(k-1)+s) : 0 \leq s < \tau(k)) : k \geq 1)$
  - $\tau(k) = \Gamma(k) - \Gamma(k-1)$, length of the kth regenerative cycle
  - $W(k) = \inf\{t \geq 0 : X(\Gamma(k-1)+t) \in \mathcal{A}\}$ first hitting to $\mathcal{A}$ after regeneration $\Gamma(k-1)$
  - $I(k) = \mathcal{I}(W(k) < \tau(k))$ with $\mathcal{I}(\cdot)$ is the indicator function
- Ratio expression: $\alpha = \dfrac{\mathbb{E}[T \wedge \tau]}{\mathbb{P}(T < \tau)}$.

Definition

Direct estimator $\alpha_1(m) = \dfrac{1}{m} \sum_{j=1}^{m} T(j)$.

Ratio estimator $\alpha_2(n) = \dfrac{(1/n)\sum_{k=1}^{n}[W(k) \wedge \tau(k)]}{(1/n)\sum_{k=1}^{n} I(k)}$.

(Known) Central limit theorems

If $p = \mathbb{P}(T < \tau) > 0$:

Proposition (Direct estimator)

$$m^{1/2}[\alpha_1(m) - \alpha] \Rightarrow \sigma_1 \mathcal{N}(0,1)$$

as $m \to \infty$*, where*

$$\sigma_1^2 = \alpha^2 + \frac{\mathbb{E}[(T \wedge \tau)^2]}{p} - 2\alpha \frac{\mathbb{E}[T\mathcal{I}(T < \tau)]}{p}.$$

Proposition (Ratio-based estimator)

$$n^{1/2}[\alpha_2(n) - \alpha] \Rightarrow \sigma_2 \mathcal{N}(0,1)$$

as $n \to \infty$*, where*

$$\sigma_2^2 = \frac{\mathbb{E}[(T \wedge \tau)^2]}{p^2} - 2\alpha \frac{\mathbb{E}[T\mathcal{I}(T < \tau)]}{p^2} + \frac{\alpha^2}{p}.$$

Question: which estimator is "more efficient"?

- Estimators $\alpha_1(m)$ and $\alpha_2(n)$ are actually very similar
- If $N(j) = \inf\{k > N(j-1) : I(k) = 1\}$ index k of the cycle corresponding to the jth cycle in which A is hit

Proposition

For $m \geq 1$, we have $\alpha_2(N(m)) = \alpha_1(m)$.

- Is an estimator more efficient than the other?

Question: which estimator is "more efficient"?

- Estimators $\alpha_1(m)$ and $\alpha_2(n)$ are actually very similar
- If $N(j) = \inf\{k > N(j-1) : I(k) = 1\}$ index k of the cycle corresponding to the jth cycle in which A is hit

Proposition

For $m \geq 1$, we have $\alpha_2(N(m)) = \alpha_1(m)$.

- Is an estimator more efficient than the other?
- Two asymptotic settings
  - Decreasing reachable sets: sequence $(\mathcal{A}_b : b \geq 1)$ of subsets of S for which $p_b \equiv \mathbb{P}(T_b < \tau) \to 0$ as $b \to \infty$
  - Highly reliable systems: fixed $\mathcal{A}$ but transitions decomposed between failures and repairs with failures getting more and more rare (index ϵ) with respect to repairs

Asymptotic result with a decreasing sequence of reachable sets

- Let $\hat{\alpha}_{1,b}(c)$ and $\hat{\alpha}_{2,b}(c)$ be the estimators obtained after c units of computational time
- To hope for consistency and CLTs, we need a computational budget t_b for which $t_b p_b \to \infty$ as $b \to \infty$

Theorem (Both estimators asymptotically identical)

Assume $\mathbb{E}[\tau^3] < \infty$. If $t_b p_b \to \infty$ as $b \to \infty$, then we have that as $b \to \infty$,

$$\sqrt{t_b p_b}\left(\frac{\hat{\alpha}_{i,b}(t_b)}{\mathbb{E}[T_b]} - 1\right) \Rightarrow \sqrt{\mathbb{E}[\tau]}\mathcal{N}(0,1), \quad i = 1, 2, \quad \textit{and}$$

$$\sqrt{t_b p_b}\left(\frac{\hat{\alpha}_{1,b}(t_b)}{\mathbb{E}[T_b]} - \frac{\hat{\alpha}_{2,b}(t_b)}{\mathbb{E}[T_b]}\right) \Rightarrow 0.$$

Numerical results for HRMS

System with 3 component types, with $n_i = 3$, failure rates ε, repair rates 1, and system is down whenever fewer than two components of any one type are operational.

Direct:

m	ϵ	Confidence Interval	Variance	CPU	Work Norm. Var.
10^7	0.1	(8.764e+00 , 8.774e+00)	5.879e+01	17.7	1.0e-04
10^7	0.01	(5.838e+02 , 5.845e+02)	3.343e+05	134	4.5e+00
10^7	0.001	(5.581e+04 , 5.588e+04)	3.117e+09	1316.5	4.1e+05

Regenerative :

n	ϵ	Confidence Interval	Variance	CPU	Work Norm. Var.
10^7	0.1	(8.762e+00 , 8.782e+00)	2.484e+02	4.283	1.1e-04
10^7	0.01	(5.788e+02 , 5.837e+02)	1.586e+07	2.917	4.6e+00
10^7	0.001	(5.459e+04 , 5.611e+04)	1.510e+12	2.800	4.2e+05

Numerical results for HRMS

System with 3 component types, with $n_i = 3$, failure rates ε, repair rates 1, and system is down whenever fewer than two components of any one type are operational.
Direct:

m	ϵ	Confidence Interval	Variance	CPU	Work Norm. Var.
10^7	0.1	(8.764e+00 , 8.774e+00)	5.879e+01	17.7	1.0e-04
10^7	0.01	(5.838e+02 , 5.845e+02)	3.343e+05	134	4.5e+00
10^7	0.001	(5.581e+04 , 5.588e+04)	3.117e+09	1316.5	4.1e+05

Regenerative :

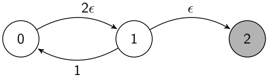
n	ϵ	Confidence Interval	Variance	CPU	Work Norm. Var.
10^7	0.1	(8.762e+00 , 8.782e+00)	2.484e+02	4.283	1.1e-04
10^7	0.01	(5.788e+02 , 5.837e+02)	1.586e+07	2.917	4.6e+00
10^7	0.001	(5.459e+04 , 5.611e+04)	1.510e+12	2.800	4.2e+05

- Similar asymptotic performance
- Direct estimator: bounded relative variance, but computational time issue
- Regenerative estimator: rather a rare event issue.

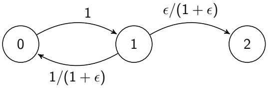
- Efficient Regenerative IS estimators extensively studied.
- Question:

  What about the direct estimator?
  Can its combination with IS yield an efficient estimator?

- We will play with the toy example:



with embedded DTMC



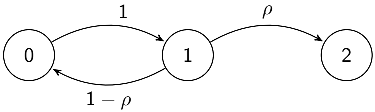
$$\mathbb{E}_\epsilon(T_\epsilon) = \sum_{n=0}^{\infty}(n+1)\left(\frac{1}{2\epsilon}+\frac{1}{1+\epsilon}\right)\left(\frac{1}{1+\epsilon}\right)^n\frac{\epsilon}{1+\epsilon} = \frac{1+3\epsilon}{2\epsilon^2}$$

$$\mathbb{E}_\epsilon[(T_\epsilon)^2] = \sum_{n=0}^{\infty}(n+1)^2\left(\frac{1}{2\epsilon}+\frac{1}{1+\epsilon}\right)^2\left(\frac{1}{1+\epsilon}\right)^n\frac{\epsilon}{1+\epsilon} = \frac{(2+\epsilon)(1+3\epsilon)^2}{4(1+\epsilon)\epsilon^4}$$

$$\mathbb{E}_\epsilon(N) = \sum_{n=0}^{\infty}(2+2n)\left(\frac{1}{1+\epsilon}\right)^n\frac{\epsilon}{1+\epsilon} = \frac{2(1+\epsilon)}{\epsilon}$$ with N:# transitions in a run.

Failure biasing

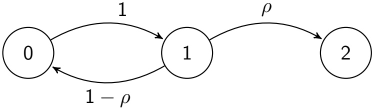
- Change the probability of making a failure transition to be ρ, independent of ϵ



- $\tilde{\mathbb{E}}_\epsilon[(T_\epsilon L)^2] = \mathbb{E}_\epsilon[(T_\epsilon)^2 L] = \sum_{n=0}^{\infty} (n+1)^2 \left(\frac{1}{2\epsilon} + \frac{1}{1+\epsilon}\right)^2 \frac{\left(\left(\frac{1}{1+\epsilon}\right)^n \frac{\epsilon}{1+\epsilon}\right)^2}{(1-\rho)^n \rho}$

Failure biasing

- Change the probability of making a failure transition to be ρ, independent of ϵ

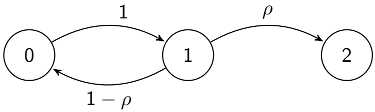


- $\tilde{\mathbb{E}}_\epsilon[(T_\epsilon L)^2] = \mathbb{E}_\epsilon[(T_\epsilon)^2 L] = \sum_{n=0}^{\infty}(n+1)^2 \left(\frac{1}{2\epsilon} + \frac{1}{1+\epsilon}\right)^2 \frac{\left(\left(\frac{1}{1+\epsilon}\right)^n \frac{\epsilon}{1+\epsilon}\right)^2}{(1-\rho)^n \rho}$
- Converging sum iff $1/((1+\epsilon)^2(1-\rho)) < 1$, i.e., ρ small enough

$$\rho < 1 - \frac{1}{(1+\epsilon)^2} = 2\epsilon - 3\epsilon^2 + o(\epsilon^2).$$

Failure biasing

- Change the probability of making a failure transition to be ρ, independent of ϵ



- $\tilde{\mathbb{E}}_\epsilon[(T_\epsilon L)^2] = \mathbb{E}_\epsilon[(T_\epsilon)^2 L] = \sum_{n=0}^{\infty}(n+1)^2\left(\frac{1}{2\epsilon}+\frac{1}{1+\epsilon}\right)^2 \frac{\left(\left(\frac{1}{1+\epsilon}\right)^n \frac{\epsilon}{1+\epsilon}\right)^2}{(1-\rho)^n\rho}$
- Converging sum iff $1/((1+\epsilon)^2(1-\rho)) < 1$, i.e., ρ small enough

$$\rho < 1 - \frac{1}{(1+\epsilon)^2} = 2\epsilon - 3\epsilon^2 + o(\epsilon^2).$$

- But $\tilde{\mathbb{E}}_\epsilon(N) = \sum_{n=0}^{\infty}(2+2n)(1-\rho)^n\rho = \frac{2}{\rho}$.
  The average simulation time for a single run will increase to infinity as $\epsilon \to 0$!

Zero-variance approximation

- For a CTMC with transition matrix $(P_{x,y})_{x,y\in S}$, if $\mathbb{E}_{\epsilon,x}$ expectation starting from x,

$$\tilde{P}_{x,y} = P_{x,y}\frac{1/\lambda(x) + \mathbb{E}_{\epsilon,y}(T_\epsilon)}{\mathbb{E}_{\epsilon,x}(T_\epsilon)}$$

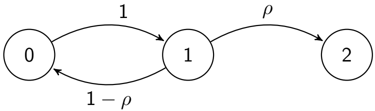
  yields an estimator with variance zero.

Zero-variance approximation

- For a CTMC with transition matrix $(P_{x,y})_{x,y\in S}$, if $\mathbb{E}_{\epsilon,x}$ expectation starting from x,

$$\tilde{P}_{x,y} = P_{x,y}\frac{1/\lambda(x) + \mathbb{E}_{\epsilon,y}(T_\epsilon)}{\mathbb{E}_{\epsilon,x}(T_\epsilon)}$$

  yields an estimator with variance zero.
- On our toy example, the only probability we can change is from 1



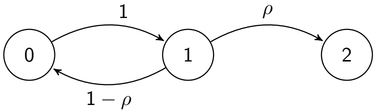
- $\rho = \frac{\epsilon}{1+\epsilon}\frac{\frac{1}{1+\epsilon}+0}{\frac{1+2\epsilon}{2\epsilon^2}} = \frac{2\epsilon^3}{(1+\epsilon)^2(1+2\epsilon)}$ yields variance 0.

Zero-variance approximation

- For a CTMC with transition matrix $(P_{x,y})_{x,y\in S}$, if $\mathbb{E}_{\epsilon,x}$ expectation starting from x,

$$\tilde{P}_{x,y} = P_{x,y}\frac{1/\lambda(x) + \mathbb{E}_{\epsilon,y}(T_\epsilon)}{\mathbb{E}_{\epsilon,x}(T_\epsilon)}$$

  yields an estimator with variance zero.
- On our toy example, the only probability we can change is from 1



- $\rho = \frac{\epsilon}{1+\epsilon}\frac{\frac{1}{1+\epsilon}+0}{\frac{1+2\epsilon}{2\epsilon^2}} = \frac{2\epsilon^3}{(1+\epsilon)^2(1+2\epsilon)}$ yields variance 0.
- But the estimation takes on average longer time, $\frac{2}{\rho} = \Theta(\epsilon^{-3})$, as ϵ gets closer to zero.
- An approximation of the zero-variance IS can be inefficient, producing an unbounded work-normalized relative variance.

Discussion on the impact of the approximation

- For $\rho = \frac{2\epsilon^3}{(1+\epsilon)^2(1+2\epsilon)}$, we retrieve a variance zero.

Discussion on the impact of the approximation

- For $\rho = \frac{2\epsilon^3}{(1+\epsilon)^2(1+2\epsilon)}$, we retrieve a variance zero.
- For $\rho = \epsilon^3$ (approximation of good asymptotic order), the variance is $\Theta(\epsilon^{-2})$, but the work-normalized relative variance is unbounded due to the computational time.

Discussion on the impact of the approximation

- For $\rho = \frac{2\epsilon^3}{(1+\epsilon)^2(1+2\epsilon)}$, we retrieve a variance zero.
- For $\rho = \epsilon^3$ (approximation of good asymptotic order), the variance is $\Theta(\epsilon^{-2})$, but the work-normalized relative variance is unbounded due to the computational time.
- For $\rho = 2\epsilon^3$ (exact first-order term), the variance is $\Theta(1)$, which is better but still not sufficient to yield a bounded work-normalized variance.

Discussion on the impact of the approximation

- For $\rho = \frac{2\epsilon^3}{(1+\epsilon)^2(1+2\epsilon)}$, we retrieve a variance zero.
- For $\rho = \epsilon^3$ (approximation of good asymptotic order), the variance is $\Theta(\epsilon^{-2})$, but the work-normalized relative variance is unbounded due to the computational time.
- For $\rho = 2\epsilon^3$ (exact first-order term), the variance is $\Theta(1)$, which is better but still not sufficient to yield a bounded work-normalized variance.

Much better than an exact first-order approximation is required. Hard to obtain in practice.

Conclusions on MTTF estimation

We have compared two standard estimators of the MTTF for regenerative processes

- a direct one expressed as the average of simulated times to failure
- one making use of the regenerative structure

1. Crude direct and ratio-based estimators are asymptotically equivalent (in two asymptotic contexts)
2. When IS is used, the regenerative expression is rather advised.

Outline

Basic idea

- Let F be the cumulative distribution function of T
- Goal: For fixed $0 < q < 1$, estimate the q-quantile $(0 < q < 1)$

$$\xi = F^{-1}(q) \equiv \inf\{t : F(t) \geq q\}$$

and the *conditional tail expectation* (CTE)

$$\gamma = E[T \mid T > \xi].$$

- Assumption: X is (classically) regenerative
with $0 = \Gamma_0 < \Gamma_1 < \Gamma_2 < \cdots$ sequence of regeneration times

Decomposition

- Using $\tau_i = \Gamma_i - \Gamma_{i-1}$ and M the number of first cycles not reaching $\mathcal{A}$

$$T = \sum_{i=1}^{M} \tau_i + T_{M+1}$$

with $T_i = \inf\{t \geq 0 : X(\Gamma_{i-1} + t) \in \mathcal{A}\}$ time to the next hit to $\mathcal{A}$ after Γ_{i-1}.

- M geometric r.v. with $P(M = k) = p(1-p)^k$ where

$$p = P(T < \tau).$$

- Recall that the regenerative structure of X allows to express

$$\alpha = E[T] = \frac{E[T \wedge \tau]}{p} \equiv \frac{\zeta}{p}.$$

Asymptotic regimes/exponential approximation

- Introduction of a rarity parameter ϵ
- Assumption: $p \equiv p_\epsilon \to 0$ as $\epsilon \to 0$.
    - Ex HRMS: Probability of reaching a failed state before coming back to the initial (perfectly working) state goes to 0 with failure rates
    - Ex GI/G/1 queue: considering a receding set of states (number of customers) $\mathcal{A} \equiv \mathcal{A}_\epsilon = \{b_\epsilon, b_\epsilon + 1, b_\epsilon + 2, \ldots\}$.

Theorem (Known result)

The scaled hitting time $T_\epsilon/\alpha_\epsilon$ converges weakly to an exponential: for each $x \geq 0$,

$$P_\epsilon(T_\epsilon/\alpha_\epsilon \leq x) \to 1 - e^{-x} \text{ as } \epsilon \to 0.$$

Quantile and CTE estimators based on the exponential approximation

From

$$F(t) = P(T \leq t) = P(T/\alpha \leq t/\alpha) \approx 1 - e^{-t/\alpha} \equiv \widetilde{F}_{\mathrm{exp}}(t),$$

we get

- $\widetilde{\xi}_{\mathrm{exp}} = \widetilde{F}_{\mathrm{exp}}^{-1}(q) = -\alpha \ln(1-q)$
- $\widetilde{\gamma}_{\mathrm{exp}} = \widetilde{\xi}_{\mathrm{exp}} + \alpha = \alpha[1 - \ln(1-q)]$.

Using the ZVA efficient estimator $\widehat{\alpha}$ of α, we get

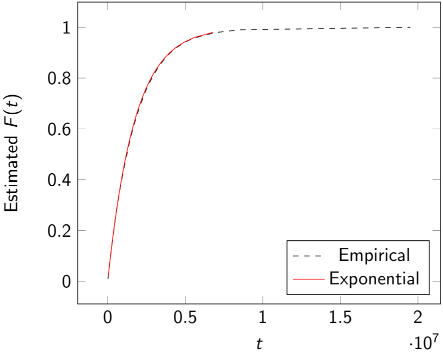
$\widehat{\xi}_{\mathrm{exp}} = \widehat{F}_{\mathrm{exp}}^{-1}(q) = -\widehat{\alpha} \ln(1-q)$ and $\widehat{\gamma}_{\mathrm{exp}} = \widehat{\xi}_{\mathrm{exp}} + \widehat{\alpha} = \widehat{\alpha}[1 - \ln(1-q)]$

- Efficient estimators
- ...but biased
- Other more involved estimators available in our WSC'2018 paper.

Numerical example

- HRMS with three component types
- five components of each type
- 15 repairmen
- system up whenever at least two components of each type work
- Each component has failure rate ϵ and repair rate 1.

With $\epsilon = 10^{-2}$



Numerical results

- Quantile estimators

ϵ	q	Empirical 95% CI	CPU	Expon. Est.	Expon. 95% CI	CPU
0.01	0.1	(1.701e+05, 1.971e+05)	890 sec	1.830e+05	(1.764e+05, 1.896e+05)	0.3 sec
0.01	0.5	(1.206e+06, 1.271e+06)	890 sec	1.204e+06	(1.161e+06, 1.247e+06)	0.3 sec
0.01	0.9	(3.958e+06, 4.135e+06)	890 sec	4.000e+06	(3.856e+06, 4.143e+06)	0.3 sec
10^{-4}	0.1	N/A	N/A	1.757e+13	(1.756e+13, 1.758e+13)	0.3 sec
10^{-4}	0.5	N/A	N/A	1.155e+14	(1.154e+14, 1.157e+14)	0.3 sec
10^{-4}	0.9	N/A	N/A	3.840e+14	(3.838e+14, 3.842e+14)	0.3 sec

- CTE estimators

ϵ	q	Empir. Est.	CPU	Expon. Est.	Expon. 95% CI	CPU
0.01	0.1	1.964e+06	890 sec	1.920e+06	(1.851e+06, 1.989e+06)	0.3 sec
0.01	0.5	3.011e+06	890 sec	2.941e+06	(2.836e+06, 3.046e+06)	0.3 sec
0.01	0.9	5.915e+06	890 sec	5.737e+06	(5.531e+06, 5.942e+06)	0.3 sec
10^{-4}	0.1	N/A	N/A	1.839e+14	(1.834e+14, 1.845e+14)	0.3 sec
10^{-4}	0.5	N/A	N/A	2.817e+14	(2.809e+14, 2.826e+14)	0.3 sec
10^{-4}	0.9	N/A	N/A	5.495e+14	(5.479e+14, 5.512e+14)	0.3 sec

  - Very efficient
  - But biased.... for small ϵ, does not *seem* a problem in practice
  - Other less biased estimators studied in our WSC'2018 paper.

References

- Mainly based on
  - P. L'Ecuyer and B. Tuffin. Approximating Zero-Variance Importance Sampling in a Reliability Setting. *Annals of Operations Research*. Vol.189, pp 277-297, Sept.2011
  - P.W. Glynn, M.K. Nakayama, and B. Tuffin. On the estimation of the mean time to failure by simulation. In the *Proceedings of the 2017 Winter Simulation Conference*, Las Vegas, NV, USA, Dec. 2017
  - P.W. Glynn, M.K. Nakayama, B.Tuffin. Using Simulation to Calibrate Exponential Approximations to Tail-Distribution Measures of Hitting Times to Rarely Visited Sets. In the *Proceedings of the 2018 Winter Simulation Conference*, Gothenburg, Sweden, Dec. 2018
- Other selected references on rare events
  - G. Rubino and B. Tuffin (eds). *Rare Event Simulation using Monte Carlo Methods*. John Wiley, 2009
  - P. L'Ecuyer, J. Blanchet, B. Tuffin, P.W. Glynn. Asymptotic Robustness of Estimators in Rare-Event Simulation. *ACM Transactions on Modeling and Computer Simulation*. Vol 20, Num. 1 Article 6, 2010
  - P. L'Ecuyer, V. Demers and B. Tuffin. Rare Events, Splitting, and Quasi-Monte Carlo. *ACM Transactions on Modeling and Computer Simulation*, Vol. 17, Num. 2, Article 9, 2007
  - P. L'Ecuyer and B. Tuffin, Approximate Zero-Variance Simulation. In *Proceedings of the 2008 Winter Simulation Conference*, 2008