Introduction to rare event simulation

Bruno Tuffin

(some works with colleagues H. Cancela, V. Demers, P. L'Ecuyer, G. Rubino)

INRIA - Centre Bretagne Atlantique, Rennes

Aussois, June 2008, AEP9

Outline

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 which occurence is rare, of probability less than 10^{-3}.
- Typical probabilities of interest are between 10^{-8} and 10^{-10}.
- This is a target for instance in nuclear plants (!)
- In most of the above problems, the mathematical model is too complicated to be solved by analytic-numeric methods because
  - the assumptions are not stringent enough,
  - the mathematical dimension of the problem is large,
  - or the state space is too large to get a result in a reasonable time.
- Then, simulation is often the only tool at hand.

Monte Carlo

- In all the above problems, the goal is to compute $\mu = \mathbb{E}[X]$ of some random variable X.
- Monte Carlo simulation (in its basic form) generates n independent copies of X, $(X_i,\ 1 \leq i \leq n)$,
- $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ approximation of μ.
- Almost sure convergence as $n \to \infty$ (law of large numbers).
- Accuracy: central limit theorem, yielding a confidence interval

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

  - α: desired confidence probability,
  - $c_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})$ with Φ is the cumulative Normal distribution function of $\mathcal{N}(0,1)$
  - $\sigma^2 = \mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, estimated by $S_n^2 = (1/(n-1))\sum_{i=1}^{n} X_i^2 - (n/(n-1))(\bar{X}_n)^2$.

Remarks on the confidence interval

- Confidence interval size: $2c_\alpha\sigma/\sqrt{n}$,
- decreasing in $1/\sqrt{n}$ independently of the mathematical dimension of the problem (advantage for large dimensions).
- Slow in the other hand: to reduce the width by 2, you need 4 times more replications.
- How to improve the accuracy? *Acceleration*
  - either by decreasing the simulation time to get a replication
  - or reducing the variance of the estimator.
- For rare events, acceleration required! (See next slide.)

Inefficiency of crude Monte Carlo

- *Crude* Monte Carlo: simulates the model directly
- Assume we want to compute the probability $\mu = \mathbb{E}[1_A] << 1$ of a rare event A.
- X_i Bernoulli r.v.: 1 if the event is hit and 0 otherwise.
- To get a single occurence, we need in average $1/\mu$ replications (10^9 for $\mu = 10^{-9}$), and more to get a confidence interval.
- $n\bar{X}_n$ Binomial with parameters (n, μ) and the confidence interval is

$$\left(\bar{X}_n - \frac{c_\alpha \sqrt{\mu(1-\mu)}}{\sqrt{n}}, \bar{X}_n + \frac{c_\alpha \sqrt{\mu(1-\mu)}}{\sqrt{n}} \right).$$

- *Relative half width* $c_\alpha \sigma/(\sqrt{n}\mu) = c_\alpha \sqrt{(1-\mu)/\mu/n} \to \infty$ as $\mu \to 0$.
- Something has to be done to accelerate the occurence (and reduce variance).

Robustness properties

- In rare-event simulation models, we often parameterize with a rarity parameter $\epsilon > 0$ such that $\mu = \mathbb{E}[X(\epsilon)] \to 0$ as $\epsilon \to 0$.
- An estimator $X(\epsilon)$ is said to have *bounded relative variance* (or *bounded relative error*) if $\sigma^2(X(\epsilon))/\mu^2(\epsilon)$ 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 $\lim_{\epsilon \to 0} \ln(\mathbb{E}[X^2(\epsilon)])/\ln(\mu(\epsilon)) = 2$.
- Other robustness measures exist (based on higher degree moments, on the Normal approximation, on simulation time...)

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

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

  - $L = d\mathbb{P}/d\tilde{\mathbb{P}}$ likelihood ratio,
  - $\tilde{\mathbb{E}}$ is the expectation associated to probability law $\mathbb{P}$.
- Required condition: $d\tilde{\mathbb{P}}(y) \neq 0$ when $h(y)d\mathbb{P}(y) \neq 0$.
- If $\mathbb{P}$ and $\tilde{\mathbb{P}}$ continuous laws, L ratio of density functions.
- If $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are discrete laws, L ratio of indiv. prob.
- Unbiased estimator: $\frac{1}{n}\sum_{i=1}^{n} h(Y_i)L(Y_i)$ with

  $(Y_i,\ 1 \leq i \leq n)$ i.i.d; copies of Y, according to $\tilde{\mathbb{P}}$.
- Goal: select probability law $\tilde{\mathbb{P}}$ such that

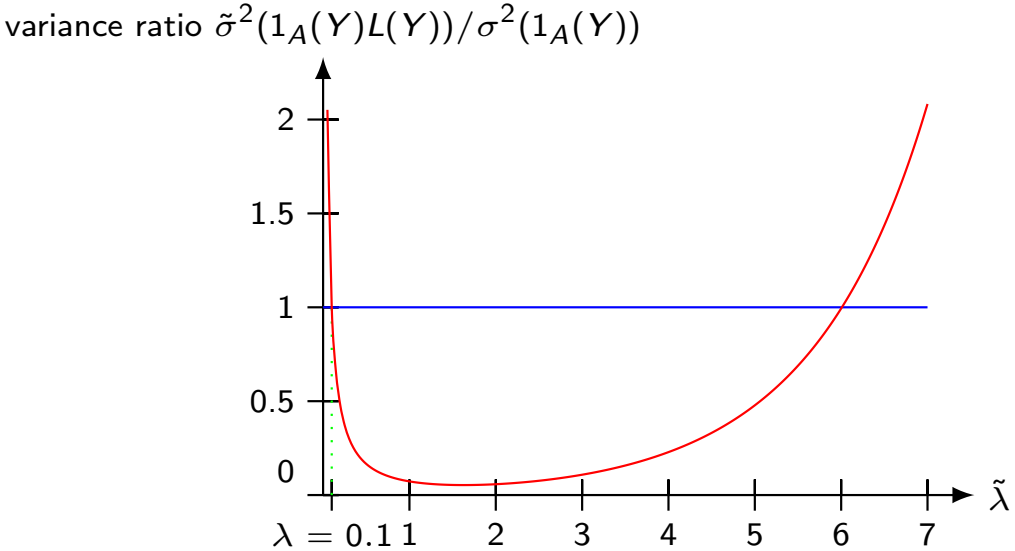
$$\tilde{\sigma}^2[h(Y)L(Y)] = \tilde{\mathbb{E}}[(h(Y)L(Y))^2] - \mu^2 < \sigma^2[h(Y)].$$

IS difficulty: system with exponential failure time

- Goal: to compute μ that the system fails before T, $\mu = \mathbb{E}[1_A(Y)] = 1 - e^{-\lambda T}$.
- Use for IS an exponential density with a different rate $\tilde{\lambda}$

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

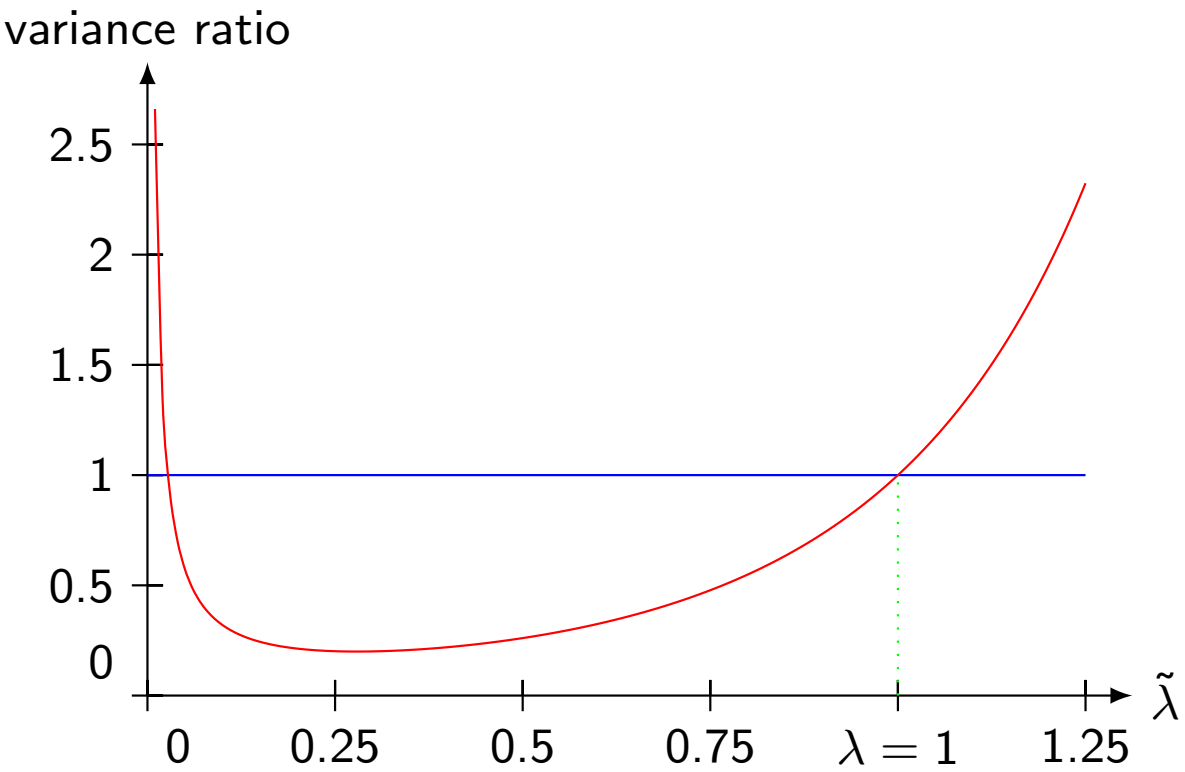
- Variance ratio for $T = 1$ and $\lambda = 0.1$:



- 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 $\lambda = 1$:



Optimal estimator for estimating $\mathbb{E}[h(Y)] = \int h(y)L(y)d\tilde{\mathbb{P}}(y)$

- Optimal change of measure:

$$\tilde{\mathbb{P}} = \frac{|h(Y)|}{\mathbb{E}[|h(Y)|]} d\mathbb{P}.$$

- *Proof:* for any alternative IS measure $\mathbb{P}'$, 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 τ, first time it hits a set Δ.
  - $\mu(y) = \mathbb{E}_y[X]$.
- 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 $\tilde{\mathbb{E}}[\tau] < \infty$.
- 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 $L(Y_0, \ldots, Y_\tau) = \dfrac{\pi_0(Y_0)}{\tilde{\pi}_0(Y_0)} \displaystyle\prod_{j=1}^{\tau-1} \frac{P(Y_{j-1}, Y_j)}{\tilde{P}(Y_{j-1}, Y_j)}$.

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_ys 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 λ and service rate $\mu > \lambda$ fits the framework with $p = \lambda/(\lambda + \mu)$.
- How to apply IS: increase the p_ys to $\tilde{p}_y$ to accelerate the occurence (but not too much again).
- Large deviation theory applies here, when B increases.
  - Strategy for an M/M/1 queue: exchange λ and μ
  - Asymptotic optimality, but no bounded relative error.

Highly Reliable Markovian Systems (HRMS)

- System with c types of components. $Y = (Y_1, \ldots, Y_c)$ with Y_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 Δ) when some combinations of components are down.
- Goal: compute $\mu(y)$ probability to hit Δ before $\mathbf{1}$.
- $\mu(\mathbf{1})$ typical measure in dependability analysis, small if ε small.
- Simulation using the embedded DTMC. Failure probabilities are $O(\varepsilon)$ (except from $\mathbf{1}$). How to improve (accelerate) this?
- Proposition: $\forall y \neq \mathbf{1}$, increase the probability of the set of failures to constant $0.5 < q < 0.9$ and use individual probabilities proportional to the original ones.
- Failures not rare anymore.

HRMS Example, and IS

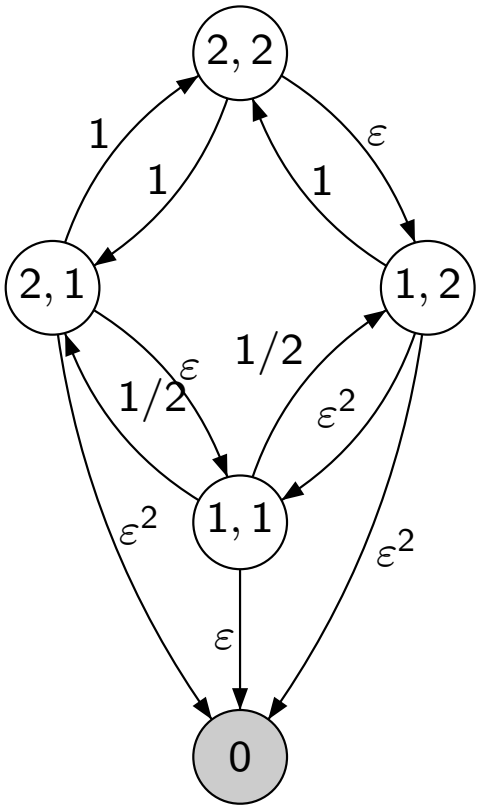


Figure: Original probabilities

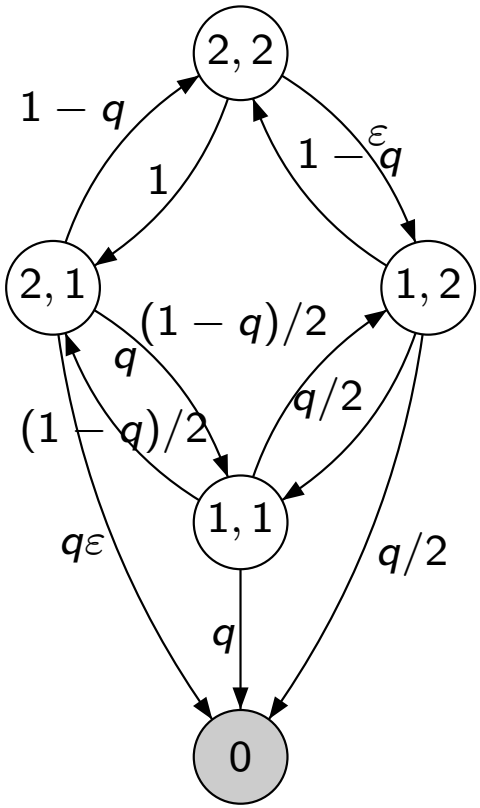


Figure: Probabilities under IS

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

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

- 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 τ, using the fact that $\mu(Y_\tau) = 0$ In that case, if $\tilde{X}$ denotes the IS estimator,

$$\begin{aligned}
\tilde{X} &= \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{P(Y_{j-1}, Y_j)}{\tilde{P}(Y_{j-1}, Y_j)} \\
&= \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{P(Y_{j-1}, Y_j)\mu(Y_{j-1})}{P(Y_{j-1}, Y_j)(c(Y_{j-1}, Y_j) + \mu(Y_j))} \\
&= \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\mu(Y_{j-1})}{c(Y_{j-1}, Y_j) + \mu(Y_j)} \\
&= \mu(Y_0)
\end{aligned}$$

- *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

$$\begin{aligned} \mu^{(j+1)}(y_j) &= (1 - a_j(y_j))\mu^{(j)}(y_j) \\ &+ a_j(y_j)\left[c(y_j, y_{j+1}) + \mu^{(j)}(y_{j+1})\right] \frac{P(y_j, y_{j+1})}{\tilde{P}^{(j)}(y_j, y_{j+1})}, \end{aligned}$$

    where $\{a_j(y),\ j \geq 0\}$, sequence of *step sizes*
  - For $\delta > 0$ constant,

$$\tilde{P}^{(j+1)}(y_j, y_{j+1}) = \max\left(P(y_j, y_{j+1})\frac{\left[c(y_j, y_{j+1}) + \mu^{(j+1)}(y_{j+1})\right]}{\mu^{(j+1)}(y_j)},\ \delta\right).$$

  - Otherwise $\mu^{(j+1)}(y) = \mu^{(j)}(y)$, $\tilde{P}^{(j+1)}(y, z) = P^{(j)}(y, z)$.
  - Normalize: $P^{(j+1)}(y_j, y) = \dfrac{\tilde{P}^{(j+1)}(y_j, y)}{\sum_z \tilde{P}^{(j+1)}(y_j, z)}$.
- 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.
- Our research direction:
  - 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 α_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 $p \longrightarrow 0$.

Illustration: HRMS

- Complicates the previous model due to the multidimensional description of a state.
- The idea is to approach $\mu(y)$ by the probability of the path from y to Δ with the largest probability
- Results (to be published):
  - Bounded Relative Error proved (as $\epsilon \longrightarrow 0$).
  - Even vanishing relative error if $\mu(y)$ contains all the paths with the smallest degree in ϵ.
- Simple version: approach $\mu(y)$ by the (sum of) probability of paths from y with only failure components of a given type.
- Results impressive with respect to the IS scheme of just increasing the probability of whole set failure transitions to q as proposed in the literature (gain of several orders of magnitudes + stability of the results).

Splitting: general principle

- Splitting is the other main rare event simulation technique.
- Assume we want to compute the probability $\mathbb{P}(D)$ of an event D.
- General idea:
  - Decompose
$$D_1 \supset \cdots \supset D_m = D,$$
  - Use $\mathbb{P}(D) = \mathbb{P}(D_1)\mathbb{P}(D_2 \mid D_1) \cdots \mathbb{P}(D_m \mid D_{m-1})$, 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.

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 $p_k = \mathbb{P}(D_k | D_{k-1})$

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 hat go a given number β 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_β. 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?

Optimal values

- 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

$$\begin{aligned} \mathrm{Var}[\hat{p}_1 \cdots \hat{p}_m] &= \prod_{k=1}^{m} \mathbb{E}[\hat{p}_k^2] - \gamma_0^2 = \left(p^2 + \frac{p(1-p)}{n}\right)^m - p^{2m} \\ &= \frac{mp^{2m-1}(1-p)}{n} + \cdots + \frac{(p(1-p))^m}{n^m}. \end{aligned}$$

- Assuming $n \gg (m-1)(1-p)/p$, $\mathrm{Var}[\hat{p}_1 \cdots \hat{p}_m] \approx mp^{2m-1}(1-p)/n \approx m\gamma_0^{2-1/m}/n$.
- The work normalized variance $\approx [\gamma_0^n m^2]/n = \gamma_0^{2-1/m} m^2$
- 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

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

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

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

- 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 jth 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:

$$\begin{aligned} h_1(x_1, x_2) &= x_2, \\ h_2(x_1, x_2) &= (x_2 + \min(0, x_2 + x_1 - L))/2, \\ h_3(x_1, x_2) &= x_2 + \min(x_1, L - x_2 - 1) \times (1 - x_2/L). \end{aligned}$$

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 $\tilde{W}_N = 10^{15} \times \hat{W}_N$.

	$N = 2^{10}$		$N = 2^{12}$		$N = 2^{14}$		$N = 2^{16}$	
	$\tilde{V}_N$	$\tilde{W}_N$	$\tilde{V}_N$	$\tilde{W}_N$	$\tilde{V}_N$	$\tilde{W}_N$	$\tilde{V}_N$	$\tilde{W}_N$
h_2, Splitting	109	120	89	98	124	137	113	125
h_2, Rus. Roul.	178	67	99	37	119	45	123	47
h_3, Splitting	93	103	110	121	93	102	107	118
h_3, Rus. Roul.	90	34	93	35	94	36	109	41

Conclusions

- Two main techniques for rare event simulation: importance sampling and splitting
- Splitting fans usually say that it has the advantage of not having to change the model's laws.
- But, requires the definition of the importance function, very similar to defining the IS change of measure.
- On the other hand, any rare event *has* to be decomposed in non-rare ones, which cannot always be done.
- Recent moves:
  - defining zero-variance approximation, yielding bounded relative error.
  - *Cross Entropy* technique: finds the optimal change of measure in a parametric family.
- Book on Rare event simulation to be published by John Wiley & Sons, by the end of the year.