SUMMER SCHOOL IN STATISTICS 2014
VIETNAM

Introduction to particle methods and Importance Sampling/Splitting methods

Part I

Agnès Lagnoux
lagnoux@univ-tlse2.fr

Contents

1 Introduction

The analysis of rare events is of great importance in many fields because of the risk associated to the event. Their probabilities are often about 10^{-9} to 10^{-12}. One can use many ways to study them: the first one is statistical analysis, based on the standard extreme value distributions but this needs a long observation period (see Aldous [1]). The second one lies on modeling and leads to estimate the rare event probability either by analytical approach (see Sadowsky [24]) or by simulation.
In this course, we focus on the simulation approach based on Monte Carlo method. Nevertheless, crude simulation is impracticable for estimating such small probabilities: to estimate probabilities of order 10^{-10} with acceptable confidence would require the simulation of at least 10^{12} events (which corresponds to the occurrence of only one hundred rare events).
To overcome these limits, fast simulation techniques are applied. In particular, importance sampling (IS) is a refinement of Monte-Carlo methods (see e.g. [20] or [21]). The main idea of IS is to make the occurrence of the rare event more frequent. More precisely IS consists in selecting a change of measure that minimizes the variance of the estimator. Another method is called splitting. The basic idea of splitting (ISp) is to partition the space-state of the system into a series of nested subsets and to consider the rare event as the intersection of a nested sequence of events (see [16]). When a given subset is entered by a sample trajectory, random retrials are generated from the initial state corresponding to the state of the system at the entry point. More refined versions of splitting as particles systems [9] or RESTART [27] have been introduced in the last decades. Unlike IS, ISp does not change the underlying dynamics of the particles.
The rest of the chapter is dedicated to basics in Monte Carlo method whereas the following one deals with IS and other methods to reduce the variance. Chapter 3 presents branching processes that will appear in Chapter 4 that considers ISp. Finally, all these techniques are illustrated on a practical in the last chapter.

1.1 Description of the Monte Carlo method

The rest of the section is largely inspired from [20] and [21]).

To use the Monte Carlo method, we must rewrite the quantity to estimate in terms of the expected value of a random variable (rv) say X. It then remains to simulate a sequence of independent and identically distributed (iid) rvs $(X_i,\ i \geqslant 1)$ distributed as X.

More precisely we want to calculate

$$I = \int_{[0,1]^d} g(u_1, \ldots, u_d) du_1 \ldots du_d.$$

Remark that if we set $X = g(U_1, \ldots, U_d)$ where $U_1, \ldots, U_d$ are uniform iid rv on $[0,1]$, we get $I = \mathbb{E}(X) = \mathbb{E}(g(U_1, \ldots, U_d))$.

For the simulation, assume that $(U_i,\ i \geqslant 1)$ is a sequence of iid uniformly distributed rvs over $[0,1]$ and set $X_1 = g(U_1, \ldots, U_d)$, $X_2 = g(U_{d+1}, \ldots, U_{2d})$ .... Then the sequence $(X_i,\ i \geqslant 1)$ is a sequence of iid rvs under the distribution X and a good approximation of I is given by

$$\frac{1}{n}(X_1 + \ldots + X_n).$$

This quantity is called the *empirical mean of the sample*.

We remark that this method is easy to program and it does not depend on the regularity of g, which can be simply measurable.

More generally, we often want to evaluate an integral in $\mathbb{R}^d$ of the form

$$I = \int_{\mathbb{R}^d} g(x_1, \ldots, x_d) f(x_1, \ldots, x_d) dx_1 \ldots dx_d$$

where $f(x)$ is positive and sums to one (i.e. $\int f(x)dx = 1$). Then I can be written in the form $\mathbb{E}(g(X))$ with X a rv valued in $\mathbb{R}^d$ having probability density function f with respect to the Lebesgue measure. We can therefore approximate I by

$$\widehat{I}_n := \frac{1}{n}(g(X_1) + \ldots + g(X_n)),$$

if $(X_i,\ i \geqslant 1)$ is sampled from the distribution $f(x)dx$.

One can easily check that

Proposition 1.1 *$\widehat{I}_n$ is an unbiased estimator of I (which means that $\mathbb{E}(\widehat{I}_n) = I$).*

Probabilistic questions
→ How and when does this method converge?
→ What can we say about the precision of the approximation i.e. what is the rate of convergence?

1.2 Limits and convergence

The answers of the previous questions are given by two of the most important probabilistic theorems.

Theorem 1.2 (Strong Law of Large Numbers) *Let $(X_i,\ i \geqslant 1)$ be a sequence of iid rvs distributed as a rv X. We assume that $\mathbb{E}(|X|) < +\infty$. Then for almost every ω*

$$\frac{1}{n}(X_1 + \ldots + X_n) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}(X).$$

This theorem then states that the empirical mean is a "good" approximation of I in the case where the function g is integrable (which is not surprising):

$$\widehat{I}_n = \frac{1}{n}(g(X_1) + \ldots + g(X_n)) \xrightarrow[n\to\infty]{} \mathbb{E}(g(X)) = \int_{\mathbb{R}^d} g(x_1, \ldots, x_d) f(x_1, \ldots, x_d) dx = I,$$

where $x = (x_1, \ldots, x_d)$.
The (random) error committed is given by

$$\epsilon_n = \mathbb{E}(g(X)) - \frac{1}{n}\left[g(X_1) + \ldots + g(X_n)\right] = I - \widehat{I}_n.$$

We want to evaluate this error. The Central Limit Theorem gives a quantity that is asymptotically equal (in distribution) to the random error ϵ_n but which is also random (standard Gaussian distributed).

Theorem 1.3 (Central Limit Theorem) *Let $(X_i,\ i \geqslant 1)$ be a sequence of iid rvs distributed as a rv X. We assume that $\mathbb{E}(X^2) < +\infty$ and denote by σ^2 the variance of X. Then*

$$\frac{\sqrt{n}}{\sigma}\epsilon_n \xrightarrow[n\to\infty]{d} G,$$

where G stands for a rv with a standard Gaussian distribution.
This means that if h is a bounded Borel function $\mathbb{E}\left(h\left(\frac{\sqrt{n}}{\sigma}\epsilon_n\right)\right)$ converges to $\mathbb{E}(h(G))$ as n goes to infinity.

Confidence intervals

We note that the Central Limit Theorem never allows us to bound the random error ϵ_n since the support of a Gaussian is equal to $\mathbb{R}$. Nevertheless it leads to a description of the error of the Monte Carlo method by giving the standard deviation $\frac{\sigma}{\sqrt{n}}$ of ϵ_n or by giving a $(1-\alpha)$%-confidence interval (CI) for the result. That means that the result is found with $(1-\alpha)$% chance in the given interval (and with α% chance of being outside). Indeed, we can deduce from the previous theorem that for all $c_1 < c_2$,

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\sigma}{\sqrt{n}}c_1 \leqslant \epsilon_n \leqslant \frac{\sigma}{\sqrt{n}}c_2\right) = \int_{c_1}^{c_2} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}.$$

In practical applications, we approximate ϵ_n by a centered Gaussian distribution with variance $\frac{\sigma^2}{n}$. In our context, we then derive the following asymptotic approximation for $\mathbb{E}(g(X))$:

$$\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}|\epsilon_n| \leqslant z_\alpha\right) = \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}|I - \widehat{I}_n| \leqslant z_\alpha\right) \approx \mathbb{P}(|G| \leqslant z_\alpha) = 1 - \alpha,$$

where z_α is the $(1-\alpha)$-quantile of the absolute value of a standard Gaussian distribution. In other words,

$$\mathbb{P}\left(\mathbb{E}(g(X)) \in \left[\widehat{I}_n - \frac{\sigma}{\sqrt{n}}z_\alpha, \widehat{I}_n + \frac{\sigma}{\sqrt{n}}z_\alpha\right]\right) \approx 1 - \alpha$$

or else the $(1-\alpha)$%-CI is given by

$$\left[\mathbb{E}(g(X)) - \frac{\sigma}{\sqrt{n}}z_\alpha, \mathbb{E}(g(X)) + \frac{\sigma}{\sqrt{n}}z_\alpha\right].$$

Estimate of the variance of the estimation

The result above shows that it is important to know the order of the size of the variance σ of the rv used in the Monte Carlo technique. The variance σ_n^2 of the estimator $\widehat{I}_n$ based on a n-sample $(X_i, i \geqslant 1)$ is given by

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(g(X_i) - \widehat{I}_n\right)^2.$$

σ_n^2 is called the *empirical variance* of the sample. Dividing by $n-1$ instead of n conduces to an unbiased estimator (which means that $\mathbb{E}(\sigma_n^2) = \sigma^2$). Although from the practical point of view it is not relevant since n will be usually large enough.

We can therefore obtain a $(1-\alpha)$%-CI by replacing σ by σ_n in the CI given by the Central Limit Theorem:

$$\left[\widehat{I}_n - \frac{\sigma_n}{\sqrt{n}}z_\alpha, \widehat{I}_n + \frac{\sigma_n}{\sqrt{n}}z_\alpha\right].$$

We therefore see that with no extra calculation (evaluating σ_n on the sample already generated) we could give a dependable estimate of the approximation error of I by $\widehat{I}_n$. It is one of the greatest advantages of the Monte Carlo method to give a realistic estimate of the error at a minimum cost.

Confidence intervals for rare event probabilities

If the quantity of interest I is a rare event probability (say less than 10^{-9}), one might be cautious studying CI for ϵ_n. In that case, it is more correct to study the relative random error instead of ϵ_n itself and by the way

$$\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\frac{|I - \widehat{I}_n|}{I} \leqslant z_\alpha\right) \text{ instead of } \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}|I - \widehat{I}_n| \leqslant z_\alpha\right).$$

We are then led to a $(1-\alpha)$%-CI of the kind

$$\left[\mathbb{E}(g(X)) - \frac{\sigma}{\sqrt{n}}z_\alpha I, \mathbb{E}(g(X)) + \frac{\sigma}{\sqrt{n}}z_\alpha I\right],$$

which means that

$$\mathbb{P}\left(\mathbb{E}(g(X)) \in \left[\widehat{I}_n - \frac{\sigma}{\sqrt{n}}z_\alpha I, \widehat{I}_n + \frac{\sigma}{\sqrt{n}}z_\alpha I\right]\right) \approx 1 - \alpha.$$

As a consequence, on the one hand, more the variance σ^2 will be small more the CI will be precise (which is true even in the general case). Hence the need to reduce the variance of the variable under concern.

On the other hand that type of crude simulation becomes inefficient to estimate rare event probabilities as showed in the following example

Example 1.1 *In telecommunication, the loss probability of a packet of information is less than* 10^{-9}. *In other words, one must simulate a billion of information packets by loss packet. To achieve a good approximation one needs the simulation of at least* 100 *billions of packets that would take hundred of days.*

Moreover since $\mathbb{P}(|G| \leqslant 1.96) \approx 0.95$,

$$\mathbb{P}\left(\frac{|I - \widehat{I}_n|}{I} \leqslant t\right) \approx \mathbb{P}\left(|G| \leqslant \frac{\sqrt{n}}{\sigma} tI\right) \approx 0.95$$

iif

$$\frac{\sqrt{n}}{\sigma} tI = \sqrt{\frac{n}{I(1-I)}} tI \approx \sqrt{nI}t \approx 1.96 \quad i.e. \quad n \approx \left(\frac{1.96}{t}\right)^2 \frac{1}{I} \approx 3.84\ t^{-2} 10^9.$$

To illustrate this let us assume that we want a normalized relative error (RE) less than 10%. *The constraint* $RE \leqslant 0.1$ *translates into* $n \geqslant 3.84\ 10^{11}$. *In other words, we need a few hundred billion experiments to get a modest* 10% *relative error with a confidence of* 95%.

If the system being simulated is complex enough, this will be impossible and something different must be done in order to provide the required estimation. Thus one needs to introduce speed-up techniques of simulation. More formally if we want to assess a fixed RE while the event probability goes to zero, we need to increase the sample size as

$$n = \left(\frac{z_\alpha}{RE}\right)^2 \frac{1}{I},$$

that is, in inverse proportion to I.

Example 1.2 *Assume that we want to calculate* $\mathbb{E}(e^{\beta G})$ *where* G *is a standard Gaussian rv. It is known that*

$$E := \mathbb{E}(e^{\beta G}) = e^{\beta^2/2}.$$

If we apply a Monte Carlo method, we take $X = e^{\beta G}$ *and the variance of* X *is then given by* $\sigma^2 = e^{2\beta^2} - e^{\beta^2}$. *Using a* n*-sample, the average relative error is of order of* $\sigma/(E\sqrt{n}) = \sqrt{(e^{\beta^2} - 1)/n}$. *If we want to achieve an order of magnitude* t *for the RE, we see that this means that we must take* $n \approx 4(e^{\beta^2} - 1)/t^2$. *If* $t = 0.1$ *and* $\beta = 5$, *we get* $n = 7\ 10^{12}$ *which is too large. The following tabular contains the results of a simulation based on* 10^5 *trials in the case* $\beta = 5$*:*

	exact value	:	268337
n=100000	estimated 95% CI	:	[-467647,2176181]
	estimated value	:	854267

This approximation is really far to be precise! But importantly the calculated CI contains the exact value. This is the reassuring aspect of the Monte Carlo method: the approximation may be mediocre but we are well aware of it. This example shows the limit of the Monte Carlo method when the variance of the rv used is large.

Example 1.3 *In financial applications, we have to calculate quantities of the type*

$$C = \mathbb{E}\left(\left(e^{\beta G} - K\right)_+\right), \tag{1}$$

G *being a standard Gaussian rv and* $x_+ = \max(0, x)$. *These quantities represent the price of an option to buy, commonly called a "call". In this case, we can give an explicit formula [19]:*

$$C = e^{\beta^2/2} N\left(\beta - \frac{\log K}{\beta}\right) - KN\left(-\frac{\log K}{\beta}\right)$$

where N *is the probability distribution of a standard Gaussian rv that is* N *is defined by* $N(x) = \int_{-\infty}^{x} e^{-u^2/2} \frac{du}{\sqrt{2\pi}}$. *We apply Monte Carlo simulation to estimate* C *and compare the exact value to results of a simulation based on various sizes of samples in the case* $\beta = K = 1$.

	exact value	:	6.72
n=100	estimated 95% CI	:	[0.08,11.39]
	estimated value	:	5.74
n=1000	estimated 95% CI	:	[4.20,10.01]
	estimated value	:	7.1
$n=10^4$	estimated 95% CI	:	[6.13,8.43]
	estimated value	:	7.28
$n=10^5$	estimated 95% CI	:	[6.59,7.69]
	estimated value	:	7.14

Let us now compare the results with those obtained when evaluating an option to sell, called "put", that is

$$P = \mathbb{E}\left(\left(K - e^{\beta G}\right)_+\right). \tag{2}$$

The explicit formula gives

$$P = KN\left(\frac{\log K}{\beta}\right) - e^{\beta^2/2}N\left(\frac{\log K}{\beta} - \beta\right).$$

We then obtain

	exact value	:	0.23842
n=100	estimated 95% CI	:	[0.166,0.276]
	estimated value	:	0.220
n=1000	estimated 95% CI	:	[0.221,0.258]
	estimated value	:	0.240
n=10000	estimated 95% CI	:	[0.232,0.244]
	estimated value	:	0.238

Here the approximation is much better than in the case of a call that can easily be proved by a calculation of the variance.

2 Methods to reduce the variance and Importance Sampling

We have seen that the rate of convergence of the Monte Carlo method is of the order of $\sigma/\sqrt{n}$. There are numerous techniques (called reduction of variance techniques) to improve this method, which try to reduce the value σ^2. The general idea is to give another representation, in the form of an expected value, of the quantity to be calculated:

$$\mathbb{E}(X) = \mathbb{E}(Y),$$

trying to reduce the variance. We are going to go through several of these methods that are applicable in practically all simulations.

See e.g. [20] and [21] for more details.

2.1 Importance Sampling

Importance Sampling (IS) is probably one of the most popular approach in rare event simulation. The general setting is as follows. Assume that we want to calculate $I = \mathbb{E}(g(X))$ where the distribution of X is given by $f(x)dx$. The quantity that we want to estimate is then

$$I = \mathbb{E}(g(X)) = \int g(x)f(x)dx.$$

In that view, we introduce a new function $\widetilde{f}$ such that $\widetilde{f} > 0$ and $\int \widetilde{f}(x)dx = 1$. Obviously, the quantity to estimate can be written as

$$\mathbb{E}(g(X)) = \int \frac{g(x)f(x)}{\widetilde{f}(x)}\widetilde{f}(x)dx = \mathbb{E}\left[\frac{g(Y)f(Y)}{\widetilde{f}(Y)}\right],$$

if Y is distributed as $\widetilde{f}(x)dx$. This means that we therefore have another method of estimating $\mathbb{E}(g(X))$ based on a n-sample $Y_1, Y_2, \ldots, Y_n$ distributed as Y, the approximation being

$$\widehat{I}_n := \frac{1}{n}\sum_{i=1}^{n}\frac{g(Y_i)f(Y_i)}{\widetilde{f}(Y_i)}.$$

This procedure will be efficient if the rv Z defined by $Z = \frac{g(Y)f(Y)}{\widetilde{f}(Y)}$ has a smaller variance than that of $g(X)$. Easily we have

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{g(Y)f(Y)}{\widetilde{f}(Y)}\right) = \int \frac{g(x)^2 f(x)^2}{\widetilde{f}(x)}dx - \mathbb{E}(g(X))^2.$$

The quantity $L(x) := \frac{f(x)}{\widetilde{f}(x)}$ is called the *likelihood ratio*.
Note that if $g > 0$, the function

$$\widetilde{f}(x) = \frac{g(x)f(x)}{\mathbb{E}(g(X))}$$

cancels the variance! This means that there is an optimal change of measure leading to a zero-variance estimator. The simulation becomes a kind of "pseudo-simulation" leading to the exact value in only one sample (unbiased estimator with variance equal to zero). Unfortunately, this result is not tractable since this optimal function $\widetilde{f}$ depends on $\mathbb{E}(g(X))$ the quantity to evaluate!

Nevertheless, this observation leads to two remarks: first there is an optimal change of measure which suggests that there are other good and even very good changes of measures. Second, it allows us to justify the following heuristic: in practice, we first choose $\widetilde{f}$ as close as possible as $|gf|$ then we proceed to a normalization to recover a probability density function.

Remark 2.1 *In order to avoid the calculation of the normalizing constant, we use the following estimate*

$$\widetilde{I}_n = \frac{\sum_{i=1}^n g(Y_i)f(Y_i)/\widetilde{f}(Y_i)}{\sum_{i=1}^n f(Y_i)/\widetilde{f}(Y_i)} = \sum_{i=1}^n g(Y_i)\omega_i$$

where $Y_1, \ldots, Y_n$ *are iid rvs with common distribution* $\widetilde{f}(x)dx$ *and the importance weights* $\omega_1, \ldots, \omega_n$ *are given by*

$$\omega_i = \frac{f(Y_i)/\widetilde{f}(Y_i)}{\sum_{i=1}^n f(Y_i)/\widetilde{f}(Y_i)}.$$

For a fixed n, $\widetilde{I}_n$ *is biased but it is asymptotically unbiased.*

Remark that $\widetilde{I}_n$ *is nothing more than the function* $g(x)$ *integrated with respect to the empirical measure*

$$\sum_{i=1}^n \omega_i \delta_{Y_i}(dy)$$

where δ_a *is the Dirac measure at* a.

Example 2.1 *We present now a first simple example to fix ideas. Assume that we want to estimate*

$$\int_0^1 \cos\left(\frac{\pi x}{2}\right) dx;$$

that corresponds to $g(x) = \cos(\frac{\pi x}{2})$ *and* X *uniformly distributed over* $[0,1]$. *We will approximate* g *by a second degree polynomial. Since* g *is even and equals to 0 at* $x = 1$ *and to 1 at* $x = 0$, *it is natural to take* $\widetilde{f}(x)$ *in the form* $\lambda(1-x^2)$ *and more precisely equals to* $\frac{3}{2}(1-x^2)$ *to satisfy the constraint* $\int \widetilde{f}(x)dx = 1$. *A calculation of the variance of* $Z = g(Y)f(Y)/\widetilde{f}(Y)$ *shows that we have reduced the variance by a factor of 100.*

Example 2.2 *Let us consider Example 1.3 again. We shall show how to apply this method in the case of the calculation of a put (2). The function* $x \mapsto e^x - 1$ *is close to* x *for small values of* x. *This suggests to rewrite* P *as*

$$P = \int_{\mathbb{R}} \frac{(K - e^{\beta x})_+}{\beta|x|} \beta|x| e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \int_0^{+\infty} \frac{(K - e^{\beta\sqrt{y}})_+ + (K - e^{-\beta\sqrt{y}})_+}{\sqrt{2\pi y}} e^{-y/2} \frac{dy}{2},$$

which means that

$$P = \mathbb{E}\left(\frac{(K - e^{\beta\sqrt{Y}})_+ + (K - e^{-\beta\sqrt{Y}})_+}{\sqrt{2\pi Y}} \right)$$

if Y *is exponentially distributed with parameter* $1/2$. *We then obtain*

	exact value	:	*0.23842*
$n = 100$	*estimated* 95% *CI*	:	*[0.239,0.260]*
	estimated value	:	*0.249*
$n = 1000$	*estimated* 95% *CI*	:	*[0.235,0.243]*
	estimated value	:	*0.239*
$n = 10^4$	*estimated* 95% *CI*	:	*[0.237,0.239]*
	estimated value	:	*0.238*

We note a significant improvement with respect to the results based on the classical Monte Carlo: for a 10^4*-sample, the RE becomes* 1% *instead of* 6%.

Rare event framework

We have seen that the basic idea of IS consists in changing the dynamics of the simulation in order that the rare event occurs more frequently, which is done by changing the underlying distribution of the system under concern. In the rare event setting, the quantity of interest writes

$$\Gamma = \mathbb{P}(X \in A) = \mathbb{E}(\mathbb{1}_A(X)) = \int_{-\infty}^{+\infty} \mathbb{1}_A(x) f(x) dx$$

and can be estimated trough the Monte Carlo method using a n-sample $X_1, \ldots, X_n$ iid with common density $f(x)dx$. It yields the following unbiased estimator

$$\widehat{\Gamma}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i).$$

As shown previously, the necessary sampling size to achieve a RE less than t with probability $1-\alpha$ should be

$$n = \left(\frac{z_\alpha \sigma_n}{t\Gamma} \right)^2,$$

i.e. proportional to the inverse of the square root of the rare event probability Γ.

Applying IS methodology, we rewrite Γ in the form

$$\Gamma = \int_{-\infty}^{+\infty} \mathbb{1}_A(x) \frac{f(x)}{\widetilde{f}(x)} \widetilde{f}(x) dx = \mathbb{E}(\mathbb{1}_A(Y) L(Y))$$

Now with a n-sample $Y_1, \ldots Y_n$ distributed following $\widetilde{f}(x)dx$, an unbiased estimate of Γ is given by

$$\widehat{\Gamma}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(Y_i) L(Y_i).$$

To find a good change of measure, one needs to have a good knowledge of the system under study. Moreover, it is possible that IS does not lead to an improvement and even with a bad change of measure, the result can be worse!

Example 2.3 *To fix ideas and better understand the difficulties to choose a good change of measure, we study the discrete-time Markov chain Y such as the one depicted in Figure 1 with state space $S = \{0, 1, 2, 3\}$ and $0 < a, b, c, d < 1$. The chain starts at 1 and we wish to evaluate the probability that it gets absorbed by state 3, that is to say $I = \mathbb{P}(X(\infty) = 3 | X(0) = 1)$. Obviously here, $I = ac/(1 - ad)$. For instance, when a and c are small, the event $\{X(\infty) = 3\}$ becomes rare.*

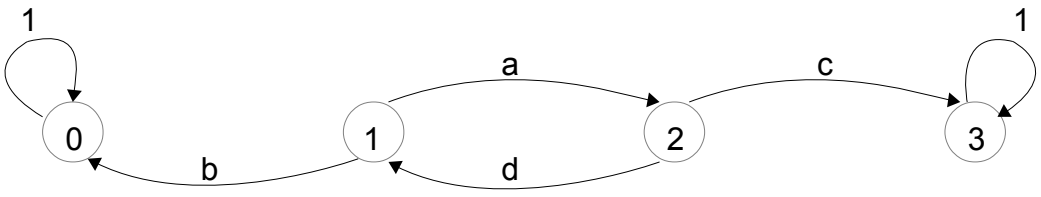

Figure 1: A small discrete-time Markov chain.

For instance, consider the case $a = c = \frac{1}{4}$ and suppose that we decide to make the event of interest $\{X(\infty) = 3\}$ more frequent by changing a to $\widetilde{a} = \frac{1}{2}$ and c to $\widetilde{c} = \frac{3}{4}$. Define $\mathcal{P}$ as the set of all possible paths of X starting at sate 1:

$$\mathcal{P} = \{\pi = (x_0, x_1, \ldots, x_K), K \geqslant 1 \text{ with } x_0 = 1,\ x_K = 0 \text{ or } 3 \text{ and } x_i \notin \{0, 3\} \text{ if } 1 \leqslant i \leqslant K-1\}$$

and $\mathcal{P}_s$ as the set of successful paths (those paths in $\mathcal{P}$ ending with state 3. Observe that

$$\mathcal{P}_s = \{\pi_k,\ k \geqslant 1\}$$

where $\pi_k = (1,(2,1)^k,2,3)$ (the notation $(2,1)^k$ meaning that the sequence $(2,1)$ is repeated k times. We have

$$\mathbb{P}(\pi_k) = (ad)^k ac = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \end{pmatrix}^k \begin{matrix} \frac{1}{4} & \frac{1}{4} \end{matrix} \quad \text{and} \quad \mathbb{P}(\widetilde{\pi}_k) = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \end{pmatrix}^k \begin{matrix} \frac{1}{2} & \frac{3}{4} \end{matrix}.$$

It can then be verified that $\mathbb{P}(\widetilde{\pi}_k) > \mathbb{P}(\pi_k)$ for $k = 0,\ldots,4$ but that $\mathbb{P}(\widetilde{\pi}_k) < \mathbb{P}(\pi_k)$ for $k \geqslant 5$. We see that even in such a simple model, finding an appropriate change of measure can be non-trivial.

Before leaving this example, consider the following IS scheme. Change a to $\widetilde{a} = 1$ and c to $\widetilde{c} = 1 - ad$. We can check that

$$L(\pi_k) = \frac{ac}{1-ad} = I$$

for all k which means that this is the optimal change of measure, the one leading to a zero-variance estimator.

The IS method is the most used in practice and in particular in the rare event context. Nevertheless there exist other strategies to reduce the variance of the estimates that we present in the following subsection.

2.2 Other related techniques

2.2.1 Control variables

The principle is the same as the one of IS: we want to evaluate $\mathbb{E}(g(X))$ that we write in the form

$$\mathbb{E}(g(X)) = \mathbb{E}(g(X) - h(X)) + \mathbb{E}(h(X))$$

where $\mathbb{E}(g(X))$ can be calculated explicitly and $\mathrm{Var}(g(X)-h(X))$ is much more smaller than $\mathrm{Var}(g(X))$. We then use a Monte Carlo method to evaluate $\mathbb{E}(g(X) - h(X))$ and a direct evaluation for $\mathbb{E}(h(X))$.

Example 2.4 *We want to compute $\int_0^1 e^x dx$. Remark that $\int_0^1 e^x dx = \mathbb{E}(g(X))$ with $g(x) = e^x$ and X uniformly distributed over $[0,1]$. In a neighborhood of 0, we have $e^x \sim 1 + x$ then we write*

$$\int_0^1 e^x dx = \int_0^1 (e^x - 1 - x)dx + \frac{3}{2}$$

and it is then easy to show that the variance of this method than reduces significantly.

Example 2.5 *We now give another example by considering the price of a call (1) of Example 1.3. It is easy to verify that the price P of the put and that of the call satisfy the relation*

$$C - P \quad = \quad \mathbb{E}\left(e^{\beta G} - K\right) = e^{\beta^2/2} - K.$$

The idea is then to write $C = P + e^{\beta^2/2} - K$ and to carry out a Monte Carlo method for P.

2.2.2 Antithetic variables

Assume that we want to calculate

$$I = \int_0^1 g(x)dx = \mathbb{E}(g(X))$$

where X is uniformly distributed over $[0,1]$.

Since $x \mapsto 1 - x$ leaves the measure dx invariant, we also have

$$I = \frac{1}{2}\int_0^1 (g(x) + g(1-x))dx$$

We can therefore apply the Monte Carlo technique to estimate I by

$$\widehat{I}_{2n} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(g(X_i)+g(1-X_i)) = \frac{1}{2n}\sum_{i=1}^{n} g(X_i)+g(1-X_i),$$

where $(X_i,\ i \geqslant 1)$ is a sequence of iid rvs uniformly distributed over $[0,1]$.
We can prove that if the function g is continuous and monotone, the quality of the approximation is improved with respect to a direct Monte Carlo method based on $2n$ realizations of the rv X. This technique can be straightforwardly generalized to higher dimensions and to other transformations preserving the distribution of the rv.

Example 2.6 *If we want to calculate the price of a put (2) of Example 1.3, we can use the fact that the distribution of G is identical to that of $-G$ and reduce the variance of a coefficient almost by 2.*

2.2.3 Method of stratification

This method is well known to statisticians and often used in surveys (see [5]). Assume that we want to calculate

$$I = \mathbb{E}(g(X)) = \int_{\mathbb{R}^d} g(x)f(x)dx$$

where X is a rv valued in $\mathbb{R}^d$ following the distribution $f(x)dx$.

We take a partition $(D_i, 1 \leqslant i \leqslant m)$ of $\mathbb{R}^d$ and we decompose the integral in the following way

$$I = \sum_{i=1}^{m}\mathbb{E}(\mathbb{1}_{D_i}(X)g(X)) = \sum_{i=1}^{m}\mathbb{E}(g(X)|X\in D_i)\mathbb{P}(X\in D_i)$$

When we know the numbers $p_i := \mathbb{P}(X \in D_i)$, we can use a Monte Carlo method to estimate the integrals $I_i := \mathbb{E}(g(X)|X \in D_i)$ by distributing optimally the n realizations. Assume that we approximate the integral I_i by $\widehat{I}_i$ based on n_i independent trials. The variance of the approximation error is then given by σ_i^2/n_i, if we denote $\sigma_i^2 := \mathrm{Var}\,(g(X)|X \in D_i)$. We then approximate I by

$$\widehat{I} := \sum_{i=1}^{m} p_i\widehat{I}_i.$$

Since the samples used to obtain the estimates $\widehat{I}_i$ are assumed to be independent, the variance of the estimate $\widehat{I}$ is obviously given by

$$\sum_{i=1}^{m} p_i^2\frac{\sigma_i^2}{n_i}.$$

It is then natural to minimize this error for a fixed number of trials $n = \sum_{i=1}^{m} n_i$. We can check that the n_i's minimizing the variance of $\widehat{I}$ are given by

$$n_i = n\frac{p_i\sigma_i}{\sum_{i=1}^{m} p_i\sigma_i}.$$

The minimum of the variance of $\widehat{I}$ then becomes

$$\frac{1}{n}\left(\sum_{i=1}^{m} p_i\sigma_i\right)^2$$

which is less than the variance obtained with n random trials by the classical Monte Carlo method. In fact, the variance becomes

$$\begin{aligned}
\mathrm{Var}(g(X)) &= \mathbb{E}(g(X)^2) - \mathbb{E}(g(X))^2 \\
&= \sum_{i=1}^{m} p_i\mathbb{E}(g(X)^2|X\in D_i) - \left(\sum_{i=1}^{m} p_i\mathbb{E}(g(X)|X\in D_i)\right)^2 \\
&= \sum_{i=1}^{m} p_i\mathrm{Var}(g(X)|X\in D_i) + \sum_{i=1}^{m} p_i\mathbb{E}(g(X)|X\in D_i)^2 - \left(\sum_{i=1}^{m} p_i\mathbb{E}(g(X)|X\in D_i)\right)^2
\end{aligned}$$

by using the definition of the conditional variance. We then use twice the convexity inequality for x^2

$$\left(\sum_{i=1}^{m} p_i a_i\right)^2 \leqslant \sum_{i=1}^{m} p_i a_i^2$$

if $\sum_{i=1}^{m} p_i = 1$ to show that

$$\mathrm{Var}(g(X)) \geqslant \sum_{i=1}^{m} p_i \mathrm{Var}(g(X)|X \in D_i) \geqslant \left(\sum_{i=1}^{m} p_i \sigma_i\right)^2,$$

which proves that provided we have an optimal strategy of trials, we can obtain by stratification, an approximation with lower variance.

Remark 2.2 *Unfortunately, note that it is possible to obtain an approximation with greater variance than the initial estimate if the assignment of the points is arbitrary. Despite this, there exist other strategies to choose the points on domains that reduce the variance. For example, if we assigns a number of points proportional to the probability of the domain: $n_i = np_i$, we then obtain an approximation with variance equals to*

$$\frac{1}{n}\sum_{i=1}^{m} p_i \sigma_i^2.$$

Now we see that $\sum_{i=1}^{m} p_i \sigma_i^2$ is a bound for $\mathrm{Var}(g(X))$. This allocation strategy is sometimes useful when we explicitly know the probabilities p_i.

2.2.4 Average value or conditioning

Assume we want to calculate

$$\mathbb{E}(g(X,Y)) = \int g(x,y) f(x,y) dxdy$$

where $f(x,y)dxdy$ is the distribution of the pair (X,Y). Let

$$h(x) = \frac{1}{\int f(x,y)dy} \int g(x,y) f(x,y) dy$$

then $\mathbb{E}(g(X,Y)) = \mathbb{E}(h(X))$. Indeed, the distribution of X is given by $m(x)dx = \left(\int f(x,y)dy\right) dx$ and thus

$$\mathbb{E}(h(X)) = \int h(X) m(x) dx = \int dx \int g(x,y) f(x,y) dy = \mathbb{E}(g(X,Y)).$$

We can recover that result noting that

$$\mathbb{E}(g(X,Y)|X) = h(X).$$

This interpretation as a conditional expectation allows us to prove that the variance of $h(X)$ is lower than that of $g(X,Y)$. If we can not calculate directly $h(x)$, we use a Monte Carlo technique for $h(X)$.

2.3 Exercises

Exercise 2.1 ***[Importance sampling]*** *Suppose we want to evaluate the integral $\mu(G) = \int G(x)\mu(dx)$ of a nonnegative and bounded potential function G with respect to some distribution μ on some measurable space $(E, \mathcal{E})$. We associate with a sequence of independent random variables $(X_i)_{i \geqslant 1}$ with common distribution μ the empirical measures $\mu^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}$.*

- *Check that $\mathbb{E}(\mu^N(G)) = \mu(G)$ and*

$$N\mathbb{E}\left(\left(\mu^N(G) - \mu(G)\right)^2\right) = \mu\left(\left(G - \mu(G)\right)^2\right) =: \sigma_\mu(G).$$

- *For any probability measure $\overline{\mu}$ such that $\mu \ll \overline{\mu}$, prove that $\mu(G) = \overline{\mu}(\overline{G})$ with $\overline{G} = G\frac{d\mu}{d\overline{\mu}}$. We let $\overline{\mu}^N := \frac{1}{N}\sum_{i=1}^N \delta_{Y_i}$ be the occupation measure associated with a sequence of N independent random variables $(Y_i)_{i\geqslant 1}$ with common distribution $\overline{\mu}$. Prove that $\mathbb{E}(\overline{\mu}^N(\overline{G})) = \mu(G)$ and*

$$\begin{aligned} N\mathbb{E}\left(\left(\overline{\mu}^N(\overline{G}) - \mu(G)\right)^2\right) &= \overline{\mu}\left(\left(\overline{G} - \mu(G)\right)^2\right) =: \sigma_{\overline{\mu}}(\overline{G}) \\ &= \sigma_\mu(G) - \mu\left(G^2\left(1 - \frac{d\mu}{d\overline{\mu}}\right)\right). \end{aligned}$$

- ***An example of potential** G. Roughly speaking, from the equation above, we see that a reduction of variance is obtained as soon as $\overline{\mu}$ is chosen such that $\frac{d\mu}{d\overline{\mu}} < 1$ on regions where G is more likely to take large values. In other words, it is judicious to choose a new reference distribution $\overline{\mu}$ so that the sampled particles $\overline{X}_i$ are more likely to visit regions with high potential. For instance, if $G = \mathbb{1}_A$ is the indicator function of some measurable set $A \in \mathcal{E}$, then prove that*

$$\sigma_{\overline{\mu}}(\overline{G}) = \sigma_\mu(G) - \mu\left(1_A\left(1 - \frac{d\mu}{d\overline{\mu}}\right)\right).$$

  If we choose $\overline{\mu}$ such that $\frac{d\mu}{d\overline{\mu}} \leqslant 1 - \delta$ for any $x \in A$, then check that

$$\overline{\mu}(A) \geqslant \mu(A)/(1-\delta) \qquad \text{and} \qquad \sigma_{\overline{\mu}}(\overline{G}) + \delta\mu(A) \leqslant \sigma_\mu(G).$$

- ***The optimal choice**. Show that the optimal distribution $\overline{\mu}$ is the Boltzmann-Gibbs measure $\overline{\mu}(dx) = \mu(G)^{-1}G(x)\mu(dx)$ in the sense that $\sigma_{\overline{\mu}}(\overline{G}) = 0$. This optimal strategy is clearly hopeless since the normalizing constant $\mu(G)$ is precisely the constant we want to estimate!*

- ***A bad choice**. Consider now the distribution $\overline{\mu}$ defined by $\overline{\mu}(dx) = \mu(G^{-2})^{-1}G^{-2}(x)\mu(dx)$, then check that*

$$\sigma_{\overline{\mu}}(\overline{G}) \geqslant \mu(G^4)/\mu(G^2) - \mu(G)^2 \geqslant \sigma_\mu(G).$$

Exercise 2.2 ***[Simple random walk]** Let $(\epsilon_n)_{n\geqslant 0}$ be independent and identically distributed random variables with common law $\mathbb{P}(\epsilon_n = 1) = 1 - \mathbb{P}(\epsilon_n = -1) = p \in (0,1)$. We consider the simple random walk X_n on $\mathbb{Z}$ defined by $X_n = \sum_{p=0}^n \epsilon_p$. Suppose we want to evaluate (using a Monte Carlo scheme) the probability that X_n enters a subset $A \subset \mathbb{N}^*$. if we have $p < 1/2$, then the random walk X_n tends to move to the left. One natural way to increase the probability that the random walk visits the set A is to change p by some $\overline{p} \in (p,1)$. In this case, the random walk Y_n defined as X_n by replacing p by $\overline{p}$ is more likely to move to the right and as a result the event $\{Y_n \in A\}$ is more likely than $\{X_n \in A\}$. The expected value of $f(X_n) = \mathbb{1}_A(X_n)$ and the particle approximation mean using the standard Monte Carlo method are given respectively by*

$$\mathbb{E}(f(X_n)) = \mathbb{P}(X_n \in A) \qquad \text{and} \qquad \overline{f(X_n)} = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_A(X_n^i)$$

where $\left(X_n^i\right)_{i\leqslant 1}$ is a collection of independent copies of X_n.

- *We let P_n be the distribution of the random sequence $(\epsilon_p)_{0\leqslant p\leqslant n} \in \{-1,+1\}^{n+1}$. Check that*

$$P_n(d(u_0,\ldots,u_n)) = (p(1-p))^{(n+1)/2}\left(p/(1-p)\right)^{\sum_{k=0}^n u_k/2}.$$

- *We let $\overline{P}_n$ be the distribution of the random sequence $(\overline{\epsilon}_p)_{0\leqslant p\leqslant n} \in \{-1,+1\}^{n+1}$ defined as $(\epsilon_p)_{0\leqslant p\leqslant n} \in \{-1,+1\}^{n+1}$ by replacing p by $\overline{p} \in (0,1)$. Deduce from the first question that $\overline{P}_n \ll P_n$ and*

$$\frac{dP_n}{d\overline{P}_n}(u_0,\ldots,u_n) = G_n\left(\sum_{k=0}^n u_k\right) \qquad \text{with} \qquad G_n(x) = \left(\frac{p(1-p)}{\overline{p}(1-\overline{p})}\right)^{\frac{n+1}{2}}\left(\frac{p(1-\overline{p})}{\overline{p}(1-p)}\right)^{\frac{x}{2}}.$$

- *Check that* $\mathbb{E}(f(X_n)) = \mathbb{E}(f(Y_n)G_n(Y_n))$ *for any* $f \in \mathcal{B}_b(\mathbb{Z})$.
- *Let* $\left(Y_n^i\right)_{i\leqslant 1}$ *be a collection of independent copies of* Y_n. *By the Central Limit Theorem, prove that the sequence of random variables*

$$\begin{aligned} W_n^N(f) &= \sqrt{N}\left(\overline{f(X_n)} - \mathbb{E}(f(X_n))\right) \\ \overline{W}_n^N(f) &= \sqrt{N}\left(\overline{f(Y_n)G_n(Y_n)} - \mathbb{E}(f(X_n))\right) \end{aligned}$$

converges in law, as $N \to \infty$, *to a pair of Gaussian random variables with mean 0 and respective variance* $\sigma_n^2(f)$ *and* $\overline{\sigma}_n^2(f)$ *defined by*

$$\begin{aligned} \sigma_n^2(f) &= \mathbb{E}\left(f(X_n)^2\right) - \mathbb{E}(f(X_n))^2 \\ \overline{\sigma}_n^2(f) &= \mathbb{E}\left(f(Y_n)^2 G_n(Y_n)^2\right) - \mathbb{E}(f(X_n))^2 \\ &= \sigma_n^2(f) + \mathbb{E}\left(f(X_n)^2\left(G_n(X_n) - 1\right)\right) \end{aligned}$$

- *Prove that for any indicator functions* $f = \mathbb{1}_A$ *with* $A \subset \{G_n \leqslant 1/a_n\}$, *for some* $a_n \geqslant 1$, *we have*

$$\overline{\sigma}_n^2(f) \leqslant a_n^{-1}\mathbb{P}(X_n \in A) - \mathbb{P}(X_n \in A)^2 \leqslant \sigma_n^2(f).$$

3 Branching processes

See e.g. Harris [13], Lyons [22] and Athreya and Ney [3].

Introduced by Galton and Watson in the XIX-th century to study the survival of a family name, branching processes constitute a mathematical model representing the development of a population along the time. Nowadays there are particularly used in demography, genetics and nuclear physics.
Each individual (object) gives birth, independently from each other, to a random number of children.

X_t (resp. X_n) represents the total number of individuals at time t (resp. at generation n) in the continuous framework (space time $T = \mathbb{R}_+$) (resp. in the discrete framework ($T = \mathbb{N}$)).

Assumptions
→ the individuals do not interact from one another;
→ the lifetime of individuals of the same type is the same or has the same distribution;
→ the reproduction distribution does not depend on time and on how many individuals are present.

These assumptions induce that the process $(X_t)_{t\geqslant 0}$ (or $(X_n)_{n\in\mathbb{N}}$) is Markovian.

Probabilistic questions
→ Determine the distribution of X_t.
→ Compute the mean size of the population at time t (and eventually the standard deviation); less precise than the distribution itself but informative however.
→ Does there exist a non-zero probability of extinction of the population?

To determine the distribution of X_t, we will compute its *probability generating function* that characterizes the distribution and is defined by

$$G_{X_t}(s) = \sum_{j=0}^{+\infty} s^j \mathbb{P}([X_t = j]) = \mathbb{E}\left(s^{X_t}\right).$$

We denote $G(s,t) = \mathbb{E}^{[X_0=1]}\left(s^{X_t}\right) = \sum_{j=0}^{+\infty} s^j p_{1,j}(t)$.

Some reminders on the probability generating function:

If X is a rv valued in $\mathbb{N}$, we define

$$G_X(s) = \sum_{k=0}^{+\infty} s^k \mathbb{P}([X = k]).$$

G_X is a power series with convergence radius greater or equal to 1 since $G_X(1) = \sum_{k=0}^{+\infty} \mathbb{P}([X = k]) = 1$.
The probability generating function characterizes the law of the rv. It is then easy to derive the expectation and the variance of X. Indeed,

$$\mathbb{E}(X) = \lim_{s\to 1_-} G'_X(s) = G'_X(1_-) \text{ and } \mathrm{Var}(X) = G''_X(1_-) + G'_X(1_-) - (G'_X(1_-))^2.$$

Moreover, if X and X' are two independent rv valued in $\mathbb{N}$, we have

$$G_{X+X'}(s) = G_X(s)\ G_{X'}(s).$$

Since $(X_t)_{t\geqslant 0}$ (discrete or continuous) is Markovian, we have the following key relation:

Proposition 3.1 $G(s, t+\tau) = G(G(s,\tau), t)$.

Proof $G(s,t+\tau)=\sum_{j=0}^{+\infty} s^j p_{1,j}(t+\tau)$.
But, as $(X_t)_{t\geqslant 0}$ is Markovian, one has $P(t+\tau)=P(t)P(\tau)$ and

$$\sum_{j=0}^{+\infty} s^j p_{1,j}(t+\tau)=\sum_{j=0}^{+\infty} s^j \sum_{k=0}^{+\infty} p_{1,k}(t)p_{k,j}(\tau)=\sum_{k=0}^{+\infty} p_{1,k}(t)\sum_{j=0}^{+\infty} s^j p_{k,j}(\tau)=\sum_{k=0}^{+\infty} p_{1,k}(t)\mathbb{E}^{[X_0=k]}(s^{X_\tau})$$

Moreover if $X_0=k$, $X_\tau=X_{\tau,1}+\ldots+X_{\tau,k}$ where $X_{\tau,i}$ represents the number of the descendents at time τ of the initial i-th individual. The $X_{\tau,i}$'s being k iid rv, we then have $\mathbb{E}^{[X_0=k]}(s^{X_\tau})=G(s,\tau)^k$ and

$$G(s,t+\tau)=\sum_{k=0}^{+\infty} G(s,\tau)^k p_{1,k}(t)=G(G(s,\tau),t).$$

□

3.1 The Galton-Watson branching process

Example 3.1 *a) Survival of a family name: we assume that only the men preserve and transmit their name.*

b) Electron multiplier: we insert on the electron trajectory a series of plates: when the electron hits a plate, it generates a random number of new electrons.

Let X_n be the number of individuals at generation n and Y be the number of direct descendents of an individual. We denote by G the probability generating function of Y and we define recursively G_n by

$$G_1=G \text{ et pour tout } n\geqslant 1,\ G_{n+1}=G_n\circ G=G\circ G_n.$$

We define also

$$m=\mathbb{E}(Y) \text{ et } \sigma^2=\mathrm{Var}(Y).$$

Theorem 3.2 $\mathbb{E}^{[X_r=k]}\left(s^{X_{r+n}}\right)=(G_n(s))^k$.

Proof By the time homogeneity, $\mathbb{E}^{[X_r=k]}\left(s^{X_{r+n}}\right)=\mathbb{E}^{[X_0=k]}\left(s^{X_n}\right)=G(s,n)^k$.
But $G(s,1)=G(s)=G_1(s)$ and, if $G(s,n)=G_n(s)$,by the key relation,

$$G(s,n+1)=G(G(s,n),1)=G(G_n(s))=G_{n+1}(s)$$

which means that $G(s,n)=G_n(s)$ for any $n\geqslant 1$.

□

As a consequence,

Corollary 3.3 *(i)* $\mathbb{E}^{[X_r=k]}(X_{n+r})=km^n$.

(ii) If $X_0=1$, *then* $\mathbb{E}(X_n)=m^n$ *et* $\mathrm{Var}(X_n)=\begin{cases}\sigma^2 m^{n-1}\frac{m^n-1}{m-1} & si\ m\neq 1\\ n\sigma^2 & si\ m=1\end{cases}$.

Proof This result can be derived recursively by deriving once and twice the function $G_{n+1}=G_n\circ G$ at $s=1$. □

Interpretation of (i)
if $m<1$, $\mathbb{E}(X_n)\to 0$, $\mathrm{Var}(X_n)\to 0$: the mean is more and more significant;
if $m>1$, $\mathbb{E}(X_n)\to+\infty$, $\mathrm{Var}(X_n)\to+\infty$: the mean is less significant;
if $m=1$, $\mathbb{E}(X_n)=1$, $\mathrm{Var}(X_n)\to+\infty$: numerous families disappear and others grows widely.

Extinction probability
We want to determine the probability that the population disappears. In that view, we must assume that $\mathbb{P}([Y=0]) \in]0,1[$, or else if $\mathbb{P}([Y=0])=1$, the population will surely disappear and if $\mathbb{P}([Y=0])=0$, extinction could not occur.

Theorem 3.4 *Let* $\pi_0 = \lim_{n\to+\infty} \mathbb{P}([X_n=0])$. *Then* $\pi_0 = 1$ *iif* $m \leqslant 1$ *(i.e.* $G'_Y(1) \leqslant 1$*) and if* $m>1$, π_0 *is the smallest positive solution of* $G(s)=s$.

Proof $\pi(n)_0 = \mathbb{P}([X_n=0]) = G_n(0)$ if $X_0=1$.
On one hand, if $X_n=0$, then $X_{n+1}=0$ and $G_n(0) \leqslant G_{n+1}(0)$. Thus $(\pi(n)_0)_n$ is increasing and upper bounded by 1: it then converges to a limit denoted π_0.
On the other hand, $\pi(n+1)_0 = G_{n+1}(0) = G(G_n(0)) = G(\pi(n)_0)$ and by the continuity of G, $\pi_0 = G(\pi_0)$ taking the limit.

If s_0 is the smallest positive fixed point of G, we have $0 \leqslant s_0$. Hence since G is increasing, $G(0) \leqslant G(s_0) = s_0$ and by successive compositions by G, $G_n(0) \leqslant s_0$ and taking the limit leads to $\pi_0 \leqslant s_0$. Then $\pi_0 = s_0$ since π_0 is a positive fixed point of G.

□

3.2 Discrete processes with a finite number of types

A first step in generalizing the simple Galton-Watson process is the consideration of processes involving several types of individuals. For example, in the reproduction of certain bacteria, the usual form may produce a mutant form that behaves differently. One can think also to the human beings classed following the right-handed and left-handed persons or even by the gender.

In this subsection we restrict ourselves to two types of individuals. Each individual of type (1) gives birth to individuals of type (1) and of type (2) and similarly each individual of type (2) gives birth to individuals of type (1) and of type (2).
At each generation n, we are interested in the size of the population, but also in the numbers of individuals of type (1) and of type (2).
We assume that an individual of type (1) generates at the end of the considered period a random number $Y^{(1)}$ of descendents of type (1) and a random number $Y^{(2)}$ of descendents of type (2). In the same way an individual of type (2) generates at the end of the same period, a random number $Z^{(1)}$ of descendents of type (1) and a random number $Z^{(2)}$ of descendents of type (2). As done in the previous subsection, we also suppose that
$\rightarrow$ the individuals do not interact from one another;
$\rightarrow$ the lifetime of each individual is the same.
Let $X_n = \left(X_n^{(1)}, X_n^{(2)}\right)$ be the number of individuals at generation n. If $X_n=(k_1,k_2)$, then $X_{n+1}^{(i)} = \left(Y_1^{(i)} + \ldots + Y_{k_1}^{(i)}\right) + \left(Z_1^{(i)} + \ldots + Z_{k_2}^{(i)}\right)$ for $i \in \{1,2\}$. We thus have

$$p_{(k_1,k_2)(j_1,j_2)}(n) = P^{[X_m=(k_1,k_2)]}([X_{m+n}=(j_1,j_2)]).$$

Let $e_1=(1,0)$, $e_2=(0,1)$ and for $i\in\{1,2\}$, $p_{(j_1,j_2)}^{(i)} = p_{e_i,(j_1,j_2)}(1)$ (the probability that an individual of type (i) has j_1 descendents of type (1) and j_2 descendents of type (2).
We also define

$$G^{(i)}(s_1,s_2) = \mathbb{E}^{[X_0=e_i]}\left(s_1^{X_1^{(1)}} s_2^{X_1^{(2)}}\right) = \sum_{j_1,j_2} s_1^{j_1} s_2^{j_2} p_{(j_1,j_2)}^{(i)} \text{ and } G_n^{(i)}(s_1,s_2) = \mathbb{E}^{[X_0=e_i]}\left(s_1^{X_n^{(1)}} s_2^{X_n^{(2)}}\right).$$

Theorem 3.5

$$\mathbb{E}^{[X_0=(k_1,k_2)]}\left(s_1^{X_n^{(1)}} s_2^{X_n^{(2)}}\right) = \left(G_n^{(1)}(s_1,s_2)\right)^{k_1}\left(G_n^{(2)}(s_1,s_2)\right)^{k_2} = \sum_{j_1,j_2} s_1^{j_1} s_2^{j_2} p_{(k_1,k_2)(j_1,j_2)}(n).$$

The probability generating function determines entirely the distribution and allows in particular to deduce the population mean size at generation n.
In that view, we define $M = (m_{ij})$, where m_{ij} represents the mean number of descendents of type (j) of an individual of type (i) i.e.

$$m_{ij} = \mathbb{E}^{[X_0=e_i]}(X_1^{(j)}) = \partial_j G^{(i)}(1,1).$$

As a consequence,

Theorem 3.6 $\mathbb{E}^{[X_r=(k_1,k_2)]}(X_{n+r}^{(1)}, X_{n+r}^{(2)}) = (k_1,k_2)M^n$.

Extinction probability

Theorem 3.7 *Let* $\pi_0 = \left(\pi_0^{(1)}, \pi_0^{(2)}\right)$ *where* $\pi_0^{(i)} = \lim_{n\to+\infty} P^{[X_0=e_i]}([X_n = (0,0)])$ *and* ρ *be the largest in absolute value eigenvalue of* $M = (m_{ij})$. *Then* $\pi_0 = (1,1)$ *iif* $\rho \leqslant 1$ *and if* $\rho > 1$, π_0 *is the smallest non negative solution of* $\begin{cases} s_1 = G^{(1)}(s_1,s_2) \\ s_2 = G^{(2)}(s_1,s_2) \end{cases}$

3.3 Markov branching processes (continuous time)

The discrete branching processes of the previous subsections are limited in the sense that the instants of generation are fixed. Even though numerous examples (e.g. in genetics, genealogy) can be modelized in such a way, a large proportion of natural reproducing processes occur in a continuous way. It is then necessary and interesting to introduce branching processes in continuous time.

Let X_t be the number of individuals at time t: $(X_t)_{t\geqslant 0}$ is a continuous Markovian process. Let $p_{kj}(t) = P^{[X_u=k]}([X_{u+t} = j])$. Then we have

$$p_{11}(h) = 1 + a_{11}h + o(h) \text{ and } p_{1j}(h) = a_{1j}h + o(h) \text{ si } j \neq 1,$$

with $a_{11} \leqslant 0$, $a_{1j} \geqslant 0$ for $j \neq 1$ and $\sum_j a_{1j} = 0$. In order to lighten notation, we write a_j instead of a_{1j}.

We have $P^{[X_t=1]}([X_{t+h} \neq 1]) = \left(\sum_{k\neq 1} a_k\right) h + o(h)$ and an individual "lives" an exponential time $\mathcal{E}(\sum_{k\neq 1} a_k)$ before transforming. We say that the individual is transformed at time time t, if at t we get 0 or a number $\geqslant 2$ of individuals. Then we consider that an individual produces at the end of his life, a number D of descendents, with D valued in $\mathbb{N}\setminus\{1\}$ ($D = 1$ is not taken into account since only the variations of the **size** of the population matter). The distribution of D is given by

$$\mathbb{P}([D = j]) = \frac{a_j}{\sum_{k\neq 1} a_k} \text{ pour } j \in \mathbb{N}\setminus\{1\}.$$

Then if $j \neq 1$, $P^{[X_t=n]}([X_{t+h} = n-1+j]) = na_j h + o(h)$ since the change can only be produced by one of the n individuals which is replaced by j new individuals.

Remark 3.8 *If* $a_0 = \mu$, $a_1 = -\lambda - \mu$, $a_2 = \lambda$ *and* $a_k = 0$ *pour* $k \geqslant 3$, *we are lead to a birth and death process with linear growth and linear diminution.*

Let $u(s) = \sum_{j=0}^{+\infty} a_j s^j$: u is sometimes called the *instantaneous probability generating function.* Then

Theorem 3.9 *u satisfies the following partial differential equations:*

$$\begin{cases} \partial_2 G(s,\tau) = u(G(s,\tau)) \\ G(s,0) = s \end{cases} \tag{3}$$

$$\begin{cases} \partial_2 G(s,t) = u(s)\partial_1(G(s,t)) \\ G(s,0) = s \end{cases}. \tag{4}$$

Proof By the key relation $G(s,t+\tau)=G(G(s,\tau),t)$ and a Taylor expansion

$$G(s,h)=\sum_{j=0}^{+\infty}s^j p_{1j}(h)=s+u(s)h+o(h).$$

Taking $t=h$ leads to $G(s,\tau+h)=G(G(s,\tau),h)=G(s,\tau)+u(G(s,\tau))h+o(h)$ and

$$\partial_2 G(s,\tau)=u(G(s,\tau)).$$

Taking $\tau=h$ leads to $G(s,t+h)=G(s+u(s)h+o(h),t)=G(s,t)+u(s)h\partial_1 G(s,t)+o(h)$ and

$$\partial_2 G(s,t)=u(s)\partial_1 G(s,t).$$

□

In particular,

Theorem 3.10 *The mean number of individuals at time t starting with 1 ancestor is*

$$\mathbb{E}^{[X_0=1]}(X_t)=e^{u'(1)t}.$$

Proof Deriving the first equation of (4) with respect to s, we get

$$\partial_1\partial_2 G(s,t)=u(s)\partial_1^2 G(s,t)+u'(s)\partial_1 G(s,t)$$

and with $s=1$ and using $u(1)=0$,

$$\partial_1\partial_2 G(s,t)=\partial_2\partial_1 G(1,t)=u'(1)\partial_1 G(1,t).$$

But $m(t)=\partial_1 G(1,t)$, then $m'(t)=\partial_2\partial_1 G(1,t)=u'(1)m(t)$, $m(t)=Ce^{u'(1)t}$ and since $m(0)=1$, we get $C=1$ that leads to the required result.

□

Interpretation

if $u'(1)<0$, then $m_t\to 0$ (this is the analog of $m<0$ of the previous subsection);
if $u'(1)=0$, then $m_t=1$ for any t;
if $u'(1)>0$, then $m_t\to+\infty$.

Extinction probability

Theorem 3.11 *Let $\pi_0=\lim_{t\to+\infty}\mathbb{P}([X_t=0])$. Then $\pi_0=1$ iif $u'(1)\leqslant 0$; π_0 is the smallest nonnegative solution of $u(s)=0$.*

Proof $\pi_0(t)=\mathbb{P}([X_t=0])=G(0,t)$ si $X_0=1$.
If $X_t=0$, then for $\tau\geqslant t$, $X_\tau=0$ and $\pi_0(t)\leqslant\pi_0(\tau)$ thus $t\mapsto\pi_0(t)$ is an increasing function upper bounded by 1; it admits a limit π_0 in $+\infty$.

$$\pi_0=\lim_{t\to+\infty}G(0,t)=\lim_{t\to+\infty}G(0,t+\tau)=\lim_{t\to+\infty}G(G(0,t),\tau)=G(\pi_0,\tau)$$

by continuity of G. Hence, $G(\pi_0,\tau)$ est independent of τ and $\partial_2 G(\pi_0,\tau)=0$.
By (4), we have $u(\pi_0)\partial_1 G(\pi_0,\tau)=0$ for any τ. In particular, for $\tau=0$, $G(s,0)=\mathbb{E}^{[X_0=1]}(s^{X_0})=s$, then $\partial_1 G(s,0)=1$and $u(\pi_0)=0$.
If s_0 is the smallest zero of u in $[0,1]$, then $G(s_0,\tau)$ is independent of τ, car $\partial_2 G(s_0,\tau)=0$. Then $G(s_0,\tau)=G(s_0,0)=s_0$.
Moreover, $s\mapsto G(s,t)=\sum_j s^j p_{1,j}(t)$ is increasing, hence $G(0,\tau)\leqslant G(s_0,\tau)=s_0$ for any τ. Taking the limit $\tau\to+\infty$, $\pi_0\leqslant s_0$, $\pi_0=s_0$ since $\pi_0\in[0,1]$ et $u(\pi_0)=0$.

□

3.4 Continuous processes with a finite number of types

In this subsection, we generalize both the passage from discrete time to continuous time and the passage of one type to two types. The notation still remain the same.
We denote $X_t^{(i)}$ the number of individuals of type (i) at time t, $e_1 = (1,0)$, $e_2 = (0,1)$,

$$p_{(k_1,k_2)(j_1,j_2)}(t) = P^{[X_u=(k_1,k_2)]}([X_{u+t} = (j_1,j_2)]) \text{ and } p_{e_i,(j_1,j_2)}(t) = p^{(i)}_{(j_1,j_2)}(t).$$

Assume that $p^{(i)}_{(j_1,j_2)}(h) = \delta_{e_i,(j_1,j_2)} + a^{(i)}_{(j_1,j_2)}h + o(h)$ with $a^{(i)}_{e_i} \leqslant 0$, $a^{(e_i)}_{(j_1,j_2)} \geqslant 0$ else, and $\sum\limits_{j_1,j_2} a^{(e_i)}_{(j_1,j_2)} = 0$, i.e. each individual of type (i) gives birth in $[t, t+h[$ to j_1 descendents of type (1) and j_2 descendent of type (2) with probability

$$p^{(i)}_{(j_1,j_2)}(h) = \delta_{e_i,(j_1,j_2)} + a^{(i)}_{(j_1,j_2)}h + o(h).$$

We define also $u^{(i)}(s_1,s_2) = \sum\limits_{j_1,j_2} s_1^{j_1}s_2^{j_2}a^{(i)}_{(j_1,j_2)}$ the instantaneous probability generating functions and $G^{(i)}(s_1,s_2,t) = \mathbb{E}^{[X_0=e_i]}\left(s_1^{X_t^{(1)}}s_2^{X_t^{(2)}}\right) = \sum\limits_{j_1,j_2} s_1^{j_1}s_2^{j_2}p^{(i)}_{(j_1,j_2)}(t)$. We then have

$$\mathbb{E}^{[X_0=(k_1,k_2)]}\left(s_1^{X_t^{(1)}}s_2^{X_t^{(2)}}\right) = \left(G^{(1)}(s_1,s_2,t)\right)^{k_1}\left(G^{(2)}(s_1,s_2,t)\right)^{k_2} = \sum_{j_1,j_2} s_1^{j_1}s_2^{j_2}p_{(k_1,k_2)(j_1,j_2)}(t).$$

Since (X_t) is Markovian, we still have a key relation:

Theorem 3.12 $G^{(i)}(s_1,s_2,t+\tau) = G^{(i)}\left(G^{(1)}(s_1,s_2,\tau), G^{(2)}(s_1,s_2,\tau), t\right)$.

from which we deduce (as done in the setting with one type) the following partial differential equations:

$$\begin{cases} \partial_3 G^{(i)}(s_1,s_2,\tau) = u^{(i)}\left(G^{(1)}(s_1,s_2,\tau), G^{(2)}(s_1,s_2,\tau)\right) \\ G^{(i)}(s_1,s_2,0) = s_i \end{cases} \tag{5}$$

$$\begin{cases} \partial_3 G^{(i)}(s_1,s_2,t) = u^{(1)}(s_1,s_2)\partial_1 G^{(i)}(s_1,s_2,t) + u^{(2)}(s_1,s_2)\partial_2 G^{(i)}(s_1,s_2,t) \\ G^{(i)}(s_1,s_2,0) = s_i \end{cases}. \tag{6}$$

Example 3.2 *a) Branching process with immigration*

Here we consider $P^{[X_t=0]}([X_{t+h} = j]) = \delta_{0,j} + b_j h + o(h)$ *with* $b_0 \leqslant 0$ *and the others* $b_j \geqslant 0$. *More precisely, we denote*

- b_j *the immigration rate of* j *individuals;*
- a_j *the birth rate of* j *individuals.*

To solve that kind of problem, we consider two types of individuals: the real ones and the fictive ones. Naturally we are only interested in the size of the real population. The fictive persons constitute the population that sends the immigrants. We can consider that it is reduced to a single individual that can gives birth to a random number of descendents. We then denote

$$a^{(1)}_{j_1,j_2} = \begin{cases} 0 & \text{if } j_2 \neq 0 \\ a_{j_1} & \text{if } j_2 = 0 \end{cases} \qquad \text{and} \qquad a^{(2)}_{j_1,j_2} = \begin{cases} 0 & \text{if } j_2 \neq 1 \\ b_{j_1} & \text{if } j_2 = 1 \end{cases}.$$

If $u(s) = \sum\limits_j a_j s^j$ *and* $v(s) = \sum\limits_j b_j s^j$, *we deduce*

$$u^{(1)}(s_1,s_2) = \sum_{j_1,j_2} s_1^{j_1}s_2^{j_2}a^{(1)}_{j_1,j_2} = \sum_{j_1} s_1^{j_1}a_{j_1} = u(s_1);$$

$$u^{(2)}(s_1,s_2) = \sum_{j_1,j_2} s_1^{j_1}s_2^{j_2}a^{(2)}_{j_1,j_2} = \sum_{j_1} s_1^{j_1}s_2 b_{j_1} = s_2 v(s_1).$$

b) Replacement of one particle with $\gamma(\lambda, 2)$ lifetime distribution by 2 particles

This example allows to solve a particular problem of continuous processes for which the lifetime of each individual is not exponential (the process is not Markovian in that case). To fix it, we introduce 2 independent and successive phases in the life of each particle, each phase being exponentially distributed $\mathcal{E}(\lambda)$.

Here $X_t^{(i)}$ represents the number of particles in phase i at time t. Obviously we are interested in $X_t^{(1)} + X_t^{(2)}$. To simplify we take $\lambda = 1$. Then

$$a^{(1)}_{j_1,j_2} = \begin{cases} -1 & \text{if } (j_1, j_2) = (1, 0) \\ 1 & \text{if } (j_1, j_2) = (0, 1) \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad a^{(2)}_{j_1,j_2} = \begin{cases} -1 & \text{if } (j_1, j_2) = (0, 1) \\ 1 & \text{if } (j_1, j_2) = (2, 0) \\ 0 & \text{else} \end{cases} .$$

We deduce $u^{(1)}(s_1, s_2) = s_2 - s_1$ and $u^{(2)}(s_1, s_2) = s_1^2 - s_2$.

3.5 Exercises

Exercise 3.1 *A culture of blood starts at time 0 with 1 red blood cell. After one minute, a red blood cell dies and is replaced, with the following probabilities, by*

- *2 red blood cells with probability 1/4;*
- *1 red and 1 white with probability 2/3;*
- *2 white blood cells with probability 1/12.*

Every blood cell lives during one minute and gives birth in the same way that its parent. Every white blood cell lives during one minute and dies without reproducing itself.

a) Evaluate the probability that no white blood cell still appeared at time $n + 1/2$ minute.

b) Evaluate the probability that the entire culture disappears.

Exercise 3.2 *A disease is modelized by a branching process with initial size N germs. At every grip of a medicine (1 a day), every germ has the probability $p = \frac{1}{2}$ to disappear. Determine the law of the duration T of the disease (or of the number of used medicine).*
Same question, when every germ lives an exponential time of average $\frac{1}{\lambda} = 2$ days.
Determine also, for $N = 3$, in every case, the mean duration of the disease.

Exercise 3.3 *We consider a population of bacteria of size X_t at time t such that $X_0 = 1$. Between t and $t + \Delta t$, every bacteria is divided in two new bacteria with probability $\lambda\Delta t + o(\Delta t)$, dies with probability $\mu\Delta t + o(\Delta t)$ where $\lambda \neq \mu$ and does not evolve with probability $1 - (\lambda + \mu)\Delta t + o(\Delta t)$.*

a) Let $G(s,t) = \mathbb{E}\left(s^{X_t}\right)$ the probability generating function of X_t. Determine a partial differential equation satisfied by satisfied by G and check that the unique solution such that $G(s,0) = s$ is

$$G(s,t) = \frac{e^{\alpha t}(1-s) - 1 + s\rho}{\rho e^{\alpha t}(1-s) - 1 + s\rho}$$

where $\rho = \frac{\lambda}{\mu}$ and $\alpha = \lambda - \mu$. Determine $\mathbb{E}(X_t)$, $p_0(t)$ and the extinction probability of the bacteria.

b) When $\mu = 0$ compare $\mathbb{E}(X_t)$ with the size of the process such that every bacterium divides every λ^{-1} units of time.

c) Determine $\mathbb{E}(X_n)$ and the extinction probability for the discrete process such that at every unit of time, a bacterium divides in two with probability $\frac{\lambda}{\lambda+\mu}$ and dies with probability $\frac{\mu}{\lambda+\mu}$.

Exercise 3.4 *We consider a population such that the number of direct descendents by individual is distributed as a binomial $\mathcal{B}(2,p)$.*

a) Assume we start with 1 individual, determine the extinction probability and the probability that there is nobody anymore, for the first time, at the third generation.

b) Assume now that number of individuals at the first generation is Poisson distributed with parameter λ. Prove that, for $p > \frac{1}{2}$, the extinction probability is $\pi = \exp[\lambda(1-2p)/p^2]$.

Exercise 3.5 *We consider a population of particles that undergo a shock every minute. Then the particle may divide in 2 (with probability p) or disappear (with probability $1-p$). We note X_n the population size after n minutes.*

a) Determine the extinction probability of the population if $X_0 = 1$ and then if $\mathbb{P}([X_0 = k]) = 1/2^k$ for any $k \in \mathbb{N}^*$.

b) We consider now that, independently for every particle, a shock occurs after an exponential time with mean 1mn. Determine the extinction probability.

c) Evaluate in every case the mean size of the population after the n-th minute.

Exercise 3.6 *We consider a population of males and females such that every female has one descendent after an exponential time with rate λ: this descendent is a female (resp. a male) with probability p (resp. $1-p$). The lifetimes of the females (resp. males) are exponential with rate μ (resp. ν).*
If X_t (resp. Y_t) represents the number of females (resp. males) at time t and if $(X_0, Y_0) = (i,j)$, check that

$$M_X(t) = \mathbb{E}(X_t) = ie^{(\lambda p - \mu)t} \text{ and } M_Y(t) = \mathbb{E}(Y_t) = \frac{i\lambda(1-p)}{\lambda p + \nu - \mu}e^{(\lambda p - \mu)t} + \left(j - \frac{i\lambda(1-p)}{\lambda p + \nu - \mu}\right)e^{-\nu t}.$$

4 Importance Splitting

An alternative way to IS is to use trajectory splitting based on a completely different idea than IS. Importance Splitting (ISp) is based on the idea that there exists some well identifiable intermediate system states that are visited much more often than the target states themselves and behave as gateway states to reach the rare event. In this model a more frequent occurrence of the rare event is achieved by performing a number of simulation retrials when the process enters regions where the chance of occurrence of the rare event is higher. In contrast to IS type algorithms, the step-by-step evolution of the system follows the original probability measure.

The principle of the algorithm is at first to run simultaneously several particles starting from the level B_i; after a while, some of them have evolved "badly", the other have "well" evolved i.e. have succeeded in reaching the threshold B_{i+1}. "Bad" particles are then moved to the position of the "good" ones and so on until A is reached. In such a way, the more promising particles are favored; unfortunately that algorithm is hard to analysis directly because of the interaction introduced between particles. Examples of this class of algorithms can be found in [2] with the " go with the winners" scheme, in [10, 15] in the context of the approximate counting and in [7, 9, 11] in a more general setting.

Nevertheless, in practice the trajectory splitting method may be difficult to apply. For example, the case of the estimation of the probability of a rare event in random dynamical systems is more complex, since the difficulty to find theoretically the optimal B_i. Furthermore, the probability to reach B_i varies generally with the state of entrance in level B_{i-1}. But it is not always the case e.g. for Markovian models (like diffusion).

A mathematical tool well adapted to study this type of algorithms is the Feynman-Kac approach developed in [9]. Asymptotic results are derived, such as LLN, CLT, and Large Deviations principles; in particular asymptotic variance of the estimator of the rare event probability is given. Non asymptotic results such as uniform Lp mean error bounds and exponential concentration inequalities with respect to the time horizon can be also found in this relevant book. Getting precise confidence intervals is more challenging. Nevertheless, all these algorithms lie on a common base, simpler to analyze and called branching splitting model. In this technique, interactions between particles are avoided and its relative simplicity allow us to derive precise results and to have better knowledge and understanding on splitting models in general.

We must precise here that we consider only one dimensional models as introduced in Garvels [12] or in a more refined version: the RESTART method [25, 26].

4.1 Importance Splitting model

The main hypothesis of splitting is that before entering the target event there exists intermediate states visited more frequently than A by the trajectory: thus define a sequence of sets of states B_i such as $A = B_{M+1} \subset B_M \subset \ldots \subset B_1$, which determines a partition of the state space into regions $B_i - B_{i+1}$ called *importance regions*.

In this model, a more frequent occurrence of the rare event is achieved by performing a number of simulation retrials when the process enters regions where the chance of occurrence of the rare event is higher. The fundamental idea consists in generating N Bernoulli $Ber(P_1)$ and check whether the subset B_1 is reached or not. If so, we duplicate the trials in R_1 retrials of Bernoulli $Ber(P_2)$ and check whether the subset B_2 is reached or not... This procedure is repeated at each level, until A is reached. If an event level is not reached, neither is A, then we stop the current retrial. Using N independent replications of this procedure, we have then considered $NR_1 \ldots R_M$ trials, taking into account for example, that if we have failed to reach a level B_i at the i-th step, the $R_i \ldots R_M$ possible retrials have failed. Clearly the particles reproduce and evolve independently.

By the Bayes formula,

$$\mathbb{P}(A) = \mathbb{P}(A|B_M)\mathbb{P}(B_M|B_{M-1}) \ldots \mathbb{P}(B_2|B_1)\mathbb{P}(B_1) := P_{M+1}P_M \ldots P_2P_1, \tag{7}$$

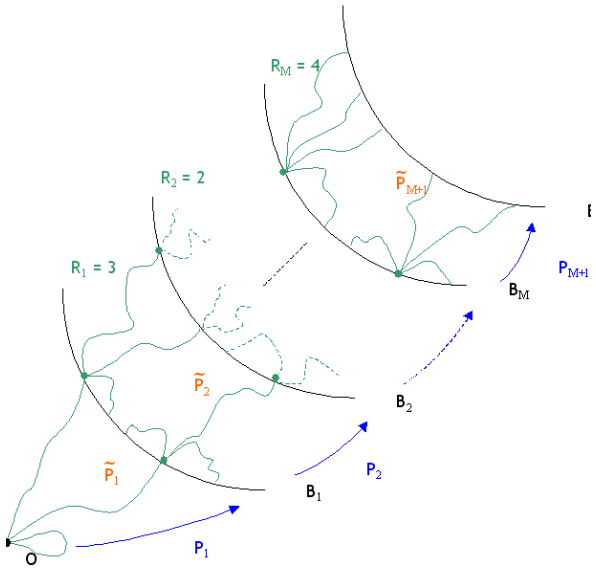

Figure 2: Splitting model

where on the right hand side, each conditioning event is "not rare" and $P_i = \mathbb{P}(B_i|B_{i-1})$. Then $\mathbb{P}(A)$ is the product of $M+1$ quantities (conditional probabilities) that are easier to estimate and with more accuracy than the probability P of the rare event itself, for a given simulation effort.

Example 4.1 *To analyze the behavior of the different implementations described above, we perform a simulation experiment using these methods. We consider a queuing network and we want to estimate the occupancy of finite buffer queuing system $M/M/1/C_0$. The results are presented in Figure 3.*
As expected and since we proceed for a given cost C ($C = 10^4$), crude simulation stops after a few iterations, the number of samples run at the beginning being not sufficient. However note that splitting simulation gives very close results to the theoretical analysis.

4.2 The estimator and its properties

A natural estimator of $\mathbb{P}(A)$ is given by the quantity

$$\widehat{P}_M = \frac{N_A}{N \prod_{i=1}^{M} R_i}, \tag{8}$$

where N_A is the total number of trajectories having reached the set A.

The estimator $\widehat{P}_M$ of $\mathbb{P}(A)$ defined in (8) can be rewritten as

$$\widehat{P}_M = \frac{1}{NR_1 \dots R_M} \sum_{i_0=1}^{N} \sum_{i_1=1}^{R_1} \dots \sum_{i_M=1}^{R_M} \mathbb{1}_{i_0} \mathbb{1}_{i_0 i_1} \dots \mathbb{1}_{i_0 i_1 \dots i_M}$$

where $\mathbb{1}_{i_0 i_1 \dots i_j}$ represents the result of the j-th trial.

4.2.1 Link with the Galton-Watson branching processes

This algorithm can be represented by N independent Galton-Watson branching processes. Indeed, consider N independent Galton-Watson branching processes $(Z_n^{(i)})_{n \geqslant 0}$, $i = 1, \dots, N$ where for any i, $Z_n^{(i)}$ is

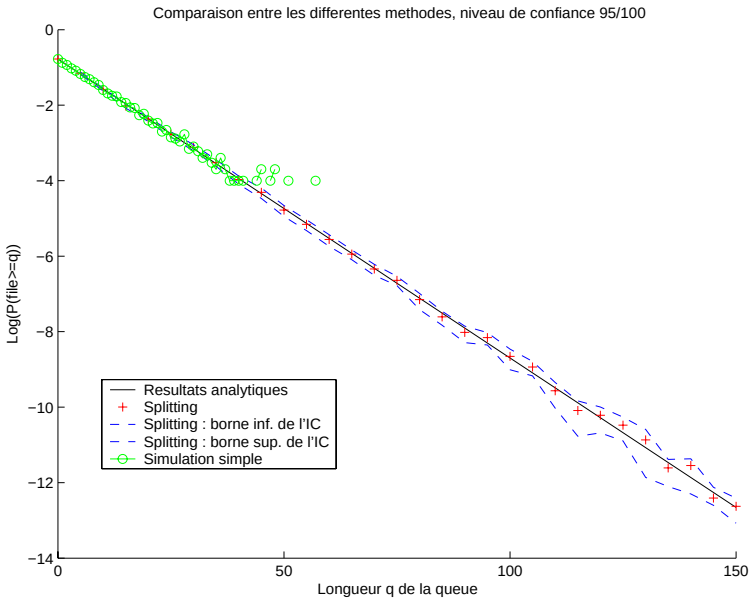

Figure 3: Comparison between the different methods-Queuing theory model

the number of particles, issued from the i-th particle ($Z_0^{(i)}$=1), that have reached threshold B_n. Then

$$\widehat{P} = \frac{1}{N}\sum_{i=1}^{N} \frac{Z_{M+1}^{(i)}}{R_1 \dots R_M}. \tag{9}$$

To lighten notation we consider $N = 1$ in the sequel and the process $(Z_n)_{n\geqslant 0}$ with $Z_0 = 1$. Then

$$Z_{n+1} = \sum_{j=1}^{Z_n} X_n^{(j)} \tag{10}$$

where for each n, the rvs $(X_n^{(j)})_{j\geqslant 1}$ are iid and binomial distributed with parameters (R_n, P_{n+1}) for $n \geqslant 1$ and Bernoulli distributed with parameter P_1 for $n = 0$.

Notice that Z_{i+1} can be written as

$$Z_{i+1} = \sum_{j=1}^{Z_i}\sum_{k=1}^{R_i} Y_i^{(j,k)}$$

where the rvs $Y_i^{(j,k)}$ are distributed as Bernoulli with parameter P_{i+1}.

4.2.2 Bias and variance of the estimator

First,

Proposition 4.1 *The estimator $\widehat{P}_M$ is unbiased.*

Proof Since

$$\mathbb{E}(\widehat{P}_M) = \mathbb{E}\left(\frac{N_A}{NR_1 \dots R_M}\right) = \frac{1}{NR_1 \dots R_M}\sum_{i_0=1}^{N}\sum_{i_1=1}^{R_1}\dots\sum_{i_M=1}^{R_M} \mathbb{E}(\mathbb{1}_{i_0}\mathbb{1}_{i_0 i_1}\dots\mathbb{1}_{i_0 i_1 \dots i_M}) = \mathbb{P}(A).$$

□

Remark 4.2 *Notice that*

$$\mathbb{E}(Z_n) = P_n R_{n-1}\mathbb{E}(Z_{n-1}).$$

Thus if $P_n R_{n-1}$ for any $n \geqslant 1$, the average number of particles at each threshold is constant and given by N the number of initial particles.

Second, as done in [25], the variance of the estimator $\widehat{P}_M$ is derived by induction and

Proposition 4.3 *The variance of $\widehat{P}_M$ is given by*

$$\mathrm{Var}(\widehat{P}_M) = \frac{\mathbb{P}(A)^2}{N}\left[\sum_{i=0}^{M}\frac{1}{r_i}\left(\frac{1}{P_{i+1|0}}-\frac{1}{P_{i|0}}\right)\right] \tag{11}$$

where $P_{i|0}$ represents the probability to reach B_i i.e. $P_{i|0} = P_1 \dots P_i$.

Proof Clearly the formula holds in straightforward simulation i.e. when $k = 0$, since $\widehat{P}_M$ is a normalized sum of iid Bernoulli variables with parameter $\mathbb{P}(A)$.

To go from k to $k+1$, assume that the variance of the estimator $\widehat{P}_M$ is derived by induction and the variance for k thresholds is given by

$$\mathrm{Var}(\widehat{P}_k) = \frac{(P_1 \dots P_{k+1})^2}{N}\left[\sum_{i=0}^{k}\frac{1}{r_i}\left(\frac{1}{P_{i+1|0}}-\frac{1}{P_{i|0}}\right)\right] \tag{12}$$

where $\widehat{P}_k$ represents the estimator of $\mathbb{P}(A)$ in a simulation with k thresholds. We want to prove that this formula holds for $k+1$ thresholds.

First of all note that for all X and Y random variables which are independent given the set B and X $\sigma(B)$-measurable we have

$$\mathrm{Var}(XY) = \mathrm{Var}(X)\mathrm{Var}(Y) + \mathrm{Var}(X)\mathbb{E}(Y)^2 + \mathrm{Var}(Y)\mathbb{E}(X)^2 \tag{13}$$

Now let

$$X_{i_0} = \mathbb{1}_{i_0}, \qquad Z_{i_0} = \frac{1}{R_1 \dots R_{k+1}}\sum_{i_1=1}^{R_1}\dots\sum_{i_{k+1}=1}^{R_{k+1}}\mathbb{1}_{i_0 i_1}\dots\mathbb{1}_{i_0 i_1 \dots i_{k+1}}$$

The random variables X_{i_0} are iid with common law $Ber(P_1)$ and conditionally to the event B_1, X_{i_0} and Z_{i_0} are independent. Note that each Z_{i_0} is the estimator of $\mathbb{P}(A)$ in a model with k thresholds, T_2 to T_{k+1} for the trajectory issued from the success of X_{i_0}. Thus

$$\mathbb{E}(Z) = P_2 \dots P_{k+2}$$

and by the induction's hypothesis,

$$\mathrm{Var}(Z) = (P_2 \dots P_{k+2})^2\left[\sum_{i=1}^{k+1}\frac{1}{R_1 \dots R_i}\left(\frac{1}{P_{i+1|1}}-\frac{1}{P_{i|1}}\right)\right].$$

So applying (13) with $X \sim Ber(P_1)$ and $Z \sim Z_{i_0}$, we have

$$\begin{aligned}\mathrm{Var}(\widehat{P}_M^{(k+1)}) &:= \frac{1}{N^2}\mathrm{Var}\left(\sum_{i_0=1}^{N}X_{i_0}Z_{i_0}\right) = \frac{P_1}{N}\left[\mathrm{Var}(Z) + (1-P_1)\mathbb{E}(Z)^2\right] \\ &= \frac{(P_1 P_2 \dots P_{k+2})^2}{N}\left[\sum_{i=0}^{k+1}\frac{1}{r_i}\left(\frac{1}{P_{i+1|0}}-\frac{1}{P_{i|0}}\right)\right]\end{aligned}$$

that concludes the proof. □

Remark 4.4 *The induction principle has a concrete interpretation: if in a simulation with M steps, the retrials generated in the first level are not taken into account except one that we call* main trial*, we have a simulation with $M-1$ steps.*

4.3 Cost of the algorithm

As said in the introduction, the aim is to minimize the variance for a fixed budget, giving optimal values for N,R_1,...,R_M, $P_1,\dots,P_{M+1}$ and M. Therefore, we have to describe the cost of a given simulation: each time a particle is launched, it generates an average cost function h. We assume that

- the cost h for a particle to reach B_i starting from B_{i-1} depends only on P_i (and not on the starting level),
- h is decreasing in x (which means that the smaller the transition probability is, the harder the transition is and the higher is the cost),
- h is non-negative,
- h converges to a constant (in general small) when x converges to 1.

The (average) cost is then

$$C_M = \mathbb{E}(Nh(P_1) + R_1N_1h(P_2) + R_2N_2h(P_3) + \ldots + R_MN_Mh(P_{M+1})) \tag{14}$$

where N_i is the number of trials that have reached threshold i. Finally,

Proposition 4.5 *The average cost of the algorithm is given by*

$$C_M = N\sum_{i=0}^{M} r_i h(P_{i+1})P_{i|0}. \tag{15}$$

Example 4.2 *We want to study the model of the simple random walk on $\mathbb{Z}$ starting from 0 that we kill as soon as it reaches the level -1 or k (success if we reach k, failure otherwise).*
So let X_n such that $X_0 = 0$ and $X_n = \sum_{i=1}^n Y_n$ where $\{Y_n\}$ is a sequence of random variables valued in $\{-1,1\}$ with $\mathbb{P}(Y_n = 1) = \mathbb{P}(Y_n = -1) = \frac{1}{2}$ and define $T_k = \inf\{n \geqslant 0 \ : \ X_n = -1 \text{ or } k\}$.
One can easily check that X_n and $X_n^2 - n$ are martingales. By the Doob's stopping theorem, $\mathbb{E}(X_{T_k}) = 0$ and $\mathbb{E}(X_{T_k}^2) = \mathbb{E}(T_k)$ which yields to

$$p := \mathbb{P}(X_{T_k} = k) = \frac{1}{k+1} \quad \text{and} \quad \mathbb{E}(T_k) = k = \frac{1}{p} - 1$$

i.e. the cost needed to reach the next level is $\frac{1}{p} - 1$ if p is the success probability.

4.4 Algorithm optimization

We now proceed to the optimization of the algorithm. To minimize the variance of $\widehat{P}_M$, the optimal values are derived in three steps:

1. The optimal values of $N, R_1, \dots, R_M$ are derived when we consider that $P_1, \dots, P_{M+1}$ are constant (i.e. the thresholds B_i are fixed).
2. Replacing these optimal values in the variance, we derive the optimal transition probabilities: $P_1, \dots, P_{M+1}$.
3. Replacing these optimal values in the variance, we derive M the optimal number of thresholds.

Optimal values for $N, R_1, \dots, R_M$. Using the method of Lagrange multipliers, we get

$$R_i = \frac{r_i}{r_{i-1}} = \sqrt{\frac{h(P_i)}{h(P_{i+1})}}\sqrt{\frac{1}{P_iP_{i+1}}}\sqrt{\frac{1-P_{i+1}}{1-P_i}} \quad i = 1,\dots,M \tag{16}$$

$$N = \frac{1}{\sqrt{h(P_1)}}\frac{C_M\sqrt{1/P_1 - 1}}{\sum_{i=1}^{M+1}\sqrt{h(P_i)}\sqrt{\frac{1}{P_i} - 1}}. \tag{17}$$

Optimal values for $P_1, \ldots, P_{M+1}$. Thus the variance becomes

$$\mathrm{Var}(\widehat{P}_M) = \frac{\mathbb{P}(A)^2}{C_M} \left[\sum_{i=1}^{M+1} \sqrt{h(P_i)} \sqrt{\frac{1}{P_i} - 1} \right]^2 .$$

Proceeding as previously under the constraint $\mathbb{P}(A) = P_1 \ldots P_{M+1}$, we obtain that all the P_i's satisfy $2\sqrt{C_M}\lambda \sqrt{h(x)(\frac{1}{x}-1)} = h'(x)(1-x) - \frac{h(x)}{x}$. If we assume that there exists a unique solution to this equation, we have $P_i = g(\lambda)$, hence $\mathbb{P}(A) = g(\lambda)^{M+1}$ and $g(\lambda) = \mathbb{P}(A)^{\frac{1}{M+1}}$. Finally

$$P_i = \mathbb{P}(A)^{\frac{1}{M+1}} \quad i = 1, \ldots, M+1. \tag{18}$$

Optimal value for M. The optimal values for $P_1, \ldots, P_{M+1}$ imply that the optimal R_i becomes $1/P_i$, $i = 1, \ldots, M$ and thus

$$\mathrm{Var}(\widehat{P}_M) = \frac{\mathbb{P}(A)^2}{C_M}(M+1)^2 h(\mathbb{P}(A)^{1/M+1})(\mathbb{P}(A)^{-1/M+1} - 1)$$

that we want to minimize in M. Remark that $R_i P_i = 1$. Let

$$f(M) = \frac{\mathbb{P}(A)^2}{C_M}(M+1)^2 h(\mathbb{P}(A)^{1/M+1})(\mathbb{P}(A)^{-1/M+1} - 1),$$

whose derivative cancels in

$$F(y) := (2(1-e^y) + y)h(e^y) - y(1-e^y)e^y h'(e^y) = 0, \text{ with } y = \frac{\ln \mathbb{P}(A)}{M+1}. \tag{19}$$

Thus the optimal number of thresholds is given by $\frac{\ln \mathbb{P}(A)}{y_0}$ where y_0 solves $F(y) = 0$. In practice we will take $M = \left[\frac{\ln \mathbb{P}(A)}{y_0}\right] - 1$ or $M = \left[\frac{\ln \mathbb{P}(A)}{y_0}\right]$ to get an integer number of thresholds.

Example 4.3 *For $h = 1$, we have to solve $y = 2(e^y - 1)$. We get $y_1 = 0$ and $y_2 \approx -1.5936$. y_2 is a minimum and the optimal value of M is*

$$M = [-0.6275 \ln \mathbb{P}(A)] - 1 \quad or \quad [-0.6275 \ln \mathbb{P}(A)]. \tag{20}$$

In this case, the variance is given by

$$var(\widetilde{P}) = \frac{\mathbb{P}(A)^2}{N}(M+1)\left(\frac{1}{P_0} - 1\right)$$

which corresponds to the asymptotic variance in the particle algorithm (see Part II of the course).

Note that M increases while $\mathbb{P}(A)$ decreases and with this value of M, each R_i and P_i become

$$R_i \approx 5 \text{ and } P_i \approx \frac{1}{5}. \tag{21}$$

Thus the optimal sampling number and the optimal transition probabilities are independent of the rare event probability.

Remark 4.6 *The solution $y = 0$ corresponds to the following optimal values*

$$M = \infty, \quad P_i = 1, \quad R_i = 1, \quad N \sim_{M\to\infty} \frac{C}{(M+1)h(1) + \ln(P)h'(1)}$$

But $P_i = 1$ implies that $P = 1$ and $R_i = 1$ means that we just perform a crude simulation.

Remark 4.7 (Link with the Galton-Watson branching processes) *Notice first that the optimal values $(R_i)_i$: $R_i = \frac{1}{P_0} := R$ and $(P_i)_i$: $P_i = P_0$ lead to*

$$R_i P_{i+1} = 1.$$

This result is not surprising since it means that the branching processes are critical Galton-Watson processes ($m = 1$). In other words, optimal values are chosen in such a way to balance the loss of variance from too little splitting and the exponential growth in computational effort from too much splitting.

4.5 Numerical applications and practical issues

Application 4.1 (Knapsack Problem)
In approximate counting, remind that the goal is to estimate the number of Knapsack solutions i.e. the cardinal of Ω *defined by*

$$\Omega := \{x \in \{0,1\}^n \ : \ \mathbf{a}.\mathbf{x} := \sum_{i=1}^{n} a_i x_i \leqslant b\}$$

for given positive real vector $\mathbf{a} = (a_i)_{i=1}^n$ *and real number* b*. We might try to apply the Markov Chain Monte-Carlo method (MCMC) [23]: construct a Markov chain* $\mathcal{M}_{Knap}$ *with state space* $\Omega = \{x \in \{0,1\}^n : \mathbf{a}.\mathbf{x} \leqslant b\}$ *and transitions from each state* $x = (x_1, \ldots, x_n) \in \Omega$ *defined by*

- *with probability* $\frac{1}{2}$ *let* $y = x$*; otherwise*
- *select* i *uniformly at random in* $\{1, \ldots, n\}$ *and let* $y' = (x_1, \ldots, x_{i-1}, 1 - x_i, x_{i+1}, \ldots, x_n)$
- *if* $a.y' \leqslant b$ *then let* $y = y'$ *else let* $y = x$

the new state is y*. This random walk on the hypercube truncated by the hyperplane* $\mathbf{a}.\mathbf{x} = b$ *converges to the uniform distribution over* Ω*. This suggests a procedure for selecting Knapsack solutions almost uniformly at random. Starting in state* $(0, \ldots, 0)$*, simulate* $\mathcal{M}_{Knap}$ *for sufficiently many steps that the distribution over states is "close"*[1] *to uniform, then return the current state. Of course sampling over* Ω *is not the same as estimating the size of* Ω*. But the first task leads to the second.*
Keep on taking the vector $\mathbf{a}$ *fixed but allow* b *to vary. Note* $\Omega(b)$ *and* $\mathcal{M}_{Knap}(b)$ *instead of* Ω *and* $\mathcal{M}_{Knap}$ *to emphasize on the dependence on* b*. Assume without loss of generality that* $a_1 \leqslant \ldots \leqslant a_n$ *and define* $b_1 = 0$ *and* $b_i = \min\{b, \sum_{i=1}^{i-1} a_j\}$*. One can check that*

$$|\Omega(b_{i-1})| \leqslant |\Omega(b_i)| \leqslant (n+1)|\Omega(b_{i-1})|.$$

Now write

$$|\Omega(b)| \ = \ |\Omega(b_{n+1})| = \frac{|\Omega(b_{n+1})|}{|\Omega(b_n)|} \frac{|\Omega(b_n)|}{|\Omega(b_{n-1})|} \cdots \frac{|\Omega(b_2)|}{|\Omega(b_1)|} |\Omega(b_1)| := \rho_n^{-1} \ldots \rho_1^{-1}.$$

The ratio $\rho_i = \frac{|\Omega(b_i)|}{|\Omega(b_{i+1})|}$ *may be estimated by sampling almost uniformly from* $\Omega(b_{i+1})$ *using the Markov chain* $\mathcal{M}_{Knap}(b_{i+1})$ *and computing the fraction of the samples that lie within* $\Omega(b_i)$*.*
Now take $a = [1, 2, 3, 4]$*,* $b = 3$*,* $h = 1$*,* $R = 5$ *and* $C = 2600$*. We chose the levels as proposed: first define* $b_1 = 0$*,* $b_2 = 1$*,* $b_3 = 3$*,* $b_4 = 3$ *and* $b_5 = b$*, secondly* $B_0 = \Omega$*,* $B_1 = \Omega(b_4)$*,* $B_2 = \Omega(b_3)$*,* $B_3 = \Omega(b_2)$ *and* $B_4 = \Omega(b_1)$*. Thus here* $M = n - 1$*,* $N = C/n$ *and* $n_{step} = 1020$*. Obviously Card*$(\Omega) = 5$*. We run 3 different simulations: the first suggested in [15] consisting in estimating the* n *ratios independently, the crude and splitting ones. We obtain different estimations for Card*(Ω)*:*

`exact value`	:	5
`estimation by crude simulation`	:	4.088
`estimation by the` n `ratios ind.`	:	5.44
`estimation by splitting simulation`	:	5.019

Even though the levels are not optimal, splitting carries out an improvement.

Application 4.2 (Tree and simulated annealing)
Here the state space is a rooted tree and the particles move from the root to the sheets of the tree. For every node v *of the tree, to each vertex* v_i *(*$i = 1 \ldots m$*) issued from* v *is associated a probability* $p(v_i|v)$ *(*$i = 1 \ldots m$*); a particle starting at* v *choose his following step according to these probabilities. The goal of the algorithm is to find a "deep" node in the tree i.e. a node at depth at least* d*. If we truncate the tree*

[1] The problem is to bound the number of steps necessary to make the Markov chain $\mathcal{M}_{Knap}(b)$ "close" to stationarity. More precisely, we need a bound of the *mixing time*:

$$\tau_{mix}(\nu) := \min\{t : \Delta_x(t') \leqslant \nu \text{ for all } t' \geqslant t\}$$

where $\Delta_x(t) = \max_{S \subset \Omega} |P^t(x, S) - \Pi(S)|$ and Π the stationary distribution. In [15], it is shown that $\mathcal{O}(n^{9/2+\nu})$ steps suffice.

at depth d, the problem amounts to evaluate the probability the the deepest node found by the algorithm is at the maximal depth d. Obviously the transition probabilities are unknown $p(v_i|v)$; nevertheless, from any node v of the tree, we know how to choose a vertex starting at v according to the probabilities $p(v_i|v)$ $(i = 1 \dots m)$.

Let us describe now the algorithm: we start with B initial particle at the root of the tree. At time i, there is a random number of particles at maximal depth i. If all the particles are in leaves, we stop the algorithm. Otherwise, we duplicate each particle that is not in a leaf in $R_i - 1$ new particles. We then continue the algorithm by choosing randomly a vertex for each of the current particles.

The cost f the simulation is given by

$$C = N \left[\sum_{i_1=1}^{r} p(v_{1i_1}|v_1) + R_1 \sum_{i_1=1}^{r} \sum_{i_2=1}^{r} p(v_{1i_1i_2}|v_{1i_1}) p(v_{1i_1}|v_1) \right.$$
$$\left. + \ldots + R_1 \ldots R_{d-1} \sum_{i_1=1}^{r} \ldots \sum_{i_d=1}^{r} p(v_{1i_1 \ldots i_d}|v_{1i_1 \ldots i_{d-1}}) \ldots p(v_{1i_1}|v_1) \right].$$

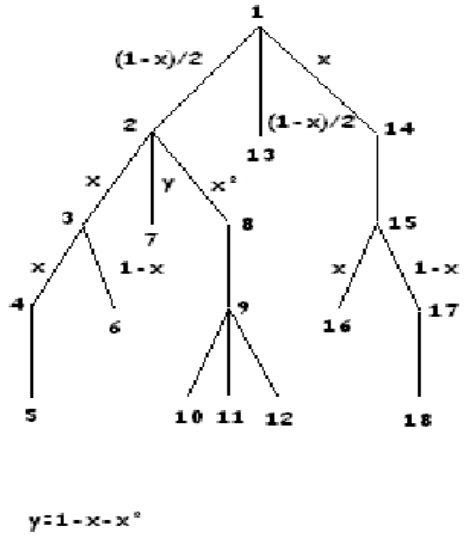

Figure 4: Tree considered in the example

Let $x \in [0, 1]$. To illustrate this algorithm, we consider the tree, represented in Figure 4, having 18 nodes such that

- *every node has in most three vertices,*
- *the probabilities that a particle at node i goes to one of the vertices from v is given by the i-th line of the matrix P given by*

$$P^t = \begin{bmatrix} (1-x)/2 & x & x & 1 & 0 & 0 & 0 & 1 & 1/3 & 0 & 0 & 0 & 0 & 1 & 1-x & 0 & 1 & 0 \\ (1-x)/2 & 1-x-x^2 & 1-x & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & x & 0 & 0 & 0 \\ x & x^2 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & x & 0 & 0 & 0 \end{bmatrix}$$

One can easily evaluate analytically the probability of interest $\mathbb{P}(A)$ to reach a leaf at maximal depth:

$$\mathbb{P}(A) = x^2(2 - x).$$

Taking for example $x = 10^{-3}$, the event A becomes rare and

$$\mathbb{P}(A) = (2 - 10^{-3})10^{-6} \approx 2\ 10^{-6}.$$

Simulation gives for equivalent costs

	exact value	:	$2\ 10^{-6}$
crude simulation	*estimated* 95%*-CI length*	:	$7\ 10^{-6}$
	estimated value	:	$2.5\ 10^{-6}$
	error	:	$5\ 10^{-7}$
splitting simulation	*estimated* 95%*-CI length*	:	$3\ 10^{-6}$
	estimated value	:	$2\ 10^{-6}$
	error	:	10^{-9}

The following discussion (simplified and approximate) highlights the interest of such a tree exploration. Let f a given function defined on S. There exists a natural way to define a tree such that his leaves correspond to local minima of f. Each leave of the tree at depth h corresponds to a connected component of $\{s\ :\ f(s) \leqslant h\}$. See Figure 5.

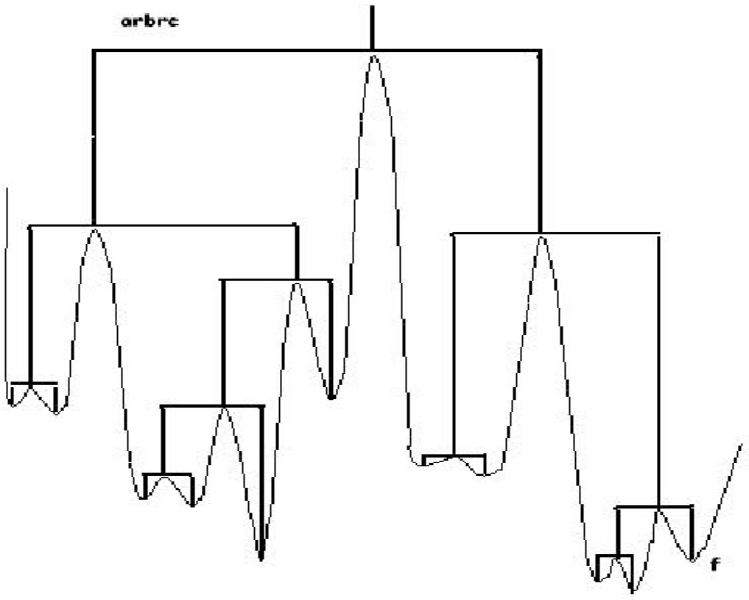

Figure 5: Correspondence between tree and simulated annealing

If a Metropolis algorithm is used to simulate the distribution

$$\pi_T(s) = c_T \exp\{-f(s)/T\},$$

we get the distribution π_T restricted to the connected component $\{s\ :\ f(s) \leqslant h(T)\}$ that contains the starting point for a given function $h(T)$. We identify this distribution to the corresponding point in the tree at depth $h(T)$. Hence one can see the progression of the simulated annealing algorithm with a decreasing temperature as a particle that evolves slowly to the back of the tree making randomly irreversible choices and stopping at a leaf.

Practical issues

1. One can also consider a variant implementation: instead of creating for every particle that have reached a new threshold a fixed number of offspring, one can create at each threshold a fixed total number of offspring. This is the strategy adopted in the particles algorithm (see [9]).

2. In practice, the adjustment of P_i close to the optimal value may be done during a first phase. The proportion of the cost devoted to this learning part is the topic of the paper [18].

3. But it soon appears that, even in the case of P_i's close to optimals, the fact that the number of replicas is not an integer destroys rapidly the accuracy of the algorithm: in such a case, one can take R_i equal to the closest integer (k or $k+1$) of the optimal value R but whatever the choice we have made, the criticality of the Galton-Watson process will be lost and the loss of precision is significant.
In [17], the author study different strategies to overcome this problem. Lead by [2], he chooses at random the sampling number with the hope of improving the simulation. In a first model (Random1), a Bernoulli rv R_i on $\{k, k+1\}$ for each particle having reached level i started from level i-1 is sampled. The parameter $p := \mathbb{P}(R_1 = k)$ is adjusted such that $m = 1$.
A second model (Random2) consists in sampling a random environmental sequence $(R_1, R_2, \ldots, R_M)$ of M iid Bernoulli random variables R_i on $\{k, k+1\}$ with common parameter p, derived by the same previous optimization approach with an additional constraint (the link between the expectations of R and its inverse). However, this problem is more complex and needs an approximate solution.
The author concludes that Random 1 provides the closest results from the optimals. Nevertheless the gain being not significant, in practice the strategy adopted is the one that choose R as the closest integer of the optimal value.

4. In dimension more than 1, the practitioner must be careful when defining the splitting surfaces. Indeed unlike in dimension 1, the optimal surfaces are not so natural.

4.6 Confidence intervals

Here, we assume that we take the optimal values of 4.4:

$$R_i = R \;\; i = 1, \ldots, M, \qquad P_i = P_0 = 1/R = \mathbb{P}(A)^{1/(M+1)} \;\; i = 1, \ldots, M+1.$$

The control of the variance of $\widehat{P}_M$ gives a crude confidence interval for $\mathbb{P}(A)$. Indeed, we get by Markov inequality

$$\begin{aligned}
\mathbb{P}\left(\frac{|\widehat{P}_M - \mathbb{P}(A)|}{\mathbb{P}(A)} \geqslant \alpha\right) &\leqslant \frac{1}{\mathbb{P}(A)^2\alpha^2}\mathbb{E}\left((\widehat{P}_M - \mathbb{E}(\widehat{P}_M))^2\right) \\
&\leqslant \frac{1}{\alpha^2 C_M}\left[(M+1)^2(\mathbb{P}(A)^{-1/M+1} - 1)h(\mathbb{P}(A)^{\frac{1}{M+1}})\right] \\
&\approx \frac{4(M+1)}{\alpha^2 N}h(\mathbb{P}(A)^{\frac{1}{M+1}})
\end{aligned}$$

which is in general useless. For example, for $h = 1$, $M = 12$ and $\alpha = 10^{-2}$, the upper bound becomes $\approx \frac{5\ 10^5}{N}$. To obtain a bound lower than 1, we need $N \geqslant 5\ 10^5$.

To improve it, we shall use Chernoff's bounding method that leads to

Proposition 4.8 *Let $\psi(\lambda)$ be the log-Laplace of W_1 and ψ^* be its Cramer transform:*

$$\psi(\lambda) = \mathbb{E}\left(e^{\lambda W_1}\right) \quad \textit{and} \quad \psi^*(\tau) = \sup_\lambda \left[\lambda\tau - \psi(\lambda)\right].$$

Thus

$$\begin{aligned}
\mathbb{P}\left(\frac{|\widehat{P}_M - \mathbb{P}(A)|}{\mathbb{P}(A)} \geqslant \alpha\right) &\leqslant e^{-N\psi^*(\mathbb{P}(A)(1-\alpha))} + e^{-N\psi^*(\mathbb{P}(A)(1+\alpha))} \qquad (22) \\
&\leqslant 2e^{-N\min(\psi^*(\mathbb{P}(A)(1-\alpha)),\psi^*(\mathbb{P}(A)(1-\alpha)))}. \qquad (23)
\end{aligned}$$

Proof For any $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}\left(\widehat{P}_M \geqslant \mathbb{P}(A)(1+\alpha)\right) &= \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^N W_i \geqslant \mathbb{P}(A)(1+\alpha)\right) \\
&= \mathbb{P}\left(e^{\lambda\sum_{i=1}^N W_i} \geqslant e^{\lambda N\mathbb{P}(A)(1+\alpha)}\right) \\
&\leqslant e^{-\lambda N\mathbb{P}(A)(1+\alpha)}\mathbb{E}(e^{\lambda W_1})^N \\
&\leqslant e^{-N[\lambda\mathbb{P}(A)(1+\alpha) - \psi(\lambda)]}
\end{aligned}$$

where $W_i = \frac{Z_{M+1}^{(i)}}{R_1 \dots R_M}$. Optimization on $\lambda > 0$ provides

$$\mathbb{P}\left(\widehat{P}_M \geqslant \mathbb{P}(A)(1+\alpha)\right) \leqslant e^{-N \sup_{\lambda>0}[\lambda \mathbb{P}(A)(1+\alpha)-\psi(\lambda)]}$$

and similarly

$$\mathbb{P}\left(\widehat{P}_M \leqslant \mathbb{P}(A)(1-\alpha)\right) \leqslant e^{-N \sup_{\lambda<0}[\lambda \mathbb{P}(A)(1-\alpha)-\psi(\lambda)]}.$$

□

So we are interested in accurate lower bounds of ψ^*.

Laplace transform of W_1

To study the Laplace transform of W_1, we turn to the theory of branching processes (see Section 3, Harris [13], Lyons [22] and Athreya and Ney [3]). More precisely we consider our splitting model as a Galton-Watson process, the thresholds representing the different generations. We straightforwardly have

$$f(s) = [P_0 s + (1 - P_0)]^R \quad \text{and} \quad \psi(\lambda) = \mathbb{E}(e^{\lambda W_1}) = g(f_M(e^{\lambda/R^M})).$$

The iterated function f_M has no explicit tractable form and we shall derive bounds for $f_M(s)$ around $s = 1$. To do this, we state a general result on the Laplace transform in critical Galton-Watson models, which we could not find in the literature.

Proposition 4.9 *Let* $\alpha_1 = \frac{f''(1)}{2} = \frac{1-P_0}{2}$.

(i) For s close to 1, $0 \leqslant s \leqslant 1$ and large n,

$$f_n(s) \leqslant 1 - \frac{(1-s)[1-\alpha_1(1-s)]}{1+\alpha_1(1-s)(n-1-\frac{\alpha_1^2(1-s)^2}{2})}. \tag{24}$$

(ii) For s close to 1 and $s \geqslant 1$ and large n,

$$f_n(s) \leqslant 1 + \frac{s-1}{1-n\alpha_1(s-1)}. \tag{25}$$

Proof (i) Using Taylor's expansion, with $f_n(s) \leqslant \theta_n \leqslant f_n(1) = 1$,

$$\begin{aligned} f_{n+1}(s) &= f(f_n(s)) = f(1) + (f_n(s)-1)f'(1) + \frac{(f_n(s)-1)^2}{2} f''(\theta_n) \\ &= f_n(s) + \frac{(f_n(s)-1)^2}{2} f''(\theta_n), \end{aligned}$$

since $f'(1) = 1$. Let $r_n = 1 - f_n(s)$, r_n satisfies

$$r_{n+1} = r_n - r_n^2 \frac{f''(\theta_n)}{2}.$$

Now let $\alpha_0 = \frac{f''(0)}{2}$. Define the decreasing sequences (a_n) and (b_n) satisfying

$$a_{n+1} = a_n - a_n^2 \alpha_1, \qquad b_{n+1} = b_n - b_n^2 \alpha_0, \qquad a_0 = b_0 = 1 - s.$$

Then

$$a_n \leqslant r_n \leqslant b_n. \tag{26}$$

1) b_n's upper bound: since $0 \leqslant b_j \leqslant 1$ we have

$$\frac{1}{b_n} = \frac{1}{b_{n-1}} + \alpha_0 \frac{1}{1-\alpha_0 b_{n-1}} = \frac{1}{b_0} + \alpha_0 \sum_{j=0}^{n-1} \frac{1}{1-\alpha_0 b_j} \geqslant \frac{1}{b_0} + n\alpha_0$$

Thus

$$b_n \leqslant \frac{1-s}{1+\alpha_0 n(1-s)}. \tag{27}$$

2) a_n's lower bound: apply (27) to a_n (α_0 becoming α_1).

$$a_n \leqslant \frac{1-s}{1+n\alpha_1(1-s)}. \tag{28}$$

By injecting (28) in $\frac{1}{a_n} = \frac{1}{a_0} + \alpha_1 \sum_{j=0}^{n-1} \frac{1}{1-\alpha_1 a_j}$, we get

$$a_n \geqslant \frac{(1-s)[1-\alpha_1(1-s)]}{1+\alpha_1(1-s)(n-1-\frac{\alpha_1^2(1-s)^2}{2})}. \tag{29}$$

Finally (26) and (29) lead to the upper bound of f_n in (24).

(ii) Let $h(s) = 1 + \frac{s-1}{1-n\alpha_1(s-1)}$. Since $f(1) = h(1) = 1$, $f'(1) = h'(1) = 1$ and $f''(1) = h''(1) = 2\alpha_1$, the sign of $f-h$ trivially depends on the sign of the third derivative of $f-h$ which is here obviously negative. Then $h \leqslant f$. Since f is increasing, we deduce (25) by induction. □

About the geometric distribution

If the law of X is such that the probabilities p_k are in a geometric proportion: $p_k = \mathbb{P}(X = k) = bc^{k-1}$ for $k = 1, 2 \ldots$ and $p_0 = 1 - p_1 - p_2 \ldots$ with $b, c > 0$ and $b \leqslant 1 - c$, then the associated g.f. is a rational function:

$$h(s) = 1 - \frac{b}{1-c} + \frac{bs}{1-cs}.$$

Taking $b = (1-c)^2$ and $c = \frac{\alpha_1}{1+\alpha_1}$ leads to

$$h(s) = 1 + \frac{s-1}{1-\alpha_1(s-1)}.$$

So we have compared the n-th functional iterate of a Binomial g.f. to the one of a geometric g.f. It suggests us to compare the importance splitting models with Binomial and with geometric laws. The second one is set in the following way: we run particles one after the other. As long as the next level is not reached we keep on generating particles, then we start again from it (the geometric distribution is the law of the first success).
This link is also underlined by Cosnard and Demangeot in [6]: for $m = 1$ and $\sigma^2 = f''(1) = 2\alpha_1$, the asymptotic behavior of f^{2^n} is the same as the one of a geometric with the same variance i.e. h.

Cramer transform of ψ

Considering the gradient of the functions, we prove that the supremum for $\lambda \geqslant 0$ is reached near 0 which justifies the use of the upper bounds for f_M obtained in Proposition 4.9. We then get

Proposition 4.10

$$\psi^*(\mathbb{P}(A)(1+\alpha)) \geqslant \min\{F(\mathbb{P}(A)(1+\alpha)), G(\mathbb{P}(A)(1-\alpha))\}$$

where

$$\begin{cases} F(x) & = \sup_{\lambda>0}[\lambda x - \ln(1 + P_0 \frac{(e^{\lambda/R^M}-1)}{1-M\alpha_1(e^{\lambda/R^M}-1)})] \\ G(x) & = \sup_{\lambda<0}[\lambda x - \ln(1 - P_0 \frac{(1-e^{\lambda/R^M})[1-\alpha_1(1-e^{\lambda/R^M})]}{u_0})]. \end{cases}$$

Finally

$$\mathbb{P}\left(\frac{|\widehat{P}-P|}{P} \geqslant \alpha\right) \leqslant 2e^{-N\min\{F(\mathbb{P}(A)(1+\alpha)), G(\mathbb{P}(A)(1-\alpha))\}}.$$

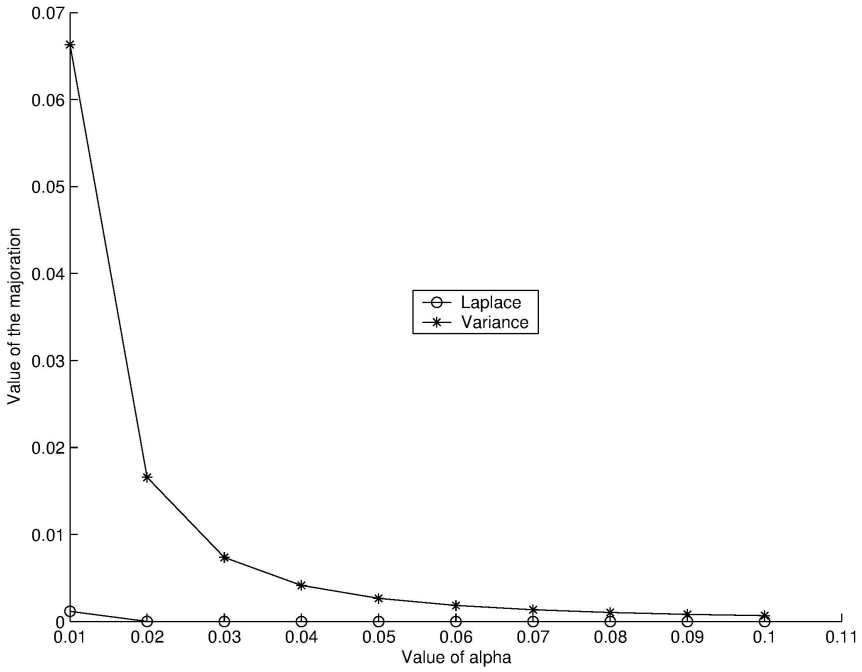

Figure 6: Upper bounds obtained by the variance and the Laplace transform

And one can easily obtain explicit but complex expressions for $F(x)$ and $G(x)$. We plot in Figure 6 the upper bounds obtained by the variance and by the Laplace transform, for different values of the prescribed error α of the CI. We take $\mathbb{P}(A) = 10^{-9}$ and the optimal values obtained above for the parameters. Note that the upper bound given by the Laplace transform is not surprisingly better than the other one (with the variance). We obtain $\mathbb{P}\left(\frac{|\widehat{P}_M - \mathbb{P}(A)}{\mathbb{P}(A)}| \geqslant \alpha\right) \leqslant L$. In the preceding example where $\mathbb{P}(A) = 10^{-9}$, if we fix $\alpha = 0.05$ and L close to 0.01, then the corresponding costs needed are $3\ 10^7$ for the variance and $3\ 10^6$ for the Laplace transform.

4.7 Exercises

Exercise 4.1 ***[An elementary gambler's ruin problem]*** *We consider a simple random walk* $X_n = x + \sum_{i=1}^n \epsilon_i$ *on* $E = \mathbb{Z}$*, starting at some* $x \in \mathbb{Z}$ *where* $(\epsilon_i)_{i \geqslant 1}$ *is a sequence of independent and identically distributed random variables with common law*

$$\mathbb{P}(\epsilon_1 = +1) = p \qquad \text{and} \qquad \mathbb{P}(\epsilon_1 = -1) = q$$

with $p,\ q \in (0,1)$ *and* $p + q = 1$*. If we use the convention* $\sum_\emptyset = 0$*, then we can interpret* X_n *as the amount of money won or lost by a player starting with* $x \in \mathbb{Z}$ *euros in a gambling game where he wins and loses 1 euro with respective probabilities* p *and* q*. If we let* $a < x < b$ *be two fixed parameters, one interesting question is to compute the probability that the player will succeed in winning* $b - x$ *euros, never losing more than* $x - a$ *euros. More formally this question becomes that of computing the probability that the chain* X_n *(starting at some* $x \in (a,b)$*) reaches the set* $B = [b, \infty)$ *before entering into the set* $C = (-\infty, a]$*. When* $p < q$ *(i.e.* $p < 1/2$*), the random walk* X_n *tends to move to the left and it becomes less and less likely that* X_n *will succeed in reaching the desired level* B*. We further assume that* $q > p$*. We introduce the stopping time*

$$R = \inf\{n \geqslant 0;\ X_n = a\}$$

as well as the first time the chain X_n *reaches one of the boundaries*

$$T = \inf\{n \geqslant 0;\ X_n \in \{a, b\}\} \leqslant R.$$

Study of $\mathbb{P}_x(R < \infty)$

- *Check that if we have* $|x - y| > n$ *or* $y - x \neq n + 2k$*, for some* $k \geqslant 1$ *then* $\mathbb{P}_x(R < \infty) = 0$*. The case where* $y - x = k - (n - k)$ *with* $0 \leqslant k \leqslant n$ *corresponds to situations where the chain has moved* k *steps to the right and* $n - k$ *steps to the left. Prove that* $\mathbb{P}_x(X_n = y) = \binom{n}{k} p^k q^{n-k}$*.*

- *Show that the function* α *defined by*

  $$x \in [a, \infty) \mapsto \alpha(x) = \mathbb{P}_x(R < \infty)$$

  is the minimal solution of the equation defined for any $x > a$ *by* $\alpha(x) = p\alpha(x+1) + q\alpha(x-1)$ *with the boundary condition* $\alpha(a) = 1$*.*

- *Whenever* $p < q$*, we recall that the general solution of the equation above has the form* $\alpha(x) = A + B(q/p)^x$ *with* $\alpha(a) = 1 = A + B(q/p)^a$*. Deduce from the above that*

  $$\alpha(x) = 1 + B[(q/p)^x - (q/p)^a] \quad \text{and} \quad \mathbb{P}_x(R < \infty) = 1 \quad \text{for any} \quad x.$$

- *Whenever* $p = q$*, we recall that the general solution of the equation above has the form* $\alpha(x) = Ax + B$ *with* $\alpha(a) = 1 = A + B(q/p)^a$*. Deduce from the above that*

  $$\alpha(x) = 1 + B[(q/p)^x - (q/p)^a] \quad \text{and} \quad \mathbb{P}_x(R < \infty) = 1 \quad \text{for any} \quad x.$$

Expectation of T

- *Check that for any* $n \geqslant 0$ *and* $\lambda > 0$*, we have*

  $$\mathbb{P}_x(R \geqslant n) = \mathbb{P}_x(X_n \geqslant a) \leqslant e^{-\lambda a} \mathbb{E}_x\left(e^{\lambda X_n}\right) = e^{-\lambda(a-x)} \left(pe^{\lambda} + qe^{-\lambda}\right)^n.$$

- *If we choose* $\lambda = \log(q/p)/2 \in (0, \infty)$*, then prove that*

  $$\mathbb{P}_x(R \geqslant n) \leqslant (p/q)^{(x-a)/2} (4pq)^{n/2}.$$

- *Deduce from the above that for* $p \neq 1/2$*,*

  $$\mathbb{E}_x(T) \leqslant \mathbb{E}_x(R) = \sum_{n \geqslant 1} \mathbb{P}_x(R \geqslant n) \leqslant \frac{(4pq)^{1/2}}{1 - (4pq)^{1/2}} (p/q)^{(x-a)/2}.$$

Study of $\mathbb{P}_x(T < R)$

- *Show that for any $a < x < b0$, the stochastic process $M_n = (q/p)^{X_n}$ is a $\mathbb{P}_x$-martingale with respect to the filtration $F_n = \sigma(X_0, \ldots, X_n)$ and if $p < q$, then $\mathbb{P}_x$-a.s. on the event $\{T \geqslant n\}$, we have that*

$$\mathbb{E}_x\left(|M_{n+1} - M_n||F_n\right) \leqslant 2(q/p)^b(q-p).$$

- *Since we have $\mathbb{E}_x(T) < \infty$ and $\mathbb{E}_x\left(|M_{n+1} - M_n||F_n\right)\mathbb{1}_{\{T \geqslant n\}} < c$ for some finite constant, prove by a well-known martingale theorem of Doob that $\mathbb{E}_x(M_T) = \mathbb{E}_x(M_0) = (q/p)^x$ and deduce that for any $x \in [a, b]$*

$$(q/p)^x = (q/p)^b\mathbb{P}_x(T < R) + (q/p)^a(1 - \mathbb{P}_x(T < R)).$$

Finally conclude that for any $p \neq q$, we have

$$\mathbb{P}_x(T < R) = \frac{(q/p)^x - (q/p)^a}{(q/p)^b - (q/p)^a}. \tag{30}$$

- *Using the strong Markov property, check that for any p and q, the function $\beta(x) = \mathbb{P}_x(T < R) = \mathbb{E}_x(\mathbb{1}_b(X_T))$ satisfies the equation*

$$\beta(x) = p\beta(x+1) + q\beta(x-1)$$

for any $x \in (a, b)$ with the boundary conditions $(\beta(a), \beta(b)) = (0, 1)$.
For $p \neq q$, check that the function (30) is the unique solution and for $p = q = 1/2$, prove that the solution is given for any $x \in [a, b]$ by

$$\mathbb{P}_x(T < R) = (x-a)/(b-a).$$

Splitting algorithm *Assume that we want to fix the intermediate thresholds B_n in such a way that the transition probability between two successive thresholds equals θ i.e.*

$$\mathbb{P}_{b_n}\left(X_{T_{n+1}} = b_{n+1}\right) = \theta,$$

where $T_n = \inf\{k \geqslant 0;\ X_k \in \{a, b_n\}\}$.

- *Show that the optimal solution is given recursively by*

$$b_{n+1} = b_n + \frac{\log\left(1 + (\theta - 1)\left(\frac{p}{q}\right)^{a-b_n}\right) - \log(\theta)}{\log(p/q)}.$$

- *Deduce that as b_n goes to infinity, if $q > p$,*

$$b_{n+1} \sim b_n - \frac{\log(\theta)}{\log(p/q)}.$$

5 Practical on Scilab

5.1 Illustrative examples

Crude Monte Carlo

Let us study Example 1.2. Propose an algorithm using the Monte Carlo scheme to evaluate

$$\mathbb{E}\left(e^{\beta G}\right)$$

with $\beta = 5$ and G a standard Gaussian rv.
Determine also a 95%-confidence interval.

Methods to reduce the variance

Let us study Example 2.4. We want to evaluate $I = \int_0^1 e^x dx$.
Propose algorithms using

1. crude Monte Carlo method
2. control variables method
3. antithetic variables method

to evaluate by several ways I.
For each method, determine also a 95%-confidence interval.

5.2 An example in finance

Let us study Example 1.3. Here we take $\beta = K = 1$.
Determine

1. crude Monte Carlo estimations of C and P;
2. an estimation of C based on control variables and the first estimation of P;
3. an estimation of P with IS method.
4. an estimation of P with antithetic variables method.

For each method, determine also a 95%-confidence interval.
Conclude.

5.3 An example in queuing theory

Let us study a M/M/1 queue. See next Section for some reminders on queuing theory.
Take for example, $\lambda = 0.1$ and $\mu = 0.12$. Determine

1. a crude Monte Carlo estimation of $\mathbb{P}(Q \geqslant L)$, $L = [\![1 : 5 : 150]\!]$;
2. a splitting estimation of $\mathbb{P}(Q \geqslant L)$, $L = [\![1 : 5 : 150]\!]$.

For each method, determine also a 95%-confidence interval.
Conclude.

5.4 Comparison between IS and Splitting on the simple random walk on $\mathbb{Z}$

In this section, we want to compare numerically on Scilab the IS and Splitting methods in the setting of the simple random walk on $\mathbb{Z}$. The goal is to estimate the probability that the line reaches length b before returning at 0.

General framework

Importance Sampling

Following exercise 2.2, we define a new random variable to simulate and the corresponding likelihood ratio.

Splitting

Following exercise 4.1, we define the optimal thresholds and run N simple random walks starting at 0. As soon as a queue reaches the next threshold before returning to 0, it is duplicated in R sub queues that evolve from this threshold and so on. The unbiased estimator of the probability under concern is then given by

$$\widehat{P}_{Splitt} = \frac{1}{N}\sum_{i_0=1}^{N}\frac{1}{R^M}\sum_{i_1=1}^{R}\cdots\sum_{i_M=1}^{R}\mathbb{1}_{i_0}\mathbb{1}_{i_0 i_1}\ldots\mathbb{1}_{i_0 i_1 \ldots i_M}$$

where $\mathbb{1}_{i_0 i_1 \ldots i_j}$ represents the result of j-th trial (i.e. it is equals to 1 if the queue reaches B_j, 0 esle).

Practical on Scilab

Write a program for both algorithms (IS and Splitting) to compare their performances (accuracy estimation, cost...) for the simple random walk.
Check also that crude simulation fails to propose an estimator in that case.

5.5 Comparison between IS and Splitting on the M/M/1 queue

In this section, we want to compare numerically on Scilab the IS and Splitting methods in queuing theory. The goal is to estimate the probability that the line reaches length L_0 before returning at 0.

General framework See [4] or [8] for more details.

A queue is constituted by

a) an **arrival flow** that represents the instants of arrival of "customers". We consider in general that the times between two successive arrivals are iid rvs. Then arrival flow is a stationary renewal process. A simple and commonly used case is the one with exponential inter arrivals; the process is then a Poisson process.

b) a **service** characterized by
* a service duration: a customer that starts his service will be immobilized a random duration with known distribution,
* a number of counter.

c) **service rules** that indicate how the service is proceeding:
* system with or without line (in a system without line, there is no queue; a customer that can not be served at his arrival is lost),
* service order: Fist In First Out (FIFO) (ex: line in the Post office), Last In First Out (LIFO) (ex: print line an the photocopier)
* several classes of customers clients (definition of priority customers)
* capacity of the queue
* at his arrival, if the line is too long, a customer may quit the line with a probability depending on the length of the queue and other parameters...

...

A queue is characterized by its Kendall notation

$$\text{A/B/C/}\dots$$

A represents the arrival flow, B the service time, C the number of counters. Then we add complementary information like policies... We use the following convention:
* M (like Markov) corresponds to a Poisson flow for the arrivals and to an exponential time service.
* D (like deterministic) corresponds to constant inter arrival times and to a fix time service for every customer.
* G (like general) corresponds to general distributions.

M/M/1 queue

This is the simplest and most studied queue.
* The arrivals correspond to a Poisson process with rate λ (the inter arrival times are iid rvs with parameter λ.
* The service time of the customers is exponentially distributed with parameter μ.
* There is a unique counter and the customers are served according to their order of arrival. There is no capacity limitation.

Let N_t be the number of customers in the queue at time t. N_t is an homogenous integer Markov process.

Proposition 5.1 *We have*
(i) $\mathbb{P}_x(N_t = x) = 1 - (\lambda + (x \wedge 1)\mu) + o(t)$;
(ii) the intensity of the process is given by

$$i(x) = \lambda + (x \wedge 1)\mu \qquad \textit{for} \;\; x \geqslant 0;$$

(iii) the transition matrix of the embedded chain is given by

$$\begin{cases} P(x, x+1) = \frac{\lambda}{\lambda+(x\wedge 1)\mu} \\ P(x, x-1) = \frac{(x\wedge 1)\mu}{\lambda+(x\wedge 1)\mu}. \end{cases}$$

The study of the transience of the process N_t amounts to that of the embedded chain. Let $\rho = \frac{\lambda}{\mu}$ the process intensity.

Proposition 5.2 *A positive measure invariant for P is given by*

$$m(x) = m(0)\frac{P(0,1)\dots P(x-1,x)}{P(1,0)\dots P(x,x-1)} = m(0)\rho^x \frac{\lambda + (x\wedge 1)\mu}{\lambda}.$$

Here we are interested by the case $\lambda < \mu$.The previous measure is then bounded and we get the existence of an invariant probability π given by

$$\pi(n) = \rho^n(1-\rho), \quad n \geqslant 0.$$

Proposition 5.3 *The performance parameters are given by*

- *the flow (arrival or departure) d is λ;*
- *the counter use rate is ρ;*
- *the average number L of customers in the system is*

$$L = \mathbb{E}_\pi(N_t) = \frac{\rho}{1-\rho};$$

- *the average number L_q of customers in the queue is*

$$L_q = \frac{\rho^2}{1-\rho};$$

- *the sojourn time in the system is*

$$W = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu} + \frac{\rho}{\mu(1-\rho)};$$

- *the sojourn time in the queue is*

$$W_q = \frac{\rho}{\mu(1-\rho)}.$$

Proof First, the arrival flow is clearly λ and $d = \lambda$. Second, if a customer enters the system with a queue of length n, its sojourn time T_q in the queue will be null if $n = 0$ and the sum of n iid exponential distributed rvs with parameter μ if $n > 0$. As a consequence,

$$\begin{aligned} \mathbb{P}(T_q \leqslant t) &= \sum_{n\geqslant 0} \mathbb{P}(T_q \leqslant t \text{ and } n \text{ customers in the queue}) \\ &= \sum_{n\geqslant 0} \mathbb{P}(T_q \leqslant t \mid n \text{ customers in the queue})\, \mathbb{P}(n \text{ customers in the queue}) \\ &= \pi(0) + \sum_{n\geqslant 1} \int_0^t \frac{\mu^n x^{n-1}}{n!} e^{-\mu x} dx \rho^n(1-\rho) \\ &= 1 - \rho + \rho(1 - e^{-\mu(1-\rho)t}) = 1 - \rho e^{-\mu(1-\rho)t} \end{aligned}$$

Then

$$W_q = \mathbb{E}(T_q) = \int_0^{+\infty} \mathbb{P}(T_q \geqslant t)dt = \frac{\rho}{\mu(1-\rho)}.$$

We then use the following relations

$$L = L_q + L_s \quad \text{and} \quad W = W_q + \frac{1}{\mu}$$

and the law of Little applied to the system, to the queue or to the counter

$$L = d\ W, \quad L_q = d\ W_q \quad \text{and} \quad L_s = \frac{1}{\mu}\ d.$$

□

Importance Sampling

From [14], the optimal change is given by

$$\begin{cases} \lambda^* = \mu \\ \mu^* = \lambda \end{cases}$$

We study N queues starting 1 according to the arrival and service rates λ^* and μ^*. The unbiased estimator of the probability under concern is then given by

$$\widehat{P}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{Y_i \geqslant L_0} L(Y_i)$$

Let

$$p_\lambda = \frac{\lambda^*}{\lambda^* + \mu^*} \frac{\lambda + \mu}{\lambda} \quad \text{and} \quad p_\mu = \frac{\mu^*}{\lambda^* + \mu^*} \frac{\lambda + \mu}{\mu}.$$

The likelihood ratio should be updated at each new event by

$$L = \begin{cases} L \times p_\lambda = L \times \frac{\lambda^*}{\lambda^* + \mu^*} \frac{\lambda + \mu}{\lambda} \\ L \times p_\mu = L \times \frac{\mu^*}{\lambda^* + \mu^*} \frac{\lambda + \mu}{\mu}. \end{cases}$$

Splitting

We define the optimal thresholds and run N queues starting at 1. As soon as a queue reaches the next threshold before returning to 0, it is duplicated in R sub queues that evolve from this threshold and so on. The unbiased estimator of the probability under concern is then given by

$$\widehat{P}_{Splitt} = \frac{1}{N} \sum_{i_0=1}^{N} \frac{1}{R^M} \sum_{i_1=1}^{R} \cdots \sum_{i_M=1}^{R} \mathbb{1}_{i_0} \mathbb{1}_{i_0 i_1} \cdots \mathbb{1}_{i_0 i_1 \ldots i_M}$$

where $\mathbb{1}_{i_0 i_1 \ldots i_j}$ represents the result of j-th trial (i.e. it is equals to 1 if the queue reaches B_j, 0 esle).

Practical on Scilab

Write a program for both algorithms (IS and Splitting) to compare their performances (accuracy estimation, cost...) on the M/M/1 queue.
For example, take $\lambda = 0.4$ and $\mu = 1$.
Check also that crude simulation fails to propose an estimator in that case.

References

[1] David Aldous. *Probability approximations via the Poisson clumping heuristic*, volume 77 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1989.

[2] Davis Aldous and Umesh V. Vazirani. "go with the winners" algorithms. *IEEE Symposium on Fundations of Computer Science*, (7):492–501, 1994.

[3] Krishna B. Athreya and Peter E. Ney. *Branching processes*. Springer-Verlag, New York, 1972. Die Grundlehren der mathematischen Wissenschaften, Band 196.

[4] Nicolas Bouleau. *Processus stochastiques et applications*, volume 1420 of *Actualités Scientifiques et Industrielles [Current Scientific and Industrial Topics]*. Hermann, Paris, 1988.

[5] William G. Cochran. *Sampling techniques*. John Wiley & Sons, New York-London-Sydney, third edition, 1977. Wiley Series in Probability and Mathematical Statistics.

[6] M. Cosnard and J. Demongeot. Théorèmes de point fixe et processus de Galton-Watson. *Ann. Sci. Math. Québec*, 8(1):5–21, 1984.

[7] Frédéric Cérou and Arnaud Guyader. Adaptive multilevel splitting for rare event analysis. *Rapport de recherche de l'INRIA - Rennes , Equipe : ASPI*. 2005.

[8] Peter Tjerk de Boer. Analysis and efficient simulation of queueing models of telecommunication systems. 2000.

[9] Pierre Del Moral. *Feynman-Kac formulae*. Probability and its Applications (New York). Springer-Verlag, New York, 2004. Genealogical and interacting particle systems with applications.

[10] Persi Diaconis and Susan Holmes. Three examples of Monte-Carlo Markov chains: at the interface between statistical computing, computer science, and statistical mechanics. In *Discrete probability and algorithms (Minneapolis, MN, 1993)*, volume 72 of *IMA Vol. Math. Appl.*, pages 43–56. Springer, New York, 1995.

[11] Arnaud Doucet, Nando de Freitas, and Neil Gordon. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pages 3–14. Springer, New York, 2001.

[12] Marnix J.J. Garvels. The splitting method in rare event simulation. *PhD thesis*. 2000.

[13] Theodore E. Harris. *The theory of branching processes*. Dover Phoenix Editions. Dover Publications Inc., Mineola, NY, 2002. Corrected reprint of the 1963 original [Springer, Berlin; MR **29** #664].

[14] P.E. Heegaard. Speed-up techniques for simulation. *Telektronikk ISSN 0085-7130*, 91(2-3):195–207, 1995.

[15] Mark Jerrum and Alistair Sinclair. *The Markov chain Monte Carlo method: an approach to approximate counting and integration*. Approximation Algorithms for NP-hard Problems. Dorit Hochbaum, 1997.

[16] Agnès Lagnoux. Rare event simulation. *Probab. Engrg. Inform. Sci.*, 20(1):45–66, 2006.

[17] Agnès Lagnoux-Renaudie. Effective branching splitting method under cost constraint. *Stochastic Process. Appl.*, 118(10):1820–1851, 2008.

[18] Agnès Lagnoux-Renaudie. A two-step branching splitting model under cost constraint for rare event analysis. *J. Appl. Probab.*, 46(2):429–452, 2009.

[19] Damien Lamberton and Bernard Lapeyre. *Introduction au calcul stochastique appliqué à la finance*. Ellipses, Édition Marketing, Paris, second edition, 1997.

[20] Bernard Lapeyre, Étienne Pardoux, and Rémi Sentis. *Méthodes de Monte-Carlo pour les équations de transport et de diffusion*. Mathématiques & Applications [Mathematics & Applications], 29. Springer-Verlag, Berlin, 1998.

[21] Pierre L'Ecuyer, François Le Gland, Pascal Lezaud, and Bruno Tuffin. Splitting techiques. In *Rare event simulation using Monte Carlo methods*, pages 39–61. Wiley, Chichester, 2009.

[22] Russell Lyons. Probability on trees and networks. *Preprint*. http://mypage.iu.edu/ rdlyons/. 2002.

[23] J. R. Norris. *Markov chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original.

[24] John S. Sadowsky. On Monte Carlo estimation of large deviations probabilities. *Ann. Appl. Probab.*, 6(2):399–422, 1996.

[25] Manuel Villén-Altamirano and José Villén-Altamirano. Restart: a method for accelerating rare event simulations. pages 71–76. North-Holland, 1991.

[26] Manuel Villén-Altamirano and José Villén-Altamirano. Restart: An efficient and general method for fast simulation of rare event. *Technical report 7*. 1997.

[27] Manuel Villén-Altamirano and José Villén-Altamirano. Analysis of restart simulation: Theoretical basis and sensitivity study. *European Transactions on Telecommunications*, 13 n°4:373–385, 2002.