On Computer-Intensive Simulation and Estimation Methods for Rare Event Analysis in Epidemic Models

Stéphan Clémençon, Anthony Cousien, Miraine Dávila Felipe, Viet Chi Tran

On Computer-Intensive Simulation and Estimation Methods for Rare Event Analysis in Epidemic Models

S. Clémençon[*] A. Cousien[†] M. Dávila Felipe[‡] V.C. Tran[§]

August 27, 2013

Abstract

This article focuses, in the context of epidemic models, on *rare events* that may possibly correspond to crisis situations from the perspective of Public Health. In general, no close analytic form for their occurrence probabilities is available and crude Monte-Carlo procedures fail. We show how recent intensive computer simulation techniques, such as *interacting branching particle methods*, can be used for estimation purposes, as well as for generating model paths that correspond to realizations of such events. Applications of these simulation-based methods to several epidemic models are also considered and discussed thoroughly.

Keywords: Stochastic epidemic model ; rare event analysis ; Monte-Carlo simulation ; importance sampling ; interacting branching particle system ; genetic models ; multilevel splitting

AMS Codes: MSC 65C35 ; 62G32 ; MSC 92D30

1 Introduction

Since the seminal contribution of [20, 6], the mathematical issues raised by the modelling and statistical analysis of the spread of communicable infectious diseases have never ceased to receive attention in the applied probability and statistics communities. Given the great diversity of situations encountered in practice (impact of demographic phenomena, presence of control strategies, endemicity, population heterogeneity, time-varying infectivity, *etc.*), a wide variety of stochastic epidemic models have been introduced in the literature, striving to incorporate more and more relevant features in order to account for real-life situations, while remaining analytically tractable. The study of the properties of the related stochastic processes (branching approximations, long-term behavior, large population asymptotics, *etc.*) and the design of efficient inference methods tailored for (generally partially observed) epidemic data are still stimulating research on mathematical epidemiology. Beyond considerations of purely academic nature, many notions and techniques developed in this field are important for practitioners. Epidemic models are used to understand and control infectious diseases and their theoretical analysis sheds some light on how to come up with figures such as the reproduction number R_0 of the epidemics (when well-defined). From a public health guidance perspective, they can be deployed in order to simulate the likeliest scenarios or compute the probability of certain events of interest, and plan control measures to stanch a disease outbreak in real-time. However, in most situations, no close analytical form is available for these probabilities and the latter are related to events that occur very rarely, for which Crude Monte-Carlo (CMC) estimation fails.

It is the main purpose of this paper to review possible techniques for rare event simulation and inference in the context of epidemic models. Motivated by practical issues in Public Health, we are concerned here with critical events such as an exceedingly long duration for an epidemic, an extremely large total number of positive diagnoses (*i.e.* large final size of the epidemic) in non endemic cases, the occurrence of a severe outbreak at a short horizon, *etc.* Here we list a number of events that may correspond to crisis situations and express the latter as excesses of a (very large) threshold by a random variable or a (randomly stopped) stochastic process for a general class of SIR epidemic models. *Importance Sampling* and *Particle Filtering* methods are next adapted to tackle the problem of estimating the occurrence probabilities of these events, as well as that of simulating realizations of the latter. Beyond the description of the methodological aspects, application of these techniques for analyzing a collection of rare events related to several numerical epidemic models, some of them being fitted from real data, is also discussed.

[*]Institut Telecom LTCI UMR Telecom ParisTech/CNRS No. 5141

[†]Inserm ATIP-Avenir: "Modélisation, Aide à la Décision, et Coût-Efficacité en Maladies Infectieuses", Lille, France

[‡]Université Pierre et Marie Curie LPMA UMR CNRS No. 7599

[§]Labo P.Painlevé UFR de Mathématiques UMR CNRS 8524, Université des Sciences et Technologies Lille 1

The article is structured as follows. Section 2 introduces a general class of epidemic models, to which the simulation/estimation techniques subsequently described apply and next review events related to these models, that may correspond to health crisis situations and generally occur very rarely. Simulation-based procedures for estimating the probability of occurrence of these events are described in Section 3, while practical applications of these techniques, based on real data sets in some cases, are considered in Section 4 for illustration purpose. Some concluding remarks are finally collected in Section 5. In this work, it is shown that crude Monte-Carlo method often fail to provide good estimates of rare events. Importance sampling methods are a well-known alternative to estimate the occurrence probabilities of rare events. However, their efficiency relies on the choice of proper instrumental distributions, which is very complicated for most probabilistic models encountered in practice. Particle systems with genealogical selection offer an efficient computationally-based tool for estimating the targeted small probabilities.

2 Background

It is the goal of this section to introduce a general class of epidemic models to which the computer-intensive estimation techniques described in the subsequent section apply. The (rare) events that shall be next statistically analyzed are formulated in terms of path properties of stochastic processes.

2.1 Epidemic models

The vast majority of (stochastic) epidemic models considered in the literature are of the *compartmental* type. They assume that the population of interest is divided into several strata or compartments, corresponding in particular to the various possible serological statuses, and stipulate a probabilistic framework that describes the transitions from one compartment to another.

The Reed-Frost model. One of the simplest epidemic models is the discrete-time chain-binomial model, generally referred to as the Reed-Frost model, that describes the spread of an infectious disease in a homogeneous and homogeneously mixing population. New infectious are assumed to occur in generations, $t = 0,\ 1,\ \ldots$ and immunity is gained by the infectives of generation t at generation $t+1$. Denoting by S_t and I_t the numbers of individuals at the t-th generation who are *susceptible* and *infective* respectively, and by $1-q$ the probability that an infective transmits the disease to a given susceptible at any generation (infections being assumed to occur independently from each other), the sequence $\{(S_t, I_t)\}_{t\in\mathbb{N}}$ with initial state $(s_0, i_0) \in \mathbb{N}^{*2}$ is a Markov chain with transitions as follows: for all $t \in \mathbb{N}$, (s_t, i_t) in $\mathbb{N}^2$ and i_{t+1} in $\{0,\ 1,\ \ldots,\ s_t\}$,

$$\mathbb{P}\left\{I_{t+1} = i_{t+1} \mid (S_t, I_t) = (s_t, i_t)\right\} = \left(\begin{array}{c} s_t \\ i_{t+1} \end{array} \right) (1 - q^{i_t})^{i_{t+1}} (q^{i_t})^{s_t - i_{t+1}} \tag{1}$$

and

$$S_{t+1} = S_t - I_{t+1}. \tag{2}$$

The set $\mathbb{N} \times \{0\}$ is *absorbing* for the Markov chain (S_t, I_t), meaning that the epidemics ceases as soon as the chain reaches this set (and then stays there forever), one may refer to [25] for an account of the Markov chain theory.

The standard stochastic SIR model. The most basic continuous-time stochastic epidemic model, generally referred to as the standard (Markovian) SIR model in a closed population of size n (see the seminal contribution of [6] for instance), counts three compartments: the *susceptible class* S, the *infective class* I and the *removed/recovered class* R. This corresponds to the situation where the epidemic is of short duration, making acceptable the assumption of a closed population, and the disease provides immunity against a possible re-infection. Fig. 1 below depicts the diagram flow of this simple epidemic model (taking $\mu = \rho \equiv 0$). For clarity, we index the events E through which the sizes $S(t)$, $I(t)$ and $R(t)$ of the three compartments that form the population evolve temporarily: we write $E = 1$ when the event that occurs is an infection, $E = 2$ when it corresponds to the removal of an infective. Taking by convention $T_0 = 0$ as time origin, the (continuous-time) dynamics of the model stipulates that all durations in competition are independent, infections and removals occur at time $t \geq 0$ with the rates $\lambda(S(t), I(t)) = \lambda S(t)I(t)/n$ and $\gamma(I(t)) = \gamma I(t)$, where $(\lambda, \gamma) \in \mathbb{R}_+^*$, respectively. Hence, the process $Z = \{(S(t), I(t), R(t))\}_{t\geq 0}$ evolves in a Markovian fashion, by jumps at random times $0 < T_1 < T_2 < \ldots$, when events $E_1,\ E_2,\ \ldots$ in $\{1,\ 2\}$ successively occur. The dynamics can be described by stochastic differential equations driven by Poisson point measures.

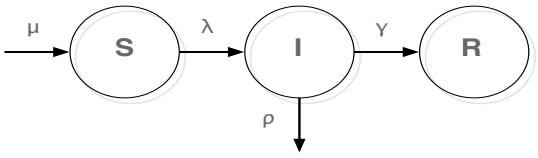

Figure 1: Diagram flow of a basic SIR stochastic model with demography.

Variants of the standard SIR model. When the epidemic under study acts on a large temporal scale, it may be necessary to incorporate additional features in the model (*cf* rates μ and ρ featured in Fig. 1) accounting for the demography of the population over which the disease spreads in an endemic manner. The number and the nature of the compartments involved in the epidemic models may also vary, depending on the infectious disease considered. For instance, the SIRS model corresponds to the situation, where immunity is lost after some time, while some AIDS epidemic models count numerous compartments, in order to account for the (non exponentially distributed) AIDS incubation period (this approach is usually referred to as *stage modelling*, see [19]). Additionally, the possible heterogeneity of the population may lead to remove the assumption of *uniform mixingness* and consider instead *multitype epidemic models* (refer for instance to Chapter 6 in [2] for a review of SIR models where the population is segmented into a finite number of subcommunities) or a population *structured by continuous variables* (see [12] for such a measure-valued stochastic process and the references therein) or spreading on random graphs which represent the underlying social network structure of the population (e.g. [14, 28]). Indeed there are many variants of the model described above, much too numerous to be listed here exhaustively. For clarity, the problem of estimating the probability of rare events related to the spread of a transmittable disease shall be addressed in the context of simple or even simplistic models, where the epidemics is described by a discrete-time Markov chain or a jump Markov process, extensions to more general situations being straightforward in most cases.

2.2 Rare/dramatic events in infectious disease epidemics

In the management of epidemics of communicable infectious diseases, the following events and quantities are of particular interest to Public Health decision makers. Here and throughout, we set $\inf \emptyset = +\infty$ by convention. The event of interest is denoted by $\mathcal{E}$. We will see that pertinent events often take the form $\mathcal{E} = \{\tau_A \leq \mathcal{T}\}$ where A is a subset of the space $\mathbb{N}^3$ where the epidemics process Z takes its values and where $\tau_A = \inf\{t \geq 0 : \ Z(t) \in A\}$ and $\mathcal{T}$ are almost-surely finite stopping times. Hence, we are interested in level-crossing probabilities of the form:

$$\mathbb{P}\{\tau_A \leq \mathcal{T}\}. \tag{3}$$

- **Duration of the epidemics.** In non endemic situations, the epidemics starts at a time arbitrarily set to $t = 0$ and ends at a short term horizon, described by the (almost-surely finite) stopping time

$$\tau = \inf\{t \geq 0 : \ I(t) = 0\}.$$

  Sharply estimating the probability $p_d(T) = \mathbb{P}\{\tau > T\}$ that the epidemics lasts more than a (very long) period of time $[0, T]$, with $0 < T < +\infty$, is an essential concern from the Public Health perspective. The computation of $1 - p_d(T)$ correspond to (3) in the case where $\mathcal{T} = T$ and $A = \mathbb{N} \times \{0\} \times \mathbb{N}$.

- **The final size of the epidemics.** The final size of the epidemics corresponds to the total number of infected individuals between times 0 and τ it is thus defined as the random variable $R(\tau)$. The probability $p_f(N_c) = \mathbb{P}\{R(\tau) \geq N_c\}$ that the size $R(\tau)$ exceeds a (critical) threshold value $N_c \geq 1$ smaller than n in the case of a closed population of total size $n \geq 1$) is of vital interest to quantify the means to be put in place (quarantine measures, supply of medications, number of hospital beds, *etc.*). Considering the stopping time $\tau_{R,N_c} = \inf\{t \geq 0 : \ R(t) \geq N_c\}$, notice that one may write:

$$p_f(N_c) = \mathbb{P}\{\tau_{R,N_c} \leq \tau\}. \tag{4}$$

  $p_f(N_c)$ reduces to (3) with $\mathcal{T} = \tau$ and $A = \mathbb{N} \times \mathbb{N} \times \{N_c, \ N_c + 1, \ \ldots\}$.

- **The incidence of the epidemics.** In order to handle in real-time a crisis situation, it is relevant to consider *time-dependent* quantities such that the probability that the (non cumulative) number of infectious individuals reaches a critical value N_I at a certain time horizon $T < \infty$. Let $\tau_{I,N_I} = \inf\{t \geq 0 : \quad I(t) \geq N_I\}$ be the corresponding stopping time, the probability one seeks to estimate is then given by:

$$p_I(T, N_I) = \mathbb{P}\left\{\tau_{I,N_I} \leq T\right\}. \tag{5}$$

  The quantity $p_I(T, N_I)$ corresponds to (3) when $\mathcal{T} = T$ and $A = \mathbb{N} \times \{N_I, N_I + 1, \dots\} \times \mathbb{N}$.

Along these lines, since Public Health decision-makers often adjust their policies, depending on the number of recently diagnosed cases, one may also be interested in the following quantity, related to removed individuals (assuming by convention that, once detected, an infected individual is removed from the subpopulation of infectives): the probability that the number of cases diagnosed between times t and $t + u$ increases by more than a threshold value $N_R \geq 1$, that is given by $\mathbb{P}\{R(t + u) - R(t) \geq N_R\}$. Although many other rare events of this type, related to an excessively duration or an exceeding of a large threshold, are of potential interest, given the wide variety of epidemic models (echoing the great diversity of real situations), methods for simulating rare events and estimating their probability of occurrence shall be investigated here through the examples listed above in the context of basic SIR models for the sake of simplicity.

3 Simulation methods for rare event analysis

The use of Monte-Carlo simulation techniques is widespread in mathematical epidemiology, see [23] for instance. However, crude Monte-Carlo methods (CMC) completely fail when applied to rare events such as those listed in Section 2.2. We first provide in §3.1 two illustrations showing the limits of CMC. An alternative in rare event simulation is known as *Importance Sampling* (IS), presented in §3.2. Roughly speaking, it consists in simulating under a different probability distribution (refered to as the *instrumental distribution*, equivalent to the original probability measure along a certain filtration) under which the event of interest $\mathcal{E}$ is more frequent. However, in absence of large deviation results for the vast majority of stochastic SIR models in the literature, proper instrumental distributions are difficult to obtain. In §3.3, we present the IBS method. We describe the method and perform in Section 4 numeric experiments.

3.1 Illustrations of the numerical inadequacy of CMC for simulating rare events

We study numerically two examples to illustrate the low quality of CMC for estimating the probabilities of rare events.

First, let us consider the basic Markovian SIR model without demography (see §2.1). For this simple model, the distribution of the final size $R(\tau)$ is proved to be the unique solution of a triangular linear system (see Theorem 2.2 in [2] for instance, or [22] for exact results of the same type in a more general framework), making the exact computation of the quantity $p_f(N_c)$ feasible (neglecting numerical stability issues, occurring even for moderate values of the population size n), whatever the threshold $N_c \geq 1$. As shown by Fig. 2, for this particular example, the accuracy of CMC estimates of the probability $p_f(N_c)$ rapidly deteriorates when N_c takes very large values (close to the total size of the population), very few (or even no) realizations of the stochastic process achieving the event $\{R(\tau) \geq N_c\}$, leading to a significant underestimation of $p_f(N_c)$, in spite of a large number of Monte-Carlo replications. Additional comments can be found in Section 4, when discussing the results.

3.2 Importance sampling

A standard approach to rare event simulation is known as *Importance Sampling*, see [9] or [4]. The (unbiased) estimate of the probability of occurrence of the rare event is obtained by multiplying the empirical frequency of the simulations under the instrumental distribution by the likelihood ratio ϕ, referred to as the *importance function*. For instance, when considering the standard Markovian SIR model described in the preceding section, a natural way of accelerating the occurrence of the events listed in §2.2 is to speed up the infection process, while slowing down the removal (*i.e.* increasing the value of the parameter λ and decreasing that of the parameter γ). More precisely, let $\mathbb{P}$ be the probability measure under which the process $\{(S(t), I(t), R(t))\}_{t\geq 0}$ is a standard Markovian SIR model with parameters $(\lambda,\ \gamma) \in \mathbb{R}_+^{*2}$ and such that $(S(0), I(0)) = (s_0, i_0) \in \mathbb{N}^{*2}$. Let $\mathbb{P}_{\text{new}}$ correspond to the pair $(\lambda_{\text{new}},\ \gamma_{\text{new}}) \in \mathbb{R}_+^{*2}$, such that $\lambda_{\text{new}} \geq \lambda$ and $\gamma_{\text{new}} \leq \gamma$. Clearly, these probability measures are absolutely continuous with respect to each other along the canonical filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 0}$ (*i.e.* $\mathcal{F}_t$ is

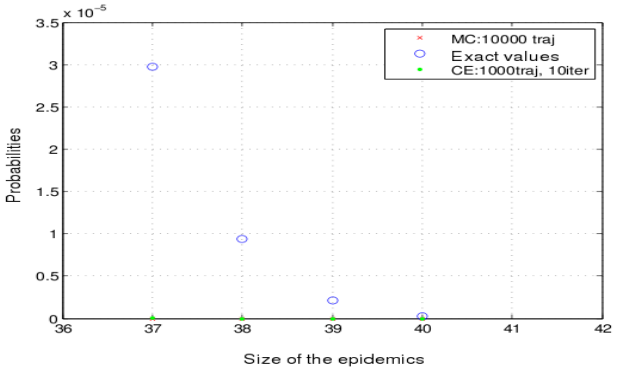

Figure 2: In a Markovian SIR model with $(s_0, i_0) = (40, 1)$ and parameters $\lambda = 1$ and $\gamma = 1$, crude Monte-Carlo estimate (based on 10 000 replicates of the epidemics process) of the probability $p_f(N_c)$ that the size of the epidemics takes a given value are plotted as a function of N_c. True values are also computed.

the σ-algebra generated by the collection of random variables $\{(S(u), I(u))\}_{u\in[0,t]}$ for all $t \geq 0$): on $\mathcal{F}_t$, the importance function (*i.e.* the likelihood ratio $d\mathbb{P}/d\mathbb{P}_{\text{new}}\,|_{\mathcal{F}_t}$) is given by:

$$\phi_t = \exp\left(-\int_0^t (\lambda - \lambda_{\text{new}})S(s)I(s)/n + (\gamma - \gamma_{\text{new}})I(s)ds\right)(\lambda/\lambda_{\text{new}})^{N(t)-R(t)}\,(\gamma/\gamma_{\text{new}})^{R(t)}\,,$$

where $N(t)$ denotes the number of events $E \in \{1,\ 2\}$ occurring between times 0 and t, and $T_{N(t)}$ is the last time when an event of this type occurs before time t. This extends to the situation where t is a $\mathcal{F}$-stopping time, such as the times of exceedance considered in §2.2. Hence, if $\mathcal{E} \in \mathcal{F}_t$, we have: $\mathbb{P}\{\mathcal{E}\} = \int \phi_t \cdot \mathbb{1}\{\mathcal{E}\}d\mathbb{P}^{\text{new}}$, denoting by $\mathbb{1}\{\mathcal{E}\}$ the indicator function of the event $\mathcal{E}$.

The success of IS crucially depends on the choice of the instrumental distribution (the specification of the instrumental parameters $(\lambda_{\text{new}},\ \gamma_{\text{new}})$ in the example previously mentioned). Ideally, it should be selected so as to reduce drastically the variance of the random variable $\phi_t \cdot \mathbb{1}\{\mathcal{E}\}$, otherwise the IS approach may completely fail. Optimal choice of probability changes can be based on large-deviation techniques, when the latter are tractable for the stochastic model considered (see Chapter 5 in [9] for further details). However, in absence of large deviation type results for the vast majority of the stochastic SIR models considered in the literature, one faces significant difficulties for selecting importance sampling estimators with small variance in practice. Recently, a number of refinements of the IS strategy have been proposed (*sequential Monte-Carlo methods* in particular), involving an iterative search of a nearly optimal instrumental distribution, see [17]. All these methods are said *intrusive*, insofar as their implementation requires to call for simulation routines related to modified versions of the distribution of interest.

Cross entropy method for IS. In the framework of estimating rare events, the *cross-entropy method* (CE) introduced in [26] can be used to modify iteratively the instrumental distribution for estimating the occurrence probability of $\mathcal{E}$, see [13, 8] or [1]. In the cases that are considered here, the law of the Markov processes depend on parameters: for instance q in the Reed-Frost model or (λ, μ) in the continuous time SIR model. Let us denote by ϕ the set of parameters and by $\mathcal{L}(Z, \phi)$ the likelihood of the path $Z = (S_t, I_t)_{t\in\mathbb{N}}$ in the Reed-Frost case or $Z = (S(t), I(t), R(t))$ in the continuous time SIR model. The idea is to choose as instrumental distribution the law $\mathcal{L}(, v)$ with the parameter v that minimises the entropy with respect to the original distribution (with parameter ϕ) conditioned on the rare event $\mathcal{E}$. We describe the algorithm in the discrete case. The methodology also applies to the standard continuous time Markovian SIR model when it comes to estimate the quantity (4). Indeed, considering the embedded Markov chain $Z = (S(T_k), I(T_k))_{k\in\mathbb{N}}$, where the T_k's denote the successive times when the epidemics process jumps, one may also write $p_f(N_c) = \mathbb{P}\{Z_{\tau_\Lambda} \in A\}$.

For clarity, we recall below the general principle of the CE method in the purpose of estimating the quantity $\theta = \mathbb{P}\{Z_{\tau_\Lambda} \in A\}$, the latter serving as a benchmark case in the experimental section, see §4.1. Here Z is a Markov chain started at z_0 and whose distribution is parameterized by ϕ and we denote by $\mathcal{L}(Z, \phi)$ its likelihood. As alternative adaptive IS methods have lead to very similar results in our experiments, they are not considered here (refer to [17]).

ADAPTIVE IMPORTANCE SAMPLING THROUGH THE CE METHOD

1. **Initialization.** Set $v^{(0)} = \phi$.

2. **Iterations.** For $k = 1, \ldots, K$,

   (a) Draw N sample paths starting from x_0 with the parameter $v^{(k-1)}$:

   $$Z^{(i)} = \left(z_0,\, Z_1^{(i)},\, \ldots,\, Z_{\tau_A^{(i)}}^{(i)} \right),\ \text{for } 1 \leq i \leq N.$$

   (b) Compute the IS estimate

   $$\widehat{\theta}_{k,N} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathcal{L}(Z^{(i)}, \phi)}{\mathcal{L}(Z^{(i)}, v^{(k-1)})} \cdot \mathbb{1}\left\{ Z_{\tau_A^{(i)}}^{(i)} \in A \right\},$$

   (c) Define the new parameter $v^{(k)}$ as the maximum in v of

   $$L(v) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left\{ Z_{\tau_A^{(i)}}^{(i)} \in A \right\} \frac{\mathcal{L}(Z^{(i)}, \phi)}{\mathcal{L}(Z^{(i)}, v^{(k-1)})} \ln \mathcal{L}(Z^{(i)}, v).$$

3. **Output.** Produce the estimate $\widehat{\theta}_{K,N}$ of the target probability.

3.3 Interacting and branching particle system methods

In contrast to the IS strategy and its variants, *Interacting Branching Particle System* methods (IBPS in abbreviated form) for rare event simulation are *non intrusive* in the sense that no modification of the code to run for simulating paths $Z = \{(S(t), I(t), R(t))\}_{t \geq 0}$ of the (epidemic) model under study is required. Roughly speaking, the IBPS principle as follows. We start with a population of N trajectories $Z^{(1)}, \ldots, Z^{(n)}$ (that we call *particles*) and modify the latter in an iterative manner: paths for which the event of interest $\mathcal{E}$ "almost occurs" (in a sense that shall be specified, depending on the nature of the event $\mathcal{E}$) are "multiplied", while the others are "killed", following in the footsteps of the celebrated ReSTART algorithm (for Repetitive Simulated Trials After Reaching Thresholds) originally introduced in the context of teletraffic data models, see [27].

So-termed *splitting techniques* (refer to [18]), thoroughly investigated in [16] (see also [11]), are fully tailored for estimating the rare event probability (3), as well as the conditional law of the epidemics process Z given the rare event of interest $\{\tau_A \leq \mathcal{T}\}$ is realized. The idea is to consider a sequence of increasing subsets of the state space, $A_0 \supset A_1 \supset A_{K+1} = A$, describing more and more difficult obstacles the process Z must pass over, before reaching the target set A. Consider the related hitting times, defined by the recurrence relation:

$$T_0 = \inf\{t \geq 0 :\ Z(t) \in A_0\} \ \text{ and } T_k = \inf\{t \geq T_{k-1} :\ Z(t) \in A_k\} \ \text{ for } k \geq 1.$$

We assume that $Z(0) \in A_0$ with probability one, so that $T_0 = 0$ almost-surely. Clearly, the rare event probability (3) factorizes the following manner,

$$\mathbb{P}\{T_{K+1} \leq \mathcal{T}\} = \mathbb{P}\{T_{K+1} \leq \mathcal{T} \mid T_K \leq \mathcal{T}\} \times \ldots \times \mathbb{P}\{T_1 \leq \mathcal{T} \mid T_0 \leq \mathcal{T}\}, \tag{6}$$

in a product of conditional probabilities of events (hopefully) much less rare and whose realizations can be more easily simulated. The technique described subsequently precisely permits to estimate each factor in (6) and build progressively epidemics paths realizing the rare event $\{\tau_A \leq \mathcal{T}\}$ as well.

In many situations, the A_k's are determined by a collection of increasing levels (the choice of the number K of intermediate levels and that of the levels themselves will be discussed later, see Remark 3.2). For instance, when it comes to estimate the probability $p_I(T, N_I)$ that the number of infectives exceeds a critical threshold value N_I before a certain time $T < \infty$, one may consider a sequence of sublevels $0 = N_I^{(0)} < \ldots < N_I^{(K+1)} = N_I$, that defines subsets $A_k = \mathbb{N} \times \{N_I^{(k)},\ N_I^{(k)} + 1,\ \ldots\} \times \mathbb{N}$ for $k = 0, \ldots, K+1$.

More precisely, the particle population model evolves according to the following genealogical structure, see [15]. At generation $k \in \{1, \ldots, K\}$, a particle Z having reached the k-th level before time $\mathcal{T}$ (i.e.

such that $T_k \leq \mathcal{T}$) are kept while the other are deleted (*selection* stage) and replaced by new particles (*mutation* stage), see Fig. 3. A new particle is a novel epidemics path Z^{new} whose path segment on $[0, T_k]$ coincides with that of a particle Z chosen randomly among the particles such that $T_k \leq \mathcal{T}$, and whose trajectory on $[T_k, \mathcal{T}]$ (or on $[T_k, T_{k+1}^{\text{new}}]$ from a practical perspective) is simply sampled from the distribution of the epidemic process when the initial condition is $Z(T_k)$. Of course, the algorithm stops (and is restarted) if no particle survives. Adaptive variants are described below. The *selection* stage is implemented by means of *weight functions* ω_k defined on the path space by $\omega_k(Z) = 1$ when $T_k \leq \mathcal{T}$ and by $\omega_k(Z) = 0$ otherwise. The method is then performed in k steps as follows.

A quite similar approach can be considered for the estimation of the probability $p_f(N_c)$ that the total size of the epidemics rises above a large threshold $N_c \geq 1$.

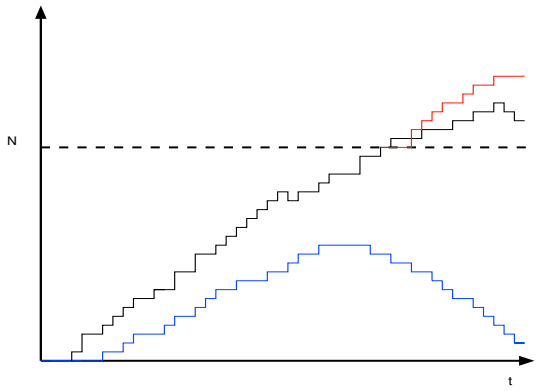

Figure 3: Multi-level splitting: the path in blue does not reach the current level N and is thus killed, while that in black does and can be selected in order to produce an *offspring*, generated by sampling from the time of exceedance (in red)

THE IBPS ALGORITHM

1. **Initialization.** Start with a collection of $N \geq 1$ simulated trajectories $Z_0^{(1)}, \ldots, Z_0^{(N)}$ of the epidemic process indexed by $i \in \{1, \ldots, N\}$, with the same initial condition $Z(0) = (s_0, i_0, 0)$, to which the weights $\omega_0^{(i)} = 1$, $1 \leq i \leq N$, are assigned. Denote by $T_0^{(i)} = 0 < T_1^{(i)} < \ldots < T_{K+1}^{(i)}$ and $\mathcal{T}^{(i)}$ the related stopping times.

2. **Iterations.** For $k = 1, \ldots, K$,

   (a) Let $\mathcal{I}_{1,k}$ be the subset of indices $i \in \{1, \ldots, N\}$ corresponding to the epidemics paths $Z_{k-1}^{(i)}$ having reached the subset A_k before time $\mathcal{T}^{(i)}$ and denote by $\#\mathcal{I}_{1,k}$ its cardinality (the algorithm is stopped and re-started if it is equal to 0). Set $\mathcal{I}_{0,k} = \{1, \ldots, N\} \backslash \mathcal{I}_{1,k}$. For each path indexed by $i \in \mathcal{I}_{1,k}$, set $Z_k^{(i)} = Z_{k-1}^{(i)}$. We also define P_k as the proportion of particles Z that have reached the subset A_k before time $\mathcal{T}$ among those which have previously reached A_{k-1}.

   (b) For each path indexed by $i \in \mathcal{I}_{0,k}$:

   - (SELECTION STEP) independently draw a particle $Z_k^{(j)}$ from distribution $\sum_{j=1}^N \omega_k^{(j)} \cdot \delta_{Z_k^{(j)}}$, with $\omega_k^{(j)} = \omega_k(Z_k^{(j)})/(\sum_{l=1}^N \omega_k(Z_k^{(l)}))$,
   - (MUTATION STEP) Define $Z_k^{(i)}$ as the path confounded with $Z_k^{(j)}$ until time $T_k^{(j)}$ and prolongate by simulation from the state $Z_k^{(j)}(T_k^{(j)})$.

   (c) Compute $P_j = \#\mathcal{I}_{1,k}/N$ and pass onto stage $k+1$.

3. **Output.** Compute the estimate of the target probability $\pi = \mathbb{P}\{\tau_A \leq \mathcal{T}\}$:
$$\widehat{\pi}_N = P_1 \times \ldots \times P_{K+1}.$$
Compute also the empirical distribution
$$\mathcal{L}_N = \frac{1}{N} \sum_{i=1}^N \delta_{Z_{K+1}^{(i)}},$$
which may serve as an estimate of the conditional law $\mathcal{L}$ of the epidemics process given the occurrence of $\{\tau_A \leq \mathcal{T}\}$.

Before showing how the IBPS performs on a variety of examples, a few remarks are in order.

Remark 3.1. (A MORE DETERMINISTIC GENETIC EVOLUTION SCHEME) It should be first underlined that alternative choices for the genealogical dynamics, different from that consisting in drawing uniformly among the surviving particles, could be possibly pertinent. As proposed in [11] (see subsection 3.2 therein), one may also consider a N-particle approximation model based on the following selection/mutation scheme: in a deterministic fashion, one keeps at each stage k all paths which have reached the k-th level, that is N_k particles say. Then the other $N - N_k$ particles are killed and replaced by a particle whose path segment on $[0, T_k]$ is chosen uniformly at random among the N_k "successfull" particles and completed by (independent) sampling on $[T_k, \mathcal{T}]$.

Remark 3.2. (TUNING PARAMETERS) Accuracy (consistency and asymptotic normality in particular) of the estimator $\widehat{\pi}_N$ produced by the IBPS algorithm has been established as the number of particles N increases to infinity in [11, 10]. However, the practical implementation requires to pick several parameters: the number of intermediate levels and the levels themselves. As explained in [21], they should be chosen, so that all factors in the product (6) are approximately of the same order of magnitude, and possibly in an adaptive way during the simulations. When applied to the problem of estimating $p_I(T, N_I)$ for instance, the adaptive variant of the multi-level splitting proposed in [10] would consist, at each step, in sorting all the simulated paths $Z^{(i)}$ by decreasing order of the quantity $\sup_{t \in [0,T]} I^{(i)}(t)$ and take the

k-th term as current intermediate level with fixed $k \in \{1, \ldots, N\}$ (hence killing at each step $N - k$ trajectories).

Remark 3.3. (PERSISTENCE OF THE EPIDEMICS) Observe also that the approach described above can be extended in order to estimate the probability that the epidemics lasts more than a (long) time $T > 0$, $p_d(T)$. Instead of stratifying the state space of the epidemics process Z (along the I- or R- axis), the idea is to write $p_D(T) = \mathbb{P}\{I(T) \geq 1\}$ and split the time axis by introducing successive durations $t_0 = 0 < t_1 < \ldots < t_{K+1} = T$ (see Fig. 4). The sequence of decreasing events is then defined by $\{I(t_k) \geq 1\}$ for $k = 0, \ldots, K+1$ and we have:

$$p_D(T) = \mathbb{P}\left\{I(t_{K+1}) \geq 1 \mid I(t_K) \geq 1\right\} \times \ldots \times \mathbb{P}\left\{I(t_1) \geq 1 \mid I(t_0) \geq 1\right\}.$$

In this case, any particle Z produces an offspring, by simulating on $[t_k, T]$ (or on $[t_k, t_{k+1}]$ in practice) a novel path segment starting from $Z(t_k)$, when it corresponds to an epidemics path that does not extinct before t_k, and is killed otherwise, see Fig. 4. A detailed description is provided in the appendix.

Remark 3.4. (DISCRETE-TIME MODELS) We point out finally that the IPBS approach can be naturally applied in a discrete-time context, so as to estimate tail probabilities $\mathbb{P}\{\sum_{k=0}^{t-1} I_k \geq N_c\}$, with $N_c \in \mathbb{N}$, at a given horizon $t \geq 1$ in a Reed-Frost model for instance. Selection/mutation steps are then performed at each intermediate time $k \in \{1, \ldots, t-1\}$: at stage k, $N \geq 1$ discrete paths are selected by means of a weight function ω_k defined on the path space and next mutate, through sampling of N independent chains from time k to time t. The crucial point naturally consists in a good choice for the weight functions used in the selection stage (which should be ideally based on an analysis of the variance of the corresponding estimates, when tractable). Typical choices are of the form $\omega_k(Z) = \exp(\alpha V(I_k))$ or $\omega_k(Z) = \exp(\alpha(V(I_k) - V(I_{k-1})))$, where $V : \mathbb{R} \to \mathbb{R}$ is a certain *potential function* and $\alpha \geq 0$, see section 4 for some examples.

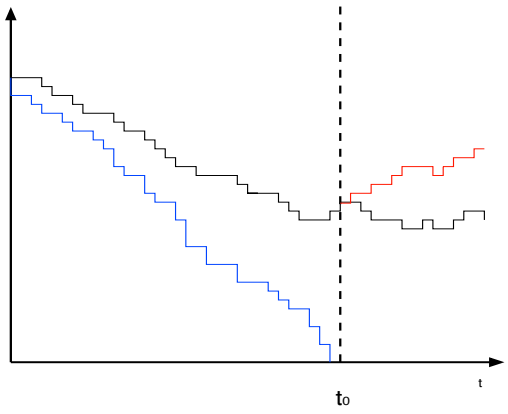

Figure 4: Time multi-level splitting: the path in blue extincts before time t_0 and is thus killed, while that in black does not and can be selected in order to produce an *offspring*, generated by sampling from time t_0 (in red)

4 Numerical experiments

Now that a comprehensive description of the IPBS approach has been given, it is the purpose of this section to provide strong empirical evidence that it is relevant in practice for rare event estimation in the context of (strongly Markovian) epidemics processes.

4.1 Toy examples

As a first go, we start with experiments based on simplistic epidemics models (see section 3 above), in order to check the accuracy of the estimates produced by IPBS methods. For comparison purposes, CMC and (adaptive) IS estimates are also displayed. Monte-Carlo replications have been generated, so as to estimate the variability of the estimators considered as well.

Reed-Frost model. In this discrete-time model, we consider the probability $\mathbb{P}(\sum_{k=0}^{t-1} I_k > N_c)$ for $t = 10$ and $N_c = 90$ or $N_c = 95$. Tables 1 and 2 below display estimates of this probability, together with

their empirical standard deviation based on $N = 1000$ Monte-Carlo replications. The IPBS approach is here implemented with two different potential functions (*cf* Remark 3.4): the method referred to as $IPBS(1)$ is based on the weight function $\omega_k(Z) = \exp(\alpha V(I_k))$ with $V(I) = I$, while that referred to as $IPBS(2)$ involves $\omega_k(Z) = \exp(\alpha(V(I_k) - V(I_{k-1})))$ with $V(I) = I$. For both IPBS methods, we test $\alpha = 0.1$ and $\alpha = 0.01$. The levels A_k appearing in the algorithms are set according to the Remark 3.2: we define these levels such that at each step, a certain proportion of paths are kept (50%, 80% or 95%) in our numerical example.

Two cases are considered, for $N_c = 90$ (Table 1) and $N_c = 95$ (Table 2). In the case $N_c = 90$ the rare event has a probability estimated by CMC of 1.44e-2, while this probability is 3.0e-4 for $N_c = 95$.

Table 1: Estimates of the tail probability $\theta = \mathbb{P}\{\sum_{k=0}^{t-1} I_k \geq N_c\}$ in a Reed-Frost model, with $N_c = 90$

Method	$\widehat{\theta}$	s.e.
CMC	1.44e-2	(3.7e-3)
CE	1.46e-2	(1.8e-3)
IPBS(1) $\alpha = 0.1$ 50%	9.1e-4	(2.8e-4)
IPBS(1) $\alpha = 0.01$ 50%	1.0e-3	(2.6e-4)
IPBS(1) $\alpha = 0.1$ 80%	1.46e-2	(2.3e-3)
IPBS(1) $\alpha = 0.01$ 80%	9.7e-3	(1.2e-3)
IPBS(1) $\alpha = 0.1$ 95%	1.42e-2	(3.1e-3)
IPBS(1) $\alpha = 0.01$ 95%	1.42e-2	(3.1e-3)
IPBS(2) $\alpha = 0.1$ 50%	1.0e-3	(2.8e-4)
IPBS(2) $\alpha = 0.01$ 50%	9.9e-4	(2.4e-4)
IPBS(2) $\alpha = 0.1$ 80%	1.0e-3	(2.8e-4)
IPBS(2) $\alpha = 0.01$ 80%	9.4e-3	(1.7e-3)
IPBS(2) $\alpha = 0.1$ 95%	1.40e-2	(3.0e-3)
IPBS(2) $\alpha = 0.01$ 95%	1.40e-2	(3.0e-3)

Table 2: Estimates of the tail probability $\theta = \mathbb{P}\{\sum_{k=0}^{t-1} I_k \geq N_c\}$ in a Reed-Frost model, with $N_c = 95$

Method	$\widehat{\theta}$	s.e.
CMC	3.0e-4	(5.5e-4)
CE	3.0e-4	(1.3e-4)
IPBS(1) $\alpha = 0.1$ 50%	2.0e-4	(8.8e-5)
IPBS(1) $\alpha = 0.01$ 50%	6.7e-5	(4.2e-5)
IPBS(1) $\alpha = 0.1$ 80%	4.1e-4	(3.4e-4)
IPBS(1) $\alpha = 0.01$ 80%	2.2e-4	(2.4e-4)
IPBS(1) $\alpha = 0.1$ 95%	3.2e-4	(4.2e-4)
IPBS(1) $\alpha = 0.01$ 95%	3.2e-4	(4.2e-4)
IPBS(2) $\alpha = 0.1$ 50%	1.0e-3	(5.6e-5)
IPBS(2) $\alpha = 0.01$ 50%	6.6e-5	(4.5e-5)
IPBS(2) $\alpha = 0.1$ 80%	2.5e-5	(2.4e-4)
IPBS(2) $\alpha = 0.01$ 80%	2.1e-4	(2.3e-4)
IPBS(2) $\alpha = 0.1$ 95%	3.1e-4	(4.3e-4)
IPBS(2) $\alpha = 0.01$ 95%	3.1e-4	(4.3e-4)

For both examples, we see that the estimation of CMC match with the estimation obtained by CE or by the IPBS methods when the levels are chosen such that at each step 95% of the paths are kept. When $N_c = 95$, standard deviation of the estimates are high and the obtained values are not always accurate.

Standard Markovian SIR model. We now consider a simple continuous-time Markovian epidemics model with no demography, as described in §2.1, in the case where the target is again the tail probability related to the epidemics size, $p_f(N_c)$ namely. We use the parameters proposed in the two examples presented in O'Neill and Roberts [24]. The first set of parameters corresponds to a toy model: $s_0 = 9$, $i_0 = 1$, $\mu \equiv 0$, $\lambda(S, I) = \lambda SI$ with $\lambda = 0.12$ and $\gamma(I, R) = \gamma I$ with $\gamma = 1$. We compared the results obtained by means of the CMC, CE and IPBS methods. Here, the method referred to as $IPBS(1)$ implements the algorithm described in the previous section, while that referred to as $IPBS(2)$ corresponds to the variant explained in Remark 3.1.

Table 3: Estimates of the tail probability $\theta = p_f(N_c)$ of the size of the epidemics in a standard Markovian SIR model without demography

Method	$\widehat{\theta}$	s.e.
CMC	2.0e-2	(4.5e-3)
CE	2.0e-2	(2.5e-3)
IPBS(1) - 1%	2.1e-2	(4.5e-3)
IPBS(1) - 5%	2.1e-2	(4.0e-3)
IPBS(1) - 20%	2.5e-2	(3.5e-3)
IPBS(2) - 1%	2.0e-2	(4.5e-3)
IPBS(2) - 5%	2.1e-2	(8.0e-3)
IPBS(2) - 20%	2.4e-2	(2.2e-2)

The second example in [24] comes from Bailey [5, p.125]. It is a smallpox outbreak in a closed community of 120 individuals in Abakaliki, Nigeria. Here the model is as above with the parameters $s_0 = 119$, $i_0 = 1$, $\lambda = 0.0008254$ and $\gamma = 0.087613$. The results are displayed in Table 4.

Table 4: Estimates of the tail probability $\theta = p_f(N_c)$ of the size of the epidemics in a standard Markovian SIR model without demography

Method	$\widehat{\theta}$	s.e.
CMC	2.5e-3	(1.6e-3)
CE	1.6e-3	(2.3e-4)
IPBS(1) - 1%	2.7e-3	(1.3e-3)
IPBS(1) - 5%	2.9e-3	(9.0e-4)
IPBS(1) - 20%	3.6e-3	(6.7e-4)
IPBS(2) - 1%	2.8e-3	(2.9e-3)
IPBS(2) - 5%	3.1e-3	(5.3e-3)
IPBS(2) - 20%	3.6e-3	(5.8e-3)

In both examples, CMC provides a good estimator of the rare probability (with 90.4% of non-zero estimates, in the second example, *i.e.* where the rare event has been observed). We take its results as a benchmark.

In Table 3, in a population of 10 individuals, we can see that every method provides a good estimate. Switching to a population of 120 individuals, one observes that CE faces difficult numerical problems related to the computation of the likelihood ratios. This method is avoided in the sequel.

The IPBS method which turns out to be the more robust is the IPBS method 1, where the levels are defined so that 1% of the paths are kept. In contrast to the Reed-Frost example, where the IPBS methods which work best correspond to a high proportion of kept trajectories (95%), here the methods that give the results which match the best CMC correspond to those where only 1% of the path at each iteration are kept. This may be explained by the number of iterations needed. IPBS for Reed-Frost model is implemented with a constant number of iterations, which is the number of time steps until t. Being too restrictive, we obtain only zero as conditional probability estimates. For the continuous time SIR model, the number of iterations is directly linked to the proportion of kept paths. The algorithm stops when the fixed proportion of best paths reaches the level N_c. When keep too many paths, the iteration becomes lengthy.

4.2 An age-structured HIV epidemic model with contact-tracing

We now consider a numerical individual-centered epidemic model, proposed and studied in the context of an asymptotically large population by [12], which is effectively used for anticipating the spread of HIV in Cuba and has been statistically fitted by the means of *Approximate Bayesian Computation* techniques (see [7] for further details) based from the HIV data repository described at length in [3]. Experiments are naturally (and fortunately) impossible in the context of epidemics. The capacity to simulate events of interest and estimate their probability of occurrence is thus of prime importance, in order to compare the effects of different control strategies for instance. Here we investigate the impact of the contact-tracing mechanism on the probability that, by means of the IPBS method described in the previous section.

As most realistic epidemics models really used by practitioners, it is more complex than the standard Markovian SIR model with demography recalled in subsection 2.1, though based on the same general concepts. Precisely, this model accounts for the effect of the contact-tracing detection system set-up since 1986 in order to control the HIV epidemics accross the island by stipulating a *structure by age* on

the class R (corresponding to the individuals diagnosed as HIV positive). The R subpopulation is hence described by a *point measure* R_t indicating the time points since each individual in the R compartment has been identified by the public health system as infected, *i.e.* $R_t([a_1, a_2])$ represents the number of positive diagnoses between times $t - a_2$ and $t - a_1$ for all $0 \leq a_1 < a_2 < +\infty$. Apart from this, the (Markovian) dynamics of the epidemics process $\{(S(t), I(t), R_t(da))\}$ is described by the flow diagram in Fig. 1 with $\mu \equiv 0$, $\lambda(S, I) = \lambda SI$ and $\gamma(I, R) = \gamma_1 I + \gamma_2 I \int_{a=0}^{+\infty} \exp(-ca) R(da)$ with $\lambda = 5.4\ 10^{-8}$, $\rho \equiv 0\ 10^{-6}$, $\gamma_1 = 0.13$, $\gamma_3 = 0.19$ and $c = 1$. The second term involved in the rate $\gamma(I, R)$ models the way detected individuals contribute to contact-tracing detection (notice incidentally that the smaller the parameter c, the more difficult the early stages of search for contact, refer to §2.1 in [12]).

Our purpose is to estimate $p_f(N_c)$ for various values of N_c: 8500, 8800 and 9000. As previously, IPBS is obtained with 1000 particles. For the CMC, 10e6 simulations have been performed. This permits to obtain a good estimate of the small probability $p_f(N_c)$ but also to compare CMC to IPBS. Indeed, if we separate the 10e6 simulations into 1000 runs of 1000 simulations, this allows us to count how many times the run provides an estimate equal to zero (the rare event has not been observed). As shown in Table 5, the CMC fails for the two last cases: whereas for $N_c = 8500$, only 2.4% of the simulations lead to an empirical probability equal to 0, this proportion is 84.4% and 98.6% for $N_c = 8800$ and $N_c = 9000$. This emphasizes the importance of the IPBS methods. CE methods do not give good results on such large populations, the computation of likelihood ratios being very sensitive numerically.

Table 5: Estimates of the tail probability $\theta = p_f(N_c)$ of the size of the age-structured epidemics model with contact-tracing for Cuban HIV epidemic

Method	$\widehat{\theta}$	(s.e.)
$N_c = 8500$		
CMC	3.4e-3	(1.8e-3)
IPBS1 - 1%	3.5e-3	(1.7e-3)
IPBS2 - 1%	3.5e-3	(3.8e-3)
$N_c = 8800$		
CMC	1.7e-4	(4.0e-4)
IPBS1 - 1%	1.5e-4	(3.0e-4)
IPBS2 - 1%	1.7e-4	(9.7e-4)
$N_c = 9000$		
CMC	1.4e-5	(1.2e-4)
IPBS1 - 1%	4.3e-6	(4.4e-5)
IPBS2 - 1%	8.4e-6	(2.1e-4)

5 Conclusion

Though (fortunately) rare, crisis situations related to the spread of a communicable infectious disease, are of great concern to public-health managers. However, proper use of simulation-based statistical methods tailored for the estimation of such rare events is not well-documented in the mathematical epidemiology literature. Indeed, the vast majority of analyses focus on the likeliest scenarios, on events occurring with large or even overwhelming probability (*e.g.* a large outbreak when the basic reproduction number is larger than one). In contrast, the present article provides an overview of recent techniques for rare event probability estimation and simulation in the context epidemics models and show how they can be used practically in order to provide efficient risk assessment tools for public-health management. The numerical results displayed in this paper provides strong empirical evidence that simulation methods based on interacting and branching particle systems are quite promising for this specific purpose.

Remark 5.1. The authors are grateful to Prof. H. de Arazoza for his helpful comments. The authors acknowledge support by the French Agency for Research under the grant funding the research project Viroscopy (ANR-08-SYSC-016-02). A.C. and V.C.T. have additional support by the Labex CEMPI (ANR-11-LABX-0007-01). The PhD of A.C. is supported by the Agence Nationale de Recherches sur le Sida et les hpatites virales (ANRS) through the project 12376.

Appendix - Temporal multilevel splitting

Here we show that the branching particle model sketched in Remark 3.3 can be used for estimating the probability $p_d(T)$ introduced in §2.2. More generally, we consider a continuous-time strong Markov process $Z = \{Z(t)\}_{t \geq 0}$ taking its values in a measurable space E with initial state $z_0 \in E$ and a Harris

recurrent set $B \subset E$. Let $\tau_B = \inf\{t > 0 : \ Z(t) \in B\}$ denote the hitting time to the set B. Our goal is here to estimate the tail probability $\pi = \mathbb{P}\{\tau_B > t\}$, *i.e.* the probability that the hitting time τ_B exceeds the (large) threshold value $t > 0$, by the means of time sublevels $t_0 = 0 < t_1 < \ldots < t_K < t_{K+1} = t$. At each stage k, the selection step simply consists in drawing with replacement among the paths Z that have not reached B before time t_k: we set $\omega_k(Z) = 1$ in this case and $\omega_k(Z) = 0$ otherwise.

TEMPORAL MULTILEVEL SPLITTING

1. **Initialization.** Start with a collection of $N \geq 1$ simulated trajectories $Z_0^{(1)}, \ \ldots, \ Z_0^{(N)}$ of the Markov process indexed by $i \in \{1, \ \ldots, \ N\}$, with the same initial condition z_0 and the same weights $\omega_0^{(i)} = 1$, $1 \leq i \leq N$. Denote by $\tau_B^{(i)}$ the corresponding hitting times.

2. **Iterations.** For $k = 1, \ \ldots, \ K$,

   (a) Let $\mathcal{I}_{1,k}$ be the subset of indices $i \in \{1, \ \ldots, \ N\}$ corresponding to the paths $Z_{k-1}^{(i)}$ which have not reached the subset B before time t_k, *i.e.* such that $\tau_B^{(i)} > t_k$, and denote by $\#\mathcal{I}_{1,k}$ its cardinality (when it is equal to 0, the algorithm is stopped and re-started). Set $\mathcal{I}_{0,k} = \{1, \ \ldots, \ N\} \setminus \mathcal{I}_{1,k}$. For each path indexed by $i \in \mathcal{I}_{1,k}$, set $Z_k^{(i)} = Z_{k-1}^{(i)}$.

   (b) For each path indexed by $i \in \mathcal{I}_{0,k}$:

   - (SELECTION STEP) independently draw a particle $Z_k^{(j)}$ from distribution $\sum_{j \in \mathcal{I}_{1,k}} \omega_k^{(j)} \cdot \delta_{Z_k^{(j)}}$, with $\omega_k^{(j)} = 1/\#\mathcal{I}_{1,k}$.

   - (MUTATION STEP) Define $Z_k^{(i)}$ as the concatenation of the path $Z_k^{(j)}$ on $[0, t_k]$ with a path simulated from the state $Z_k^{(j)}(t_k)$ for times larger than t_k.

   (c) Compute $P_j = \mathcal{I}_{1,k}\#/N$ and pass onto stage $k+1$.

3. **Output.** Compute the estimate of the target probability $\pi = \mathbb{P}\{\tau_B > t\}$:

$$\widehat{\pi}_N = P_1 \times \ldots \times P_{K+1},$$

   where P_{K+1} is defined as the proportion of particles Z that have not reached the subset B before time t among those which had not reached A before time t_K.
   Compute also the empirical distribution

$$\mathcal{L}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{Z_{K+1}^{(i)}},$$

   which may serve as an estimate of the conditional law $\mathcal{L}$ of the epidemics process given the event $\{\tau_B > t\}$ occurs.

We highlight the fact that the probability $\mathbb{P}\{\tau_B > t\}$ is actually of the same form as (3). Indeed, this corresponds to the situation of the bivariate Markov process $\{(Z(t), t)\}_{t \geq 0}$ with the (rare) set $A = \mathbb{N}^* \times [T, \ +\infty[$ and $\mathcal{T}$ as the extinction time τ. Therefore, works by [10] may be adapted to prove consistence and asymptotic normality when the number of particles N tends to infinity. In particular, an adaptive variant of the temporal multilevel splitting is as follows.

Adaptive variant. The method described above requires to fix in advance the number of time points and the time-points themselves, whereas, ideally, they should be determined in an adaptive fashion. We start by running N independent paths of the epidemics and rank them by decreasing durations $\mathcal{T}^{(i)}$, $1 \leq i \leq N$. The first threshold t_1 can be chosen as the duration of the $k-1$-th longest epidemics, so that k paths are kept and $N-k$ are killed. For each killed path, we resample from the k paths that have been kept and resimulate the part of the path after t_1. This allows to define recursively a system of longer and longer epidemic paths.

References

[1] Ahamed, T., Borkar, V., Juneja, S.: Adaptive importance sampling technique for Markov chains using stochastic approximation. Op. Res. **54**(3), 489–504 (2006)

[2] Andersson, H., Britton, T.: Stochastic Epidemic models and Their Statistical Analysis, *Lecture Notes in Statistics*, vol. 151. Springer, New York (2000)

[3] de Arazoza, H., Joanes, J., Lounes, R., Legeai, C., Clémençon, S., Perez, J., Auvert, B.: The HIV/AIDS epidemic in Cuba: description and tentative explanation of its low prevalence. BMC Disease (2007)

[4] Asmussen, S., Glynn, P.: Stochastic Simulation: Algorithms and Analysis. Springer (2007)

[5] Bailey, N.T.J.: The mathematical theory of infectious diseases and its applications, second edn. Hafner Press [Macmillan Publishing Co., Inc.] New York (1975)

[6] Bartlett, M.: Some evolutionary stochastic processes. J. Roy. Statist. Soc. B **11**, 211–229 (1949)

[7] Blum, M., Tran, V.: HIV with contact tracing: a case study in approximate Bayesian computation. Biostatistics **11**(4), 644–660 (2010)

[8] Boer, P.T.D., Kroese, D., Mannor, S., Rubinstein, R.: A tutorial on the cross-entropy method. Annals of Operations Research **134**, 19–67 (2005)

[9] Bucklew, J.: An Introduction to Rare Event Simulation. Springer Series In Statistics. Springer (2004)

[10] Cérou, F., Guyader, A.: Adaptive multilevel splitting for rare event analysis. Stoch. Anal. Proc. **25**(2), 417–443 (2007)

[11] Cérou, F., Moral, P.D., LeGland, F., Lézaud, P.: Genetic Genealogical Models in Rare Event Analysis. Alea **1**, 181–203 (2006)

[12] Clémençon, S., Tran, V., Arazoza, H.D.: A stochastic SIR model with contact-tracing: large population limits and statistical inference. Journal of Biological Dynamics **2**(4), 392–414 (2008)

[13] DeBoer, P., Nicola, V., Rubinstein, R.: Adaptive importance sampling simulation of queuing networks method, pp. 646–655. Proceedings of the 2000 Winter Simulation Conference (2000)

[14] Decreusefond, L., Dhersin, J.S., P.Moyal, Tran, V.: Large graph limit for an SIR process in random network with heterogeneous connectivity. Annals of Applied Probability (2012). In press

[15] DelMoral, P.: Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Probability and its applications. Springer (2004)

[16] DelMoral, P., Miclo, L.: Branching and Interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering, pp. 1–145. Séminaire de Probabilités XXXIV. Lecture Notes in Mathematics No. 1729. J. Azéma, M. Emery, M. Ledoux and M. Yor (Eds) (2000)

[17] Doucet, A., de Freitas, N., Gordon, N.: Sequential Monte Carlo: Methods and Practice. Statistics for Engineering and Information Science. Springer (2001)

[18] Glasserman, P., Heidelberger, P., Shahabuddin, P., Zajic, T.: Multilevel splitting for estimating rare event probabilities. Oper. Res. **47**(4), 585–600 (1999)

[19] Isham, V.: Stochastic models for epidemics with special reference to AIDS. Annals of Applied Probability **3**(1), 1–27 (1993)

[20] Kermack, W., McKendrick, A.: A contribution to the mathematical theory of epidemics. Proc. Roy. Soc. Lond. A **115**, 700–721 (1927)

[21] Lagnoux, A.: Rare event simulation. Probability in the Egineering ans Informational Sciences **20**(1), 45–66 (2006)

[22] Lefèvre, C., Picard, P.: A non-standard family of polynomials and the final size distribution of reed-frost epidemic processes. Adv. Appl. Prob. **22**, 25–48 (1990)

[23] Mode, C., Sleeman, C.: Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases and Computers. World Scientific (2000)

[24] O'Neill, P., Roberts, G.: Bayesian inference for partially observed stochastic epidemics. J.R. Statist. Soc. A **162**(1), 121–129 (1999)

[25] Revuz, D.: Markov Chains. North-Holland (1984)

[26] Rubinstein, R.: Optimization of computer simulation models with rare events. European Jour. Op. Res. **99**, 89–112 (1996)

[27] Villén-Altamirano, M., Villén-Altamirano, J.: RESTART: a method for accelerating rare events simulation, pp. 71–76. Proceedings of the 13-th International Teletraffic Conference (1991)

[28] Volz, E.: SIR dynamics in random networks with heterogeneous connectivity. J. Math. Biol. **56**(3), 293–310 (2008)