Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

ON A GENERALIZED SPLITTING METHOD FOR SAMPLING FROM A CONDITIONAL DISTRIBUTION

Pierre L'Ecuyer

DIRO, GERAD, and CIRRELT
Pavillon Aisenstadt, Université de Montréal
Montréal (Québec), CANADA

Zdravko I. Botev

School of Mathematics and Statistics
The University of New South Wales
Sydney, NSW 2052, AUSTRALIA

Dirk P. Kroese

School of Mathematics and Physics
The University of Queensland
Brisbane, QLD 4072, AUSTRALIA

ABSTRACT

We study the behavior of a generalized splitting method for sampling from a given distribution conditional on the occurrence of a rare event. The method returns a random-sized sample of points such that *unconditionally* on the sample size, each point is distributed exactly according to the original distribution conditional on the rare event. In addition, for any measurable cost function which is nonzero only when the rare event occurs, the method provides an unbiased estimator of the expected cost. On the other hand, the distribution of these points depends on the (random) number of points in the sample, and these points are not independent. So if we select at random one of the returned points, its distribution differs in general from the exact conditional distribution given the rare event. But if we repeat the algorithm n times and select one of the returned points at random, the distribution of the selected point converges to the exact one in total variation when n increases. The empirical distribution of the set of all points returned over all n replicates also converges to the conditional distribution given the rare event. The method also provides consistent confidence intervals for conditional expectations given that the rare event occurs.

1 INTRODUCTION

We consider the problem of estimating a conditional expectation when the conditioning is on the occurrence of a rare event, and more generally, how to draw a sample from the corresponding conditional distribution. This has many applications in various areas. For example, in finance, the expected shortfall, defined as the expected loss given that the loss exceeds a given quantile of its distribution (Yamai and Yoshiba 2005), has this form. As another example, for bridge regression in a Bayesian setting (Polson, Scott, and Windle 2014), there is a constraint (upper bound) on the norm of the vector β of regression coefficients, and the estimation of β requires resampling it repeatedly from a distribution truncated to the area where β satisfies the constraint.

A naive rejection method for sampling from the conditional distribution is to repeatedly draw independent samples from the original distribution until the rare event occurs, and return the realizations where this happens, until a sufficiently large sample has been obtained. But when the event is very rare, for example smaller than 10^{-10}, this simple method would be too inefficient.

We study an alternative approach based on the *generalized splitting* (GS) algorithm introduced in Botev and Kroese (2012a), which is itself a modification of the classical multilevel splitting methodology for rare-event simulation (Kahn and Harris 1951; Ermakov and Melas 1995; Glasserman et al. 1999; Garvels

et al. 2002; L'Ecuyer et al. 2006; L'Ecuyer et al. 2007; L'Ecuyer et al. 2009). The general idea of GS is to define a discrete-time Markov chain whose state (at any given step) represents a realization of all the random variables involved in the simulation. This chain evolves by resampling these random variables (at each step) under a conditional distribution that pushes the chain toward a state that corresponds to the rare event of interest. This GS algorithm has been seen empirically to be very effective to estimate extremely small rare-event probabilities by simulation, for various applications; see for example Botev and Kroese (2012b), Botev et al. (2013), Botev et al. (2016). This includes certain large problems for which no other effective method is known; for example reliability estimation in very large networks. A good understanding of its properties is therefore warranted and important.

It was suggested in Botev and Kroese (2012a) that the final states of the set of trajectories that reach the rare event in the GS algorithm are "approximately" (and not exactly) distributed according to the distribution conditional on the rare event, but no proof or counterexample was provided, and it was also stated that GS provides an unbiased estimator of the rare-event probability. The prime purpose of the present paper is to explain this apparent contradiction and study more formally the behavior of the GS algorithm.

In a nutshell, what goes on is that for each run of GS there is a random number M of trajectories that reach the rare event, and the distribution of the terminal states of these trajectories depends on M. As a result, if we pick at random one of those M terminal states from a given run of GS, assuming that $M > 1$, in general this state does not obey the conditional distribution of the state given the rare event. On the other hand, if we run GS n times, independently, and collect the terminal states of all the trajectories that have reached the rare event over the n runs, their empirical distribution converges at an $\mathcal{O}(n^{-1/2})$ rate to the conditional distribution given the rare event. This can be used to provide a consistent estimator for an expectation conditional on the rare event, based on the GS algorithm. This conditional expectation can be written as a ratio of expectations, and a confidence interval for this ratio can be obtained based on GS combined with standard technology for estimating ratios of expectations.

The remainder is organized as follows. In Section 2 we define the problem considered in this paper and recall the GS algorithm. In Section 3 we prove various properties of the algorithm. Section 4 gives a small numerical illustration whose aim is to provide insight into the (sometimes surprising) behavior of the algorithm. Conclusions and directions for future research are given in Section 5.

2 THE GENERALIZED SPLITTING METHOD

We recall the GS method and explain how it applies to our problem. We consider a random vector $\mathbf{Y}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in some measurable space. For concreteness we assume that this measurable space is $(\mathbb{R}^d, \mathcal{E})$, where $\mathcal{E}$ is the Borel σ-algebra, and that $\mathbf{Y}$ has an absolutely continuous distribution with respect to the Lebesgue measure, with density f, from which we can easily sample independent realizations of $\mathbf{Y}$. Adaptation to discrete (or other kinds of) distributions is straightforward.

Suppose we are given a measurable set $B \in \mathcal{E}$ for which $p = \mathbb{P}[\mathbf{Y} \in B]$ is unknown and can be very small. Our goal is to sample $\mathbf{Y}$ from its distribution conditional on $\mathbf{Y} \in B$. This could be for the purpose of estimating the expectation $\mathbb{E}[h(\mathbf{Y})\mathbb{I}(\mathbf{Y} \in B)]$ or the conditional expectation $\mathbb{E}[h(\mathbf{Y}) \mid \mathbf{Y} \in B]$, for some real-valued cost function h, or for another reason.

A naive way of doing this is to sample $\mathbf{Y}$ from its unconditional distribution repeatedly and independently, and return the first realization of $\mathbf{Y}$ that belongs to B. This standard rejection method returns a $\mathbf{Y}$ that obeys exactly the desired conditional distribution, but it becomes impractical (much too inefficient) when p is very small, because the expected number of trials is $1/p$.

As an alternative approach, we consider here the GS algorithm introduced in Botev and Kroese (2012a), which we recall with some simplifications. To apply the GS algorithm, we first need to choose:

1. an *importance function* $S : \mathbb{R}^d \to (0, \infty)$ for which $\{\mathbf{y} \in \mathbb{R}^d : S(\mathbf{y}) > \gamma^*\} = B$ for some $\gamma^* > 0$,
2. an integer $s \geq 2$, called the *splitting factor*, and

3. an integer $\tau > 0$ and real numbers $0 = \gamma_0 < \gamma_1 < \cdots < \gamma_\tau = \gamma^*$ for which

$$\rho_t \stackrel{\text{def}}{=} \mathbb{P}[S(\mathbf{Y}) > \gamma_t \mid S(\mathbf{Y}) > \gamma_{t-1}] \approx 1/s \tag{1}$$

for $t = 1, \ldots, \tau$ (except for ρ_τ, which can be larger than $1/s$). These γ_t's represent the τ levels of the splitting algorithm.

For each level γ_t we construct a Markov chain whose stationary density is equal to the density of $\mathbf{Y}$ conditional on $S(\mathbf{Y}) > \gamma_t$ (a truncated density), given by

$$f_t(\mathbf{y}) \stackrel{\text{def}}{=} f(\mathbf{y}) \frac{\mathbb{I}(S(\mathbf{y}) > \gamma_t)}{\mathbb{P}[S(\mathbf{Y}) > \gamma_t]}. \tag{2}$$

We denote by $\kappa_t(\mathbf{y} \mid \mathbf{x})$ the density of the transition kernel of this Markov chain; that is, the density of the next state $\mathbf{Y}$ conditional on the current state $\mathbf{x}$. There are many ways of constructing this Markov chain and κ_t.

At the first stage (or level), generating $\mathbf{Y}$ conditional on $S(\mathbf{Y}) > \gamma_0$ means that we generate it from its original (unconditional) density, so $f_0 = f$. When a generated state $\mathbf{Y}$ at the first level satisfies $S(\mathbf{Y}) > \gamma_1$, then its density is obviously that of $\mathbf{Y}$ conditional on $S(\mathbf{Y}) > \gamma_1$, which is f_1. At the t-th stage, if a Markov chain starts from a state having density f_{t-1} and evolves according to the kernel κ_{t-1}, then each visited state also has density f_{t-1}, which is the stationary density for the Markov chain with this kernel.

Algorithm 1 Generalized Splitting

Require: $s, \tau, \gamma_1, \ldots, \gamma_\tau$

```
Generate a vector Y from its unconditional density f.
if S(Y) ≤ γ_1 then
    return  𝒴_τ = ∅ and M = 0.
else
    𝒴_1 ← {Y}
    for t = 2 to τ do
        𝒴_t ← ∅      // list of states that have reached the level γ_t
        for all Y ∈ 𝒴_{t-1} do
            set Y_0 = Y
            for j = 1 to s do
                sample Y_j from the density κ_{t-1}(· | Y_{j-1})
                if S(Y_j) > γ_t then
                    add Y_j to 𝒴_t
return  the list 𝒴_τ and its cardinality M = |𝒴_τ|.
```

Algorithm 1 describes this procedure with a single starting point. In the algorithm, $\mathscr{Y}_t$ denotes the set of states that have reached the level γ_t at step t. These are the states $\mathbf{Y}_j$ visited by Markov chain trajectories that start from some state $\mathbf{Y}_0 \in \mathscr{Y}_{t-1}$ and for which $S(\mathbf{Y}_j) > \gamma_t$. Indentation delimits the scope of the **if**, **else**, and **for** statements. This GS algorithm returns a list $\mathscr{Y}_\tau$ of states that belong to B (this list is a multiset, in the sense that it may contain the same state more than once) as well as the size of this list. Note that $\mathscr{Y}_t$ and M are random and their distributions depend on the choice of importance function S and transition kernel densities κ_t. We will sometimes use the notation $\mathbb{P}_{\text{GS}}$ and $\mathbb{E}_{\text{GS}}$ to denote probabilities and expectations in which those random objects are involved. Sometimes, for simplicity, we just use the generic $\mathbb{P}$ and $\mathbb{E}$, assuming implicitly that these random objects are defined on the same probability space as $\mathscr{Y}$.

In many applications there is a natural choice for the importance function S. Good values for s, τ, and the levels $\{\gamma_t\}$ can typically be found via an (independent) adaptive pilot algorithm, as explained in Botev and Kroese (2012a). Based on empirical investigations, taking $s = 2$ is usually the best choice.

3 PROPERTIES OF THE GS METHOD

In this section, we prove some properties of the GS algorithm, examine in what sense it approximates the distribution of the state conditional on B, and show how it can be used to estimate conditional expectations.

3.1 Unbiased Sampling at Each Level

The collection of all Markov chain trajectories in the GS algorithm forms a branching tree, with $\tau-1$ levels of branching, and where the branching (or splitting) of each chain at each level (when it occurs) is by the factor s. There is a total of s^{t-1} *potential trajectories* up to level t, for each t. In Algorithm 1, only $|\mathscr{Y}_t|$ of these s^{t-1} trajectories are actually *kept alive* up to level t; they are those trajectories for which the state at step t is placed in $\mathscr{Y}_t$. The other ones are discarded, either at step t or earlier. But for the purpose of studying unbiasedness properties, we can imagine a modified version of Algorithm 1 that keeps all the trajectories at all levels. In this imaginary modified version, we remove the statement "**if** $S(\mathbf{Y}_j) > \gamma_t$" from Algorithm 1, and we replace everywhere $\mathscr{Y}_t$ by $\overline{\mathscr{Y}}_t$. That is, we always add $\mathbf{Y}_j$ to $\overline{\mathscr{Y}}_t$, so $\overline{\mathscr{Y}}_t$ will eventually contain s^{t-1} states, which correspond to the s^{t-1} (potential) trajectories up to level t. Thus, $\overline{\mathscr{Y}}_1$ always has 1 element (drawn from f), and $\overline{\mathscr{Y}}_t$ has s^{t-1} elements, for $t = 2,\ldots,\tau$. To each of these s^{t-1} trajectories we assign an index $(1, j_2, \ldots, j_t)$, where for each $t \geq 2$, $j_t \in \{1,\ldots,s\}$ is the value of j that corresponds to this trajectory in the inner loop of the modified algorithm when the outer loop counter is at value t. Let us denote by $\mathbf{Y}(1, j_2, \ldots, j_t)$ the state at step t for the trajectory with index $(1, j_2, \ldots, j_t)$. The set $\overline{\mathscr{Y}}_t$ contains all these s^{t-1} states. The trajectories that are kept alive up to level t in the original algorithm are those for which the event

$$\mathscr{E}_t := \mathscr{E}_t(1, j_2, \ldots, j_t) := \{\mathbf{Y}(1) > \gamma_1, \ldots, \mathbf{Y}(1, j_2, \ldots, j_t) > \gamma_t\}$$

occurs. Proposition 1 tells us that the states for which this event occurs, which are those retained in $\mathscr{Y}_t$ by Algorithm 1, have exactly the correct conditional density. This means that the GS sampling at each level is unbiased.

Proposition 1 For any fixed level t and index $(1, j_2, \ldots, j_t)$, conditional on $\mathscr{E}_t(1, j_2, \ldots, j_t)$, the state $\mathbf{Y}(1, j_2, \ldots, j_t)$ has density f_t (exactly). For $t = \tau$, this is the density of $\mathbf{Y}$ conditional on $\{\mathbf{Y} \in B\}$.

Proof. The proof is by induction on t. At the first stage we generate $\mathbf{Y} = \mathbf{Y}(1)$ from its original density f. Conditional on $\mathscr{E}_1(1)$, $\mathbf{Y}(1)$ satisfies $S(\mathbf{Y}(1)) > \gamma_1$, and thus $\mathbf{Y}(1) \in \mathscr{Y}_1$. Its density is obviously that of $\mathbf{Y}$ conditional on $S(\mathbf{Y}) > \gamma_1$, so the result holds for $t = 1$. Let us now assume that the result holds for $t-1$, which means that conditional on $\mathscr{E}_{t-1}(1, j_2, \ldots, j_{t-1})$, the state $\mathbf{Y}(1, j_2, \ldots, j_{t-1})$ has density f_{t-1}. The states $\mathbf{Y}(1, j_2, \ldots, j_{t-1}), \mathbf{Y}(1, j_1, \ldots, j_{t-1}, 1), \ldots, \mathbf{Y}(1, j_1, \ldots, j_{t-1}, s)$ are the successive states visited by a Markov chain with kernel κ_{t-1} and stationary density f_{t-1}, and whose initial state $\mathbf{Y}(1, j_2, \ldots, j_{t-1})$ has already the stationary density f_{t-1}. Therefore, conditional on $\mathscr{E}_{t-1}(1, j_2, \ldots, j_{t-1})$, all these states also have the same density f_{t-1}. This implies that for any $j \in \{1,\ldots,s\}$, conditional on $\mathscr{E}_t(1, j_2, \ldots, j_{t-1}, j)$, the density of $\mathbf{Y}(1, j_2, \ldots, j_{t-1}, j)$ is the density f_{t-1} for $\mathbf{Y}$, conditional on $S(\mathbf{Y}) > \gamma_t$, which is f_t. This completes the induction. □

Proposition 2 For any measurable function $h : \mathbb{R}^d \to \mathbb{R}$ and any measurable subset $A \subseteq B$, we have

$$\mathbb{E}_{\mathrm{GS}}\left[\sum_{\mathbf{Y}\in\mathscr{Y}_\tau} h(\mathbf{Y})\mathbb{I}(\mathbf{Y} \in A)\right] = s^{\tau-1}\mathbb{E}[h(\mathbf{Y})\mathbb{I}(\mathbf{Y} \in A)], \tag{3}$$

where the first expectation is with respect to $\mathscr{Y}_\tau$ defined in Algorithm 1 and the second expectation is with respect to $\mathbf{Y}$ having its original density f.

Proof. We can write

$$\begin{aligned}
&\mathbb{E}_{\mathrm{GS}}\left[\sum_{\mathbf{Y}\in\mathscr{Y}_\tau} h(\mathbf{Y})\mathbb{I}(\mathbf{Y}\in A)\right] \\
&= \mathbb{E}_{\mathrm{GS}}\left[\sum_{\mathbf{Y}=\mathbf{Y}(1,j_2,\ldots,j_\tau)\in\overline{\mathscr{Y}}_\tau} h(\mathbf{Y})\mathbb{I}(\mathbf{Y}\in A)\mathbb{I}(\mathscr{E}_\tau(1,j_2,\ldots,j_\tau))\right] \\
&= s^{\tau-1}\mathbb{E}_{\mathrm{GS}}\left[h(\mathbf{Y}(1,j_2,\ldots,j_\tau))\mathbb{I}(\mathbf{Y}(1,j_2,\ldots,j_\tau)\in A)\mathbb{I}(\mathscr{E}_\tau(1,j_2,\ldots,j_\tau))\right] \\
&= s^{\tau-1}\mathbb{E}[h(\mathbf{Y})\mathbb{I}(\mathbf{Y}\in A)],
\end{aligned}$$

in which the sum in the second expectation is over all $s^{\tau-1}$ terminal states $\mathbf{Y}(1,j_2,\ldots,j_\tau)$ in the modified algorithm, and the second and third equalities follow from Proposition 1 with $t=\tau$, which tells us that the density of $\mathbf{Y}(1,j_2,\ldots,j_\tau)$ conditional on $\mathbf{Y}(1,j_2,\ldots,j_\tau)\in A$ and $\mathscr{E}_\tau(1,j_2,\ldots,j_\tau)$ does not depend on the index $(1,j_2,\ldots,j_\tau)$ and is equal to f_τ conditional on $\mathbf{Y}(1,j_2,\ldots,j_\tau)\in A$. Since $A\subseteq B$, the latter is the original density f of $\mathbf{Y}$ conditional on $\mathbf{Y}\in A$. □

Let $H(A)=|\mathscr{Y}_\tau\cap A|$, the number of states $\mathbf{Y}\in\mathscr{Y}_\tau$ returned by the GS algorithm that belong to A. By taking h as the identity function in (3), we obtain

$$\mathbb{E}_{\mathrm{GS}}[H(A)] = \mathbb{E}_{\mathrm{GS}}\left[\sum_{\mathbf{Y}\in\mathscr{Y}_\tau}\mathbb{I}(\mathbf{Y}\in A)\right] = s^{\tau-1}\mathbb{P}[\mathbf{Y}\in A]. \tag{4}$$

That is, the expected number of states $\mathbf{Y}\in\mathscr{Y}_\tau$ that belong to A is proportional to $\mathbb{P}[\mathbf{Y}\in A]$, and therefore proportional to the conditional probability $\mathbb{P}[\mathbf{Y}\in A\mid\mathbf{Y}\in B]=\mathbb{P}[\mathbf{Y}\in A]/\mathbb{P}[\mathbf{Y}\in B]$, with a known proportionality constant. In particular $\mathbb{E}_{\mathrm{GS}}[M]=s^{\tau-1}\mathbb{P}[\mathbf{Y}\in B]$. We thus have $\mathbb{E}_{\mathrm{GS}}[H(A)]/\mathbb{E}_{\mathrm{GS}}[M]=\mathbb{P}[\mathbf{Y}\in A\mid\mathbf{Y}\in B]$. On the other hand, under the empirical distribution determined by the set $\mathscr{Y}_\tau$, the probability of A given B is $H(A)/M$, and $\mathbb{E}_{\mathrm{GS}}[H(A)/M]\neq\mathbb{P}[\mathbf{Y}\in A\mid\mathbf{Y}\in B]$ in general.

3.2 Estimating the Conditional Expectation

Each run of the GS algorithm returns a sample of random size, $\mathbf{Y}_1,\ldots,\mathbf{Y}_M$, that satisfies the previous propositions. One might be tempted to conclude that to generate a single realization of $\mathbf{Y}$ from its distribution conditional on $\mathbf{Y}\in B$, it suffices to pick $\mathbf{Y}$ at random uniformly from $\mathbf{Y}_1,\ldots,\mathbf{Y}_M$, conditional on $M\geq 1$ (if $M=0$ we just retry), and that this could be used in particular to produce an unbiased estimator $h(\mathbf{Y})$ of the conditional expectation $\mathbb{E}[h(\mathbf{Y})\mid\mathbf{Y}\in B]$. But with this scheme, in general, the returned $\mathbf{Y}$ does not have the correct distribution (with pdf $f(\mathbf{y})/\mathbb{P}(\mathbf{Y}\in B)$ for all $\mathbf{y}\in B$), because the $\mathbf{Y}_j$'s and M are generally dependent. In fact, if we define $H=\sum_{\mathbf{Y}\in\mathscr{Y}_\tau}h(\mathbf{Y})$, we can write (in what follows, we just use $\mathbb{E}$ for $\mathbb{E}_{\mathrm{GS}}$, for simplicity):

$$\mathbb{E}[h(\mathbf{Y})\mid\mathbf{Y}\in B] = \frac{\mathbb{E}[h(\mathbf{Y})\mathbb{I}(\mathbf{Y}\in B)]}{\mathbb{P}[\mathbf{Y}\in B]} = \frac{\mathbb{E}[h(\mathbf{Y})\mathbb{I}(\mathbf{Y}\in B)]s^{\tau-1}}{\mathbb{E}[M]} = \frac{1}{\mathbb{E}[M]}\mathbb{E}\left[\sum_{\mathbf{Y}\in\mathscr{Y}_\tau}h(\mathbf{Y})\right] = \frac{\mathbb{E}[H]}{\mathbb{E}[M]} \stackrel{\mathrm{def}}{=} \nu,$$

in which the 4-th equality comes from (4) with $A=B$. This is a ratio of expectations. If we pick $\mathbf{Y}^*$ uniformly from $\mathscr{Y}_\tau=\{\mathbf{Y}_1,\ldots,\mathbf{Y}_M\}$ and compute $h(\mathbf{Y}^*)$ (if $M=0$ we try again independently until $M>0$), the expectation is

$$\mathbb{E}[h(\mathbf{Y}^*)] = \mathbb{E}\left[\frac{1}{M}\sum_{\mathbf{Y}^*\in\mathscr{Y}_\tau}h(\mathbf{Y}^*)\right] = \mathbb{E}[H/M],$$

This is the expectation of a ratio.

To estimate $\mathbb{E}[h(\mathbf{Y}) \mid \mathbf{Y} \in B]$, we can use standard techniques for the estimation of a ratio of expectations. Such techniques are widely used. For example, for regenerative simulation (Asmussen and Glynn 2007). A standard approach is to generate n independent replicates of the GS estimator, let $\mathscr{Y}_{\tau,1}, \ldots, \mathscr{Y}_{\tau,n}$ be the n sets $\mathscr{Y}_\tau$ obtained from these realizations, and let $M_i = |\mathscr{Y}_{\tau,i}|$ and $H_i = \sum_{\mathbf{Y} \in \mathscr{Y}_{\tau,i}} h(\mathbf{Y})$ be the realizations of M and H for replicate i, for $i = 1, \ldots, n$. The classical estimator of the conditional expectation is

$$\hat{\nu}_n = \frac{H_1 + \cdots + H_n}{M_1 + \cdots + M_n} = \frac{\bar{H}_n}{\bar{M}_n}, \tag{5}$$

where $\bar{H}_n$ and $\bar{M}_n$ are the averages of the n realizations of H and M, respectively, and when $\bar{M}_n = 0$ we can either define $\hat{\nu}_n = 0$ or increase the sample size until $M_i > 0$ for some i (asymptotically, this makes no difference).

Let us assume that $\mathrm{Var}[H] < \infty$. Then, $\hat{\nu}_n \to \nu$ with probability 1 when $n \to \infty$ (strong law of large numbers) and the estimator $\hat{\nu}_n$ also obeys a central limit theorem as follows. Define the empirical variances and covariances

$$\begin{aligned}
\hat{\sigma}_H^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (H_i - \bar{H}_n)^2, \\
\hat{\sigma}_M^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (M_j - \bar{M}_n)^2, \\
\hat{\sigma}_{HM} &= \frac{1}{n-1} \sum_{i=1}^{n} (H_i - \bar{H}_n)(M_i - \bar{M}_n),
\end{aligned}$$

and let

$$\hat{\sigma}_{\nu,n}^2 = (\hat{\sigma}_H^2 + \hat{\sigma}_M^2 \hat{\nu}_n^2 - 2\hat{\sigma}_{HM} \hat{\nu}_n)/(\bar{M}_n)^2. \tag{6}$$

By applying the delta method and Slutsky's theorem, one has that

$$\frac{\sqrt{n}(\hat{\nu}_n - \nu)}{\hat{\sigma}_{\nu,n}} \xrightarrow{\mathrm{d}} \mathsf{N}(0,1) \qquad \text{as } n \to \infty, \tag{7}$$

where $\mathsf{N}(0,1)$ means a standard normal random variable. This central limit theorem permits one to compute a consistent confidence interval for the ratio of expectations, based on n replicates of GS. Alternative (often more accurate) confidence intervals can also be computed via bootstrap methods; see Choquet, L'Ecuyer, and Léger (1999).

3.3 Sampling from the Conditional Distribution

To generate independent realizations of $\mathbf{Y}$ *approximately* from its conditional distribution given B, we can pick the realizations at random with replacement from $\mathscr{Y}_\cup = \mathscr{Y}_{\tau,1} \cup \cdots \cup \mathscr{Y}_{\tau,n}$, assuming that this set is nonempty. If it is empty, we try again with n fresh runs of GS, until $\mathscr{Y}_\cup$ is nonempty. Let $\widehat{\mathbb{Q}}_n$ be the (empirical) distribution of $\mathbf{Y}$ obtained in this way, defined by $\widehat{\mathbb{Q}}_n[A] = |\mathscr{Y}_\cup \cap A|/|\mathscr{Y}_\cup|$, let $\mathbb{Q}_n$ be the distribution defined by $\mathbb{Q}_n[A] = \mathbb{E}[\widehat{\mathbb{Q}}_n[A]]$, and let $\mathbb{Q}[\cdot] = \mathbb{P}[\cdot \mid B]$ denote the true conditional distribution. If $\mathbf{Y}^*$ is defined as in Section 3.2, $\mathbb{Q}_n[A]$ represents the prior probability that $\mathbf{Y}^* \in A$ before we run GS, while $\widehat{\mathbb{Q}}_n[A]$ is the posterior probability that $\mathbf{Y}^* \in A$ after we have run GS. We are interested in the convergence of $\widehat{\mathbb{Q}}_n$ and $\mathbb{Q}_n$ to $\mathbb{Q}$ when $n \to \infty$.

For any measurable set $A \subseteq B$, by taking $h(\mathbf{Y}) = \mathbb{I}[\mathbf{Y} \in A]$, we obtain

$$H_i = H_i(A) = \sum_{\mathbf{Y} \in \mathscr{Y}_{\tau,i}} \mathbb{I}(\mathbf{Y} \in A) = |\mathscr{Y}_{\tau,i} \cap A|,$$

the number of samples returned by run i of GS and that belong to A. We also have $p = \mathbb{P}[A]$, $m = \mathbb{E}[M] = ps^{\tau-1}$, $\nu = \mathbb{Q}[A] = \mathbb{P}[A \mid B]$, and we define $\hbar(A) = m\mathbb{Q}[A] = \mathbb{E}[H_i(A)]$. When we generate an observation $\mathbf{Y}$ from $\widehat{\mathbb{Q}}_n$, it belongs to A with probability $\bar{H}_n(A)/\bar{M}_n$ (assuming that $\bar{M}_n > 0$). Note that $\hbar(A) \le m$ and $m \approx 1$ in the context of GS. However, M_i and H_i take their values in $\{0, 1, \ldots, s^{\tau-1}\}$. We show that the empirical distribution $\widehat{\mathbb{Q}}_n$ converges to $\mathbb{Q}$ in the following large-deviation sense:

Proposition 3 For any $\varepsilon \in (0, p/2]$, we have

$$\sup_{A \subseteq B} \mathbb{P}\left[|\widehat{\mathbb{Q}}_n[A] - \mathbb{Q}[A]| > 2\varepsilon/p\right] \le 4e^{-2n\varepsilon^2}. \tag{8}$$

Proof. If we divide M_i and H_i by $s^{\tau-1}$ and apply Hoeffding's inequality (Hoeffding 1963), we obtain that for all $\varepsilon > 0$ and any measurable $A \subseteq B$,

$$\begin{aligned} \mathbb{P}\left[|\bar{M}_n - m| > \varepsilon'\right] &\le 2e^{-2n\varepsilon^2} \qquad \text{and} \\ \mathbb{P}\left[|\bar{H}_n(A) - \hbar(A)| > \varepsilon'\right] &\le 2e^{-2n\varepsilon^2}, \end{aligned}$$

where $\varepsilon' = \varepsilon s^{\tau-1} = m\varepsilon/p$. Therefore, with probability at least $1 - 4e^{-2n\varepsilon^2}$, we have that both $|\bar{M}_n - m| \le \varepsilon'$ and $|\bar{H}_n(A) - \hbar(A)| \le \varepsilon'$. When these two inequalities hold, for $\varepsilon' \le m/2$, we have

$$\begin{aligned} \frac{\bar{H}_n(A)}{\bar{M}_n} - \mathbb{Q}[A] &\le \frac{\hbar(A)+\varepsilon'}{m-\varepsilon'} - \frac{\hbar(A)}{m} = \frac{m(\hbar(A)+\varepsilon') - \hbar(A)(m-\varepsilon')}{m(m-\varepsilon')} \\ &= \frac{(m-\hbar(A))\varepsilon'}{m(m-\varepsilon')} \le \frac{\varepsilon'}{m-\varepsilon'} \le \frac{2\varepsilon'}{m} = \frac{2\varepsilon}{p} \end{aligned}$$

and

$$\begin{aligned} \mathbb{Q}[A] - \frac{\bar{H}_n(A)}{\bar{M}_n} &\le \frac{\hbar(A)}{m} - \frac{\hbar(A)-\varepsilon'}{m+\varepsilon'} = \frac{\hbar(A)(m+\varepsilon') - m(\hbar(A)-\varepsilon')}{m(m+\varepsilon')} \\ &= \frac{(\hbar(A)+m)\varepsilon')}{m(m+\varepsilon')} \le \frac{2\varepsilon'}{m} = \frac{2\varepsilon}{p}. \end{aligned}$$

This completes the proof. □

Based on this bound, in the next proposition we prove converge in total variation of $\mathbb{Q}_n$ to $\mathbb{Q}$. Note that total variation convergence of $\widehat{\mathbb{Q}}_n$ to $\mathbb{Q}$ cannot hold, because once $\widehat{\mathbb{Q}}_n$ is known, one can always select a set A of measure zero that contains all the data points, and therefore $\sup_{A \subseteq B} |\widehat{\mathbb{Q}}_n[A] - \mathbb{Q}[A]|$ is always 1.

Proposition 4

$$\lim_{n\to\infty} \sup_{A \subseteq B} |\mathbb{Q}_n[A] - \mathbb{Q}[A]| = 0. \tag{9}$$

Proof. We have

$$\begin{aligned} |\mathbb{Q}_n[A] - \mathbb{Q}[A]| &= \left|\mathbb{E}\left[\widehat{\mathbb{Q}}_n[A] - \mathbb{Q}[A]\right]\right| \\ &\le \mathbb{E}\left[|\widehat{\mathbb{Q}}_n[A] - \mathbb{Q}[A]|\right] \\ &\le \int_0^\infty \mathbb{P}\left[|\widehat{\mathbb{Q}}_n[A] - \mathbb{Q}[A]| > x\right] \mathrm{d}x \\ &\le \int_0^\infty \mathbb{P}\left[|\widehat{\mathbb{Q}}_n[A] - \mathbb{Q}[A]| > \varepsilon\right] (2/p)\mathrm{d}\varepsilon \qquad \text{(by taking } x = 2\varepsilon/p\text{)} \\ &\le \frac{8}{p}\int_0^\infty e^{-2n\varepsilon^2}\mathrm{d}\varepsilon = \frac{2\sqrt{2\pi}}{p\sqrt{n}}. \end{aligned}$$

□

3.4 Discussion on the Bounds

We emphasize that Proposition 3 holds regardless of the choice of importance function S and transition kernels κ_t for the Markov chains. It is a worst-case result, and for this reason it is rather weak in the situation where B is a rare event. Indeed, when p is very small, for ε/p to be small one must take $\varepsilon \ll p$, and then n must be huge for the exponential upper bound in (8) to be small. That is, the rare event problem remains. But this is for the worst possible implementation of GS. To get a better bound, we need stronger assumptions on the effectiveness of the GS algorithm.

In fact, we may consider n independent trials with the naive rejection method mentioned at the beginning of Section 2, define $M_i = \mathbb{I}[\mathbf{Y}_i \in B]$ and $H_i(A) = \mathbb{I}[\mathbf{Y}_i \in A]$ where $\mathbf{Y}_i$ is the replicate of $\mathbf{Y}$ on the ith trial, and then estimate $\mathbb{Q}[A]$ by $\bar{H}_n(A)/\bar{M}_n$ for each A. This is the same as approximating $\mathbb{Q}$ by the empirical distribution of the $\mathbf{Y}_i$ that belong to B. By applying Hoeffding's inequality to the binary variables M_i and $H_i(A)$, and mimicking the above argument for the ratio, we find exactly the same inequality as in (8). This naive rejection method with n trials corresponds (in terms of distribution) to a worst-case GS that would always give either $M_i = 0$ or $M_i = s^{\tau-1}$. In this case, observing $M_i > 0$ remains a rare event. This would be an extreme situation in which the ingredients of the GS method have been chosen in the worst possible way. Even then, GS cannot do worst than the naive rejection method (if we do not account for the different amount of work per sample).

In the worst case we just examined, there is a maximum amount of dependence between the chains. At the other (most optimistic) extreme, there is the situation in which M and all the states in $\mathscr{Y}_\tau$ are assumed to be independent. In this situation, $\mathscr{Y}_\tau$ turns out to be an independent sample drawn exactly from distribution $\mathbb{Q}$. In practical applications, we are usually far from these two extremes, and somewhere in between.

4 A SIMPLE BIVARIATE UNIFORM EXAMPLE

The purpose of this small example is to illustrate the main ideas in a simple setting where we know exactly the conditional distribution and how to generate from it, so we can compare the approximate distributions with the exact ones.

Suppose that $\mathbf{Y} = (Y_1, Y_2)$ has the uniform distribution over the two-dimensional unit square $\mathscr{Y} = [0,1]^2$. Define $S(\mathbf{y}) = S(y_1, y_2) = \max(y_1, y_2)$ and let $B = \{\mathbf{y} \in \mathscr{Y} : S(\mathbf{y}) \geq 1-\delta\}$, where $\delta > 0$ is a small constant. The density of $\mathbf{Y}$ conditional on $\mathbf{Y} \in B$ is obviously uniform over B, whose surface is $2\delta - \delta^2 = \delta(2-\delta)$, so the conditional density is $1/(\delta(2-\delta))$ over B, and it is easy to generate $\mathbf{Y}$ directly from it. We want to compare the conditional density for a sample returned by GS with this uniform density.

For the GS algorithm we take S as the importance function, we select integers $s \geq 2$ and $\tau \geq 2$, and we select the levels γ_t so that $\gamma_t^2 = 1 - s^{-t}$ for $t = 1, \ldots, \tau$, and $1 - \delta = \gamma_\tau$. With these choices, we have $\mathbb{P}[S(\mathbf{Y}) > \gamma_t \mid S(\mathbf{Y}) > \gamma_{t-1}] = 1/s$ and $\mathbb{P}[S(\mathbf{Y}) > 1-\delta] = s^{-\tau} = \delta(2-\delta)$ exactly.

For the Markov chain kernel κ_t, we shall consider two cases. The first is a *symmetric* Gibbs resampling scheme, which always resamples the two coordinates one after the other, in random order, conditional on $S(\mathbf{Y}) > \gamma_{t-1}$. That is, when resampling a coordinate Y_i of $\mathbf{Y}$, we erase and forget its current value resample it from its distribution conditional on $S(\mathbf{Y}) > \gamma_{t-1}$, given the other coordinates of $\mathbf{Y}$, so the chain will never again go below the level γ_{t-1} that we have already reached. In our case, whenever Y_1 is resampled, if $Y_2 > \gamma_{t-1}$ we resample Y_1 uniformly over $(0,1)$, otherwise we resample Y_1 uniformly over $(\gamma_{t-1}, 1)$. The procedure is symmetrical when we resample Y_2. Our second sampling scheme, to be described later, will be asymmetric.

We use Figure 1 to illustrate and discuss the behavior of GS for this example. The figure is for $s = 2$ and $\tau = 2$, but the discussion is for general s. GS first generates $\mathbf{Y}$ uniformly in the square. With probability $(s-1)/s$, $\mathbf{Y}$ falls in the white square and no point is returned, otherwise $\mathbf{Y}$ falls in the colored areas and then $\mathbf{Y}$ is resampled s times (each resample starts from the previous one) with coordinates resampled in random order and conditional on $S(\mathbf{Y}) > \gamma_1$ as described earlier. This produces a Markov chain trajectory over s steps, in this colored region. Out of these s resampled states, we retain those that fall in the set

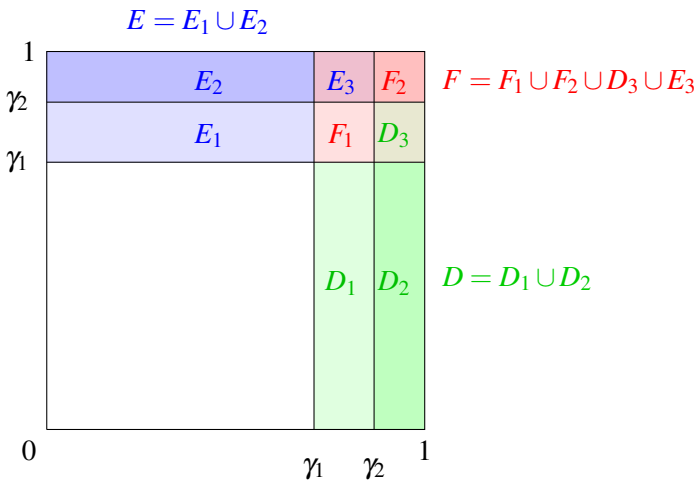

Figure 1: Sampling (approximately) conditionally on $\max(Y_1, Y_2) > \gamma_2$ via GS when (Y_1, Y_2) is uniform over the unit square. The picture is drawn for $s = 2$ and $\tau = 2$. Here $B = D_2 \cup D_3 \cup E_2 \cup E_3 \cup F_2$, $\gamma_1 = \sqrt{1/2} \approx 0.70710678$ and $\gamma_2 = \sqrt{3/4} \approx 0.86602540$.

$B = S(\mathbf{Y}) > \gamma_2$ and they form the multiset $\mathscr{Y}_2$. This GS procedure is repeated n times independently and the n realizations of $\mathscr{Y}_2$ (in case $\tau = 2$) are merged in a single multiset $\mathscr{Y}_\cup$, as in Section 3.3. One random point $\mathbf{Y}^*$ selected uniformly from $\mathscr{Y}_\cup$ has distribution $\widehat{\mathbb{Q}}_n$. And the prior distribution of $\mathbf{Y}^*$ before applying GS is $\mathbb{Q}_n$, which is the expectation of $\widehat{\mathbb{Q}}_n$. Two questions of interest here are how close is $\mathbb{Q}_n$ to $\mathbb{Q}$ when n is not too large, and what is the impact of the resampling strategy on the difference? Note that $\mathbb{Q}$ here is uniform with density s^τ over B.

In this simple example, given the initial distribution of $\mathbf{Y}$ and the resampling scheme, it is easy to see that $\widehat{\mathbb{Q}}_n$ and $\mathbb{Q}_n$ are uniform over each of the five colored regions in Figure 1 that comprise B. We will estimate the density of $\mathbb{Q}_n$ over each of these regions by a simulation experiment, with various values of n, and compare with the uniform density over B. We will replicate the following experiment $r = 10^7$ times, independently. For each replicate, we perform n independent runs of GS and construct the random multiset $\mathscr{Y}_\cup$. If $\mathscr{Y}_\cup$ is empty, this replicate has no contribution and we move to the next replicate. Otherwise, we compute the proportion of states in $\mathscr{Y}_\cup$ that fall in each of the five regions D_2, E_2, D_3, E_3, and F_2, and divide each proportion by the area of the corresponding region, to obtain a conditional density given $\mathscr{Y}_\cup$. To estimate the exact densities $d(D_2)$, $d(E_2)$, $d(D_3)$, $d(E_3)$, and $d(F_2)$ over the five regions for the distribution of the retained state under in this setting, we simulated this process r times and averaged the conditional densities over the R_0 replications for which $\mathscr{Y}_\cup$ was nonempty. We did this with $n = 1$, 10, 100, 1000. Table 1 reports the results for $s = 2$ and $s = 10$ (the given digits are roughly the significant digits of the density estimates). The exact density of $\mathbf{Y}$ conditional on B is s^2 over B, so it is 4 for $s = 2$ and 100 for $s = 10$. The results show that $\mathbb{Q}_n$ converges to $\mathbb{Q}$ very quickly with n and is already quite close even with $n = 1$.

The symmetric resampling of $\mathbf{Y}$ just examined works nicely. We now try our second resampling scheme (a different way to define κ_t) deliberately chosen to be bad. Instead of resampling the two coordinates of $\mathbf{Y}$, we resample only the first coordinate Y_1 conditional on $S(\mathbf{Y}) > \gamma_1$, and do not resample Y_2. We call this *one-way* resampling. In this case, all points $\mathbf{Y} \in \mathscr{Y}_\tau$ returned by GS on a given run have the same value of Y_2. We repeated the same simulation experiment with this poor resampling scheme, still with $r = 10^7$. The results are in Table 2 The bias in $\mathbb{Q}_n$ is now much larger than for the symmetric resampling case, and is larger for $s = 10$ than for $s = 2$. A larger s amplifies the bias because it creates more dependence. We

Table 1: Density estimates in each region with $r = 10^7$ independent replicates, with n independent runs of GS per replicate, for $s = 2$ and 10.

s	n	R_0	D_2	E_2	D_3	E_3	F_2
2	1	3756298	3.98	3.98	4.06	4.06	4.05
2	10	9909982	3.994	3.995	4.016	4.019	4.017
2	100	10000000	4.000	3.999	4.001	4.002	4.001
2	1000	10000000	4.000	4.000	4.000	4.000	4.000
10	1	651966	99.8	100.1	101.2	100.5	99.2
10	10	4902162	100.0	100.0	99.8	100.3	100.0
10	100	9987952	100.0	100.0	100.0	100.0	100.0
10	1000	10000000	100.0	100.0	100.0	100.0	99.9

can observe the convergence to the uniform density when n increases and this convergence is also slower when s is larger.

Table 2: Density estimates in each region, with $r = 10^7$ independent replicates and n independent runs of GS per replicate, for the one-way resampling, for $s = 2$ and 10.

s	n	R_0	D_2	E_2	D_3	E_3	F_2
2	1	3199391	4.82	3.12	5.84	3.13	3.13
2	10	9788291	4.30	3.68	4.67	3.68	3.68
2	100	10000000	4.022	3.977	4.049	3.978	3.978
2	1000	10000000	4.002	3.998	4.005	3.998	3.998
10	1	385940	170.5	25.9	254	25.9	26.4
10	10	3251227	167.8	28.9	244	28.9	29.2
10	100	9803488	140.8	58.1	166.7	58.1	57.9
10	1000	10000000	104.3	95.7	105.2	95.7	95.6

5 CONCLUSION

We have analyzed the convergence of a GS method for sampling from a conditional distribution, conditional on a rare event. This method has several applications in simulation and statistics. We proved convergence in total variation to the exact conditional distribution when the number n of replicates goes to infinity. The convergence was illustrated numerically. We also proved that the method provides an unbiased estimator of the corresponding conditional expectation, for any measurable cost function. In further work, we are interested in designing versions that achieve improved convergence rates for the total variation distance between the sampling conditional distribution and the exact one.

ACKNOWLEDGMENTS

This work has been supported by a discovery grant from NSERC-Canada, a Canada Research Chair, and an Inria International Chair, to P. L'Ecuyer. It was also supported by the Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers, under Grant Number CE140100049. Much of this paper was written in 2013 when the first author was visiting D. P. Kroese and Z. I. Botev in Brisbane and Sydney.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation*. New York: Springer-Verlag.

Botev, Z. I., and D. P. Kroese. 2012a. "Efficient Monte Carlo Simulation via the Generalized Splitting Method". *Statistics and Computing* 22(1):1–16.

Botev, Z. I., and D. P. Kroese. 2012b. "Efficient Monte Carlo Simulation via the Generalized Splitting Method". *Statistics and Computing* 22(1):1–16.

Botev, Z. I., P. L'Ecuyer, G. Rubino, R. Simard, and B. Tuffin. 2013. "Static Network Reliability Estimation Via Generalized Splitting". *INFORMS Journal on Computing* 25(1):56–71.

Botev, Z. I., P. L'Ecuyer, and B. Tuffin. 2016. "Static Network Reliability Estimation under the Marshall-Olkin Copula". *ACM Transactions on Modeling and Computer Simulation* 26(2):Article 14.

Choquet, D., P. L'Ecuyer, and C. Léger. 1999. "Bootstrap Confidence Intervals for Ratios of Expectations". *ACM Transactions on Modeling and Computer Simulation* 9(4):326–348.

Ermakov, S. M., and V. B. Melas. 1995. *Design and Analysis of Simulation Experiments*. Dordrecht, The Netherlands: Kluwer Academic.

Garvels, M. J. J., D. P. Kroese, and J.-K. C. W. Van Ommeren. 2002. "On the Importance Function in Splitting Simulation". *European Transactions on Telecommunications* 13(4):363–371.

Glasserman, P., P. Heidelberger, P. Shahabuddin, and T. Zajic. 1999. "Multilevel Splitting for Estimating Rare Event Probabilities". *Operations Research* 47(4):585–600.

Hoeffding, W. 1963. "Probability Inequalities for Sums of Bounded Random Variables". *Journal of the American Statistical Association* 58:13–29.

Kahn, H., and T. E. Harris. 1951. "Estimation of Particle Transmission by Random Sampling". *National Bureau of Standards Applied Mathematical Series* 12:27–30.

L'Ecuyer, P., V. Demers, and B. Tuffin. 2006. "Splitting for Rare-Event Simulation". In *Proceedings of the 2006 Winter Simulation Conference*, 137–148: IEEE Press.

L'Ecuyer, P., V. Demers, and B. Tuffin. 2007. "Rare-Events, Splitting, and Quasi-Monte Carlo". *ACM Transactions on Modeling and Computer Simulation* 17(2):Article 9.

L'Ecuyer, P., F. LeGland, P. Lezaud, and B. Tuffin. 2009. "Splitting Techniques". In *Rare Event Simulation Using Monte Carlo Methods*, edited by G. Rubino and B. Tuffin, 39–62. Wiley. Chapter 3.

Polson, N. G., J. G. Scott, and J. Windle. 2014. "The Bayesian bridge". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(4):713–733.

Yamai, Y., and T. Yoshiba. 2005. "Value-at-risk versus expected shortfall: A practical perspective". *Journal of Banking and Finance* 29(4):997–1015.

AUTHOR BIOGRAPHIES

PIERRE L'ECUYER is Professor in the Département d'Informatique et de Recherche Opérationnelle, at the Université de Montréal, Canada. He also holds an International Chair at Inria, in Rennes, France. He is a member of the CIRRELT and GERAD research centers. His main research interests are random number generation, quasi-Monte Carlo methods, efficiency improvement via variance reduction, sensitivity analysis and optimization of discrete-event stochastic systems, and discrete-event simulation in general. He has published over 270 scientific articles, book chapters, and books, and has developed software libraries and systems for random number generation and stochastic simulation (SSJ, TestU01, RngStreams, Lattice Builder, etc.). He has been a referee for 145 different scientific journals. He has served as Editor-in-Chief for *ACM Transactions on Modeling and Computer Simulation* from 2010 to 2013. He is currently Associate Editor for *ACM Transactions on Mathematical Software*, *Statistics and Computing*, and *International Transactions in Operational Research*. More information can be found on his web page: http://www.iro.umontreal.ca/~lecuyer.

ZDRAVKO BOTEV is Senior Lecturer at the School of Mathematics and Statistics at the University of New South Wales in Sydney, Australia. Previously he was an Australian Research Council fellow and a postdoctoral fellow at the University of Montreal. Currently, he is serving as an associate editor of the INFORMS Journal of Computing. His research interests include splitting and adaptive importance sampling

methods for rare-event simulation. He is the author the most accurate nonparametric density estimation software, which has been widely used in the applied sciences. For more information, visit his webpage: http://web.maths.unsw.edu.au/~zdravkobotev.

DIRK P. KROESE is a Professor of Mathematics and Statistics at The University of Queensland. He is the co-author of several influential monographs on simulation and Monte Carlo methods, including Handbook of Monte Carlo Methods and Simulation and the Monte Carlo Method, (3rd Edition). Dirk is a pioneer of the well-known Cross-Entropy method-an adaptive Monte Carlo technique, invented by Reuven Rubinstein, which is being used around the world to help solve difficult estimation and optimization problems in science, engineering, and finance. His personal website can be found under https://people.smp.uq.edu.au/DirkKroese/.