Sequential Monte Carlo Samplers for Rare Events

Adam M. Johansen, Pierre Del Moral, and Arnaud Doucet

Abstract. We present novel sequential Monte Carlo (SMC) algorithms for the simulation of two broad classes of rare events which are suitable for the estimation of tail probabilities and probability density functions in the regions of rare events, as well as the simulation of rare system trajectories. These methods have some connection with previously proposed importance sampling (IS) and interacting particle system (IPS) methodologies, particularly those of [**8, 4**], but differ significantly from previous approaches in a number of respects: especially in that they operate directly on the path space of the Markov process of interest.

1. Introduction

The problem of estimating rare event probabilities has attracted a great deal of attention in recent times. Here we propose novel algorithms which are applicable to two types of *rare events*, both of which are defined in terms of the canonical Markov chain:

$$\left(\Omega = \prod_{n=0}^{\infty} E_n, \mathcal{F} = \prod_{n=0}^{\infty} \mathcal{F}_n, (X_n)_{n\in\mathbb{N}}, \mathbb{P}_{\eta_0} \right),$$

where the law $\mathbb{P}_{\eta_0}$ is defined by its finite dimensional distributions, with the notation $x_{0:N} \triangleq (x_0, x_1, \ldots, x_N)$:

$$\mathbb{P}_{\eta_0} \circ X_{0:N}^{-1}(dx_{0:N}) = \eta_0(dx_0) \prod_{i=1}^{N} M_i(x_{i-1}, dx_i).$$

In the first instance we consider *static rare events*, which correspond to the probability that the *trajectory* of the Markov chain over a particular, deterministic time interval lies in some set, $\mathcal{T} \subset \prod_{p=0}^{P} E_p$, which is rare, $\mathbb{P}_{\eta_0}(X_{0:P} \in \mathcal{T}) << 1$. This technique, which is described in section 2, is applicable to problems such as those considered by [**8**].

In section 3 we consider what we term *dynamic rare events*, and these correspond to the probability that a homogeneous Markov chain on a space $(E, \mathcal{E})$ enters some rare set, $\mathcal{T} \subset E$, before it next enters some recurrent set, $\mathcal{R}$; i.e. $\mathbb{P}_{\eta_0}(X_\tau \in \mathcal{T})$ where the stopping time is defined as $\tau = \inf\{t : X_t \in \mathcal{R} \cup \mathcal{T}\}$. This corresponds to the classes of problems considered by RESTART (see [**10**] for a review), multilevel splitting [**12**] and the approaches of [**4, 5**].

In both instances, we define a sequence of distributions over the *path space* of these Markov chains – which in the dynamic case is clearly a trans-dimensional distribution in the sense that the dimension of the state of interest is a random variable: see [**14**]. The first of these distributions corresponds to the law of the Markov chain (up to a stopping time in the dynamic case) and subsequent distributions are distorted according to a sequence of potentials which ultimately causes the distributions to concentrate their mass on the rare events of interest. This allows us to estimate probabilities and related quantities via sequential Monte Carlo. This iterative approach makes it possible to obtain weighted samples with weights of low variance from the target distribution which would otherwise be extremely difficult to sample from.

This approach dramatically ameliorates the sample diversity relative to methods which iteratively extend the path and apply importance resampling, a strategy which will inevitably lead to degeneracy at the beginning of the path [**8**]. Furthermore, as noted by [**1**] if the transition kernel of the Markov chain admits heavy tails, then rare events are likely to be driven by single large shocks rather than an accumulation of small ones and consequently working on the path space is likely to produce much better results in such settings.

1.1. Sequential Monte Carlo Samplers. The sequential Monte Carlo samplers framework [**7**] is a general framework which allows us to sample, sequentially from a sequence of distributions defined upon essentially arbitrary spaces. We use this framework here to allow us to obtain samples from the sequences of distributions described above.

Given some sequence of distributions $\{\pi_t\}_{t\geq 1}$ on a sequence of measurable spaces $(E_t, \mathcal{F}_t)_{t\geq 1}$ (which we assume admit a density with respect to some suitable dominating measures, dx_t, denoted $\pi_t(dx_t) = \pi_t(x_t)dx_t$) from which we wish to obtain sets of weighted samples, we construct a sequence of distributions on a sequence of spaces of increasing dimension which admit the distributions of interest as marginals, by defining:

$$\widetilde{\pi}_t(dx_{1:t}) = \pi_t(dx_t) \prod_{s=t-1}^{1} L_s(x_{s+1}, dx_s),$$

where L_s is an arbitrary Markov kernel from space E_{s+1} to E_s (these act, in some sense, backwards in time). It is clear that standard SMC methods can now be applied on this space, by propagating samples forward from one distribution to the next according to a sequence of Markov kernels, $(K_t)_{t\geq 1}$, and correcting for the discrepancy between the proposal and the target distribution by importance sampling. Although a full discussion of the selection of appropriate kernels is beyond the scope of this paper, we note that one can typically employ MCMC kernels with invariant distribution π_t for this purpose, and obtain good results when the discrepancy between π_{t-1} and π_t is small.

As always it is important to ensure that a significant fraction of the particle set have non-negligible weights. The effective sample size (ESS), introduced by [**18**], is an approximation obtained by Taylor expansion of a quantity which describes the effective number of iid samples to which the set corresponds. The ESS is defined as $ESS = \left[\sum_{i=1}^{N} {W^{(i)}}^{-2}\right]^{-1}$ where $\left\{W^{(i)}\right\}$ are the normalized importance weights. This approximation, of course, fails if the particle set does not accurately represent the support of the distribution of interest. Resampling should be carried out after any iteration which causes the ESS to fall below a reasonable threshold (typically around half of the total number of particles), to prevent the sample becoming degenerate with a small number of samples having very large weights.

The rapidly increasing dimension raises the concern that the variance of the importance weights will be extremely high. It can be shown (again, see [**7**]) that the optimal form for the Markov kernels L_s – in the sense that they minimise the variance of the importance weights if resampling occurs at every time step – depends upon the distributions of interest and the importance sampling proposal kernels K_t in the following way, with the fraction being understood as the Radon-Nikodým derivative of $K_{t+1}(x_t, \cdot)$ with respect to $\int \pi_t(dx)K_{t+1}(x, \cdot)$:

$$L_t^{opt}(x_{t+1}, dx_t) = \pi_t(dx_t)\frac{dK_{t+1}(x_t, \cdot)}{d\int \pi_t(dx)K_{t+1}(x, \cdot)}(x_{t+1}). \tag{1}$$

In practice it is important to choose a sequence of kernels which are as close to the optimal case as possible to prevent the variance of the importance weights from becoming extremely large.

Sampling from a complex distribution by direct importance sampling will typically lead to very high variance importance weights, unless the structure of the problem permits the construction of an importance distribution which is very close to the target distribution. Making use of the SMC samplers approach, with a number of intermediate distributions, allows us to avoid the difficulties which one would ordinarily associate with importance sampling strategies on a space as large as the ones which we consider: employing the sequence of distributions allows us to ensure that each iteration of the algorithm amounts to importance sampling with a *good* importance distribution – providing that the proposal, K_t is well chosen – and the weights of our samples can be kept close to unity. This can produce a substantial reduction in the variance of the importance weights, relative to that obtained by the direct approach.

2. Static Rare Event Estimation

The approach which we propose is to employ a sequence of intermediate distributions which move smoothly from the "simple" distribution $\mathbb{P}_{\eta_0} \circ X_{0:P}^{-1}$ to the target distribution $\mathbb{P}_{\eta_0} \circ X_{0:P}^{-1}(\cdot | X_{0:P} \in \mathcal{T})$ and to obtain samples from these distributions using SMC methods. This approach has an interpretation as a mean field approximation to a Feynman-Kac flow in distribution space, and many theoretical results – including a central limit theorem – are consequently available [**6**]; stability has been established in a slightly different setting by [**15**] and work is ongoing to extend those results. By operating directly upon the path space, we obtain a number of advantages. It provides more flexibility in constructing the importance distribution than methods which consider only the time marginals, and allows us to take complex correlations into account. Later, we will show that it also allows us to consider the dynamic case, which is normally treated as a stopping time problem, in terms of trans-dimensional inference.

We can, of course, cast the probability of interest as the expectation of an indicator function over the rare set, and the conditional distribution of interest in a similar form as:

$$\begin{aligned}
\mathbb{P}_{\eta_0}\left(X_{0:P} \in \mathcal{T}\right) &= \mathbb{E}_{\eta_0}\left[\mathbb{I}_{\mathcal{T}}(X_{0:P})\right], \\
\mathbb{P}_{\eta_0}\left(dx_{0:p} \,|X_{0:P} \in \mathcal{T}\right) &= \frac{\mathbb{P}_{\eta_0}\left(dx_{0:p} \cap \mathcal{T}\right)}{\mathbb{E}_{\eta_0}\left[\mathbb{I}_{\mathcal{T}}(X_{0:P})\right]}.
\end{aligned}$$

We concern ourselves with those cases in which the rare set of interest can be characterised by some measurable function, $V : E_{0:P} \to \mathbb{R}$, which has the properties that:

$$\begin{array}{rccl}
V: & \mathcal{T} & \to & [\hat{V}, \infty), \\
V: & E_{0:P} \setminus \mathcal{T} & \to & (-\infty, \hat{V}).
\end{array}$$

In this case, it makes sense to consider a sequence of distributions defined by a potential function which is proportional to their Radon-Nikodỳm derivative with respect to the law of the Markov chain, namely:

$$g_\theta(x_{0:p}) = \left(1 + \exp\left(-\alpha(\theta)\left(V(x_{0:P}) - \hat{V}\right)\right)\right)^{-1}$$

where $\alpha(\theta) : [0,1] \to \mathbb{R}_+$ is a differentiable monotonically increasing function such that $\alpha(0) = 0$ and $\alpha(1)$ is sufficiently large that this potential function approaches the indicator function on the rare set as we move through the sequence of distributions defined by this potential function at the parameter values $\theta \in \{t/T : t \in \{0, 1, \ldots, T\}\}$.

Let $\left\{\pi_t(dx_{0:P}) \propto \mathbb{P}_{\eta_0}(dx_{0:P}) g_{t/T}(x_{0:P})\right\}_{t=0}^{T}$ be the sequence of distributions which we use. The SMC samplers framework allows us to obtain a set of samples from each of these distributions in turn via a sequential importance sampling and resampling strategy. Note that each of these distributions is over the first $P+1$ elements of a Markov chain: they are defined upon a common space.

In order to estimate the expectation which we seek, make use of the identity:

$$\mathbb{E}_{\eta_0}\left[\mathbb{I}_{\mathcal{T}}(X_{0:P})\right] = \pi_T\left[\frac{Z_1}{g_1(X_{0:P})}\mathbb{I}_{\mathcal{T}}(X_{0:P})\right],$$

where $Z_\theta = \pi_0(g_\theta)$ and use the particle approximation of the right hand side of this expression. Similarly, the subset of particles representing samples from π_T which hit the rare set can be interpreted as samples from the conditional distribution of interest.

We use the notation $(Y_t^{(i)})_{i=1}^N$ to describe the particle set at time t and $Y_t^{(i,j)}$ to describe the j^{th} state in the Markov chain described by particle i at time t. We further use $Y_t^{(i,-p)}$ to refer to every state in the Markov chain described by particle i at time t except the p^{th}, and similarly, $Y_t^{(i,-p)} \cup Y' \triangleq \left(Y_t^{(i,0:p-1)}, Y', Y_t^{(i,p+1:P)}\right)$, i.e., it refers to the Markov chain described by the same particle, with the p^{th} state of the Markov chain replaced by some quantity Y'.

2.1. Path Sampling Approximation. The estimation of the normalising constant associated with our potential function can be achieved by a Monte Carlo approximation to the *path sampling* formulation given by [**11**]. Given a parameter θ such that a potential function $g_\theta(x)$ allows a smooth transition from a reference distribution to a distribution of interest, as some parameter increases from zero to one, one can estimate the logarithm of the ratio of their normalising constants via an integral relationship.

In our cases, we can describe our sequence of distributions in precisely this form via a discrete sequence of intermediate distributions parameterized by a sequence of values of θ:

$$\begin{aligned}
\frac{\mathrm{d}\log g_\theta}{\mathrm{d}\theta}(x) &= \frac{(V(x)-\hat{V})}{\exp(\alpha(\theta)(V(x)-\hat{V}))+1}\frac{\mathrm{d}\alpha}{\mathrm{d}\theta} \\
\Rightarrow \log\left(\frac{Z_{t/T}}{Z_0}\right) &= \int_0^{t/T}\mathbb{E}_\theta\left[\frac{(V(\cdot)-\hat{V})}{\exp(\alpha(\theta)(V(\cdot)-\hat{V}))+1}\right]\frac{\mathrm{d}\alpha}{\mathrm{d}\theta}d\theta \\
&= \int_0^{\alpha(t/T)}\mathbb{E}_{\frac{\alpha(t/T)}{\alpha(1)}}\left[\frac{(V(\cdot)-\hat{V})}{\exp(\alpha(V(\cdot)-\hat{V}))+1}\right]d\alpha,
\end{aligned}$$

where $\mathbb{E}_\theta$ is used to denote the expectation under the distribution associated with the potential function at the specified value of its parameter.

The SMC sampler provides us with a set of weighted particles obtained from a sequence of distributions suitable for approximating this integral. At each α_t we can obtain an estimate of the expectation within the integral via the usual importance sampling estimator, and this integral can then be approximated via a trapezoidal integration. As we know that $Z_0 = 0.5$ we are then able to estimate the normalising constant of the final distribution and then use an importance sampling estimator to obtain the probability of hitting the rare set.

2.2. Density Estimation. Algorithm 1 provides pointwise estimates of the probabilities of rare events; one can use importance sampling to obtain estimates of the probabilities of similar events. In certain applications, once is actually interested in the estimation of PDFs (probability density functions). We remark that the algorithm presented here can be easily adapted to this task, and propose the following approach for one dimensional PDFs. The generalisation to the multivariate case is straightforward, although it becomes more difficult to accurately describe the particle localisation.

Assume that we wish to estimate the PDF of a function $f : E_{0:P} \to \mathbb{R}$ and have a set of points R at which an estimate is required (typically a grid of some sort in a region of interest). Assume that it is possible to describe the region in which the particles are located at each time step and that $\bar{R}_t \subset R$ is the set of points from R which lie inside this region at time t.

Let Δ denote the width of a window which is used to estimate the PDF at a point, ideally, this should be sufficiently small that the PDF is close to linear across regions of this width and large enough that a reasonable number of particles typically lie within ranges of this size.

In order to obtain an estimate of the PDF, we first attempt to obtain estimates, for each $r \in R$, of the probabilities:

$$\begin{aligned}
\mathbb{P}_{\eta_0}\left(f(Y_0)\in\left[r-\frac{\delta}{2}, r+\frac{\delta}{2}\right]\right) &= \mathbb{E}_0\left[\mathbb{I}_{[r-\frac{\delta}{2},r+\frac{\delta}{2}]}(f(Y_0))\right] \\
&= \mathbb{E}_{\frac{\alpha(t/T)}{\alpha(1)}}\left[\mathbb{I}_{[r-\frac{\delta}{2},r+\frac{\delta}{2}]}(f(Y_0))\frac{\mathrm{d}\pi_0}{\mathrm{d}\pi_t}(Y_t)\right]
\end{aligned}$$

At time t, each of these may be estimated using the particle set at that time, by:

$$\begin{aligned}
\hat{p}_t(r) &= \sum_{i=1}^N W_{t,i}\mathbb{I}_{[r-\frac{\Delta}{2},r+\frac{\Delta}{2}]}(f(Y_{t,i}))\widehat{\frac{\mathrm{d}\pi_0}{\mathrm{d}\pi_t}}(Y_{t,i}) \\
&= \sum_{i=1}^N W_{t,i}\mathbb{I}_{[r-\frac{\Delta}{2},r+\frac{\Delta}{2}]}(f(Y_{t,i}))\hat{Z}_t/g_{t/T}(Y_{t,i})
\end{aligned}$$

Algorithm 1 An SMC algorithm for static rare events.

At $t = 0$.
for $i = 1$ to N **do**
Sample $Y_0^{(i)} \sim \nu$ for some importance distribution ν.
Set $W_0^{(i)} \propto \frac{\pi_0(Y_0^{(i)})}{\nu(Y_1^{(i)})}$ such that $\sum_{j=1}^N W_0^{(j)} = 1$.
end for
for $t = 1$ to T **do**
if ESS < threshold **then** resample $\left\{W_{t-1}^{(i)}, Y_{t-1}^{(i)}\right\}_{i=1}^N$ using stratified resampling [**3**] to obtain $\left\{\hat{W}_{t-1}^{(i)}, \hat{Y}_{t-1}^{(i)}\right\}_{i=1}^N$ **else** let $\left\{\hat{W}_{t-1}^{(i)}, \hat{Y}_{t-1}^{(i)}\right\}_{i=1}^N = \left\{W_{t-1}^{(i)}, Y_{t-1}^{(i)}\right\}_{i=1}^N$.
If desired, apply a Markov kernel, $\tilde{K}_{t-1}$ of invariant distribution π_{t-1} to improve sample diversity, for each particle, sample $\tilde{Y}_{t-1}^{(i)} \sim \tilde{K}_{t-1}(\hat{Y}_{t-1}^{(i)}, \cdot)$. Otherwise, let $\left\{\tilde{Y}_{t-1}^{(i)}\right\}_{i=1}^N = \left\{\hat{Y}_{t-1}^{(i)}\right\}_{i=1}^N$.
for $i = 1$ to N **do**
Sample $Y_t^{(i)} \sim K_t(\tilde{Y}_{t-1}^{(i)}, \cdot)$.
Weight $W_t^{(i)} \propto \hat{W}_{t-1}^{(i)} \omega_t^{(i)}$ where the *incremental importance weight*, $\omega_t^{(i)}$ is defined through $\omega_t^{(i)} = \frac{\pi_t(Y_t^{(i)}) L_{t-1}(Y_t^{(i)}, \tilde{Y}_{t-1}^{(i)})}{\pi_{t-1}(\tilde{Y}_{t-1}^{(i)}) K_t(\tilde{Y}_{t-1}^{(i)}, Y_t^{(i)})}$, and $\sum_{j=1}^N W_t^{(j)} = 1$.
end for
end for
Approximate the path sampling identity to estimate the normalising constant:

$$\hat{Z}_1 = \frac{1}{2} \exp\left[\sum_{t=1}^T (\alpha(t/T) - \alpha((t-1)/T)) \frac{\hat{E}_{t-1} + \hat{E}_t}{2}\right]$$

$$\hat{E}_t = \frac{\sum_{j=1}^N W_t^{(j)} \frac{V\left(Y_t^{(j)}\right) - \hat{V}}{1 + \exp\left(\alpha_t\left(V\left(Y_t^{(j)}\right) - \hat{V}\right)\right)}}{\sum_{j=1}^N W_t^{(j)}}$$

Estimate the rare event probability using importance sampling:

$$p^\star = \hat{Z}_1 \frac{\sum_{j=1}^N W_T^{(j)} \left(1 + \exp(\alpha(1)(V\left(Y_T^{(j)}\right) - \hat{V}))\right) \mathbb{I}_{(\hat{V}, \infty]}\left(V\left(Y_T^{(j)}\right)\right)}{\sum_{j=1}^N W_T^{(j)}}.$$

where $\hat{Z}_t$ is the path sampling estimator of the normalising constant of the distribution $\pi_t(dx) \propto g_{t/T}(x)\pi_0(dx)$. At every point r, taking the mean of all of the individual estimators which were obtained from the region of support of the particle set at the appropriate time yields:

$$\hat{p}(r) = \frac{\sum_{t=1}^T \hat{p}_t(r) \mathbb{I}_{\bar{R}_t}(r)}{\sum_{t=1}^T \mathbb{I}_{\bar{R}_t}(r)}$$

For sufficiently small Δ, under suitably continuity conditions, this provides us with the density estimate which we seek.

2.3. Examples. We now provide some simple examples of algorithm 1.

2.3.1. *A Gaussian Random Walk.* It is useful to consider a simple example for which it is possible to obtain analytic results for the rare event probability. The tails of a Gaussian distribution serve well in this context, and we borrow the example of [**8**]. We consider a homogeneous Markov chain defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for which the initial distribution is a standard Gaussian distribution and each kernel is a standard Gaussian distribution centred on the previous position:

$$\eta_0(dx) = \mathcal{N}(dx; 0, 1) \quad \forall n > 0 : M_n(x, dy) = \mathcal{N}(dy; x, 1).$$

Threshold, $\hat{V}$	True log probability	SMC Mean	SMC Variance	k	T
5	-2.32	-2.30	0.016	2	333
10	-5.32	-5.30	0.028	4	667
15	-9.83	-9.81	0.026	6	1000
20	-15.93	-15.94	0.113	10	2000
25	-23.64	-23.83	0.059	12.5	2500
30	-33.00	-33.08	0.106	14	3500
$9\sqrt{15}$	-43.63	-43.61	0.133	12	3600
$10\sqrt{15}$	-53.23	-53.20	0.142	11.5	4000

TABLE 1. Means and variances of the estimates produced by 10 runs of the proposed algorithm using 100 particles at each threshold value for the Gaussian random walk example.

The function $V(x_{0:P}) := x_P$ corresponds to a canonical coordinate operator and the rare set $\mathcal{T} := E^P \times [\hat{V}, \infty)$ is simply a Gaussian tail probability: the distribution of X_P is simply $\mathcal{N}(0, P+1)$ as the sum of $P+1$ iid standard Gaussian random variables.

Sampling from π_0 is trivial. We employ an importance kernel which moves position i of the chain by $ij\delta$ where j is a random variable sampled from a distribution which makes the probability of proposing each possible change from a grid proportional to the associated target probability, and δ is an arbitrary scale parameter. The operator, $\mathcal{O}_\chi$, defined by $\mathcal{O}_\chi Y_t^{(i)} = \left(Y_t^{(i,p)} + p\chi\right)_{p=0}^{P}$, where χ is interpreted as a parameter, is used for notational convenience. This forward kernel can be written as:

$$K_t(Y_{t-1}^{(i)}, Y_t^{(i)}) = \sum_{j=-S}^{S} w_t(Y_{t-1}^{(i)}, Y_t^{(i)}) \delta_{\mathcal{O}_{\delta j} Y_{t-1}^{(i)}}(Y_t^{(i)}),$$

where the probability of each of the possible moves is given by

$$w_t(Y_{t-1}^{(i)}, Y_t^{(i)}) = \frac{\pi_t(Y_t^{(i)})}{\sum_{j=-S}^{S} \pi_t(\mathcal{O}_{\delta j} Y_{t-1}^{(i)})}.$$

This leads to the following optimal auxiliary kernel, as given by (1):

$$L_{t-1}(Y_t^{(i)}, Y_{t-1}^{(i)}) = \frac{\pi_{t-1}(Y_{t-1}^{(i)}) \sum\limits_{j=-S}^{S} w_t(Y_{t-1}^{(i)}, Y_t^{(i)}) \delta_{\mathcal{O}_{\delta j} Y_{t-1}^{(i)}}(Y_t^{(i)})}{\sum\limits_{j=-S}^{S} \pi_{t-1}(\mathcal{O}_{-\delta j} Y_t^{(i)}) w_t(\mathcal{O}_{-\delta j} Y_t^{(i)}, Y_t)}.$$

The incremental importance weight is consequently:

$$\omega_t(Y_{t-1}^{(i)}, Y_t^{(i)}) = \frac{\pi_t(Y_t^{(i)})}{\sum\limits_{j=-S}^{S} \pi_t(\mathcal{O}_{-\delta j} Y_t^{(i)}) w_t(\mathcal{O}_{-\delta j} Y_t^{(i)}, Y_t^{(i)}) \delta_{\mathcal{O}_{\delta j} Y_{t-1}^{(i)}}(Y_t^{(i)})}.$$

As the calculation of the integrals involved in the incremental weight expression tend to be analytically intractable in general, we have made use of a discrete grid of proposal distributions as proposed by [**21**]. This naturally impedes the exploration of the sample space. Consequently, we make use of a Metropolis-Hastings kernel of the correct invariant distribution at each time step (whether resampling has occurred, in which case this also helps to prevent sample impoverishment, or not). We make use of a linear schedule $\alpha(\theta) = k\theta$ and show the results of our approach (using a chain of length 15, a grid spacing of $\delta_t = 0.025$ and $S = 12$ in the sampling kernel) in table 1. For the purposes of comparison, we also implemented the algorithm of [**8**] and show its performance in table 2.

2.3.2. *Polarisation Mode Dispersion.* As a more realistic example, we consider the so-called “outage” probability due to polarisation mode dispersion (PMD) in single-mode optical fibres. This problem has recently been considered by a number of sources, including [**2, 8, 9**] who obtained good results with their methods. The advantage of the SMC

Threshold	True	IPS(1,000)		IPS(20,000)	
$\hat{V}$		Mean	Variance	Mean	Variance
5	-2.32	-2.28	0.008	-2.174	0.003
10	-5.32	-5.27	0.016	-5.320	0.003
15	-9.83	-9.81	0.086	-9.887	0.003
20	-15.93	-16.00	0.224	-15.92	0.004
25	-23.64	-23.38	2.510	-23.40	0.143
30	- 33.00	-29.33 [9]	-	-32.02 [5]	0.209

TABLE 2. Results of 10 runs of the algorithm of [**8**] with 1,000 particles and 20,000 particles, respectively. Numbers in square brackets indicate the number of runs which failed to hit the rare set at all.

samplers framework when applied to this problem is predominantly that it is able to obtain estimates of much smaller rare event probabilities than the methods which have been proposed in the literature, see figure 1, although it requires more knowledge on the part of the user than the method of [**8**].

We have a sequence of rotation vectors, r_n which evolve according to the equation:

$$r_n = R(\theta_n, \phi_n) r_{n-1} + \frac{1}{2}\Omega(\theta_n)$$

where ϕ_n is a random variable distributed uniformly on the interval $[-\pi, \pi]$ and θ_n is a random variable taking values in $[-\pi, \pi]$ such that $\cos(\theta_n)$ is uniformly distributed on $[-1, 1]$, $s_n = \mathrm{sgn}(\theta_n)$ is uniformly distributed on $\{-1, +1\}$, the vector $\Omega(\theta) = (\cos(\theta), \sin(\theta), 0)$ and $R(\theta, \phi)$ is the matrix which describes a rotation of ϕ about axis $\Omega(\theta)$.

It is convenient for our purposes to consider this as a Markov chain on the 6 dimensional space $E_n = \mathbb{R}^3 \times [0, 2\pi] \times [-1, 1] \times \{-1, +1\}$. Where $X_n = \{r_n, \phi_n, c_n = \cos(\theta_n), s_n = \mathrm{sgn}(\theta_n)\}$ We assume that $r_0 = (0, 0, 0)$ (This corresponds to the simulation performed in [**2**] – the "squares" in figure 2 therein) and then the finite dimensional distributions are given by:

$$\mathbb{P} \circ X_{0:n}^{-1}(r_{0:n}, \theta_{1:n}, c_{1:n}, s_{1:n}) = \delta_{(0,0,0)}(r_0) \prod_{i=1}^{N} \frac{1}{2\pi} \frac{1}{4} \left[\delta_{R(\theta_n, s_n \cos^{-1}(c_n)) r_{n-1} + \Omega(\theta_n)/2}(r_n) \right]$$

As $\{r_n\}$ can be obtained deterministically from $\{\phi_n, c_n, s_n\}$ we shall henceforth think of the distribution as a three dimensional one over just these variables. The magnitude of r is termed the differential group delay (DGD) and is the quantity of interest.

One option for a proposal distribution is an update move, which does not adjust the state associated with a particle at all, but does correct the particle weights to account for the change in distribution from one time step to the next. This has incremental weight equal to the ratio of the densities of the distribution at time t to that at time $t-1$. A priori this would lead very fast to sample degeneracy, and so we also apply a Markov kernel of the correct invariant distribution to maintain sample diversity. We employed a Metropolis-Hastings kernel with a proposal which randomly selects two indices uniformly between 1 and n and proposes replacing the ϕ and c values between those two indices with values drawn from the uniform distribution over $[-\pi, \pi] \times [-1, 1]$. This proposal is then accepted with the usual Metropolis acceptance probability. Results of the approach proposed here and that of [**8**] are illustrated in figure 1.

2.3.3. *Counting Problems.* Counting the number of ways of filling a knapsack with unique combinations of objects of known sizes is one of the classic counting problems of computer science, see for example, [**16**]. In its simplest form, the problem is this one: given a vector, a of n object sizes and a *knapsack* of capacity b we wish to estimate the number of different combinations of objects which will fit within the knapsack, which clearly corresponds to the number of unique $0-1$ vectors x for which the following inequality is satisfied:

$$a \cdot x = \sum_{i=1}^{P} a_i x_i \leq b. \tag{2}$$

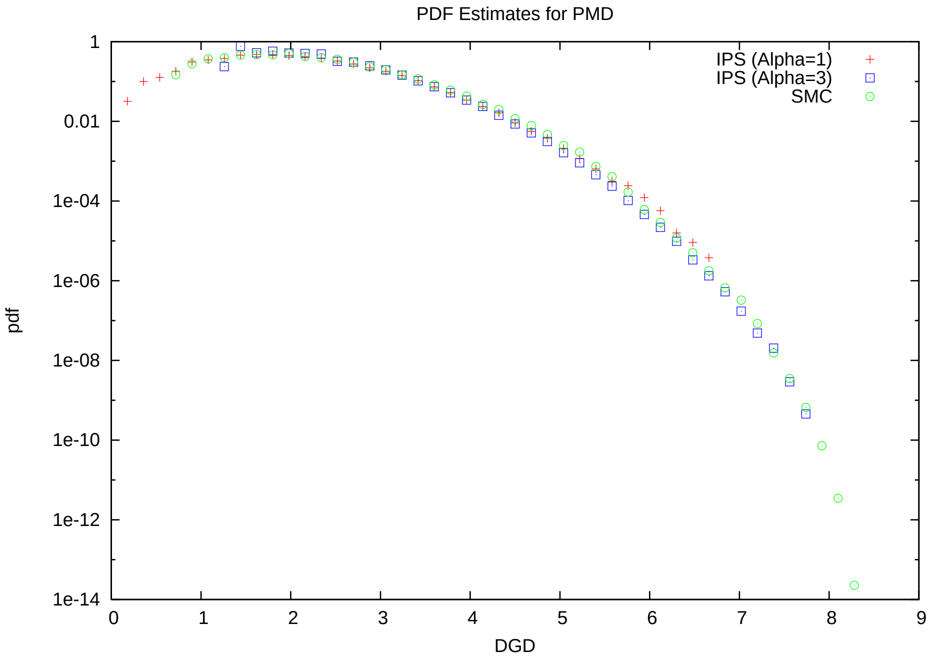

FIGURE 1. PDF estimates obtained with the SMC samplers methodology and the IPS approach of [**8**]. The SMC Sampler used $N = 100, T = 8000, c = 250$. The IPS used $N = 20,000, \alpha = 1$ and $N = 20,000, \alpha = 3$, respectively.

As observed by [**20**] the approximate solution of this problem can be cast as a rare event problem[1]. If one can determine the probability of a random vector sampled uniformly from the vertices of the P-dimensional unit hypercube satisfying the inequality, then the number of such valid vectors is clearly 2^P times that probability.

We define $V(x) := -a \cdot x$ such that inequality (2) can clearly be expressed as $V(x) \geq -(b + \min_{i \neq j} |a_i - a_j|)$. This approach turns the "rare" sets of interest into the level sets of our function and leads to a sequence of distributions which decreasing mass in states which violate the inequality. More precisely, we define the sequence of distributions from which we wish to sample as:

$$\pi_t(x) \propto \left[1 + \exp\left(\alpha\left(\frac{t}{T}\right)\left(a \cdot x - b - \min_{i \neq j} |a_i - a_j|\right)\right)\right]^{-1}.$$

It is trivial to sample from π_0 which is simply the uniform distribution over the vertices of the n-dimensional unit hypercube, and providing that T is *sufficiently* large, the discrepancy between successive distributions defined in this was will be small.

An obvious choice of move is an adjustment move which selects one element of the state vector and proposes a value of 0 or 1 with probabilities proportional to the probability of the resulting state under the current distribution, so the forward kernel is:

$$K_t(Y_{t-1}^{(i)}, Y_t^{(i)}) = \frac{1}{P} \sum_{p=1}^{P} \delta_{Y_{t-1}^{(i,-p)}} \left(Y_t^{(i,-p)}\right) \left[w_0^{pi} \delta_0(Y_t^{(i,p)}) + w_1^{pi} \delta_1(Y_t^{(i,p)})\right]$$

with

$$w_y^{pi} \propto \pi_t \left(Y_{t-1}^{(i,-p)} \cup y\right).$$

[1]Although the event of a particular set of objects fitting into the knapsack need not be a particularly rare one in some configurations, that does not pose any particular problems for the methodology described here.

N	Runs	Threshold, $\hat{V}$,	Mean	Variance
100	50	2.5	-12.889	0.04727
100	50	10.5	-10.214	0.03647
100	50	75.5	-2.0413	0.00501
1000	5	2.5	-12.811	0.04805
1000	5	10.5	-10.207	0.00652
1000	5	75.5	-2.0026	0.00027

TABLE 3. SMC results for the counting problem of section 2.3.3 obtained using a linear schedule for α increasing from 0 to 1 over 800 steps.

Which has the following optimal backward kernel (see equation (1)) associated with it:

$$L_{t-1}(Y_t^{(i)}, Y_{t-1}^{(i)}) = \frac{\pi_{t-1}(Y_{t-1}^{(i)}) \sum_{p=1}^{P} \delta_{Y_{t-1}^{(i,-p)}}\left(Y_t^{(i,-p)}\right)\left[w_0^{pi}\delta_0(Y_t^{(i,p)}) + w_1^{pi}\delta_1(Y_t^{(i,p)})\right]}{\sum_{p=1}^{P}\sum_{y\in\{0,1\}} \pi_{t-1}\left(Y_t^{(i,-p)} \cup y\right)}$$

Leading to the incremental weight expression:

$$\omega_t(Y_{t-1}^{(i)}, Y_t^{(i)}) = \frac{P\pi_t(Y_t^{(i)})}{\sum_{p=1}^{P}\sum_{y\in\{0,1\}} w_y^p \pi_{t-1}\left(Y_t^{(i,-p)} \cup y\right)}$$

This is essentially a random scan Gibbs sampler kernel, with an associated importance weight to compensate for the fact that the particles available from the previous time step are distributed according to π_{t-1} rather than π_t.

Results are shown in table 3 for $a = (1, 2, \ldots, 20)$ for knapsacks with capacities of 2,10 and 75; for comparison at similar computational cost (using 100,000 samples in total, compared with 80,000 for the SMC algorithm with 100 particles and 800 intermediate distributions) crude Monte Carlo simulation gave, over 50 runs, answers of 2.7263 with variance 30.562 and 44.040 with variance 453.27 in the two less rare instances. The logarithms of these means are -12.87 and -10.08, respectively, but many runs failed to produce any satisfactory solutions and a direct comparison is not possible due to the enormous variance. The *true* values for the thresholds given here were -12.764, -10.102 and -1.9756, respectively. Although a comparison with the approach of [**19**] would be possible, this is a much more challenging problem than that considered there in which $a = (1, \ldots, 4)$ and a threshold of $b = 3$ was employed.

3. Dynamic Rare Event Estimation

Length constraints prevent us from doing more than illustrating the applicability of our approach to the dynamic rare event estimation problem, and providing a simple toy example by way of illustration. Further details will be presented in [**17**].

Considering the space on which the paths of interest (i.e. those starting in the support of η_0 and then evolving according to the law of the Markov chain until they hit $\mathcal{R} \cup \mathcal{T}$) exist

$$F = \bigcup_{i=2}^{\infty} \{i\} \times \mathrm{supp}(\eta_0) \times (E \setminus (\mathcal{R} \cup \mathcal{T}))^{i-2} \times \mathcal{R} \cup \mathcal{T},$$

where we assume that the support of the initial distribution does not include either the rare set or the recurrent set – $\mathrm{supp}(\eta_0) \cap (\mathcal{R} \cup \mathcal{T}) = \emptyset$ – for notational convenience, we can see that this is actually a trans-dimensional estimation problem, as follows:

$$\mathbb{P}_{\eta_0}(X_\tau \in \mathcal{T}) = \sum_{p=2}^{\infty} \int \mathbb{P}_{\eta_0}(dx_{0:p}) \mathbb{I}_{\mathcal{T}}(x_p) \prod_{s=0}^{p-1} \mathbb{I}_{E\setminus(\mathcal{R}\cup\mathcal{T})}(x_s).$$

In common with many techniques for solving this problem, we employ a decreasing sequence of sets which concentrate themselves on the rare set of interest: $\mathcal{T} = \mathcal{T}_T \subset \mathcal{T}_{T-1} \ldots \mathcal{T}_2 \subset \mathcal{T}_1$. Our approach differs slightly, in that we endeavour to arrange these sets

Algorithm 2 An SMC algorithm for dynamic rare event simulation.

$t = 1$. Initialise an ensemble of N weighted particle: **for** $i = 1$ to N **do** sample $Y_1^{(i)}$ from the law of the Markov chain until it hits either $\mathcal{T}_1$ or $\mathcal{R}$, at stopping time $\tau_1^{(i)}$; set $W_1^{(i)} = \mathbb{I}_{\mathcal{T}_1}\left(X^{(i)}_{\tau_1^{(i)}}\right)$. **end for**

for $t = 2$ to T **do**

Resample, to obtain $\{\frac{1}{N}, \hat{Y}_{t-1}^{(i)}\}_{i=1}^N$. If desired, apply a Markov kernel (typically a reversible-jump kernel which may include dimension-changing moves [**13**]), $\tilde{K}_{t-1}$ of invariant distribution π_{t-1}: for each path-particle sample $\tilde{Y}_{t-1}^{(i)} \sim \tilde{K}_{t-1}(\hat{Y}_{t-1}^{(i)}, \cdot)$. Otherwise, let $\left\{\tilde{Y}_{t-1}^{(i)}\right\}_{i=1}^N = \left\{\hat{Y}_{t-1}^{(i)}\right\}_{i=1}^N$.

Propose a revised estimate for each path-particle from the proposal kernel K_t, which should correspond to extending the path if necessary until it hits either $\mathcal{T}_t$ or $\mathcal{R}$, and reweight the particle ensemble using $W_t^{(i)} = \mathbb{I}_{\mathcal{T}_t}\left(X_{\tau_t^{(i)}}\right)$ (for convenience we assume that K_t is the law of the Markov chain conditioned upon hitting $\mathcal{T}_{T-1}$ before $\mathcal{R}$).

end for

We can now estimate the quantity of interest: $p^\star = \prod_{t=1}^{T} \hat{Z}_t$ with $Z_t = \frac{1}{N}\sum_{i=1}^{N} W_t^{(i)}$.

---

such that the majority of paths reaching $\mathcal{T}_t$ before $\mathcal{R}$ also reach $\mathcal{T}_{t+1}$ before $\mathcal{R}$. That is, the sets are somehow closer together than is usually the case with splitting approaches. For simplicity we construct a sequence of distributions which place all of their mass on one of these sets, although it is easy to envisage situations in which potential functions more like that employed in the static case could produce better results. We define our synthetic distributions as:

$$\pi_t(X_{1:\tau_t}) = \mathbb{P}_{\eta_0}\left(X_{1:\tau_t} | X_{\tau_t} \in \mathcal{T}_t\right) = \mathbb{P}_{\eta_0}\left(X_{1:\tau_t}, X_{\tau_t} \in \mathcal{T}_t\right)/Z_t$$

with the stopping times $\tau_t = \inf\{t : X_t \in \mathcal{T}_t \cup \mathcal{R}\}$ and the normalising constant $Z_t = \mathbb{P}_{\eta_0}\left(X_{\tau_t} \in \mathcal{T}_t\right)$.

As in the static case, providing that we are able to obtain samples from this sequence of distributions, we can obtain an estimate of the ratio of normalising constants. Using such a 0-1 potential function makes it impossible to employ the path sampling [**11**] identity as the logarithm of the potential function no longer has a well defined derivative. However, we may still obtain an estimate of the ratio of normalising constants by the more naïve approach of taking the product of the particle system estimates of the ratio of normalising constants from one time step to the next. We could also employ a smooth potential function to allow us to employ the path sampling approach.

Algorithm 2 provides a fairly general framework for rare event probability estimation within this framework. We remark that, in full generality, it is simple to amend the algorithm to allow the proposal kernel, K_t, to modify that part of the path which has already sampled: this simply alters the importance weight.

3.1. Example. Consider a simple random walk over the integers, starting from $X_0 = 0$, defined by transition kernel:

$$M(x_n, dx_{n+1}) = (1-p)\delta_{x_n - 1}(dx_{n+1}) + p\delta_{x_n+1}(dx_{n+1})$$

If we defined $\mathcal{R} = (-\infty, -a]$ and $\mathcal{T} = [b, \infty)$ for two integers, $0 < a < b$, it is trivial to see that $X_\tau \in \{-a, b\}$ and it is straightforward to verify that

$$\mathbb{P}_{\eta_0}(X_\tau \in \mathcal{T}) = \begin{cases} \frac{a}{a+b} & \text{if } p = \frac{1}{2} \\ \frac{1-\left(\frac{1-p}{p}\right)^a}{1-\left(\frac{1-p}{p}\right)^{a+b}} & \text{otherwise} \end{cases}$$

As an illustration, consider the case where $p < 0.5$, with $a = 1$ and $b = 10$. Table 4 summarises the results of 100 runs of the SMC algorithm using ten intermediate distributions and various numbers of particles. In all cases, the proposal distribution K_t corresponds to extending the path until it hits either $\mathcal{T}_t = [t, \infty)$ or $\mathcal{R}$.

N	$p = 0.1$		$p = 0.2$	
	Mean	Standard Deviation	Mean	Standard Deviation
100	2.71×10^{-10}	2.76×10^{-10}	6.86×10^{-7}	5.10×10^{-7}
500	2.36×10^{-10}	0.98×10^{-10}	6.82×10^{-7}	1.86×10^{-7}
2000	2.55×10^{-10}	0.58×10^{-10}	7.29×10^{-7}	0.92×10^{-7}
5000	2.53×10^{-10}	0.38×10^{-10}	7.17×10^{-7}	0.50×10^{-7}

Table 4. Simulation results for the simple random walk example with $p = 0.1$ and $p = 0.2$, and in both cases $a = 1$ and $b = 10$. The true values are 2.55×10^{-10} and 7.15×10^{-7}, respectively.

4. Conclusions and Ongoing Work

We have presented two novel algorithms for estimating the probability of rare events, and the distribution of Markov chains conditioned upon the occurrence of such events. Work is ongoing in this area. Methodologically, we are interested in determining how to select the sequence of distributions to be employed, and whether this can be done adaptively (perhaps in a manner similar to [**5**]), and how the competing demands on computational power of using a large number of particles and using a large number of intermediate distributions can be best balanced; theoretically, it would be interesting to establish reasonable conditions under which the particle system is stable, as well as establishing computable bounds on the variance and bias of the estimates which it produces.

References

[1] S. Asmussen, K. Binswanger, and B. Højgaard. Rare events simulation for heavy-tailed distributions. *Bernoulli*, 6(2):303–322, 2000.
[2] G. Biondini, W. L. Kath, and C. R. Menyuk. Importance sampling for polarization-mode distortion. *IEEE Photonic Technology Letters*, 14(2):310–312, 2002.
[3] J. Carpenter, P. Clifford, and P. Fearnhead. An improved particle filter for non-linear problems. *IEEE Proceedings on Radar, Sonar and Navigation*, 146(1):2–7, 1999.
[4] F. Cérou, P. Del Moral, F. Le Gland, and P. Lezaud. Genetic genealogical models in rare event analysis. Prepublication, Laboratoire de Statistiques et Probabilités, Université Paul Sabatier, 118 route de Narbonne, 21062, Toulouse cedex, France, 2002.
[5] F. Cérou and A. Guyader. Adaptive multilevel splitting for rare event analysis. Research Report 5710, INRIA, October 2005.
[6] P. Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications.* Probability and Its Applications. Springer Verlag, New York, 2004.
[7] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society B*, 63(3):411–436, 2006.
[8] P. Del Moral and J. Garnier. Genealogical particle analysis of rare events. *Annals of Applied Probability*, 15(4):2496–2534, November 2005.
[9] J. Garnier and P. Del Moral. Simulations of rare events in fibre optics by interacting particle systems. *Elsevier Science*, 2006. Submitted.
[10] M. J. J. Garvels and D. P. Kroese. A comparison of RESTART implementations. In D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, editors, *Proceedings of the 1998 Winter Simulation Conference*, 1998.
[11] A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.
[12] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47(4):585–600, July 1999.
[13] P. J. Green. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.
[14] P. J. Green. Trans-dimensional Markov chain Monte Carlo. In Peter J Green, Nils Lid Hjort, and Sylvia Richardson, editors, *Highly Structured Stochastic Systems*, Oxford Statistical Science Series, chapter 6. Oxford University Press, 2003.
[15] A. Jasra and A. Doucet. Stability of sequential Monte Carlo samplers via the Foster-Lyapunov condition. *Submitted*, 2006.
[16] M. Jerrum and A. Sinclair. The Markov chain Monte Carlo method: an approach to approximate counting and integration. In D. S. Hochbaum et al., editors, *Approximation Algorithms for NP-hard Problems*, pages 482–520. PWS Publishing, Boston, 1996.
[17] A. M. Johansen, P. Del Moral, and A. Doucet. Sequential Monte Carlo samplers for rare event estimation. Technical Report CUED/F-INFENG/TR-543, University of Cambridge, Department of Engineering, Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, 2006. In preparation.
[18] A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, March 1994.
[19] A. Lagnoux. Rare event simulation. Preprint, Laboratoire de Statistique et Probabilités, Univeristé Paul Sabatier, 118, Route de Narbonne, 31062 Toulouse, Cedex 4, France, 2005. To appear in PEIS.

[20] A. Lagnoux. Rare event simulation. *Probability in the Engineering and Informational Sciences*, 20(1):43–66, 2006.
[21] G. W. Peters. Topics in sequential Monte Carlo samplers. M.Sc. thesis, University of Cambridge, Department of Engineering, January 2005.

Adam Johansen, University of Cambridge Department of Engineering, Trumpington Street, Cambridge, CB2 1PZ, UK

Pierre Del Moral, Laboratoire Jean Alexandre Dieudonné, Unité Mixte de Recherche CNRS 6621, Université Nice Sophia Antipolis, France

Arnaud Doucet, University of British Columbia Department of Statistics, 333-6356 Agricultural Road,Vancouver, BC, V6T 1Z2, Canada