Adam Michael Johansen
Christ's College

Some Non-Standard Sequential Monte Carlo Methods and Their Applications

A thesis submitted to the University of Cambridge
in partial fulfilment of the requirements of the degree
Doctor of Philosophy

December 15, 2006

Signal Processing Laboratory
Department of Engineering
University of Cambridge

For Louise, Maureen and Michael.

"What the world needs is not dogma but an attitude of scientific inquiry combined with a belief that the torture of millions is not desirable, whether inflicted by Stalin or by a Deity imagined in the likeness of the believer."

– Bertrand Russell

Declaration This dissertation is the result of work carried out by myself between October 2002 and December 2006. It includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It contains approximately 56,900 words and does not exceed the limit of 65,000 words. It contains 10 figures, which does not exceed the limit of 150.

Signed:

Adam Johansen

Acknowledgements There are a great many people without whom the writing of this thesis would not have been possible.

I would like to thank my supervisor, Bill Fitzgerald. I would also like to extend my deepest gratitude to Arnaud Doucet, with whom I have worked closely, and who has been tremendously helpful throughout my Ph.D..

Thanks are due to all those in the signal processing group in Cambridge for many stimulating discussions over coffee – and to the group for providing the coffee itself. I would especially like to thank in no particular order: Matthew Orton, Gareth Peters, Edmund Jackson, Jonathon Cameron, Dave Rimmer, Frédéric Desobry, James Ng, Ryan Anderson, Mark Miller, John Earl, Maurice Fallon, Nick Whitely and Mark Briers who were always ready to discuss the most obscure of issues.

For their extraordinary patience, I must thank my proofreaders, especially, Louise Birch, Edmund Jackson and Nick Whitely whose tireless efforts have improved the presentation of this work significantly.

Outside of Cambridge, I must offer thanks to Pierre Del Moral for making me feel extremely welcome in Nice and to those that made my visits to Vancouver a pleasant and fruitful one – especially Luis Montesano, Nando de Freitas and Raphael Gottardo. I also wish to thank Manuel Davy for many stimulating discussions.

I would also like to thank my parents and Louise Birch for their constant support, without which this would never have been written.

Summary

Over the course of the past five decades Monte Carlo methods have progressed from their infancy to a set of powerful and widely used computational tools. Scarcely any field which involves complex integration or optimisation problems has not been influenced by this rapid development. The end of the twentieth century was marked by an explosive development in Markov chain Monte Carlo methods. In recent years, there has been a great deal of development in the field of population-based Monte Carlo methods. These include a number of developments of the sequential Monte Carlo methodology (which has been used as a technique for the approximate solution of the optimal filtering problem for somewhat longer) allowing it to be applied to a much broader range of sampling problems.

This thesis comprises three novel contributions to the field. The first of these is somewhat theoretical in character: it is proven that one particular piece of methodology (the sequential Monte Carlo implementation of the probability hypothesis density filter) which has recently been proposed in the literature converges almost surely, and obeys a central limit theorem with a particular variance expression.

The other contributions are of a methodological nature. One of which is to develop algorithms for maximum likelihood estimation for latent variable models using a population-based sampling technique. This approach allows for the – possibly unknown – marginal likelihood to be maximised via a data augmentation strategy. The final contribution is a method by which rare event probabilities can be estimated using another population-based simulation technique employing a sequence of artificial distributions. Illustrative examples of these techniques are also presented – as are comparisons to alternative approaches, where this is appropriate.

Table of Contents

List of Abbreviations

"Brevity is the soul of lingerie."

– Dorothy Parker

a.e.	*almost every*
a.s.	*almost surely*
AIS	Annealed Importance Sampling
AMS	Adaptive Multi-level Splitting
APF	Auxiliary Particle Filter
BDMCMC	Birth-and-Death Markov Chain Monte Carlo
cdf	cumulative distribution function
CFTP	Coupling From The Past
CLT	Central Limit Theorem
EM	Expectation Maximisation
ESS	Effective Sample Size
HMM	Hidden Markov Model
iid	independent, identically distributed
IPS	Interacting Particle System
MAP	Maximum a Posteriori
MCEM	Monte Carlo EM
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings
ML	Maximum Likelihood
MMAP	Marginal Maximum a Posteriori
MML	Marginal Maximum Likelihood
pdf	probability density function
pgfl	probability generating functional
PHD	Probability Hypothesis Density
PMC	Population Monte Carlo
RESTART	Repetititive Simulation Trials After Reaching Thresholds
RFS	Random Finite Set
RJMCMC	Reversible Jump Markov Chain Monte Carlo
SA	Simulated Annealing
SAEM	Stochastic Approximation EM
SAME	State Augmentation for Marginal Estimation
SEM	Stochastic EM
SIR	Sequential Importance Resampling
SIS	Sequential Importance Sampling
SISR	Sequential Importance Sampling / Resampling
SLLN	Strong Law of Large Numbers
SMC	Sequential Monte Carlo
s.t.	such that

List of Algorithms

List of Figures

1. Introduction

"We could, of course, use any notation we want; do not laugh at notations; invent them, they are powerful. In fact, mathematics is, to a large extent, invention of better notations."

– Richard P. Feynman, "The Feynman Lectures on Physics"

1.1 Context

In recent years Monte Carlo methods have become a widely used and powerful tool in many fields from statistics to biology, finance and even computer games. Chapter 2 summarises many of the main developments in this field, particularly where they are relevant to the work presented in this thesis. Many of the ideas which were adopted and widely used in the last few decades were originally proposed in the 1950s. However, the development of fast, cheap computers was an obvious prerequisite for the wide applicability of such computationally intensive methods. Now that such computing power is widely available there remains work to be done on the development of efficient and widely applicable algorithms. It is to this end that this thesis has been largely concerned with the development of novel algorithms which can be applied to, and might be expected to perform well in, a wide variety of scenarios with minimal problem-specific "tuning" being required.

We have also been concerned with addressing certain theoretical problems associated with existing algorithms to provide some insight into the situations in which such algorithms might perform well and to provide some guidance to applications specialists about the use of such algorithms in real-world problems. Particularly, convergence results and central limit theorems provide guidance about the circumstances in which algorithms converge and asymptotic variance expressions provide useful information about the design of particular algorithms.

1.2 Notation

It is useful to summarise the notation which is used throughout this thesis before becoming involved in the details. Wherever possible, notation is used consistently throughout the thesis, although the particular requirements of certain chapters are such that there is inevitably some conflict between this requirement and the

desire to be consistent with standard usage within the literature. Wherever there is any likelihood of confusion this has been indicated in the text.

$\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ are used to denote the fields of natural numbers, integers and real numbers, respectively and the $+$ subscript is used to denote the *non-negative* subsets of these fields where appropriate. Given a real quantity, $x \in \mathbb{R}$, we define the floor, ceiling and remainder of x as:

$$\begin{aligned} \lfloor x \rfloor &\triangleq \sup\{y \in \mathbb{Z} \ : y \leq x\}, \\ \lceil x \rceil &\triangleq \inf\{y \in \mathbb{Z} \ : y \geq x\}, \\ \text{and } x^\sharp &\triangleq x - \lfloor x \rfloor. \end{aligned}$$

We make use of $x \wedge y$ and $x \vee y$ to denote the minimum and maximum, respectively, of x and y. The cardinality of a set, A, is denoted $|A|$; that of an explicitly defined set by $\#\{a_1, \ldots, a_N\} = N$. The empty set is denoted $\emptyset$ and we adopt the usual conventions for the sum and product of any function over it ($\sum_{\emptyset} = 0$ and $\prod_{\emptyset} = 1$), and its upper and lower bounds ($\inf_\emptyset = \infty$ and $\sup_\emptyset = -\infty$).

Allow $x_{p:q}$ to denote the vector comprising components $x_p, x_{p+1}, \ldots, x_q$. Given a vector, $x = x_{1:d}$ we use $x_{-k} = x_{1:k-1,k+1:d}$ to refer to all but the k^{th} element of that vector.

Where it is necessary to describe matrices in terms of their components, we write $A = [a_{ij}]$ where a_{ij} is the expression for component i, j of matrix A. Given a general measurable space $(E, \mathcal{E})$ we refer to the set of all σ-finite measures on that space, as $\mathcal{M}(E)$. The set of all probability measures on $(E, \mathcal{E})$ is denoted $\mathcal{P}(E) \subset \mathcal{M}(E)$, and the Dirac measured located at e is denoted $\delta_e(\cdot)$. We denote the Banach space of bounded measurable functions on E (endowed with the supremum norm, $||\xi||_\infty = \sup_{u \in E} |\xi(u)|$ for any $\xi : E \to \mathbb{R}$), $\mathcal{B}_b(E)$. On any space, $\mathbf{0}, \mathbf{1}$ and Id denote the zero function, unit function and identity operator, respectively.

Given a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ and an $\mathcal{E}/\mathcal{F}$-measurable random variable X, we allow $\mathbb{P} \circ X^{-1}$ to denote the measure on $(E, \mathcal{E})$ corresponding to the law of X. For example, given a random process $(X_n)_{n \in \mathbb{N}}$ we denote the law of the first N elements $\mathbb{P} \circ X_{1:N}^{-1}$. That is, X^{-1} denotes the inverse image associated with the random variable X.

Give a second general measurable space $(F, \mathcal{F})$, we define a kernel from E to F, K, to be a function $K : E \times \mathcal{F} \to \mathbb{R}_+$ such that:

$$\begin{aligned} \forall e \in E, K(e, \cdot) \in \mathcal{M}(F), \\ \text{and } \forall df \in \mathcal{F}, K(\cdot, df) \in \mathcal{E}. \end{aligned}$$

Such a kernel induces two operators, one on $\mathcal{M}(E)$:

$$\mu K(\cdot) = \int_E \mu(de) K(e, \cdot) \forall \mu \in \mathcal{M}(E),$$

and one on the $\mathcal{F} - measurable$ functions on F:

$$K(\xi)(\cdot) = \int_F K(\cdot, df)\xi(f) \forall \xi \in \mathcal{B}_b(F).$$

When we wish to consider the joint distribution induced over $(E, \mathcal{E}) \times (F, \mathcal{F})$ by a measure π on $(E, \mathcal{E})$ and such a kernel we use the notation $\pi \otimes K(de, df) = \pi(de)K(e, df)$ and given a sequence of measurable spaces $(E_n, \mathcal{E}_n)_{n\in\mathbb{N}}$ and a collection of kernels $K_p : E_{p-1} \times \mathcal{E}_p \to \mathbb{R}_+$, we write

$$K_{p:q}^{\otimes}(e_p, de_{p+1:q}) = \prod_{j=p+1}^{q} K_j(e_{j-1}, de_j).$$

We note that this allows us to relate the tensor product of kernels to their convolution in a concise manner:

$$\forall x_p \in E_p, A_q \in \mathcal{E}_q, \quad K_{p:q}(x_p, A_q) = \int K_{p:q}^{\otimes}(x_p, dx_{p+1:q-1} \otimes A_q).$$

The terms *Markov kernel* and *transition kernel* will be reserved for those kernels with the additional property that

$$\forall e \in E, K(e, \cdot) \in \mathcal{P}(F).$$

Given a kernel from E to E, we define the n-fold application of that kernel inductively as:

$$K^n(e, de') = K^{n-1}K(de')(e); K^0(e, de') = \delta_e(de').$$

The following notations are used to describe various probability distributions: $\mathcal{B}er(p)$ describes the Bernoulli distribution with success probability p, $\mathcal{D}i(\alpha)$ the Dirichlet distribution with parameter vector α, $\mathcal{N}(\mu, \sigma^2)$ describes a normal of mean μ and variance σ^2, $\mathcal{G}a(\alpha, \beta)$ a gamma distribution of shape α and rate β, $\mathcal{IG}(\alpha, \beta)$ the inverse gamma distribution associated with $\mathcal{G}a(\alpha, \beta)$, $\mathcal{L}ogistic(\mu, s)$ the logistic distribution with location μ and scale s and $\mathcal{KS}$ refers to the Kolmogorov-Smirnov distribution. The measures and densities associated with these measures are indicated in the same way, with an additional argument separated from the parameters by a semi-colon.

Finally, we note that the thesis is written first person plural in accordance with the conventions of technical literature. No intention to indicate collaboration or the involvement of other parties is attached to this usage, and collaborative work is indicated explicitly within the text where necessary.

1.3 Outline

This thesis is concerned with some recent developments in the theory and methodology of Sequential Monte Carlo (SMC). It begins, in the next chapter, with a

survey of the Monte Carlo literature before moving on to the novel contributions of this thesis, which are:

– Chapter 3: provides an asymptotic analysis of the SMC implementation of the Probability Hypothesis Density (PHD) filter, including convergence of the particle approximation and a central limit theorem.
– Chapter 4: presents a novel SMC algorithm for obtaining marginal parameter estimates.
– Chapter 5: introduces a novel SMC approach to the estimation of the probability of rare events.

Finally, some potentially interesting areas of further work are proposed.

2. Monte Carlo Methods

"Fortunately, the future is unpredictable and also – because of quantum effects – uncertain."

– Andrei Dmitrievich Sakharov

We shall take *Monte Carlo Methods* to be the class of algorithms which fundamentally make use of random samples from some distribution to achieve their result. In practice, of course, almost all computer simulations make use of pseudo-random number generators, we shall not concern ourselves with that subtlety. Typical applications include approximateintegration of some function with respect to that distribution, or the optimisation of a related function. This is an area with a rich literature, with the recent interest in its use with modern computing machines going back at least as far as [112] and we cannot hope to provide a comprehensive discussion here. An excellent, and recent, book length review of the subject is provided by [130]; historical commentaries on the early days of the Monte Carlo method at Los Alamos National Laboratory are provided by [50, 110]. It is the intention of this section to provide an overview of the major methods in use at present, particularly those which the work presented later depends upon, and to set them in context.

By way of motivation, we note that Monte Carlo methods – which a few decades ago were of only specialist interest – are now one of the most broadly used computational techniques. Indeed, the Metropolis algorithm has been named one of the ten most influential algorithms of the twentieth century by the editors of *Computing in Science and Engineering* [44] – a more detailed commentary on their choices was presented by [26]. It would not be possible to give any meaningful overview of the areas in which such methods have found applications within the last few decades, but these include signal processing [135], mechanical engineering [22, 23], communications [129], statistics [19, 142], finance [61, 68, 70], many areas of physics including optics [9, 115, 56], cosmology [143, 45] and the analysis of spin glasses [120, 6], biology [103, 53], chemistry [10] and others.

We shall consider only Monte Carlo integration, although many of the same principles apply equally to Monte Carlo optimisation. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we allow π to describe the law of some random variable $X : \Omega \to E$.

Throughout this section, π shall denote a probability measure on a measurable space $(E, \mathcal{E})$ with respect to which we wish to integrate a measurable function $\xi : E \to \mathbb{R}$

2.1 Perfect Monte Carlo

The terms *perfect*, *naïve* and *crude* Monte Carlo simulation all refer to those methods which involve obtaining a set of independent, identically distributed (*iid*) samples from the distribution of interest and using these samples to perform integration of a function with respect to that distribution. It is this approach which was originally referred to as *the Monte Carlo method* by [112] but in the intervening decades the term has come to encompass a much broader class of approaches. The generation of random variates themselves is a complex subject. Many approaches to the generation of uniform bits are described in [95, chapter 3]; a wide variety of approaches to obtaining non-uniform random variates are considered in [43]. For the purposes of this thesis, it is assumed that it is possible to draw the samples which are required by the approaches described.

When a set of *iid* samples $\{X_i\}_{i=1}^N$ drawn from π is available, the Monte Carlo estimator of the integral of the function ξ under that measure can be expressed as:

$$\hat{I}(\xi) := \frac{1}{N} \sum_{i=1}^{N} \xi(X_i), \tag{2.1}$$

which may alternatively be interpreted as the integral of ξ under the empirical measure of the sample set, $\hat{\pi}_{MC}^N$:

$$\hat{\pi}_{MC}^N(\cdot) := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}(\cdot) \quad \hat{I}(\xi) = \hat{\pi}_{MC}^N(\xi). \tag{2.2}$$

The Strong Law of Large Numbers (SLLN) for *iid* random variables (see, for example, [140, p. 391]) ensures the almost sure convergence of this estimator to the true value as the number of samples tends to infinity, provided only that they have finite expectation,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \xi(X_i) \overset{a.s.}{\to} \mathbb{E}\left[\xi(X_1)\right], \tag{2.3}$$

and the central limit theorem (CLT) further ensures that the distribution of the estimator is well behaved in the same limit [140, p. 335], providing the variance is finite,

$$\lim_{N\to\infty} \sqrt{N} \left(\hat{\pi}_{MC}^N(\xi) - \pi(\xi)\right) \overset{d}{\to} \mathcal{N}\left(0, \mathrm{Var}_\pi\left(\xi\right)\right). \tag{2.4}$$

There are a number of methods for obtaining *iid* samples from distributions from which it is not possible to sample directly. As these methods underly the

more sophisticated methods upon which we will later rely, some of the common approaches are summarised below.

Inversion Sampling. If π is a distribution over the real numbers which admits a density with respect to Lebesgue measure, and it is possible to invert its cumulative distribution function (*cdf*), F, then it is possible to transform a sample, U, from a uniform distribution over $[0, 1]$ into a sample, X, from π by making use of the following transformation:

$$X = F^{-1}(U).$$

Actually, it suffices to obtain a *generalised inverse* of F, a function with the property that, $F^{-1}(U) = \inf_x \{F(x) \geq U\}$ as the image of the set of points over which the true inverse is multi-valued is π-null. More details are provided by [130] and [43, chapter 1] which also discusses some extensions to the method. This approach had been considered by Ulam prior to 1947 [50].

2.1.1 Rejection Sampling

Another approach is available if there exists a distribution, μ, from which it is possible to obtain samples, and with respect to which π is absolutely continuous with bounded Radon-Nikodým derivative. In this case, we can simply draw samples, X, from μ and accept them as samples from π if an independent sample, U, from a uniform distribution over $[0, 1]$ lies below $\frac{1}{M}\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X)$ for some majorising constant, $M \geq \sup_x \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)$. Algorithm 2.1 gives a formal description of the algorithm. Intuitively, this approach simply makes use of the fact that sampling uniformly from the area beneath the probability density function (*pdf*) (where it exists) of a distribution and discarding the irrelevant coordinate provides samples from the distribution itself (a result known as the fundamental theorem of simulation). The first suggestion of this technique appears to have been in a 1947 letter from Von Neumann to Ulam [50].

Algorithm 2.1 Rejection Sampling

Ensure: $M \geq \sup_x \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)$
1: Sample $X \sim \mu$
2: Sample $U \sim \mathcal{U}[0, 1]$
3: **if** $MU \leq \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X)$ **then**
4: Accept X as a sample from π
5: **else** {Reject this sample}
6: Go to step 1
7: **end if**

The expected proportion of samples which are accepted are given by:

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_{[0,\frac{1}{M}\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X)]}(U)\right] &= \int \frac{1}{M}\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)\mu(dx) \\ &= \frac{1}{M}\end{aligned}$$

which makes it clear that this approach can only be efficient when a distribution close to π is available for use as a sampling distribution – otherwise there will be regions in which the Radon-Nikodým derivative is large: consequently, a large value of M is required, and many samples will be discarded for every one which is retained. The correctness of the method is easy to verify by considering the distribution of *accepted* samples:

$$\begin{aligned}\mathbb{P}\left(X \in dx \,|MU \leq \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X)\right) &= \frac{\mathbb{P}\left(X \in dx, MU \leq \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X)\right)}{\mathbb{P}\left(MU \leq \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X)\right)} \\ &= \mu(dx)\frac{1}{M}\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)/\frac{1}{M} \\ &= \pi(dx).\end{aligned}$$

[109] noted that it is possible to obtain greater computational efficiency if there exist cheap-to-evaluate functions which tightly bound the density of π with respect to μ, and termed such bounds *squeezing functions*. More recently, a mechanism for adaptively obtained such bounds from the samples themselves, leading to a sampling technique termed *adaptive rejection sampling*, has been proposed in the case of log-concave densities [67, 62] (where it is possible to bound the log density by considering functions which are piecewise linear between those points at which it has already been necessary to evaluate the function) and subsequently for general densities [65] (although the latter makes use of techniques from the field of Markov chain Monte Carlo, described below, and does not produce independent samples).

2.2 Importance Sampling

Rejection sampling seems in some senses rather wasteful, as it necessarily discards some samples without making any use of them. Importance sampling, in contrast, makes use of every sample but *weights* each one according to the degree of similarity between the target and instrumental distributions. Algorithm 2.2 describes the approximation of $\pi(\xi)$ using N samples from μ. This is essentially an application of the result which [130] terms the *importance sampling fundamental identity*: $\pi(\xi) = \mu\left(\xi\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)$ provided that $\pi << \mu$.

In many interesting cases, the target measure is known only up to a normalising constant. In this case, a common strategy is to divide the *unnormalised* estimate of the quantity of interest by the unnormalised estimate of the integral of the

Algorithm 2.2 Importance Sampling

1: **for** $i = 1$ to N **do**
2: Sample $X_i \sim \mu$
3: Set $W_i = \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X_i)$
4: **end for**
5: $\hat{\pi}^N_{IS,1}(\cdot) = \frac{1}{N}\sum_{i=1}^N W_i \delta_{X_i}(\cdot) \Rightarrow \hat{\pi}^N_{IS,1}(\xi) = \frac{1}{N}\sum_{i=1}^N W_i \xi(X_i)$

unit function. The integral of the unit function under an unnormalised measure, is exactly the normalising constant required to turn that measure into a probability measure. This strategy, as it makes use of the ratio of two unbiased estimators, introduces a bias for finite samples. It is, however, asymptotically consistent, and can provide lower variance estimates than standard importance sampling [130]. Algorithm 2.3 provides a formal description of this approach.

Algorithm 2.3 Self-Normalised Importance Sampling

1: **for** $i = 1$ to N **do**
2: Sample $X_i \sim \mu$
3: Set $W_i = Z\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X_i)$, where Z is some unknown normalising constant
4: **end for**
5: $\hat{\pi}^N_{IS,2}(\cdot) = \sum_{i=1}^N W_i \delta_{X_i}(\cdot) / \sum_{i=1}^N W_i \Rightarrow \hat{\pi}^N_{IS,2}(\xi) = \sum_{i=1}^N W_i \xi(X_i) / \sum_{i=1}^N W_i$

This approach is justified by a SLLN and Central Limit Theorem (CLT) in the same way as perfect Monte Carlo [61].

It has been noted that rejection sampling can be interpreted as importance sampling on the space $\Omega \times [0,1]$ on which $X \times U$ are distributed using a particular importance function [20]. The same study demonstrates that importance sampling has a lower variance than rejection sampling using the importance function as the proposal distribution, with a suitable majorising constant (when one considers using N samples from that proposal distribution to estimate the integral of a function under that distribution). As this argument shows that importance sampling is essentially a Rao-Blackwellised [130, section 4.2] version of rejection sampling, this is what would be expected.

A comparison between the two techniques is also provided by [130, section 3.3.3], who consider a slightly difference case. Their analysis considers drawing n samples by rejection sampling, and producing in the process a random number of samples which are used within the importance sampling estimator, which introduces a stopping-time. This makes the comparison more complex, but it is possible to say that there exists an instrumental distribution from which importance samples can be drawn which will lead to an estimator which dominates the rejection sampling case. This is perhaps a more relevant comparison, although for large n one would expect the differences to be negligible. However, if one actually requires

iid samples from π, rather than an approximation to an integral, then importance sampling cannot be used. Further details concerning Rao-Blackwellisation in general, and in the particular case of rejection sampling, can be found in [15]. Their Rao-Blackwellised version of rejection sampling is slightly more subtle than the traditional importance sampling estimator, as it takes into account the stopping time corresponding to accepting the n^{th} sample.

It has also been demonstrated that the rejection sampling estimator is dominated by one which makes use of the rejected samples to produce an unbiased estimator of zero which has negative covariance with the rejection sampling estimator [121].

2.3 Markov Chain Monte Carlo (MCMC)

In many cases it is extremely difficult to obtain large numbers of *iid* samples from a distribution of interest. The principle behind Markov Chain Monte Carlo (MCMC) methods is that, if one can construct an ergodic Markov chain which has π as its stationary distribution, then the samples obtained from a *sufficiently long* simulation of that Markov chain will correspond to a large set of dependent samples from π. This is a complex and interesting area which extends far beyond the scope of this thesis. An extremely good review of the theory of Markov chains is given by [113], or more briefly by [119]; more approachable introductions to the area are provided by [130, chapter 6] and [132, 146], and a good, if slightly dated, reference for the application of MCMC methods is [66]. Many more recent developments are summarised in [130].

It has recently been demonstrated that it is possible to consider a set of weighted samples as a jump-Markov process with sojourn times corresponding to their weights [108]. This is an exciting development which should in time lead to the transfer of many results between the rich literatures of Markov Chains and importance sampling.

2.3.1 Discrete Time Markov Chains

It is not possible to completely avoid reference to the theory of Markov Chains, whilst adequately summarising MCMC. This section contains a few essential concepts which motivate and justify the techniques which are described below. Simulation algorithms typically make use of discrete time Markov chains, and the continuous time case (also referred to as *Markov processes*) will not be considered. We assume, without loss of generality, that the index set is $\mathbb{N}$.

Consider a (possibly inhomogeneous) Markov chain, $(X_n)_{n\in\mathbb{N}}$, which takes its values in a sequence of measurable spaces $(E_n, \mathcal{F}_n)_{n\in\mathbb{N}}$ with initial distribution

η_0 and elementary transitions given by the set of Markov kernels $(M_n)_{n\geq 1}$, $M_n : E_{n-1} \to \mathcal{P}(E_n)$, where $\mathcal{P}(E_n)$ denotes the class of probability measures on space E_n, i.e. the canonical Markov chain:

$$\left(\Omega = \prod_{n=0}^{\infty} E_n, \mathcal{F} = (\mathcal{F}_n)_{n\geq 0}, (X_n)_{n\in\mathbb{N}}, \mathbb{P}_{\eta_0}\right), \tag{2.5}$$

where $\{\mathcal{F}_n\}$ denotes the natural filtration of $\{X_n\}$ and the law $\mathbb{P}_{\eta_0}$ of the chain, with initial distribution η_0, is defined by its finite dimensional distributions:

$$\mathbb{P}_{\eta_0} \circ X_{0:N}^{-1}(dx_{0:N}) = \eta_0(dx_0) \prod_{i=1}^{N} M_i(x_{i-1}, dx_i). \tag{2.6}$$

The defining property of a Markov chain is the *(weak) Markov property*: for any deterministic times $m \leq n$, for any $\eta_0 \in \mathcal{P}(E_0)$:

$$\mathbb{P}_{\eta_0}(X_{m:n} \in dx_{m:n} | X_{0:m-1}) = \mathbb{P}_{\eta_0}(X_{m:n} \in dx_{m:n} | X_{m-1}).$$

For the remainder of this section we shall consider only time homogeneous Markov chains for which $E_i = E$ and $M_i = M$ at all times. Such Markov chains are sufficient for the purposes of considering MCMC algorithms, but the full generality of the above definition is required when we come to consider the Feynman-Kac flows which underpin the theory of many other methods. It is useful to consider certain fundamental properties of Markov chains, as they apply to the homogeneous case. Here, the *strong Markov property* extends its weak counterpart to include finite stopping times, so that, for any $\mathbb{P}$-*almost surely* (a.s.) finite stopping time T, and any function $\xi : E^{\infty} \to \mathbb{R}$, whenever the expectations exist:

$$\mathbb{E}(\xi(X_{T+1}, X_{T+2}, \dots) | X_{1:T}) = \mathbb{E}(\xi(X_{T+1}, X_{T+2}, \dots) | X_T), \ \mathbb{P} - a.s..$$

The strong Markov property holds for all discrete time Markov chains [113], and this can be straightforwardly proved by conditioning upon the possible values of the stopping time and applying the weak Markov property.

Irreducibility. Loosely speaking, a Markov chain is irreducible if (almost) all states communicate; the property corresponds to the existence of a path of positive probability from (almost) any point in the space to (almost) any measurable set. A Markov chain of the sort described is ψ-irreducible for some $\psi \in \mathcal{P}(E)$ if, for *ψ-almost every* (a.e.) $A \in \mathcal{E}$, the following holds:

$$\exists n \in \mathbb{N} \text{ s.t. } \forall e \in E, M^n(e, A) > 0.$$

The term *strongly ψ-irreducible* is used to refer to chains for which this holds with $n = 1$. Note that, in the discrete state space case, it is possible to term a chain *irreducible* if there is a finite probability of moving from any state to any other state in finite time, i.e. $\forall x, y \in E, \mathbb{P}_{\delta_x}(\inf\{n : X_n = y\} < \infty) > 0$. This concept clearly does not generalise to continuous state spaces.

Aperiodicity. In the discrete state space case, a Markov chain is aperiodic if there exist no cycles of length greater than one, where a cycle is defined as the greatest common denominator of the length of all routes of positive probability between two states. In the general case, a little more care is required. The introduction of so-called *small sets* provides a suitable analogue for the individual states of the countable state space case, see for example [119, chapter 2] or [113, chapter 5]. In essence a small set is one from which a minorisation condition holds, i.e. a set C is small if,

$$\forall x \in C \exists m \in \mathbb{N}, \delta \in \mathbb{R}_+, \nu \in \mathcal{M}(E) \text{ s.t. } \forall B \in \mathcal{E} : M^m(x, B) \geq \delta\nu(B). \tag{2.7}$$

A Markov chain has a cycle of length d if, there exists a small set C, for which:

$$d = \gcd\left\{m \geq 1 : \forall x \in C, B \in \mathcal{E} \exists \delta_m > 0, \nu_m \in \mathcal{M}(E) \text{ s.t. } M^m(x, B) \geq \delta_m \nu_m(B)\right\}. \tag{2.8}$$

In full generality, if the longest cycle associated with a Markov chain is of length one then that chain is aperiodic.

Recurrence. A recurrent Markov chain is, roughly, one which is expected to visit every important state infinitely often. A Markov chain of the form described above is *recurrent* if there exists some $\psi \in \mathcal{P}(E)$ for which it is ψ-irreducible and $\mathbb{E}[\#\{X_i \in A\}] = \infty$ for ψ-*a.e.* A.

In considering the convergence of MCMC algorithms, a stronger form of recurrence is useful. A set A is *Harris recurrent* if $\forall e \in E, \mathbb{P}_{\delta_e}(\#\{X_i \in A\} = \infty) = 1$; a ψ-irreducible chain is Harris recurrent if ψ-*a.e.* set A is Harris recurrent. This form of recurrence was shown to be sufficient to guarantee the existence of a unique invariant distribution for the chain [76].

Ergodicity. The ergodic theorem [113, chapter 13] tells us that, for any Harris recurrent Markov chain, $\{X_n\}$, with stationary distribution π, and any $\mathbb{L}_{1,\mu}$-integrable function ξ:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \xi(X_i) = \int \xi(x)\pi(dx).$$

Numerous theoretical results from the theory of general state space Markov chains can be applied to the particular cases employed in MCMC. It is neither feasible nor desirable to summarise them here, however, [145] provides a summary of some of the more useful convergence results with particular emphasis on the Metropolis-Hastings algorithm. See [119, chapter 7] for limit theorems of Harris recurrent Markov Chains and [113, part III] for various forms of convergence.

2.3.2 Metropolis-Hastings (MH)

The seminal paper of Metropolis *et al.* [111] is widely regarded as being the first in this field, and introduced the principal ideas behind the most widely used MCMC algorithm to date, the Metropolis-Hastings (MH) algorithm. The key feature of the algorithm is that it provides a mechanism for taking a symmetric proposal kernel and producing a new Markov kernel which has the desired stationary distribution. An alternative approach (which differs only in the form of the acceptance probability) was proposed by [7]. The principal innovations of Hastings [77] were to modify the methodology to permit asymmetric proposal distributions and to generalise the form of the acceptance probability in such a way as to include both of the earlier algorithms as special cases. The approach of Metropolis, as extended to asymmetric proposal distributions by Hastings, was shown to be optimal, at least in the case of discrete state spaces, in the sense of asymptotic variance [122]. Algorithm 2.4 describes the procedure for obtaining a sequence of samples from a Markov chain of invariant distribution π by making use of a Markov transition kernel K, both of which are assumed to admit a density with respect to a suitable dominating measure, λ. For simplicity, we use the same symbols to refer to the densities and their associated measures throughout the remainder of this section. A more detailed historical survey is provided by [78].

Algorithm 2.4 Metropolis-Hastings

1: **repeat**
2: Sample $\hat{X}_t \sim K(X_t, \cdot)$.
3: Calculate the MH acceptance probability, $\alpha = \min\left(1, \frac{\pi(\hat{X}_t)K(\hat{X}_t, X_t)}{\pi(X_t)K(X_t, \hat{X}_t)}\right)$.
4: Sample $U \sim \mathcal{U}[0, 1]$.
5: **if** $U \leq \alpha$ **then**
6: Accept this proposal, and set $X_{t+1} = \hat{X}_t$.
7: **else**
8: Reject this proposal, and set $X_{t+1} = X_t$.
9: **end if**
10: **until** Sufficiently many samples have been obtained.

As mentioned above, the key feature of this algorithm is that it produces a Markov kernel which has the desired invariant distribution. Consider the Markov transition kernel, $M : E \to \mathcal{P}(E)$ under whose influence the sample sequence, X_t, described in algorithm 2.4 evolves:

$$\begin{aligned} M(x, dy) &= \alpha(x, y)K(x, y)\lambda dy + (1 - \alpha(x, y))\delta_x(dy) \\ &= \left[1 \wedge \frac{\pi(y)K(y, x)}{\pi(x)K(x, y)}\right] K(x, dy) + \left[0 \vee 1 - \frac{\pi(y)K(y, x)}{\pi(x)K(x, y)}\right] \delta_x(dy). \end{aligned}$$

It can easily be shown that this kernel satisfies the detailed balance condition for π. Consider the case in which $\alpha(x, y) \leq 1$:

$$\begin{aligned}\pi(x)M(x,y) &= \pi(x)\left[\frac{\pi(y)K(y,x)}{\pi(x)K(x,y)}K(x,y) + \left(1 - \frac{\pi(y)K(y,x)}{\pi(x)K(x,y)}\right)\delta_x(y)\right] \\ &= \pi(y)\left[K(y,x) + \left(\frac{\pi(x)}{\pi(y)} - \frac{K(y,x)}{K(x,y)}\right)\delta_x(y)\right] \\ &= \pi(y)K(y,x) = \pi(y)M(y,x).\end{aligned}$$

where the final equality holds because $\alpha(x, y) \leq 1 \Rightarrow \alpha(y, x) = 1$. Note that result must also hold in the converse case by symmetry.

Further, it is clear that detailed balance for π is sufficient to make π a stationary measure for M:

$$\pi(x)K(x,y) = \pi(y)K(y,x)$$
$$\int \pi(x)K(x,y)\lambda(dx) = \int \pi(y)K(y,x)\lambda(dx) = \pi(y).$$

It is precisely this property which makes single MH steps integral components of many other Monte Carlo algorithms: it provides a mechanism for introducing diversity into a set of samples from the target distribution without changing the distribution of those samples.

2.3.3 Gibbs Sampling

The Gibbs sampler [59] produces a Markov Chain by updating one component of the state vector during each iteration. The value of each element at time t is sampled from the distribution of that element conditional upon the values of all the other parameters at time $t-1$ and those parameters which have already been update at time t. An applied introduction the Gibbs sampler is provided by [103, chapter 6]; a more theoretical approach is taken by [130, chapters 8-10].

The original Gibbs sampler updated each parameter in sequence, as described in algorithm 2.5. Consequently, the deterministic-scan Gibbs update is not reversible, although each individual component of it is. Another commonly used approach is the random scan Gibbs sampler, which is described in algorithm 2.6.

Algorithm 2.5 Gibbs Sampler (Deterministic-Scan)

1: {This is a single step of the deterministic Gibbs Sampler}
2: Given a sampler X_{t-1} from π (which contains D components)
3: **for** $d = 1$ to D **do**
4: Sample $X_{t,d} \sim \mu\left(\cdot|\sigma\left(X_{t,1}, \ldots X_{t,d-1}, X_{t-1,d+1}, \ldots, X_{t-1,D}\right)\right)$
5: **end for**

Algorithm 2.6 Gibbs Sampler (Random Scan)

1: {This is a single step of the random scan Gibbs Sampler}
2: Given a sampler X_{t-1} from π (which contains D components)
3: Sample, from a suitable distribution (typically one which is uniform over all permutations of $\{1,\ldots,D\}$), a set of indices $\{n_1,\ldots,n_D\}$ which is isomorphic to $\{1,\ldots,D\}$.
4: **for** $d = 1$ to D **do**
5: $\quad X_{t,n_d} \sim \mu\left(\cdot|\sigma\left(X_{t,n_1},\ldots X_{t,n_{d-1}}, X_{t-1,n_{d+1}},\ldots,X_{t-1,n_D}\right)\right)$
6: **end for**

Gibbs sampling is often viewed as an algorithm in its own right. It is, however, simply a special case of the Metropolis-Hastings algorithm in which a single component is updated during each step, and the proposal distribution which is used is the true conditional distribution of that parameter given the present values of the others. That is, a single iteration of the Gibbs sampler corresponds to the application of D successive Metropolis-Hastings steps, with the relevant conditional distribution used as the proposal kernel for each step. The consequence of this special choice of proposal distribution is that the MH acceptance probability is always one, and rejection never occurs. This connection was recognised, to at least some degree, from the beginning, and [59] describes the algorithm as *a "heat bath" version of the Metropolis algorithm* and further notes the equivalence of the two approaches in particular cases in section 10.

It is straightforward to verify that the Metropolis-Hastings acceptance probability is uniformly one whenever the conditional distribution of the component being updated is used as the proposal distribution. Consider the MH acceptance probability associated with a move which updates component k of a D-component vector using a proposal distribution corresponding to its conditional distribution under the target measure given the present values of its other elements:

$$\begin{aligned}\alpha(x,(x_{1:k-1},y_k,x_{k+1:D}) &= \frac{\pi((x_{1:k-1},y_k,x_{k+1:D}))\pi(x_k|x_{-k})}{\pi(x)\pi(y_k|x_{-k})} \\ &= \frac{\pi(x_{-k})\pi(y_k|x_{-k})\pi(x_k|x_{-k})}{\pi(x_k|x_{-k})\pi(x_{-k})\pi(y_k|x_{-k})} = 1\end{aligned}$$

Although Gibbs sampling has been widely used, it has two significant weaknesses:

– its performance is very heavily dependent upon the parameterisation of the system being explored, as it is only able to make axis-aligned moves, and,
– it must be possible to sample from the conditional distributions of the target measure.

2.3.4 Reversible Jump MCMC

Reversible Jump Markov Chain Monte Carlo (RJMCMC) is essentially a mechanism by which the Metropolis-Hastings algorithm can be extended to allow the

exploration of a space comprising the disjoint union of subspaces of differing dimensions, a defining feature of which is that any proposed move between dimensions must be *reversible* [73, 75]. By comparing measures directly in this way it is possible to construct a Markov chain which explores the spaces of different dimension with the desired invariant distribution.

It is possible to relate RJMCMC to Birth-and-Death Markov Chain Monte Carlo (BDMCMC) [144] in which a continuous time jump-Markov process is simulated to produce a marked point process with the desired invariant distribution. It is interesting to note that the BDMCMC approach can be interpreted as a limit for RJMCMC [12]. This comparison notes no clear improvements due to the use of continuous time implementations, and suggests that their increased computational requirements more than outweigh potential advantages.

2.3.5 Perfect Sampling

Despite the similarities in their names, *perfect Monte Carlo* and *perfect sampling* refer to different things. A perfect sampling algorithm shall be taken to mean any one of several algorithms for making use of a Markov Chain to obtain *iid* samples from the target distribution. The MCMC techniques described above cannot be used to do this, as they would formally need to be allowed to iterate for an infinite time to obtain a single sample from the true distribution.

The difficulty with using MCMC to obtain *iid* samples from the target distribution is, essentially, that it is rarely possible to determine how long it takes for the chain to *forget* its initial conditions. One approach to overcoming this problem would be to construct the chain in such a way that it is possible to determine when a chain started from *any* position would have reached precisely the same state. This is essentially the Coupling From The Past (CFTP) algorithm proposed by [127] and later expanded upon in [128]. There are, however, some subtleties: it is not possible to consider an ensemble of *coupled*[1] Markov Chains initialised at every possible starting point evolving until they coalesce, as the distribution of the chain at this stopping time need not correspond to the stationary distribution. Indeed, it is possible to construct transition kernels for which the state of the chain at the point of coalescence is always the same!

The CFTP approach uses the intuition that under suitable conditions, a chain initialised at time $-\infty$ will correspond to a sample from the stationary distribution at time 0. It is possible to obtain samples from such a chain by considering starting chains at $T = -1$ and then checking for coalescence prior to time 0. If it is not

[1] Coupled, in the sense that their evolution is determined by the same set of random variables. For example, the transition made by each chain at each time being determined by inversion sampling make use of *the same* $\mathcal{U}[0, 1]$ random variable.

found, then the starting time is made earlier and the process repeated, ensuring that the same collection of random variables are used to control the evolution from time -1 onwards. The process is repeated until such a time as the chains have all coalesced by time 0 at which point, the past has been forgotten as chains started from anywhere have reached the same state *and* the final state is obtained at a deterministic time.

It is problematic that the time to coalescence of the chains considered in the CFTP algorithm is a (typically unbounded) random variable which is not independent of the final state of the chain. Consequently, bias is introduced by an implementation which places any upper bound upon the time taken for a sample to be obtained. The algorithm presented in [54] is intended to overcome this problem. The approach employed is related to rejection sampling, and constructs a chain *backwards* from an *arbitrarily selected* state at an arbitrary future time. Having done this, the values of the $\mathcal{U}[0,1]$ random variables which govern each transition are sampled, conditioned upon this path of the chain. If all possible starting points at time 0 would have coalesced to this state under the action of these random variables (and by construction, if they have coalesced then this must be the state at that time), then whatever state is reached by the backward chain at time 0 is a sample from the stationary distribution; otherwise rejection occurs and a new attempt must be made. This approach requires the selection of the future time which must be of the right magnitude to give a substantial acceptance probability without taking an unduly long time to provide each candidate sample.

It is a perennial problem in the field of MCMC that it is extremely difficult to determine how long it is necessary to allow the simulated chain to run for in order for the samples obtain to correspond to the stationary distribution. This step is completely avoided by the perfect sampling approach as *iid* samples from the target distribution are provided. A good tutorial on the subject is provided by [14]. However, attractive though this paradigm is, it suffers from two major difficulties: it is not straightforward to construct CFTP algorithms on infinite (or even large) state spaces unless they have some (partial) ordering *and* the transition kernel has a monotone structure and, even in cases where such chains can be constructed, the time to coalescence can be enormous, making each sample extremely expensive. At least at the present time, there seem to be many problems in which we are compelled to use either approximate samples from the distribution of interest or to employ importance sampling techniques.

2.4 Feynman-Kac Methods

Whilst it is something of a departure from the usual classification, it is useful to consider another family of Monte Carlo algorithms: those based upon measure-valued Feynman-Kac flows, rather than Markov chains or independent samples.

2.4.1 Discrete Time Feynman-Kac Flows

As with Markov chains, the general theory of Feynman-Kac flows falls far outside the remit of this thesis. An excellent monograph on the subject has recently been published [34], and contains a great many useful results ranging from comprehensive semi-group analyses to propagation of chaos estimates and central limit theorems. Some elementary aspects of the theory must be introduced here as it underpins the work done in several of the following chapters. Again, only discrete time flows will be considered here.

We begin with the canonical Markov chain, expression 2.5. In addition to the collection of Markov kernels, $\{M_n\}_{n\in\mathbb{N}}$ we employ a collection of potential functions $\{G_n\}_{n\in\mathbb{N}}$ where $G_n : E_n \to [0,\infty)$ and use these to define the law of a stochastic process in terms of that of the Markov chain. A given Feynman-Kac flow produces two closely related stochastic processes, the unnormalised *prediction flow*,

$$\mathbb{Q}_{\eta_0,n}(dx_{0:n}) = \frac{1}{Z_n}\prod_{i=0}^{n-1} G_i(x_i)\mathbb{P}_{\eta_0}\circ X_{0:n}^{-1}\left(dx_{0:n}\right), \tag{2.9}$$

and its associated *update flow*,

$$\begin{aligned}\hat{\mathbb{Q}}_{\eta_0,n}(dx_{0:n}) &= \frac{1}{\hat{Z}_n}\prod_{i=0}^{n} G_i(x_i)\mathbb{P}_{\eta_0}\circ X_{0:n}^{-1}\left(dx_{0:n}\right)\\ &= \frac{Z_n}{\hat{Z}_n}\mathbb{Q}_{\eta_0,n}(dx_{0:n})G_n(x_n),\end{aligned}$$

where Z_n and $\hat{Z}_n$ are normalising constants.

It is useful to associate four additional sets of distributions with this flow: the predictive and updated, unnormalised and normalised time marginals. The unnormalised distributions are defined weakly, for $\mathbb{L}_{1,\mathbb{Q}_{\eta_0,n}}$ functions $f_n : E_n \to \mathbb{R}$:

$$\gamma_n(f_n) = \int \prod_{i=0}^{n-1} G_i(x_i)f_n(x_n)\mathbb{P}_{\eta_0}\left(dx_{0:n}\right),$$

$$\text{and } \hat{\gamma}_n(f_n) = \int \prod_{i=0}^{n} G_i(x_i)f_n(x_n)\mathbb{P}_{\eta_0}\left(dx_{0:n}\right) = \gamma_n(f_nG_n),$$

whilst the normalised forms are, as one might expect:

$$\eta_n(f_n) = \frac{\gamma_n(f_n)}{\gamma_n(\mathbf{1})},$$
$$\text{and } \hat{\eta}_n(f_n) = \frac{\hat{\gamma}_n(f_n)}{\hat{\gamma}_n(\mathbf{1})}.$$

For our purposes it suffices, at this stage, to note that distributions of this form provide an invaluable framework for sequential estimation and integration by Monte Carlo methods. By considering the historical process, $X'_n = X_{1:n}$ associated with many sequential importance sampling type algorithms, one obtains a situation which can be described as the mean field approximation of such a flow. Indeed, this framework underlies, amongst other things, the whole of the sequential Monte Carlo approach – including the celebrated *particle filter*.

2.4.2 Sequential Monte Carlo (SMC)

Any Monte Carlo scheme which makes use of an Interacting Particle System (IPS) associated with the mean-field interpretation of a Feynman-Kac flow [34] – or a similar measure-valued flow – shall be considered to be a SMC method in this work. This is perhaps slightly more general than the usual interpretation, but includes the traditional definition, which is loosely the approximate solution of the optimal filtering equation by propagating the empirical measure associated with a set of particles through time according to the filtering recursion.

Filtering and Particle Filters. It is useful to look at the various particle methods for approximating the optimal filtering equations which have been derived over the past two decades and which comprise the standard SMC methods. Good tutorials on the subject are available, notably [48] and the book length review [47]. A more recent summary which includes some elements which are introduced in section 2.4.4 is [130, chapter 14].

We follow the formulation of [28], which includes a detailed proof that the recursion results presented here lead to the correct conditional measures in the case where the state and observation spaces correspond to finite dimensional real spaces. We consider the canonical Markov chain, $\{X_n\}_{n\geq 0}$ as introduced in section 2.3.1, which is termed the signal process, and an associated observation process $Y_n = h(n, X_n) + W_n$, $n > 0$ for some measurable function h and a sequence of independent random variables which are independent of X_n. W_n is assumed to admit a density g_n with respect to some measure $\lambda(dx)$. The *filtering problem* is to determine the conditional distribution of the signal process, given the σ-algebra generated by the observation process. Typically, it is necessary to do this sequentially in time in order to obtain the distribution of some signal variable given a sequence of observations obtained at times up to the present.

We remark that, strictly, in general, it is necessary to consider the expectations of regular functions integrated with respect to these measures as the *generalised Bayes theorem* provides only a weak description of these measures [140, p230]). Such a formulation is provided by many sources, particularly with reference to IPS in [29, 34].

It is convenient to consider the sequence of measures defined weakly, for bounded measurable $\xi_n : E_{0:n} \to \mathbb{R}$, by:

$$\begin{aligned}\pi_n(\xi_n) &= \mathbb{P}_{\eta_0}\left(\xi_n(X_{0:n})|Y_{0:n-1} = y_{0:n-1}\right)\\ \hat{\pi}_n(\xi_n) &= \mathbb{P}_{\eta_0}\left(\xi_n(X_{0:n})|Y_{0:n} = y_{0:n}\right),\end{aligned}$$

which are related, recursively via the following relationships:

$$\begin{aligned}\pi_n(\xi_n) &= \hat{\pi}_{n-1}M_n(\xi_n)\\ \hat{\pi}_n(\xi_n) &= \frac{\pi_n(\xi_n g_n^{y_n})}{\pi_n(g_n^{y_n})}\end{aligned}$$

where $g_n^{y_n}(x) := g_n(y_n - h(n,x))$. The first of these steps amounts to prediction of the state of the system at time n based upon knowledge of its state at time $n-1$ and the dynamics of the system and the second to updating that belief given an indirect observation of the system at time n. Looking at this recursive structure, it is clear that the filtering distributions amount to the normalised predicted and updated Feynman-Kac measures, with the collection of Markov kernels determined by the system dynamics and the potential functions by the measurement functions, g_n.

Particle Filters. We shall use the term *particle filters* to refer to SMC methods for approximating the optimal filter by numerical integration techniques, as proposed by [92]. Although we consider SMC somewhat more generally than this, for the purposes of considering this important model, it is useful to consider the recursion described above in terms of the sequence of densities which provide a version of the conditional expectations of interest – see [46] for details. Loosely, one assumes that all of the distributions of interest admit densities with respect to a suitable dominating measure and, using p to denote these densities, one has the probability of state x_n generating an observation y_n given by $p(y_n|x_n)$ and the probability of a state transition from x_n to x_{n+1} by $p(x_{n+1}|x_n)$ which allows us to recursively express the density of the joint distribution of the state sequence from time 0 to t given the associate observations using:

$$\begin{aligned}p(x_{1:n}|y_{1:n-1}) &= p(x_n|x_{n-1})p(x_{n-1}|y_{1:n-1})\\ p(x_{1:n}|y_{1:n}) &= \frac{p(y_n|x_n)p(x_{1:n}|y_{1:n-1})}{\int p(y_n|x'_n)p(x'_n|y_{1:n-1})dx'_n}.\end{aligned}$$

The first method which is usually considered a SMC algorithm for approximate filtering is the Sequential Importance Sampling (SIS) approach which is described in algorithm 2.7 (Bayesian importance sampling was proposed by [61], see [48] for the sequential formulation). The essence of this approach is to extend importance sampling to a sequential framework by extending the length of the path associated with each particle at each time, according to its marginal distribution conditioned upon its present value.

Algorithm 2.7 Sequential Importance Sampling (SIS)

1: Set $t = 1$.
2: **for** $i = 1$ to N **do**
3: $X_1^{(i)} \sim \mu_1$
{where μ_1 is an instrumental distribution.}
4: $W_1^{(i)} \propto \frac{\mathrm{d}\pi_1}{\mathrm{d}\mu_1}(X_1^{(i)})$
5: **end for**
6: $t \leftarrow t + 1$
7: **for** $i = 1$ to N **do**
8: $X_t^{(i)} \sim \mu_t\left(\cdot|X_{1:t-1}^{(i)}\right)$
{$\mu_t(\cdot|X_{1:t-1})$ is some instrumental distribution which may depend upon the particle history.}
9: $W_t^{(i)} \propto W_{t-1}^{(i)} \times \frac{\mathrm{d}\pi_t(\cdot|X_{1:t-1}^{(i)})}{\mathrm{d}\mu_t(\cdot|X_{1:t-1}^{(i)})}(X_t^i)$.
10: **end for**
11: Go to step 6.

The importance weighting step will, unless the proposal distribution is extremely close to the true distribution, lead over a number of iterations to a small number of particles with very large weights and ultimately a single particle will have a weight very close to unity, with the rest being essentially zero. This is the problem of particle weight degeneracy. In order to address this problem, the Sequential Importance Resampling (SIR) algorithm was applied to filtering [72, 93], following a proposal to use a SIR algorithm for missing data problems [136]. This approach (which is referred to as the *bootstrap* filter in those instances in which the prior is used as the proposal distribution) is described formally by algorithm 2.8. Resampling is a method by which a weighted set of samples from some distribution are replaced with an unweighted set of samples from the same distribution by replicating those particles with large weights and eliminating those with small weights. As resampling leads to a set of equally weighted particles at each time, the degeneracy problem is alleviated to some extent. However, resampling does reduce the number of distinct paths with particle representations, and any attempt to perform integrations over the path space will suffer from this form of degeneracy. There are three resampling schemes in common use amongst the particle

filtering community: multinomial resampling as used by [72], residual resampling [102] and stratified resampling [13].

Algorithm 2.8 Sequential Importance Resampling (SIR)

1: Set $t = 1$.
2: **for** $i = 1$ to N **do**
3: $\hat{X}_1^{(i)} \sim \mu_1$
{where μ_1 is an instrumental distribution.}
4: $W_1^{(i)} \propto \frac{\mathrm{d}\pi_1}{\mathrm{d}\mu_1}(\hat{X}_1^{(i)})$
5: **end for**
6: Resample $\left\{\hat{X}_t^i, W_t^{(i)}\right\}$ to obtain $\{X_t^{(i)}\}$.
7: $t \leftarrow t + 1$
8: **for** $i = 1$ to N **do**
9: $\hat{X}_t^{(i)} \sim \mu_t\left(\cdot|X_{1:t-1}^{(i)}\right)$
{$\mu_t(\cdot|X_{1:t-1})$ is some instrumental distribution which may depend upon the particle history.}
10: $W_t^{(i)} \propto \frac{\mathrm{d}\pi_t(\cdot|X_{1:t-1}^{(i)})}{\mathrm{d}\mu_t(\cdot|X_{1:t-1}^{(i)})}(X_t^i)$.
11: **end for**
12: Go to step 6.

Multinomial resampling is the simplest approach: a new particle set is sampled with replacement from the discrete distribution provided by the original particle set. Residual resampling has lower variance than multinomial resampling [104] and differs in that particles are deterministically replicated according to the integer part of their weight, and then multinomial resampling is performed, using the fractional part of the particle weights to obtain the remainder of the particle set. Stratified resampling is the minimum variance unbiased resampling technique [13] and can be interpreted as stratified sampling from a mixture distribution in which each particle provides one mixture component.

In practice it is not desirable to resample after every iteration of a sequential algorithm as the resampling process can only increase the Monte Carlo variance of the current particle set – as a Rao-Blackwellisation argument makes clear. The Effective Sample Size (ESS) [97] is often used as a measure of sample impoverishment, with resampling carried out whenever the effective sample size falls below some threshold. The ESS, which is generally approximated by a simple function of the importance weights, is defined (when one is using weighted samples from μ to approximate π) by:

$$\mathrm{ESS}(t) \triangleq \frac{N}{1 + \mathrm{Var}_\mu\left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)} \approx \frac{N^2}{\sum_{i=1}^N \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X^{(i)})^2}$$

is obtained by considering the ratio of the variance of the integral of an arbitrary function under the empirical measure of the particle set obtained by importance sampling to that which would be obtained with the same number of *iid* samples

from the true distribution. This expression is obtained by applying the delta-method (see, for example, [32, p. 33]) which amounts to a second order Taylor expansion of the function about its mean justified by the asymptotic normality of the distribution, and approximating the variance of the weights with their sample variance. For this reason, some danger is involved: if the current empirical distribution is *bad enough* then none of the particles will have much weight, but it is entirely possible that the variance of the weights will be small. However, if it could be evaluated as an expectation under the true distribution, the small possibility of extremely large values of the weight would show that in fact the representation is extremely poor.

Although it has the verisimilitude of a degeneracy reduction technique, as noted by [24], resampling does not really alleviate the problem of sample degeneracy – it simply leads to a number of identical particles with equal weights rather than a single particle with a large weight. Although this mechanism can allow the marginal distribution of the state X_t to be estimated well, the path space distribution rapidly becomes degenerate. This is illustrated in a rare event estimation context in chapter 5, along with a novel algorithm which avoids this problem.

One approach to avoid sample degeneracy is provided by the *Resample-Move* algorithm [63, 64]. A particular case termed *Gibbs iteration within SIS* was proposed by [105]. The innovation of this approach is to move each particle according to a Markov kernel of the appropriate distribution after the resampling step. This is a useful degeneracy reduction technique in a much more general setting. In principle, it is possible to apply such a kernel to the path space on which the particles exist, but this is computationally infeasible as the space grows at every iteration and obtaining fast mixing kernels with acceptable computational cost is not possible. Sampling approaches which operate directly on the path space are becoming possible as a result of the work described in section 2.4.4.

It is not uncommon for an MCMC step to be included, either after resampling or after every iteration, to help reduce the rate at which the sample becomes impoverished. Thus, a general particle filter corresponds to something like the SIS algorithm, with resampling according to the current weight distribution carried out whenever the effective sample size falls below a suitable threshold and in some instance with a Markov move of suitable invariant distribution applied to each particle after each iteration or resampling step.

2.4.3 Auxiliary Variable Methods: The APF

We mention the Auxiliary Particle Filter (APF) proposed by [125, 126] and subsequently enhanced by [3]. This approach allows a weighting to be applied to par-

ticles based upon how well they explain the next observation prior to resampling, leading to a better set of particles at the next time-step. We assume that the densities mentioned above when describing the optimal filter exist, and that we have a further density, $\hat{p}(y_t|x_{t-1})$ which is an approximation of $\int p(y_t|x'_t)p(x'|x_{t-1})dx'$. This approach is described in algorithm 2.9.

Algorithm 2.9 Auxiliary Particle Filter (APF)

1: $t = 0$
2: **for** $i = 1$ to N **do**
3: $\hat{X}_0^{(i)} \sim \pi_0$
{where π_0 is an instrumental distribution.}
4: $W_0^{(i)} \propto \frac{\mathrm{d}\pi_0}{\mathrm{d}\pi_0}(X_0^i)$
{ π_0 is the marginal distribution of X_0}
5: **end for**
6: $t \leftarrow t + 1$
7: **for** $i = 1$ to N **do**
8: Set $\lambda_t^{(i)} \propto W_{t-1}^{(i)} \times \hat{p}(y_t|X_{t-1}^{(i)})$
{Where $\hat{p}$ is an analytically tractable approximation to the likelihood.}
9: **end for**
10: Resample $\left\{X_{t-1}^{(i)}, \lambda_t^{(i)}\right\}$ to obtain $\left\{X_t'^{(i)}, \frac{1}{N}\right\}$.
11: **for** $i = 1$ to N **do**
12: Sample $X_t^{(i)} \sim q(\cdot|X_t'^{(i)})$.
13: Set $W_t^{(i)} \propto \frac{p(y_t|X_t)p(X_t^{(i)}|X_{t-1}^{(i)})}{\hat{p}(y_t|X_{t-1}^{(i)})q(X_t^{(i)}|X_{t-1}^{(i)})}$
14: **end for**
15: Go to step 6.

Feynman-Kac Formulation. It is convenient to express this algorithm in terms of a mean field approximation to a Feynman-Kac flow, an approach which has proved extremely fruitful for standard particle filtering [34] (notation should correspond to that used there). As far as we are aware, this reformulation which allows a central limit theorem to be obtained straightforwardly has not previously been proposed in the literature.

Consider a sequence of random variables, $\{X_n\}_{n\geq 1}$, each of which is a vector, $X_n := X_{n,1:n} \in E_n$, and corresponds to the full sequence of hidden states from time 1 to time n. Defining a sequence of potential functions, $G_n : E_n \to (0, \infty)$ as:

$$G_1(x_1) = \frac{p(y_1|x_{1,1})p(x_{1,1})}{q_1(x_{1,1})} \times \hat{p}(y_2|x_{1,1}) \tag{2.10}$$

$$G_n(x_n) = \frac{p(y_n|x_{n,n})p(x_{n,n}|x_{n,n-1})}{\hat{p}(y_n|x_{n,n-1})q_n(x_{n,n}|y_n, x_{n,n-1})} \times \hat{p}(y_{n+1}|x_{n,n}) \tag{2.11}$$

and a sequence of Markov kernels, $M_n : E_{n-1} \to \mathcal{P}(E_n)$, as:

$$M_n(x_{n-1}, dx_n) = \prod_{p=1:n-1} \delta_{x_{n-1,p}}(dx_{n,p})q_n(dx_{n,n}|y_n, x_{n-1,n}) \tag{2.12}$$

we obtain a Feynman-Kac flow in distribution space, whose normalised time-marginal prediction measures allow us to view the auxiliary particle filter described previously as a mean field approximation of this distribution.

There is a slight subtlety as we wish to weight the terminal particle set slightly differently, but this can be easily achieved by considering the integral of a function under the empirical distribution of the Auxiliary Particle Filter (APF) as the integral of the product of that function and a suitable weighting function under the empirical measure associated with an N-particle mean field approximation to this Feynman-Kac flow. We define this sequence of weight functions as:

$$W_1(x_1) \propto \frac{p(y_1|x_{1,1})p(x_{1,1})}{q_1(x_{1,1})} \tag{2.13}$$

$$W_n(x_n) \propto \frac{p(y_n|x_{n,n})p(x_{n,n}|x_{n,n-1})}{\hat{p}(y_n|x_{n,n-1})q_n(x_{n,n}|y_n, x_{n,n-1})}. \tag{2.14}$$

It is clear that if $\eta_1 := q_1$ then $\eta_n(W_n \times f_n)/\eta_n(W_n) = \int f_n(x_{n,1:n})p(dx_{1:n}|y_{1:n})$. This formulation amounts to viewing the APF as a mechanism by which a Feynman-Kac flow is produced which has the property that its mean field interpretation will place more particles in the correct place than that associated with the distribution of interest; this flow is then used as an importance sampling instrumental distribution for that of interest.

In order to obtain a mean field approximation which corresponds directly to the APF, it is necessary to make use of the correct McKean interpretation of the flow. This amounts to defining a sequence of non-linear kernels, $K_{n,\eta} : E_{n-1} \times \mathcal{P}(E_{n-1}) \to \mathcal{P}(E_n)$ such that $\eta_{n-1}K_{n,\eta_{n-1}} = \eta_n$. Using selection and mutation operations, with the selection operation corresponding to a Boltzmann-Gibbs operator with potential function G_{n-1} corresponds to precisely the case which we require.

So, we have: $K_{n,\eta} := S_{n-1,\eta}M_n$, where $S_{n-1,\eta}(dx) = \frac{G_{n-1}(x)\eta(dx)}{\eta(G_{n-1})}$. A mean field interpretation of this flow corresponds to algorithm 2.9 without the final weighting stage. Consequently, we have that: $\int p(dx_{1:n}|y_{1:n})f(x_{1:n}) = \eta_n(W_n \times f)/\eta_n(W_n)$ and the behaviour of the particle system can be analysed using the techniques pioneered in [34].

Using results from [34, chapter 9], we know that under suitable conditions:

$$\lim_{N\to\infty} \gamma_n^N(W_n) \to \gamma_n(W_n) \tag{2.15}$$

and we know that we seek a central limit theorem for the quantity:

$$\frac{\eta_n^N(W_n f)}{\eta_n^N(W_n)} = \frac{\gamma_n^N(W_n f)}{\gamma_n^N(W_n)}. \tag{2.16}$$

Now,

$$\frac{\gamma_n^N(W_n f)}{\gamma_n^N(W_n)} - \frac{\gamma_n(W_n f)}{\gamma_n(W_n)} = \frac{\gamma_n(W_n)\gamma_n^N(W_n f) - \gamma_n^N(W_n)\gamma_n(W_n f)}{\gamma_n(W_n)\gamma_n^N(W_n)} \tag{2.17}$$

$$= \frac{\gamma_n(W_n)}{\gamma_n^N(W_n)}\left(\frac{\gamma_n^N(W_n f)}{\gamma_n(W_n)} - \frac{\gamma_n^N(W_n)\gamma_n(W_n f)}{\gamma_n(W_n)\gamma_n(W_n)}\right) \tag{2.18}$$

$$= \frac{\gamma_n(W_n)}{\gamma_n^N(W_n)}\left(\frac{1}{\gamma_n(W_n)}\gamma_n^N\left(W_n f - W_n\frac{\gamma_n(W_n f)}{\gamma_n(W_n)}\right)\right) \tag{2.19}$$

$$= \frac{\gamma_n(W_n)}{\gamma_n^N(W_n)}\gamma_n^N\left(W_n\frac{f - \frac{\gamma_n(W_n f)}{\gamma_n(W_n)}}{\gamma_n(W_n)}\right). \tag{2.20}$$

By employing exactly the approach which [34] uses to obtain a central limit theorem for the normalised flow, we are able to note that $\lim_{N\to\infty}\frac{\gamma_n(W_n)}{\gamma_n^N(W_n)} \to 1$ and by applying Slutzky's theorem, we can make use of the central limit theorem which applies to γ_n^N.

We know from [34, chapter 9] that:

$$\sqrt{N}\left(\gamma_n^N(f) - \gamma_n(f)\right) \xrightarrow{d} \mathcal{W}_n(f), \tag{2.21}$$

where $\mathcal{W}_n(f)$ is a centred Gaussian field with variance given by:

$$\sigma^2(\mathcal{W}_n(f)) = \frac{1}{N}\sum_{q=1}^{N}\gamma_q(1)^2\left[\eta_{q-1}K_{q,\eta_{q-1}}\left[Q_{q,n}(f) - K_{q,\eta_{q-1}}(Q_{q,n}(f))\right]^2\right]. \tag{2.22}$$

Thus, by applying Slutzky's lemma, and defining $f_n' := W_n\frac{f - \frac{\gamma_n(W_n f)}{\gamma_n(W_n)}}{\gamma_n(W_n)}$ we obtain:

$$\sqrt{N}\left(\frac{\gamma_n^N(W_n f)}{\gamma_n^N(W_n)} - \frac{\gamma_n(W_n f)}{\gamma_n(W_n)}\right) \xrightarrow{d} \mathcal{W}_n(f_n'). \tag{2.23}$$

Work is ongoing to make use of this result to obtain guidelines upon the use of auxiliary variable approaches within SMC methods, as well as to allow the comparison of this method, the marginalised particle filter [94] and more standard particle filtering techniques in realistic scenarios.

2.4.4 SMC Samplers

We consider here the methods developed in [36, 37, 38, 123]. The motivation for this approach is that it would be extremely useful to have a generic technique for sampling from a sequence of distributions defined upon arbitrary spaces which are somehow related. It is not possible to employ the standard SMC framework for this approach as this approach makes no changes to the history of the process, only to new states which are added, and can only be applied to a sequence of distributions defined on a strictly increasing sequence of spaces with the same conditional independence properties as the optimal filter.

The principal innovation of the SMC sampler approach is to construct a sequence of synthetic distributions with the necessary properties. Given a collection

of measurable spaces $(E_n, \mathcal{E}_n)_{n\in\mathbb{N}}$, upon which the sequence of probability measures from which we wish to sample, $(\pi_n)_{n\in\mathbb{N}}$ is defined, it is possible to construct a sequence of distributions $(\tilde{\pi}_n)_{n\in\mathbb{N}}$ upon the sequence of spaces $(\prod_{p=1}^{n} E_n)_{n\in\mathbb{N}}$ endowed with the product σ-algebras, which have the target at time n as a marginal distribution at that time. As we are only interested in this marginal distribution, there is no need to adjust the position of earlier states in the chain and standard SMC techniques can be employed.

However, this approach suffers from the obvious deficiency that it involves conducting importance sampling upon an extremely large space, whose dimension is increasing with time. In order to ameliorate the situation, the synthetic distributions are constructed as:

$$\tilde{\pi}_n(dx_{1:n}) = \pi_n(dx_n) \prod_{p=1:n-1} L_p\left(x_{p+1}, dx_p\right), \tag{2.24}$$

where $(L_n)_{n\in\mathbb{N}}$ is a sequence of Markov kernels *from* E_n *into* E_{n-1}.

With this structure, an importance sample from $\tilde{\pi}_n$ is obtained by taking the path $x_{1:n-1}$, a sample from $\tilde{\pi}_{n-1}$, and extending it with a Markov kernel $K_n : E_{n-1} \to \mathcal{P}(E_n)$, which leads to the importance weight[2]:

$$\begin{aligned}\frac{\mathrm{d}\tilde{\pi}_n}{\mathrm{d}\left[\tilde{\pi}_{n-1} \otimes K_n\right]}(x_{1:n}) &= \frac{\mathrm{d}\left[\pi_n \otimes L_{n:1}^{\otimes}\right]}{\mathrm{d}\left[\pi_{n-1} \otimes L_{n-1:1}^{\otimes} \otimes K_n\right]}(x_{1:n}) \\ &= \frac{\mathrm{d}\left[\pi_n \otimes L_{n-1}\right]}{\mathrm{d}\left[\pi_{n-1} \otimes K_n\right]}(x_{n-1}, x_n).\end{aligned}$$

Where densities exist, this may be written in the rather more intuitive form:

$$\frac{\pi_n(x_n) L_{n-1}(x_n, x_{n-1})}{\pi_{n-1}(x_{n-1}), K_n(x_{n-1}, x_n)},$$

and this perhaps makes it clearer than the previous expression that the construction is chosen such that the distribution of the beginning of the path, conditional upon the terminal values is the same under successive distributions due to this Markovian structure. This approach ensures that the importance weights at time n depend only upon x_n and x_{n-1}. This also allows the algorithm to be constructed in such a way that the full path of the sampler is not stored, in much the same manner as standard SMC. Algorithm 2.10 describes the basic algorithm.

This approach is clearly very flexible, and is perhaps *too general*: in addition to the choice of the proposal kernels K_n, it is now necessary to select a sequence of auxiliary kernels L_n. The appearance of these kernels in the weight expression makes it clear that it will be extremely important to select these carefully. In fact, the central limit theorem presented in [37] demonstrates that the variance of

[2] With the obvious abuse of notation, the ordering of the arguments is unambiguous. We refer the reader to chapter 1 for an explanation of the tensor product notation.

Algorithm 2.10 SMC Sampler

1: $t = 1$
2: **for** $i = 1$ to N **do**
3: $X_1^{(i)} \sim \mu_1$
{where μ_1 is an instrumental distribution.}
4: $W_1^{(i)} \propto \frac{\mathrm{d}\pi_1}{\mathrm{d}\mu_1}(X_1^{(i)})$
5: **end for**
6: $t \leftarrow t + 1$
7: **for** $i = 1$ to N **do**
8: $X_t^{(i)} \sim K_t(X_{t-1}^{(i)}, \cdot)$
9: $W_t^{(i)} \propto W_{t-1}^{(i)} \frac{\pi_t(X_t^{(i)}) L_{t-1}(X_t^{(i)}, X_{t-1}^{(i)})}{\pi_{t-1}(X_{t-1}^{(i)}) K_t(X_{t-1}^{(i)}, X_t^{(i)})}$
10: **end for**
11: Resampling can be conducted at this stage.
12: Sample rejuvenation can be conducted at this stage by allowing the particles to evolve under the action of a Markov kernel of the invariant distribution π_t.

the estimator is strongly dependent upon the choice of these kernels. The same source obtains an expression for the optimal auxiliary kernels (in the sense that they minimise the variance of the importance weights if resampling is conducted at every step):

$$L_{n-1}^{opt}(x_n, dx_{n-1}) = \frac{\mathrm{d}K_n(x_{n-1}, \cdot)}{\mathrm{d}\pi_{n-1}K_n}(x_n)\pi_{n-1}(dx_{n-1}). \tag{2.25}$$

Using the optimal kernel, one finds that the weight expression reduces to the following form – where densities are assumed to exist for simplicity of presentation:

$$\begin{aligned}
&\frac{\pi_n(x_n)L_{n-1}^{opt}(x_n, x_{n-1})}{\pi_{n-1}(x_{n-1}), K_n(x_{n-1}, x_n)} \\
=&\frac{\pi_n(x_n)\pi_{n-1}(x_{n-1})K_n(x_{n-1}, x_n)/\pi_{n-1}K_n(x_n)}{\pi_{n-1}(x_{n-1}), K_n(x_{n-1}, x_n)} \\
=&\frac{\pi_n(x_n)}{\pi_{n-1}K_n(x_n)},
\end{aligned}$$

and this has the intuitive interpretation that this is the kernel which amounts to integrating out the effect of x_{n-1} and thus performing importance sampling directly on the space of interest.

Whenever it is possible to use this kernel, one should do so. However, in many instances the integral $\pi_{n-1}K_n$ will prove intractable, and an approximation to the optimal kernel must be used instead[3]. A number of strategies are suggested in [38]. In instances in which all distributions are defined on a common space $(E, \mathcal{E})$, K_n has invariant distribution π_n and the target distributions do not differ too much from one another. One particularly interesting variant is:

$$L_{n-1}^{tr}(x_n, dx_{n-1}) = \frac{\mathrm{d}K_n(x_{n-1}, \cdot)}{\mathrm{d}\pi_n}(x_n)\pi_n(dx_{n-1}), \tag{2.26}$$

[3] We note that it is *approximately optimal* rather than an algorithmic approximation: the algorithm remains exact, only the estimator variance suffers as a result of the approximation.

which is the time-reversal kernel associated with K_n. Using this kernel amounts to approximating π_{n-1} by π_n which, in these circumstances, is likely to be a very good approximation if $\pi_{n-1} \approx \pi_n$. We note that if this auxiliary kernel is employed then the importance weight is, in the case where a density exists, given by:

$$\begin{aligned}&\frac{\pi_n(x_n)L_{n-1}^{tr}(x_n, x_{n-1})}{\pi_{n-1}(x_{n-1}), K_n(x_{n-1}, x_n)}\\ =&\frac{\pi_n(x_{n-1})}{\pi_{n-1}(x_{n-1})}.\end{aligned}$$

In this instance, the weighting of each particle is independent of the final state and it becomes sensible to perform resampling *before* sampling that state.

It can be seen that some other methodologies which have been developed recently can be interpreted as special cases of SMC Samplers. Annealed Importance Sampling (AIS) [117] can be interpreted as a special case of the SMC samplers approach, employing 2.26 as the auxiliary kernel in which no resampling is performed. The *particle filter for static parameters* of [24] is a resample move approach with the same implicit backward kernel, this time with resampling. Population Monte Carlo (PMC) [11] can be interpreted as another special case of this framework, with $\pi_n = \pi$ and $L_{n-1}(x, x') = \pi(x')$.

A wide range of convergence results for these methods have been obtained, including a central limit theorem [38] and stability results via a Foster-Lyapunov approach [82].

2.4.5 Feynman-Kac Metropolis Sampling

Another area in which Feynman-Kac flows seem likely to make a major impact is that of *interacting Markov Chains*. By allowing a particle set to evolve according to a Feynman-Kac flow with a particular stationary distribution, one essentially produces an interacting set of "non-linear Markov Chains" (in the sense that their evolution depends upon one another's states, and this can dramatically improve the rate of convergence to equilibrium.

The Feynman-Kac Metropolis model described by [34, chapter 5] and the approach of [35], can be seen as the logical extension of MCMC from the realm of Markov chains to that of Feynman-Kac flows. The drawback of these methods is that the nonlinearity of the normalised Feynman-Kac flows is such that it is not possible to simulate them exactly. A particle approximation scheme is proposed in these works, together with a selection of stability results. The usual mean field particle interpretation convergence results apply. As noted in the original articles, a major advantage of these models over more traditional MCMC techniques is that the rate of decay to equilibrium is independent of the limiting measure.

2.5 Summary

We have presented a brief survey of the Monte Carlo literature, concentrating upon recent developments in the SMC field. It is in this area that this thesis makes its contribution, and we begin in the next chapter with an asymptotic analysis of a recently developed interacting particle system which is unusual in that it does not have a Feynman-Kac interpretation.

3. The SMC Implementation of the PHD Filter

"In some sort of crude sense, which no vulgarity, no humour, no overstatement can quite extinguish, the physicists have known sin; and this is a knowledge which they cannot lose."

– J. Robert Oppenheimer

The work done in sections 3.3 and particularly 3.4 was done in collaboration with Sumeetpal Singh, Arnaud Doucet and Ba-Ngu Vo. An early version of these results was published as [90] and the following sections closely resemble [91].

3.1 Introduction

In a standard Hidden Markov Model (HMM), the state and measurement at time k are two vectors of possibly different dimensions, belonging to E and F respectively. These vectors evolve randomly over time but their dimensions are fixed. The aim is to compute recursively in time the distribution of the hidden state given all the observations that have been received so far. In *multiple-object filtering*, recently introduced and studied by the data-fusion and tracking community [71, 107], the aim is to perform filtering when the state and observation variables are the finite subsets of E and F. Conceptually, this problem can be thought of as that of performing filtering when the state and observation spaces are the disjoint unions, $\uplus_{i=0}^{\infty} E^i$ and $\uplus_{i=0}^{\infty} F^i$, respectively. We remark that developing efficient computational tools to propagate the posterior density is extremely difficult in this setting – see, for example, [60].

An alternative which is easier to approximate computationally, the Probability Hypothesis Density (PHD) filter, has recently been proposed [107]. The PHD filter is a recursive algorithm that propagates the first moment, also referred to as the *intensity* [31], of the multiple-object posterior. The first moment is an appropriately defined measure on E (although the term is also used to refer to the Radon-Nikodým derivative of this measure with respect to some appropriately defined dominating measure on the same space). While the first moment is now a function on E, i.e. the dimension of the "state space" is now fixed, the PHD filter recursion still involves multiple integrals that have no closed form expressions in general. An SMC implementation of the PHD filter was proposed in [147].

The aim of this chapter is to analyse the convergence of the SMC implementation of the PHD filter proposed in [147]. Although numerous convergence results and central limit theorems have been obtained for particle systems which approximate Feynman-Kac flows [5] (including the optimal filtering equations) as mentioned in section 2.4, the PHD filter, being a first moment of the multiple-object posterior, is an unnormalised density that does not obey the standard Bayes recursion. Thus, convergence results and central limit theorems which have been derived for Feynman-Kac flows do not apply to the SMC approximation of the PHD filter. Our contribution is to extend existing results to this system which has a number of added difficulties, particularly that the total mass of the filter is a time-varying quantity and the recursions are non-standard.

3.2 Background and Problem Formulation

3.2.1 Notation and Conventions

It is convenient, at this stage, to summarise the notation required to deal with random sets and the PHD filter and the conventions which have been adopted throughout the remainder of this chapter. Except where otherwise specified, this is consistent with usage throughout the rest of the thesis, however some additional notation is needed to deal with random sets and related concepts, and that is summarised here rather than in section 1.2 as it is not required elsewhere in the thesis.

It is assumed throughout that the particle system first introduced in section 3.2.3 is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. All expectations and probabilities which are not explicitly associated with some other measure are taken with respect to $\mathbb{P}$. We have assumed throughout this chapter that all measures admit a density with respect to some dominating measure, $\lambda(dx)$, and used the same symbol to represent a density and its associated measure, i.e. for some measure $\mu \in \mathcal{M}(E)$,

$$\mu(dx) = \mu(x)\lambda(dx).$$

Given a measure μ, the integral of a function, f, with respect to μ is denoted $\mu(f)$.

Given two transition kernels K and L which admit a density with respect to a suitable dominating measure, where L is absolutely continuous with respect to K (i.e. $K >> L$), we define $\frac{L}{K}(u,v) = \frac{dL(u,\cdot)}{dK(u,\cdot)}(v)$ (i.e. the Radon-Nikodým derivative). Given a transition kernel K and a non-negative function $g : E \times E \to \mathbb{R}^+$ we define the new kernel $K \times g$ by $K \times g(u, dv) = K(u, dv)g(u, v)$. Similarly, for two measures μ and ν on E, we define $\frac{\mu}{\nu}(u)$ to be $\frac{d\mu}{d\nu}(u)$. If μ and ν both admit densities with respect to the same dominating measure λ then $\frac{d\mu}{d\nu}(u)$ is simply the ratio of those

densities evaluated at the point u. For any two functions $f, g : E \to \mathbb{R}$ we write fg for the standard multiplication of these functions.

When dealing with random finite sets, the convention in the literature is to use capital Greek letters to refer to a random set, a capital Roman letter to refer to a realisation of such a set and a lower case Roman letter to refer to an element of a realisation. We have followed this convention wherever possible.

Finally, we have considered the evolution of the PHD as an unnormalised density on a general space E. It is usual, but not entirely necessary, to assume that $E = \mathbb{R}^d$ and that the dominating measure $\lambda(dx)$ is Lebesgue measure. For the target tracking application described in section 3.2.4, this is, indeed, the case.

3.2.2 Multiple Object Filtering

We remark that although the description below is removed from any particular application, the model is popular with the data fusion and target tracking community [60, 71, 107]. Our intention in giving this abstract presentation is to emphasise the generality of the model with the intention of arousing the interest of other scientific communities.

The multiple-object state evolves over time in a Markovian fashion and at each time k, a multiple-object measurement is generated based upon the state at time k only. The multiple-object state and multiple-object measurement at time k are naturally represented as finite subsets $X_k \subset E$ and $Z_k \subset F$ respectively. For example, at time k, let X_k have $M(k)$ elements, i.e.,

$$X_k = \{x_{k,1}, \ldots, x_{k,M(k)}\} \in \mathcal{T}(E),$$

where $\mathcal{T}(E)$ denotes the collection of all finite subsets of the space E. Similarly, if $N(k)$ observations $z_{k,1}, \ldots, z_{k,N(k)}$ from F are received at time k, then

$$Z_k = \{z_{k,1}, \ldots, z_{k,N(k)}\} \in \mathcal{T}(F),$$

is the multiple-object measurement. Analogous to the standard HMM case, in which uncertainty is characterised by modelling the states and measurements by random vectors, uncertainty in a multiple-object system is characterised by modelling multiple-object states and multiple-object measurements as Random Finite Set (RFS) Ξ_k and Σ_k in E and F respectively. We denote particular realisations of Ξ_k and Σ_k by X_k and Z_k, respectively. Conditioned upon a realisation X_{k-1} of the state at time $k-1$, Ξ_{k-1} the state evolution satisfies

$$\Xi_k = \Xi_k^S(X_{k-1}) \cup \Xi_k^B(X_{k-1}) \cup \Gamma, \tag{3.1}$$

where $\Xi_k^S(X_{k-1})$ denotes the RFS of elements that have 'survived' to time k and the other terms are RFSs of new elements, which are decomposed as $\Xi_k^B(X_{k-1})$

of elements spawned (spawning is a term used in the tracking literature for the process by which a large target, such as an aircraft carrier, emits a number of smaller targets, such as aircraft) from X_{k-1} and the RFS Γ_k of elements that appear spontaneously at time k. Note that the state evolution model incorporates individual element motion, element birth, death and spawning, as well as interactions between the elements. Similarly, given a realisation X_k of Ξ_k at time k, the observation Σ_k is modelled by

$$\Sigma_k = \Theta_k(X_k) \cup \Lambda_k, \tag{3.2}$$

where $\Theta_k(X_k)$ denotes the RFS of measurements generated by X_k, and Λ_k denotes the RFS of measurements that do not originate from any element in X_k, such as false measurements due to sensor noise or objects other than the class of objects of interest. The observation process so defined can capture element measurement noise, element-dependent probability of occlusions and false measurements.

The multiple-object filtering problem concerns the estimation of the multiple-object state X_k at time step k given the collection $Z_{1:k} \equiv (Z_1, ..., Z_k)$ of all observations up to time k. The object of interest is the posterior probability density of Ξ_k.

The above description of the dynamics of $\{\Xi_k\}$ and $\{\Sigma_k\}$ was a constructive one, while in filtering one needs to specify the state transition and observation density, that is, the densities of the following measures,

$$P(\Xi_k \in A | \Xi_{k-1} = X_{k-1}),$$
$$P(\Sigma_k \in B | \Xi_k = X_k),$$

where $A \subset \mathcal{T}(E)$ and $B \subset \mathcal{T}(F)$ are the measurable sets of their respective spaces. As this chapter is concerned with the propagation of the first moment of the filtering density, we refer the reader to [147, 107] for details on the state transition and observation densities. We have also omitted details on how the RFSs of survived elements $\Xi_k^S(X_{k-1})$, spawned elements $\Xi_k^B(X_{k-1})$ and spontaneously spawned elements Γ_k are constructed. Similarly, details on the RFSs of true (or element generated) observations $\Theta_k(X_k)$ and false measurements Λ_k were omitted. Naturally, the construction of these sets are application specific and a simple numerical example provided in Section 3.2.4 aims to clarify the ideas. We refer the reader to [71, 107] for the constructions for applications in target tracking.

Although much work is being done in the area of multiple target tracking, including attempts at developing practical SMC algorithms which operate on the multiple target state spaces [98, 118], it is extremely difficult to perform the computations involved in the filtering problem for this type of model when the number of targets is large. SMC methods cannot operate efficiently when direct importance

sampling on a very high dimensional space is involved. Thus it is important to consider computationally tractable principled approximations. One such approximation is the PHD filter, one such approximation which has become popular among the tracking community [107, 147].

3.2.3 The PHD Filter

The PHD filter is a method of updating a measure, $\tilde{\alpha}_{k-1}$ given a random set of observations, Z_k, which can be interpreted as a first moment approximation of the usual Bayesian filtering equation. Within this framework, the quantity of interest is the intensity measure of a point process. Whilst it can be described by a measure, it is not in general a probability measure and it is necessary to maintain an estimate of both the total mass and the distribution of that mass.

Before summarising the mathematical formulation of the PHD filtering recursion, we briefly explain what is meant by the *first moment* of a random finite set. A finite subset $X \in \mathcal{T}(E)$ can also be equivalently represented by the counting measure, N_X, on the measurable subsets of E defined, for all such sets, A, by $N_X(A) = \sum_{x \in X} \mathbf{1}_A(x) = |A \cap X|$. Consequently, the random finite set Ξ can also be represented by a random counting measure N_Ξ defined by $N_\Xi(A) = |\Xi \cap A|$. This representation is commonly used in the point process literature [31].

The first moment of a random vector is simply the expectation of that random vector under a suitable probability measure. As there is no concept of set addition, an exact analogue of this form of moment is not possible in the RFS case. However, using the random counting measure representation, the 1st moment or *intensity measure* of a RFS Ξ is the first moment of its associated counting measure, i.e.,

$$\tilde{\alpha}(A) = \mathbb{E}\left[N_\Xi(A)\right].$$

The intensity measure of a set A gives the expected number of elements of Ξ that are in A. Although the intensity measure $\tilde{\alpha}$ is an integral of the counting measures, it is not itself a counting measure and hence does not necessarily have a finite set representation.

The density of the intensity measure with respect to a suitable dominating measure λ, when it exists, is also denoted $\tilde{\alpha}$ and is termed the *intensity function*[1]. In the tracking literature, $\tilde{\alpha}$ is also known as the PHD.

The PHD is the first moment of a RFS and hence tells us, for any region, the expected number of elements within that region. In the context of multiple-object filtering, the PHD recursion described below propagates the density of the intensity measure $\tilde{\alpha}_k(A) := \mathbb{E}\left[N_{\Xi_k}(A) | Z_1, \ldots, Z_k\right]$ for $k \geq 0$. This is clearly

[1] As a reminder, we use the same notation for a measure and its density throughout this chapter.

a useful representation for multiple-object filtering and other applications, as it provides a simultaneous description of the number of elements of Ξ_k within the space, and their locations.

The PHD recursion can be described in terms of *prediction* and *update* steps, just as the optimal filtering recursion can. The derivation of the update step cannot be reproduced here due to space constraints, but the most elegant approach involves considering the evolution of the probability generating functional associated with a Poisson process under the action of the update step. It is not possible to reproduce this derivation here, but it is presented accessibly in [107]:

$$\alpha_k(dx) = (\Phi_k\tilde{\alpha}_{k-1})(dx) = (\tilde{\alpha}_{k-1}\phi_k)(dx) + \gamma_k(dx) \tag{3.3}$$

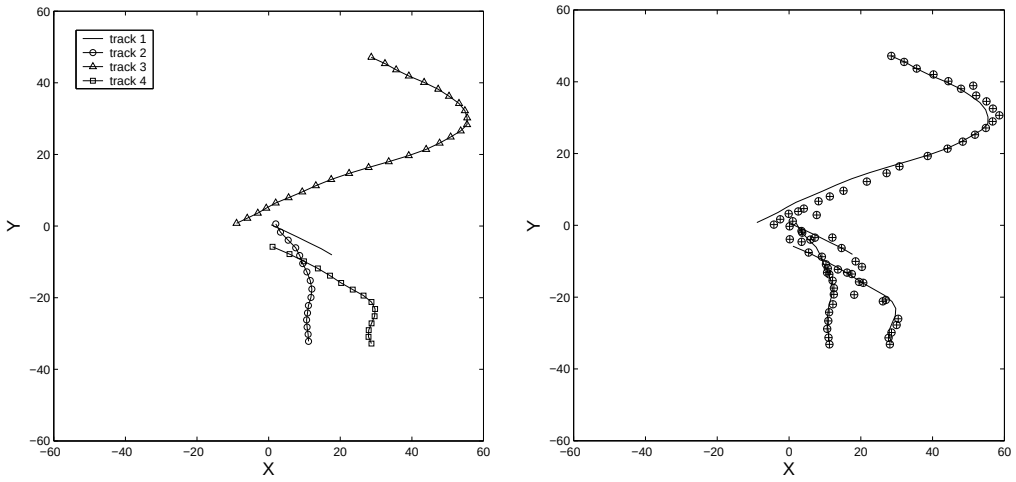
$$\tilde{\alpha}_k(dx) = (\Psi_k\alpha_k)(dx) = \left(\nu_k(x) + \sum_{z\in Z_k} \frac{\psi_{k,z}(x)}{\kappa_k(z) + \alpha_k(\psi_{k,z})}\right)\alpha_k(dx). \tag{3.4}$$

The prediction operator Φ_k is described in terms of a kernel, ϕ_k, which does not in general integrate to 1, and an additive measure, γ_x. The prediction kernel, ϕ_k describes the dynamics of existing elements and can be decomposed as: $\phi_k(x, dy) = e_k(x)f_k(x, dy) + b_k(x, dy)$ where $e_k(x)$ is the probability of an element at x at time $k-1$ surviving to time k, $f_k(x, dy)$ is a Markov kernel which describes the dynamics of the surviving elements and $b_k(x, dy)$ is a "spawning" kernel which describes the probability of an element at x at time $k-1$ giving rise to a new element in a neighbourhood dy at time k.

The update operator Ψ_k is a nonlinear operator which resembles a linear combination of Boltzmann-Gibbs operators (one of which describes the update equation of Bayesian filtering) with different associated potentials. However, there are some subtle differences which prove to be significant. The term Z_k denotes the random set of observations at time k and $\psi_{k,z}$ is the "likelihood" function associated with an observation at z at time k. $\kappa_k(z)$ is the intensity of the false measurement process at z. Finally, $\nu_k(x)$ is the probability of failing to observe an element at x at time k.

Note the correspondence between the terms in the PHD recursion and the sets in the constructive description of the multiple-object filtering problem in section 3.2.2. The pairing of the terms are as follows: (Ξ_k^B, b_k) describe object birth including spawning, (Γ_k, γ_k) describe spontaneous births, and $(\Xi_k^S, e_k f_k)$ describe the dynamics of surviving objects. The measurement model has a more subtle relationship, Θ_k incorporates all of the information of $\psi_{k,z}$ and ν_k while the effect of Λ_k on the first moment is described by κ_k.

An SMC Implementation of the PHD Filter. We consider algorithm 3.1, which is essentially that proposed in [147], which describes a sequential Monte



(a) Ground truth (b) PHD filter output estimates ($\oplus$)

Fig. 3.1. PHD Filter Example: plots of 4 superimposed tracks over 40 time steps. Taken from [147].

Carlo method for approximating the evolution of the PHD filter. Alternative SMC implementations of the PHD filter have also been proposed [141, 148], the first of which is substantially different from that considered here as different particle sets are used to deal with differing target numbers whilst the second is more similar in character. It is assumed that the filter is initialised at time zero by sampling a set of L_0 particles from the true PHD (or, rather, the probability measure obtained by appropriately normalising it) and weighting them according to the total mass at time zero such that each particle has weight $w_0^{(i)} = \tilde{\alpha}_0(\mathbf{1})/L_0$. The following recursion is then used to predict the configuration of the particle set at the next time step and then to update the estimate based upon the next observation set, just as in the standard filtering case. It is understood that the importance densities used may be conditioned upon the current observation set in addition to the previous particle position. We omit the dependency on the observation set in our notation.

3.2.4 A Motivating Example

We present a brief example (which is borrowed directly, with permission, from [147]) to illustrate the utility of the multiple-object filtering framework and the SMC implementation of the PHD filter. Consider the problem of tracking an unknown number of targets that evolve in $\mathbb{R}^4$. For instance, in a two dimensional tracking example, each target could be described by its x and y coordinates as well as its velocity in these directions. Existing targets can leave the surveillance area and new targets can enter the scene. At time k, a realisation of the state is $X_k = \{x_{k,1}, \ldots, x_{k,M(k)}\} \subset \mathbb{R}^4$. As for the observations, each target generates one observation with a certain probability (i.e. each target generates at most one observation) and the sensors can measure false observations that are not asso-

Algorithm 3.1 An SMC implementation of the PHD filter

Assume that a particle approximation consisting of L_{k-1} weighted particles is available at time $k-1$, with associated empirical measure $\tilde{\alpha}_{k-1}^{L_{k-1}}$. Then the following sequence of steps can be applied to provide such a particle approximation at time k:

Prediction:

Propagate forward the particles which survived the previous iteration to account for the dynamics of existing objects. For $i = 1, \ldots, L_{k-1}$, sample $Y_k^{(i)}$ from some importance distribution $q_k(X_{k-1}^{(i)}, \cdot)$ and calculate the importance weights

$$\tilde{w}_k^{(i)} = \frac{\phi_k\left(X_{k-1}^{(i)}, Y_k^{(i)}\right)}{q_k\left(X_{k-1}^{(i)}, Y_k^{(i)}\right)} w_{k-1}^{(i)} \tag{3.5}$$

Generate some new particles to account for spontaneous births. For $i = L_{k-1}+1, \ldots, L_{k-1}+J_k$, sample $Y_k^{(i)}$ from some importance distribution $p_k(\cdot)$ and calculate the importance weights

$$\tilde{w}_k^{(i)} = \frac{1}{J_k} \frac{\gamma_k(Y_k^{(i)})}{p_k(Y_k^{(i)})} \tag{3.6}$$

Update:

Compute the empirical estimate of the normalising constant associated with each observation,

$$C_k(z) = \kappa_k(z) + \sum_{i=1}^{L_{k-1}+J_k} \tilde{w}_k^{(i)} \psi_{k,z}(Y_k^{(i)})$$

Adjust the particle weights to reflect the most recent observations. Update all the particle weights with:

$$\hat{w}_k^{(i)} = \left[\nu\left(Y_k^{(i)}\right) + \sum_{z \in Z_k} \frac{\psi_{k,z}\left(Y_k^{(i)}\right)}{C_k(z)}\right] \tilde{w}_k^{(i)}$$

Resampling:

Estimate the total mass: $\hat{N}_k = \sum_{j=1}^{L_{k-1}+J_k} \hat{w}_k^{(j)}$

Resample to reduce sample impoverishment (that is, the presence of a large (and increasing in time) number of particles with very small weights) and to prevent exponential growth of the size of the particle ensemble. Starting from the particle/weight pairs $\left\{\frac{\hat{w}_k^{(i)}}{\hat{N}_k}, Y_k^{(i)}\right\}_{i=1}^{L_{k-1}+J_k}$ sample L_k particles from the empirical probability distribution obtained by suitably normalising it, to obtain a set of L_k particles of equal weight $\left\{w_k^{(i)}/\hat{N}_k, X_k^{(i)}\right\}_{i=1}^{L_k}$

Rescale the weights to reflect the total mass of the system (i.e. multiply the particle weights by a factor of $\hat{N}_k$) giving the particle/weight ensemble $\left\{w_k^{(i)}, X_k^{(i)}\right\}_{i=1}^{L_k}$ which defines $\tilde{\alpha}_k^{L_k}$.

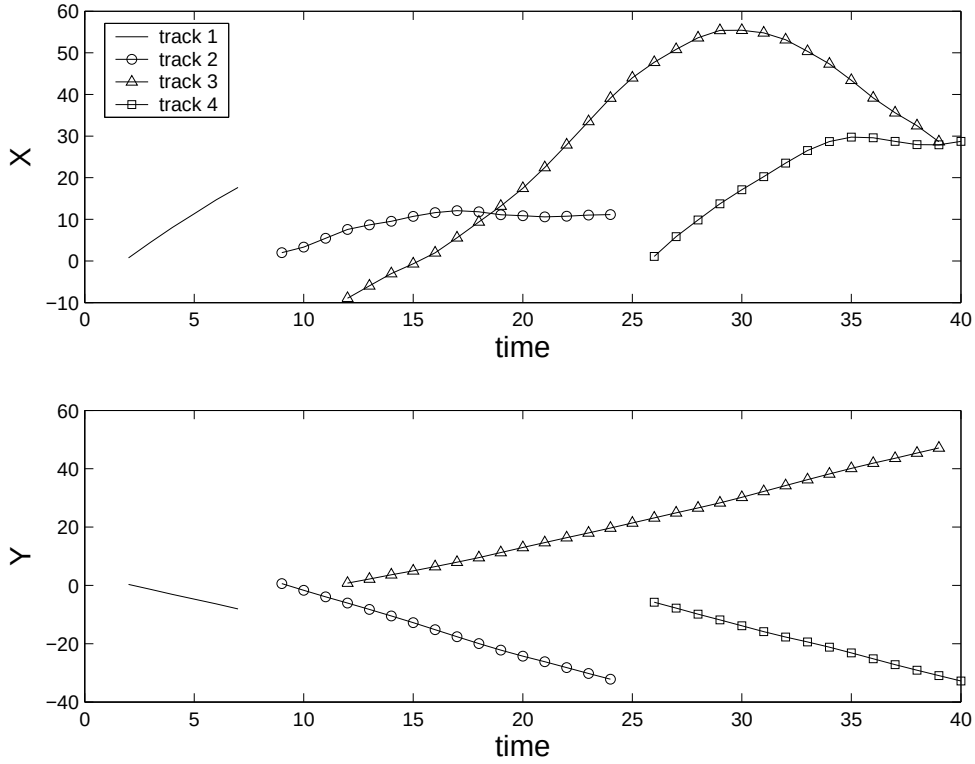
ciated with any target, i.e., clutter. Assume that sensors measure a noisy value of the x and y coordinate of a target. A realisation of the observation would be $Z_k = \{z_{k,1}, \ldots, z_{k,N(k)}\} \subset \mathbb{R}^2$ where measurement $z_{k,i}$ could either correspond to an element in X_k or be a false measurement. Note that the number of observations need not coincide with the number of targets.

We now demonstrate the results of tracking the targets using the SMC implementation of the PHD filter. In our example each target moves according to a standard linear Gaussian model. Each existing target has a probability of survival that is independent of its position and velocity. In this example, a target at time $k-1$ survives to time k with probability 0.95. For simplicity no spawning is considered. At each time k, new targets can appear spontaneously according to a Poisson point process with an intensity function γ_k set to $0.2\mathcal{N}(\cdot; \bar{x}, Q)$, where $\mathcal{N}(\cdot; \bar{x}, Q)$ denotes a normal density with mean $\bar{x}$ and uncertainty corresponding to the covariance, Q. This corresponds to one new target being created every five time steps around a location $\bar{x}$ with covariance Q. As for the observations, each target generates a noisy observation of its position with certain probability. Additionally, false measurements are generated according to a Poisson point process with a uniform intensity function.

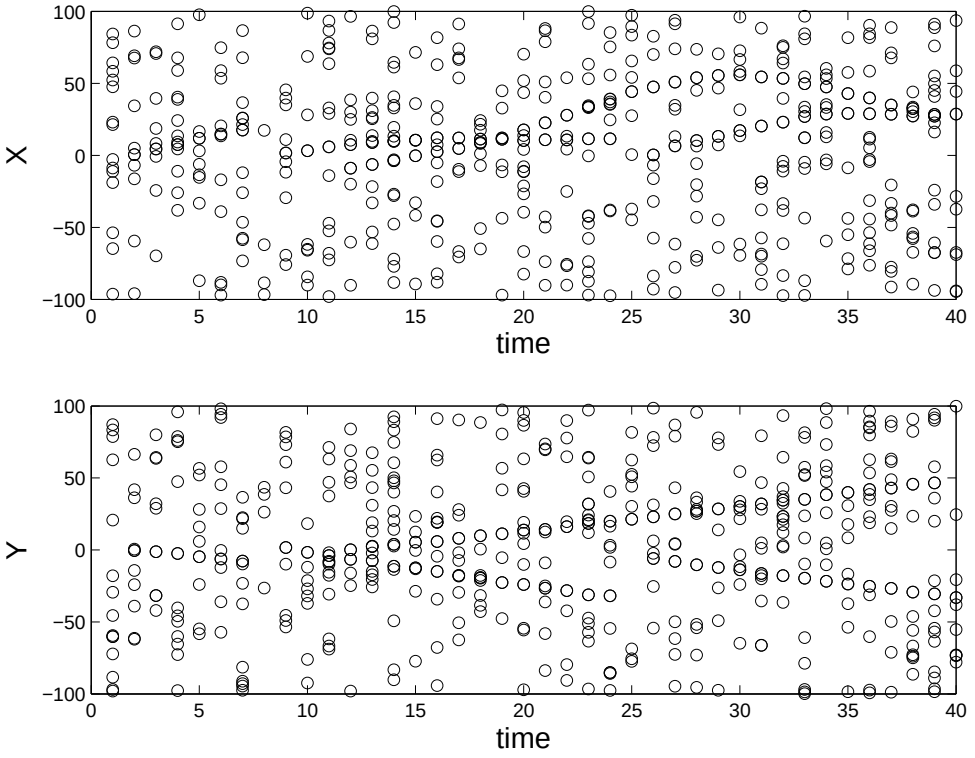
The peaks of $\tilde{\alpha}_k$ are points in E with the highest local concentration of the expected number of targets, and hence may be used to generate estimates for the location of the elements of Ξ. Since the total mass of the intensity measure gives the expected number of targets, the simplest approach is to round the particle estimate of this quantity to the closest integer, $\hat{N}_k$ and then to select the $\hat{N}_k$ largest peaks as target locations. This was the approach adopted in this numerical example, for which the positions of 4 targets over 40 time steps are displayed in Figure 3.1(b). These 4 targets start in the vicinity of the origin and move radially outwards. The start and finish times of each target can be seen from Figure 3.2(a), which plots the individual x and y components of each track against time. The x and y coordinates of the observations Z_k for all 40 time steps are shown in Figure 3.2(b). Figure 3.1(b) shows the position estimates superimposed on the true tracks over the 40 time steps. Observe the close proximity of the estimated positions to the true tracks even though the tracks of the targets were not strictly generated according to the assumed model.

3.3 Convergence Study

It is shown in this section that the integral of any bounded test function under the SMC approximation of the PHD filter converges to the integral of that function under the true PHD filter in mean of order p (for all integer p) and hence



(a) Ground truth: plots of x and y components of the 4 true tracks against time, showing the different start and finish times of the tracks.



(b) x and y components of position observations immersed in clutter of rate $r = 10$.

Fig. 3.2. PHD filter example: true target positions and generated observations as a function of time. Taken from [147].

almost surely. The restriction that test functions must be bounded seems more reasonable in the context of the PHD filter than the standard optimal filter as one is typically interested in the integrals of indicator functions. The result is shown to hold recursively by decomposing the evolution of the filter into a number of steps at each time. A number of additional points need to be considered in the present case. We assume throughout that the observation record $\{Z_k\}_{k\geq 0}$ is fixed and generates the PHD recursion.

We note that the results of this section are somewhat similar to those derived concurrently by [27]. Both approaches are obtained by considering a modification of the approach of [30] to take into account the differences between the PHD filter and the standard particle filter. Our approach to obtaining almost sure convergence is somewhat different to that employed by either of these papers. The central limit theorem presented below has no analogue in [27].

Remark 3.3.1. As a preliminary, we need to show that both the true and approximate filters have finite mass at all times. In the case of the true filter this follows by assuming that the mass is bounded at time zero and that $||\phi_k||_\infty$ is finite. Proceeding by induction we have:

$$\begin{aligned} \tilde{\alpha}_k(\mathbf{1}) &= \Psi_k \Phi_k \tilde{\alpha}_{k-1}(\mathbf{1}) \\ \Phi_k \tilde{\alpha}_{k-1}(\mathbf{1}) &\leq \gamma_k(\mathbf{1}) + ||\phi_k||_\infty \tilde{\alpha}_{k-1}(\mathbf{1}) \\ \tilde{\alpha}_k(\mathbf{1}) &\leq |Z_k| + \gamma_k(\mathbf{1}) + ||\phi_k||_\infty \tilde{\alpha}_{k-1}(\mathbf{1}) \end{aligned} \tag{3.7}$$

whilst, in the case of the particle approximation, it can always be shown to hold from the convergence towards the true filter at the previous time. Note that, whenever we have a result of the form (3.10) or (3.11) together with (3.7) the total mass of the approximate filter must be finite with probability one and a finite upper bound upon the mass can be obtained immediately (consider the $\mathbb{L}_1$ convergence result obtained by setting $p = 1$ in (3.10) or (3.11)).

We make extensive use of [34, Lemma 7.3.3], the relevant portion of which is reproduced here.

Lemma 3.3.1 (Del Moral, 2004). *Given a sequence of probability measures $(\mu_i)_{i\geq 1}$ on a given measurable space $(E, \mathcal{E})$ and a collection of independent random variables, one distributed according to each of those measures, $(X_i)_{i\geq 1}$, where $\forall i, X_i \sim \mu_i$, together with any sequence of measurable functions $(h_i)_{i\geq 1}$ such that $\mu_i(h_i) = 0$ for all $i \geq 1$, we define for any $N \in \mathbb{N}$,*

$$m_N(X)(h) = \frac{1}{N}\sum_{i=1}^{N} h_i(X_i) \quad \textit{and} \quad \sigma_N^2(h) = \frac{1}{N}\sum_{i=1}^{N} \left(\sup(h_i) - \inf(h_i)\right)^2$$

If the h_i have finite oscillations (i.e., $\sup(h_i) - \inf(h_i) < \infty \; \forall i \geq 1$) then we have:

$$\sqrt{N}\mathbb{E}\left[|m_N(X)(h)|^p\right]^{1/p} \leq d(p)^{1/p}\sigma_N(h)$$

with, for any pair of integers n, p *such that* $n \geq p \geq 1$*, denoting* $(n)_p = n!/(n-p)!$*:*

$$d(2n) = (2n)_n 2^{-n} \quad \textit{and} \quad d(2n-1) = \frac{(2n-1)_n}{\sqrt{n-\frac{1}{2}}} 2^{-(n-\frac{1}{2})}$$

We begin by showing that as the number of particles used to approximate the PHD filter tends towards infinity, the estimate of the integral of any bounded measurable function under the empirical measure associated with the particle approximation converges towards the integral under the true PHD filter in terms of $\mathbb{L}_p$ norm and that the two integrals are $\mathbb{P}-$a.s. equal in the limit of infinitely many particles. The principal result of this section is theorem 3.3.1 which establishes the first result and leads directly to the second.

Throughout this section we assume that a particle approximation consisting of L_{k-1} weighted particles is available at time $k-1$, with associated empirical measure $\tilde{\alpha}_{k-1}^{L_{k-1}}$. These particles are propagated forwards according to algorithm 3.1, and an additional J_k particles are introduced to account for the possibility of new objects appearing at time k. This gives us a $M_k = J_k + L_{k-1}$ particle approximation, denoted $\alpha_k^{M_k}$, to the PHD filter at time k, which is subsequently re-weighted (corresponding to the update step of the exact algorithm) and resampled to provide a sample of L_k particles at this time, $\tilde{\alpha}_k^{L_k}$. This leads to a recursive algorithm and provides a convenient decomposition of the error introduced at each time-step into quantities which can be straightforwardly bounded. We assume that J_k and M_k are chosen in a manner independent of the evolution of the particle system, but which may be influenced by such factors as the number of observations.

3.3.1 Conditions

As a final precursor to the convergence study, we present a number of weak conditions which are sufficient for the convergence results below to hold. The following conditions are assumed to hold throughout:

- The particle filter is initialised with some finite mass by *iid* sampling from a tractable distribution $\tilde{\alpha}_0$.
- The observation set is finite, $|Z_k| < \infty \forall k$.
- All of the importance ratios are bounded above:

$$\sup_{(x,y)\in E\times E}\left|\frac{\phi_k(x,y)}{q_k(x,y)}\right| < R_1 < \infty \quad \sup_{x\in E}\left|\frac{\gamma_k(x)}{p_k(x)}\right| < R_2 < \infty \tag{3.8}$$

  and that at least one of these ratios is also strictly positive.
- The individual object likelihood function is bounded above and strictly positive:

$$0 < \psi_{k,z}(x) < R_3 < \infty \tag{3.9}$$

– The number of particles used at each time step are not dependent upon the particle approximation at that time step. In the case of the convergence results we allow for fairly general behaviour, requiring only that the number of particles at each stage is proportional to the number used at the previous step in the algorithm, $L_k \propto M_k = L_{k-1} + J_k$ and $J_k \propto L_{k-1}$; in the central limit theorem we assume that N particles are propagated forward at each time step and some additional fraction η_k are introduced at each time k to describe the spontaneous birth density (this is done for convenience rather than through necessity).
– Resampling is done according to a multinomial scheme, i.e. the number of representatives of each particle which survives is sampled from a multinomial distribution with parameters proportional to the particle weights.

The first of these conditions simply constrain the initialisation of the particle approximation, the next is a weak finiteness requirement placed upon the true system, the next two are implementation issues and are required to ensure the importance weights and that the filter density remains finite. The penultimate condition prevents unstable interactions between the filter mass and the particle approximation.

3.3.2 $\mathbb{L}_p$ Convergence and Almost Sure Convergence

The following theorem is the main result of this section and is proved by induction. It is shown that each step of the algorithm introduces an error (in the $\mathbb{L}_p$ sense) whose upper bound converges to zero as the number of particles tends to infinity and that the errors accumulated by the evolution of the algorithm have the same property.

Theorem 3.3.1 ($\mathbb{L}_p$ Convergence). *Under the conditions specified in section 3.3.1, there exist finite constants such that for any $\xi \in \mathcal{B}_b(E)$, $\xi : E \to \mathbb{R}$ the following holds for all times k:*

$$\mathbb{E}\left[\left|\alpha_k^{M_k}(\xi) - \alpha_k(\xi)\right|^p\right]^{1/p} \leq \bar{c}_{k,p}\frac{||\xi||_\infty}{\sqrt{M_k}} \tag{3.10}$$

$$\mathbb{E}\left[\left|\tilde{\alpha}_k^{L_k}(\xi) - \tilde{\alpha}_k(\xi)\right|^p\right]^{1/p} \leq c_{k,p}\frac{||\xi||_\infty}{\sqrt{L_k}} \tag{3.11}$$

Convergence in an $\mathbb{L}_p$ sense directly implies convergence in probability, so we also have:

$$\begin{aligned} \alpha_k^{M_k}(\xi) &\xrightarrow{p} \alpha_k(\xi) \\ \tilde{\alpha}_k^{L_k}(\xi) &\xrightarrow{p} \tilde{\alpha}_k(\xi) \end{aligned}$$

Furthermore, by a Borel-Cantelli argument, the particle approximation of the integral of any function with finite fourth moment converges almost surely to the

integral under the true PHD filter as the number of particles tends towards infinity.

Proof. Equation (3.11) holds at time 0 by lemma 3.3.2.

Now, if equation (3.11) holds at time $k-1$ then, by lemmas 3.3.3 and 3.3.4, equation (3.10) holds at time k.

Similarly, if equation (3.10) holds at time k then by lemmas 3.3.5 and 3.3.6, equation (3.11) also holds at time k.

The theorem follows by induction. □

Lemma 3.3.2 (Initialisation). *If, at time zero, the particle approximation, $\tilde{\alpha}_0^{L_0}$, is obtained by taking L_0 iid samples from $\tilde{\alpha}_0/\tilde{\alpha}_0(\mathbf{1})$ and weighting each by $\tilde{\alpha}_0(\mathbf{1})/L_0$, then there exists a finite constant $c_{0,p}$ such that, for all $p \geq 1$ and for all test functions ξ in $\mathcal{B}_b(E)$:*

$$\mathbb{E}\left[\left|\tilde{\alpha}_0^{L_0}(\xi) - \tilde{\alpha}_0(\xi)\right|^p\right]^{1/p} \leq c_{0,p}\frac{||\xi||_\infty}{\sqrt{L_0}}$$

Proof. This can be seen to be true directly by applying lemma 3.3.1. □

Lemma 3.3.3 (Prediction). *If, for some finite constant $c_{k-1,p}$, and all test functions ξ in $\mathcal{B}_b(E)$:*

$$\mathbb{E}\left[\left|\tilde{\alpha}_{k-1}^{L_{k-1}}(\xi) - \tilde{\alpha}_{k-1}(\xi)\right|^p\right]^{1/p} \leq c_{k-1,p}\frac{||\xi||_\infty}{\sqrt{L_{k-1}}}$$

Then there exists some finite constant $\hat{c}_{k,p}$ such that, for all test functions ξ in $\mathbb{L}_p(E)$:

$$\mathbb{E}\left[\left|\Phi_k\tilde{\alpha}_{k-1}^{L_{k-1}}(\xi) - \Phi_k\tilde{\alpha}_{k-1}(\xi)\right|^p\right]^{1/p} \leq \hat{c}_{k,p}\frac{||\xi||_\infty}{\sqrt{L_{k-1}}}$$

Proof. From the definition of the prediction operator:

$$\begin{aligned}
& \mathbb{E}\left[\left|\Phi_k\tilde{\alpha}_{k-1}^{L_{k-1}}(\xi) - \Phi_k\tilde{\alpha}_{k-1}(\xi)\right|^p\right]^{1/p} \\
= \; & \mathbb{E}\left[\left|\tilde{\alpha}_{k-1}^{L_{k-1}}\phi_k(\xi) - \tilde{\alpha}_{k-1}\phi_k(\xi)\right|^p\right]^{1/p} \\
= \; & \mathbb{E}\left[\left|\left(\tilde{\alpha}_{k-1}^{L_{k-1}} - \tilde{\alpha}_{k-1}\right)\phi_k(\xi)\right|^p\right]^{1/p}
\end{aligned}$$

Hence, by the assumption of the lemma:

$$\begin{aligned}
\mathbb{E}\left[\left|\Phi_k\tilde{\alpha}_{k-1}^{L_{k-1}}(\xi) - \Phi_k\tilde{\alpha}_{k-1}(\xi)\right|^p\right]^{1/p} &\leq c_{k-1,p}\frac{\sup_\zeta|\phi_k(\zeta,\xi)|}{\sqrt{L_{k-1}}} \\
&\leq c_{k-1,p}\frac{\sup_{\zeta,x}\phi_k(\zeta,x)\,||\xi||_\infty}{\sqrt{L_{k-1}}}
\end{aligned}$$

Which gives us the claim of the lemma with: $\hat{c}_{k,p} = c_{k-1,p}\sup_{\zeta,x}\phi_k(x,\zeta)$ □

Lemma 3.3.4 (Sampling). *If, for some finite constant, $\hat{c}_{k,p}$:*

$$\mathbb{E}\left[\left|\Phi_k\tilde{\alpha}_{k-1}^{L_{k-1}}(\xi) - \Phi_k\tilde{\alpha}_{k-1}(\xi)\right|^p\right]^{1/p} \leq \hat{c}_{k,p}\frac{||\xi||_\infty}{\sqrt{L_{k-1}}}$$

Then, there exists a finite constant $\tilde{c}_{k,p}$ such that:

$$\mathbb{E}\left[\left|\alpha_k^{M_k}(\xi) - \alpha_k(\xi)\right|^p\right]^{1/p} \leq \tilde{c}_{k,p}\frac{||\xi||_\infty}{\sqrt{M_k}}$$

Proof. Let $\mathcal{G}_{k-1}$ be the $\sigma-$field generated by the set of all particles until time $k-1$. By a conditioning argument on $\mathcal{G}_{k-1}$, we may view $\left(Y_k^{(i)}\right)_{i\geq 1}$ as independent samples with respective distributions $\left(q_k\left(Y_k^{(i)},\cdot\right)\right)_{i\geq 1}$. Let $\ddot{\alpha}_k^{L_{k-1}}$ be the empirical measure associated with the particles $\left(Y_k^{(i)}\right)_{i\geq 1}$ after the re-weighting step in equation (3.5), i.e.,

$$\ddot{\alpha}_k^{L_{k-1}} = \sum_{i=1}^{L_{k-1}} \tilde{w}_k^{(i)}\delta_{Y_k^{(i)}}$$

and define the sequence of functions $h_i(\cdot) = \frac{\phi_k(X_{k-1}^{(i)},\cdot)\xi(\cdot)}{q_k(X_{k-1}^{(i)},\cdot)} - \phi_k(\xi)\left(X_{k-1}^{(i)}\right)$ and associated measures $\mu_i(\cdot) = q_k(X_k^{(i)},\cdot)$ such that $\mu_i(h_i) = 0$. It is clear that:

$$\frac{\ddot{\alpha}_k^{L_{k-1}}(\xi) - \tilde{\alpha}_k^{L_{k-1}}\phi_k(\xi)}{\tilde{\alpha}_{k-1}^{L_{k-1}}(\mathbf{1})} = \sum_{i=1}^{L_{k-1}}\frac{w_{k-1}^{(i)}h_i(Y_k^{(i)})}{\tilde{\alpha}_{k-1}^{L_{k-1}}(\mathbf{1})}$$

Which allows us to write:

$$\begin{aligned}
& \mathbb{E}\left[\left|\ddot{\alpha}_k^{L_{k-1}}(\xi) - \tilde{\alpha}_k^{L_{k-1}}\phi_k(\xi)\right|^p\right] \\
= \;& \mathbb{E}\left[\left|\tilde{\alpha}_{k-1}^{L_{k-1}}(\mathbf{1})\right|^p \mathbb{E}\left[\left|\sum_{i=1}^{L_{k-1}}\frac{w_{k-1}^{(i)}h_i(Y_k^{(i)})}{\tilde{\alpha}_{k-1}^{L_{k-1}}(\mathbf{1})}\right|^p \middle| \mathcal{G}_{k-1}\right]\right] \\
\leq \;& \mathbb{E}\left[\left|\tilde{\alpha}_{k-1}^{L_{k-1}}(\mathbf{1})\right|^p\right]\frac{2^p d(p)\left(\left|\left|\frac{\phi_k}{q_k}\right|\right|_\infty ||\xi||_\infty\right)^p}{\left(\sqrt{L_{k-1}}\right)^p}
\end{aligned}$$

where the final inequality follows from an application of lemma 3.3.1. This gives us the bound:

$$\mathbb{E}\left[\left|\ddot{\alpha}_k^{L_{k-1}}(\xi) - \tilde{\alpha}_k^{L_{k-1}}\phi_k(\xi)\right|^p\right]^{1/p} \leq \frac{2d(p)^{1/p}C_{k,p}^{\tilde{\alpha}}R_1\,||\xi||_\infty}{\sqrt{L_{k-1}}}$$

Where $C_{k,p}^{\alpha}$ is the finite constant which bounds $\mathbb{E}\left[\left|\tilde{\alpha}_{k-1}^{L_{k-1}}(\mathbf{1})\right|^p\right]^{1/p}$ (see remark 3.3.1).

If we allow $\mathring{\alpha}_k^{J_k}$ be the particle approximation to γ_k obtained by importance sampling from p_k then it is straightforward to verify that, for some finite constant $\hat{B}_k^p$ obtained by using lemma 3.3.1 once again:

$$\mathbb{E}\left[\left|\mathring{\alpha}_k^{J_k}(\xi) - \gamma_k(\xi)\right|^p\right]^{1/p} \leq 2d(p)^{1/p}\left|\left|\frac{\gamma_k}{p_k}\right|\right|_\infty\frac{||\xi||_\infty}{\sqrt{J_k}}$$

And noting that $\alpha_k^{M_k} = \mathring{\alpha}_k^{J_k} + \ddot{\alpha}_k^{L_{k-1}}$ we can apply Minkowski's inequality to obtain:

$$\begin{aligned}\mathbb{E}\left[\left|\alpha_k^{M_k}(\xi) - \Phi_k\tilde{\alpha}_{k-1}^{L_k}\right|^p\right]^{1/p} &\leq 2C_{k,P}^{\alpha}d(p)^{1/p}\frac{\left|\left|\frac{\phi_k}{q_k}\right|\right|_\infty ||\xi||_\infty}{\sqrt{L_{k-1}}} \\ &\quad +2d(p)^{1/p}\left|\left|\frac{\gamma_k}{p_k}\right|\right|_\infty \frac{||\xi||_\infty}{\sqrt{J_k}}\end{aligned}$$

Defining $l_{k-1} = L_{k-1}/M_k$ and $j_k = J_k/M_k$ for convenience, we arrive at the result of the lemma with (making use of (3.8)):

$$\tilde{c}_{k,p} = 2d(p)^{1/p}C_{k,p}^{\alpha}\frac{R_1}{\sqrt{l_{k-1}}} + 2d(p)^{1/p}\frac{R_2}{\sqrt{j_k}}$$

□

Lemma 3.3.5 (Update). *If for some finite constant $\tilde{c}_{k,p}$:*

$$\mathbb{E}\left[\left|\alpha_k^{M_k}(\xi) - \alpha_k(\xi)\right|^p\right]^{1/p} \leq \tilde{c}_{k,p}\frac{||\xi||_\infty}{\sqrt{M_k}}$$

Then there exists a finite constant $\bar{c}_{k,p}$ such that:

$$\mathbb{E}\left[\left|\Psi_k\alpha_k^{M_k}(\xi) - \tilde{\alpha}_k(\xi)\right|^p\right]^{1/p} \leq \bar{c}_{k,p}\frac{||\xi||_\infty}{\sqrt{M_k}}$$

Proof. The proof follows by expanding the norm and using Minkowski's inequality to bound the overall norm. The individual constituents are bounded by the assumption of the lemma.

$$\begin{aligned}&\mathbb{E}\left[\left|\Psi_k\alpha_k^{M_k}(\xi) - \tilde{\alpha}_k(\xi)\right|^p\right]^{1/p} \\ =\ &\mathbb{E}\left[\left|\alpha_k^{M_k}\left(\nu_k\xi + \sum_{z\in Z_k}\frac{\psi_{k,z}\xi}{\kappa_k(z) + \alpha_k^{M_k}(\psi_{k,z})}\right)\right.\right. \\ &\qquad\left.\left.-\alpha_k\left(\nu_k\xi + \sum_{z\in Z_k}\frac{\psi_{k,z}\xi}{\kappa_k(z) + \alpha_k(\psi_{k,z})}\right)\right|^p\right]^{1/p} \\ \leq\ &\mathbb{E}\left[\left|\alpha_k^{M_k}(\nu_k\xi) - \alpha_k(\nu_k\xi)\right|^p\right]^{1/p} + \\ &\sum_{z\in Z_k}\mathbb{E}\left[\left|\alpha_k^{M_k}\left(\frac{\psi_{k,z}\xi}{\kappa_k(z) + \alpha_k^{M_k}(\psi_{k,z})}\right) - \alpha_k\left(\frac{\psi_{k,z}\xi}{\kappa_k(z) + \alpha_k(\psi_{k,z})}\right)\right|^p\right]^{1/p}\end{aligned}$$

Noting that ν_k is a probability, the first term is trivially bounded by the assumption of the lemma:

$$\mathbb{E}\left[\left|\alpha_k^{M_k}(\nu_k\xi) - \alpha_k(\nu_k\xi)\right|^p\right]^{1/p} \leq \tilde{c}_{k,p}\frac{||\nu_k||_\infty\,||\xi||_\infty}{\sqrt{M_k}} \leq \tilde{c}_{k,p}\frac{||\xi||_\infty}{\sqrt{M_k}}$$

In order to bound the second term a little more effort is required, consider a single element of the summation:

$$
\begin{aligned}
& \mathbb{E}\left[\left|\alpha_k^{M_k}\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k^{M_k}(\psi_{k,z})}\right)-\alpha_k\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k(\psi_{k,z})}\right)\right|^p\right]^{1/p} \\
\leq\ & \mathbb{E}\left[\left|\alpha_k^{M_k}\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k^{M_k}(\psi_{k,z})}\right)-\alpha_k^{M_k}\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k(\psi_{k,z})}\right)\right|^p\right]^{1/p}+ \\
& \mathbb{E}\left[\left|\alpha_k^{M_k}\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k(\psi_{k,z})}\right)-\alpha_k\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k(\psi_{k,z})}\right)\right|^p\right]^{1/p} \\
\leq\ & \mathbb{E}\left[\left|\frac{\alpha_k^{M_k}\left(\psi_{k,z}\xi\right)\left[\left(\kappa_k(z)+\alpha_k^{M_k}(\psi_{k,z})\right)-\left(\kappa_k(z)+\alpha_k(\psi_{k,z})\right)\right]}{\left(\left(\kappa_k(z)+\alpha_k^{M_k}(\psi_{k,z})\right)\right)\left(\left(\kappa_k(z)+\alpha_k(\psi_{k,z})\right)\right)}\right|^p\right]^{1/p}+ \\
& \mathbb{E}\left[\left|\alpha_k^{M_k}\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k(\psi_{k,z})}\right)-\alpha_k\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k(\psi_{k,z})}\right)\right|^p\right]^{1/p} \\
\leq\ & 2\frac{\mathbb{E}\left[\left|\alpha_k^{M_k}(\psi_{k,z})-\alpha_k(\psi_{k,z})\right|^p\right]^{1/p}||\xi||_\infty}{\kappa_k(z)+\alpha_k(\psi_{k,z})}
\end{aligned}
$$

Where the final line follows from the positivity assumptions placed upon one of the weight ratios and the likelihood function. This allows us to assert that:

$$
\begin{aligned}
& \sum_{z\in Z_k}\mathbb{E}\left[\left|\alpha_k^{M_k}\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k^{M_k}(\psi_{k,z})}\right)-\alpha_k\left(\frac{\psi_{k,z}\xi}{\kappa_k(z)+\alpha_k(\psi_{k,z})}\right)\right|^p\right]^{1/p} \\
\leq\ & 2Z_k\,||\xi||_\infty\sup_z\frac{\mathbb{E}\left[\left|\alpha_k^{M_k}(\psi_{k,z})-\alpha_k(\psi_{k,z})\right|^p\right]^{1/p}}{\kappa_k(z)+\alpha_k(\psi_{k,z})} \\
\leq\ & \frac{2Z_k\tilde{c}_{k,p}\,||\xi||_\infty}{\sqrt{M_k}}\sup_z\frac{||\psi_{k,z}||_\infty}{\kappa_k(z)+\alpha_k(\psi_{k,z})}
\end{aligned}
$$

Combining this with the previous result and assumption (3.9) gives the result of the lemma with:

$$
\bar{c}_{k,p}=1+2Z_k\tilde{c}_{k,p}\sup_z\frac{R_3}{\kappa_k(z)+\alpha_k(\psi_{k,z})}
$$

□

Lemma 3.3.6 (Resampling). *If, for some finite constant, $\bar{c}_{k,p}$:*

$$
\mathbb{E}\left[\left|\Psi_k\alpha_k^{M_k}(\xi)-\tilde{\alpha}_k(\xi)\right|^p\right]^{1/p}\leq\bar{c}_{k,p}\frac{||\xi||_\infty}{\sqrt{M_k}}
$$

and the resampling scheme is multinomial, then there exists a finite constant $c_{k,p}$ such that:

$$
\mathbb{E}\left[\left|\tilde{\alpha}_k^{L_k}(\xi)-\tilde{\alpha}_k(\xi)\right|^p\right]^{1/p}\leq c_{k,p}\frac{||\xi||_\infty}{\sqrt{L_k}}
$$

Proof. By Minkowski's inequality,

$$
\begin{aligned}
& \mathbb{E}\left[\left|\tilde{\alpha}_k^{L_k}(\xi)-\tilde{\alpha}_k(\xi)\right|^p\right]^{1/p} \\
\leq\ & \mathbb{E}\left[\left|\tilde{\alpha}_k^{L_k}(\xi)-\Psi_k\tilde{\alpha}_k^{M_k}(\xi)\right|^p\right]^{1/p}+\mathbb{E}\left[\left|\Psi_k\tilde{\alpha}_k^{M_k}(\xi)-\tilde{\alpha}_k(\xi)\right|^p\right]^{1/p}
\end{aligned}
$$

By the assumption of the lemma:

$$\mathbb{E}\left[\left|\Psi_k\tilde{\alpha}_k^{M_k}(\xi)-\tilde{\alpha}_k(\xi)\right|^p\right]^{1/p}\leq \bar{c}_{k,p}\frac{||\xi||_\infty}{\sqrt{M_k}}$$

We can bound the remaining term by taking the expectation conditioned upon the sigma algebra generated by the particle ensemble prior to resampling, noting that the resampled particle set is *iid* according to the empirical distribution before resampling:

$$\mathbb{E}\left[\left|\tilde{\alpha}_k^{L_k}(\xi)-\Psi_k\tilde{\alpha}_k^{M_k}(\xi)\right|^p\right]^{1/p}\leq C_{k,p}^D C_{k,p}^R\frac{||\xi||_\infty}{\sqrt{L_k}}$$

where $C_{k,p}^D$ is the upper bound on $\mathbb{E}\left[\left|\tilde{\alpha}_k^{M_k}\right|^p\right]^{1/p}$ approximation at the resampling stage (again, this must exist by remark 3.3.1) and $C_{k,p}^R$ is a constant given by Del Moral's $\mathbb{L}_p$-bound lemma, lemma 3.3.1.

Thus we have the result of the lemma with:

$$c_{k,p}=C_{k,p}^D C_{k,p}^R+\bar{c}_{k,p}\sqrt{\frac{M_k}{L_k}}$$

□

It would be convenient to establish time-uniform convergence results and the stability of the filter with respect to its initial conditions. However, the tools pioneered by [34] and subsequently [100, 25] are not appropriate in the present case: the PHD filter is not a Feynman-Kac flow and decoupling the "prediction" and "update" steps of the filter is not straightforward due to the inherent nonlinearity and the absence of a linear unnormalised flow. It is not obvious how to obtain such results under realistic assumptions.

3.4 Central Limit Theorem

A number of people have published central limit theorems for SMC Methods [41, 34, 25, 100]. As the PHD filtering equations are somewhat different to the standard Bayesian filtering recursion, a number of significant differences need to be addressed in this case. Firstly, the total mass of the filter is variable and unknown rather than fixed at unity and secondly, two importance sampling steps are required at each time. The other main result of this chapter is theorem 3.4.1 which shows that a central limit theorem holds for the SMC approximation of the PHD filter. We adopt an inductive approach to demonstrating that a central limit theorem applies to estimates of the integral of an arbitrary test function under the random measure associated with the particle approximation to the PHD filter.

3.4.1 Formulation

It is convenient to write the PHD in a slightly different form to that given by equations (3.3) and (3.4) for the purposes of considering the central limit theorem.

It is useful to describe the evolution of the PHD filter in terms of selection and mutation operations to allow the errors introduced at each time to be divided into the error propagated forward from earlier times and that introduced by sampling at the present time-step. The formulation used is similar to that employed in the analysis of Feynman-Kac flows [34] under an interacting-process interpretation.

We introduce a potential function, $G_{k,\alpha_k} : E \to \mathbb{R}$ and its associated selection operator $S_{k,\alpha_k} : E \times E \to \mathbb{R}$ and as the selection operator which we employ updates the measure based upon the full distribution at the previous time, we may define the measure $\hat{S}_{k,\alpha_k}(x) = \alpha_k(S_{k,\alpha_k}(\cdot, x))\alpha_k(G_{k,\alpha_k})/\alpha_k(\mathbf{1})$ which is obtained by applying the selection operator to the measure and renormalising to correctly reflect the evolution of the mass of the filter:

$$\begin{aligned} G_{k,\alpha_k}(\cdot) &= \nu_k(\cdot) + \sum_{z \in Z_k} \frac{\psi_{k,z}(\cdot)}{\kappa_k(z) + \alpha_k(\psi_{k,z})} \\ S_{k,\alpha_k}(x, y) &= \frac{\alpha_k(y) G_{k,\alpha_k}(y)}{\alpha_k(G_{k,\alpha_k})} \\ \hat{S}_{k,\alpha_k}(\cdot) &= \alpha_k(\cdot) G_{k,\alpha_k}(\cdot) \end{aligned}$$

For clarity of exposition, we have assumed in this section that N particles are propagated forward from each time step to the next and that $\eta_k N$ particles are introduced to account for spontaneous births at time k (i.e., in the notation of the previous section, $L_k = N$ and $J_k = \eta_k N$). The notation $N_k = (1 + \eta_k)N$ is also used for notational convenience.

The interpretation of this formulation is slightly different and perhaps more intuitive. Update and resampling occur simultaneously and comprise the selection step, while prediction follows as a mutation operation. Here we use α_k to refer to the predicted filter as in (3.3), and it is not necessary to make any reference to the updated filter. We separate the spontaneous birth component of the measure from that which depends upon the past and write the PHD recursion as:

$$\begin{aligned} \alpha_k(\xi) &= \hat{\alpha}_k(\xi) + \mathring{\alpha}_k(\xi) \\ \hat{\alpha}_k(\xi) &= \hat{S}_{k-1,\alpha_{k-1}} \phi_k(\xi) \\ \mathring{\alpha}_k(\xi) &= \gamma_k(\xi), \end{aligned}$$

we note that the form of $\hat{S}_{k-1,\alpha_{k-1}}$ is such that this is a recursive description.

The Particle Approximation. Within this section, the particle approximation described previously can be restated as the following iterative procedure, algorithm 3.2. This provides an alternative view of algorithm 3.1 given in section 3.2.3, with the additional assumption that the number of particles propagated forward at each time step is constant, with no explicit reference to $\tilde{\alpha_k}^{L_k}$. As we are concerned with asymptotic results the increased clarity more than compensates for the slight reduction in generality.

Algorithm 3.2 A reformulation of the SMC PHD.

Let the particle approximation prior to resampling at time $k-1$ be of the form

$$\alpha_{k-1}^{N_{k-1}} = \frac{1}{N}\sum_{i=1}^{N_{k-1}} \tilde{w}_{k-1}^{(i)} \delta_{X_{k-1}^{(i)}}$$

Sample N particles to propagate forward via the selection operator:

$$\left\{\text{Sample } Y_k^{(i)} \sim S_{k-1,\alpha_{k-1}^{N_{k-1}}}(\cdot)\right\}_{i=1}^{N}$$

Mutate these N particles.

$$\left\{\text{Sample } X_k^{(i)} \sim q_k(Y_k^{(i)},\cdot)\right\}_{i=1}^{N}$$

Introduce $\eta_k N$ particles to account for the possibility of births.

$$\left\{\text{Sample } X_k^{(i)} \sim p_k(\cdot)\right\}_{i=N+1}^{N_k}$$

Define the particle approximation at time k as $\alpha_k^{N_k} = \hat{\alpha}_k^N + \mathring{\alpha}_k^{\eta_k N}$ where:

$$\hat{\alpha}_k^N = \frac{1}{N}\sum_{i=1}^{N} \tilde{w}_k^{(i)} \delta_{X_k^{(i)}} \text{ and } \mathring{\alpha}_k^{\eta_k N} = \frac{1}{N}\sum_{i=N+1}^{N_k} \tilde{w}_k^{(i)} \delta_{X_k^{(i)}}$$

and the weights are given by:

$$\tilde{w}_k^{(i)} = \begin{cases} \alpha_{k-1}^{N_{k-1}}\left(G_{k-1,\alpha_{k-1}^{N_{k-1}}}\right)\frac{\phi_k(Y_k^{(i)},X_k^{(i)})}{q_k(Y_k^{(i)},X_k^{(i)})} & i \in \{1,\dots,N\} \\ \frac{1}{\eta_k}\frac{\gamma_k(X_k^{(i)})}{p_k(X_k^{(i)})} & i \in \{N+1,\dots,N_k\} \end{cases}$$

3.4.2 Variance Recursion

Theorem 3.4.1 (Central Limit Theorem). *The particle approximation to the PHD filter follows a central limit theorem with some finite variance for all continuous bounded test functions $\xi : E \to \mathbb{R}^d$, at all times $k \geq 0$:*

$$\lim_{N\to\infty} \sqrt{N}\left[\alpha_k^{N_k}(\xi) - \alpha_k(\xi)\right] \xrightarrow{d} \mathcal{N}\left(0, \sigma_k^2(\xi)\right)$$

provided that the result holds at time 0, which it does, for example, if the filter is initialised by obtaining samples from a normalised version of the true filter by importance sampling and weighting them correctly.

In all cases we prove the case for scalar-valued test functions and the generalisation to the vector-valued case follows directly via the Cramer-Wold device [8, p.397].

Proof. By assumption, the result of the theorem holds at time 0. Using induction the result can be shown to hold for all times by the sequence of lemmas, lemma 3.4.1-3.4.4, that follow.

The core of the proof is the following decomposition:

$$\begin{aligned}\alpha_k^{N_k}(\xi) - \alpha_k(\xi) &= \hat{\alpha}_k^N(\xi) - \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k \times \frac{\phi_k}{q_k}\right)(\xi) + \\ &\quad \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k \times \frac{\phi_k}{q_k}\right)(\xi) - \hat{\alpha}_k(\xi) + \\ &\quad \mathring{\alpha}_k^{(\eta_k N)}(\xi) - \mathring{\alpha}_k(\xi)\end{aligned}$$

Consistent with the notation defined in section 3.2.1, $\hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k \times \frac{\phi_k}{q_k}\right)(\xi)$ is to be understood as

$$\int \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}(du) \int q_k(u, dv) \frac{\phi_k(u,v)}{q_k(u,v)} \xi(v)$$

i.e., $q_k \times \frac{\phi_k}{q_k}$ defines a new transition kernel from E to E.

The first term in this decomposition accounts for errors introduced at time k by using a particle approximation of the prediction step and this is shown to converge to a centred normal distribution of variance $\hat{V}_k(\xi)$ in lemma 3.4.1. The second term describes the errors propagated forward from previous times, and is shown to follow a central limit theorem with variance $\ddot{V}_k(\xi)$ in lemma 3.4.2. The final term corresponds to sampling errors in the spontaneous birth components of the filter and this is shown to follow a central limit theorem with variance $\mathring{V}_k(\xi)$ in lemma 3.4.3.

Lemma 3.4.4 shows that the result of combining the three terms of the decomposition is a random variable which itself follows a central limit theorem with variance:

$$\sigma_k^2(\xi) = \hat{V}_k(\xi) + \ddot{V}_k(\xi) + \mathring{V}_k(\xi)$$

which is precisely the result of the theorem for scalar test functions.

In the case of vector test functions, the result follows by the Cramer-Wold device, applied to any linear combination of their components, and the covariance matrix is denoted $\Sigma_k(\xi) = [\Sigma_k(\xi_i, \xi_j)]$. □

Lemma 3.4.1 (Selection-prediction Sampling Errors). *The selection-prediction sampling error (due to steps 2 and 3) at time k converges to a normally distributed random variable of finite variance as the size of the particle ensemble tends towards infinity:*

$$\lim_{N\to\infty} \sqrt{N}\left(\hat{\alpha}_k^N(\xi) - \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k \times \frac{\phi_k}{q_k}\right)(\xi)\right) \xrightarrow{d} \mathcal{N}\left(0, \hat{V}_k(\xi)\right)$$

Proof. Consider the term under consideration:

$$\begin{aligned}&\hat{\alpha}_k^N(\xi) - \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k \times \frac{\phi_k}{q_k}\right)(\xi) \\ =\ & \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{w}_k^{(i)}\xi(X_k^{(i)}) - \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k \times \frac{\phi_k}{q_k}\right)(\xi)\right) \\ =\ & \frac{1}{\sqrt{N}}\sum_{i=1}^{N} U_{k,i}^N\end{aligned}$$

where

$$U_{k,i}^N = \frac{\tilde{w}_k^{(i)}\xi(X_k^{(i)}) - \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k \times \frac{\phi_k}{q_k}\right)(\xi)}{\sqrt{N}}$$

Let $\mathcal{H}_k^N = \sigma\left(\left\{X_n^{(i)}, \tilde{w}_n^{(i)}\right\}_{i=1}^{N_n} : n = 0, \ldots, k\right)$ be the sigma algebra generated by the particle ensembles occurring at or before time k and further let $\mathcal{H}_{k,j}^N = \sigma\left(\mathcal{H}_{k-1}^N, \left\{X_k^{(i)}, \tilde{w}_k^{(i)}\right\}_{i=1}^{j}\right)$.

It is evident that conditioned upon $\mathcal{H}_{k-1}^N$, $\left\{Y_k^{(i)}, X_k^{(i)}\right\}_{i=1}^N$ are *iid* samples from the product distribution $S_{k-1,\alpha_{k-1}^{N_{k-1}}}(y)q_k(y,x)$ and, therefore:

$$\mathbb{E}\left[U_{k,i}^N \middle| \mathcal{H}_{k,i-1}^N\right] = \mathbb{E}\left[U_{k,i}^N \middle| \mathcal{H}_{k-1}^N\right] = 0$$

Furthermore, conditionally, $U_{k,i}^N$ has finite variance, which follows from assumption (3.8) and the assumption that the observation set and the initial mass of the filter are finite:

$$\begin{aligned}
\mathbb{E}\left[\left(U_{k,i}^N\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(U_{k,i}^N\right)^2 \middle| \mathcal{H}_{k,i-1}^N\right]\right] \\
&= \frac{1}{N}\mathbb{E}\left[\left(\tilde{w}_k^{(i)}\xi(X_k^{(i)})\right)^2 - \left(\hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k \times \frac{\phi_k}{q_k}\right)(\xi)\right)^2\right] \\
&< \frac{2\mathbb{E}\left[\left(\alpha_{k-1}^{N_{k-1}}(\mathbf{1}) + |Z_{k-1}|\right)^2\right] R_1^2 \,||\xi||_\infty^2}{N} < \infty
\end{aligned}$$

Noting that, by the $\mathbb{L}_p$ convergence result of theorem 3.3.1 the expectation may be bounded above uniformly in N. We have that $\forall t \in [0,1], \epsilon > 0$:

$$\lim_{N\to\infty} \sum_{i=1}^{\lfloor Nt \rfloor} \mathbb{E}\left[\left(U_{k,i}^N\right)^2 \mathbb{I}_{|U_{k,i}^N|>\epsilon} \middle| \mathcal{H}_{k,i-1}^N\right] \xrightarrow{p} 0 \tag{3.12}$$

By noting that the following convergence result holds (and this can be seen by expanding each term and using theorem 3.3.1, noting that if two sequences of bounded random variables converge to two finite limits, then the product of those sequences converges to the product of their respective limits and that for nonzero random variables the same is true of the quotient of those sequences)

$$\alpha_{k-1}^{N_{k-1}}\left(G_{k,\alpha_{k-1}^{N_{k-1}}}\right)\alpha_{k-1}^{N_{k-1}}\left(G_{k,\alpha_{k-1}^{N_{k-1}}}\left(\phi_k \times \frac{\phi_k}{q_k}\right)(\xi)\right) - \alpha_{k-1}^{N_{k-1}}\left(G_{k,\alpha_{k-1}^{N_{k-1}}}\phi_k(\xi)\right)^2 \tag{3.13}$$

$$\xrightarrow{p} \alpha_{k-1}\left(G_{k,\alpha_{k-1}}\right)\alpha_{k-1}\left(G_{k,\alpha_{k-1}}\left(\phi_k \times \frac{\phi_k}{q_k}\right)(\xi)\right) - \alpha_{k-1}\left(G_{k,\alpha_{k-1}}\phi_k(\xi)\right)^2 \tag{3.14}$$

it is apparent (as (3.13) is equal to $\frac{N}{\lfloor Nt \rfloor}$ times (3.15) and (3.14) to $\frac{1}{t}$ times (3.16)) that

$$\sum_{k=1}^{\lfloor Nt \rfloor} \mathbb{E}\left[\left(U_{k,i}^N\right)^2 \middle| \mathcal{H}_{k,i-1}^N \right] = \frac{\lfloor Nt \rfloor}{N} \alpha_{k-1}^{N_{k-1}} \left(G_{k-1,\alpha_{k-1}^{N_{k-1}}} \right)^2 \times \left[S_{k-1,\alpha_{k-1}^{N_{k-1}}} \left(\phi_k \times \frac{\phi_k}{q_k} \right) (\xi^2) - S_{k-1,\alpha_{k-1}^{N_{k-1}}} (\phi_k(\xi))^2 \right] \tag{3.15}$$

$$\xrightarrow{p} t\alpha_{k-1} \left(G_{k-1,\alpha_{k-1}} \right)^2 \times \left[S_{k-1,\alpha_{k-1}} \left(\phi_k \times \frac{\phi_k}{q_k} \right) (\xi^2) - S_{k-1,\alpha_{k-1}} (\phi_k(\xi))^2 \right] \tag{3.16}$$

From this, it can be seen that for each N, the sequence

$$\left(U_{k,i}^N, \mathcal{H}_{k,i}^N\right), \quad 1 \leq i \leq N$$

is a square-integrable martingale difference which satisfies the Lindeberg condition (3.12) and hence a martingale central limit theorem may be invoked (see, for example, [140, page 543])) to show that:

$$\lim_{N\to\infty} \sqrt{N} \left(\hat{\alpha}_k^N(\xi) - \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}} \left(q_k \times \frac{\phi_k}{q_k} \right) (\xi) \right) \xrightarrow{d} \mathcal{N}\left(0, \hat{V}_k(\xi)\right)$$

where,

$$\begin{aligned} \hat{V}_k(\xi) &= \alpha_{k-1} \left(G_{k-1,\alpha_{k-1}} \right)^2 \left[S_{k-1,\alpha_{k-1}} \left(\phi_k \times \frac{\phi_k}{q_k} \right) (\xi^2) - S_{k-1,\alpha_{k-1}} (\phi_k(\xi))^2 \right] \\ &= \alpha_{k-1} (G_{k-1,\alpha_{k-1}}) \hat{S}_{k-1,\alpha_{k-1}} \left(\phi_k \times \frac{\phi_k}{q_k} \right) (\xi^2) - \hat{S}_{k-1,\alpha_{k-1}} (\phi_k(\xi))^2 \end{aligned}$$

□

Lemma 3.4.2 (Propagated Errors). *The error resulting from propagating the particle approximation forward rather than the true filter has an asymptotically normal distribution with finite variance.*

$$\hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}} \left(q_k \times \frac{\phi_k}{q_k} \right) (\xi) - \hat{\alpha}_k(\xi) \xrightarrow{d} \mathcal{N}\left(0, \ddot{V}_k(\xi)\right)$$

Proof. Direct expansion of the potential allows us to express this difference as:

$$\begin{aligned} & \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}} \left(q_k \times \frac{\phi_k}{q_k} \right) (\xi) - \hat{\alpha}_k(\xi) \\ = \; & \hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}} \phi_k(\xi) - \hat{S}_{k-1,\alpha_{k-1}} \phi_k(\xi) \\ = \; & \alpha_{k-1}^{N_{k-1}} \left(\phi_k(\xi) G_{k-1,\alpha_{k-1}^{N_{k-1}}} \right) - \alpha_{k-1} \left(\phi_k(\xi) G_{k-1,\alpha_{k-1}} \right) \\ = \; & \alpha_{k-1}^{N_{k-1}} \left(\phi_k(\xi) \nu_{k-1} \right) - \alpha_{k-1} \left(\phi_k(\xi) \nu_{k-1} \right) + \\ & \sum_{z \in Z_{k-1}} \frac{\alpha_{k-1}^{N_{k-1}}(\Delta_{k-1,z}) - \alpha_{k-1}(\Delta_{k-1,z})}{\kappa_{k-1}(z) + \alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z})} \end{aligned}$$

where $\Delta_{k-1,z} = \psi_{k-1,z}\phi_k(\xi) - \frac{\alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))}{\kappa_{k-1}(z)+\alpha_{k-1}(\psi_{k-1,z})}\psi_{k-1,z}$ and the final equality can be shown to hold by considering a single term in the summation thus:

$$\begin{aligned}
&\frac{\alpha_{k-1}^{N_k}(\psi_{k-1,z}\phi_k(\xi))}{\kappa_{k-1}(z)+\alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z})} - \frac{\alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))}{\kappa_{k-1}(z)+\alpha_{k-1}(\psi_{k-1,z})} \\
= &\frac{1}{\kappa_{k-1}(z)+\alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z})}\left[\alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z}\phi_k(\xi)) - \alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))\right. \\
&\left.+\alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi)) - \frac{\kappa_{k-1}(z)+\alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z})}{\kappa_{k-1}(z)+\alpha_{k-1}(\psi_{k-1,z})}\alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))\right] \\
= &\frac{1}{\kappa_{k-1}(z)+\alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z})}\left[\alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z}\phi_k(\xi)) - \alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))\right. \\
&\left.+\frac{\alpha_{k-1}(\psi_{k-1,z}) - \alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z})}{\kappa_{k-1}(z)+\alpha_{k-1}(\psi_{k-1,z})}\alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))\right] \\
= &\frac{\alpha_{k-1}^{N_{k-1}}\left(\psi_{k-1,z}\phi_k(\xi) - \frac{\alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))\psi_{k-1,z}}{\kappa_{k-1}(z)+\alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))\psi_{k-1,z}}\right)}{\kappa_{k-1}(z)+\alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z})} \\
&-\frac{\alpha_{k-1}\left(\psi_{k-1,z}\phi_k(\xi) - \frac{\alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))\psi_{k-1,z}}{\kappa_{k-1}(z)+\alpha_{k-1}(\psi_{k-1,z}\phi_k(\xi))\psi_{k-1,z}}\right)}{\kappa_{k-1}(z)+\alpha_{k-1}^{N_{k-1}}(\psi_{k-1,z})}
\end{aligned}$$

If we set

$$\Delta_{k-1} = \left[\nu_k\phi_k(\xi), \Delta_{k-1,Z_{k-1,1}}, \ldots, \Delta_{k-1,Z_{k-1,|Z_{k-1}|}}\right]$$

where Z_{k-1}^i denotes the i^{th} element of the set Z_{k-1}, and,

$$\rho_{k-1}^{N_{k-1}} = \left[1, \frac{1}{\kappa_{k-1}(Z_{k-1,1}) + \alpha_{k-1}^{N_{k-1}}(\psi_{k-1,Z_{k-1,1}})}, \ldots, \right. \\ \left. \frac{1}{\kappa_{k-1}(Z_{k-1,|Z_{k-1}|}) + \alpha_{k-1}^{N_{k-1}}(\psi_{k-1,Z_{k-1,|Z_{k-1}|}})}\right]^T$$

Then the quantity of interest may be written as an inner product:

$$\left\langle \rho_{k-1}^{N_{k-1}}, \alpha_{k-1}^{N_{k-1}}\left(\Delta_{k-1}\right) - \alpha_{k-1}\left(\Delta_{k-1}\right)\right\rangle$$

We know from theorem 3.3.1 that $\rho_{k-1}^{N_{k-1}} \xrightarrow{p} \rho_{k-1}$, where

$$\rho_{k-1} = \left[1, \frac{1}{\kappa_{k-1}(Z_{k-1,1}) + \alpha_{k-1}(\psi_{k-1,Z_{k-1,1}})}, \ldots, \right. \\ \left. \frac{1}{\kappa_{k-1}(Z_{k-1,|Z_{k-1}|}) + \alpha_{k-1}(\psi_{k-1,Z_{k-1,|Z_{k-1}|}})}\right]^T$$

And furthermore, we know by the induction assumption that each $\alpha_{k-1}^{N_{k-1}}\left(\Delta_{k-1}\right) - \alpha_{k-1}\left(\Delta_{k-1}\right)$ is asymptotically normal with zero mean and some known variance, $\Sigma_{k-1}(\Delta_{k-1})$. By Slutzky's theorem, therefore, the quantity of interest converges to a normal distribution of mean zero and variance $\ddot{V}_k(\xi) = \rho_{k-1}^T\Sigma_{k-1}(\Delta_{k-1})\rho_{k-1}$.

□

Lemma 3.4.3 (Spontaneous Births). *The error in the particle approximation to the spontaneous birth element of the PHD converges to a normal distribution with finite variance:*

$$\lim_{N\to\infty} \sqrt{N}\left[\mathring{\alpha}_k^{\eta_k N}(\xi) - \mathring{\alpha}_k(\xi)\right] \xrightarrow{d} \mathcal{N}\left(0, \mathring{V}_k(\xi)\right)$$

Proof.

$$\mathring{\alpha}_k^{\eta_k N}(\xi) - \mathring{\alpha}_k(\xi) = \frac{\gamma_k(\mathbf{1})}{\eta_k N} \sum_{j=N+1}^{N_k} \left(\frac{\gamma_k(X_k^{(j)})}{\gamma_k(\mathbf{1}) p_k(X_k^{(j)})} \xi(X_k^{(j)}) - \frac{\gamma_k(\xi)}{\gamma_k(\mathbf{1})} \right)$$

Of course, the particles appearing within this sum are *iid* according to p_k and this corresponds to $\gamma_k(\mathbf{1})$ multiplied by the importance sampling estimate giving us the standard result:

$$\sqrt{\eta_k N}\left[\frac{\mathring{\alpha}_k^{\eta_k N}(\xi) - \mathring{\alpha}_k(\xi)}{\gamma_k(\mathbf{1})}\right] \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}_{p_k}\left(\frac{\gamma_k}{\gamma_k(\mathbf{1}) p_k}\xi\right)\right)$$

which is precisely the result of the lemma with:

$$\mathring{V}_k(\xi) = \frac{1}{\eta_k}\left[\gamma_k\left(\frac{\gamma_k}{p_k}\xi^2\right) - \gamma_k(\xi)^2\right]$$

□

Lemma 3.4.4 (Combining Terms). *Using the results of lemmas 3.4.1–3.4.3 it follows that $\alpha_k^{N_k}(\xi) - \alpha_k(\xi)$ satisfies the central limit theorem:*

$$\lim_{N\to\infty} \sqrt{N}\left(\alpha_k^{N_k}(\xi) - \alpha_k(\xi)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_k(\xi)\right)$$

where the asymptotic variance is given by $\sigma_k(\xi) = \hat{V}_k(\xi) + \ddot{V}_k(\xi) + \mathring{V}_k(\xi)$.

Proof. The proof follows the method of [100]. The characteristic function of the random variable of interest is

$$\Upsilon_k(t) = \mathbb{E}\left[\exp\left(it\sqrt{N}\left(\alpha_k^{N_k}(\xi) - \alpha_k(\xi)\right)\right)\right]$$

As the particles associated with the spontaneous birth term of the PHD are independent of those propagated forward from the previous time we can write:

$$\begin{aligned}\Upsilon_k(t) &= \mathbb{E}\left[\exp\left(it\sqrt{N}\left(\mathring{\alpha}_k^{\eta_k N}(\xi) - \mathring{\alpha}_k(\xi)\right)\right)\right] \\ &\quad \times \mathbb{E}\left[\exp\left(it\sqrt{N}\left(\hat{\alpha}_k^{N}(\xi) - \hat{\alpha}_k(\xi)\right)\right)\right]\end{aligned}$$

The first term of this expansion is the characteristic function of a normal random variable, so all that remains is to show that the same is true of the second term. Using the same decomposition as above, we may write:

$$
\begin{aligned}
& \mathbb{E}\left[\exp\left(it\sqrt{N}\left(\hat{\alpha}_k^N(\xi)-\hat{\alpha}_k(\xi)\right)\right)\right] \\
= \; & \mathbb{E}\left[\overbrace{\mathbb{E}\left[\exp\left(it\sqrt{N}\left\{\hat{\alpha}_k^N(\xi)-\hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k\times\frac{\phi_k}{q_k}\right)(\xi)\right\}\right)\middle|\,\mathcal{H}_{k-1}^N\right]}^{\mathsf{A}}\right. \\
& \left.\underbrace{\times\exp\left(it\sqrt{N}\left\{\hat{S}_{k-1,\alpha_{k-1}^{N_{k-1}}}\left(q_k\times\frac{\phi_k}{q_k}\right)(\xi)-\hat{\alpha}_k(\xi)\right\}\right)}_{\mathsf{B}}\right] \\
= \; & \mathbb{E}\left[\left(\mathsf{A}-\exp\left(-\frac{t^2\hat{V}_k(\xi)}{2}\right)\right)\mathsf{B}\right]+\exp\left(-\frac{t^2\hat{V}_k(\xi)}{2}\right)\mathbb{E}[\mathsf{B}]
\end{aligned}
$$

All that remains is to show that the first term in this expansion vanishes and we will have shown that the characteristic function of interest $\Upsilon_k(t)$ corresponds to a Gaussian distribution as it can be expressed as the product of three Gaussian characteristic functions. Furthermore, it must have variance equal to the sum of the variances of the three constituent Gaussians which is exactly the result which we wish to prove.

By the conditionally *iid* nature of the particles, we can write:

$$
\mathsf{A}=\mathbb{E}\left[\exp\left(it\sum_{j=1}^{N}U_{k,j}^N\right)\middle|\,\mathcal{H}_{k-1}^N\right]=\mathbb{E}\left[\exp\left(itU_{k,1}^N\right)\middle|\,\mathcal{H}_{k-1}^N\right]^N
$$

Hence:

$$
\left|\mathsf{A}-\exp\left(-\frac{t^2\hat{V}_k(\xi)}{2}\right)\right|=\left|\mathbb{E}\left[\exp\left(itU_{k,1}^N\right)\middle|\,\mathcal{H}_{k-1}^N\right]^N-\exp\left(-\frac{t^2\hat{V}_k(\xi)/N}{2}\right)^N\right|
$$

Using the same result as [100] (i.e. that $|u^N-v^N|\leq N|u-v|\forall|u|\leq 1,|v|\leq 1$) we obtain:

$$
\left|\mathsf{A}-\exp\left(-\frac{t^2\hat{V}_k(\xi)}{2}\right)\right|\leq N\left|\mathbb{E}\left[\exp\left(itU_{k,1}^N\right)\middle|\,\mathcal{H}_{k-1}^N\right]-\exp\left(-\frac{t^2\hat{V}_k(\xi)/N}{2}\right)\right|
$$

The following decomposition can be used to show that this difference converges to zero as $N\to\infty$:

$$
\begin{aligned}
& \mathbb{E}\left[\exp\left(itU_{k,1}^N\right)-\exp\left(-\frac{t^2\hat{V}_k(\xi)/N}{2}\right)\middle|\,\mathcal{H}_{k-1}^N\right] \\
= \; & \mathbb{E}\left[\exp\left(itU_{k,1}^N\right)\middle|\,\mathcal{H}_{k-1}^N\right]-\left(1-\frac{t^2(U_{k,1}^N)^2}{2}\right)+ && (3.17)\\
& \left(1-\frac{t^2(U_{k,1}^N)^2}{2}\right)-\exp\left(-\frac{t^2}{2}\mathbb{E}\left[\left(U_{k,1}^N\right)^2\middle|\,\mathcal{H}_{k-1}^N\right]\right)+ && (3.18)\\
& \exp\left(-\frac{t^2}{2}\mathbb{E}\left[\left(U_{k,1}^N\right)^2\middle|\,\mathcal{H}_{k-1}^N\right]\right)-\exp\left(-\frac{t^2\hat{V}_k(\xi)/N}{2}\right) && (3.19)
\end{aligned}
$$

We now show that the product of N and the expectation of each of these terms converges to zero. First, consider (3.17). We can represent e^{iy} as $1+iy-\frac{y^2}{2}+\frac{|y|^3}{3!}\theta(y)$ for some suitable function $\theta(y), |\theta| < 1$. Thus, as $U^N_{k,1}$ is a martingale increment:

$$\begin{aligned} & \left| \mathbb{E}\left[\exp\left(itU^N_{k,1}\right) - \left(1 - \frac{t^2(U^N_{k,1})^2}{2}\right) \middle| \mathcal{H}^N_{k-1} \right] \right| \\ \leq & \frac{t^3}{6}\mathbb{E}\left[\left|U^N_{k,1}\right|^3 \middle| \mathcal{H}^N_{k-1} \right] \\ \leq & \frac{t^3}{6N^{3/2}} 2^3 \left|\left|\xi\right|\right|^3_\infty \left(\alpha^{N_{k-1}}_{k-1}(\mathbf{1})R_1 + |Z_{n-1}|R_1 \right)^3 \end{aligned}$$

and N times the expectation of this quantity converges to zero as $N \to \infty$.

To deal with (3.18) note that $1 - u \leq exp(-u) \leq 1 - u + u^2 \quad \forall u \geq 0$. Setting

$$u = \frac{t^2}{2}\mathbb{E}\left[\left(U^N_{k,1}\right)^2 \middle| \mathcal{H}^N_{k-1} \right]$$

one obtains:

$$\begin{aligned} & \mathbb{E}\left[\left| \left(1 - \frac{t^2(U^N_{k,1})^2}{2}\right) - \exp\left(-\frac{t^2}{2}\mathbb{E}\left[\left(U^N_{k,1}\right)^2 \middle| \mathcal{H}^N_{k-1} \right] \right) \right| \middle| \mathcal{H}^N_{k-1} \right] \\ \leq & \frac{t^4}{4}\mathbb{E}\left[\left(U^N_{k,1}\right)^2 \middle| \mathcal{H}^N_{k-1} \right]^2 \\ \leq & \frac{t^4}{4}\frac{1}{N^2} 4 \left|\left|\xi\right|\right|^4_\infty \left(\alpha^{N_{k-1}}_{k-1}(\mathbf{1})R_1 + |Z_{k-1}|R_1 \right)^4 \end{aligned}$$

and once again, the expectation of N times the quantity of interest converges to zero.

Finally, (3.19) can be shown to vanish by considering the following exponential bound. For $v \geq u \geq 0$, we can write $|e^{-u} - e^{-v}| \leq |u - v|$ by employing the intermediate value theorem, and this yields:

$$\begin{aligned} & \left| \exp\left(-\frac{t^2}{2}\mathbb{E}\left[\left(U^N_{k,1}\right)^2 \middle| \mathcal{H}^N_{k-1} \right] \right) - \exp\left(-\frac{t^2\hat{V}_k(\xi)/N}{2} \right) \right| \\ \leq & \frac{t^2}{2N} \left| N\mathbb{E}\left[\left(U^N_{k,1}\right)^2 \middle| \mathcal{H}^N_{k-1} \right]^2 - \hat{V}_k(\xi) \right| \end{aligned}$$

which can be exploited by noting that:

$$\begin{aligned} N\mathbb{E}\left[\left(U^N_{k,i}\right)^2 \middle| \mathcal{H}^N_{k,i-1} \right] = & \left[\alpha^{N_{k-1}}_{k-1}(G_{k-1,\alpha^{N_{k-1}}_{k-1}})\hat{S}_{k-1,\alpha^{N_{k-1}}_{k-1}}\left(\phi_k \times \frac{\phi_k}{q_k}\right)(\xi^2) \right. \\ & \left. -\hat{S}_{k-1,\alpha^{N_{k-1}}_{k-1}}(\phi_k(\xi))^2 \right] \qquad (3.20) \\ \hat{V}_k(\xi) = & \left[\alpha_{k-1}(G_{k-1,\alpha_{k-1}})\hat{S}_{k-1,\alpha_{k-1}}\left(\phi_k \times \frac{\phi_k}{q_k}\right)(\xi^2) \right. \\ & \left. -\hat{S}_{k-1,\alpha_{k-1}}(\phi_k(\xi))^2 \right] \qquad (3.21) \end{aligned}$$

As (3.20) converges to (3.21) in probability (cf. lemma 3.4.2) and (3.20) is bounded above, (3.20) converges to (3.21) in $\mathbb{L}_1$ and the result we seek follows. Consequently, (3.19) vanishes and we have the result of the lemma. □

3.5 Dynamic Clustering with a PHD Filter

One of the principle limitations of the PHD filter is the assumption that each target generates at most one observation. For the purpose of dynamic clustering or group tracking, it would be useful to associate large numbers of observations with a single "target" whose parameters describing the entire group. An attempt to generalise the PHD recursion to such a measurement model was made by Mahler [106]. However, there is a particularly strong assumption within that derivation, which appears to prevent the methodology from being applicable in any realistic situation. An attempt is made below to obtain a more general recursion for a first moment approximation of the dynamic clustering problem, however, no computationally feasible algorithm is obtained.

3.5.1 Background and Formulation

To enable easy comparison with [106] we make use of broadly the same formulation, however, the notation used here has been chosen to be consistent with the remainder of this thesis.

The probability generating functional (*pgfl*) associated with the intensity of a Poisson process is used to obtain the recursion. The *pgfl* is defined as a complex-valued function, ξ, which acts upon a suitable class of complex-valued measurable functions, $\mathcal{C} = \{\xi : E \to \mathbb{C} : ||\xi||_\infty \leq 1\}$ on the state space of the point process, with the property that it completely characterises the intensity of a function. We make use of the property described, for example, by [107]: loosely speaking, the functional derivative of a *pgfl* at a point, evaluated at the unit function corresponds to the first moment density at that point.

We seek a recursion which, analogously to Bayesian filtering, allows us to *predict* the state of a set of cluster parameters at a later time based upon an estimate of the present state, combined with knowledge of the target dynamics and, furthermore, to *update* this prediction based upon a subsequent observation. This amounts to a variation upon equations (3.3-3.4) which are suitable for a measurement model which allows each object to generate a set of observations.

As in the case of the standard PHD filter described above, we make use of the intensity of a point process to describe our knowledge of the state of the system. At time k, we assume we have some measure α_k which contains the estimate of this intensity at that time. We wish to predict the intensity at time $k+1$ based upon this knowledge, and this can be done by the approach described in section 3.5.2 and we denote this predicted intensity $\hat{\alpha}_k$. Having obtained this prediction, we wish to update it to take into account the set of measurements obtained at time

$k+1$ to obtain a new measure α_{k+1} and approaches to doing this are described in section 3.5.3

3.5.2 The Prediction Equation

The difference between the algorithm presented here and that of [106] lies entirely within the update step, and the prediction equation which we require is exactly that which Mahler obtained:

$$\hat{\alpha}_k(dx = d(a,u)) = \alpha_k \phi_k(dx = d(a,u)) + \gamma_k(dx = d(a,u)), \tag{3.22}$$

where, for convenience, the state is decomposed as $x = (a,u)$ where a denotes the size of a cluster and u describes its parameters.

This is essentially identical to the standard PHD prediction step 3.3, and each term has a meaning analogous to that in the standard PHD filter, although in this case the parameters are those of a cluster of objects, rather than an individual object and the dynamic model applies to clusters rather than individual objects.

3.5.3 The Update Equation

It is convenient to use the probability generating functional approach to obtain the recursion. We do not distinguish between functions and functionals by using different styles of bracket as this would not be consistent with usage elsewhere in this thesis. The *pgfl* of the PHD at time $k+1$ after prediction, but before the data update is given by:

$$\hat{G}_k(h) = \sum_{i=0}^{\infty} \frac{1}{i!} \int h(x_1) \dots h(x_i) \hat{p}_k(\{x_1, \dots, x_i\} | Z_{1:k}) \lambda^i \left(\mathrm{d}(x_1, \dots, x_i)\right)$$

where $\hat{p}_k$ is a density with respect to the unnormalized Poisson process over the space of interest and $x_i = (a_i, u_i)$ contains the size and the parameters of the i^{th} of its points. In our case, these correspond to the parameters of one of the clusters which is to be tracked.

At this point Mahler makes the rather strong assumption that all of the clusters are *completely correlated* and hence that the joint distribution $\hat{p}_k$ can be reduced to the rather simple form:

$$\hat{p}_k(X|Z_{1:k}) = |X|! \sigma_{|X|} \delta_{x_1}(x_i) \dots \delta_{x_1}(x_2) \hat{s}_k(x_1)$$

where, $X = \left\{x_1, \dots, x_{|X|}\right\}$, and,

$$\hat{s}_k(x_1) = \hat{\alpha}_k(x_1) / \hat{\alpha}_k(\mathbf{1})$$

and σ_i is the probability that i clusters are present, assuming that the predicted PHD is the intensity of a Poisson process from which the clusters were drawn (a quantity for which there is, in general, no closed form expression).

Alternatively, one might suppose that the clusters are independent (*completely uncorrelated* as Mahler puts it in [106]) and have identically distributed parameters and furthermore that the number of clusters which are present are proportional to a Poisson distribution with a parameter equal to the total mass of the PHD estimate at the previous time-step, which we shall term N_k (i.e. $\hat{\alpha}_k(dx) = N_k \hat{s}_k(dx)$ for some probability density $\hat{s}_k$).

$$\hat{p}_k(X|Z_{1:k}) = \exp(-N_k) \prod_{x \in X} N_k \hat{s}_k(x)$$

Making this assumption, we obtain, as desired, the *pgfl* of a Poisson point process:

$$\begin{aligned}
\hat{G}_k(h) &= \sum_{i=0}^{\infty} \frac{1}{i!} \int h(x_1) \dots h(x_i) \hat{p}_k(\{x_1, \dots, x_i\}|Z_{1:k}) \lambda^i \left(\mathrm{d}(x_1, \dots, x_i)\right) \\
&= \exp(-N_k) \sum_{i=0}^{\infty} \frac{1}{i!} \int h(x_1) N_k \hat{s}_k(x_1) \dots h(x_i) N_k \hat{s}_k(x_i) \lambda^i \left(\mathrm{d}(x_1, \dots, x_i)\right) \\
&= \exp(-N_k) \sum_{i=0}^{\infty} \frac{1}{i!} \prod_{j=1}^{i} \int h(x_j) \hat{\alpha}_k(x_j) \lambda \left(\mathrm{d}x_j\right) \\
&= \exp\left(N_k \left(\int h(x) \hat{s}_k(dx) - 1\right)\right) \\
&= \exp\left(\hat{\alpha}_k(h) - N_k\right)
\end{aligned}$$

A Taylor Expansion Approach. Having expressed $\hat{G}_k(h) = \exp\left(\hat{\alpha}_k(h) - N_k\right)$ we are able to Taylor expand this expression to an order of our choosing. This amounts to assuming that the number of clusters present is at most equal to the order at which the expression is truncated (this can be seen by viewing the approximation as directly truncating the sum from which the exponential was obtained, rather than as a Taylor expansion of that exponential). In a sense, the approach of Mahler is similar to a first order truncation here, but his approach appears to maintain some meaningful value for the total mass of the PHD which a first order truncation does not.

We also have, following Mahler, that the joint *pgfl* of the predicted state and the observation set is given by $F_{k+1}(g, h) = \hat{G}_k(h\theta_g)$ with $\theta_g(x = (a, u)) = \exp(-a - af(g|u))$. And hence, that $\frac{\delta \theta_g}{\delta z}(a, u) = af(z|u)exp(-a + af(g|u)) = af(z|u)\theta_g(a, u)$. That the numerator and denominator of the updated PHD can be expressed in terms of functional derivatives of F evaluated at $(\mathbf{0}, \mathbf{1})$ follows from the usual considerations and has been shown in some detail by [107]. That is, the update equation may be written in the form:

$$\hat{\alpha}_{k+1}(x=(b,v)) = \left(\frac{\delta^m F_{k+1}}{\delta z_m \dots \delta z_1}(\mathbf{0},\mathbf{1})\right)^{-1} \frac{\delta^{m+1} F_{k+1}}{\delta(x=(b,v))\delta z_m \dots \delta z_1}$$

A First Order Expansion. The most naïve approach is to expand G to 1st order:

$$\hat{G}_k(h) = \exp(-N_k)(1+\hat{\alpha}_k(h)) + \mathcal{O}\left(\hat{\alpha}_k(h)^2\right)$$

This gives us:

$$F_{k+1}(g,h) = \exp(-N_k)(1+\hat{\alpha}_k(h\theta_g))$$

And hence that:

$$\begin{aligned}
\frac{\delta F_{k+1}}{\delta z_1}(g,h) &= \exp(-N_k)\hat{\alpha}_k(hf(z_1|u)\theta_g(a,u)a) \\
\frac{\delta^2 F_{k+1}}{\delta z_2 \delta z_1}(g,h) &= \exp(-N_k)\hat{\alpha}_k(hf(z_1|u)f(z_2|u)\theta_g(a,u)a^2) \\
&\vdots \\
\frac{\delta^m F_{k+1}}{\delta z_m \dots \delta z_1}(g,h) &= \exp(-N_k)\hat{\alpha}_k\left(ha^m\theta_g(a,u)\prod_{i=1}^m f(z_i|u)\right)
\end{aligned}$$

It is simple to see that:

$$\frac{\delta^{m+1} F_{k+1}}{\delta(x=(b,v))\delta z_m \dots \delta z_1}(g,h) = \exp(-N_k)b^m\theta_g(b,v)\prod_{i=1}^m f(z_i|v)\hat{\alpha}_k(b,v)$$

We then have:

$$\begin{aligned}
\hat{\alpha}_{k+1}(x=(b,v)) &= \left(\frac{\delta^m F_{k+1}}{\delta z_m \dots \delta z_1}(\mathbf{0},\mathbf{1})\right)^{-1} \frac{\delta^{m+1} F_{k+1}}{\delta(x=(b,v))\delta z_m \dots \delta z_1} \\
&= \frac{b^m e^{-b} f(z_m|v)\dots f(z_1|v)\hat{\alpha}_k(b,v)}{\hat{\alpha}_k(a^m e^{-a} f(z_m|u)\dots f(z_1|u))} \\
&= \frac{b^m e^{-b} f(z_m|v)\dots f(z_1|v)\hat{\alpha}_k(b,v)}{\hat{\alpha}_k(a^m e^{-a} f(z_m|u)\dots f(z_1|u))}
\end{aligned} \tag{3.23}$$

A Second Order Expansion. The next order of expansion yields another special case and gives some insight into the general result and the problems to which it leads:

$$\hat{G}_k(h) = \exp(-N_k)(1+\hat{\alpha}_k(h)+\frac{1}{2}\hat{\alpha}_k(h)^2) + \mathcal{O}\left(\hat{\alpha}_k(h)^3\right)$$

In this case we obtain:

$$
\begin{aligned}
F_{k+1}(g,h) &= \exp(-N_k)(1+\hat{\alpha}_k(h\theta_g)+\frac{1}{2}\hat{\alpha}_k(h\theta_g)^2)\\
\frac{\delta F_{k+1}}{\delta z_1}(g,h) &= \exp(-N_k)\frac{\delta}{\delta z_1}\left(1+\hat{\alpha}_k(h\theta_g)+\frac{1}{2}\hat{\alpha}_k(h\theta_g)^2\right)\\
&= \exp(-N_k)\left(1+\hat{\alpha}_k(h\theta_g)\right)\frac{\delta\hat{\alpha}_k(h\theta_g)}{\delta z_1}\\
&= \exp(-N_k)\left(1+\hat{\alpha}_k(h\theta_g)\right)\times\\
&\quad \frac{\delta}{\delta z_1}\left(\int\int \hat{\alpha}_k(a,u)e^{-a+af(g|u)}\lambda\left(\mathrm{d}a\right)\lambda\left(\mathrm{d}u\right)\right)\\
&= \exp(-N_k)\left(1+\hat{\alpha}_k(h\theta_g)\right)\times\\
&\quad \int\int \hat{\alpha}_k(a,u)(af(z_1|u)))e^{-a+af(g|u)}\lambda\left(\mathrm{d}a\right)\lambda\left(\mathrm{d}u\right)\\
&= \exp(-N_k)\left(1+\hat{\alpha}_k(h\theta_g)\right)\hat{\alpha}_k(a\theta_g f(z_1|u))
\end{aligned}
$$

The functional differentiation can be carried out iteratively, and becomes progressively more cumbersome, but there is a closed form expression for the m^{th} derivative of the generating functional.

Lemma 3.5.1. *In the case being considered here, with the pgfl obtained by truncating the exponential at second order, the following holds:*

$$
\begin{aligned}
\frac{\delta^m F_{k+1}}{\delta z_m\ldots\delta z_1}(g,1) = \exp(-N_k)\{&(1+\hat{\alpha}_k(\theta_g))\hat{\alpha}_k(a^m\theta_g f(z_1|u)\ldots f(z_m|u))+\\
&\sum_{i=1}^{\lfloor m/2\rfloor}\sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}}\hat{\alpha}_k\left(a^i\theta_g\prod_{j\in p}f(z_j|u)\right)s\left(a^{m-i}\theta_g\prod_{j\in q}f(z_j|u)\right)\Bigg\}
\end{aligned}
$$

where $\lfloor m/2\rfloor$ is the largest integer smaller than $m/2$ and $\mathfrak{P}_{(2)}^{m,i}$ is the set of all distinct partitions of the first m natural numbers into two partitions, one of size i and the other of size $m-i$. For our purposes, this set can be thought of as containing elements which are pairs of sets, one containing i natural numbers and the other containing $m-i$ natural numbers which cover $\{1,\ldots,m\}$.

Proof. The simplest approach is to use induction.

The first step is straightforward: we know that the result holds in the case $m=1$ as this corresponds to the result obtained on the previous page ($\lfloor 1/2\rfloor = 0$).

Now, if we assume that the result holds for m we need it to hold for $m+1$ as well. This is slightly more subtle than the previous step, but can be shown by direct differentiation:

$$
\begin{aligned}
& e_k^N \frac{\delta^{m+1} F_{k+1}}{\delta z_{m+1} \dots \delta z_1}(g,1) \\
& = \frac{\delta}{\delta z_{m+1} y} \{(1+\hat{\alpha}_k(\theta_g))\hat{\alpha}_k(a^m \theta_g f(z_1|u) \dots f(z_m|u)) + \\
& \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^i \theta_g \prod_{j\in p} f(z_j|u)\right) \hat{\alpha}_k\left(a^{m-i} \theta_g \prod_{j\in q} f(z_j|u)\right) \Bigg\} \\
& = (1+\hat{\alpha}_k(\theta_g))\hat{\alpha}_k(a^{m+1} \theta_g f(z_1|u) \dots f(z_{m+1}|u)) + \\
& \hat{\alpha}_k(\theta_g a f(z_{m+1}|u))\hat{\alpha}_k(a^m \theta_g f(z_1|u) \dots f(z_m|u)) + \\
& \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^i \theta_g \prod_{j\in p} f(z_j|u)\right) \hat{\alpha}_k\left(a^{m+1-i} \theta_g f(z_{m+1}|u) \prod_{j\in q} f(z_j|u)\right) + \\
& \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^{i+1} \theta_g f(z_{m+1}|u) \prod_{j\in p} f(z_j|u)\right) \hat{\alpha}_k\left(a^{m-i} \theta_g \prod_{j\in q} f(z_j|u)\right)
\end{aligned}
$$

What we require, in order for the recursion we seek to hold, is:

$$
\begin{aligned}
& (1+\hat{\alpha}_k(\theta_g))\hat{\alpha}_k(a^{m+1} \theta_g f(z_1|u) \dots f(z_{m+1}|u)) + \\
& \sum_{i=1}^{\lfloor (m+1)/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m+1,i}} \hat{\alpha}_k\left(a^i \theta_g \prod_{j\in p} f(z_j|u)\right) \hat{\alpha}_k\left(a^{m+1-i} \theta_g \prod_{j\in q} f(z_j|u)\right) \\
& = (1+\hat{\alpha}_k(\theta_g))\hat{\alpha}_k(a^{m+1} \theta_g f(z_1|u) \dots f(z_{m+1}|u)) + \\
& \hat{\alpha}_k(\theta_g a f(z_{m+1}|u))\hat{\alpha}_k(a^m \theta_g f(z_1|u) \dots f(z_m|u)) + \\
& \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^i \theta_g \prod_{j\in p} f(z_j|u)\right) \hat{\alpha}_k\left(a^{m+1-i} \theta_g f(z_{m+1}|u) \prod_{j\in q} f(z_j|u)\right) + \\
& \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^{i+1} \theta_g f(z_{m+1}|u) \prod_{j\in p} f(z_j|u)\right) \hat{\alpha}_k\left(a^{m-i} \theta_g \prod_{j\in q} f(z_j|u)\right)
\end{aligned}
$$

The first line on either side of these equations is clearly identical and can be removed, leaving:

$$
\begin{aligned}
& \sum_{i=1}^{\lfloor (m+1)/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m+1,i}} \hat{\alpha}_k\left(a^i \theta_g \prod_{j\in p} f(z_j|u)\right) \hat{\alpha}_k\left(a^{m+1-i} \theta_g \prod_{j\in q} f(z_j|u)\right) \\
& = \hat{\alpha}_k(\theta_g a f(z_{m+1}|u))\hat{\alpha}_k(a^m \theta_g f(z_1|u) \dots f(z_m|u)) + \\
& \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^i \theta_g \prod_{j\in p} f(z_j|u)\right) \hat{\alpha}_k\left(a^{m+1-i} \theta_g f(z_{m+1}|u) \prod_{j\in q} f(z_j|u)\right) + \\
& \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^{i+1} \theta_g f(z_{m+1}|u) \prod_{j\in p} f(z_j|u)\right) \hat{\alpha}_k\left(a^{m-i} \theta_g \prod_{j\in q} f(z_j|u)\right)
\end{aligned}
$$

Whilst this looks somewhat difficult to prove, it amounts to a simple truism: any partition of $\{1,\ldots,m+1\}$ into two non-empty subsets must be one of the following:

– $\{\{m+1\},\{1,\ldots,m\}\}$ which corresponds to the first term in the expression
– Some partition of $\{1,\ldots,m+1\}$ with $m+1$ added to the first subset which corresponds to the third term in the expression
– Some partition of $\{1,\ldots,m+1\}$ with $m+1$ added to the second subset which corresponds to the second term in the expression

Considering the two expressions in this way, we see immediately that the two sides match and the lemma holds by induction. □

The previous lemma also gives us the result we need to obtain the other functional derivative of the generating functional we require.

$$
\begin{aligned}
& e_k^N \frac{\delta^{m+1} F_{k+1}}{\delta(x=(b,v))\delta z_m \ldots \delta z_1}(g,h) \\
&= \frac{\delta}{\delta x=(b,v)} \{(1+\hat{\alpha}_k(\theta_g))\hat{\alpha}_k(a^m\theta_g f(z_1|u)\ldots f(z_m|u))+ \\
&\sum_{i=1}^{\lfloor m/2\rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^i\theta_g\prod_{j\in p} f(z_j|u)\right)\hat{\alpha}_k\left(a^{m-i}\theta_g\prod_{j\in q} f(z_j|u)\right)\Bigg\} \\
&= \hat{\alpha}_k(b,v)\theta_g(b,v)\hat{\alpha}_k\left(a^m\theta_g h f(z_1|u)\ldots f(z_m u)\right)+ \\
&(1+\hat{\alpha}_k(\theta_g h))(b^m\theta_g(b,v)f(z_1|v)\ldots f(z_m|v))\hat{\alpha}_k(b,v)+ \\
&\sum_{i=1}^{\lfloor m/2\rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} b^i\theta_g(b,v)\prod_{j\in p} f(z_j|v)\hat{\alpha}_k\left(a^{m-i}\theta_g h\prod_{j\in q} f(z_j|u)\right)\hat{\alpha}_k(b,v)+ \\
&\sum_{i=1}^{\lfloor m/2\rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^i\theta_g(a,u)\prod_{j\in p} f(z_j|v)\right) b^{m-i}\theta_g(b,v)\prod_{j\in q} f(z_j|v)\hat{\alpha}_k(b,v)
\end{aligned}
$$

Then evaluating these two expressions at $g=\mathbf{0}, h=\mathbf{1}$ we obtain:

$$\begin{aligned}
\hat{\alpha}_{k+1}((b,v)) =& \frac{\mathcal{N}}{\mathcal{D}}\hat{\alpha}_k(b,v) \\
\mathcal{N} =& e^{-b}\hat{\alpha}_k(a^m e^{-a} f(z_1|u)\dots f(z_m|u)) \\
&+ (1+\hat{\alpha}_k(e^{-a}))b^m e^{-b} f(z_1|v)\dots f(z_m|v) \\
&+ \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} b^i e^{-b} \prod_{j\in p} f(z_j|v)\hat{\alpha}_k\left(a^{m-i}e^{-a}\prod_{j\in q} f(z_j|u)\right) \\
&+ \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^i e^{-a}\prod_{j\in p} f(z_j|v)\right) b^{m-i}e^{-b}\prod_{j\in q} f(z_j|v) \\
\mathcal{D} =& (1+\hat{\alpha}_k(e^{-a}))\hat{\alpha}_k(a^m e^{-a} f(z_1|u)\dots f(z_m|u)) + \\
&+ \sum_{i=1}^{\lfloor m/2 \rfloor} \sum_{(p,q)\in\mathfrak{P}_{(2)}^{m,i}} \hat{\alpha}_k\left(a^i e^{-a}\prod_{j\in p} f(z_j|u)\right)\hat{\alpha}_k\left(a^{m-i}e^{-a}\prod_{j\in q} f(z_j|u)\right).
\end{aligned}$$

A General n^{th} Order Expansion. The obvious next step, having obtained results for these special cases is to attempt to find a general result for arbitrary orders of expansion. In order to make use of the relationship between one order of expansion and the next, it is useful to make the following definitions:

$$\begin{aligned}
\tilde{F}_i(g,h) &= \frac{\hat{\alpha}_k(\theta_g h)^i}{i!} \\
F_i(g,h) &= \exp(-N_k)\sum_{j=1}^{i}\tilde{F}_i
\end{aligned}$$

There are, actually, three regimes for each $\tilde{F}_i$:

– $m < n$: fewer observations than clusters

– $m = n$: equal numbers of observations and clusters

– $m > n$: more observations than clusters.

Whilst the last of these is obviously the most important from an applications point of view, we need to calculate iterated derivatives of F which respect to the observation set and hence the first two of these regimes must be considered for the purposes of obtaining a derivation, even if they are not of interest in most applications.

Lemma 3.5.2. *For $m < n$:*

$$\frac{\delta^m \tilde{F}_n}{\delta z_m \dots \delta z_1} = \sum_{i=1}^{m} \tilde{F}_{n-i} \sum_{(p_1,\dots,p_i)\in\mathbf{P}_{<i>}^{(m)}} \prod_{j=1}^{i} A_{p_j}$$

where some additional notation has been defined for convenience.

$\mathbf{P}_{<i>}^{(m)}$ *is the set of all non-trivial (in the sense that all of the sets comprising the partition containing at least one element) partitions of the first m natural numbers into i partitions.*

$A_X = \hat{\alpha}_k \left(a^{|X|} \prod_{i \in X} f(u|z_i)\theta_g h \right)$ *for some set of natural numbers,* X*.*

This notation can disguise the complexity of the expressions being manipulated, but rather shortens the expressions which we need to manipulate.

Proof. Again, induction is the simplest method to obtain this result.

First consider a few simple cases:

$$\begin{aligned}\frac{\delta \tilde{F}_n}{\delta z_1} &= \frac{\delta}{\delta z_1} \frac{\hat{\alpha}_k (\theta_g h)^n}{n!} \\ &= \frac{n \hat{\alpha}_k (\theta_g h)^{n-1}}{n!} \frac{\delta}{\delta z_1} \hat{\alpha}_k (\theta_g h) \\ &= \tilde{F}_{n-1} A_{\{1\}}\end{aligned}$$

thus we have that the result holds for $m = 1$. To see how these entities behave, it is useful to consider a few more special cases before progressing to the general result.

$$\begin{aligned}\frac{\delta^2 \tilde{F}_n}{\delta z_2 \delta z_1} &= \frac{\delta}{\delta z_2} \left(\frac{\delta \tilde{F}_n}{\delta z_1} \right) \\ &= \frac{\delta}{\delta z_2} \left(\tilde{F}_{n-1} A_{\{1\}} \right) \\ &= \tilde{F}_{n-2} A_{\{1\}} A_{\{2\}} + \tilde{F}_{n-1} A_{\{1,2\}}\end{aligned}$$

$$\begin{aligned}\frac{\delta^3 \tilde{F}_n}{\delta z_3 \delta z_2 \delta z_1} &= \frac{\delta}{\delta z_3} \left(\frac{\delta^2 \tilde{F}_n}{\delta z_2 \delta z_1} \right) \\ &= \frac{\delta}{\delta z_3} \left(\tilde{F}_{n-2} A_{\{1\}} A_{\{2\}} + \tilde{F}_{n-1} A_{\{1,2\}} \right) \\ &= \tilde{F}_{n-3} A_{\{1\}} S_{\{2\}} A_{\{3\}} + \\ &\quad \tilde{F}_{n-2} \left(A_{\{1\}} S_{\{2,3\}} + A_{\{1,3\}} S_{\{2\}} + A_{\{1,2\}} A_{\{3\}} \right) + \tilde{F}_{n-1} S_{\{1,2,3\}}\end{aligned}$$

Hence the result holds in the $m = 2, m = 3$ cases as well. All that remains is to show that it holds for general $m < n$. As we have three particular cases at the start of the sequence of interest, all that we actually need to prove is that the result holds at $m + 1$ if it holds at m.

Under the induction assumption:

$$
\begin{aligned}
\frac{\delta^{m+1}\tilde{F}_n}{\delta z_{m+1}\dots\delta z_1} &= \frac{\delta}{\delta z_{m+1}}\left(\frac{\delta^m \tilde{F}_n}{\delta z_m\dots\delta z_1}\right) \\
&= \frac{\delta}{\delta z_{m+1}}\left(\sum_{i=1}^{m}\tilde{F}_{n-i}\sum_{(p_1,\dots,p_i)\in\mathbf{P}^{(m)}_{<i>}}\prod_{j=1}^{i}A_{p_j}\right) \\
&= \sum_{i=1}^{m}\left(\tilde{F}_{n-i-1}A_{\{m+1\}}\sum_{(p_1,\dots,p_i)\in\mathbf{P}^{(m)}_{<i>}}\prod_{j=1}^{i}A_{p_j}+\right. \\
&\qquad \left.\tilde{F}_{n-i}\sum_{(p_1,\dots,p_i)\in\mathbf{P}^{(m)}_{<i>}}\sum_{k=1}^{i}S_{p_k\bigcup\{m+1\}}\prod_{j\neq k}A_{p_j}\right) \\
&= \sum_{i=2}^{m}\left(\tilde{F}_{n-i}A_{\{m+1\}}\sum_{(p_1,\dots,p_{i-1})\in\mathbf{P}^{(m)}_{<i-1>}}\prod_{j=1}^{i-1}A_{p_j}+\right. \\
&\qquad \left.\tilde{F}_{n-i}\sum_{(p_1,\dots,p_i)\in\mathbf{P}^{(m)}_{<i>}}\sum_{k=1}^{i}S_{p_k\bigcup\{m+1\}}\prod_{j\neq k}A_{p_j}\right)+ \\
&\quad \tilde{F}_{n-(m+1)}A_{\{m+1\}}\sum_{(p_1,\dots,p_m)\in\mathbf{P}^{(m)}_{<m>}}\prod_{j=1}^{m}A_{p_j}+ \\
&\quad \tilde{F}_{n-1}\sum_{p_1\in\mathbf{P}^{(m)}_{<1>}}A_{p_1\bigcup\{m+1\}} \\
&= \sum_{i=2}^{m}\left(\tilde{F}_{n-i}\sum_{(p_1,\dots,p_i)\in\mathbf{P}^{(m+1)}_{<i>}}\prod_{j=1}^{i}A_{p_j}\right)+ \\
&\quad \tilde{F}_{n-(m+1)}\prod_{j=1}^{m+1}A_{\{j\}}+\tilde{F}_{n-1}A_{\{1,\dots,m+1\}}
\end{aligned}
$$

Looking at the final line of this expression, which follows by noting that there is precisely one non-trivial partition of $\{1,\dots,m\}$ into m sets and also into a single set, as well as by using the same logic as in the $m=2$ case to simplify the expressions within the summation, we can compare the elements it contains with those which we are expecting.

A little consideration shows that the first term, the last term and all the intermediate terms match those which this lemma asserts should be the case and the result thus holds by an induction argument. □

The same result clearly also holds in the $n=m$ case, with the additional point that $\tilde{F}_{n-n}=1$. Hence, we obtain, for this particular case:

$$\frac{\delta^n \tilde{F}_n}{\delta z_n \dots \delta z_1} = \sum_{i=1}^{n} \frac{\hat{\alpha}_k(\theta_g h)^{n-i}}{(n-i)!} \sum_{(p_1,\dots,p_i)\in \mathbf{P}^{(n)}_{<i>}} \prod_{j=1}^{i} A_{p_j}$$

Lemma 3.5.3. *For $m \geq n$:*

$$e\frac{\delta^m \tilde{F}_n}{\delta z_m \dots \delta z_1} = \sum_{i=1}^{n} \frac{\hat{\alpha}_k(\theta_g h)^{n-i}}{(n-i)!} \sum_{(p_1,\dots,p_i)\in \mathbf{P}^{(m)}_{<i>}} \prod_{j=1}^{i} A_{p_j} \tag{3.24}$$

Proof. We have the result we seek for the $n = m$ case by the argument above.

In order to prove that it holds in generality, an induction argument requires only that we prove that, if the result holds for m it also holds for $m + 1$.

Under the induction assumption:

$$e\frac{\delta^{m+1} \tilde{F}_n}{\delta z_{m+1} \dots \delta z_1} = \sum_{i=1}^{n-1} \left(\frac{\hat{\alpha}_k(\theta_g h)^{n-i-1}}{(n-i-1)!} \sum_{(p_1,\dots,p_i)\in \mathbf{P}^{(m)}_{<i>}} \prod_{j=1}^{i} A_{p_j} \right) +$$
$$\sum_{i=1}^{n} \frac{\hat{\alpha}_k(\theta_g h)^{n-i}}{(n-i)!} \sum_{(p_1,\dots,p_i) in \mathbf{P}^{(m)}_{<i>}} \sum_{k=1}^{i} A_{p_k \bigcup \{m+1\}} \prod_{j\neq k} A_{p_j}$$

A little manipulation of this result suffices to prove the lemma. □

In order to obtain a general result we also require an expression for the first derivative with respect to δx of this expression. Obtaining the update equation is simply a matter of summing the derivatives of the $\tilde{F}_n$ expressions to obtain those of the F_n that we are really interested in, evaluating them at the appropriate values of g and h and then taking their ratios.

In actual fact the additional derivative can be obtained from the result of this lemma directly:

$$\begin{aligned}
&e\frac{\delta^{m+1} \tilde{F}_n}{\delta(b,v)\delta z_m \dots \delta z_1} \\
&= \sum_{i=1}^{n} \left(\frac{\delta}{\delta(b,v)} \left(\frac{\hat{\alpha}_k(\theta_g h)^{n-i}}{(n-i)!} \right) \sum_{(p_1,\dots,p_i)\in \mathbf{P}^{(m)}_{<i>}} \prod_{j=1}^{i} A_{p_j} \right. \\
&\quad \left. + \frac{\hat{\alpha}_k(\theta_g h)^{n-i}}{(n-i)!} \sum_{(p_1,\dots,p_i)\in \mathbf{P}^{(m)}_{<i>}} \frac{\delta}{\delta(b,v)} \left(\prod_{j=1}^{i} A_{p_j} \right) \right) \\
&= \sum_{i=1}^{n-1} \frac{\hat{\alpha}_k(\theta_g h)^{n-i-1}}{(n-i-1)!} \theta_g(b,v)\hat{\alpha}_k(b,v) \sum_{(p_1,\dots,p_i)\in \mathbf{P}^{(m)}_{<i>}} \prod_{j=1}^{i} A_{p_j} + \\
&\sum_{i=1}^{n} \frac{\hat{\alpha}_k(\theta_g h)^{n-i}}{(n-i)!} \sum_{(p_1,\dots,p_i)\in \mathbf{P}^{(m)}_{<i>}} \sum_{k=1}^{i} \hat{\alpha}_k(b,v)\theta_g(b,v) b^{|p_k|} \prod_{l\in p_k} f(z_l|v) \prod_{j\neq k}^{i} A_{p_j}
\end{aligned}$$

A little rearrangement yields:

$$= \left(\sum_{i=1}^{n-1} \frac{\hat{\alpha}_k (\theta_g h)^{n-i-1}}{(n-i-1)!} \sum_{p \in \mathbf{P}^{(m)}_{<i>}} \left(\prod_{j=1}^{i} A_{p_j} + \frac{\hat{\alpha}_k(\theta_g h)}{(n-i)} \sum_{k=1}^{i} b^{|p_k|} \prod_{l \in p_k} f(z_l|v) \prod_{j \neq k} A_{p_j} \right) \right.$$

$$\left. + \sum_{p \in \mathbf{P}^{(m)}_{<n>}} \sum_{k=1}^{n} b^{|p_k|} \prod_{l \in p_k} f(z_l|v) \prod_{j \neq k} A_{p_j} \right) \theta_g(b,v) \hat{\alpha}_k(b,v) \tag{3.25}$$

$$= \left(\sum_{i=1}^{n-1} \frac{\hat{\alpha}_k (\theta_g h)^{n-i-1}}{(n-i-1)!} \sum_{p \in \mathbf{P}^{(m)}_{<i>}} \left(\sum_{k=1}^{i} \prod_{j \neq k} A_{p_j} \left(A_{p_k} + \frac{\hat{\alpha}_k(\theta_g h)}{(n-i)} b^{|p_k|} \prod_{l \in p_k} f(z_l|v) \right) \right) \right.$$

$$\left. + \sum_{p \in \mathbf{P}^{(m)}_{<n>}} \sum_{k=1}^{n} b^{|p_k|} \prod_{l \in p_k} f(z_l|v) \prod_{j \neq k} A_{p_j} \right) \theta_g(b,v) \hat{\alpha}_k(b,v) \tag{3.26}$$

Theorem 3.5.1. *General Form of the PHD Update Equation If we truncate the exponential form of the generating function at an order N, we obtain as the PHD update equation:*

$$\alpha_{k+1}(b,v) = \left(\sum_{n=1}^{N} \sum_{i=1}^{n} \frac{\hat{\alpha}_k (e^{-a})^{n-i}}{(n-i)!} \sum_{(p_1,\ldots,p_i) \in \mathbf{P}^{(m)}_{<i>}} \prod_{j=1}^{i} \bar{A}_{p_j} \right)^{-1} \times$$

$$\left(\sum_{n=1}^{N} \sum_{i=1}^{n-1} \frac{\hat{\alpha}_k (e^{-a})^{n-i-1}}{(n-i-1)!} \sum_{p \in \mathbf{P}^{(m)}_{<i>}} \left(\sum_{k=1}^{i} \prod_{j \neq k} \bar{A}_{p_j} \left(\bar{A}_{p_k} + \frac{\hat{\alpha}_k(e^{-a})}{(n-i)} b^{|p_k|} \prod_{l \in p_k} f(z_l|v) \right) \right) \right.$$

$$\left. + \sum_{p \in \mathbf{P}^{(m)}_{<n>}} \sum_{k=1}^{n} b^{|p_k|} \prod_{l \in p_k} f(z_l|v) \prod_{j \neq k} \bar{A}_{p_j} \right) e^{-b} \hat{\alpha}_k(b,v)$$

where $m = |Z_{k+1}|$ *and*

$$\bar{A}_Z = \hat{\alpha}_k \left(e^{-a} \prod_{z \in Z} (a f(z|u)) \right)$$

Proof. We know that:

$$\alpha_{k+1}(b,v) = \left(\frac{\delta^m F_N(\mathbf{0},\mathbf{1})}{\delta z_m \ldots \delta z_1} \right)^{-1} \frac{\delta^{m+1} F_N(\mathbf{0},\mathbf{1})}{\delta(b,v) \delta z_m \ldots \delta z_1}$$

$$= \left(\sum_{n=1}^{N} \frac{\delta^m \tilde{F}_n(\mathbf{0},\mathbf{1})}{\delta z_m \ldots \delta z_1} \right)^{-1} \sum_{n=1}^{N} \frac{\delta^{m+1} \tilde{F}_n(\mathbf{0},\mathbf{1})}{\delta(b,v) \delta z_m \ldots \delta z_1}$$

By direct substitution into equation 3.24 we have:

$$\frac{\delta^m \tilde{F}_n}{\delta z_m \ldots \delta z_1}(\mathbf{0},\mathbf{1}) = \sum_{i=1}^{n} \frac{\hat{\alpha}_k (e^{-a})^{n-i}}{(n-i)!} \sum_{(p_1,\ldots,p_i) \in \mathbf{P}^{(m)}_{<i>}} \prod_{j=1}^{i} \hat{\alpha}_k \left(e^{-a} \prod_{k in p_j} (a f(z_k|u)) \right) \tag{3.27}$$

and by equation 3.26 we have as an expression for e times the numerator:

$$\left(\sum_{i=1}^{n-1}\frac{\hat{\alpha}_k(e^{-a})^{n-i-1}}{(n-i-1)!}\sum_{p\in\mathbf{P}^{(m)}_{<i>}}\left(\sum_{k=1}^{i}\prod_{j\neq k}\bar{A}_{p_j}\left(\bar{A}_{p_k}+\frac{\hat{\alpha}_k(e^{-a})}{(n-i)}b^{|p_k|}\prod_{l\in p_k}f(z_l|v)\right)\right)\right.$$
$$\left.+\sum_{p\in\mathbf{P}^{(m)}_{<n>}}\sum_{k=1}^{n}b^{|p_k|}\prod_{l\in p_k}f(z_l|v)\prod_{j\neq k}\bar{A}_{p_j}\right)e^{-b}\hat{\alpha}_k(b,v)$$

combining these, we obtain:

$$\alpha_{k+1}(b,v)=\left(\sum_{n=1}^{N}\sum_{i=1}^{n}\frac{\hat{\alpha}_k(e^{-a})^{n-i}}{(n-i)!}\sum_{(p_1,\ldots,p_i)\in\mathbf{P}^{(m)}_{<i>}}\prod_{j=1}^{i}\bar{A}_{p_j}\right)^{-1}\times$$
$$\left(\sum_{n=1}^{N}\sum_{i=1}^{n-1}\frac{\hat{\alpha}_k(e^{-a})^{n-i-1}}{(n-i-1)!}\sum_{p\in\mathbf{P}^{(m)}_{<i>}}\left(\sum_{k=1}^{i}\prod_{j\neq k}\bar{A}_{p_j}\left(\bar{A}_{p_k}+\frac{\hat{\alpha}_k(e^{-a})}{(n-i)}b^{|p_k|}\prod_{l\in p_k}f(z_l|v)\right)\right)\right.$$
$$\left.+\sum_{p\in\mathbf{P}^{(m)}_{<n>}}\sum_{k=1}^{n}b^{|p_k|}\prod_{l\in p_k}f(z_l|v)\prod_{j\neq k}\bar{A}_{p_j}\right)e^{-b}\hat{\alpha}_k(b,v)$$

□

3.5.4 Approaches to Computation

It is immediately apparent that the expressions obtained in the previous section do not produce reasonable computable estimators, and it is not obvious how to approximate them in a computationally tractable manner.

There are a number of possibilities for using a PHD-filter type approach to perform dynamic clustering / cluster tracking.

1. Cluster first and then track: rather than attempting to perform both steps simultaneously, perform a clustering at each time step to obtain an estimate of the cluster parameters and then, using the cluster parameters as a set of meta-observations, the PHD filter could be used to track these parameters.
2. Track first and then cluster. Similar in spirit to the first approach, this method would involve tracking the observation set using the PHD filter and then attempting to estimate the cluster parameters from the PHD.
3. Modify the approach to track each cluster separately in some sense. One way of achieving this is discussed in more depth below.

The first two of these approaches seem unlikely to be particularly interesting – they each involve the computation of a batch clustering at each time step and whilst an estimate can be obtained from the clustering at the previous time step, this seems likely to be cumbersome as well as being limited by the performance of both the PHD and a separate clustering algorithm.

The third approach seems more likely to be interesting, although it poses quite a number of theoretical difficulties.

The major problem with the approach proposed in [106] is that it makes the assumption that there are some unknown number of clusters which all have identical parameters in order to derive the update equation.

The update described in equation 3.23 is somewhat limited in that it amounts to explicitly making the assumption that there exists a single cluster. The obvious approach to obtaining a tractable clustering PHD which permits clusters which have disjoint supports[2] is to limit the set of observations considered when calculating the update for each point in the PHD to those within the *main part* of the region of support of the cluster with the parameters described by that point. Although this seems rather heuristic, it may be possible to obtain some reasonable properties in this way.

The first step of this approach is to define a region in the observation space which will include the majority of points which are associated with a given cluster type. This is related to the concept of *gating* which is widely used in the multiple-target tracking community (see, for example, [5]). The most intuitive approach to doing this would appear to be to select the region of minimal volume which encloses some fraction (say 99 %) of the likelihood associated with a cluster with the parameters described by a cluster with parameters x. This amounts to making use of the region enclosed by one of the *minimal volume multivariate quantiles* of [51].

$$Q(x;t) = \arg\inf \left\{ \lambda(A) : \int_A g(z|x)\lambda dz > t \right\} \tag{3.28}$$

Where $t \in (0,1)$ is some threshold and $g(z|x)$ is the likelihood function. Using an approximation of the type described here, the update equation becomes something of the form:

$$\alpha_{k+1}^{<t>}(b,v) = \frac{b^{|Z_{k+1} \bigcap Q(v;t)|} e^{-b} \prod_{z \in Z_{k+1}} \left(f(z|v)\mathbb{I}_{Q(v;t)}(z) + (1 - \mathbb{I}_{Q(x;t)}(z)) \right) \hat{\alpha}_k(b,v)}{\int a^{|Z_{k+1} \bigcap Q(u;t)|} e^{-a} \prod_{z \in Z_{k+1}} \left(f(z|u)\mathbb{I}_{Q(u;t)}(z) + (1 - \mathbb{I}_{Q(u;t)}(z)) \right) \hat{\alpha}_k(da,du)} \tag{3.29}$$

Having obtained an expression of this sort for the update equation, a straightforward modification of the particle approach to obtaining an approximation of the standard PHD filter described in [147] provides an obvious way to implement the algorithm. Indeed, preliminary experiments suggest that such an approach is viable and perform as one would expect. However, the algorithm seems not to be

[2] As well, of course, as those with overlapping supports.

sufficiently interesting that more rigorous experimentation would be justified at this stage.

Obvious Deficiencies of the Proposed Approximation Scheme.

It is a heuristic with little theoretical justification. This is largely true, and although there are a number of factors which motivate an approximation of this sort, it does not seem likely to be widely useful. The most obvious concern is that it is likely to be extremely difficult to obtain convergence results for the approximation.

Some motivation comes from the fact that in the limit of extremely well separated clusters one might wish to run separate tracking algorithms for each cluster, and other than the fact that a simple particle approximation to this scheme allows for resampling to redistribute particles between separated clusters, that is exactly what this approach amounts to doing. Another benefit is that it prevents the problem obtained with Mahler's methodology in which points tend to be dragged towards the centre of distinct clusters in the update stage, and has a "natural" method for considering only those points which are reasonably associated with a given cluster when calculating the intensity for particular cluster parameters.

Interpretation of the total mass $\hat{\alpha}_{k+1}(\mathbf{1})$ is unclear. Considering the form of the expression it is obvious that, if there is a single "compact" cluster then the total mass of the filter must be exactly one, as the update equation gives an expression which clearly corresponds to a probability density. Indeed, if there are N well-spaced, "compact" clusters then the total mass must be exactly N as those parts of D which are not within the support of a given cluster do not contribute to the denominator of the updated D and hence each cluster amounts to a probability density.

As clusters become "close together" in the sense that the $Q(x;t)$ regions associated with points within each cluster begin to overlap there is a tendency to reduce this value somewhat as the denominator for x in the support of each cluster is increased due to intensity within the other cluster. This is, approximately at least, exactly the behaviour which is desired.

Problems with interpretation and behaviour for poorly separated clusters. This could potentially be a problem, but it is a pathology of the approach suggested in [106] and it is always difficult to deal with poorly separated clusters.

Computationally Intensive for an Approximation. If there are M observations and N particles are used to obtain an approximation to the update equation then the worst case performance appears to be $\mathcal{O}\left((MN)^2\right)$ which, whilst not ideal, is tractable for problems of a reasonable scale on modern hardware. Cleverer approximation schemes may reduce the complexity somewhat.

3.5.5 Further Issues

Other concerns which could be raised about existing work on the PHD filter, which it would be nice to address include the following:

1. There is the assumption that all observations are generated by a cluster element; this precludes the possibility of clutter. This is unfortunate as the 1st moment interpretation would make including two different types of cluster, one for clutter and one for true clusters of interest rather cumbersome and difficult to interpret, particularly as the real distribution would ideally have precisely one clutter class.
2. Set integrals correspond to integrals with respect to unnormalized Poisson processes. To allow the densities to be specified in a cleaner fashion it would be nicer to use a normalised form of this expression.
3. It would be nice to relax the assumption that the distribution of the number of points attributable to each cluster at each time step should follow a Poisson distribution.

3.6 Summary

We have presented almost sure, and $\mathbb{L}_p$ convergence results for the SMC implementation of the PHD filter under very reasonable conditions. A central limit theorem has also been shown to hold, and an asymptotic variance expression obtained. These are interesting theoretical results, and should reassure practitioners that the method is sound.

An attempt was made at generalising the methodology to the group tracking and dynamic clustering problem as previous attempts at doing so have suffered from serious deficiencies. No practical algorithm was obtained, and this provides another example of an attempt at generalising the PHD filter to systems other than that for which it was first derived meeting with little success. It seems likely that the PHD approach will meet with limited success outside the domain in which it was first proposed.

4. Marginal Parameter Estimation via SMC

"On the other hand, it is impossible for a cube to be written as a sum of two cubes or a fourth power to be written as a sum of two fourth powers or, in general, for any number which is a power greater than the second to be written as a sum of two like powers. I have a truly marvelous demonstration of this proposition which this margin is too narrow to contain."

– Pierre de Fermat

A short version of the work presented in this chapter was published as [87], and a longer version is in preparation.

4.1 Introduction

Maximum Likelihood (ML) parameter estimation is a well established technique which is widely employed for obtaining point estimates of the parameters of probabilistic models; within a Bayesian framework Maximum a Posteriori (MAP) estimation fulfills a very similar role. Although there is a marked preference for the use of posterior means within much of the Bayesian literature, in situations in which the parameters are not identifiable this does not always make sense as the posterior means of all exchangeable parameters is, of course, the same [16]. Although it is often possible to adopt approaches such as re-ordering of the parameters and imposing identifiability constraints to allow the use of such approach, there do exist situations in which MAP estimation may be preferred. Furthermore, within a non-Bayesian context, the MAP estimator has an interpretation as a penalised maximum likelihood estimator, which can be applied to systems within unbounded likelihood functions [74].

4.1.1 Problem Formulation

The situation in which we are interested is that in which one has some likelihood function $p(y, z|\theta)$ in which $(y, z) \in \mathcal{Y} \times \mathcal{Z}$ for observed data y and latent variables (often called "hidden data"), z. This joint likelihood is known, but as z is not observed, the *marginal likelihood*,

$$p(y|\theta) = \int p(y, z|\theta) dz, \tag{4.1}$$

is the quantity of interest. As this integral is generally not tractable, it is not straightforward to maximise it with respect to the parameters to obtain the

(marginal) maximum likelihood (ML) estimator:

$$\hat{\theta}^{ML} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(y|\theta). \tag{4.2}$$

4.1.2 Notation

It is convenient at this stage to remind the reader of certain aspects of the notation used below. Given a real quantity, $x \in \mathbb{R}$, we define the floor, ceiling and remainder of x as:

$$\begin{aligned} \lfloor x \rfloor &\triangleq \sup\{y \in \mathbb{Z} \ : y \leq x\}, \\ \lceil x \rceil &\triangleq \inf\{y \in \mathbb{Z} \ : y \geq x\}, \\ \text{and } x^\sharp &\triangleq x - \lfloor x \rfloor. \end{aligned}$$

The following notations are used to describe various probability distributions: $\mathcal{B}er\,(p)$ describes the Bernoulli distribution with success probability p, $\mathcal{D}i\,(\alpha)$ the Dirichlet distribution with parameter vector α, $\mathcal{N}\left(\mu, \sigma^2\right)$ describes a normal of mean μ and variance σ^2, $\mathcal{G}a\,(\alpha, \beta)$ a gamma distribution of shape α and rate β, $\mathcal{IG}\,(\alpha, \beta)$ the inverse gamma distribution associated with $\mathcal{G}a\,(\alpha, \beta)$, $\mathcal{L}ogistic\,(\mu, s)$ the logistic distribution with location μ and scale s and $\mathcal{KS}$ refers to the Kolmogorov-Smirnov distribution.

With some abuse of notation, we assume throughout this chapter that all measures admit a density with respect to some dominating measure, λ and use the same notation to refer to that density and the measure itself, i.e. $\pi(dx) = \pi(x)\lambda(dx)$ with the understanding that $\pi(x) = \frac{\mathrm{d}\pi}{\mathrm{d}\lambda}(x)$.

4.1.3 Previous Approaches

When the marginal likelihood $p(y|\theta)$ can be evaluated, the classical approach to problems of this sort is the Expectation Maximisation (EM) algorithm [42], which is a numerically well-behaved algorithm. For complex models, $p(y|\theta)$ cannot be computed analytically and typically the expectation step of the EM algorithm cannot be performed in closed-form either. In such scenarios, Monte Carlo variants of EM have been proposed – including Stochastic EM (SEM), Monte Carlo EM (MCEM) and stochastic approximation Stochastic Approximation EM (SAEM). See [130] for a comparative summary of these approaches. Note that all of these approaches are susceptible to trapping in local modes.

An alternative approach related to Simulated Annealing (SA) is to build a sequence of distributions which concentrates itself on the set of maxima of the likelihood. Let $p(\theta)$ be an instrumental prior distribution whose support includes the ML estimate then the distributions,

$$p_\gamma^{ML}(\theta|y) \propto p\,(\theta)\, p(y|\theta)^\gamma, \tag{4.3}$$

concentrate themselves on the set of ML estimates as $\gamma \to \infty$. Indeed asymptotically the contribution from this instrumental prior vanishes as shown in section 4.2.2. The term $p(\theta)$ is only present to ensure that the distributions $\left\{p_\gamma^{ML}(\theta|y)\right\}$ are integrable – it may be omitted in those instances in which this is already the case. To sample from these distributions, one would like to use MCMC methods. Unfortunately, this is impossible whenever $p(y|\theta)$ is not known pointwise up to a normalizing constant.

To circumvent this problem, it has been proposed in [49] (in a MAP, rather than ML, setting) to build a sequence of artificial distributions known up to a normalizing constant, which admit as a marginal distribution the target distribution $p_\gamma^{ML}(\theta|y)$ for an integer power γ greater than one. A similar scheme was subsequently proposed by [81, 55] in the ML setting. This is achieved by simulating a number of replicates of the missing data where one defines,

$$p_\gamma(\theta, z_{1:\gamma}|y) \propto p(\theta) \prod_{i=1}^{\gamma} p(Y = y, Z = z_i|\theta), \tag{4.4}$$

with $z_{i:j} = (z_i, ..., z_j)$. Indeed it is easy to check that:

$$\int p_\gamma(\theta, z_{1:\gamma}|y) dz_{1:\gamma} = p_\gamma^{ML}(\theta|y).$$

The approach of [49], termed State Augmentation for Marginal Estimation (SAME), is to construct an inhomogeneous Markov chain which produces samples from a sequence of such distributions for increasing values of γ. Just as in SA, this concentrates the mass on the set of global maxima of $p(y|\theta)$ as γ becomes large. Another approach proposed by [81] is to construct a homogeneous Markov chain whose invariant distribution corresponds to such a distribution for a predetermined value of γ. It can be theoretically established that these methods converge asymptotically towards the set of estimates of interest if γ grows slowly enough to ∞. However in practice, these approaches suffer from two major weaknesses. First, they allow only integer values for γ. Second, unless a very slow annealing schedule is used, the MCMC chain tends to become trapped in local modes.

We note that a specialised version of this approach intended for Bayesian optimal design was also proposed [116] and, following the publication of the work presented here, an SMC version of this has also been suggested [99].

We propose another approach to sampling from $p_\gamma(\theta, z_{1:\gamma}|y)$. We sample from this sequence of distributions using SMC. SMC methods have been used primarily to solve optimal filtering problems in signal processing and statistics. They are used here in a completely different framework which requires extensions of the

methodology described further below. Broadly speaking, the distributions of interest are approximated by a collection of random samples termed *particles* which evolve over time using sampling and resampling mechanisms. The population of samples employed by this method makes it much less prone to trapping in local maxima, and the framework naturally allows for the introduction of bridging densities between target distributions, say, $p_\gamma(\theta, z_{1:\gamma}|y)$ and $p_{\gamma+1}(\theta, z_{1:\gamma+1}|y)$ for some integer γ – something which is essential to obtain good results in cases with sharply peaked target distributions.

At first glance, the algorithm appears very close to mutation-selection schemes employed in the genetic algorithms literature. However, there are two major differences with these algorithms. First, they require the function being maximized to be known pointwise, whereas we do not. Second, convergence results for our method follow straightforwardly from general results on Feynman-Kac flows [34].

4.2 An SMC Sampler Approach

4.2.1 Methodology

We will consider the use of the sampling methodology described in the previous section for marginal ML estimation – noting that the method can be easily adapted to Bayesian marginal MAP setting by considering a slightly different sequence of target distributions. We also note that such an approach, with suitably diffuse priors has an interpretation as a penalised maximum likelihood approach in non-Bayesian settings, and can be meaningfully applied to systems with unbounded likelihoods, unlike direct maximum likelihood estimation. The target distribution which we propose as generally admissible for this task is (where we have supressed the dependence of these distributions upon the observed data, which is assumed fixed, for convenience):

$$\pi_t(\theta, z_{1:\lceil\gamma_t\rceil}) \propto p(\theta)p(z_{\lceil\gamma_t\rceil}|\theta)^{\gamma_t\sharp}\prod_{i=1}^{\lfloor\gamma_t\rfloor} p(z_i|\theta). \tag{4.5}$$

This additional term allows us to introduce a sequence with non-integer elements, whilst having the same form as (4.4). Clearly, we have $\pi_t(\theta, z_{1:\lceil\gamma_t\rceil}|y) = p_{\gamma_t}(\theta, z_{1:\gamma_t}|y)$ for any integer γ_t. Again, an increasing sequence $(\gamma_t)_{t\geq 1}$ is required, corresponding in some sense to the annealing schedule of SA. To simplify notation we will denote $Z^{(i)}_{t,1:\lceil\gamma_t\rceil}$ – the values of $z_{1:\lceil\gamma_t\rceil}$ simulated at time t for the i^{th} particle – by $Z^{(i)}_t$.

We propose obtaining weighted sets of samples from distributions of the form of (4.5) with a monotonically increasing sequence $\{\gamma_t\}_{t=1}^T$, by employing an SMC sampler. Algorithm 4.1 describes the general framework which we propose. In

order to make use of this framework, it is necessary to specify an initial importance distribution, ν, as well as forward and backward transition kernels (K_t and L_t) for each step of the algorithm. To that end, we now go on to describe two particular cases of this algorithm – one which is applicable to a limited, but common class of models and another which is much more widely applicable.

Algorithm 4.1 A general SMC algorithm for MML estimation.

Initialisation: $t = 1$:
Sample, $\left\{\left(\theta_1^{(i)}, Z_1^{(i)}\right)\right\}_{i=1}^N$ independently from $\nu(\cdot)$.
Calculate importance weights $W_1^{(i)} \propto \frac{\pi_1(\theta_1^{(i)}, Z_1^{(i)})}{\nu(\theta_1^{(i)}, Z_1^{(i)})}$.
for $t = 2$ to T **do**
  if ESS $<$ Threshold, **then**
    resample.
  end if
  Sample, $\left\{\left(\theta_t^{(i)}, Z_t^{(i)}\right)\right\}_{i=1}^N$ such that $\left(\theta_t^{(i)}, Z_t^{(i)}\right) \sim K_t\left(\left(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}\right), \cdot\right)$.
  Set importance weights, $\frac{W_t^{(i)}}{W_{t-1}^{(i)}} \propto \frac{\pi_t(\theta_t^{(i)}, Z_t^{(i)}) L_{t-1}\left(\left(\theta_t^{(i)}, Z_t^{(i)}\right), \left(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}\right)\right)}{\pi_{t-1}(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}) K_t\left(\left(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}\right), \left(\theta_t^{(i)}, Z_t^{(i)}\right)\right)}$.
end for

Initially, it is interesting to consider an analytically convenient special case, which leads to a particularly elegant algorithm, 4.2, below. When we are able to sample from particular conditional distributions, and evaluate the marginal likelihood pointwise, it is possible to evaluate the optimal auxiliary kernels as given by (2.25).

Algorithm 4.2 A special case SMC algorithm for MML estimation.

Initialisation: $t = 1$:
Sample, $\left\{\tilde{\theta}_1^{(i)}\right\}_{i=1}^N$ independently from $\nu(\cdot)$.
Calculate importance weights $\tilde{W}_1^{(i)} \propto \frac{\pi_1(\tilde{\theta}_1^{(i)})}{\nu(\tilde{\theta}_1^{(i)})}$.
if ESS $<$ Threshold, **then**
  Resample $\left\{\theta_1^{(i)}\right\}_{i=1}^N \sim \sum_{j=1}^N \tilde{W}_1^{(j)} \delta_{\tilde{\theta}_1^{(j)}}(\cdot)$ and set $\left\{W_1^{(i)} = 1/N\right\}_{i=1}^N$.
else
  Let $\left\{\theta_1^{(i)}, W_1^{(i)}\right\}_{i=1}^N = \left\{\tilde{\theta}_1^{(i)}, \tilde{W}_1^{(i)}\right\}_{i=1}^N$.
end if
Sample, $\left\{Z_1^{(i)}\right\}_{i=1}^N$ from the conditional distribution $p(Z_1^{(i)}|\theta_1^{(i)})$.
for $t = 2$ to T **do**
  Sample, $\left\{\tilde{\theta}_t^{(i)}\right\}_{i=1}^N$ such that $\tilde{\theta}_t^{(i)} \sim \pi_t(\cdot|Z_{t-1}^{(i)})$.
  Calculate importance weights: $\tilde{W}_t^{(i)} \propto W_{t-1}^{(i)} p(y|\tilde{\theta}_t^{(i)})^{\gamma_t - \gamma_{t-1}}$.
  if ESS $<$ Threshold, **then**
    Resample $\left\{\theta_t^{(i)}\right\}_{i=1}^N \sim \sum_{j=1}^N \tilde{W}_t^{(j)} \delta_{\tilde{\theta}_t^{(j)}}(\cdot)$ and set $\left\{W_t^{(i)} = 1/N\right\}_{i=1}^N$.
  else
    Let $\left\{\theta_t^{(i)}, W_t^{(i)}\right\}_{i=1}^N = \left\{\tilde{\theta}_t^{(i)}, \tilde{W}_t^{(i)}\right\}_{i=1}^N$.
  end if
  Sample $Z_{t,1:\lceil\gamma_t\rceil}^{(i)} \sim \pi_t(Z_t^{(i)}|\theta_t^{(i)})$.
end for

This algorithm fits directly into the framework of algorithm 4.1 making use of an auxiliary kernel which combines a time-reversal Markov kernel for those elements which were present at the previous time step, and an "optimal" component for those elements which are newly generated,

$$L_{t-1}(\theta_t, Z_t; \theta_{t-1}, Z_{t-1}) = \frac{\pi_{t-1}(\theta_{t-1}, Z_{t-1})K_t(\theta_{t-1}, Z_{t-1}; \theta_t, Z_t)}{\pi_{t-1}(\theta_t, Z_{t,1:\lfloor\gamma_t\rfloor})\pi_t(Z_{\lfloor\gamma_t\rfloor+1:\lceil\gamma_t\rceil}|\theta_t)},$$

which leads to the weight expression $W_t = \pi_t(\theta_t)/\pi_{t-1}(\theta_t)$. This has a convenient interpretation as a resample-move algorithm [63, 64] on a state space of fixed dimension which allows for a simpler theoretical analysis. Essentially, one simply considers the sampling of the auxiliary variables to be a way to employ a kernel defined on Θ by $K_t(\theta_{t-1}, \theta_t) = \int \pi_t(\theta_t|z_t)\pi_t(z_t|\theta_{t-1})dz_t$ and everything follows directly.

Finally, we present a generic form of the algorithm which can be applied to a broad class of problems, although it will often be less efficient to use this generic formulation than to construct a dedicated sampler for a particular class of problems. We assume that a collection of Markov kernels $(\mathcal{K}_t)_{t\geq 1}$ with invariant distributions corresponding to $(\pi_t)_{t\geq 1}$ is available, using these as a component of the proposal kernels allows the evaluation of the optimal auxiliary kernel. That is, we set,

$$\begin{aligned} L_{t-1}(\theta_t, z_t; \theta_{t-1}, z_{t-1}) =& \frac{\pi_{t-1}(\theta_{t-1}, z_{t-1})K_t(\theta_{t-1}, z_{t-1}; \theta_t, z_t)}{\pi_{t-1}K_t(\theta_t, z_t)} \\ =& \frac{\pi_{t-1}(\theta_t, z_t)\mathcal{K}(\theta_{t-1}, z_{t-1}; \theta_t, Z_{t,1:\lfloor\gamma_{t-1}\rfloor})}{\pi_{t-1}(\theta_t, Z_{1:\lfloor\gamma_{t-1}\rfloor})}. \end{aligned}$$

We assume that good importance distributions for the conditional probability of the variables being marginalised can be sampled from and evaluated, $q(\cdot|\theta)$, and, if the annealing schedule is to include non-integer inverse temperatures, then we have appropriate importance distributions for distributions proportional to $p(z|\theta)^\alpha, \alpha \in (0, 1)$, which we denote $q_\alpha(z|\theta)$. We remark that this is not the most general possible approach, but is one which should work acceptably for a broad class of problems. It is also possible to incorporate MCMC moves into the algorithm, using the associated time-reversal kernel as the auxiliary kernel if the optimal form cannot be obtained, in order to assist mixing if required although this has not been necessary with the examples considered here.

There are a number of possible estimators associated with these algorithms. When the marginal likelihood cannot readily be evaluated, we recommend that the estimate is taken to be the first moment of the empirical distribution induced by the final particle ensemble; this may be justified by the asymptotic (in the inverse temperature) normality of the target distribution (see, for example, [130,

Algorithm 4.3 A generic SMC algorithm for MML estimation.

Initialisation: $t = 1$:
Sample, $\left\{\left(\theta_1^{(i)}, Z_1^{(i)}\right)\right\}_{i=1}^{N}$ independently from $\nu(\cdot)$.
Calculate importance weights $W_1^{(i)} \propto \frac{\pi_1(\theta_1^{(i)}, Z_1^{(i)})}{\nu(\theta_1^{(i)}, Z_1^{(i)})}$.
for $t = 2$ to T **do**
  if ESS $<$ Threshold, **then**
    resample.
  end if
  Sample, $\left\{\left(\theta_t^{(i)}, Z_t^{(i)}\right)\right\}_{i=1}^{N}$ such that $\left(\theta_t^{(i)}, Z_{t,1:\lfloor\gamma_{t-1}\rfloor}^{(i)}\right) \sim \mathcal{K}_{t-1}\left(\theta_{t-1}^{(i)}, Z_{t-1}^{(i)}; \cdot\right)$, for $j = \lfloor\gamma_{t-1}\rfloor + 1$ to $\lfloor\gamma_t\rfloor$, $Z_{t,j}^{(i)} \sim q(\cdot|\theta_t^{(i)})$ and $Z_{t,\lceil\gamma_t\rceil}^{(i)} \sim q_{\gamma_t - \lfloor\gamma_t\rfloor}(\cdot|\theta_t^{(i)})$
  Set importance weights: $\frac{W_t^{(i)}}{W_{t-1}^{(i)}} \propto \prod_{j=\lfloor\gamma_{t-1}\rfloor+1}^{\lfloor\gamma_t\rfloor} \frac{p(y, Z_j^{(i)}|\theta^{(i)}) p(y, Z_{\lceil\gamma_t\rceil}^{(i)}|\theta^{(i)})^{\gamma_t - \lfloor\gamma_t\rfloor}}{q(Z_j^{(i)}|\theta^{(i)}) q_{\gamma_t - \lfloor\gamma_t\rfloor}(Z_{\gamma_t}^{(i)}|\theta^{(i)})}$
end for

p. 203]). This is the approach which we took in section 4.3. In those cases in which the cheap evaluation of the marginal likelihood (or posterior, where appropriate) is possible it would also be possible to choose the parameters associated with the particle with the largest value of this likelihood, although there seems to be little advantage in doing so, except perhaps for systems in which multiple global optima exist and are simultaneously explored by the particle set.

4.2.2 Convergence

In order to demonstrate that the estimates obtained by the algorithms proposed in section 4.2.1 converge to the maximum likelihood estimator there are two elements which need to be proved. First, it is necessary to demonstrate that the mean of the sequence of the target distributions converges to the maximum likelihood estimator in question; secondly, it is necessary to demonstrate that the particle system estimate of this quantity converges to its true value. In the interests of simplicity, we consider only the case in which the parameter space is some subset of $\mathbb{R}^n$, i.e. $\Theta \subset \mathbb{R}^n$. We denote the set of maximum likelihood estimates:

$$\Theta_{ML} \triangleq \left\{\theta : p(y|\theta) = \sup_{\theta' \in \Theta} p(y|\theta')\right\}.$$

As a preliminary, we demonstrate that under very weak conditions, the maxima of the synthetic distribution 4.5 coincide with those of the likelihood function.

Theorem 4.2.1. *Provided that the likelihood is bounded above and the following conditions hold:*

$$\inf_{\theta \in \Theta_{ML}} p(\theta) = \alpha > 0,$$

$$\textit{and } \sup_{\theta' \in \Theta} p(\theta) < \infty,$$

we have the following result:

$$\forall \theta \in \Theta_{ML}, \theta' \in \Theta \setminus \Theta_{ML} : \lim_{\gamma \to \infty} \pi_\gamma(\theta) - \pi_\gamma(\theta') > 0$$

Proof. Take some $\theta' \in \Theta \setminus \Theta_{ML}$ and any $\theta_{ML} \in \Theta_{ML}$; let $\beta = p(\theta') < \infty$ and $\epsilon = p(y|\theta_{ML}) - p(y|\theta') > 0$ where positivity follows from the fact that θ' is not a member of the set of maximum likelihood estimators. Here, we have:

$$\frac{\pi_\gamma(\theta_{ML})}{\pi_\gamma(\theta')} \geq \frac{\alpha}{\beta}(1+\epsilon)^\gamma,$$

which exceeds unity providing that:

$$\gamma > \frac{\log(\beta/\alpha)}{\log(1+\epsilon)}.$$

For every $\epsilon > 0$, there exists some $\gamma < \infty$ such that this condition holds, and hence we have the result of the theorem. □

Convergence to the ML Estimate. We begin by demonstrating that the distribution 4.3 concentrates itself, as $\gamma \uparrow \infty$, on the set of maximum likelihood estimates under some weak assumptions. The following assumptions are sufficient to obtain the result which we require:

Assumption 1. $p(\theta)$ and $p(y|\theta)$ are α-Lipschitz continuous in θ for some $\alpha > 0$, and furthermore both $\log(p(\theta)) \in \mathcal{C}^3(\mathbb{R}^n)$ and $\log p(y|\theta) \in \mathcal{C}^3(\mathbb{R}^n)$.

Assumption 2. Θ_{ML} is a non-empty, countable set which is nowhere dense; $p(\theta)$ is assumed to be bounded above and to be non-zero on the points of Θ_{ML}; and $p(y|\theta)$ is assumed to be bounded above.

Assumption 3. The dominating measure has no mass at the points of the set of maximum likelihood solutions, $\lambda(\{t : t \in \Theta_{ML}\}) = 0$. In practice, this is ensured by assumption 2 and 5; if one wishes to relax assumption 5 then this explicit assumption is required.

Assumption 4. There exists a non-empty level set of $p(y|\theta)$ which is compact. For some $k < \sup p(y|\theta)$, $\{\theta : p(y|\theta) \geq k\}$ is compact.

Intuitively, these assumptions amount to requiring that there are not *too many* parameter values for which the maximum of the likelihood is obtained, the density is reasonably well behaved in terms of continuity properties and the support of the measure is compact if we are able to neglect the tails. These assumptions are very reasonable considering the problem at hand. However, verifying assumption 1 could be problematic in general as one does not typically have an analytic expression for $p(y|\theta)$.

We make the following assumption, which is somewhat stronger than that which is required in order to simplify the presentation without unduly restricting the applicability of our result:

Assumption 5. The dominating measure with respect to which π_γ admits a distribution is the Lebesgue measure on $\mathbb{R}^n$, denoted $\lambda(dx)$.

Theorem 4.2.2. *Under assumptions 1 to 5 the measure of interest converges to a distribution which is singular on the points of maximum likelihood, weighted by the Jacobian of the transformation implied by its density with respect to Lebesgue measure:*

$$\lim_{\gamma\to\infty} \pi_\gamma(dt) \propto \sum_{\theta_{ml}\in\Theta_{ML}} \alpha(\theta_{ml})\delta_{\theta_{ml}}(dt), \tag{4.6}$$

$$\alpha(\theta_{ml}) = \det\left[- \left.\frac{\partial^2 \log p(y|\theta)}{\partial\theta_m \partial\theta_n}\right|_{\theta=\theta_{ml}} \right]^{-1/2} \tag{4.7}$$

Proof. Writing $\pi_{\gamma,\vartheta}(\theta) \propto \exp\left(\vartheta\left[\log(p(\theta)^{1/\gamma}) + \log(p(y|\theta))\right]\right)$ noting that we have $\pi_\gamma = \pi_{\gamma,\vartheta}|_{\vartheta=\gamma}$, assumptions 1 to 5 are sufficient to ensure that conditions (A1) to (A5) of [80, Theorem 2.1] hold for all γ, and so:

$$\lim_{\vartheta\to\infty} \pi_{\gamma,\vartheta}(dt) \propto \sum_{\theta_{ml}\in\Theta_{ML}} \alpha_\gamma(\theta_{ml})\delta_{\theta_{ml}}(dt)$$

$$\alpha_\gamma(\theta_{ml}) = \det\left[- \left.\frac{\partial^2 \left[\frac{1}{\gamma}\log p(\theta) + \log p(y|\theta)\right]}{\partial\theta_m \partial\theta_n}\right|_{\theta=\theta_{ml}} \right]^{-1/2}$$

the result follows directly by taking the limit as $\gamma \to \infty$. □

Convergence of the Particle System. Having demonstrated that our target distribution converges to the set of maximum likelihood estimates, we now wish to determine the circumstances under which the empirical measure associated with the interacting particle systems proposed in section 4.2.1 converges to that target distribution. It is convenient throughout this section to use the symbol x_t to refer to the full set of variables sampled at time t, i.e. $x_t = \{\theta_t, z_{t,1:\gamma_t}\}$.

As a starting point, we note that under suitable conditions the following central limit theorem holds [37]:

Theorem 4.2.3 (Del Moral et al.). *Under the weak integrability conditions given in [34, Chapter 9, p300–306] and [25, Theorem 4], the following central limit applies for all $2 \leq t$, providing that multinomial resampling is applied after every iteration:*

$$\lim_{N\to\infty} \sqrt{N}\left[\pi_t^N(\psi) - \pi_t(\psi)\right] \xrightarrow{d} \mathcal{N}\left(0, \sigma_t(\psi)^2\right) \tag{4.8}$$

where π_t^N denotes the empirical measure associated with a system of N particles at time t, ψ is a sufficiently regular test function (see [25] for details) and the variance may be expressed as:

$$\sigma_t(\psi)^2 = \int \frac{\tilde{\pi}_t(x_1)^2}{\nu(x_1)} \left(\int \psi(x_t)\tilde{\pi}_t(x_t|x_1) - \mathbb{E}_{\pi_t}(\psi) \right)^2 dx_1 + \sum_{i=1}^{t-1} \int \frac{(\tilde{\pi}_t(x_k)L_{k-1}(x_k, x_{k-1}))^2}{\pi_{k-1}(x_{k-1})K_k(x_{k-1}, x_k)} \left(\int \psi(x_t)\tilde{\pi}_t(x_t|x_k)dx_t - \mathbb{E}_{\pi_t}(\psi) \right)^2 dx_{k-1:k} + \int \frac{(\pi_t(x_t)L_{t-1}(x_t, x_{t-1}))^2}{\pi_{t-1}(x_{t-1})K_t(x_{t-1}, x_t)} (\psi(x_t) - \mathbb{E}_{\pi_t}(\psi))^2 dx_{t-1:t} \quad (4.9)$$

where we have defined the following distributions for notational convenience:

$$\tilde{\pi}_t(x_k) = \int \tilde{\pi}_n(x_{1:t})dx_{1:k-1.k+1:t}$$
$$\tilde{\pi}_t(x_n|x_k) = \int \tilde{\pi}_n(x_{1:t})dx_{1:k-1.k+1:t-1}/\tilde{\pi}_t(x_k).$$

Finally, we present a stability result for the particle system. Under some strong assumptions it is possible to bound the variance expression 4.9 by a finite quantity at all finite times. This bound tends to infinity as the number of iterations does, but it is clearly far from tight and much work remains to be done on the stability of such systems. We present this result more as a proof of concept than as a practically applicable stability argument. For convenience and clarity, we consider only cases in which the annealing schedule takes integer values commencing from unity, although there should be no difficulty in generalising the result.

Assumption 6. The instrumental prior, the complete likelihood and the conditional distribution of the latent variables admit a density which is bounded above and below:

$$\begin{array}{ccccccccc} 0 & < & \epsilon_\theta & \leq & p(\theta) & \leq & \epsilon_\theta^{-1} & < & \infty, \\ 0 & < & \epsilon_y & \leq & p(y|z,\theta) & \leq & \epsilon_y^{-1} & < & \infty, \\ 0 & < & \epsilon_z & \leq & p(z|\theta) & \leq & \epsilon_z^{-1} & < & \infty. \end{array}$$

Assumption 7. The kernel, $\mathcal{K}$ employed to increase sample diversity in algorithm 4.3 admits a density which is bounded above and below[1], as does the proposal distribution used to sample latent variables:

$$\begin{array}{ccccccccc} 0 & < & \underline{K}_k & \leq & \mathcal{K}_k & \leq & \overline{K}_k & < & \infty, \\ 0 & < & \delta_k & \leq & q_k & \leq & \delta_k^{-1} & < & \infty, \end{array}$$

and the initial sampling distribution has the same property:

$$0 < \delta_1 \leq \nu \leq \delta_1^{-1} < \infty.$$

Theorem 4.2.4. *If we employ algorithm 4.3 and assumptions 6 and 7 hold, in addition to those required by theorem 4.2.3, then the asymptotic variance given*

[1] In practice, this condition is unlikely to be satisfied – the Metropolis-Hastings kernel which is typically employed for this purpose does not admit a density.

by expression 4.9 is finite for all $2 \leq t < \infty$ *and is bounded by the following expression:*

$$\begin{aligned}\sigma_t(\psi)^2 \leq & \frac{\overline{K}_2}{\underline{K}_2}\left(\delta_1\delta_2^{2(\gamma_2-1)}\epsilon_\theta\epsilon_y^2\epsilon_z\right)^{-1} + \\ & \sum_{k=1}^{t-1}\prod_{i=k-1}^{t-1}\left[\frac{\overline{K}_{i+1}}{\underline{K}_{i+1}}\delta_{i+1}^{-2(\gamma_{i+1}-\gamma_i)}\right](\epsilon_\theta\epsilon_y^{2\gamma_k}\epsilon_z^{\gamma_k}\underline{K}_k\delta_k^{\gamma_k-\gamma_{k-1}})^{-1}\operatorname{osc}(\psi)^2 + \\ & \left(\epsilon_\theta\epsilon_y^{2\gamma_t}\epsilon_z^{\gamma_t}\delta_t^{\gamma_t-\gamma_{t-1}}\underline{K}_t\right)^{-1}\operatorname{osc}(\psi)^2,\end{aligned}$$

where $\operatorname{osc}(\psi) \triangleq \sup_{x,x'}|\psi(x)-\psi(x')|$.

Proof. The proof employs lemmas 4.2.1 to 4.2.3 which follow this theorem.
The first term in the variance expression is bounded directly by lemma 4.2.2.
Considering the term:

$$\begin{aligned}&\int \frac{(\tilde{\pi}_t(x_k)L_{k-1}(x_k,x_{k-1}))^2}{\pi_{k-1}(x_{k-1})K_k(x_{k-1},x_k)}\left(\int\psi(x_t)\tilde{\pi}_t(x_t|x_k)dx_t - \mathbb{E}_{\pi_t}(\psi)\right)^2 dx_{k-1:k} \\ =&\int \frac{\tilde{\pi}_t(x_k)L_{k-1}(x_k,x_{k-1})}{\pi_{k-1}(x_{k-1})K_k(x_{k-1},x_k)}\left(\int\psi(x_t)\tilde{\pi}_t(x_t|x_k)dx_t - \mathbb{E}_{\pi_t}(\psi)\right)^2 \tilde{\pi}_t(x_k)L_{k-1}(x_k,x_{k-1})dx_{k-1:k} \\ \leq &\prod_{i=k-1}^{t-1}\left[\frac{\overline{K}_{i+1}}{\underline{K}_{i+1}}\delta_{i+1}^{-2(\gamma_{i+1}-\gamma_i)}\right](\epsilon_\theta\epsilon_y^{2\gamma_k}\epsilon_z^{\gamma_k}\underline{K}_k\delta_k^{\gamma_k-\gamma_{k-1}})^{-1}\times \\ &\int\left(\int\psi(x_t)\tilde{\pi}_t(x_t|x_k)dx_t - \mathbb{E}_{\pi_t}(\psi)\right)^2\tilde{\pi}_t(x_k)L_{k-1}(x_k,x_{k-1})dx_{k-1:k},\end{aligned}$$

where the final line follows by lemma 4.2.3. It is then apparent that we may bound this term by:

$$\prod_{i=k-1}^{t-1}\left[\frac{\overline{K}_{i+1}}{\underline{K}_{i+1}}\delta_{i+1}^{-2(\gamma_{i+1}-\gamma_i)}\right](\epsilon_\theta\epsilon_y^{2\gamma_k}\epsilon_z^{\gamma_k}\underline{K}_k\delta_k^{\gamma_k-\gamma_{k-1}})^{-1}\operatorname{osc}(\psi)^2.$$

The final term in the variance expression is bounded by:

$$\left(\epsilon_\theta\epsilon_y^{2\gamma_t}\epsilon_z^{\gamma_t}\delta_t^{\gamma_t-\gamma_{t-1}}\underline{K}_t\right)^{-1}\operatorname{osc}(\psi)^2,$$

by very similar arguments.

□

Lemma 4.2.1. *Under assumption 6 we may bound the normalising constant associated with the target distributions:*

$$\begin{aligned}\mathcal{Z}_k &= \int p(\theta_k)\prod_{l=1}^{\gamma_k}p(z_{k,l}|\theta_k)p(y|z_{k,l},\theta_k)d\theta_k dz_k \\ &\leq \int p(\theta_k)\prod_{l=1}^{\gamma_k}p(z_{k,l}|\theta_k)\epsilon_y^{-1}d\theta_k dz_k && \leq \epsilon_y^{-\gamma_k}, \\ \textit{and similarly: } \mathcal{Z}_k &\geq \int p(\theta_k)\prod_{l=1}^{\gamma_k}p(z_{k,l}|\theta_k)\epsilon_y d\theta_k dz_k && \geq \epsilon_y^{\gamma_k}.\end{aligned}$$

.

Lemma 4.2.2. *The following bound applies under assumptions 6 and 7:*

$$\frac{\tilde{\pi}_t(x_1)}{\nu(x_1)} \leq \frac{\overline{K}_2}{\underline{K}_2}\left(\delta_1 \delta_2^{2(\gamma_2-1)} \epsilon_\theta \epsilon_y^2 \epsilon_z\right)^{-1}.$$

Proof. The result follows directly from the assumptions:

$$\begin{aligned}
\frac{\tilde{\pi}_t(x_1)}{\nu(x_1)} &= \frac{\int dx_{2:n} \pi_n(x_n) \prod_{j=1}^{n-1} L_j(x_{j+1}, x_J)}{\nu(x_1)} \\
&\leq \sup_{x_2} \frac{L_1(x_2, x_1)}{\nu(x_1)} \\
&= \sup_{x_2} \frac{\pi_1(x_1) K_2(x_1, x_2)}{\pi_1 K_2(x_2) \nu(x_1)} \\
&\leq (\delta_1 \epsilon_\theta \epsilon_y^2 \epsilon_z)^{-1} \frac{\overline{K}_2}{\underline{K}_2} \delta_2^{-2(\gamma_2-1)}
\end{aligned}$$

□

Lemma 4.2.3. *Under assumption 6 and 7 we may bound the following ratio of densities:*

$$\frac{\tilde{\pi}_t(x_k) L_{k-1}(x_k, x_{k-1})}{\pi_{k-1}(x_{k-1}) K_k(x_{k-1}, x_k)} \leq \prod_{i=k-1}^{t-1} \left[\frac{\overline{K}_{i+1}}{\underline{K}_{i+1}} \delta_{i+1}^{-2(\gamma_{i+1}-\gamma_i)}\right] (\epsilon_\theta \epsilon_y^{2\gamma_k} \epsilon_z^{\gamma_k} \underline{K}_k \delta_k^{\gamma_k - \gamma_{k-1}})^{-1}$$

Proof.

$$\begin{aligned}
\frac{\tilde{\pi}_t(x_k) L_{k-1}(x_k, x_{k-1})}{\pi_{k-1}(x_{k-1}) K_k(x_{k-1}, x_k)} &= \frac{\int dx_{t:k+1} \pi_t(x_t) \prod_{i=k}^{t-1} L_i(x_{i+1}, x_i) L_{k-1}(x_k, x_{k-1})}{\pi_{k-1}(x_{k-1}) K_k(x_{k-1}, x_k)} \\
&= \frac{\int dx_{t:k+1} \pi_t(x_t) \prod_{i=k-1}^{t-1} \left[\frac{\pi_i(x_i) K_{i+1}(x_i, x_{i+1})}{\pi_i K_{i+1}(x_{i+1})}\right]}{\pi_{k-1}(x_{k-1}) K_k(x_{k-1}, x_k)} \\
&= \frac{\int dx_{t:k+1} \prod_{i=k-1}^{t} \pi_i(x_i) \prod_{i=k-1}^{t-1} \left[\frac{K_{i+1}(x_i, x_{i+1})}{\pi_i K_{i+1}(x_{i+1})}\right]}{\pi_{k-1}(x_{k-1}) K_k(x_{k-1}, x_k)}.
\end{aligned}$$

Noting that we may bound the terms in the second product,

$$\begin{aligned}
\frac{K_{i+1}(x_i, x_{i+1})}{\pi_i K_{i+1}(x_{i+1})} &\leq \sup_{x_i, x_i'} \frac{K_{i+1}(x_i, x_{i+1})}{K_{i+1}(x_i', x_{i+1})} \\
&\leq \frac{\overline{K}_{i+1}}{\underline{K}_{i+1}} \delta_{i+1}^{-2(\gamma_{i+1}-\gamma_i)},
\end{aligned}$$

we obtain:

$$\frac{\tilde{\pi}_t(x_k)L_{k-1}(x_k,x_{k-1})}{\pi_{k-1}(x_{k-1})K_k(x_{k-1},x_k)} \leq \frac{\int dx_{t:k+1} \prod_{i=k-1}^{t} \pi_i(x_i) \prod_{i=k-1}^{t-1} \left[\frac{\overline{K}_{i+1}}{\underline{K}_{i+1}}\delta_{i+1}^{-2(\gamma_{i+1}-\gamma_i)}\right]}{\pi_{k-1}(x_{k-1})K_k(x_{k-1},x_k)}$$
$$= \prod_{i=k-1}^{t-1}\left[\frac{\overline{K}_{i+1}}{\underline{K}_{i+1}}\delta_{i+1}^{-2(\gamma_{i+1}-\gamma_i)}\right]\frac{\pi_{k-1}(x_{k-1})\pi_k(x_k)}{\pi_{k-1}(x_{k-1})K_k(x_{k-1},x_k)}$$
$$= \prod_{i=k-1}^{t-1}\left[\frac{\overline{K}_{i+1}}{\underline{K}_{i+1}}\delta_{i+1}^{-2(\gamma_{i+1}-\gamma_i)}\right]\frac{\pi_k(x_k)}{K_k(x_{k-1},x_k)}.$$

Now, we may bound the final term:

$$\frac{\pi_k(x_k)}{K_k(x_{k-1},x_k)} = \frac{p(\theta_k)\prod_{l=1}^{\gamma_k} p(z_{k,l}|\theta_k)p(y|z_{k,l},\theta_k)/\mathcal{Z}_k}{\mathcal{K}_k\left(x_{k-1}, x_k \setminus z_{k,\gamma_{k-1}+1:\gamma_k}\right)\prod_{l=\gamma_{k-1}+1}^{\gamma_k} q_k(z_{k,l}|\theta_k)}$$
$$\leq (\epsilon_\theta \epsilon_y^{\gamma_k}\epsilon_z^{\gamma_k}\underline{K}_k\delta_k^{\gamma_k-\gamma_{k-1}}\mathcal{Z}_k)^{-1}$$
$$\leq (\epsilon_\theta \epsilon_y^{2\gamma_k}\epsilon_z^{\gamma_k}\underline{K}_k\delta_k^{\gamma_k-\gamma_{k-1}})^{-1}$$

where the final line follows from lemma 4.2.1 and this gives us the result of the lemma. □

4.3 Examples and Results

To demonstrate the effectiveness of the proposed algorithms, and to allow comparison with other techniques, we now show results on a simple toy example and a number of challenging models. We begin with a one dimensional example in section 4.3.1, followed by a Gaussian mixture model in section 4.3.2, a non-linear non-Gaussian state space model which is widely used in financial modelling in section 4.3.3 and finally an auxiliary variable formulation of the logistic regression problem in section 4.3.4.

For the purpose of comparing algorithms on an equal footing, it is necessary to employ some measure of computational complexity. In the interests of simplicity we note that almost all of the computational cost associated with each of the algorithms considered here come from either sampling the latent variables or determining their expectation. In fact, evaluating the parameters of the distributions from which they are sampled or the expressions from which their expectations may be obtained is by far the largest cost. We introduce the quantity χ defined as the total number of complete replicates of the latent variable vector which needs to be simulated or estimated in one complete run of an algorithm (note that in the case of SAME and the SMC algorithm, this figure depends upon the annealing

schedule as well as the final temperature and the number of particles in the SMC case).

4.3.1 Toy Example

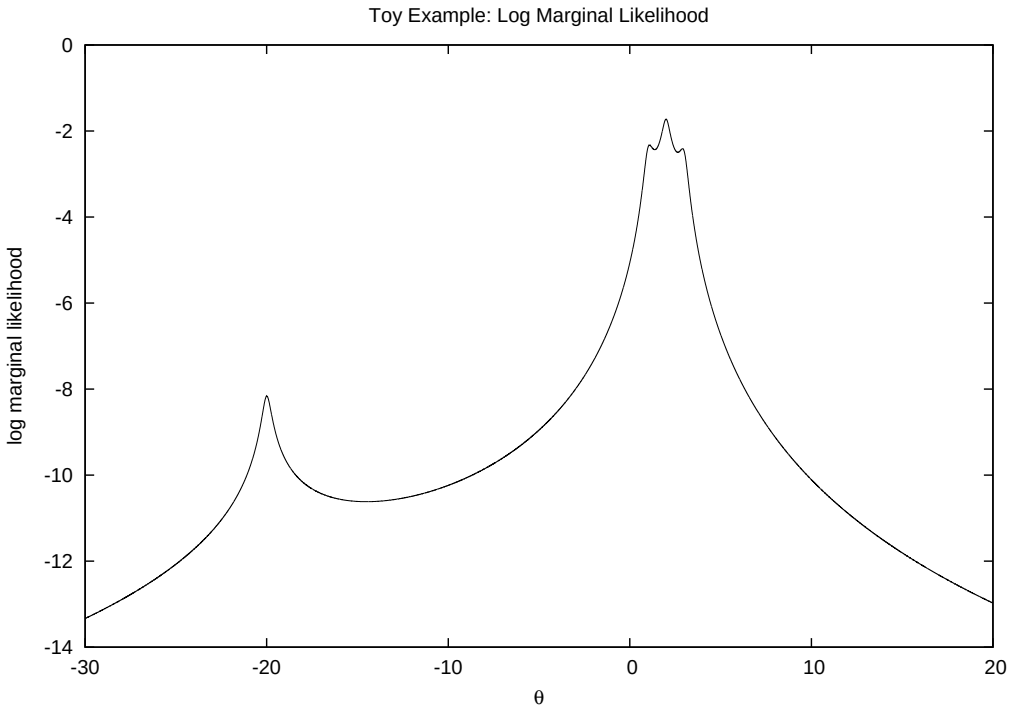


Fig. 4.1. The log marginal likelihood of the toy example of section 4.3.1.

We consider first a toy example in one dimension for which we borrow example 1 of [55]. The model consists of a student t-distribution of unknown location parameter θ with 0.05 degrees of freedom. Four observations are available, $y = (-20, 1, 2, 3)$. The logarithm of the marginal likelihood in this instance is given by:

$$\log p(y|\theta) = -0.525 \sum_{i=1}^{4} \log\left(0.05 + (y_i - \theta)^2\right)$$

which is not susceptible to analytic maximisation. However, global the maximum is known to be located at 1.997, and local maxima exist at $\{-19.993, 1.086, 2.906\}$ as illustrated in figure 4.1. We can complete this model by considering the student t-distribution as a scale-mixture of Gaussians and associating a latent precision parameter Z_i with each observation. The log likelihood is then:

$$\log p(y, z|\theta) = -\sum_{i=1}^{4} \left[0.475 \log z_i + 0.025 z_i + 0.5 z_i (y_i - \theta)^2\right]$$

In the interest of simplicity, we make use of a linear temperature scale, $\gamma_t = t$, which takes only integer values. We are able to evaluate the marginal likelihood function pointwise, and can sample from the conditional distributions:

N	T	Mean	Std. Dev.	Min	Max
50	15	1.992	0.014	1.95	2.03
100	15	1.997	0.013	1.97	2.04
20	30	1.958	0.177	1.09	2.04
50	30	1.997	0.008	1.98	2.01
100	30	1.997	0.007	1.98	2.01
20	60	1.998	0.015	1.91	2.02
50	60	1.997	0.005	1.99	2.01

Table 4.1. Simulation results for the toy problem. Each line summarises 50 simulations with N particles and final temperature T. Only one simulation failed to find the correct mode.

$$\pi_t(z_{1:\gamma_t}|\theta, y) = \prod_{i=1}^{\gamma_t} \mathcal{G}a\left(z_i \middle| 0.525, 0.025 + \frac{(y_i - \theta)^2}{2}\right), \tag{4.10}$$

$$\pi_t(\theta|z_{1:\gamma_t}) \propto \mathcal{N}\left(\theta \middle| \mu_t^{(\theta)}, \Sigma_t^{(\theta)}\right), \tag{4.11}$$

where the parameters,

$$\Sigma_t^{(\theta)} = \left[\sum_{i=1}^{t}\sum_{j=1}^{4} z_{i,j}\right]^{-1} = \left[1/\Sigma_{t-1}^{(\theta)} + \sum_{j=1}^{4} z_{t,j}\right]^{-1}, \tag{4.12}$$

$$\mu_t^{(\theta)} = \Sigma_t^{(\theta)} \sum_{i=1}^{t} y^T z_i \quad = \Sigma_t^{(\mu)}\left(\mu_{t-1}^{(\theta)}/\Sigma_{t-1}^{(\theta)} + y^T z_t\right), \tag{4.13}$$

may be obtained recursively. Consequently, one can make use of algorithm 4.2 to solve this problem. We use an instrumental uniform $[-50, 50]$ prior distribution over θ. Some simulation results are given in table 4.1. The estimate is taken to be the first moment of the empirical distribution induced by the final particle ensemble.

4.3.2 A Finite Gaussian Mixture Model

To allow comparison with other techniques, and to illustrate the strength of the method proposed here in avoiding local maxima, we consider a finite Gaussian Mixture model. A set of observations $\{y_i\}_{i=1}^{P}$ is assumed to consist of P *iid* samples from a distribution of the form:

$$p(y_i|\pi, \mu\sigma^2) = \sum_{s=1}^{S} \pi_s \times \mathcal{N}\left(y_i|\mu_s, \sigma_s^2\right) \tag{4.14}$$

where $0 < \pi_s < 1; \sum_{s=1}^{S} \pi_s = 1$ are the weights of each mixture component and $\{\mu_s, \sigma_s^2\}_{s=1}^{S}$ is the set of their means and variances. As is usual with such mixtures, it is convenient to introduce auxiliary allocation variables, Z_i which allow us to assign each observation to one of the mixture components, then we may write the distribution in the form:

$$p(y_i|\pi, \mu, \sigma^2, Z_i = z_i) = \mathcal{N}\left(y_i|\mu_{z_i}, \sigma_{z_i}^2\right), \quad p(Z_i = z_i) = \pi_{z_i}.$$

It is both well known and somewhat obvious, that the maximum likelihood estimate of all parameters of this model is not well defined as the likelihood is not bounded. However, the inclusion of prior distributions over the parameters has a bounding effect and makes MAP estimation possible [131]. We consequently show the results of all algorithms adapted for MAP estimation by inclusion of diffuse priors, which are as follows:

$$\begin{aligned}\pi &\sim \mathcal{D}i\left(\chi\right)\\ \sigma_i^2 &\sim \mathcal{IG}\left(\frac{\lambda_i+3}{2},\frac{\beta_i}{2}\right)\\ \mu_i|\sigma_i^2 &\sim \mathcal{N}\left(\alpha_i,\sigma_i^2/\lambda_i\right)\end{aligned}$$

It is straightforward to adjust our algorithm 4.2 to deal with the MAP, rather than ML case. For this application it is possible to sample from all of the necessary distributions and to evaluate the marginal posterior pointwise and so we employ such an algorithm.

At iteration t of the algorithm, for particle i we sample the parameter estimates, conditioned upon the previous values of the latent variables according to the conditional distributions:

$$\begin{aligned}\pi &\leftarrow \mathcal{D}i\left(\gamma_t(\chi-1)+1+n(\lfloor\gamma_t\rfloor)+{\gamma_t}^\sharp\Delta n(\lceil\gamma_t\rceil)\right),\\ \sigma_i^2 &\leftarrow \mathcal{IG}\left(A_i,B_i\right)\\ \mu_i|\sigma_i^2 &\leftarrow \mathcal{N}\left(\frac{\gamma_t\lambda_i\alpha_i+\overline{y}\left(\lfloor\gamma_t\rfloor\right)_i+{\gamma_t}^\sharp\Delta\overline{y}\left(\lceil\gamma_t\rceil\right)_i}{\gamma_t\lambda_i+n\left(\lfloor\gamma_t\rfloor\right)_i+{\gamma_t}^\sharp\Delta n\left(\lceil\gamma_t\rceil\right)_i},\frac{\sigma_i^2}{\gamma_t\lambda_i+n\left(\lfloor\gamma_t\rfloor\right)_i{\gamma_t}^\sharp\Delta n\left(\lceil\gamma_t\rceil\right)_i}\right)\end{aligned}$$

where we have defined the following quantities for convenience:

$$\begin{aligned}n\left(i\right)_j &= \sum_{l=1}^{i}\sum_{p=1}^{P}\mathbb{I}_j(Z_{l,p}) & \Delta n\left(i\right)_j &= n\left(i\right)_j - n\left(i-1\right)_j\\ \overline{y}\left(i\right)_j &= \sum_{l=1}^{i}\sum_{p=1}^{P}\mathbb{I}_j(Z_{l,p})y_j & \Delta\overline{y}\left(i\right)_j &= \overline{y}\left(i\right)_j - \overline{y}\left(\left(i-1\right)\right)_j\\ \overline{y^2}\left(i\right)_j &= \sum_{l=1}^{i}\sum_{p=1}^{P}\mathbb{I}_j(Z_{l,p})y_j^2 & \Delta\overline{y^2}\left(i\right)_j &= \overline{y^2}\left(i\right)_j - \overline{y^2}\left(i-1\right)_j,\end{aligned}$$

and the parameters for the inverse gamma distribution from which the variances are sampled from are:

$$\begin{aligned}A_i &= \frac{\gamma_t(\lambda_i+1)+n\left(\lfloor\gamma_t\rfloor\right)_i+{\gamma_t}^\sharp\Delta n\left(\lceil\gamma_t\rceil\right)_i}{2}+1\\ B_i &= \frac{1}{2}\left(\gamma_t(\beta_i+\lambda_i\alpha_i^2)+\overline{y^2}\left(\lfloor\gamma_2\rfloor\right)_i+{\gamma_t}^\sharp\Delta\overline{y^2}\left(\lceil\gamma_t\rceil\right)_i-\right.\\ &\qquad\left.\sum_{g=1}^{\lfloor\gamma_t\rfloor}\frac{\left(\Delta\overline{y}\left(g\right)_i+\lambda_i\alpha_i\right)^2}{\lambda_i+\Delta n\left(g\right)_i}-{\gamma_t}^\sharp\frac{\left(\Delta\overline{y}\left(\lceil\gamma_t\rceil\right)_i+\lambda_i\alpha_i\right)^2}{\lambda_i+\Delta n\left(\lceil\gamma_t\rceil\right)_i}\right)\end{aligned}$$

Then we sample all of the allocation variables from the appropriate distributions, noting that this is equivalent to augmenting them with the new values and applying an MCMC move to those persisting from earlier iterations.

Simulated Data. We present results first from data simulated according to the model. 100 data were simulated from a distribution of the form of 4.14, with parameters $\pi = [0.2, 0.3, 0.5]$, $\mu = [0, 2, 3]$ and $\sigma = \left[1, \frac{1}{4}, \frac{1}{16}\right]$. The same simulated data set was used for all runs, and the log posterior density of the generating parameters was -155.87. Results for the SMC algorithm are shown in table 4.2 and for the other algorithms in table 4.3 – two different initialisation strategies were used for these algorithms, that described as "Prior" in which a parameter set was sampled from the prior distributions, and "Hull" in which the variances were set to unity, the mixture weights to one third and the means were sampled uniformly from the convex hull of the observations.

Two annealing schedules were used for the SAME algorithm, one, denoted SAME (6), involved keeping the number of replicates of the augmentation data fixed to 1 for the first half of the iterations and then increasing linearly to a final maximum value of 6; the other, denoted SAME (50), keeping it fixed to one for the first 250 iterations, and then increasing linearly to 50. The annealing schedule for the SMC algorithm was of the form $\gamma_t = Ae^{bt}$ for suitable constants to make $\gamma_1 = 0.01$ and $\gamma_T = 6$. This is motivated by the intuition that when γ is small, the effect of increasing it by some amount $\Delta\gamma$ is to change its form somewhat more than would be the case for a substantially larger value of γ. Varying the forms of the annealing schedules did not appear to substantially affect the results.

N	T	χ	**Mean**	**Std. Dev.**	**Min**	**Max**
25	25	1325	-154.39	0.55	-155.76	-153.64
25	50	2125	-153.88	0.13	-154.18	-153.59
50	50	4250	-153.80	0.08	-153.93	-153.64
100	50	8500	-153.74	0.07	-153.91	-153.59
250	50	21250	-153.70	0.07	-153.90	-153.54
1000	50	85000	-153.64	0.04	-153.71	-153.57
100	100	20300	-153.73	0.08	-153.92	-153.61

Table 4.2. Mean and standard deviation of final log posterior estimated by 50 runs of the SMC Algorithm on simulated data from a finite Gaussian mixture with varying numbers of particles, N, and intermediate distributions, T.

Galaxy Data. We also applied these algorithms, with the same parameters to the galaxy data of [133]. This data set consists of the velocities of 82 galaxies, and it has been suggested that it consists of a mixture of between 3 and 7 distinct components – for example, see [134] and [52]. For our purposes we have estimated the parameters of a 3 component Gaussian mixture model from which we assume

Algorithm	Init.	T	χ	Mean	Std. Dev.	Min	Max
EM	Prior	500	500	-169.79	8.50	-181.16	-160.70
EM	Hull	500	500	-158.06	3.23	-166.39	-153.85
EM	Prior	5000	5000	-168.24	8.41	-181.02	-153.83
EM	Hull	5000	5000	-157.73	3.83	-165.81	-153.83
SAME(6)	Prior	4250	8755	-155.45	0.82	-157.56	-154.06
SAME(6)	Hull	4250	8755	-155.32	0.87	-157.35	-154.03
SAME(50)	Prior	4250	112522	-154.91	0.81	-156.22	-153.94
SAME(50)	Hull	4250	112522	-155.05	0.82	-156.11	-153.98

Table 4.3. Performance of the EM and SAME Algorithm on simulated data from a finite Gaussian mixture. Means and standard deviations of the log posterior of the final estimates of 50 runs of each algorithm are shown.

the data was drawn. Results for the SMC algorithm are shown in table 4.4 and for the other algorithms in table 4.5.

N	T	χ	Mean	Std. Dev.	Min	Max
25	25	1325	-44.21	0.13	-44.60	-43.96
50	25	2650	-44.18	0.10	-44.48	-43.95
25	50	2125	-44.14	0.10	-44.32	-43.92
50	50	4250	-44.07	0.07	-44.22	-43.96
100	50	8500	-44.05	0.06	-44.18	-43.94
250	50	21250	-44.00	0.05	-44.10	-43.91
1000	50	85000	-43.96	0.03	-44.02	-43.92
100	100	20300	-44.03	0.05	-44.15	-43.94

Table 4.4. Mean and standard deviation of final log posterior estimated by 50 runs of the SMC Algorithm on the galaxy dataset of [133] from a finite Gaussian mixture with varying numbers of particles, N, and intermediate distributions, T.

Algorithm	Init.	T	χ	Mean	Std. Dev.	Min	Max
EM	Hull	500	500	-46.54	2.92	-54.12	-44.32
EM	Hull	5000	5000	-46.91	3.00	-56.68	-44.34
SAME(6)	Hull	4250	8755	-45.18	0.54	-46.61	-44.17
SAME(50)	Hull	4250	112522	-44.93	0.21	-45.52	-44.47

Table 4.5. Performance of the EM and SAME Algorithm on the galaxy data of [133] from a finite Gaussian mixture. Means and standard deviations of the log posterior of the final estimates of 50 runs of each algorithm are shown.

Discussion. The results obtained from the simulated data experiments illustrate that EM is prone to becoming trapped in local modes, which is supported by the results obtained on the real data – even at greater computational costs it is not able to perform as well as the SA-related algorithms. In contrast, both the SAME and SMC algorithms do much better – and it is clear that, for given computational cost, the SAME algorithm does not perform as well as the population-based method proposed here.

We note that precisely the same moves were used for the SAME algorithm and the SMC algorithm.

4.3.3 Stochastic Volatility

We take this more complex example from [81]. We consider the following model:

$$Z_i = \alpha + \delta Z_{i-1} + \sigma_u u_i \qquad Z_1 \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$
$$Y_i = \exp\left(\frac{Z_i}{2}\right)\epsilon_i$$

where u_i and ϵ_i are uncorrelated standard normal random variables, and $\theta = (\alpha, \delta, \sigma_u)$. The marginal likelihood of interest $p(\theta|y)$ is available only as a high dimensional integral over the latent variables, Z and this integral cannot be computed.

In this case we are unable to use algorithm 4.2, and employ a variant of algorithm 4.3. The serial nature of the observation sequence suggests introducing blocks of the latent variable at each time, rather than replicating the entire set at each iteration. This is motivated by the same considerations as the previously discussed sequence of distributions, but makes use of the structure of this particular model. Thus, at time t, given a set of M observations, we have a sample of $M\gamma_t$ volatilities, $\lfloor\gamma_t\rfloor$ complete sets and $M(\gamma_t - \lfloor\gamma_t\rfloor)$ which comprises a partial estimate of another replicate. That is, we use target distributions of the form:

$$p_t(\alpha, \delta, \sigma, z_t) \propto p(\alpha, \delta, \sigma) \prod_{i=1}^{\lfloor\gamma_t\rfloor} p(y, z_{t,i} | \alpha, \delta, \sigma)\, p\left(y_{1:M(\gamma_t - \lfloor\gamma_t\rfloor)}, z_{t,i}^{1:M(\gamma_t - \lfloor\gamma_t\rfloor)} \middle| \alpha, \delta, \sigma\right),$$

where $z_{t,i}^{1:M(\gamma_t - \lfloor\gamma_t\rfloor)}$ denotes the first $\gamma_t - \lfloor\gamma_t\rfloor$ volatilities of the i^{th} replicate at iteration t.

Making use of diffuse conjugate prior distributions[2] for θ ensures that the prior distributions are rapidly "forgotten", leading to a maximum likelihood estimate. Our sampling strategy at each time is to sample (α, δ) from their joint conditional distribution, then to sample σ from a Gibbs sampling kernel before proposing new volatilities using a Kalman-smoother as the proposal distribution.

Simulated Data. We consider a sequence of 500 observations generated from a stochastic volatility model with parameter values of $\alpha = -0.363$, $\delta = 0.95$ and $\sigma = 0.26$ (suggested by [81] as being *consistent with empirical estimates for financial equity return time series*). A linear annealing schedule increasing from $4/T$ to 4 in T steps was employed.

[2] i.e. uniform over the $(-1, 1)$ stability domain for δ, standard normal for α and square-root inverse gamma with parameters $\alpha = 1$, $\beta = 0.1$ for σ.

N	T	γ_T	α	δ	σ
1,000	250	4	-0.45 ± 0.19	0.939 ± 0.026	0.36 ± 0.09
1,000	500	4	-0.59 ± 0.27	0.919 ± 0.037	0.43 ± 0.11
1,000	1,000	4	-0.21 ± 0.02	0.973 ± 0.003	0.25 ± 0.02
5,000	250	4	-0.33 ± 0.06	0.954 ± 0.008	0.31 ± 0.04

Table 4.6. SMC Sampler results for simulated stochastic volatility data with generating parameters of $\delta = 0.95$, $\alpha = -0.363$ and $\sigma = 0.26$, estimates were obtained over 50 runs of each configuration.

Discussion. This is a difficult problem and it has not proved straightforward to find competitive approaches for ML point estimation in this problem – although it may be of interest to consider a MCEM approach using MCMC to perform the E-step. Although using simulated data provides knowledge of the generating parameter, it is not obvious what the *correct* answer is: as we are unable to evaluate the marginal likelihood in this case, we cannot compare its value given the generating parameters and those estimated by the algorithm. Pragmatically, one might expect the generating parameters to be close to the ML estimate given a reasonably large amount of data. Looking at the estimates shown in table 4.6, reasonable values for the persistence parameter δ are more readily obtained than the other parameters. The mean reversion parameter, α is strongly negatively correlated with δ and this is reflected in the particle set which is produced by the algorithm and indeed the final estimates. Estimating the variance parameter, σ is the most difficult problem as this is closely related to the particular sequences of latent variables which are considered.

The algorithm proposed here is able to obtain estimates of all three parameters which appear reasonable given knowledge of the generating parameters, but a substantial computational cost is attached to obtaining those estimates. In practice, better performance would be obtained by applying MCMC moves to the previously estimates volatilities in order to provide better mixing properties. Applying such moves, at least at the beginning of the sampling, might lead to substantially improved performance even at a given computational cost and the trade-off between the increased computational cost and improved mixing related to the application of such moves is something which requires investigation in SMC in general.

4.3.4 Bayesian Logistic Regression

As a final example, we consider the auxiliary variable formulation of Bayesian logistic regression proposed by [79]. As usual we are interested in estimating β given a set of noisy observations $\{y_i\}_{i=1}^n$ and exactly known covariates $\{x_i\}_{i=1}^n$ given the model:

$$
\begin{aligned}
y_i &\sim \mathcal{B}er\left(g^{-1}(\eta_i)\right) \\
\eta_i &= x_i \cdot \beta \\
\beta &\sim \mathcal{N}\left(b, v\right),
\end{aligned}
$$

where in our case the link function is taken to be the logit function ($\log(p/(1-p))$) allowing us to formulate the problem via appropriate auxiliary variables as:

$$
\begin{aligned}
y_i &= \mathbb{I}_{(0,\infty]}(z_i) & & \\
z_i &= x_i \cdot \beta + \epsilon_i & \epsilon_i &\sim \mathcal{N}\left(0, \lambda_i\right) \\
\lambda_i &= (2\psi_i)^2 & \psi_i &\sim \mathcal{KS} \\
\beta_i &\sim \mathcal{N}\left(b, v\right), & &
\end{aligned}
$$

where ϵ_i has the form of a scale mixture of normals with a marginal logistic distribution [2].

We treat β as a parameter to be estimated, and $\{\lambda_i, z_i\}_{i=1}^n$ as latent variables to be marginalised. In this instance we can sample everything from the appropriate conditional distributions. In the integer γ case, the distribution over λ_i is non-standard, but straightforward to sample from using the techniques suggested in [79] – and the others are simply:

$$
\begin{aligned}
\beta | z, \lambda, y &\sim \mathcal{N}\left(B, V\right) \\
z_i | \beta, x_i, y_i &\sim \begin{cases} \mathcal{L}ogistic\left(\beta x_i, 1\right) \mathbb{I}_{(0,\infty]}(z_i) & \text{if } y_i = 1 \\ \mathcal{L}ogistic\left(\beta x_i, 1\right) \mathbb{I}_{[-\infty,0]}(z_i) & \text{otherwise.} \end{cases}
\end{aligned}
$$

Rejection sampling could be used to permit non-integer γ. The parameters of the normal distribution over β are given by,

$$
\begin{aligned}
V &= \left[\gamma_t v^{-1} + \sum_{i=1}^{n} x_i x_i^T \left(\sum_{g=1}^{\gamma_t} \lambda_{gi}^{-1}\right)\right]^{-1} \\
B &= V\left[\gamma_t v^{-1} b + \sum_{i=1}^{n} x_i \sum_{g=1}^{\gamma_t} \frac{z_{gi}}{\lambda_{gi}}\right].
\end{aligned}
$$

We then weight the particles according to algorithm 4.2, which is straightforward as the marginal posterior is simply:

$$
p\left(\beta | \, y\right) = \prod_{i=1}^{n} \left[\frac{\mathbb{I}_0(y_i) + e^{\beta \cdot x_i} \mathbb{I}_1(y_i)}{1 + e^{\beta \cdot x_i}}\right] \mathcal{N}\left(\beta \, | b, v\right).
$$

It is well known that standard optimisation techniques can perform well for logistic regression as the marginal posterior can be evaluated exactly. [114] considers eight such techniques and advocates the use of conjugate gradient or quasi-Newton methods. Our attempt to compare our algorithm with the conjugate gradient approach which [114] suggests was frustrated by its poor convergence if initialised outside a neighbourhood of a good solution.

Real Data. We consider the data used in section 3.1 of [21], which corresponds to 200 eight dimensional vectors describing various characteristics related, it is supposed, to the probability that a female spouse will form part of the work force.

Table 4.7 summarises the results obtained by the SMC algorithm with various numbers of intermediate distributions and particles, always with a final value of $\gamma_T = 10$; and table 4.8 the results obtained by applying the SAME algorithm with various final temperatures. Figure 4.2 compares their performance at various computational costs. Again, a linear annealing schedule was employed.

Discussion. It is interesting to note that those cases in which too few particles/intermediate distributions were used to allow an adequate characterisation of the density, the median likelihoods lie very close to the best answer found in any circumstances. In these cases the failure mode is clearly the trapping of the entire particle set in a sub-optimal local mode, as one would expect in an annealing type algorithm with too fast an annealing schedule.

N	**T**	$\chi/1,000$	**Mean**	**Std. Dev.**	**Min**	**Median**	**Max**
500	50	137.5	-116.69	22.40	-265.26	-112.29	-112.29
1000	50	275.0	-113.27	4.32	-138.02	-112.29	-112.29
50	100	27.5	-116.51	24.92	-286.32	-112.30	-112.29
100	100	55.0	-113.92	7.71	-161.11	-112.30	-112.29
250	100	137.5	-112.29	0.002	-112.30	-112.29	-112.29
500	100	275.0	-112.29	0.001	-112.29	-112.29	-112.29
1000	100	550.0	-112.29	0.001	-112.29	-112.29	-112.29
25	200	27.5	-112.32	0.018	-112.39	-112.32	-112.30
50	200	55.0	-112.30	0.008	-112.33	-112.30	-112.29
100	200	110.0	-112.30	0.004	-112.31	-112.30	-112.29

Table 4.7. SMC performance on the data set of [21]. These are the results from 50 independent runs of the algorithm.

T	γ_T	$\chi/1,000$	**Mean**	**Std. Dev.**	**Min**	**Median**	**Max**
10,000	10	55.0	-112.74	0.267	-113.84	-112.68	-112.38
25,000	10	137.5	-112.69	0.250	-112.65	-112.41	-112.41
50,000	10	275.0	-112.76	0.195	-113.56	-112.68	-112.42
1,000	50	27.5	-120.24	55.61	-505.60	-112.37	-112.31
1,200	50	33.0	-112.36	0.026	-112.41	-112.36	-112.31
1,500	50	41.2	-112.38	0.043	-112.48	-112.37	-112.31
2,000	50	55.0	-112.37	0.042	-112.49	-112.36	-112.31
10,000	50	275.0	-112.37	0.040	-112.48	-112.38	-112.32
1,000	100	55.0	-112.33	0.022	-112.39	-112.33	-112.30
5,000	100	275.0	-112.33	0.021	-112.39	-112.33	-112.29

Table 4.8. SAME performance on the data set of [21]. These are the results from 50 independent runs of the algorithm.

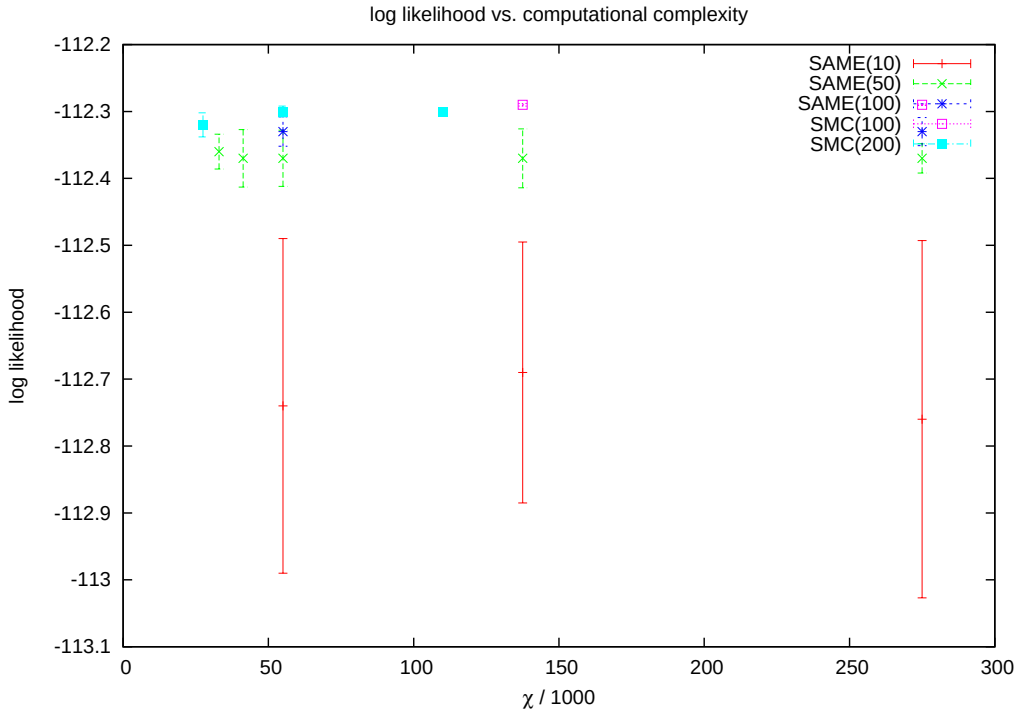


Fig. 4.2. Estimated log marginal likelihood vs. computational cost for the SMC and SAME algorithms. Error-bars denote the standard deviation, and points the mean, over 50 independent runs of each algorithm.

As is illustrated in figure 4.2, at a given computational cost, the SMC algorithm proposed above outperforms the SAME algorithm. It is not straightforward to verify that the algorithm has found the true global mode. However, the fact that the best performance found at ever increasing computational costs are the same lends some weight to the hypothesis that this is in a neighbourhood of either a global maximum, or an extremely attractive local maximum.

4.4 Summary

We have presented a collection of novel, population-based annealing algorithms for obtaining marginal parameter estimates within latent variable models. After a demsontration of the convergence of the proposed estimator under certain conditions, these algorithms were applied to three challenging problems, and performed well – on the examples presented here, they outperformed standard techniques at comparable computational costs.

5. Rare Event Simulation via SMC

"There are trivial truths, and there are great truths. The opposite of a trivial truth is plainly false. The opposite of a great truth is also true."

– Neils Bohr

A short version of this chapter was presented as [89] and an extended version is in preparation [88].

5.1 Introduction

The problem of estimating rare event probabilities has attracted a great deal of attention in recent times – see, for example, the reviews provided by [57, 69, 139]. 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:

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

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 5.3, is applicable to problems such as those considered by [40].

In section 5.4 we consider what we term *dynamic rare events*, and these correspond to the probability that a homogeneous Markov chain on a space $(E, \mathcal{F})$ 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 through $\tau = \inf\{t : X_t \in \mathcal{R} \cup \mathcal{T}\}$. We note that the recurrence of $\mathcal{R}$ is required only to makke the stopping time τ almost surely finite, and that we assume that $\mathcal{R} \cap \mathcal{T} = \emptyset$. This corresponds to the classes of problems considered by Repetititive Simulation Trials After Reaching

Thresholds (RESTART) (see [57] for a review), multi-level splitting [69] and the approaches of [17, 18].

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 [75]. The first of these distributions corresponds to the law of the Markov chain (up to a stopping time in the dynamic case, and a deterministic time in the static case) and subsequent distributions are distorted according to a sequence of potentials which ultimately cause 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 from which it would otherwise be extremely difficult to sample.

This approach dramatically ameliorates the sample diversity relative to that of the samples obtained by methods which iteratively extend the path and apply importance resampling, which inevitably leads to degeneracy at the beginning of the path [40]. Furthermore, as noted by [4], 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.

5.2 Classical and Recent Approaches

We begin with a survey of elementary techniques which can be applied to the rare event estimation problem, and attempt to illustrate why these techniques are not universally suitable, or require a large amount of application specific adjustment to obtain good results.

5.2.1 General

Initially, we consider the two obvious approaches to rare event estimation by Monte Carlo methods, highlighting the deficiencies of these approaches.

Crude Monte Carlo. The crude Monte Carlo approach can be used to estimate the probability of an event simply by sampling many evolutions of the system and using as an estimator the probability of that event under the empirical measure induced by those samples. That is, given some measure, π, sample a large collection of random variables, $\{X_i\}_{i=1}^N$ from that measure and the probability of an event $\mathcal{T}$ is simply:

$$\hat{p}^{MC}(\mathcal{T}) \triangleq \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{\mathcal{T}}(X_i).$$

Unfortunately, although the expected number of samples hitting the rare set is clearly $N\pi(\mathcal{T})$, the variance of the number is $N\pi(\mathcal{T})(1-\pi(\mathcal{T}))$ (which can be seen by considering the number of times that the rare set is reached as a binomial random variable with success probability $\pi(\mathcal{T})$) and the ratio of the standard deviation to the mean is consequently $\sqrt{\frac{1-\pi(\mathcal{T})}{N\pi(\mathcal{T})}}$ which for extremely rare events is approximately $\sqrt{\frac{1}{N\pi(\mathcal{T})}}$ which can be extremely large, even for large N. With $1/\pi(\mathcal{T})$ samples, the standard deviation of the estimator is equal to its mean and for many rare events of interest, even this modest requirement would necessitate the use of a billion or more samples. This is clearly not a practical approach.

Importance Sampling. It is theoretically possible to make use of any distribution, μ, with respect to which π is absolutely continuous to obtain estimates whilst increasing the number of occurrences of the rare events using the usual importance sampling identity, given N samples from μ, $\{X_i\}_{i=1}^N$ we have the estimator

$$\hat{p}^{IS}(\mathcal{T}) \triangleq \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X_i)\mathbb{I}_{\mathcal{T}}(X_i).$$

It is well known (see, for example, [130]) that the optimal importance distribution, μ, is given by:

$$\mu(dx) = \frac{\pi(dx)\mathbb{I}_{\mathcal{T}}(x)}{\pi(\mathbb{I}_{\mathcal{T}})},$$

and as the function whose expectation is being calculated is positive, samples from this distribution would yield an unbiased estimator of zero variance (see, for example [48]). As usual, it is impossible to use the optimal importance distribution as its normalising constant corresponds to precisely the quantity which is to be estimated.

Developing good importance distributions which provide low variance estimators can be a difficult and application-specific problem, which usually involves either an analysis of the physical or mathematical system underlying the problem, or a more formal approach, such as a large deviations analysis.

The Cross-Entropy Method. A closely related approached, which has been termed the *cross-entropy* method is essentially a scheme for adaptively constructing a sequence of distributions which provide increasingly good importance sampling proposal distributions for the estimation of the rare event of interest. The approach was first proposed in [137] and more recent tutorials provide a clear explanation [33, 138]. It applies to rare events which may be described as the level set of a suitable potential function[1].

[1] It has recently been proposed that the method can be extended to a number of settings, including several which do not involve rare event simulation but this is beyond the scope of this thesis.

The approach which is taken is, in brief, as follows. Given the rare event of interest, a parametric family of importance sampling distributions if first proposed. It is important that this family is sufficiently flexible that there exists a suitable set of parameter values which will provide good importance distributions for the rare event of interest. Starting with some arbitrary parameter value, one then samples a collection of points from some initial distribution and then selects only those samples which exceed the $1-\rho$ quantile of the potential function under the empirical distribution induced by the sample (i.e. take the proportion ρ of the samples which have the highest values of the potential function). Using the selected particles, one finds the parameter value which minimises the cross-entropy between the optimal importance function for estimating the probability of exceeding the level set of the potential function implied by the $1-\rho$ quantile found above, and a sampling distribution within the parametric family which was selected previously. This is done by using the importance sampling estimate of the integral provided by the empirical measure associated with the selected sample points. This procedure is carried out iteratively until the $1-\rho$ quantile which is found either ceases to increase of reaches that associated with the rare event of interest.

The choice of cross-entropy as the distance measure is a pragmatic one motivated by computational considerations. In practice, one wishes to minimise the estimator variance but this operation is not computationally tractable and there are a broad range of problems in which the cross-entropy can be minimised analytically.

There has been much interest in this method, which is able to produce good solutions to moderately complex problems with a reasonable computational cost.

5.2.2 Static Rare Event Simulation

In this section we briefly summarise some rare event simulation techniques from the literature intended to handle the static case described above.

Genealogical Interacting Particle Systems. A method is proposed in [40] to make use of interacting particle systems with a Feynman-Kac interpretation to obtain samples which are biased towards the rare set of interest. This approach is essentially a technique for performing importance sampling in an almost automatic fashion.

The approach consists of defining a potential function, $V : E \to \mathbb{R}$, on the state space of a homogeneous Markov chain, which describes the rare event in the sense that either rare trajectories tend to have large values of this function at *every* step in the Markov chain (which in the case of rare events characterised by the sum of light tailed distributions being large, for example, seems very reasonable), or that

the increase in this potential function from one state in the chain to the next tends to be large. Feynman-Kac potentials are then defined which correspond to either $G_n^{(\beta)}(x_{0:n}) = \exp(\beta V(x_n))$ or $G_n^{(\alpha)}(x_{0:n}) = \exp\left(\alpha\left[V(x_n) - V(x_{n-1})\right]\right)$, where α and β are tuning parameters which determine how rare the generated trajectories are likely to become.

Trajectories of the system are then simulated via algorithm 5.1 – which deals with the second type of Feynman-Kac potential; that which is recommended in the original paper, and that which is used for comparative purposes in later sections. We remark that the presentation here follows that in the original paper, including the simulation of an additional set of random variables $\{X_{-1}^{(i)}\}_{i=1}^N$. This is done for mathematical convenience in the analysis of the algorithm as it allows the use of the same recursion and potential function at every step in the algorithm; in practice, of course, there would be no advantage in actually simulating this set of random variables and making use of them could only increase the Monte Carlo variance of the estimator. The simulations below do not employ such a set of random variables. Similarly, the selection step as described corresponds to the application of multinomial resampling – which simplifies analysis of the algorithm but increases the Monte Carlo variance; the use of stratified resampling is always preferred in practice.

This seems like a sensible approach, and one which is likely to perform well in a number of situations. In section 5.3.3 we detail the principal differences between this technique and the one which we propose, algorithm 5.3. We remark that, unlike algorithm 5.3, this approach is only suitable for estimating the probability that the *final state* of the Markov chain lies in a particular rare set. Whilst it might be possible to adapt this approach to estimating rare event probabilities which depend upon the full path in some instances, it seems implausible that good performance could be obtained in general.

5.2.3 Dynamic Rare Event Simulation

Several techniques have recently been developed specifically for simulating dynamic rare events, we summarise these approaches here.

Multi-Level Splitting and RESTART. Multi-level splitting and RESTART refer to the same class of techniques – which were originally proposed in the 1950s and subsequently rediscovered on a number of occasions; see [69] and references therein. The *hybrid subset method* which has been recently proposed in the field of mechanical engineering [22, 23] appears to be another rediscovery of this approach. A number of techniques which are being developed in the physics and chemistry literature (see, [1] for an example of three) are similarly, based upon

Algorithm 5.1 An interacting particle system for static rare event estimation.

Initialise the particle ensemble:

for $i = 1$ to N **do**

Sample $X_{-1}^{(i)} \sim \eta_0$ and $\hat{X}_0^{(i)} \sim \eta_0$.

end for

for $p = 0$ to P **do**

Estimate the normalising constant of the importance distribution:

$$\hat{Z}_p^N = \frac{1}{N} \sum_{i=1}^{N} \exp\left(\alpha \left[V(\hat{X}_p^{(i)}) - V(X_{p-1}^{(i)})\right]\right)$$

Apply a selection step (termed resampling in the SMC literature) according to the potential.

for $i = 1$ to N **do**

$$X_p^{(i)} \sim \frac{1}{N\hat{Z}_p^N} \sum_{j=1}^{N} \exp\left(\alpha \left[V(\hat{X}_p^{(j)}) - V(X_{p-1}^{(j)}\right]\right) \delta_{\hat{X}_p^{(j)}}(\cdot)$$

end for

Apply a mutation step. **for** $i = 1$ to N **do**

$$\hat{X}_{p+1}^{(i)} \sim M_{p+1}(X_p^{(i)}, \cdot)$$

end for

end for

The rare event probability can be estimated as:

$$\hat{P}^{IPS}(X_P \in \mathcal{T}) = \prod_{p=0}^{P-1} \hat{Z}_p^N \times \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\mathcal{T}}\left(\hat{X}_P\right)$$

exactly the techniques described here. These examples of unnecessarily duplicated effort perhaps show the importance of communication between methodologists and applications specialists!

The principle behind *splitting* approaches to dynamic rare event simulation is to produce a decreasing sequence of sets, and simulate a large number of particle trajectories under the dynamic of the system. Whenever one of these particles reaches a set which is "rarer" than any it has entered before, it is duplicated and the weight associated with each "child" particle is set to half of that of its parent. Particles are killed whenever they hit the recurrent set. This approach is reviewed in [69, 57], along with many variations.

Variants of these approaches are numerous, but all are based upon a decomposition of the rare event into a product of conditional events. Given a rare set $\mathcal{T}$ and a recurrent set $\mathcal{R}$, a decreasing sequence of sets, $\mathcal{T}_1 \supset \mathcal{T}_2 \supset \cdots \supset \mathcal{T}_T = \mathcal{T}$ is constructed, along with the associated stopping times $\tau_i = \inf\{t : X_t \in \mathcal{T}_i \cup \mathcal{R}\}$ (which are almost surely finite by the recurrence of $\mathcal{R}$) and the following decomposition of the rare event probability is used:

$$\mathbb{P}\left(X_{\tau_t} \in \mathcal{T}_t\right) = \prod_{s=1}^{t} \mathbb{P}\left(X_{\tau_s} \in \mathcal{T}_s | X_{\tau_{s-1}} \in \mathcal{T}_{s-1}\right)$$

with the conventions that $\tau_0 = 0$ and $\mathcal{T}_0 = E$. The algorithm estimates each of these probabilities via a mean field approximation, propagating a number of paths until they hit each set, and then resampling by replicating those paths which did not return to the recurrent set whilst simultaneously estimating the conditional probability as the fraction of paths which survived this stage. i.e. commencing from a set of N paths which hit $\mathcal{T}_{s-1}$, $\{X^{(i)}_{0:\tau^{(i)}_{s-1}}\}_{i=1}^N$, each path is extended by sampling from the law of the Markov chain until it hits either $\mathcal{T}_s$ or $\mathcal{R}$. It is then possible to estimate:

$$\mathbb{P}\left(X_{\tau_s} \in \mathcal{T}_s | X_{\tau_{s-1}} \in \mathcal{T}_{s-1}\right) \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{\mathcal{T}_s}\left(X^{(i)}_{\tau_s}\right)$$

and sample (with replacement) N paths from those which successfully hit $\mathcal{T}_s$ for use in the next iteration[2].

Adaptive Multi-level Splitting (AMS). A recent variation of the splitting type method was proposed by [18]. In order to avoid one of the principle difficulties with this approach, that of selecting the sequence of nested sets, an adaptive approach was proposed.

Only the one dimensional, continuous time case with continuous sample paths in a state space $E = \mathbb{R}$ was considered. The approach was to simulate a collection of paths until they hit $\mathcal{R}$, determine the point closest to the rare set reached by each path. Having done so, find the largest of these distances, $d^\star$, obtained by at least some number k of *successful* paths, and estimate the probability of hitting that set before the recurrent set as S/N. Starting from $d^\star$, sample N paths until they hit $\mathcal{R}$ and repeat this process until $\mathcal{T}$ is reached.

Although it was suggested by [18] that the generalisation to more complex problems should be straightforward, it does not seem clear that this is the case for discrete time processes or continuous time processes with potentially discontinuous sample paths even in one dimension – or indeed for continuous time processes which cannot be simulated exactly: if it is possible to reach a point closer to the rare set than some state $x^\star$ without passing through that state then the conditioning argument does not appear to hold, and this seems particularly difficult to deal with in multi-dimensional state spaces. It seems likely that these problems will be overcome to a greater or lesser extent in the fullness of time and that this approach will become a valuable tool in the estimation of rare event probabilities and perhaps in other areas of adaptive importance sampling.

[2] This is what [57] describes as *fixed effort* RESTART; in the *fixed splitting* form, the number of paths is random with a fixed number of replicates of each successful path being propagated forward.

Algorithm 5.2 Adaptive multi-level splitting for dynamic rare event estimation

1: Initialise the particle set by sampling from the law of the Markov chain until stopping time χ, at which they hit the recurrent set $\mathcal{R}$.
2: Set the iteration count, $T = 0$.
3: **for** $i = 1$ to N **do**
4: Sample $X^{(i)}_{0:\chi^{(i)}} \sim \mathbb{P}_{\eta_0}(\cdot)$ where $\chi^{(i)} = \inf\left\{t : X^{(i)}_t \in \mathcal{R}\right\}$
5: Calculate $S^{(i)} = \sup\left\{V(X^{(i)}_t) : t \leq \chi^{(i)}\right\}$, the best state reached by each path.
6: **end for**
7: Let K be the set of indices of the k particles with the largest associated values of $S^{(i)}$:

$$\text{i.e. } \forall i \in K, j \notin K : S^{(i)} \geq S^{(j)}.$$

8: Let $q = \inf_{i \in K} S^{(i)}$.
9: **if** $V(q) \geq \hat{V}$ **then**
10: Estimate the rare event probability using:

$$\hat{p}^{AMS} = r\left(\frac{k}{N}\right)^T$$

where r denotes the fraction of the current set of particles which hits the rare set.
11: **else**
12: **for** $i = 1$ to N **do**
13: **if** $i \notin K$ **then**
14: Let $X^{(i)}_0 = V^{-1}(q)$ and sample the rest of the path according to the law of the Markov chain until the recurrent set is hit.
15: **else**
16: Let $s^{(i)} = \inf\{t : X^{(i)}_t = S^{(i)}\}$. Let $X^{(i)} = X^{(i)}_{s^{(i)}:\chi^{(i)}}$, $\chi^{(i)} = \chi^{(i)} - s^{(i)}$.
17: **end if**
18: Calculate $S^{(i)} = \sup\left\{V(X^{(i)}_t) : t \leq \chi^{(i)}\right\}$, the best state reached by each path.
19: **end for**
20: Increment T and **goto** step 7.
21: **end if**

The one dimensional algorithm applied to a general space in which a function $V : E \to \mathbb{R}$ increases towards the rare set and has the property that $V(x) \geq \hat{V} \iff x \in \mathcal{T}$ is shown in algorithm 5.2.

Genealogical Interacting Particle Systems. An interacting particle system interpretation of a Feynman-Kac flow has been proposed as a generic solution to this problem [17]. In some sense, this approach is essentially the same as some variants of the multi-level splitting algorithm provided with a rigorous theoretical interpretation. Casting the approach into a Feynman-Kac form allows a wealth of recently obtained theoretical results [34] to be applied.

Again, a decreasing sequence of sets which concentrates itself upon the rare set of interest is employed. In this case a multi-level Feynman-Kac interpretation is required with one level corresponding to a skeleton of stopping times and a lower level corresponding to the path from one set to the next. The algorithm amounts to sampling a particle approximation to a Feynman-Kac flow in which the state at time t corresponds to a realisation of a Markov chain which begins in $\mathcal{T}_t$ and finishes when it first enters either $\mathcal{T}_{t+1}$ or $\mathcal{R}$. A zero-one potential function is applied at each iteration to kill those paths which return to $\mathcal{R}$. The normalising constant of the flow when the particle set reaches $\mathcal{T}$ corresponds to the rare event probability of interest.

5.3 Static Rare Event Estimation

In order to solve the problem of computing, for deterministic P, $\mathbb{P}_{\eta_0}(X_{0:P} \in \mathcal{T})$, we propose employing a SMC sampling approach. We remark that this class of problems includes those cases in which we wish to determine whether some property of a collection of *iid* variables fulfill a condition which is rarely satisfied, as well as a wide range of cases involving properties of Markov chains.

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 [34]; stability has been established in a slightly different setting by [82] and work is ongoing to extend those results.

By operating directly upon the path space, we gain 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, $V : E_{0:P} \rightarrow \mathbb{R}$, which has the properties that:

$$\begin{aligned} V: \quad \mathcal{T} &\rightarrow [\hat{V}, \infty), \\ V: \quad E_{0:P} \setminus \mathcal{T} &\rightarrow (-\infty, \hat{V}). \end{aligned}$$

In this case, it makes sense to consider a sequence of distributions defined by a potential function which is proportional to their Radon-Nikodým 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 $\alpha(\theta) : [0,1] \rightarrow \mathbb{R}_+$ 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, we make use of the identity:

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

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'.

5.3.1 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 [58]. 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 case, we can describe our sequence of distributions in precisely this form via a discrete sequence of intermediate distributions parameterised by a sequence of values of θ:

$$\begin{aligned}
\frac{\mathrm{d}\log g_\theta}{\mathrm{d}\theta}(x) &= \frac{(V(x)-\hat{V})}{\exp(\alpha(\theta)(V(x)-\hat{V}))+1}\frac{\mathrm{d}\alpha}{\mathrm{d}\theta} \\
\Rightarrow \log\left(\frac{Z_{t/T}}{Z_0}\right) &= \int_0^{t/T} \mathbb{E}_\theta\left[\frac{(V(\cdot)-\hat{V})}{\exp(\alpha(\theta)(V(\cdot)-\hat{V}))+1}\right]\frac{\mathrm{d}\alpha}{\mathrm{d}\theta}d\theta \\
&= \int_0^{\alpha(t/T)} \mathbb{E}_{\frac{\alpha'}{\alpha(1)}}\left[\frac{(V(\cdot)-\hat{V})}{\exp(\alpha'(V(\cdot)-\hat{V}))+1}\right]d\alpha',
\end{aligned}$$

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 a series of values of $\theta \in [0,1]$ 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 subsequently use an importance sampling estimator to obtain the probability of hitting the rare set.

Some theoretical justification for the intuition that the path sampling approach is likely to lead to lower variance estimators is provided in appendix A.1 in which two approaches are compared for a very particular case. More general results are being investigated. Similarly, appendix A.2 illustrates the improvement in variance which can be obtained by using this approach, rather than simple Monte Carlo, again in a particular case.

Algorithm 5.3 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 $W_0^{(i)} \propto \frac{\pi_0(Y_0^{(i)})}{\nu(Y_0^{(i)})}$ 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 [13] 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 $W_t^{(i)} \propto \hat{W}_{t-1}^{(i)} \frac{\pi_t(Y_t^{(i)}) L_{t-1}(Y_t^{(i)}, \tilde{Y}_{t-1}^{(i)})}{\pi_{t-1}(\tilde{Y}_{t-1}^{(i)}) K_t(\tilde{Y}_{t-1}^{(i)}, Y_t^{(i)})}$ with $\sum_{j=1}^N W_t^{(j)} = 1$.

end for

end for

Approximate the path sampling identity to estimate the normalising constant:

$$\begin{aligned} \hat{Z}_1 &= \frac{1}{2} \exp\left[\sum_{t=1}^{T} (\alpha(t/T) - \alpha((t-1)/T)) \frac{\hat{E}_{t-1} + \hat{E}_t}{2}\right] \\ \hat{E}_t &= \frac{\sum_{j=1}^N W_t^{(j)} \frac{V\left(Y_t^{(j)}\right) - \hat{V}}{1 + \exp\left(\alpha_t\left(V\left(Y_t^{(j)}\right) - \hat{V}\right)\right)}}{\sum_{j=1}^N W_t^{(j)}} \end{aligned}$$

Estimate the rare event probability using importance sampling:

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

5.3.2 Density Estimation

Algorithm 5.3 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, one is actually interested in the estimation of the *pdf* itself. We remark that the algorithm presented here can be easily adapted to this task, and propose a number of possible approaches to doing so. The one which is recommended, and which is likely to perform well in general (at least when the *pdf* is that of a one dimensional variable) is described first. We then go on to describe a number of alternatives which might be preferred in particular circumstances.

Historical process estimation. This approach makes use of all of the samples obtained by the algorithm. The following explanation describes an approach to the estimation of a one dimensional *pdf*. 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.

The examples presented below allow us to provide a number of definite examples. We note that in the *pdf* estimation cases described here, the function V itself (that is the terminal position of the Markov chain in the Gaussian tail case and the length of the polarisation vector in the PMD case) was the function of interest and this set was taken to be the interval between the 1st decile and the 9th decile of the particle set at the time sorted appropriately i.e. If $\bar{Y}_t$ is the set of particles at time t sorted such that $V(Y_t^{(i)}) \leq V(Y_t^{(i+1)})$ then $\bar{R}_t = R \cap \left[V(\bar{Y}_{\lfloor 0.1N \rfloor}), V(\bar{Y}_{\lceil 0.9N \rceil})\right]$.

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:

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

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

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

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.

Importance Sampling. The methodology described here provides, for a given potential function threshold value, a collection of weighted samples from a distribution function close to the product of an indicator function on the associated level set of the indicator function and the original distribution function, we are able to make use of those samples as samples from an importance distribution and by re-weighting them appropriately. Hence, we may use them to estimate a *pdf* with reasonable accuracy in a region with potential bounded below by the threshold value.

Whilst this approach is intuitively appealing due to its simplicity of implementation, it seems unlikely that it will ever be an efficient way to perform density estimation.

Dedicated SMC Sampler Density Estimation. In situations in which one is generally interested in estimating the *pdf* of some complicated random variable over a particular set of interest, say $\mathcal{S}$, a set of pointwise evaluations of the *pdf* distributed approximately uniformly over that set would seem to be the ideal

tool[3]. In those cases in which it is not possible to obtain such by direct means, one method which could be employed, and should be applicable fairly generally is to make use of the SMC samplers framework to progress via a smooth sequence of intermediate distributions from the law of the random variable to a distribution which is uniform over $\mathcal{S}$ and elsewhere zero. This would be an efficient method for density estimation within the SMC samplers framework.

Smooth Particle Filter Approaches. In the case where the state space of the Markov Chain is $\mathbb{R}$, and no Metropolis moves are employed, we could in principle perform one simulation for an arbitrary threshold, replacing the resampling step with a smoothed form just as in [124]. By observing the changing rare event probability of exceeding the threshold as a function of threshold we can obtain the *cdf* of the distribution.

5.3.3 Comparison with the IPS Approach

The method presented in [40] performed well in the applications described therein. Naïvely, one would expect it to require less computational time than the method described above which involves resampling on the path space of the Markov chain of interest and requires the introduction of intermediate distributions. However, there are other factors which suggest that the method described here can perform better in various senses in many circumstances. The following are all factors which should be considered:

– Although the SMC samplers approach described above involves more complex sampling and a number of intermediate distributions, it can obtain good results with many fewer particles.
– As observed in [4], extreme values in the sums of random variables with light-tailed distributions are predominantly the result of a large number of moderately extreme values within the terms of that sum, but in the heavy-tailed case it is much more likely for a single extreme value within the sum to lead to the extreme behaviour overall. In principle the framework proposed above should be able to operate within both regimes. One might expect the method described within [40] to have difficulty, in practice, as it applies a weighting at every time step which depends upon the present terminal and penultimate states of the Markov chain. Consequently, after a few time steps only those chains which have moved towards a reasonably large value of the potential function will remain, whilst

[3] It is clearly possible to envisage situations in which the set of interest and the law of the random variable are sufficiently complex that it is not straightforward to simply establish a grid and evaluate the analytically known *pdf* over that set, and we shall assume that these are the situations in which Monte Carlo density estimates are of interest.

the true distribution is (in the homogeneous case, at least) equally likely to have a single extreme value at any point in the chain.

– The algorithm described above works directly upon the path space of the Markov chain of interest and should lead to a good description of the distribution over this space. However, the algorithm of [40] works on the state space of the Markov chain and one would expect the frequent resampling involved to lead to degeneracy in the path space (see figure 5.2 for an illustration of this effect).
– The value of the constant α within the method of [40] must be large enough to push a substantial part of the particle set to the rare set, but not so large that it eliminates the particles which hit the less rare parts of that set. In contrast, the construction of our potential function is such that the terminal value of α must be large enough, but no problems should occur if it is much larger than the minimum value needed to push part of the particle set into the rare set.
– In the case of [40] the number of particles required for reasonable estimation is related to the rarity of the event to be estimated as it requires that the true one-step transition kernel produces moves at each individual time-step which are sufficiently extreme that a sum of the most extreme values found at each individual time-step hits the rare set. This essentially prevents the algorithm from being used with extremely rare events as a prohibitive number of particles would be required.

Of course, the SMC sampler approach also suffers from certain limitations and the following factors should also be considered:

– The independence of the potential function and the threshold has the advantage that the IPS approach can provide estimates of the probability of lying in the rare set at each point in the evolution of the Markov chain, although this will only be accurate within the region in which a reasonable number of particles have hit the rare set.
– Inspection of the final particle paths provides a straightforward way to verify that a reasonable number of particles have hit the rare set and that the rare set is reasonably well covered by particles in the case of the IPS approach, whilst this cannot be determined straightforwardly in the SMC case. However, this diagnostic is of limited value as it is not straightforward to verify that all *paths* to the rare set are well represented in this way without detailed knowledge of the evolution of the chain.
– Although the dependence upon the rarity of the event is more subtle in the SMC case, there is still an increasing computational cost associated with estimating progressively rare events. As events become rarer, it will become increasingly difficult to move smoothly from the initial distribution to the target, and this

will necessitate the introduction of an increasing number of intermediate distributions.

5.3.4 Examples

We now provide a number of examples of algorithm 5.3, together with a comparison to the IPS algorithm of [40] and crude Monte Carlo as appropriate. We begin with a simple Gaussian random walk which allows us to demonstrate the accuracy of the approach, before moving on to the optical problem studied by [40]. Finally, we demonstrate a slightly different use of the algorithm to obtain approximate solutions to counting problems.

A Toy Example: The 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 [40]. 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:

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

The function $V(x_{0:P}) \triangleq x_P$ corresponds to a canonical coordinate operator and the rare set $\mathcal{T} \triangleq E^P \times [\hat{V}, \infty)$ 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

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

This leads to the following optimal auxiliary kernel (2.25):

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

The incremental importance weight is consequently:

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

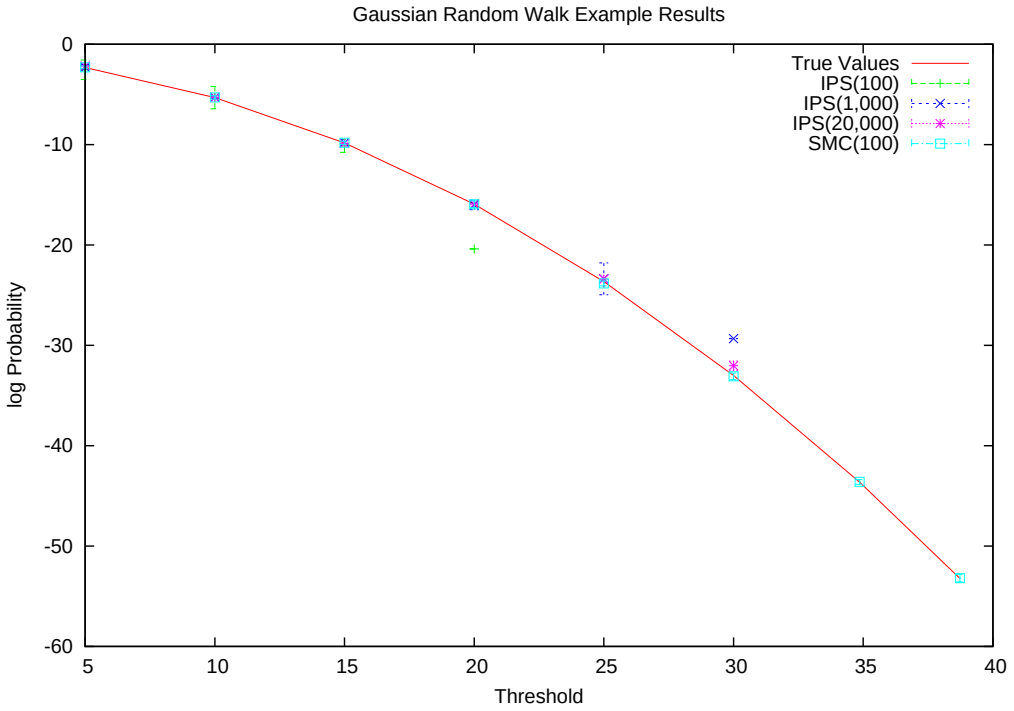


Fig. 5.1. The results shown in table 5.1 and 5.2. Numbers in brackets denote the size of the particle set. The error-bars indicate the estimator standard deviation.

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
$9\sqrt{15} \approx 34.9$	-43.63	-43.61	0.133	12	3600
$10\sqrt{15} \approx 38.7$	-53.23	-53.20	0.142	11.5	4000

Table 5.1. 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.

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 [123]. 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).

Threshold, $\hat{V}$	True	IPS(1,000)		IPS(20,000)	
		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 5.2. Results of 10 runs of the algorithm of [40] 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.

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 5.1. For the purposes of comparison, we also implemented the algorithm of [40] and show its performance in table 5.2. The performance of both algorithms is illustrated by figure 5.1, and figure 5.2 illustrates the diversity of the samples obtained by the different approaches.

Intermediate Distributions. One obvious requirement when using this method to estimate rare event probabilities is the specification of the number and spacing of intermediate distributions and what final distribution to use. In the framework provided here, that amounts to deciding on a sequence of values of $\alpha(\theta_t)$, with the terminal value of this sequence providing the terminal distribution. This is done indirectly by selecting a functional form for α and a sequence of values of θ_t between 0 and 1. Without loss of generality, we may assume that $\theta_t = t/T$ and transfer all of the freedom to the function α. In general this is not completely trivial, but should be no more arduous than the selection of an annealing schedule in SA or the choice of sequences of sets in multi-level splitting and related algorithms.

In the case of the homogeneous Markov chain with Gaussian increments that we have been considering in this section, some straightforward analysis suggests itself – we note that this is not intended to provide any more than some loose guidelines motivated by those properties of the sequence of distributions that seem desirable. If we first consider the case in which a linear sequence of values is used, $\alpha(\theta_t) = kt$ for some constant k, we are essentially left with two questions: how many distributions are required and what should be the value of k.

In this case, as in many others, it is straightforward to check the estimates generated against existing known results (by choosing an event which is much less rare than that of interest and estimating the probability by crude Monte Carlo, if necessary). This allows the verification of the results produced with particular values of k and numbers of intermediate distributions. Having done this, all that

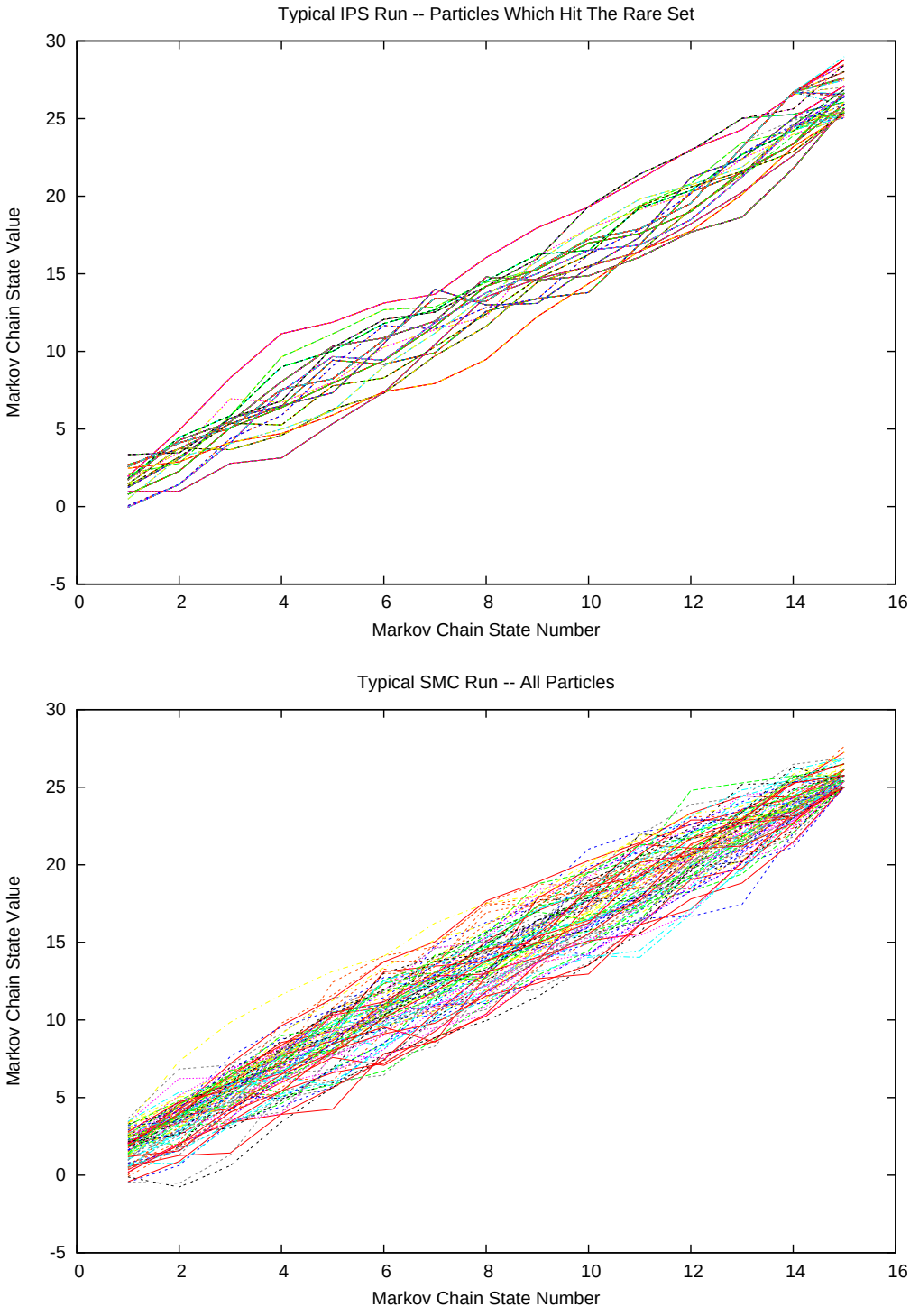


Fig. 5.2. A pair of typical plots of the path-space values of the terminal estimates of the two algorithms applied to the Gaussian tail probability estimation problem. Note that the IPS plot shows the 187 particles (from 20,000) which exceeded 25 in element 15, and suffers from substantial degeneracy in the earlier states, whilst the SMC samplers plot shows all of the 100 particles which were used.

remains is determining how the number of distributions and the spacing between them should depend upon the threshold value being considered. However, caution must always be used when attempting to project results from simple situations to more complex ones, and there is some danger inherent in this approach: it is always possible that the behaviour in more complex problems will be qualitatively different to that in simple, easy to analyse examples.

Number of Distributions. When considering a number of different runs of the algorithm with different threshold values, it would be preferable to arrange for roughly the same proportion of the particle set to be pushed into the rare regime in each case. Essentially, our aim is to keep constant the ratio of the product of the true density and the potential function at some distance ϵ past the threshold value, to that at the origin, for different threshold values, $\hat{V}$:

$$\frac{1+e^{-\alpha(1)(0-\hat{V})}}{1+e^{-\alpha(1)\epsilon}}\exp(-(\hat{V}+\epsilon\sigma)^2) = -k'$$

for some constant k'. Upon taking logarithms and neglecting lower order terms, this leads fairly rapidly to the conclusion that $\alpha(1) \propto \hat{V}$.

Spacing of Distributions. Consider the rate of change of the potential function with α, evaluated at $X_P = 0$ for simplicity. If $\hat{V}$ denotes the threshold value, then we have:

$$\left.\frac{dG}{dt}\right|_{X_P=0} = -k\hat{V}\frac{e^{\alpha(\theta_t)\hat{V}}}{\left(1+e^{\alpha(\theta_t)\hat{V}}\right)^2}$$

and when near the origin, we would assume that $\alpha(\theta_t)\hat{V}$ is still small, suggesting that we want $k \propto \hat{V}^{-1}$ if we are going to have the same rate of motion of the distribution, in some sense, for all thresholds.

This leads to $\theta_t \propto \hat{V}^{-1}$ and $\alpha(1) \propto \hat{V}$ which would suggest that the number of distributions required scales as $\hat{V}^2$. However, in practice this does not appear to be necessary and these considerations are, to say the least, extremely rough guidelines.

A Physical Example: 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 [9, 40, 56] who obtained good results with their methods. The advantage of the SMC 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, although it requires more knowledge on the part of the user than the method of [40].

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 $[-1, 1]$, $s_n = \text{sgn}(\theta_n)$ 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 $X_n = \{r_n, \phi_n, c_n = \cos(\theta_n), s_n = \text{sgn}(\theta_n)\}$. We assume that $r_0 = (0, 0, 0)$ (This corresponds to the simulation performed in [9] – the "squares" in figure 2 therein) and then the finite dimensional distributions are given by:

$$\mathbb{P} \cdot X_{0:n}^{-1}(r_{0:n}, \theta_{1:n}, c_{1:n}, s_{1:n}) = \delta_{(0,0,0)}(r_0) \prod_{i=1}^{n} \frac{1}{8\pi} \left[\delta_{R(\theta_i, s_i \cos^{-1}(c_i)) r_{i-1} + \Omega(\theta_i)/2}(r_i)\right]$$

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 this 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 an 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 rapidly 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 [40] are illustrated in figure 5.3.

We now highlight the importance of carefully designing proposal kernels. It is important to develop approaches which improve the distribution of particles, without dramatically increasing the computational cost. Otherwise, one may as well use a simple update move followed by a MCMC move of the correct invariant distribution with a much large particle set.

There is a temptation to propose a state adjustment move which selects an index from the discrete uniform distribution and proposes a grid based set of

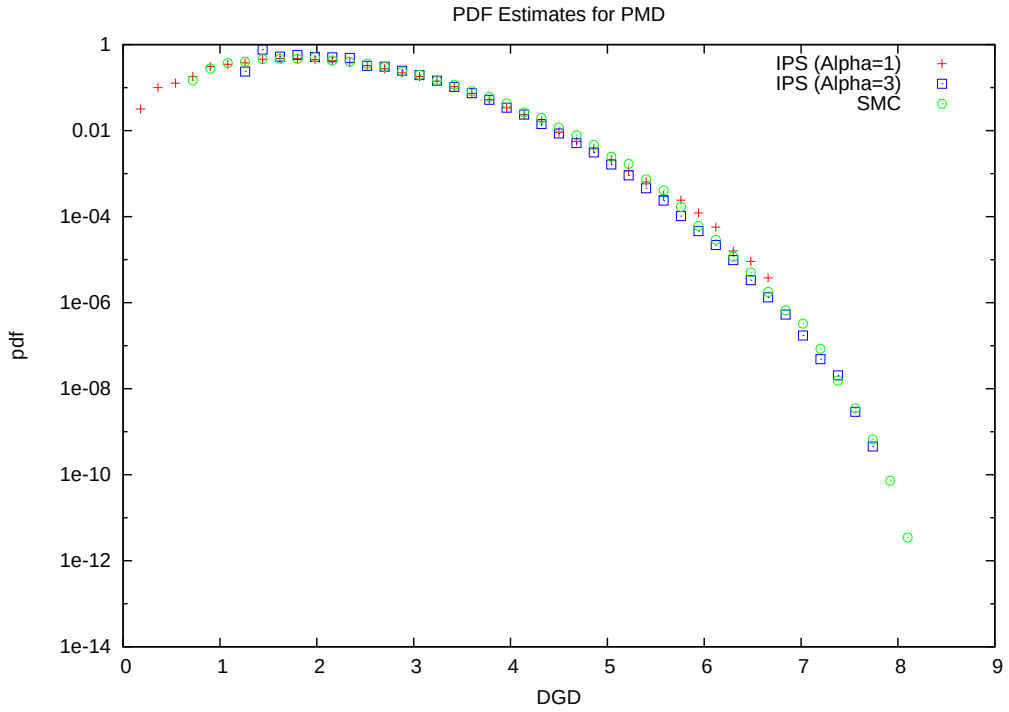


Fig. 5.3. The *pdf* estimates obtained with the SMC samplers methodology and the IPS approach of [40]. The SMC Sampler used $N = 100, T = 8,000, c = 250$. The IPS used $N = 20,000, \alpha = 1$ and $N = 20,000, \alpha = 3$, respectively. The example contains sixteen segments, each of length 0.5 in random orientations.

changes to the value of $\{\theta, c, s\}$ before updating the value of r at all subsequent positions in the chain. If we implement such a move together with the optimal reverse kernel then the calculation becomes extremely expensive.

Another is to approximate the optimal situation and make use of a mixture of distinct moves with equal probability of selection, each of which has an associated optimal reverse kernel, and then make use of the computationally preferable approximation suggested in [37]. If we do this then we obtain as a proposal kernel for p selected from the discrete random distribution over $1, \ldots, P$:

$$K_t^p(Y_{t-1}, Y_t) = \sum_{j=-S}^{S} \sum_{k=-R}^{R} w_t^p(Y_{t-1}, Y_t) \delta_{\phi_{t-1}^p + j\delta_\phi}(\phi_t^p) \delta_{c_t^p + j\delta_c}(c_{t+1}^p)$$

where the individual move weights w_t^p are chosen to be proportional to the density at time t. This leads to an incremental weight of:

$$W_t^p = \frac{\pi_t(Y_t)}{\sum_{j=-S}^{S} \sum_{k=-R}^{R} \pi_{t-1}(Y_t^p \cup \{\phi_t^p + j\delta_\theta, c_t^p + k\delta_c\}) w_t^p(Y_t^p \cup \{\phi_t^p + j\delta_\phi, c_t^p + k\delta_c\}, Y_t)}$$

However, the computational cost of making such moves is still quite substantial, largely due to the need to recompute the sequence of polarisation vectors for each set of angles under consideration, and the performance is not significantly better than that obtained with update moves and MCMC steps alone.

A Computer Science Example: Counting Problems. Counting the number of solutions to combinatorial constraint satisfaction problems is one of the classic problems of computer science – see, for example, [86]. Indeed, the sorting problem can be mapped to the problem of counting the number of items in the list to be sorted which appear before or after each of the other items in the list [96, page 75]. One such problem is determining the number of unique combinations of objects of known sizes which will fit into a knapsack of a particular capacity. 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 corresponds to the number of unique $0-1$ vectors x for which the following inequality is satisfied:

$$a \cdot x = \sum_{i=1}^{P} a_i x_i \leq b. \tag{5.1}$$

As observed by [101] the approximate solution of this problem can be cast as a rare event problem[4]. 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 2^P times that probability.

We define $V(x) := -a \cdot x$ such that inequality (5.1) can clearly be expressed[5] as $V(x) \geq -\left(b + \min_{i \neq j} |a_i - a_j|\right)$. 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 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:

[4] 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.

[5] In the interest of symmetry, it seems preferable to use a threshold on the potential function which lies between the largest possible set of items which fits in the knapsack and the smallest possible set which is too large; one approach would be to use a threshold of $\hat{b} = b + \min_{i \neq j} |a_i - a_j|$ rather than b directly.

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

with

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

This has the following optimal backward kernel (2.25) 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 weight expression:

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

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 5.3 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 [101] 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.

N	**Runs**	**Threshold:** $b + 0.5$	**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 5.3. SMC results for the counting problem of section 5.3.4 obtained using a linear schedule for α increasing from 0 to 1 over 800 steps. Estimates are of the log probabilities.

As a complement to the conclusion of the PMD example, we now show that, in this instance, simply using update and MCMC moves leads to rather poorer

performance than making use of a well designed SMC sampler proposal. We employ 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 $t-1$. We utilise an MCMC proposal after every iteration, consisting of a symmetric transition kernel in which two indices are selected randomly in the vector and all elements lying between those two points (cyclically, in that if the second vector is smaller than the first then all points outside that range are considered) are replaced with samples from a $\mathcal{B}er\,(0.5)$ distribution, and the proposal is subsequently accepted with the usual Metropolis-Hastings acceptance ratio. The results of this approach, and of using a combination of adjustment and update moves is shown in table 5.4.

N	**Runs**	**Threshold:** $b+0.5$	p_1	**Mean**	**Variance**
100	50	2.5	1.0	-12.889	0.04727
100	50	2.5	0.5	-13.039	0.09726
100	50	2.5	0.0	-13.361	0.66333
100	50	10.5	1.0	-10.214	0.03647
100	50	75.5	1.0	-2.0413	0.00501
1000	5	2.5	1.0	-12.811	0.04805
1000	5	2.5	0.5	-12.890	0.04566
1000	5	2.5	0.0	-12.806	0.24853
1000	5	10.5	1.0	-10.207	0.00652
1000	5	75.5	1.0	-2.0026	0.00027

Table 5.4. Further results, again using a linear schedule for α. p_1 represents the probability that a random scan adjustment move, rather than update move is applied at each time. Again, results correspond to log probabilities.

By way of a conclusion of this point, we note that the random scan Gibbs sampler, which was proposed above, could equally well have been employed as a deterministic scan Gibbs sampler, with different elements of the vector update at each time. This naturally simplifies the calculation of the importance weights and leads to a kernel of the form:

$$K_t(Y_{t-1}, Y_t) = \delta_{Y_{t-1}^{(1:p-1,p+1:n)}}\left(Y_t^{(1:p-1,p+1:n)}\right)\left[w_0^p\delta_0(Y_t^{(p)}) + w_1^p\delta_1(Y_t^{(p)})\right]$$

with

$$w_i^p \propto \pi_t\left(\left(Y_{t-1}^{(1:p-1)}, i, Y_{t-1}^{(p+1:n)}\right)\right)$$

and $p \triangleq (t \bmod n) + 1$. Leading to the incremental weight expression:

$$w_t(Y_{t-1}, Y_t) = \frac{\pi_t(Y_{t+1})}{\sum_{y\in\{0,1\}} w_y^p \pi_{t-1}\left(\left(Y_t^{(1:p-1)}, y, Y_t^{(p+1:n)}\right)\right)}$$

The results of employing this proposal are illustrated in table 5.5. In this case we consider using twice as many intermediate distributions, illustrate that this

can produce a substantial improvement when dealing with proposal distributions which mix less well. This provides a demonstration that using a proposal distribution which explores the space more slowly but which allows for much faster evaluation of the importance weights can lead to better performance at a given computational cost.

N	**Runs**	T	**Threshold,** $b + 0.5$	**Mean**	**Variance**
100	50	800	2.5	-12.978	0.04834
100	50	1600	2.5	-12.846	0.02258
100	50	800	10.5	-10.250	0.02661
100	50	1600	10.5	-10.173	0.01243

Table 5.5. Some results obtained using a linear schedule for α and a deterministic scan adjustment move. We show results with both 800 and 1600 intermediate distributions.

5.4 Dynamic Rare Event Estimation

Here we illustrate that it is possible to employ our approach for solving the same class of problem as the various multi-level splitting algorithms. We employ a Feynman-Kac formulation which is very different to that used by [17]: in our case the flow is entirely synthetic, whereas the evolution of the flow is fundamentally related to the dynamical structure of the chain of interest in the previously proposed algorithm.

Consider 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, for notational convenience, we assume that the support of the initial distribution does not include either the rare set nor the recurrent set – $\mathrm{supp}(\eta_0) \cap (\mathcal{R} \cup \mathcal{T}) = \emptyset$. It becomes apparent from this representation that this is actually a trans-dimensional estimation problem, as follows:

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

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 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

Algorithm 5.4 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 $W_1^{(i)} = \mathbb{I}_{\mathcal{T}_1}\left(X_{\tau_1^{(i)}}^{(i)}\right)$.
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 [73]), $\tilde{K}_{t-1}$ 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 $Z_t = \frac{1}{N}\sum_{i=1}^{N} W_t^{(i)}$.

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:

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

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

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 zero-one valued potential function makes it impossible to employ the path sampling identity of [58] 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 5.4 provides a fairly general framework for rare event probability estimation. However, in full generality, one might wish to consider situations in which the proposal kernels, $\{K_t\}$ are able to modify that part of the path which has been proposed thus far in addition to extending it. This is trivial to accomplish, and simply leads to a slightly more complex weight expression. In the interests of clarity we present only the simpler case here.

5.4.1 Example

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

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

Defining $\mathcal{R} = (-\infty, -a]$ and $\mathcal{T} = [b, \infty)$ for two integers, $-a < 0 < b$, 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 5.6 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	**S.D.**	**Mean**	**S.D.**
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 5.6. 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.

5.5 Summary

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. Examples are presented which illustrate the effectiveness of these algorithms in a number of applications. Work is ongoing in both the theoretical analysis and methodological development of these algorithms.

Particularly in the dynamic rare event estimation case, the approach which has been initiated here requires a great deal more investigation. In particular, methodological developments determining which sequence of distributions it makes sense to use in this context and how proposal distributions should be designed warrants a great deal of further work. The algorithm presented here provides little

more than a re-interpretation of multi-level splitting as a SMC sampler, but the flexibility of this sampling scheme makes it possible to apply much more general sampling strategies.

Methodologically, we are interested in determining how to select the sequence of distributions to be employed, whether this can be done adaptively (perhaps in a manner similar to [18]), 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 to determine computable bounds upon the variance and bias associated with the SMC algorithms presented above (we note that the bias may be controlled by combining results for trapezoidal integration [130] and Feynman-Kac formulae [39]) and the variance by using the techniques pioneered by [34].

Although we have not performed a theoretical analysis of the algorithms presented here, it is possible to establish a broad range of theoretical results by the application of the techniques pioneered in [34]. Noting that we can obtain a central limit theorem using these techniques, and that it will have the form presented in [37], it is clear that the asymptotic variance of the estimates produced by the algorithms presented here will be closely related to the mixing properties of the proposal kernels used to move around the space.

6. Conclusions

"And so we beat on, boats against the current, borne back ceaselessly into the past."

– F. Scott Fitzgerald, "The Great Gatsby"

This thesis was concerned with SMC methods of a non-standard character. These fall into two categories: those which use interacting particle systems to approximate non-linear flows which do not admit a Feynman-Kac representation and those in which a more usual Feynman-Kac representation is available but the intention is to perform some inference which does not resemble filtering, prediction or smoothing.

6.1 Contributions

This thesis has included contributions to some areas of the theory of SMC, together with some applications of those methods for illustrative purposes.

In chapter 3 an asymptotic analysis of the SMC approximation of the PHD filter was conducted. It has been shown that under weak assumptions the integral of all bounded test functions under the empirical measure associated with the particle approximation converges almost surely to their values under the exact filter, and, under similar conditions, a central limit theorem holds, with a variance as given in section 3.4. It should be noted that, whilst this provides some theoretical justification for the particle approximation if one accepts the validity of the PHD recursion itself, it provides no information about the accuracy of the PHD recursion or how it relates to the optimal filter on the full multiple object space.

Chapter 4 presents a novel algorithm for marginal parameter estimation using SMC techniques. After convergence results, illustrating that the algorithm converges to the maximisers of the likelihood (or posterior) under appropriate regularity conditions, applications to a number of particular cases are presented. It is shown that the algorithm outperforms a number of competing techniques, at comparable computational cost, when applied to a finite mixture model. Good performance is also observed from applications to logistic regression and stochastic volatility modelling.

In chapter 5 a framework for the estimation of rare event probabilities and approximate counting problems was proposed and illustrated with several examples. Two broad classes of rare events are considered: whether a finite section of the trajectory of a (potentially inhomogeneous) Markov chain lies within a particular set of small measure, and whether a homogeneous Markov chain hits a particular set before its first entry to some recurrent set. Algorithms for the estimation of both types of rare event probability, as well as the *pdf* are provided. In the first case, several examples are considered and comparisons to competing techniques are presented. The second category is presented together with a simple example, but work remains to be done in this area.

6.2 Future Directions

Considering direct extensions of the work presented in this thesis, a number of potentially interesting areas for further research exist.

Concerning the PHD filter, it would be interesting to generalise the methodology in a number of directions – particularly allowing higher order moments to be considered. Preliminary work in this direction has shown that it is non-trivial to do so whilst obtaining a computationally tractable algorithm.

The marginal estimation algorithm is reasonably self-contained; further theoretical results would be of interest – particularly regarding stability of the asymptotic variance under weaker conditions than those used here – and the application of this approach to such problems as Bayesian optimal design could be of interest.

The work presented in chapter 5 admits numerous possible avenues of inquiry: strategies for deciding upon the balance between the number of particles and the number of intermediate distributions used are needed, as are methods for designing sensible proposal kernels; theoretically, various rates of convergence and stability results would be of interest, particularly, the restrictions upon the annealing schedule which are required to guarantee convergence would be interesting. Clearly, the dynamic rare event simulation work is far from exhausted; we have not yet begun to address the problems which will occur when attempting to design sensible proposal kernels for complex problems or to conduct a thorough theoretical analysis of the asymptotic properties of the estimator.

Concurrent with the work presented in this thesis, a number of developments have been made in the field of sequential Monte Carlo and population based sampling techniques – particularly theoretical and methodological results. Further work remains to be done in these directions, and is perhaps of more interest than direct extensions or applications of the work presented within this thesis. Of particular interest are the stratified SMC technique [83], some work on population-

based RJMCMC methods [84, 85] and various techniques for the construction of self-interacting Markov chains with desired invariant distributions. Further developments of population-based sampling techniques over the coming years will hopefully lead to more versatile sampling strategies with faster convergence to the distributions of interest.

6.3 Summary

This thesis has developed novel SMC techniques and gone some way towards illustrating the versatility of such methods. The use of population-based Monte Carlo techniques depending upon sequential importance sampling and resampling to solve general sampling problems is an as yet under-developed technique, and one which shows great promise for the future. Although much remains to be done, and many questions remain unanswered, it is apparent that intelligent implementation of SMC algorithms can address a wide variety of problems.

References

"By other men's labours we are led to the sight of things most beautiful that have been wrested from darkness and brought into light"

– Senenca, "The Shortness of Life"

[1] R. J. Allen, D. Frenkel, and P. R. ten Wolde. Simulating rare events in equilibrium or non-equilibrium stochastic systems. *Journal of Chemical Physics*, 124(024102):1–16, 2006.

[2] D. F. Andrews and C. L. Mallow. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B*, 36(1):99–103, 1974.

[3] C. Andrieu, M. Davy, and A. Doucet. Improved auxiliary particle filtering: Applications to time-varying spectral analysis. In *Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing*, pages 309–312, Singapore, 2001.

[4] S. Asmussen, K. Binswanger, and B. Højgaard. Rare events simulation for heavy-tailed distributions. *Bernoulli*, 6(2):303–322, 2000.

[5] Y. Bar-Shalom and X. R. Li. *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS Publishing, Storrs, CT, 1995.

[6] P. H. R. Barbosa, E. P. Raposo, and M. D. Coutinho-Filho. Monte Carlo studies of the spin-glass-like phase in $Fe_{0.25}Zn_{0.25}F_2$. *Journal of Applied Physics*, 87(9):6531–6533, May 2000.

[7] A. A. Barker. Monte Carlo calculations of the radial distribution function for a proton-electron plasma. *Australian Journal of Physics*, 18:119–133, 1965.

[8] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, second edition, 1986.

[9] G. Biondini, W. L. Kath, and C. R. Menyuk. Importance sampling for polarization-mode distortion. *IEEE Photonic Technology Letters*, 14(2):310–312, 2002.

[10] L. M. Birch. *Computational Insights into Protein-Ligand Interactions*. Ph.D. thesis, University of Cambridge, Department of Chemistry, Lensfield Road, Cambridge, UK, August 2005.

[11] O. Cappé, A. Guillin, Jean-Michel Marin, and Christian P. Robert. Population Monte Carlo. Technical Report 0234, CEREMADE, Université Paris Dauphine, Université Paris IX, Paris, 2002.

[12] O. Cappé, C. P. Robert, and T. Rydén. Reversible jump, birth-and-death and more general continuous time MCMC samplers. *Journal of the Royal Statistical Society*, B 65(3):679–700, 2003.

[13] 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.

[14] G. Casella, M. Lavine, and C. P. Robert. Explaining the perfect sampler. *American Statistician*, 55:299–305, 2001.

[15] G. Casella and C. P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

[16] G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, September 2000.

[17] 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.

[18] F. Cérou and A. Guyader. Adaptive multilevel splitting for rare event analysis. Research Report 5710, INRIA, October 2005.

[19] H. P. Chan and T. L. Lai. Importance sampling for generalized likelihood ratio procedures in sequential analysis. *Sequential Analysis*, 24:259–278, 2005.

[20] Y. Chen. Another look at rejection sampling through importance sampling. Working Paper 04-30, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708, USA, 2004.

[21] S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolois-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

[22] J. Ching, S. K. Au, and J. L Beck. Reliability estimation using dynamical systems subject to stochastic excitation using subset simulation with splitting. *Computer Methods in Applied Mechanics and Engineering*, 194:1557–1579, 2005.

[23] J. Ching, J. L Beck, and S. K. Au. Hybrid subset simulation method for reliability estimation of dynamical systems subject to stochastic excitation. *Probabilistic Engineering Mechanics*, 20(3):199–214, July 2005.

[24] N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–551, 2002.

[25] N. Chopin. Central limit theorem for sequential Monte Carlo methods and its applications to Bayesian inference. *Annals of Statistics*, 32(6):2385–2411, December 2004.

[26] B. A. Cipra. The best of the 20th century: Editors name top 10 algorithms. *SIAM News*, 33(4), 2000.

[27] D. E. Clark and J. Bell. Convergence results for the particle PHD filter. *IEEE Transactions on Signal Processing*, 54(7), 2005.

[28] D. Crisan. Particle filters – a theoretical perspective. In Doucet et al. [47], pages 17–41.

[29] D. Crisan, P. Del Moral, and T. Lyons. Discrete filtering using branching and interacting particle systems. *Markov Processes and Related Fields*, 5(3):293–318, 1999.

[30] D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746, March 2002.

[31] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*, volume I: Elementary Theory and Methods of *Probability and Its Applications*. Springer, New York, second edition, 2003.

[32] A. C. Davison. *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, The Edinburgh Building, Cambridge, CB2 2RU, 2003.

[33] T. Homem de Mello and R. Y. Rubinstein. Rare event estimation for static models via cross-entropy and importance sampling. Manuscript, Technion, Haifa, Israel, 2002.

[34] P. Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Probability and Its Applications. Springer Verlag, New York, 2004.

[35] P. Del Moral and A. Doucet. On a class of genealogical and interacting Metropolis models. *Lecture Notes in Mathematics*, 1832:415–446, 2003.

[36] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers (2005 revision). Technical Report CUED/F-INFENG/TR-443, University of Cambridge, Department of Engineering, 2005. Revised version of a report from 2002.

[37] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo methods for Bayesian Computation. In *Bayesian Statistics 8*. Oxford University Press, 2006.

[38] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society B*, 63(3):411–436, 2006.

[39] P. Del Moral, A. Doucet, and G. Peters. Asymptotic and increasing propagation of chaos expansions for genealogical particle models. *Proceedings of the LSP*, 2004. Submitted.

[40] P. Del Moral and J. Garnier. Genealogical particle analysis of rare events. *Annals of Applied Probability*, 15(4):2496–2534, November 2005.

[41] P. Del Moral and A. Guionnet. Central limit theorem for non linear filtering and interacting particle systems. *Annals of Applied Probability*, 9(2):275–297, 1999.

[42] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39:2–38, 1977.

[43] L. Devroye. *Non-Uniform Random Variate Generation*. Springer Verlag, New York, 1986.

[44] J. Dongarra and F. Sullivan. The top 10 algorithms. *Computing in Science & Engineering*, 6(1):22–23, January 2000.

[45] M. Doran and C. M. Müller. Analyse this! a cosmological constraint package for cmbeasy. *ArXiv Mathematics e-prints*, astro-ph(0311311v2), 2004.

[46] A. Doucet, N. de Freitas, and N. Gordon. *An Introduction to Sequential Monte Carlo Methods*, pages 3–14. In *Statistics for Engineering and Information Science* [47], 2001.

[47] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer Verlag, New York, 2001.

[48] A. Doucet, S. Godsill, and C. Andrieu. On sequential simulation-based methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[49] A. Doucet, S. J. Godsill, and C. P. Robert. Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12:77–84, 2002.

[50] R. Eckhardt. Stan Ulam, John von Neumann and the Monte Carlo method. *Los Alamos Science*, 15:131–, 1987.

[51] J. H. J. Einmahl and D. M. Mason. Generalized quantile processes. *Annals of Statistics*, 20(2):1062–1078, June 1992.

[52] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, June 1995.

[53] P. Fearnhead. Perfect simulation from non-neutral population genetic models: Variable population size and population sub-division. *Genetics*, 2006. To appear.

[54] J. A. Fill. An interruptible algorithm for perfect sampling via Markov Chains. *Annals of Applied Probability*, 8(1):131–162, 1998.

[55] C. Gaetan and J.-F. Yao. A multiple-imputation Metropolis version of the EM algorithm. *Biometrika*, 90(3):643–654, 2003.

[56] J. Garnier and P. Del Moral. Simulations of rare events in fibre optics by interacting particle systems. *Elsevier Science*, 2006. Submitted.

[57] 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.

[58] 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.

[59] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[60] I. Gentil, B. Rémillard, and P. Del Moral. Filtering of images for detecting multiple targets trajectories. In *Statistical Modeling and Analysis for Complex Data Problems*. Springer Verlag, April 2005.

[61] J. Geweke. Bayesian inference in econometrics models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339, November 1989.

[62] W. R. Gilks. Derivative-free adaptive rejection sampling for Gibbs sampling. In J. M. Bernado, J. O. Berger, A. P. David, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 641–649. Oxford University Press, 1992.

[63] W. R. Gilks and C. Berzuini. Following a moving target – Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society B*, 63:127–146, 2001.

[64] W. R. Gilks and Carlo Berzuini. RESAMPLE-MOVE filtering with Cross-Model jumps. In Doucet et al. [47], pages 117–138.

[65] W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 44(4), 1995.

[66] W. R. Gilks, S. Richardson, and D. J. Spieghalter, editors. *Markov Chain Monte Carlo In Practice*. Chapman and Hall, first edition, 1996.

[67] W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348, 1992.

[68] P. Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer Verlag, Heidelberg, 2003.

[69] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47(4):585–600, July 1999.

[70] P. Glasserman and Jingyi Li. Importance sampling for portfolio credit risk. Available at "http://www2.gsb.columbia.edu/faculty/pglasserman/Other/is_credit.pdf", December 2003.

[71] I. R. Goodman, R. P. S. Mahler, and H. T. Nguyen. *Mathematics of Data Fusion*. Kluwer Academic Publishers, 1997.

[72] N. J. Gordon, S. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, April 1993.

[73] P. J. Green. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[74] P. J. Green. Penalized likelihood. In *Encyclopaedia of Statistical Science, Update Volume*, volume 2. 1998.

[75] P. J. Green. Trans-dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, Oxford Statistical Science Series, chapter 6. Oxford University Press, 2003.

[76] T. E. Harris. The existence of stationary measures for certain Markov processes. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 113–124. University of California Press, 1956.

[77] W. K. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 52:97–109, 1970.

[78] D. B. Hitchcock. A history of the Metropolis-Hastings algorithm. *The American Statistician*, 57:254–257, 2003.

[79] C. C. Holmes and L. Held. Bayesian auxiliary variable models for binary and polychotomous regression. *Bayesian Analysis*, 2006. To Appear.

[80] C.-R. Hwang. Laplace's method revisited: Weak convergence of probability measures. *The Annals of Probability*, 8(6):1177–1182, December 1980.

[81] E. Jacquier, M. Johannes, and N. Polson. MCMC maximum likelihood for latent state models. *Journal of Econometrics*, 2005. To Appear.

[82] A. Jasra and A. Doucet. Stability of sequential Monte Carlo samplers via the Foster-Lyapunov condition. *Submitted*, 2006.

[83] A. Jasra, A. Doucet, D. A. Stephens, and C. C. Holmes. Stratified sequential Monte Carlo samplers for trans-dimensional simulation. 2006. Submitted.

[84] A. Jasra, D. A. Stephens, and C. C. Holmes. On population-based simulation for static inference. 2006. Submitted.

[85] A. Jasra, D. A. Stephens, and C. C. Holmes. Population-based reversible jump Markov chain Monte Carlo. 2006. In Preparation.

[86] 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.

[87] A. M. Johansen, A. Doucet, and M. Davy. Maximum likelihood parmeter estimation for maximum likelihood models using sequential Monte Carlo. In *Proceedings of ICASSP*, volume III, pages 640–643. IEEE, May 2006.

[88] 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.

[89] A. M. Johansen, P. Del Moral, and A. Doucet. Sequential Monte Carlo samplers for rare events. In *Proceedings of the 6th International Workshop on Rare Event Simulation*, Bamberg, Germany, October 2006.

[90] A. M. Johansen, S. S. Singh, A. Doucet, and B.-N. Vo. Convergence of the SMC implementation of the PHD filter. Technical Report CUED/F-INFENG/TR-517, University of Cambridge, Department of Engineering, Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, April 2005.

[91] A. M. Johansen, S. S. Singh, A. Doucet, and B.-N. Vo. Convergence of the SMC implementation of the PHD filter. *Methodology and Computing in Applied Probability*, 8(2):265–291, June 2006.

[92] G. Kitagawa. Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032–1041, December 1987.

[93] G. Kitagawa. Monte Carlo filter and smoother for Non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5:1–25, 1996.

[94] M. Klass, N. de Freitas, and A. Doucet. Towards practical n^2 Monte Carlo: The marginal particle filter. In *Proceedings of Uncertainty in Artificial Intelligence*, 2005.

[95] D. E. Knuth. *The Art of Computer Programming*, volume 2: Seminumerical Algorithms. Addison-Wesley, Boston, third edition, 1998.

[96] D. E. Knuth. *The Art of Computer Programming*, volume 3: Sorting and Searching. Addison-Wesley, Boston, second edition, 1998.

[97] 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.

[98] C. Kreucher, K. Kastella, and A. Hero. Tracking multiple targets using a particle filter representation of the joint multitarget probability density. In O. E. Drummond, editor, *Signal and Data POrocessing of Small Targets*, volume 5204 of *Proceedings of SPIE*, pages 258–269, 2003.

[99] H. Kück, N. de Freitas, and A. Doucet. SMC samplers for Bayesian optimal nonlinear design. In *Proceedings of the Nonlinear Statistical Signal Processing Workshop*, Cambridge, 2006.

[100] H. R. Künsch. Recursive Monte Carlo filters: Algorithms and theoretical analysis. *Annals of Statistics*, 33(5):1983–2021, 2005.

[101] A. Lagnoux. Rare event simulation. *Probability in the Engineering and Informational Sciences*, 20(1):43–66, 2006.

[102] J. Liu and R. Chen. Blind decovolution via sequential imputation. *Journal of the American Statistical Association*, 90(430):567–576, June 1995.

[103] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer Verlag, New York, 2001.

[104] J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, September 1998.

[105] S. N. MacEachern, M. Clyde, and J. S. Liu. Sequential importance sampling for nonparametric Bayes models: The next generation. *Canadian Journal of Statistics*, 27(2):251–267, 1999.

[106] R. P. S. Mahler. Bayesian cluster detection and tracking using a generalized Cheeseman approach. In *Proceedings of the 2003 Aerosense Conference*, volume 5096. SPIE, 2003.

[107] R. P. S. Mahler. Multitarget Bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1152, October 2003.

[108] S. Malefaki and G. Iliopoulos. On convergence of importance sampling and other properly weighted samples to the target distribution. *ArXiv Mathematics e-prints*, ST(0505045), May 2005.

[109] G. Marsaglia. The squeeze method for generating gamma variates. *Computers and Mathematics with Applications*, 3:321–325, 1977.

[110] N. Metropolis. The beginnings of the Monte Carlo method. *Los Alamos Science*, 15:125–130, 1987.

[111] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[112] N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, September 1949.

[113] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Verlag, 1993.

[114] T. Minka. A comparison of numerical optimizers for logistic regression. Technical Report 758, CMU Statistics, 2003. Revised version.

[115] R. O. Moore, G. Biondinin, and W. L. Kath. Importance sampling for noise-induced amplitude and timing jitter in soliton transmission systems. *Optics Letters*, 28(2):105–107, January 2003.

[116] P. Müller, B. Sansó, and M. de Iorio. Optimum Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, 99:788–798, 2004.

[117] R. M. Neal. Annealed importance sampling. Technical Report 9805, University of Toronto, Department of Statistics, 1998.

[118] W. Ng, J. Li, S. Godsill, and J. Vermaak. A review of recent results in multiple target tracking. In *Proceedings of ISPA*, 2005.

[119] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Number 83 in Cambridge Tracts in Mathematics. Cambridge University Press, 1st paperback edition, 1984.

[120] M. Palassini and S. Caracciolo. Monte Carlo simulation of the three-dimensional Ising spin glass. In D. P. Landau, S. P. Lewis, and H. B. Schuetter, editors, *Proceedings of the XIIth Workshop on Computer Simulation Studies in Condensed Matter Physics*, 1999.

[121] F. Perron. Beyond accept-reject sampling. *Biometrika*, 86(4):803–813, December 1999.

[122] P. H. Peskun. Optimum Monte-Carlo sampling using Markov Chains. *Biometrika*, 60:607–612, 1973.

[123] G. W. Peters. Topics in sequential Monte Carlo samplers. M.Sc. thesis, University of Cambridge, Department of Engineering, January 2005.

[124] M. K. Pitt. Smooth particle filters for likelihood evaluation and maximisation. Warwick Economic Research Papers 651, University of Warwick, Department of Economics, 2002.

[125] M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

[126] M. K. Pitt and N. Shephard. Auxiliary variable based particle filters. In Doucet et al. [47], chapter 13, pages 273–293.

[127] J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov Chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252, 1996.

[128] J. G. Propp and D. B. Wilson. How to get a perfectly random sample from a generic Markov Chain and generate a random spanning tree of a directed graph. *Journal of Algorithms*, 27:170–217, 1998.

[129] E. Punskaya, A. Doucet, and W. J. Fitzgerald. On the use and misuse of particle filtering in digital communications. In *Proceedings of EUSIPCO*, Toulouse, 2002.

[130] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Verlag, New York, second edition, 2004.

[131] C. P. Robert and D. M. Titterington. Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing*, 8:145–158, 1998.

[132] G. Roberts. Markov Chain concepts related to sampling algorithms. In Gilks et al. [66], chapter 3, pages 45–54.

[133] K. Roeder. Density estimation with cofidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association*, 85(411):617–624, September 1990.

[134] K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902, September 1997.

[135] J. J. K. Ó Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Statistics and Computing. Springer Verlag, 1996.

[136] D. B. Rubin. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few impotations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82(298):543–546, June 1987.

[137] R. Rubinstein. Optimization of computer simulations of rare events. *European Jouranl of Operations Research*, 99:89–112, 1997.

[138] R. Rubinstein and D. Kroese. *The Cross-Entropy Method*. Springer Verlag, 2004.

[139] P. Shahabuddin. Rare event simulation in stochastic models. In C. Alexopoulos, W. R. Lielegdon, and D. Goldsman, editors, *Proceedings of the Winter Simulation Conference*, pages 178–185, 1995.

[140] A. N. Shiryaev. *Probability*. Number 95 in Graduate Texts in Mathematics. Springer Verlag, New York, second edition, 1995.

[141] H. Sidenbladh. Multi-target particle filtering for the probability hypothesis density. In *6th International Conference on Information Fusion*, pages 800–806, Cairns, Australia, 2003.

[142] D. Siegmund. Importance sampling in Monte Carlo study of sequential tests. *Annals of Statistics*, 4(4):673–684, July 1976.

[143] A. Slosar and M. P. Hobson. An improved Markov-chain monte carlo sampler for the estimation of cosmological parameters from CMB data. *ArXiv Mathematics e-prints*, astro-ph(0307219v1), 2003.

[144] M. Stephens. Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74, 2000.

[145] L. Tierney. Markov Chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1762, 1994.

[146] L. Tierney. Introduction to general state space Markov Chain theory. In Gilks et al. [66], chapter 4, pages 59–74.

[147] B. Vo, S.S. Singh, and A. Doucet. Sequential Monte Carlo methods for multi-target filtering with random finite sets. *IEEE Transactions on Aerospace and Electronic Systems*, 41(4):1223–1245, October 2005.

[148] T. Zajic and R. P. S. Mahler. A particle-systems implementation of the PHD multitarget tracking filter. In Ivan Kadar, editor, *Signal Processing, Sensor Fusion and Target Recognition XII*, volume 5096 of *Proceedings of SPIE*, pages 291–299, 2003.

A. Rare Event Simulation via SMC

A.1 Comparison of path sampling with the naïve method

For simplicity we consider only some simple Gaussian measures defined upon $\mathbb{R}$.

In this case we consider estimating the normalising constant of a distribution $\pi_1(\lambda(dx)) = \frac{1}{Z_1} g_1(x)\pi_0(\lambda(dx)) \propto g_1(x)\pi_0(\lambda(dx)) = f_1(x)$ where $\pi_0(\lambda(dx)) = \mathcal{N}\left(\lambda(dx); \mu, \sigma^2\right)$ and $g_1(x) = \exp(\alpha x)$ is a potential function. This is, of course, equivalent to estimating the ratio of normalising constants, but in this case we have the luxury of knowing that $Z_0 = 1$.

Naïve Importance Sampling. Perhaps the simplest approach is to obtain N *iid* samples, $\{X_i\}_{i=1}^N$ from π_0 and use the following relationship to estimate Z_1:

$$\begin{aligned} \pi_0\left(\frac{\mathrm{d}f_1}{\mathrm{d}\pi_0}\right) &= \pi_0\left(Z_1 \frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_0}\right) \\ &= \pi_1(Z_1) = Z_1, \end{aligned}$$

which suggests the naïve importance sampling (IS) estimator,

$$\hat{Z}_1^{IS} = \frac{1}{N}\sum_{i=1}^{N} \frac{f_1(X_i)}{\pi_0(X_i)}. \tag{A.1}$$

It is clear that this estimator is consistent:

$$\mathbb{E}\left(\hat{Z}_1^{IS}\right) = \pi_0\left(\frac{\mathrm{d}f_1}{\mathrm{d}\pi_0}\right) = Z_1,$$

and the variance is given by:

$$\begin{aligned} \mathrm{Var}\left(\hat{Z}_1^{IS}\right) &= \frac{1}{N}\left[\pi_0\left(\left[\frac{\mathrm{d}f_1}{\mathrm{d}\pi_0}\right]^2\right) - Z_1^2\right] \\ &= \frac{1}{N} Z_1^2 \left[\pi_1\left(\frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_0}\right) - 1\right]. \end{aligned}$$

The measure π_1 is a Gaussian with mean $\mu + \sigma^2\alpha$ and the same variance as the original measure. Thus, the variance expression is:

$$\begin{aligned}\mathrm{Var}\left(\hat{Z}_1^{IS}\right) &= \frac{1}{N}\left[\exp\left(\frac{1}{2}\alpha^2\sigma^2+\alpha\mu\right)\right]^2\left[\int_{-\infty}^{\infty}\frac{\mathcal{N}\left(x;\mu+\sigma^2\alpha,\sigma^2\right)^2}{\mathcal{N}\left(x;\mu,\sigma^2\right)}\lambda\left(dx\right)-1\right]\\ &= \frac{1}{N}\left[\exp\left(\frac{1}{2}\alpha^2\sigma^2+\alpha\mu\right)\right]^2\left[\exp\left(\frac{\left(\mu+\sigma^2\alpha-\mu\right)^2}{\sigma^2}\right)-1\right].\end{aligned}$$

This gives us as the variance of the importance sampling estimator:

$$\mathrm{Var}\left(\hat{Z}_1^{IS}\right)=\frac{1}{N}\exp\left(\alpha^2\sigma^2+2\alpha\mu\right)\left[\exp(\alpha^2\sigma^2)-1\right]. \tag{A.2}$$

In order to compare this expression with those obtained from path sampling it is convenient to obtain an estimate for $\mathrm{Var}\left(\log(\hat{Z}_1^{IS})\right)$, which can be done by the delta-method [8, p. 368].

$$\begin{aligned}\mathrm{Var}\left(\log(\hat{Z}_1^{IS})\right) &\approx \left[\frac{\mathrm{d}\log Z_1^{IS}}{\mathrm{d}Z_1^{IS}}\right]^2_{\mathbb{E}Z_1^{IS}}\mathrm{Var}\left(\hat{Z}_1^{IS}\right)\\ &= \frac{1}{N}\left[\frac{1}{Z_1}\right]^2\mathrm{Var}\left(\hat{Z}_1^{IS}\right)\\ &= \frac{1}{N}\left[\exp(\alpha^2\sigma^2)-1\right].\end{aligned} \tag{A.3}$$

The result follows from the consistency of the estimator of $\hat{Z}_1^{IS}$ and the above results.

Approximate Path Sampling via Importance Sampling. We use a Monte Carlo approximation of [58] to obtain an alternative estimator of $\log Z_1$ which shall be denoted by $\widehat{\log Z_1}^{PS(IS)}$.

The standard path sampling formula tells us that given a parameterised form, q_θ selected such that $q_0=\pi_0$ and $q_1=g_1$ then:

$$\log Z_1 = \int_0^1 \mathbb{E}\left[\left.\frac{\mathrm{d}\log q(\cdot|\theta)}{\mathrm{d}\theta}\right|_{\theta=\theta'}\right]d\theta'.$$

Our approach is to discretise the integral over θ by evaluating it at a set of T points located at $\theta_{t_1=0}<\theta_{t_2}<\cdots<\theta_{t_{T-1}}<\theta_{t_T}=1$ The expectation can then be evaluated at each of these values of theta via a Monte Carlo approach. In our case, we consider estimating each of these expectations by importance sampling using π_0 as the importance distribution. Thus we have:

$$\widehat{\log Z_1}^{PS(IS)}=\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{\lfloor N/T\rfloor}\left.\frac{\mathrm{d}\log q_\theta}{\mathrm{d}\theta}\right|_{\theta=\theta_t}(X_{t,n})W_{t,n}$$

where the random variables $X_{t,n}$ are all *iid* samples from π_0 and the importance weights $W_{t,n}$ are estimated according to:

$$W_{t,n}=\frac{q_\theta(X_{t,n})/\pi_0(X_{t,}n)}{\sum_{m=1}^{\lfloor N/T\rfloor}q_\theta(X_{t,m})/\pi_0(X_{t,m})}.$$

However, as this estimator is biased and the ratio of two correlated estimators, it is not trivial to obtain an analytic expression for the variance of this estimator.

Path Sampling without Importance Sampling. If, instead, we assume that we are able to obtain samples from $\pi_\theta \propto q_\theta$ then we are free to use the alternative estimator:

$$\widehat{\log Z_1}^{PS(S)} = \frac{1}{T\lfloor N/T \rfloor} \sum_{t=1}^{T} \sum_{n=1}^{\lfloor N/T \rfloor} \left. \frac{\mathrm{d} \log q_\theta}{\mathrm{d}\theta} \right|_{\theta=\theta_t} (X_{t,n}),$$

where, the random variables $X_{t,n}$ are all independent samples from π_{θ_t}.

Using the independence of the random variables, the variance of this expression can be decomposed as:

$$\begin{aligned} \mathrm{Var}\left(\widehat{\log Z_1}^{PS(S)}\right) &= \frac{1}{T^2} \sum_{t=1}^{T} \mathrm{Var}\left(\frac{1}{\lfloor N/T \rfloor} \sum_{n=1}^{\lfloor N/T \rfloor} \left. \frac{\mathrm{d} \log q_\theta}{\mathrm{d}\theta} \right|_{\theta=\theta_t} (X_{t,n}) \right) \\ &= \frac{1}{(\lfloor N/T \rfloor T)^2} \sum_{t=1}^{T} \sum_{n=1}^{\lfloor N/T \rfloor} \mathrm{Var}\left(\left. \frac{\mathrm{d} \log q_\theta}{\mathrm{d}\theta} \right|_{\theta=\theta_t} (X_{t,n}) \right). \end{aligned}$$

In our particular case, we know that $\left. \frac{\mathrm{d} \log q_\theta}{\mathrm{d}\theta} \right|_{\theta=\theta_t} (X) = \alpha X$, which allows us to obtain a closed form expression for this variance under these simplifying assumptions. This leads us to the following expression:

$$\begin{aligned} \mathrm{Var}\left(\widehat{\log Z_1}^{PS(S)}\right) &= \frac{\alpha^2}{(\lfloor N/T \rfloor T)^2} \sum_{t=1}^{T} \sum_{n=1}^{\lfloor N/T \rfloor} \mathrm{Var}\left(X\right) \\ &= \frac{\alpha^2 \sigma^2}{\lfloor N/T \rfloor T}, \end{aligned}$$

which follows from the fact that the variance expression proves to be independent of t and all of $\{X_{t,n}\}_{n=1}^{\lfloor N/T \rfloor}$ are identically distributed for a given t.

Importance Sampling with Intermediate Distributions. Using the same formulation as previously used for the path sampling framework, it is possible to consider the following importance sampling estimator of the normalising constant of interest, where the $X_{t,n}$ are independently drawn from $\pi_{\theta_{t-1}}$:

$$\hat{Z}_1^{ISS} = \prod_{t=1}^{T} \frac{1}{\lfloor N/T \rfloor} \sum_{n=1}^{\lfloor N/T \rfloor} \frac{q_{\theta_t}(X_{t,n})}{q_{\theta_{t-1}}(X_{t,n})}.$$

To compare the variance of this estimator with that obtained from path sampling it is useful to obtain the logarithm of this estimator and to consider its variance via the delta method. This also allows us to consider the variances of the individual likelihood ratios and sum them.

By precisely the same arguments as in the naïve importance sampling section we can write the variance of any one of these importance ratios as:

$$\mathrm{Var}\left(\log \hat{Z}_1^{ISS}\right) = \sum_{t=1}^{T} \mathrm{Var}\left(\log \left[\frac{1}{\lfloor N/T \rfloor} \sum_{n=1}^{\lfloor N/T \rfloor} \frac{q_{\theta_t}(X_{t,n})}{q_{\theta_{t-1}}(X_{t,n})} \right] \right).$$

It is clear that:

$$\mathrm{Var}\left(\frac{1}{\lfloor N/T\rfloor}\sum_{n=1}^{\lfloor N/T\rfloor}\frac{q_{\theta_t}(X_{t,n})}{q_{\theta_{t-1}}(X_{t,n})}\right) = \frac{1}{\lfloor N/T\rfloor}\left(\frac{Z_{\theta_t}}{Z_{\theta_{t-1}}}\right)^2\left(\pi_{\theta_t}\left(\frac{\mathrm{d}\pi_{\theta_t}}{\mathrm{d}\pi_{\theta_{t-1}}}\right)-1\right),$$

and we know that:

$$\begin{aligned}\pi_{\theta_t}\left(\frac{\mathrm{d}\pi_{\theta_t}}{\mathrm{d}\pi_{\theta_{t-1}}}\right) &= \int\frac{\left(\mathcal{N}\left(x;\mu+\alpha\theta_t\sigma^2,\sigma^2\right)\right)^2}{\mathcal{N}\left(x;\mu+\alpha\theta_{t-1}\sigma^2,\sigma^2\right)}\lambda\left(dx\right)\\ &= e^{\alpha(\theta_t-\theta_{t-1})\sigma^2}.\end{aligned}$$

Hence:

$$\mathrm{Var}\left(\frac{1}{\lfloor N/T\rfloor}\sum_{n=1}^{\lfloor N/T\rfloor}\frac{q_{\theta_t}(X_{t,n})}{q_{\theta_{t-1}}(X_{t,n})}\right) = \frac{1}{\lfloor N/T\rfloor}\left(\frac{Z_{\theta_t}}{Z_{\theta_{t-1}}}\right)^2\left(e^{\alpha^2(\theta_t-\theta_{t-1})^2\sigma^2}-1\right),$$

and employing the delta method again, we have:

$$\begin{aligned}\mathrm{Var}\left(\log\left[\frac{1}{\lfloor N/T\rfloor}\sum_{n=1}^{\lfloor N/T\rfloor}\frac{q_{\theta_t}(X_{t,n})}{q_{\theta_{t-1}}(X_{t,n})}\right]\right) &= \frac{\left(e^{\alpha^2(\theta_t-\theta_{t-1})^2\sigma^2}-1\right)}{\lfloor N/T\rfloor}\\ \mathrm{Var}\left(\log\hat{Z}_1^{ISS}\right) &= \frac{\sum_{t=1}^{T}\left(e^{\alpha^2(\theta_t-\theta_{t-1})^2\sigma^2}-1\right)}{\lfloor N/T\rfloor}.\end{aligned}$$

If we assume that $\theta_t = t/T$ then we obtain the homogeneous result:

$$\mathrm{Var}\left(\log\hat{Z}_1^{ISS}\right) = \frac{\sum_{t=1}^{T}\left(e^{\frac{\alpha^2\sigma^2}{T^2}}-1\right)}{\lfloor N/T\rfloor}. \tag{A.4}$$

Comparison. We present in figure A.1 the variance of 100 repetitions of three of the estimators described above with a total number of particles ranging from 100 to 100,000 spaced in equal logarithmic increments. Figure A.2 shows the absolute difference between the mean of each of these sets of 100 estimates and the true value (50). This illustrates that the importance sampling from the prior approach is doomed to fail in cases where the mass of the prior and the target distribution are substantially differently distributed. It also shows that the path sampling approaches introduced above perform significantly better than the naïve methods.

It is instructive to consider the theoretical relationship between the variance of the two estimators which do perform adequately. If we consider sampling exactly from a sequence of intermediate distributions, then the variance of the importance sampling estimator is:

$$\mathrm{Var}\left(\log\hat{Z}_1^{ISS}\right) = \frac{\sum_{t=1}^{T}\left(e^{\frac{\alpha^2\sigma^2}{T^2}}-1\right)}{\lfloor N/T\rfloor}$$

whilst that of the equivalent path sampling estimator is

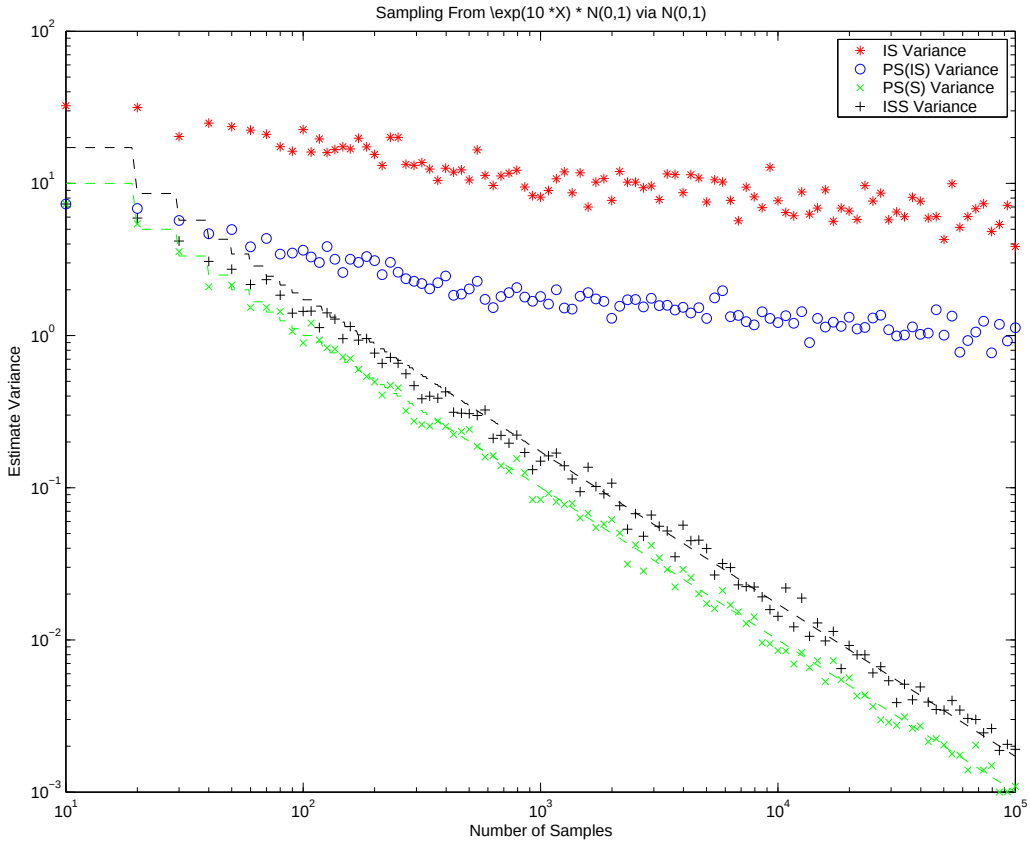


Fig. A.1. Variance of four estimators used to determine the normalising constant of a distribution proportional to $\mathcal{N}(x; 0, 1) \exp(10x)$ estimated from 50 applications of each estimator. Dashed lines correspond to the theoretical values where appropriate. The theoretical variance of the naïve IS estimator is enormous. This hasn't been realised here because it is due to extremely rare samples with tremendously large weights which were not, in fact, produced in any of these samples – it would not be useful to include the theoretical variance of this estimator.

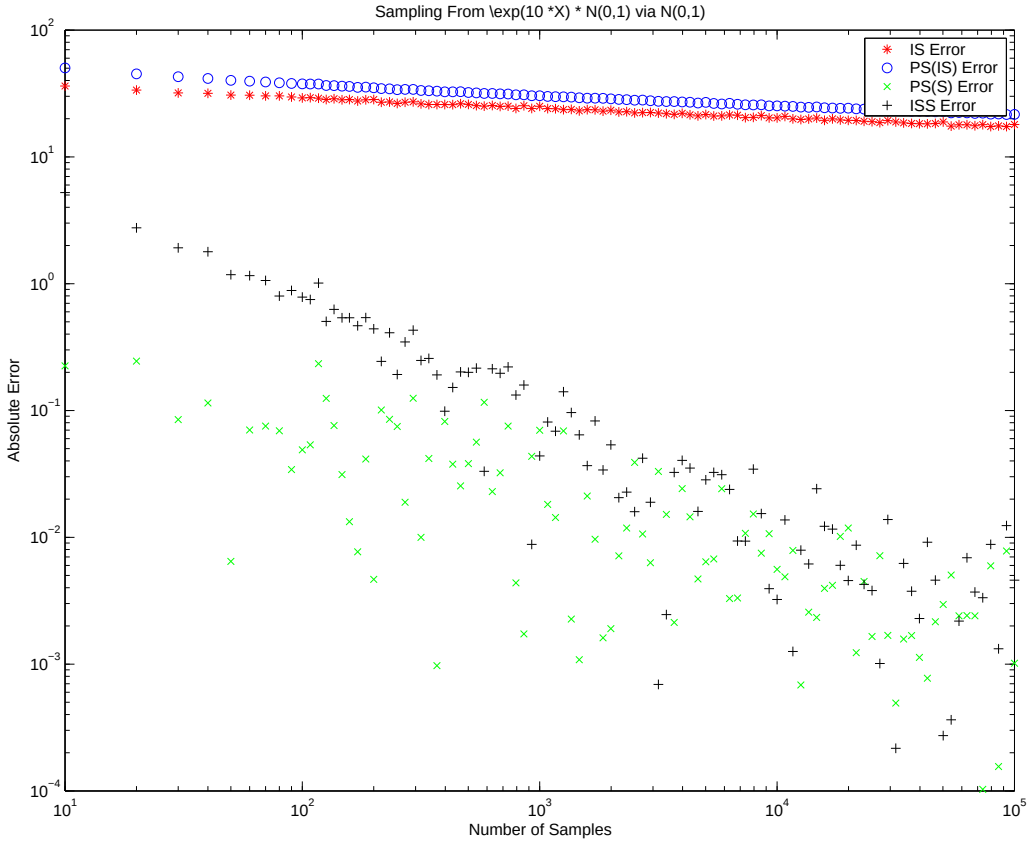


Fig. A.2. Absolute error of four estimators used to determine the normalising constant of a distribution proportional to $\mathcal{N}(x; 0, 1)\exp(10x)$ estimated from the mean of 50 applications. For the purpose of comparison, the true value of $\log(Z_1)$ is 50.

$$\mathrm{Var}\left(\widehat{\log Z_1}^{PS(S)}\right) = \frac{\alpha^2\sigma^2}{\lfloor N/T \rfloor T}.$$

We can see that:

$$\begin{aligned}
\frac{\mathrm{Var}\left(\log \hat{Z}_1^{ISS}\right)}{\mathrm{Var}\left(\widehat{\log Z_1}^{PS(S)}\right)} &= \frac{\frac{\sum_{t=1}^{T}\left(e^{\frac{\alpha^2\sigma^2}{T^2}}-1\right)}{\lfloor N/T \rfloor}}{\frac{\alpha^2\sigma^2}{\lfloor N/T \rfloor T}} \\
&= \frac{T^2}{\alpha^2\sigma^2}\left(e^{\frac{\alpha^2\sigma^2}{T^2}}-1\right) \\
&= \left(\frac{T}{\alpha\sigma}\right)^2 \sum_{i=1}^{\infty}\frac{1}{i!}\left(\frac{\alpha\sigma}{T}\right)^{2i} \\
&\geq 1+\frac{1}{2}\left(\frac{\alpha\sigma}{T}\right)^2,
\end{aligned}$$

hence the path-sampling estimator dominates the equivalent direct importance sampling estimator in this framework.

A.2 Variance of a Special Case

In order to illustrate the advantages of the methods proposed here over the crude Monte Carlo approach, it is useful to consider a particularly simple special case.

Let $\pi_k(dx_k) = \frac{1}{Z_k}\mathbb{I}_{T_k}(x_k)$ and let the proposal kernel be simply $K_k(x_{k-1}, dx_k) = \pi_k(dx_k)$. Further, assume that we are able to use the time-reversal Markov Kernel for the backward case, giving:

$$L_{k-1}(x_k, x_{k-1}) = \frac{\pi_k(x_{k-1})K_k(x_{k-1}, x_k)}{\pi_k(x_k)} = \pi_k(x_{k-1}).$$

In this case we obtain, using $\tilde{\pi}_k$ for the full extended space distribution of the SMC sampler:

$$\tilde{\pi}_n(x_k) = \int \pi_n(x_n)\prod_{j=k}^{n-1} L_j(x_{j+1}, x_j)dx_{k+1:n} = \pi_{k+1}(x_k).$$

The naïve estimator of the ratio of normalising constants obeys the following central limit theorem [37]:

Theorem A.2.1 (Del Moral et al.). *The naïve estimate of the normalising constant obeys the following central limit theorem.*

$\lim_{N\to\infty}\sqrt{N}\left(\log\frac{\widehat{Z_n}}{Z_1} - \log\frac{Z_n}{Z_1}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{SMC,n}\right)$,

where

$$\sigma^2_{SMC,n} = \int \frac{\tilde{\pi}_n(x_1)^2}{\eta_1(x_1)} dx_{n-1:n} - 1 \tag{A.5}$$

$$+ \sum_{k=2}^{n-1} \left(\int \frac{\left(\tilde{\pi}_k(x_k) L_{k-1}(x_k, x_{k-1})\right)^2}{\pi_{k-1}(x_{k-1}) K_k(x_{k-1}, x_k)} dx_{k-1:k} - 1 \right) \tag{A.6}$$

$$+ \int \frac{\left(\pi_n(x_n) L_{n-1}(x_n, x_{n-1})\right)^2}{\pi_{n-1}(x_{n-1}) K_n(x_{n-1}, x_n)} dx_{n-1:n} - 1. \tag{A.7}$$

In our special case, assuming that $\eta_1 = \pi_1$, we can straightforwardly evaluate each of the terms in this expression, obtaining for A.5:

$$\begin{aligned}
\int \frac{\tilde{\pi}_n(x_1)^2}{\eta_1(x_1)} dx_{n-1:n} - 1 &= \int \frac{\pi_2(x_1)^2}{\pi_1(x_1)} dx_1 - 1 \\
&= \frac{Z_1}{Z_2^2} \int \mathbb{I}_{T_2}(x_1) / \mathbb{I}_{T_1}(x_1) dx_1 - 1 \\
&= \frac{Z_1}{Z_2} - 1,
\end{aligned}$$

and for each term in the summation of A.6 we have:

$$\begin{aligned}
\int \frac{\left(\tilde{\pi}_k(x_k) L_{k-1}(x_k, x_{k-1})\right)^2}{\pi_{k-1}(x_{k-1}) K_k(x_{k-1}, x_k)} dx_{k-1:k} - 1 &= \int \frac{\left(\pi_{k+1}(x_k) \pi_k(x_{k-1})\right)^2}{\pi_{k-1}(x_{k-1}) \pi_k(x_k)} dx_{k-1:k} - 1 \\
&= \frac{Z_{k-1} Z_k}{Z_{k+1}^2 Z_k^2} \int \mathbb{I}_{T_{k+1}}(x_k) \mathbb{I}_{T_k}(x_{k-1}) dx_{k-1:k} - 1 \\
&= \frac{Z_{k-1} Z_k}{Z_{k+1} Z_k} - 1 = \frac{Z_{k-1}}{Z_{k+1}} - 1.
\end{aligned}$$

Finally, we obtain for A.7:

$$\begin{aligned}
\int \frac{\left(\pi_n(x_n) L_{n-1}(x_n, x_{n-1})\right)^2}{\pi_{n-1}(x_{n-1}) K_n(x_{n-1}, x_n)} dx_{n-1:n} - 1 &= \int \frac{\left(\pi_n(x_n) \pi_n(x_{n-1})\right)^2}{\pi_{n-1}(x_{n-1}) \pi_n(x_n)} dx_{n-1:n} - 1 \\
&= \int \frac{\pi_n(x_n) \pi_n(x_{n-1})^2}{\pi_{n-1}(x_{n-1})} dx_{n-1:n} - 1 \\
&= \frac{Z_{n-1}}{Z_n^2} \int \mathbb{I}_{T_n}(x_{n-1}) / \mathbb{I}_{T_{n-1}}(x_{n-1}) dx_{n-1} - 1 \\
&= \frac{Z_{n-1}}{Z_n} - 1.
\end{aligned}$$

Combining all of these, we see that the asymptotic variance can be expressed as:

$$\sigma^2_{SMC,n} = \frac{Z_1}{Z_2} + \sum_{k=2}^{n-1} \frac{Z_{k-1}}{Z_{k+1}} + \frac{Z_{n-1}}{Z_n} - n.$$

The variance of $\frac{\hat{Z}_n}{Z_1}$ can be obtained from that of $\log \frac{\hat{Z}_n}{Z_1}$ via the delta method:

$$\begin{aligned}
\mathrm{Var}\left(\frac{\hat{Z}_n}{Z_1}\right) &\approx \left[\frac{\partial \frac{\hat{Z}_n}{Z_1}}{\partial \log \frac{\hat{Z}_n}{Z_1}}\right]^2_{\log \frac{Z_n}{Z_1}} \mathrm{Var}\left(\log \frac{\hat{Z}_n}{Z_1}\right) \\
&= \left(\frac{Z_n}{Z_1}\right)^2 \frac{1}{N} \sigma^2_{SMC,n}.
\end{aligned}$$

The variance of crude Monte Carlo estimates obtained from $N \times n$ samples drawn from π_1 will be $Z_n/Z_1(1 - Z_n/Z_1)/Nn$.

It is convenient to further simplify things at this stage by assuming that $Z_{i-1}/Z_i = \eta$ with $\eta = \sqrt[n]{Z_1/Z_n}$. In this case,

$$\begin{aligned}\frac{\sqrt{\mathrm{Var}\left(\frac{\hat{Z}_n}{Z_1}\right)}}{Z_n/Z_1} &= \sqrt{\frac{\eta + \sum_{k=2}^{-1}\eta^2 + \eta - n}{N}} \\ &< \sqrt{\frac{n\left(\eta^2-1\right)}{N}} \\ &= \sqrt{\frac{n}{N}}\sqrt{\left(\frac{Z_1}{Z_n}\right)^{\frac{2}{n}} - 1} \\ &= \sqrt{\frac{n}{N}}\sqrt{\frac{Z_1}{Z_n}^{\frac{2}{n}}}\sqrt{1 - \frac{Z_n}{Z_1}^{2/n}} < \sqrt{\frac{n}{N}}\sqrt[n]{\frac{Z_1}{Z_n}}.\end{aligned}$$

If we consider the ratio of this bound to the variance of the crude Monte Carlo estimator, we obtain:

$$\frac{\sqrt{\frac{n}{N}}\sqrt[n]{\frac{Z_1}{Z_n}}}{Z_n/Z_1(1 - Z_n/Z_1)/Nn} = n\left(\frac{Z_1}{Z_n}\right)^{\frac{2-n}{2n}}\left(1 - \frac{Z_n}{Z_1}\right)^{-1/2}.$$

As shown in figure A.3, the SMC estimator dominates the naïve approach in terms of variance – excluding a very small number of configurations of the SMC system in which an unrealistic number (fewer than 3, say) intermediate distributions are employed.

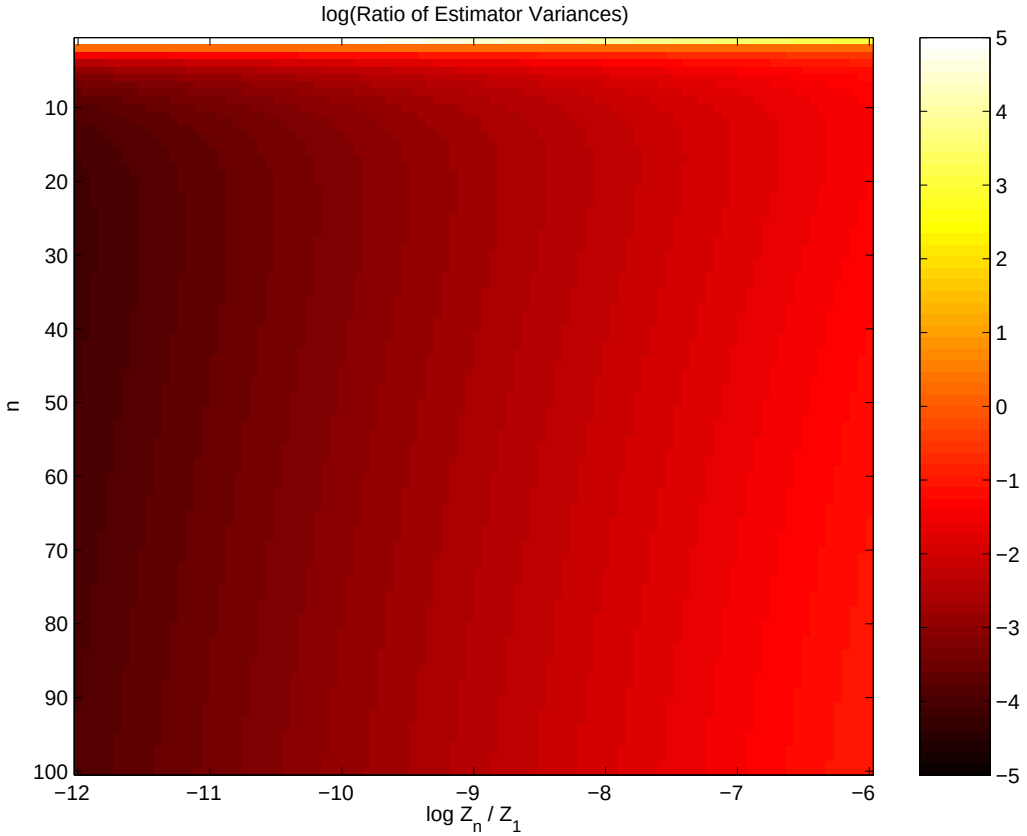


Fig. A.3. The log of the ratio of the variance bound of the SMC estimator to the variance of crude Monte Carlo for various numbers of intermediate distributions and event probabilities. As would be expected, for very rare events, with a reasonable number of intermediate distributions the method dramatically outperforms the crude Monte Carlo counterpart.

No. Even now I can't altogether believe that any of this has really happened...

Christopher Isherwood, "Goodbye to Berlin"