1

An Introduction to Sequential Monte Carlo Methods

Arnaud Doucet
Nando de Freitas
Neil Gordon

1.1 Motivation

Many real-world data analysis tasks involve estimating unknown quantities from some given observations. In most of these applications, prior knowledge about the phenomenon being modelled is available. This knowledge allows us to formulate Bayesian models, that is prior distributions for the unknown quantities and likelihood functions relating these quantities to the observations. Within this setting, all inference on the unknown quantities is based on the posterior distribution obtained from Bayes' theorem. Often, the observations arrive sequentially in time and one is interested in performing inference on-line. It is therefore necessary to update the posterior distribution as data become available. Examples include tracking an aircraft using radar measurements, estimating a digital communications signal using noisy measurements, or estimating the volatility of financial instruments using stock market data. Computational simplicity in the form of not having to store all the data might also be an additional motivating factor for sequential methods.

If the data are modelled by a linear Gaussian state-space model, it is possible to derive an exact analytical expression to compute the evolving sequence of posterior distributions. This recursion is the well known and widespread *Kalman filter*. If the data are modelled as a partially observed, finite state-space Markov chain, it is also possible to obtain an analytical solution, which is known as the hidden Markov model *HMM filter*. These two filters are the most ubiquitous and famous ones, yet there are a few other dynamic systems that admit finite dimensional filters (Vidoni 1999, West and Harrison 1997).

The aforementioned filters rely on various assumptions to ensure mathematical tractability. However, real data can be very complex, typically involving elements of non-Gaussianity, high dimensionality and nonlinearity, which conditions usually preclude analytic solution. This is a problem of fundamental importance that permeates most disciplines of science. According to the field of interest, the problem appears under many different names, including Bayesian filtering, optimal (nonlinear) filtering, stochastic filtering and on-line inference and learning. For over thirty years, many approximation schemes, such as the extended Kalman filter, Gaussian sum approximations and grid-based filters, have been proposed to surmount this problem. The first two methods fail to take into account all the salient statistical features of the processes under consideration, leading quite often to poor results. Grid-based filters, based on deterministic numerical integration methods, can lead to accurate results, but are difficult to implement and too computationally expensive to be of any practical use in high dimensions.

Sequential Monte Carlo (SMC) methods are a set of simulation-based methods which provide a convenient and attractive approach to computing the posterior distributions. Unlike grid-based methods, SMC methods are very flexible, easy to implement, parallelisable and applicable in very general settings. The advent of cheap and formidable computational power, in conjunction with some recent developments in applied statistics, engineering and probability, have stimulated many advancements in this field.

Over the last few years, there has been a proliferation of scientific papers on SMC methods and their applications. Several closely related algorithms, under the names of *bootstrap filters, condensation, particle filters, Monte Carlo filters, interacting particle approximations* and *survival of the fittest*, have appeared in several research fields. This book aims to bring together the main exponents of these algorithms with the goal of introducing the methods to a wider audience, presenting the latest algorithmic and theoretical developments and demonstrating their use in a wide range of complex practical applications. For lack of space, it has unfortunately not been possible to include all the leading researchers in the field, nor to address the theoretical and practical issues with the depth they deserve.

The chapters in the book are grouped in three parts. In the first part, a detailed theoretical treatment of various SMC algorithms is presented. The second part is mainly concerned with outlining various methods for improving the efficiency of the basic SMC algorithm. Finally, the third part discusses several applications in the areas of financial modelling and econometrics, target tracking and missile guidance, terrain navigation, computer vision, neural networks, time series analysis and forecasting, machine learning, robotics, industrial process control and population biology. In each of these parts, the chapters are arranged alphabetically by author.

The chapters are to a large extent self-contained and can be read independently. Yet, for completeness, we have added this introductory chapter to allow readers unfamiliar with the topic to understand the fundamentals and to be able to implement the basic algorithm. Here, we describe a general probabilistic model and the Bayesian inference objectives. After outlining the problems associated with the computation of the posterior distributions, we briefly mention standard approximation methods and point out some of their shortcomings. Subsequently, we introduce Monte Carlo methods, placing particular emphasis on describing the simplest – but still very useful – SMC method. This should enable the reader to start applying the basic algorithm in various contexts.

1.2 Problem statement

For sake of simplicity, we restrict ourselves here to signals modelled as Markovian, nonlinear, non-Gaussian state-space models, though SMC can be applied in a more general setting[1]. The unobserved signal (hidden states) $\{\mathbf{x}_t; t \in \mathbb{N}\}$, $\mathbf{x}_t \in \mathcal{X}$, is modelled as a *Markov process* of initial distribution $p(\mathbf{x}_0)$ and transition equation $p(\mathbf{x}_t | \mathbf{x}_{t-1})$. The observations $\{\mathbf{y}_t; t \in \mathbb{N}^*\}$, $\mathbf{y}_t \in \mathcal{Y}$, are assumed to be conditionally independent given the process $\{\mathbf{x}_t; t \in \mathbb{N}\}$ and of marginal distribution $p(\mathbf{y}_t | \mathbf{x}_t)$. To sum up, the model is described by

$$\begin{aligned} &p(\mathbf{x}_0) \\ &p(\mathbf{x}_t | \mathbf{x}_{t-1}) \quad \text{for } t \geq 1 \\ &p(\mathbf{y}_t | \mathbf{x}_t) \quad \text{for } t \geq 1. \end{aligned}$$

We denote by $\mathbf{x}_{0:t} \triangleq \{\mathbf{x}_0, ..., \mathbf{x}_t\}$ and $\mathbf{y}_{1:t} \triangleq \{\mathbf{y}_1, ..., \mathbf{y}_t\}$, respectively, the signal and the observations up to time t.

Our aim is to estimate recursively in time the *posterior distribution* $p(\mathbf{x}_{0:t} | \mathbf{y}_{1:t})$, its associated features (including the marginal distribution $p(\mathbf{x}_t | \mathbf{y}_{1:t})$, known as the *filtering distribution*), and the expectations

$$I(f_t) = \mathbb{E}_{p(\mathbf{x}_{0:t} | \mathbf{y}_{1:t})}[f_t(\mathbf{x}_{0:t})] \triangleq \int f_t(\mathbf{x}_{0:t}) \, p(\mathbf{x}_{0:t} | \mathbf{y}_{1:t}) \, d\mathbf{x}_{0:t}$$

for some function of interest $f_t : \mathcal{X}^{(t+1)} \to \mathbb{R}^{n_{f_t}}$ integrable with respect to $p(\mathbf{x}_{0:t} | \mathbf{y}_{1:t})$. Examples of appropriate functions include the conditional mean, in which case $f_t(\mathbf{x}_{0:t}) = \mathbf{x}_{0:t}$, or the conditional covariance of $\mathbf{x}_t$ where $f_t(\mathbf{x}_{0:t}) = \mathbf{x}_t \mathbf{x}_t^{\mathrm{T}} - \mathbb{E}_{p(\mathbf{x}_t | \mathbf{y}_{1:t})}[\mathbf{x}_t] \, \mathbb{E}^{\mathrm{T}}_{p(\mathbf{x}_t | \mathbf{y}_{1:t})}[\mathbf{x}_t]$.

[1]For simplicity, we use $\mathbf{x}_t$ to denote both the random variable and its realisation. Consequently, we express continuous probability distributions using $p(d\mathbf{x}_t)$ instead of $\Pr(\mathbf{X}_t \in d\mathbf{x}_t)$ and discrete distributions using $p(\mathbf{x}_t)$ instead of $\Pr(\mathbf{X}_t = \mathbf{x}_t)$. If these distributions admit densities with respect to an underlying measure μ (usually counting or Lebesgue), we denote these densities by $p(\mathbf{x}_t)$. To make the material accessible to a wider audience, we shall allow for a slight abuse of terminology by sometimes referring to $p(\mathbf{x}_t)$ as a distribution.

At any time t, the posterior distribution is given by *Bayes' theorem*

$$p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t})\, p(\mathbf{x}_{0:t})}{\int p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t})\, p(\mathbf{x}_{0:t})\, d\mathbf{x}_{0:t}}.$$

It is possible to obtain straightforwardly a recursive formula for this joint distribution $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$,

$$p(\mathbf{x}_{0:t+1}|\mathbf{y}_{1:t+1}) = p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) \frac{p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1})\, p(\mathbf{x}_{t+1}|\mathbf{x}_t)}{p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t})}. \quad (1.1)$$

The marginal distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ also satisfies the following recursion.

$$\textit{Prediction}: p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})\, p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\, d\mathbf{x}_{t-1}; \quad (1.2)$$

$$\textit{Updating: } p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)\, p(\mathbf{x}_t|\mathbf{y}_{1:t-1})}{\int p(\mathbf{y}_t|\mathbf{x}_t)\, p(\mathbf{x}_t|\mathbf{y}_{1:t-1})\, d\mathbf{x}_t}. \quad (1.3)$$

These expressions and recursions are deceptively simple because one cannot typically compute the normalising constant $p(\mathbf{y}_{1:t})$, the marginals of the posterior $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$, in particular $p(\mathbf{x}_t|\mathbf{y}_t)$, and $I(f_t)$ since they require the evaluation of complex high-dimensional integrals.

This is why, from the mid-1960s, a great many papers and books have been devoted to obtaining approximations for these distributions, including, as discussed in the previous section, the extended Kalman filter (Anderson and Moore 1979, Jazwinski 1970), the Gaussian sum filter (Sorenson and Alspach 1971) and grid-based methods (Bucy and Senne 1971). Other interesting work in automatic control was done during the 60s and 70s based on SMC integration methods (see (Handschin and Mayne 1969)). Most likely because of the modest computers available at the time, these last algorithms were overlooked and forgotten. In the late 1980s, the great increase of computational power made possible rapid advances in numerical integration methods for Bayesian filtering (Kitagawa 1987).

1.3 Monte Carlo methods

To address the problems described in the previous section, many scientific and engineering disciplines have recently devoted a considerable effort towards the study and development of *Monte Carlo* (MC) integration methods. These methods have the great advantage of not being subject to any linearity or Gaussianity constraints on the model, and they also have appealing convergence properties.

We start this section by showing that, when one has a large number of samples drawn from the required posterior distributions, it is not difficult to approximate the intractable integrals appearing in equations (1.1)-(1.3). It is, however, seldom possible to obtain samples from these distributions directly. One therefore has to resort to alternative MC methods, such as importance sampling. By making this general MC technique recursive, one obtains the *sequential importance sampling* (SIS) method. Unfortunately, it can easily be shown that SIS is guaranteed to fail as t increases. This problem can be surmounted by including an additional selection step. The introduction of this key step in (Gordon, Salmond and Smith 1993) led to the first operationally effective method. Since then, theoretical convergence results for this algorithm have been established. See, for example, (Del Moral 1996) and the chapters in this book by Crisan and Del Moral and Jacod.

1.3.1 Perfect Monte Carlo sampling

Let us assume that we are able to simulate N independent and identically distributed (i.i.d.) random samples, also named particles, $\left\{\mathbf{x}_{0:t}^{(i)}; i=1,...,N\right\}$ according to $p\left(\mathbf{x}_{0:t}\right|\mathbf{y}_{1:t})$. An empirical estimate of this distribution is given by

$$P_N\left(d\mathbf{x}_{0:t}\right|\mathbf{y}_{0:t}) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\mathbf{x}_{0:t}^{(i)}}\left(d\mathbf{x}_{0:t}\right),$$

where $\delta_{\mathbf{x}_{0:t}^{(i)}}\left(d\mathbf{x}_{0:t}\right)$ denotes the delta-Dirac mass located in $\mathbf{x}_{0:t}^{(i)}$. One obtains straightforwardly the following estimate of $I\left(f_t\right)$

$$I_N\left(f_t\right) = \int f_t\left(\mathbf{x}_{0:t}\right) P_N\left(d\mathbf{x}_{0:t}\right|\mathbf{y}_{1:t}) = \frac{1}{N}\sum_{i=1}^{N} f_t\left(\mathbf{x}_{0:t}^{(i)}\right).$$

This estimate is unbiased and, if the posterior variance of $f_t\left(\mathbf{x}_{0:t}\right)$ satisfies $\sigma_{f_t}^2 \triangleq \mathbb{E}_{p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})}\left[f_t^2\left(\mathbf{x}_{0:t}\right)\right] - I^2\left(f_t\right) < +\infty$, then the variance of $I_N\left(f_t\right)$ is equal to $var\left(I_N\left(f_t\right)\right) = \frac{\sigma_{f_t}^2}{N}$. Clearly, from the strong law of large numbers,

$$I_N\left(f_t\right) \underset{N\to+\infty}{\overset{a.s.}{\longrightarrow}} I\left(f_t\right),$$

where $\overset{a.s.}{\longrightarrow}$ denotes almost sure convergence. Moreover, if $\sigma_{f_t}^2 < +\infty$, then a central limit theorem holds

$$\sqrt{N}[I_N\left(f_t\right) - I\left(f_t\right)] \underset{N\to+\infty}{\Longrightarrow} \mathcal{N}\left(0, \sigma_{f_t}^2\right),$$

where $\Longrightarrow$ denotes convergence in distribution. The advantage of this perfect MC method is clear. From the set of random samples $\left\{\mathbf{x}_{0:t}^{(i)}; i=1,...,N\right\}$, one can easily estimate any quantity $I\left(f_t\right)$ and the rate of convergence of this estimate is *independent of the dimension of the integrand*. In contrast, any deterministic numerical integration method has a rate of convergence that decreases as the dimension of the integrand increases.

Unfortunately, it is usually impossible to sample efficiently from the posterior distribution $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$ at any time t, $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$ being multivariate, non-standard, and only known up to a proportionality constant. In applied statistics, Markov chain Monte Carlo (MCMC) methods are a popular approach to sampling from such complex probability distributions (Gilks, Richardson and Spiegelhalter 1996, Robert and Casella 1999). However, MCMC methods are iterative algorithms unsuited to recursive estimation problems. So, alternative methods have to be developed.

1.3.2 Importance sampling

Importance sampling

An alternative classical solution consists of using the *importance sampling* method, see for example (Geweke 1989). Let us introduce an arbitrary so-called *importance sampling distribution* (also often referred to as the proposal distribution or the importance function) $\pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$[2]. Assuming that we want to evaluate $I(f_t)$, and provided that the support of $\pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$ includes the support of $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$, we get the identity

$$I(f_t) = \frac{\int f_t(\mathbf{x}_{0:t})\, w(\mathbf{x}_{0:t})\, \pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})\, d\mathbf{x}_{0:t}}{\int w(\mathbf{x}_{0:t})\, \pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})\, d\mathbf{x}_{0:t}},$$

where $w(\mathbf{x}_{0:t})$ is known as the *importance weight* ,

$$w(\mathbf{x}_{0:t}) = \frac{p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})}{\pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})}.$$

Consequently, if one can simulate N i.i.d. particles $\left\{\mathbf{x}_{0:t}^{(i)}, i = 1, ..., N\right\}$ according to $\pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$, a possible Monte Carlo estimate of $I(f_t)$ is

$$\widehat{I}_N(f_t) = \frac{\frac{1}{N}\sum_{i=1}^{N} f_t\left(\mathbf{x}_{0:t}^{(i)}\right) w\left(\mathbf{x}_{0:t}^{(i)}\right)}{\frac{1}{N}\sum_{j=1}^{N} w\left(\mathbf{x}_{0:t}^{(j)}\right)} = \sum_{i=1}^{N} f_t\left(\mathbf{x}_{0:t}^{(i)}\right) \widetilde{w}_t^{(i)},$$

where the *normalised importance weights* $\widetilde{w}_t^{(i)}$ are given by

$$\widetilde{w}_t^{(i)} = \frac{w\left(\mathbf{x}_{0:t}^{(i)}\right)}{\sum_{j=1}^{N} w\left(\mathbf{x}_{0:t}^{(j)}\right)}. \tag{1.4}$$

For N finite, $\widehat{I}_N(f_t)$ is biased (ratio of two estimates) but asymptotically, under weak assumptions, the strong law of large numbers applies, that is, $\widehat{I}_N(f_t) \underset{N\to+\infty}{\overset{a.s.}{\longrightarrow}} I(f_t)$. Under additional assumptions, a central limit theorem with a convergence rate still independent of the dimension of the integrand

[2]We underline the (possible) dependence of $\pi(\cdot)$ on $\mathbf{y}_{1:t}$ by writing $\pi(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$.

can be obtained (Geweke 1989). It is clear that this integration method can also be interpreted as a sampling method where the posterior distribution $p(\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t})$ is approximated by

$$\widehat{P}_N\left(d\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t}\right) = \sum_{i=1}^{N} \widetilde{w}_t^{(i)} \delta_{\mathbf{x}_{0:t}^{(i)}}\left(d\mathbf{x}_{0:t}\right), \tag{1.5}$$

and $\widehat{I}_N(f_t)$ is nothing but the function $f_t(\mathbf{x}_{0:t})$ integrated with respect to the empirical measure $\widehat{P}_N(d\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t})$:

$$\widehat{I}_N(f_t) = \int f_t(\mathbf{x}_{0:t})\,\widehat{P}_N\left(d\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t}\right).$$

Importance sampling is a general Monte Carlo integration method. However, in its simplest form, it is not adequate for *recursive estimation.* That is, one needs to get all the data $\mathbf{y}_{1:t}$ before estimating $p(\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t})$. In general, each time new data $\mathbf{y}_{t+1}$ become available, one needs to recompute the importance weights over the entire state sequence. The computational complexity of this operation increases with time. In the following section, we present a strategy for overcoming this problem.

Sequential Importance Sampling

The importance sampling method can be modified so that it becomes possible to compute an estimate $\widehat{P}_N(d\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t})$ of $p(\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t})$ without modifying the past simulated trajectories $\left\{\mathbf{x}_{0:t-1}^{(i)}; i=1,...,N\right\}$. This means that the importance function $\pi(\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t})$ at time t admits as marginal distribution at time $t-1$ the importance function $\pi(\mathbf{x}_{0:t-1}|\,\mathbf{y}_{1:t-1})$, that is

$$\pi\left(\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t}\right) = \pi\left(\mathbf{x}_{0:t-1}|\,\mathbf{y}_{1:t-1}\right)\pi\left(\mathbf{x}_t|\,\mathbf{x}_{0:t-1},\mathbf{y}_{1:t}\right).$$

Iterating, one obtains

$$\pi\left(\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t}\right) = \pi\left(\mathbf{x}_0\right)\prod_{k=1}^{t}\pi\left(\mathbf{x}_k|\,\mathbf{x}_{0:k-1},\mathbf{y}_{1:k}\right).$$

It is easy to see that this importance function allows us to evaluate recursively in time the importance weights (1.4). Indeed, one has

$$\widetilde{w}_t^{(i)} \propto \widetilde{w}_{t-1}^{(i)} \frac{p\left(\mathbf{y}_t|\,\mathbf{x}_t^{(i)}\right)p\left(\mathbf{x}_t^{(i)}\middle|\,\mathbf{x}_{t-1}^{(i)}\right)}{\pi\left(\mathbf{x}_t^{(i)}\middle|\,\mathbf{x}_{0:t-1}^{(i)},\mathbf{y}_{1:t}\right)}. \tag{1.6}$$

An important particular case of this framework arises when we adopt the prior distribution as importance distribution

$$\pi\left(\mathbf{x}_{0:t}|\,\mathbf{y}_{1:t}\right) = p\left(\mathbf{x}_{0:t}\right) = p\left(\mathbf{x}_0\right)\prod_{k=1}^{t}p\left(\mathbf{x}_k|\,\mathbf{x}_{k-1}\right).$$

In this case, the importance weights satisfy $\widetilde{w}_t^{(i)} \propto \widetilde{w}_{t-1}^{(i)} p\left(\mathbf{y}_t \mid \mathbf{x}_t^{(i)}\right)$. In the following section, we restrict ourselves to the use of the prior distribution as importance sampling distribution. However, it is important to keep in mind that the method is far more general than this.

SIS is an attractive method, but it is nothing but a constrained version of importance sampling. Unfortunately, it is well known that importance sampling is usually inefficient in high-dimensional spaces (Gilks et al. 1996, Robert and Casella 1999). So, as t increases, this problem will arise in the SIS setting.

1.3.3 The Bootstrap filter

The problem encountered by the SIS method is that, as t increases, the distribution of the importance weights $\widetilde{w}_t^{(i)}$ becomes more and more skewed. Practically, after a few time steps, only one particle has a non-zero importance weight. The algorithm, consequently, fails to represent the posterior distributions of interest adequately. To avoid this degeneracy, one needs to introduce an additional selection step.

Notion

The key idea of the bootstrap filter is to eliminate the particles having low importance weights $\widetilde{w}_t^{(i)}$ and to multiply particles having high importance weights (Gordon et al. 1993). More formally, we replace the weighted empirical distribution $\widehat{P}_N\left(d\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right) = \sum_{i=1}^{N} \widetilde{w}_t^{(i)} \delta_{\mathbf{x}_{0:t}^{(i)}}\left(d\mathbf{x}_{0:t}\right)$ by the unweighted measure

$$P_N\left(d\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right) = \frac{1}{N} \sum_{i=1}^{N} N_t^{(i)} \delta_{\mathbf{x}_{0:t}^{(i)}}\left(d\mathbf{x}_{0:t}\right),$$

where $N_t^{(i)}$ is the number of offspring associated to particle $\mathbf{x}_{0:t}^{(i)}$; it is an integer number such that $\sum_{i=1}^{N} N_t^{(i)} = N$. If $N_t^{(j)} = 0$, then the particle $\mathbf{x}_{0:t}^{(j)}$ dies. The $N_t^{(i)}$ are chosen such that $P_N\left(d\mathbf{x}_{0:t} \mid \mathbf{y}_{0:t}\right)$ is close to $\widehat{P}_N\left(d\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right)$ in the sense that, for any function f_t,

$$\int f_t\left(\mathbf{x}_{0:t}\right) P_N\left(d\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right) \approx \int f_t\left(\mathbf{x}_{0:t}\right) \widehat{P}_N\left(d\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right). \tag{1.7}$$

After the selection step, the surviving particles $\mathbf{x}_{0:t}^{(i)}$, that is the ones with $N_t^{(i)} > 0$, are thus approximately distributed according to $p\left(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right)$. There are many different ways to select the $N_t^{(i)}$, the most popular being the one introduced in (Gordon et al. 1993). Here, one obtains the surviving particles by sampling N times from the (discrete) distribution $\widehat{P}_N\left(d\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right)$; this is equivalent to sampling the number of offspring $N_t^{(i)}$ according to a multinomial distribution of parameters $\widetilde{w}_t^{(i)}$. Equation (1.7) is satisfied in the sense that one can check easily that, for any bounded function f_t with $\|f_t\| = \sup_{\mathbf{x}_{0:t}} |f_t(\mathbf{x}_{0:t})|$, there exists C such that

$$\mathbb{E}\left[\left(\int f_t(\mathbf{x}_{0:t})\, P_N(d\mathbf{x}_{0:t}|\, \mathbf{y}_{1:t}) - \int f_t(\mathbf{x}_{0:t})\, \widehat{P}_N(d\mathbf{x}_{0:t}|\, \mathbf{y}_{1:t})\right)^2\right] \leq \frac{C\,\|f_t\|^2}{N}.$$

Algorithm description

We can now specify the algorithm in detail as follows.

Bootstrap Filter

1. *Initialisation*, $t = 0$.

- For $i = 1, ..., N$, sample $\mathbf{x}_0^{(i)} \sim p(\mathbf{x}_0)$ and set $t = 1$.

2. *Importance sampling step*

- For $i = 1, ..., N$, sample $\widetilde{\mathbf{x}}_t^{(i)} \sim p\left(\mathbf{x}_t|\, \mathbf{x}_{t-1}^{(i)}\right)$ and set $\widetilde{\mathbf{x}}_{0:t}^{(i)} = \left(\mathbf{x}_{0:t-1}^{(i)}, \widetilde{\mathbf{x}}_t^{(i)}\right)$.
- For $i = 1, ..., N$, evaluate the importance weights

$$\widetilde{w}_t^{(i)} = p\left(\mathbf{y}_t|\, \widetilde{\mathbf{x}}_t^{(i)}\right). \tag{1.8}$$

- Normalise the importance weights.

3. *Selection step*

- Resample with replacement N particles $\left(\mathbf{x}_{0:t}^{(i)}; i = 1, \ldots, N\right)$ from the set $\left(\widetilde{\mathbf{x}}_{0:t}^{(i)}; i = 1, \ldots, N\right)$ according to the importance weights.
- Set $t \leftarrow t + 1$ and go to step 2.

Note that in equation (1.8), $\widetilde{w}_{t-1}^{(i)}$ does not appear because the propagated particles $\mathbf{x}_{0:t-1}^{(i)}$ have uniform weights after the resampling step at time $t-1$. Also, we do not need to store the paths of the particles from time 0 to time t if we are only interested in estimating $p(\mathbf{x}_t|\, \mathbf{y}_{1:t})$. A graphic representation of the algorithm is shown in Figure 1.1.

The bootstrap filter has several attractive properties. Firstly, it is very quick and easy to implement. Secondly, it is to a large extent modular. That is, when changing the problem one need only change the expressions for the importance distribution and the importance weights in the code. Thirdly, it can be straightforwardly implemented on a parallel computer. Finally, the resampling step is a black box routine that only requires as inputs the

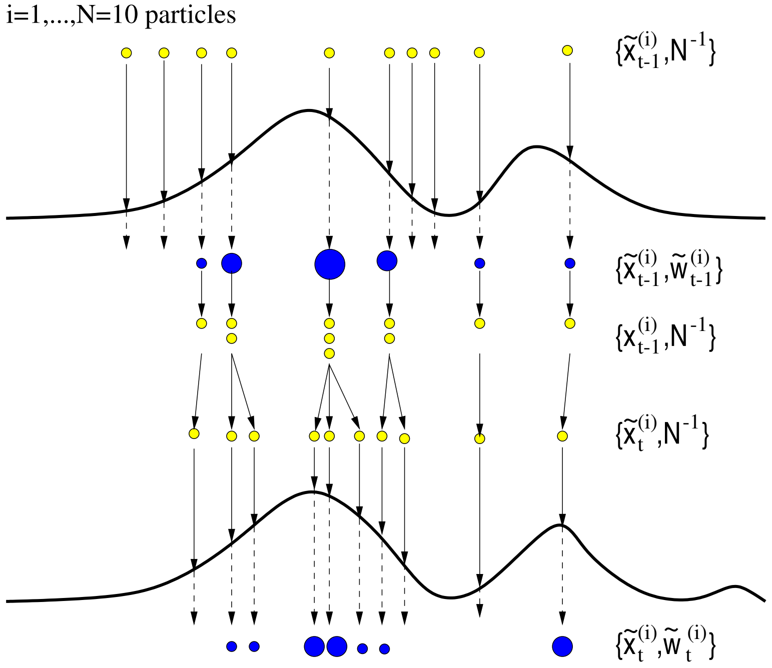

Figure 1.1. In this example, the bootstrap filter starts at time $t-1$ with an unweighted measure $\{\widetilde{\mathbf{x}}_{t-1}^{(i)}, N^{-1}\}$, which provides an approximation of $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-2})$. For each particle we compute the importance weights using the information at time $t-1$. This results in the weighted measure $\{\widetilde{\mathbf{x}}_{t-1}^{(i)}, \widetilde{w}_{t-1}^{(i)}\}$, which yields an approximation $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$. Subsequently, the resampling step selects only the fittest particles to obtain the unweighted measure $\{\widetilde{\mathbf{x}}_{t-1}^{(i)}, N^{-1}\}$, which is still an approximation of $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$. Finally, the sampling (prediction) step introduces variety, resulting in the measure $\{\widetilde{\mathbf{x}}_{t}^{(i)}, N^{-1}\}$, which is an approximation of $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$.

importance weights and indices (both being one-dimensional quantities). This enables one to easily carry out sequential inference for very complex models.

An illustrative example

For demonstration purposes, we apply the bootstrap algorithm to the following nonlinear, non-Gaussian model (Gordon et al. 1993, Kitagawa 1996):

$$\begin{aligned} x_t &= \frac{1}{2}x_{t-1} + 25\frac{x_{t-1}}{1+x_{t-1}^2} + 8\cos(1.2t) + v_t \\ y_t &= \frac{x_t^2}{20} + w_t, \end{aligned}$$

where $x_1 \sim \mathcal{N}\left(0, \sigma_1^2\right)$, v_t and w_t are mutually independent white Gaussian noises, $v_k \sim \mathcal{N}\left(0, \sigma_v^2\right)$ and $w_k \sim \mathcal{N}\left(0, \sigma_w^2\right)$ with $\sigma_1^2 = 10$, $\sigma_v^2 = 10$ and $\sigma_w^2 = 1$. The estimated filtering distributions are shown in Figure 1.2.

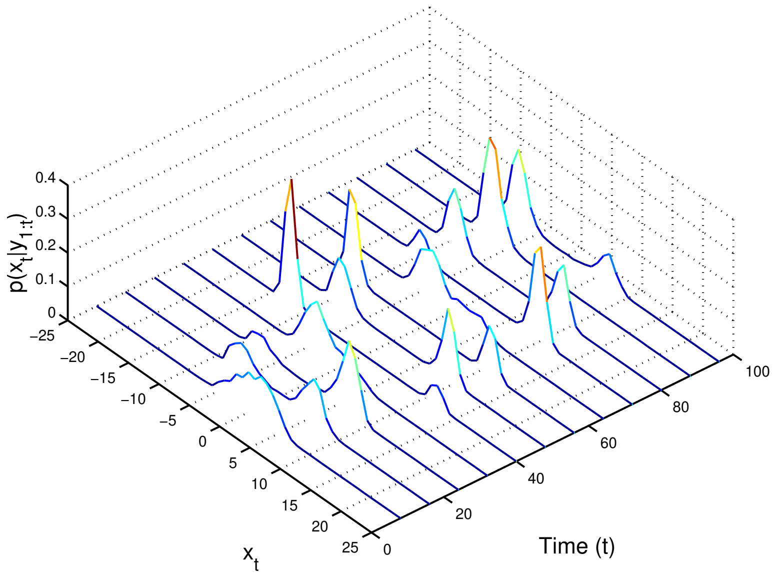

Figure 1.2. Estimated filtering distribution using 1000 particles.

Notice that for this model the minimum mean square estimates could be misleading because they do not provide enough information about the shape of the distribution. Indeed, one of the advantages of Monte Carlo methods is that they provide a complete description of the posterior distribution, not just a single point estimate.

1.4 Discussion

The aim of this chapter was to motivate the use of SMC methods to solve complex nonlinear, non-Gaussian on-line estimation problems. We also provided a brief introduction to SMC methods by describing one of the most basic particle filters. We hope we have convinced the reader of the enormous potential of SMC.

The algorithm we described is applicable to a very large class of models and is straightforward to implement. The price to pay for this simplicity is computationally inefficiency in some application domains.

The following chapters aim to develop and improve upon many of the ideas sketched here, and to propose new ones. Firstly, Chapters 2 and 3 provide a rigorous theoretical basis for SMC methods. Secondly, Chapters 4 to 14 describe numerous algorithmic developments which allow significant performance improvements over standard methods. Finally, Chapters 15 to 26 demonstrate the relevance of SMC to a wide range of complex practical applications.

The extended Kalman filter and related recursive sub-optimal estimation methods have been widely used for over 30 years in numerous applications. It is our belief that SMC methods can not only improve estimation performance in most of these applications, but also allow us to deal with more complex models that were out of reach a few years ago. We hope that this book constitutes a valuable and unified source of information on the topic.