**Uniwersytet Warszawski**
Wydział Matematyki, Informatyki i Mechaniki

**Agnieszka Borowska**

Nr albumu: 291512

# Sequential Monte Carlo: Selected topics

**Praca licencjacka**
**na kierunku MATEMATYKA**

Praca wykonana pod kierunkiem
**dra Pawła Wolffa**
Zakład Teorii Prawdopodobieństwa

Wrzesień 2015

## Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

## Oświadczenie autorki pracy

Świadoma odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązujacymi przepisami.
Oświadczam również, że przedstawiona praca nie byla wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.
Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autorki pracy

## Streszczenie

W pracy rozważany jest problem wnioskowania statystycznego o ukrytym sygnale rządzącym dynamiką danego systemu na podstawie obserwacji podlegających szumowi. System reprezentowany jest poprzez model zmiennych stanu z uwagi na szerokie spektrum problemów możliwych do analizy przy jego pomocy. Ponieważ dla badanego problemu w ogólności nie istnieje analityczne rozwiązanie, rozważamy klasę metod do przybliżania rozkładów a posteriori dla zmiennych stanów, tzw. Sekwencyjne Monte Carlo. Opierają się one na miarach Diraca konstruowanych na podstawie próbek losowych (cząsteczek) z rozkładów pochodzących z poprzednich iteracji. Szczególna uwaga poświęcona jest problemowi filtracji, polegającemu na estymacji bieżącego stanu systemu na podstawie bieżących pomiarów. Wyprowadzane są formuły charakteryzujące filtry cząsteczkowe, które służą do konstrukcji algorytmów przydatnych w analizie numerycznej. Omówiony jest problem degeneracji, immanentny dla sekwencyjnego losowania istotnego, jak również wybrane metody do radzenia sobie z nim. Zaprezentowane są podstawowe wyniki dotyczące zbieżności filtrów cząsteczkowych. Pracę zamyka cześć ilustrująca trzy zastosowania filtrów cząsteczkowych.



## Abstract

We analyse the problem of inference about a latent signal governing the dynamics of a system given only the observed noisy data. We adopt the discrete-time state space approach due to the wide range of problems it can capture. Because in general no closed-form solution are available in this framework, we discuss the class of methods used for approximating of the posterior state distributions, called Sequential Monte Carlo. These methods are based on the Dirac-measures which stem from the draws (particles) from the distribution constructed in the previous iteration. A special attention is devoted to the filtering problem, where one is interested in the estimation of the current state of the system given the current system measurements. We derive theoretical forms of the particle filters, which we then use to construct algorithms suitable for numerical analysis. We discuss the degeneracy problem, inherent to the sequential importance sampling and selected methods to tackle it. The basic convergence results in the context of particle filters are presents. Finally, we consider three numerical application.

## Słowa kluczowe

Sekwencyjne Monte Carlo; Filtry cząsteczkowe; Filtracja Bayesowska; Zbieżność miar; Modele przestrzeni stanów.

## Keywords

Sequential Monte Carlo; Particle filters; Bayesian filtering; Convergence of measures; State space models.

## Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.1 Matematyka

## Klasyfikacja tematyczna

60. Probability theory and stochastic processes
60G. Stochastic processes
60G35. Signal detection and filtering
62. Statistics
62F. Parametric inference
62F15. Bayesian inference
62M. Inference from stochastic processes

## Tytuł pracy w języku polskim

Sekwencyjne Metody Monte Carlo: Wybrane Zagadnienia

# Contents

# Chapter 1

# Introduction

The problem of inference about a latent signal governing the dynamics of the system under study, given only the observed noisy data, is ubiquitous in science. In many real-life applications, considered e.g. in applied statistics, engineering, physics or economics, this question is approached using the so called *state space models* (cf. Durbin and Koopman, 2012, for a detailed exposition in the context of time series analysis). This class of models provides a versatile tool to analyse time series due to its explicit focus on the state vector. This flexibility, however, comes in general at the price of the lack of a closed-form solution for these models. Hence, one needs to resort to simulation-based techniques.

The notion of *Sequential Monte Carlo* (SMC) refers to a class of simulation-based methods relying on the point mass (or particle) representation of probability distributions, which are used for (Bayesian) statistical inference for dynamic systems (cf. Doucet et al., 2001, for a comprehensive study). The key problem approached using SMC is the *optimal filtering problem*, consisting in the estimation of the current (unobserved) state of the system, based on the observations. Both, the state and the observation processes are stochastic, as in the Hidden Markov Models (HMM) framework (cf. Cappé et al., 2006). The aim is to determine the *posterior* distribution of the states, i.e. conditional on the observations. This is obtained by applying the Bayes Theorem, which allows for combining the *prior* beliefs about the state of the system, with the information coming from the data. The former may come from the computations in the previous iterations, while the latter is quantified using the likelihood function.

Importantly, in the SMC techniques the computations can be performed sequentially, considerably reducing the problem's dimensionality. This constitutes the crucial difference of the framework under consideration as compared to the standard, non-sequential Markov Chain Monte Carlo (MCMC) methods. To obtain draws from a distribution of interest, which in general is non-standard and difficult to sample from, one uses *importance sampling*, where the draws from a candidate distribution are assigned the so called *importance weights* to correct for their "fit" to the target. Unfortunately, the basic sequential setting suffers from two intrinsic problems related to the fact that often after just a few iterations only a very small part of the particles has non-negligible weights. This theoretically grounded phenomenon of the *weight degeneracy* (cf. Kong et al., 1994) has been devoted a noticeable attention in the literature (cf. Li et al., 2014, for a survey of recent approaches). The most straightforward solution it to perform *resampling* every iteration, so that the weights of the particles are reset.

In out thesis, the main emphasis in put on *particle filters*, with the aim to cover the most important aspects related to this subject. We derive their theoretical forms, which we then use to construct algorithms suitable for numerical analysis. A special attention is paid to the above-mentioned weight degeneracy problem: we provide a theoretical motivation for the occurrence of this phenomenon and we discuss some methods to tackle it. Next, we present the basic convergence results in the context of particle filters, as

they are essentially approximations to the nonlinear filtering equations. Since particle filters are based on interacting, hence statistically dependent, samples, one cannot show their convergence by referring to the standard Monte Carlo theorems. Finally, we consider three numerical application, which aim to demonstrate the theoretical results of previous sections. Throughout the thesis we adopt the state-space approach, where we restrict ourselves to the discrete-time setup.

The structure of this thesis is as follows. Chapter 2 specifies the modelling framework based on the state space models and introduces the problem of the optimal filtering. The main concepts of particle filtering and the related algorithms are presented in Chapter 3. In Chapter 4 we define the convergence concepts for measure-valued random variables and discusses the basic convergence results for particle filters. Chapter 5 presents three applications of particle filers to different types of models, starting with the most basic one, the *local level model*, to finally move to the nonlinear non-Gaussian *stochastic volatility model*. Chapter 6 concludes and presents the topics for further research. The notation used within the thesis is shown in Appendix A.

# Chapter 2

# Model specification and the filtering problem

In this Chapter we establish the general framework for this thesis and formulate the basic filtering problem. Below, we consider a general *dynamic model*, which is a mathematical representation of time-varying phenomena in different fields of science and engineering. The fundamental components of such a dynamic model are the *dynamics model* and the *measurement model*. The former describes the evolution of a *state* process, which itself is latent but affects the *measurement* process in an observed way, characterised by the measurement model. These types of models are ubiquitous in signal processing and related fields. Moreover, very often one restricts attention to the class of the so called *state space models*, where the state is assumed to follow the first order Markov process and the whole system can be described for instance by a set of first-order differential or difference equations (rather than by a set of $n$-th-order ones). Therefore, we first formally characterise the general state space model.

As it has been already pointed out in the Introduction, we aim to perform inference on the unknown state process within a Bayesian approach, i.e. to combine the prior knowledge of the system with the information delivered by the data (measurement). In the most general setting one is interested in the whole *posterior* distribution of the state process, given the data. However, a very common practise is limiting of attention to the *filtering problem*, which consists in the recursive estimation of the most recent latent state (or a measurable function thereof). This problem is also know as *Bayesian* or *optimal filtering* problem and it will be the subject this thesis mainly focuses on.

Two framework conventions will be used in this and in the subsequent chapters, depending on the current purposes, which is a common practice in the literature (cf. Cappé et al. (2006), Doucet et al. (2001), Crisan and Doucet (2002), Durbin and Koopman (2012)). When introducing the general state space model and its components in Section 2.1, as well as when analysing the convergence of particle filters in Chapter 4, we will adopt a probabilistic representation, more appropriate for a theoretical analysis, with the notation based mainly on Crisan and Doucet (2002) and Crisan (2001). To formulate the Bayesian filtering problem in Sections 2.2.1 and 2.2.2 and to discuss the particle filters in Chapter 3, we will follow a signal processing convention, more suitable for the application purposes, following e.g. Doucet and Johansen (2009) and Kong et al. (1994).

## 2.1 General state space model

We consider the latent state (or signal) process $X = \{X_t\}_{t\in\mathbb{N}}$ and the measurement (or observation) process $Y = \{Y_t\}_{t\in\mathbb{N}}$ (a more detailed exposition of these processes is given in the next two Subsections).

The state process together with the observation process constitute the following general *state space model* (SSM)

$$X_t = F_t(X_{t-1}, V_t), \tag{2.1.1}$$

$$Y_t = G_t(X_t, W_t), \tag{2.1.2}$$

$$X_0 \sim p(dx_0). \tag{2.1.3}$$

In the above specification $F_t$ and $G_t$ are known, possibly nonlinear, measurable functions of the state and of the white noise processes $W = \{W_t\}_{t\in\mathbb{N}}$ and $V = \{V_t\}_{t\in\mathbb{N}}$, and $p(dx_0)$ is the initial distribution of the signal process. The disturbance processes are assumed to be mutually independent and independent of $X_0$. In general, one does not make any additional distributional assumptions on both noise processes, hence the introduced framework is often referred to as *nonlinear, non-Gaussian* SSM. Moreover, we assume that all the stochastic processes are defined on $(\Omega, \mathcal{F}, \mathbb{P})$, which is a fixed probability space.

Since below we describe the signal and measurement processes in more detail, here let us just briefly comment on the above specification. The transition equation (2.1.1) defines the first-order Markov process $X$ with the equivalent probabilistic description of the signal evolution being $p(dx_{t+1}|x_t)$, i.e. the transition distribution. The observation equation (2.1.2) characterises the likelihood of the observations $Y$, given the state $X$, so it relates the recorded measurement to the latent state vector. An equivalent probabilistic model for $Y$ is $p(dy_t|x_t)$, i.e. the observation distribution. Figure 2.1.1 graphically presents the dependencies between states and observations in the general SMM.

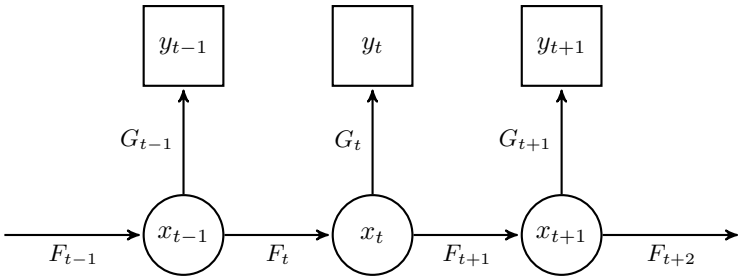



Figure 2.1.1: A graphical representation of the general SSM.

The SMM is a class which includes many models of interest and provides a very flexible tool to model many real-life problems. It is because nonlinear dynamic systems frequently arise in science, engineering, economics and a number of other fields, while non-Gaussian distributions are either natural or practical in applications. Finally, it is worth noting that in many fields the generic SSM is also known as the *hidden Markov model* (HMM, cf. Cappé et al. (2006)).

### 2.1.1 The signal process

As it has already been pointed out, the transition equation (2.1.1) defines the first-order Markov signal process $X = \{X_t\}_{t\in\mathbb{N}}$ which is a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{R}^{n_x}$. The $\sigma$-algebra generated by $X$ is denoted by $\mathcal{F}_t^X$,

$$\mathcal{F}_t^X = \sigma\left(X_s,\, s \in \{0, 1, \ldots, t\}\right).$$

Then, the Markovianity assumption on $X$ means that $\forall t$

$$\mathbb{P}\left(X_t \in A | \mathcal{F}_{t-1}^X\right) = \mathbb{P}(X_t \in A | X_{t-1}) \quad \text{a.s.}, \quad \forall A \in \mathcal{B}(\mathbb{R}^{n_x}).$$

The *transition kernel* $K_t(\cdot,\cdot)$ of the Markov chain $X$ is a function on $\mathbb{R}^{n_x} \times \mathcal{B}(\mathbb{R}^{n_x})$ satisfying

$$K_t(x, A) = \mathbb{P}(X_t \in A | X_{t-1} = x), \quad \forall t \in \mathbb{N}, \forall x \in \mathbb{R}^{n_x}. \tag{2.1.4}$$

For a fixed $x \in \mathbb{R}^{n_x}$ and $t \in \mathbb{N}$, $K_t(x,\cdot)$ is a probability measure on $\mathbb{R}^d$; for a fixed $A \in \mathcal{B}(\mathbb{R}^{n_x})$ and $t \in \mathbb{N}$, $K_t(\cdot, A)$ is a bounded, $\mathcal{B}(\mathbb{R}^{n_x})$-measurable function. In the further part of the work, we will restrict ourselves to *Feller transition kernels*, i.e. transition kernels satisfying the *Feller property*:

$$f \in \mathcal{C}_b(\mathbb{R}^d) \quad \Rightarrow \quad K_t f \in \mathcal{C}_b(\mathbb{R}^d),$$

where we define the *function* $K_t f$ as follows

$$K_t f(x) := \int_{\mathbb{R}^{n_x}} f(y) K_t(x, dy).$$

The transition kernel together with the initial distribution $p(dx_0)$ of a Markov process $X$ uniquely determine its distribution.

Next, for a fixed $t$, consider the distribution of a random variable $X_t$, denoted by $p_t(dx_t)$ and given by

$$p_t(A) := \mathbb{P}(X_t \in A), \qquad \forall A \in \mathcal{B}(\mathbb{R}^{n_x}).$$

This distribution is often referred to as the *prior* distribution for $X_t$, because it depends only on $X_t$[1]. Making use of the transition kernel formula (2.1.4), one can recursively express $p_t$ as

$$p_t = p_{t-1} K_{t-1} = p_0 \prod_{i=0}^{t-1} K_i,$$

where for a $\mathbb{R}^{n_x}$-valued distribution $q_t$ on $(\Omega, \mathcal{F}, \mathbb{P})$ we define the measure $q_{t-1}K_{t-1}$ as

$$(q_{t-1} K_t)(A) := \int_{\mathbb{R}^{n_x}} K_t(x, A) q_{t-1}(dx).$$

For simplicity, we assume that $K_t(x_{t-1}, dx_t)$ and $p(dx_0)$ admit densities with respect to the corresponding Lebesgue measures, i.e.

$$\begin{aligned} K_t(x_{t-1}, dx_t) &= \mathbb{P}(X_t \in dx_t | X_{t-1} = x_{t-1}) = p(x_t | x_{t-1}) dx_t, \\ p(dx_0) &= \mathbb{P}(X_0 \in dx_0) = p(x_0) dx_0. \end{aligned}$$

### 2.1.2 The observation process

The observation process $Y = \{Y_t\}_{t \in \mathbb{N}}$, characterised by (2.1.2), is a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{R}^{n_y}$, and it is assumed to be *conditionally independent* of the signal $X$ (cf. Figure 2.1.1).

**Definition 2.1.1** (Conditional independence of stochastic processes)**.** *Let $Z = \{Z_t\}_{t \in \mathbb{N}}$ and $U = \{U_t\}_{t \in \mathbb{N}}$ be stochastic processes on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $\mathbb{R}^{n_z}$ and $\mathbb{R}^{n_u}$, respectively. Next, let $\left\{\mathcal{F}_t^Z\right\}_{t \in \mathbb{N}}$, $\left\{\mathcal{F}_t^U\right\}_{t \in \mathbb{N}}$ be the filtrations generated by $Z$ and $U$, respectively, with $\mathcal{F}_t^Z = \sigma\left(Z_s : s \in [0,t]\right)$ and $\mathcal{F}_t^U = \sigma\left(U_s : s \in [0,t]\right)$. We say that $Z$ is conditionally independent of $U$ if*

$$\mathbb{P}(Z_t \in B | \sigma(\mathcal{F}_s^Z, \mathcal{F}_t^U)) = \mathbb{P}(Z_t \in B | U_t), \qquad \forall s < t,\ t, s \in \mathbb{N}, \quad \forall B \in \mathcal{B}(\mathbb{R}^{n_z}).$$

[1] For this reason it can be seen as the *marginal* distribution of the *joint* distribution of the process $X$ up to $t$, as it will be formally specified in Subsections 2.2.1 and 2.2.2.

This means that for a given $t$ the distribution of $Y_t$ depends on $X_t$ only, so that its marginal distribution can be expressed as

$$\mathbb{P}(Y_t \in B | X_t = x_t) = \int_B p(dy_t|x_t), \qquad B \in \mathcal{B}(\mathbb{R}^{d_y}).$$

Notice, that in the parlance of the general SSM, the conditional independence of $Y$ given $X$ amounts to the noise processes $V$ and $W$ in (2.1.1) and (2.1.2) being serially and mutually independent[2]. Similarly as for the signal process, we assume for simplicity that $p(dy_t|x_t)$ admits a density with respect to the corresponding Lebesgue measure, i.e.

$$p(dy_t|x_t) = p(y_t|x_t)dy_t$$

and that this density is bounded and continuous. Finally, we put $Y_0 = 0$.

### 2.1.3 Additive example

To illustrate the link between the probabilistic and signal processing approaches, following Crisan and Doucet (2002), consider a scalar dynamic model

$$X_t = f(X_{t-1}) + V_t,$$
$$Y_t = g(X_t) + W_t,$$

where $V = \{V_t\}_{t\in\mathbb{N}}$ and $W = \{W_t\}_{t\in\mathbb{N}\setminus\{0\}}$ are both independent and identically distributed (i.i.d.) sequences, also mutually independent. This is a special case of the general SSM (2.1.1)-(2.1.3), where the noises additively affect the signal and observations. We take

$$\mathbb{P}(V_t \in C) = \int_C p_V(dv) = \int_C p_V dv, \qquad \forall C \in \mathcal{B}(\mathbb{R}),$$
$$\mathbb{P}(W_t \in D) = \int_D p_W(dw) = \int_D p_W dw, \qquad \forall D \in \mathcal{B}(\mathbb{R}),$$

which yields

$$K(x_{t-1}, x_t) = p_V(x_t - f(x_{t-1})),$$
$$p(y_t|x_t) = p_W(y_t - g(x_t)).$$

## 2.2 Problem formulation

The key feature of the SSM representation is its recursive structure, which makes this class of dynamic models perfectly suited for the sequential (on-line) analysis. Given the current knowledge about the system, one may be interested in updating this knowledge when new information (observations) become available. This is a natural framework for the Bayesian analysis, where the *prior* distribution of the unknown variables is combined with the *likelihood* function, describing the probability of the data given the hidden variables, to yield the *posterior* distribution for the latent variables. Below, in Subsection we derive the recursive formula for the posterior distribution, which is additionally split into two steps: *prediction* and *updating*.

[2] An alternative, yet similar definition of conditional independence is provided by Cappé et al. (2006) within the HMM framework. First, they define the HMM as a bivariate discrete time process $\{X_t, Y_t\}_{t\in\mathbb{N}}$ where $\{X_t\}_{t\in\mathbb{N}}$ is a Markov chain, and, conditional on $\{X_t\}_{t\in\mathbb{N}}$, $\{Y_t\}_{t\in\mathbb{N}}$ is a sequence of independent variables such that the conditional distribution of $Y_t$ depends only on $X_t$.

### 2.2.1 Bayesian recursion

Denote by $X_{0:t} = \{X_0, \dots, X_t\}$ and by $Y_{0:t} = \{Y_0, \dots, Y_t\}$ the path up to time $t$ of the signal and of the observation process, respectively, with the corresponding realisation up to time $t$ denoted by $x_{0:t} = \{x_0, \dots, x_t\}$ and by $y_{0:t} = \{y_0, \dots, y_t\}$. We wish to determine the *posterior* probability distribution

$$p(dx_{0:t}|y_{0:t}) := \mathbb{P}(X_{0:t} \in dx_{0:t}|Y_{0:t} = y_{0:t}),$$

which describes the probability of all the transitions over time given all the observations, up to time $t$. Since we have assumed that the required distributions admit densities, from now on we will write

$$p(dx_{0:t}|y_{0:t}) = p(x_{0:t}|y_{0:t})dx_{0:t}. \tag{2.2.1}$$

Then, the Bayes' theorem yields

$$p(x_{0:t}|y_{0:t}) = \frac{p(y_{0:t}|x_{0:t})p(x_{0:t})}{\int_{(\mathbb{R}^{n_x})^{t+1}} p(y_{0:t}|x_{0:t})p(x_{0:t})dx_{0:t}}. \tag{2.2.2}$$

The aim is to derive a *recursive* formula for (2.2.2), i.e. to express $p(x_{0:t}|y_{0:t})$ in the following way

$$p(x_{0:t}|y_{0:t}) = f_t\left(p(x_{0:t-1}|y_{0:t-1}), y_t\right), \qquad \forall t,$$

where $f_t$ is a function to be determined, possibly time-dependent. In other words, we want to propagate over time the joint and the marginal distributions $p(x_{0:t}|y_{0:t})$ and $p(x_t|y_{0:t})$. We have

$$\begin{aligned}
p(x_{0:t}|y_{0:t}) &= \frac{p(y_{0:t}|x_{0:t})p(x_{0:t})}{p(y_{0:t})} \\
&= \frac{p(y_t, y_{0:t-1}|x_{0:t})p(x_{0:t})}{p(y_t, y_{0:t-1})} \\
&\overset{(*)}{=} \frac{p(y_t|y_{0:t-1}, x_{0:t})p(y_{0:t-1}|x_{0:t})p(x_{0:t})}{p(y_t|y_{0:t-1})p(y_{0:t-1})} \\
&\overset{B}{=} \frac{p(y_t|y_{0:t-1}, x_{0:t})p(x_{0:t}|y_{0:t-1})p(y_{0:t-1})p(x_{0:t})}{p(y_t|y_{0:t-1})p(y_{0:t-1})p(x_{0:t})} \\
&= \frac{p(y_t|y_{0:t-1}, x_{0:t})p(x_{0:t}|y_{0:t-1})}{p(y_t|y_{0:t-1})} \\
&\overset{OI}{=} \frac{p(y_t|x_t)p(x_t, x_{0:t}|y_{0:t-1})}{p(y_t|y_{0:t-1})} \\
&\overset{(*)}{=} \frac{p(y_t|x_t)p(x_t|x_{0:t-1}, y_{0:t-1})p(x_{0:t-1}|y_{0:t-1})}{p(y_t|y_{0:t-1})} \\
&\overset{MC}{=} \frac{p(y_t|x_t)p(x_t|x_{t-1})p(x_{0:t-1}|y_{0:t-1})}{p(y_t|y_{0:t-1})},
\end{aligned}$$

where B denotes application of Bayes' law, OI refers to observations independence and MC to the Markov nature of the state process. With $(*)$ we denoted the following straightforward identity

$$p(x, y|z) = p(x|y, z)p(y|z). \tag{*}$$

Hence, finally, we arrive at the recursive formula for the posterior

$$p(x_{0:t}|y_{0:t}) = p(x_{0:t-1}|y_{0:t-1})\frac{p(y_t|x_t)p(x_t|x_{t-1})}{p(y_t|y_{0:t-1})}. \tag{2.2.3}$$

The computation of (2.2.3) can be split into two subsequent steps: prediction and updating, which is of

a significant practical importance. The former consists in deriving $p(x_{0:t}|y_{0:t-1})$, i.e. estimation of the posterior distribution of $X$ up to time $t$ given only past observations $y_{0:t-1}$, i.e. without the most recent realisation of $Y_t$. The latter refers to re-evaluation of our beliefs regarding the distribution of $X_{0:t}$ after having observed $y_t$, the recent realisation of $Y_t$. The recursion steps are as follows

$$p(x_{0:t}|y_{0:t-1}) = p(x_{0:t-1}|y_{0:t-1})p(x_t|x_{t-1}), \tag{2.2.4}$$
$$p(x_{0:t}|y_{0:t}) = \tilde{C}_t^{-1} p(y_t|x_t)p(x_{0:t}|y_{0:t-1}), \tag{2.2.5}$$

where $\tilde{C}_t$ is a normalising constant, to ensure that (2.2.5) is a proper density function, equal to

$$\begin{aligned}\tilde{C}_t &= \int_{(\mathbb{R}^{n_x})^{t+1}} p(y_t|x_t)p(x_{0:t}|y_{0:t-1})dx_{0:t} \\ &= p(y_t|y_{0:t-1}).\end{aligned}$$

One can see that the recursive system (2.2.4) and (2.2.5) has a straightforward Bayesian interpretation. The predictive distribution becomes the prior distribution in the updating equation, which combined with the likelihood of the new observation delivers the posterior distribution.

### 2.2.2 The filtering problem

One is often interested not exactly in $p(x_{0:t}|y_{0:t})$ itself, but rather in its marginal distribution $p(x_t|y_{0:t})$, referred to as *filtering distribution* and given by

$$p(x_t|y_{0:t}) = \int_{(\mathbb{R}^{n_x})^t} p(x_{0:t}|y_{0:t})dx_{0:t-1}. \tag{2.2.6}$$

The filtering distribution describes the current state of the system given all the observations up to now. In other words, the filtering problem consists of computing the conditional distribution of the signal given the $\sigma$-algebra generated by the observation process from time 0 to the current time $t$.

Then, the filtering counterparts of the Bayesian recursion (2.2.4) and (2.2.5) are given by

$$p(x_t|y_{0:t-1}) = p(x_{t-1}|y_{0:t-1})p(x_t|x_{t-1}), \tag{2.2.7}$$
$$p(x_t|y_{0:t}) = C_t^{-1} p(y_t|x_t)p(x_t|y_{0:t-1}), \tag{2.2.8}$$

with the normalising constant

$$\begin{aligned}C_t &= \int_{\mathbb{R}^{n_x}} p(y_t|x_t)p(x_t|y_{0:t-1})dx_t \\ &= p(y_t|y_{0:t-1}).\end{aligned}$$

The simplicity of the above formulae is misleading, as these densities are usually unavailable in closed form. A noticeable exception is the linear Gaussian case, which admits the analytical solution in the form of the celebrated *Kalman filter* (cf. e.g. Durbin and Koopman (2012)). For completeness, we recall it briefly in the next subsection.

### 2.2.3 Linear Gaussian example

Consider the linear Gaussian SSM, given by

$$\begin{aligned} X_{t+1} &= T_t X_t + R_t V_t, \qquad & V_t &\overset{i.i.d}{\sim} \mathcal{N}(0, H_t), \\ Y_t &= Z_t X_t + W_t, \qquad & W_t &\overset{i.i.d}{\sim} \mathcal{N}(0, Q_t), \\ X_1 &\sim \mathcal{N}(\bar{X}_{1|0}, P_{1|0}), \end{aligned}$$

where $V_t$, $W_s$ are independent for all $t, s$, and independent from $X_1$ (cf. Durbin and Koopman, 2012). The system matrices $T_t, Z_t, R_t, Q_t, H_t$ are deterministic and have known forms.

Then the unobserved state $X_t$ can be recursively estimated from the observations with the Kalman filter as follows

$$\begin{aligned} \nu_t &= Y_t - Z_t X_t, \\ F_t &= Z_t P_t Z_t^T + H_t, \\ K_t &= T_t P_t Z_t^T F_t^{-1}, \\ \bar{X}_{t+1} &= T_t \bar{X}_t + K_t \nu_t, \\ P_{t+1} &= T_t P_t T_t^T + R_t Q_t R_t^T - K_t F_t K_t^T, \end{aligned}$$

for $t = 1, \ldots, n$, and starting with given values for $\bar{X}_1$ and $P_1$. The derivation of the Kalman filter is based on the lemma for multivariate normal regression theory. The proof can be found in numerous sources, e.g. Durbin and Koopman (2012), and will be omitted as not directly related to the main subject of this thesis.

In the above system $\bar{X}_{t+1|t}$ and $P_{t+1|t}$ characterise the optimal state *predictions*, i.e. the conditional expectation and variance, respectively, of the random variable $X_{t+1}$, given the observations up to $t$, i.e.

$$\begin{aligned} \bar{X}_{t+1|t} &= \mathbb{E}[X_{t+1}|Y_t], \\ P_{t+1|t} &= \mathrm{Var}[X_{t+1}|Y_t]. \end{aligned}$$

The *filtered* estimates, which are defined as

$$\begin{aligned} \bar{X}_{t|t} &= \mathbb{E}[X_t|Y_t], \\ P_{t|t} &= \mathrm{Var}[X_t|Y_t], \end{aligned}$$

are then given by

$$\begin{aligned} \bar{X}_{t|t} &= \bar{X}_{t|t-1} + M_t \nu_t, \\ P_{t|t} &= P_{t|t-1} - M_t F_t M_t^T, \end{aligned}$$

where $M_t = P_{t|t-1} Z_t^T F_t^{-1}$. Notice, that because the model under consideration is linear Gaussian, the distribution of the latent state is fully specified by its first two moments. Therefore, $\bar{X}_{t|t}$ and $P_{t|t}$ describe what one is truly interested in, i.e. the *filtering distribution* of $X_t$

$$p(x_t|y_{0:t}) = \phi(x_t; \bar{X}_{t|t}, P_{t|t}).$$

# Chapter 3

# Particle filtering

Particle filters are sequential Monte Carlo integration techniques based on *particle*, i.e. point mass, representation of probability densities. They are used for on-line, or sequential, inference problems and can be applied to general state-space model, as they are not based on any assumptions regarding the functional form of the model (e.g. linear) nor require disturbances to follow any particular distribution (e.g. Gaussian).

## 3.1 Monte Carlo integration

For further reference let us first shortly recall the basic *crude Monte Carlo* integration method before describing in more detail the importance sampling and sequential importance sampling techniques. Suppose we are interested in estimation of the (conditional) mean of an arbitrary measurable function $f_t : \mathcal{X}^{t+1} \to \mathbb{R}$, given by

$$\begin{aligned} \bar{f}_t &= \mathbb{E}\left[f_t(X_{0:t})|Y_{0:t}\right] \\ &= \int f_t(x_{0:t})p(x_{0:t}|y_{0:t})dx_{0:t}, \end{aligned} \tag{3.1.1}$$

where the conditional distribution $p(x_{0:t}|y_{0:t})$ is given by (2.2.2). To estimate (3.1.1), we can generate independently $N$ random sequences $X_{0:t}^{(1)}, \ldots, X_{0:t}^{(N)} \overset{iid}{\sim} p(x_{0:t}|y_{0:t})$ and consider the sample average

$$\hat{f}_t^{MC} = \frac{1}{N}\sum_{i=1}^{N} f(X_{0:t}^{(i)}). \tag{3.1.2}$$

Indeed, by the Strong Law of Large Numbers, (3.1.2) is *strongly consistent*, i.e.

$$\hat{f}_t^{MC} \underset{N\to\infty}{\overset{a.s.}{\to}} f_t.$$

Moreover, by the Central Limit Theorem, the MC estimator is asymptotically normal since

$$\begin{aligned} \sqrt{N}\left(\hat{f}_t^{MC} - \bar{f}_t\right) =& \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(f_t(X_{0:n}^{(i)}) - \bar{f}_t\right) \\ &\overset{n\to\infty}{\to} N\left(0, \mathrm{Var}_p f(X)\right), \end{aligned}$$

provided that the variance $\mathrm{Var}_p f(X)$ exists.

## 3.2 Importance sampling

Since the basic building-block of particle filtering techniques is importance sampling, we first recall this algorithm, before discussing its recursive version, the *sequential* importance sampling. Importance Sampling is an effective Monte Carlo integration algorithm, used to estimate (3.1.1) in practice, since it is difficult to sample from $p(x_{0:t}|y_{0:t})$. Therefore, one usually resorts to drawing from the so called *importance density* $q(x_{0:t}|y_{0:t})$, with the support including the one of the state posterior. It is assumed the sampling from $q(x_{0:t}|y_{0:t})$ is relatively easy and inexpensive.

To start with, notice that we can then express (3.1.1) using $q(x_{0:t}|y_{0:t})$ in the following way

$$\begin{aligned}\bar{f}_t &= \int f_t(x_{0:t})\frac{p(x_{0:t}|y_{0:t})}{q(x_{0:t}|y_{0:t})}q(x_{0:t}|y_{0:t})dx_{0:t} \\ &= \mathbb{E}_q\left[f_t(X_{0:t})\frac{p(X_{0:t}|Y_{0:t})}{q(X_{0:t}|Y_{0:t})}\right] \\ &= \mathbb{E}_q\left[f_t(X_{0:t})\tilde{W}_t\right], \end{aligned} \tag{3.2.1}$$

where $\mathbb{E}_q$ stands for expectation with respect to density $q$ and

$$\tilde{W}_t(x_{0:t}) = \frac{p(x_{0:t}|y_{0:t})}{q(x_{0:t}|y_{0:t})} \tag{3.2.2}$$

is known as the *importance weight* function. Notice, that since the importance weight function is defined as the likelihood ratio, it is the Radon-Nikodým derivative of the true distribution $p(\cdot)$ with respect to the importance distribution $q(\cdot)$. Generally, it depends on $x_{0:t}$ and $y_{0:t}$, however, in the remaining part of the work we skip the arguments for notational convenience. Hence, with some abuse of notation, we will use the same symbols to denote the weight functions as functions of stochastic processes and of real vectors.

Since the joint density factorises as follows

$$p(x_{0:t}, y_{0:t}) = p(x_{0:t}|y_{0:t})p(y_{0:t}),$$

one can express (3.2.2) as

$$\begin{aligned}\tilde{W}_t &= \frac{1}{p(y_{0:t})}\frac{p(x_{0:t}, y_{0:t})}{q(x_{0:t}|y_{0:t})} \\ &= \frac{1}{p(y_{0:t})}\tilde{w}_t, \end{aligned} \tag{3.2.3}$$

with

$$\tilde{w}_t = \frac{p(x_{0:t}, y_{0:t})}{q(x_{0:t}|y_{0:t})} \tag{3.2.4}$$

so that $\tilde{W}_t$ is $\tilde{w}_t$ corrected for the (unconditional) observation density. Then, (3.2.1) becomes

$$\bar{f}_t = \frac{1}{p(y_{0:t})}\mathbb{E}_q\left[f_t(X_{0:t})\tilde{w}_t\right]. \tag{3.2.5}$$

Notice, that by taking $f_t \equiv \mathbf{1}$ one can obtain from (3.2.5) that

$$\mathbb{E}_q\left[\tilde{w}_t\right] = p(y_{0:t}), \tag{3.2.6}$$

which implies that (3.2.5) can be rewritten as follows

$$\bar{f}_t = \frac{\mathbb{E}_q\left[f_t(X_{0:t})\tilde{w}_t\right]}{\mathbb{E}_q\left[\tilde{w}_t\right]}. \tag{3.2.7}$$

Next, we can estimate (3.2.7) using a random sample $x_{0:t}^{(1)}, \ldots, x_{0:t}^{(N)}$ drawn from the importance distribution $q(x_{0:t}|y_{0:t})$. The required estimate has the form

$$\hat{f}_t = \frac{N^{-1}\sum_{i=1}^{N} f_t(x_{0:t}^{(i)})\tilde{w}_t^{(i)}}{N^{-1}\sum_{i=1}^{N}\tilde{w}_t^{(i)}} \tag{3.2.8}$$

$$= \sum_{i=1}^{N} f_t(x_{0:t}^{(i)}) w_t^{(i)}, \tag{3.2.9}$$

where

$$\tilde{w}_t^{(i)} = \frac{p(x_{0:t}^{(i)}, y_{0:t})}{q(x_{0:t}^{(i)}|y_{0:t})}, \qquad w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{i=1}^{N}\tilde{w}_t^{(i)}},$$

so that $w_t^{(i)}$ are the *normalised importance weights*. Notice that the normalisation justifies focusing on the corrected importance weights $w_t^{(i)}$ instead of the original ones, $W_t^{(i)}$, defined in (3.2.3). Since the former are obtained from the latter by correcting by the term $1/p(y_{0:t})$, which is the same among all the particles, one would obtain that $W_t^{(i)} = w_t^{(i)}$, $\forall i$, with $W_t^{(i)} = \tilde{W}_t^{(i)} / \sum_{j=1}^{N}\tilde{W}_t^{(j)}$.

Hence, we obtained a normalised Monte Carlo estimate $\hat{f}_t$, which, as a ratio of two estimates, is *biased*. Indeed, in Appendix B we show that

$$\mathbb{E}_q\hat{f}_t = \bar{f}_t - \frac{1}{N}\mathrm{Cov}_q\left[\tilde{W}_t, f(X_{0:t})\tilde{W}_t\right] + \frac{\bar{f}_t}{N}\mathrm{Var}_q\left[\tilde{W}_t\right]. \tag{3.2.10}$$

which means the bias of the IS estimator is of order $1/N$. Nevertheless, it is *strongly consistent*, as by the Strong Law of Large Numbers in (3.2.8) the nominator converges to $\bar{f}_t$, while the denominator converges to 1, both almost surely.

One can interpret the estimate $\hat{f}_t$ as an integral with respect to the empirical measure

$$\hat{P}_t(dx_{0:t}|y_{0:t}) = \sum_{i=1}^{N} w_t^{(i)}\delta_{x_{0:t}^{(i)}}(dx_{0:t}) \tag{3.2.11}$$

of the function of interest $f_t$, given by

$$\hat{f}_t = \int f_t(x_{0:t})\hat{P}_t(x_{0:t}|y_{0:t}).$$

Then, the empirical measure resulting from the above integration method can be used to approximate the *posterior* distribution $p(x_{0:t}|y_{0:t})$. This is indeed the key idea behind particle filtering, which amounts to representing the posterior distribution by a weighted average of randomly chosen particles (samples). Notice, however, that to compute the estimate (3.2.9), one requires to draw new samples of $x_{0:t}^{(i)}$ from $q(x_{0:t}|y_{0,t})$ at each point in time $t$, so each time new data $y_t$ becomes available. This means that for long time series the computational complexity of this endeavour becomes extremely high. Therefore, for on-line inference more appropriate methods have been designed, which we will discuss in the next section.

## 3.3 Sequential importance sampling

To make the importance sampling algorithm suitable for on-line inference problems one needs to be able to modify this method in such a way that it becomes possible to compute the estimate of the posterior distribution without modifying the past simulated trajectories. Then, each time new information $y_t$ becomes available, one needs only to simulate one set of current particles $x_t^{(i)}$, $i = 1, \ldots, N$, and not a set of the whole sequences of them, $x_{0:t}^{(i)}$. Hence, at each time $t$, the existing samples are simply augmented with the new samples yielding the updated samples.

### 3.3.1 Recursion for the importance density

To allow for this, putting of some restrictions on the importance density are required. We suppose that at each $t$, the importance density $q(x_{0:t}|y_{0:t})$ factorises in the following way

$$\begin{aligned} q(x_{0:t}|y_{0:t}) &= q(x_{0:t-1}|y_{0:t-1})q(x_t|x_{0:t-1}, y_{0:t}) \\ &= q(x_0)\prod_{k=1}^{t} q(x_k|x_{0:k-1}, y_{0:k}). \end{aligned} \tag{3.3.1}$$

Hence, the posterior is computed by updating using the last term in (3.3.1) of the previously computed density $q(x_{0:t-1}|y_{0:t-1})$, which plays a role of a prior density.

As pointed out in Durbin and Koopman (2012), the above assumption is not as demanding as it may seem at first glance. Notice that one can write without any additional assumptions

$$\begin{aligned} q(x_{0:t}|y_{0:t}) &= \frac{q(x_{0:t}, y_{1:t})}{q(y_{0:t})} \\ &= \frac{q(x_t|x_{0:t-1}, y_{0:t})q(x_{0:t-1}, y_{0:t})}{q(y_{0:t})} \\ &= \frac{q(x_t|x_{0:t-1}, y_{0:t})q(x_{0:t-1}|y_{0:t})q(y_{0:t})}{q(y_{0:t})} \\ &= q(x_{0:t-1}|y_{0:t})q(x_t|x_{0:t-1}, y_{0:t}). \end{aligned}$$

Then, to obtain (3.3.1) one only needs to allow for

$$q(x_{0:t-1}|y_{0:t-1}) \equiv q(x_{0:t-1}|y_{0:t}),$$

which means that augmenting of the set of conditional variables (i.e. $y_{0:t-1}$) in the (already computed) importance density $q(x_{0:t-1}|y_{0:t-1})$ by $y_t$ shall not play a role. Taking into account that $x_{0:t-1}$ has already been generated based on the past observations $y_{0:t-1}$, and that the new value of $y_t$ is independent of everything but the current state $x_t$ (so is independent of simulated sequence $x_{0:t-1}$), the above assumption does not seem particularly restrictive.

From (3.3.1) follows that for the $i$-th particle, $i = 1, \ldots, N$, the importance density updating equation takes the form

$$q(x_{0:t}^{(i)}|y_{0:t}) = q(x_{0:t-1}^{(i)}|y_{0:t-1})q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{0:t}),$$

which is the fundamental particle filtering recursion.

### 3.3.2 Recursion for the importance weights

Plugging (3.3.1) into (3.2.4), one obtaines that the importance weights become

$$\begin{aligned}\tilde{w}_t &= \frac{p(x_{0:t}, y_{0:t})}{q(x_{0:t}|y_{0:t})} \\ &= \frac{p(x_t, y_t|x_{0:t-1}, y_{0:t-1})p(x_{0:t-1}, y_{0:t-1})}{q(x_t|x_{0:t-1}, y_{0:t})q(x_{0:t-1}|y_{0:t-1})} \\ &= \frac{p(x_t, y_t|x_{0:t-1}, y_{0:t-1})}{q(x_t|x_{0:t-1}, y_{0:t})}\frac{p(x_{0:t-1}, y_{0:t-1})}{q(x_{0:t-1}|y_{0:t-1})} \\ &= \frac{p(x_t, y_t|x_{0:t-1}, y_{0:t-1})}{q(x_t|x_{0:t-1}, y_{0:t})}\tilde{w}_{t-1} \\ &\overset{\substack{MC\\CI}}{=} \frac{p(x_t|x_{t-1})p(y_t|x_t)}{q(x_t|x_{0:t-1}, y_{0:t})}\tilde{w}_{t-1},\end{aligned}$$

where by MC and CI we mean that by the Markov property of the state process and conditional independence of the observation process one can write

$$p(x_t, y_t|x_{0:t-1}, y_{0:t-1}) = p(y_t|x_t)p(x_t|x_{t-1}).$$

For the $i$-th particle, $i = 1, \ldots, N$, the importance weight updating equation takes the form

$$\tilde{w}_t^{(i)} = \frac{p(x_t^{(i)}|x_{t-1}^{(i)})p(y_t|x_t^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{0:t})}\tilde{w}_{t-1}^{(i)} \tag{3.3.2}$$

with its normalised counterpart given by

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^N \tilde{w}_t^{(j)}}$$

In (3.3.2), the term

$$\frac{p(x_t^{(i)}|x_{t-1}^{(i)})p(y_t|x_t^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{0:t})}$$

is called the *incremental weight* and it indicates how well a certain augmented particle predicts the next observation.

For further reference, notice that since for the "normalised" weights we have

$$W_t^{(i)} = \frac{1}{p(y_{0:t})}w_t^{(i)},$$

we get the following recursion

$$\tilde{W}_t^{(i)} = \frac{1}{p(y_t|y_{0:t-1})}\frac{p(x_t^{(i)}|x_{t-1}^{(i)})p(y_t|x_t^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{0:t})}\tilde{W}_{t-1}^{(i)}. \tag{3.3.3}$$

### 3.3.3 SIS algorithm

A formal specification of the SIS procedure is given by Algorithm 1. The computational complexity to generate $N$ particles approximating $p(x_{0:T}, y_{0:T})$ is $O(N(T+1))$, where, for notational convenience, it is assumed that one is interested in performing of the computations up to a fixed time $T$.

**Algorithm 1** SIS algorithm

**Step 1:** initialisation

**for** $i := 1, \dots, N$ **do**

Sample $x_0^{(i)} \sim q(x_0)$.

**end for**

**Step 2:** importance sampling

**for** $t := 1, \dots, T$ **do**

**for** $i := 1, \dots, N$ **do**

Sample $x_t^{(i)} \sim q(x_t | x_{t-1}^{(i)}, y_{1:t})$.

Set $x_{0:t}^{(i)} := \left(x_{0:t-1}^{(i)}, x_t^{(i)}\right)$.

Compute the corresponding importance weights

$$\tilde{w}_t^{(i)} = \frac{p(x_t^{(i)} | x_{t-1}^{(i)}) p(y_t | x_t^{(i)})}{q(x_t^{(i)} | x_{0:t-1}^{(i)}, y_{0:t})} \tilde{w}_{t-1}^{(i)}.$$

**end for**

Normalise the importance weights

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^{N} \tilde{w}_t^{(j)}}$$

Approximate the filtering density

$$\hat{p}(x_t | y_{0:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{x_t^{(i)}}(x_t).$$

Compute the required estimate

$$\hat{f}_t = \sum_{i=1}^{N} f(x_t^{(i)}) w_t(i).$$

**end for**

## 3.4 Degeneracy and resampling

The clever SIS weight recursion formula (3.3.2) unfortunately leads to the *degeneracy problem*, i.e. the problem of the distribution of the weights becoming highly skewed when $t$ increases. With time there remain only few particles with non-negligible weights, so that most of the probability mass is concentrated on them, while the other particles are of almost zero weight. In consequence, the SIS algorithm fails to represent adequately the posterior distributions of interest. In fact, this problem is inevitable as the *unconditional* variance of the importance weights is non-decreasing over time, as originally shown by Kong et al. (1994). Before presenting their key theorem, let us recall some basic properties of martingale sequences.

**Lemma 3.4.1** (Non-decreasing variance of a martingale sequence)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\{M_n\}_{n \geq 1}$ be a martingale sequence on this space with respect to the filtration $\{\mathcal{F}_n\}$ of $\mathcal{F}$ and let $\{\Delta_n\}$ be the martingale difference sequence for $\{M_n\}$, i.e.*

$$\Delta_n = M_n - M_{n-1}.$$

*Then for* $n \geq 0$

$$\mathbb{E}M_n = \mathbb{E}M_0.$$

*Moreover, if* $\mathbb{E}M_n^2 < \infty$ *for some* $n \geq 1$*, then for* $j \leq n$ *the random variables* $\Delta_j$ *are uncorrelated and square-integrable implying*

$$\mathbb{E}M_n^2 = \mathbb{E}M_0^2 + \sum_{j=1}^{n} \mathbb{E}\Delta_j^2.$$

*Proof.* The first part follows immediately from the martingale nature of $M_n$ and the Law of Iterated Expectation, as we have

$$\begin{aligned} \mathbb{E}\left[M_n|\mathcal{F}_0\right] &= \mathbb{E}M_0, \\ \mathbb{E}\left[\mathbb{E}\left[M_n|\mathcal{F}_0\right]\right] &= \mathbb{E}\left[\mathbb{E}M_0\right], \\ \mathbb{E}M_n &= \mathbb{E}M_0. \end{aligned}$$

For the second part, notice that for $n \geq 1$

$$\begin{aligned} \mathbb{E}\Delta_n &= \mathbb{E}\left[M_n - M_{n-1}\right] \\ &= \mathbb{E}M_n - \mathbb{E}M_{n-1} \\ &= 0. \end{aligned}$$

Next, without loss of generality, suppose that $M_0 = 0$. Since we assumed that the random variable $M_n$ is square-integrable, then for $k \leq n$ also $\mathbb{E}M_k^2 < \infty$, by the Jensen inequality for conditional expectation and $X_k = \mathbb{E}\left[M_n|\mathcal{F}_k\right]$.

Thus, for $k \leq n$ also $\mathbb{E}\Delta_k^2 < \infty$, as it is a difference of two square-integrable random variables, which, by the Cauchy-Schwartz inequality, yields that for $k \leq m \leq n$ also $\mathbb{E}\Delta_k\Delta_m < \infty$. Since $\Delta_k$ is then $\mathcal{F}_k$-measurable, we have

$$\begin{aligned} \mathbb{E}\Delta_k\Delta_{m+1} &= \mathbb{E}\left[\mathbb{E}\Delta_k\Delta_{m+1}|\mathcal{F}_k\right] \\ &= \mathbb{E}\Delta_k\left[\mathbb{E}\Delta_{m+1}|\mathcal{F}_k\right] \\ &= \mathbb{E}\Delta_k \cdot 0 \\ &= 0. \end{aligned}$$

Because we can decompose $M_n$ as $M_n = \sum_{j=1}^{n} \Delta_j$, i.e. into a sum of uncorrelated random variables, we can calculate the variance of $M_n$ as if it was a sum of mean-zero dependent random variables. Then, we

obtain

$$\begin{aligned}
\operatorname{Var}[M_n] &= \mathbb{E}M_n^2 \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{n}\Delta_j\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{n}\sum_{k=1}^{n}\Delta_j\Delta_k\right] \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}\Delta_j\Delta_k \\
&= \sum_{j=1}^{n}\mathbb{E}\Delta_j^2 + 2\sum_{j=1}^{n}\sum_{k>j}^{n}\mathbb{E}\Delta_j\Delta_k \\
&= \sum_{j=1}^{n}\mathbb{E}\Delta_j^2 + 0 \\
&= \sum_{j=1}^{n}\mathbb{E}\Delta_j^2,
\end{aligned}$$

which completes the proof. □

The above lemma allows us to argue that the variance of the importance weights is non-decreasing in time by simply showing that they form a martingale sequence. This was indeed the key idea behind the theorem shown in Kong et al. (1994), which we below extend to the analysed case of the importance density given by (3.3.1).

**Theorem 3.4.1** (Kong-Liu-Wong)**.** *The importance weight $w_t^{(i)}$ is a martingale sequence in $t$ with both, the sample $x_{0:t}^i$ and the observations $y_{0:t}$ treated as random. This implies that its variance is an increasing function of $t$.*

*Proof.* Consider the $i$-particle $x_{0:t-1}^{(i)}$ and let

$$\mathcal{F}_{t-1}^i = \sigma\left\{x_{0:t-1}^{(i)}, y_{0:t-1}\right\},$$

be a $\sigma$-algebra generated by all the observations and this particle, up to time $t-1$. Recall that the recursion formula (3.3.3) for the "corrected" weights (3.2.3) reads

$$\tilde{W}_t^{(i)} = \frac{1}{p(y_t|y_{0:t-1})}\frac{p(x_t^{(i)}|x_{t-1}^{(i)})p(y_t|x_t^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{0:t})}\tilde{W}_{t-1}^{(i)},$$

which is due to the assumed form of the importance function, which factorises according to (3.3.1)

$$q(x_{0:t}^{(i)}|y_{0:t}) = q(x_{0:t-1}^{(i)}|y_{0:t-1})q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{0:t}).$$

Moreover, recall that the formula for the posterior has the following form

$$p(x_{0:t}^{(i)}|y_{0:t}) = p(x_{0:t-1}^{(i)}|y_{0:t-1})\frac{p(x_t^{(i)}|x_{t-1}^{(i)})p(y_t|x_t^{(i)})}{p(y_t|y_{0:t-1})}.$$

Now, let us treat both, $y_t$ and $x_t^{(i)}$, as realizations of random variables, with $x_t^{(i)} \sim q(x_t|x_{0:t-1}^{(i)}, y_{0:t})$.

Then, the joint conditional density is given by

$$p(x_t^{(i)}, y_t | x_{0:t-1}^{(i)}, y_{0:t-1}) = q(x_t | x_{0:t-1}^{(i)}, y_{0:t}) p(y_t | x_{0:k-1}^{(i)}, y_{0:k-1}). \tag{3.4.1}$$

Notice that the elements $x_0^{(i)}, x_1^{(i)}, \ldots, x_{t-1}^{(i)}$ of $x_{0:t-1}^{(i)}$ were generated based on $y_{0:t-1}$, so one can treat $y_t$ as conditionally independent of $x_{0:t-1}^{(i)}$. This allows us to rewrite (3.4.1) as

$$p(x_t^{(i)}, y_t | x_{0:t-1}^{(i)}, y_{0:t-1}) = q(x_t | x_{0:t-1}^{(i)}, y_{0:t}) p(y_t | y_{0:k-1}). \tag{3.4.2}$$

Next, consider the conditional expectation of $W_t^{(i)}$, given $x_{0:t-1}^{(i)}$, $y_{0:t-1}$, with $x_t^{(i)}$ and $y_t$ treated as random. By the weight recursion formula we have

$$\begin{aligned}
\mathbb{E}\left[W_t^{(i)} | \mathcal{F}_{t-1}^i\right] &= \int\int W_t^{(i)} p(x_t^{(i)}, y_t | x_{0:t-1}^{(i)}, y_{0:t-1}) dx_t^{(i)} dy_t \\
&= W_{t-1}^{(i)} \int\int \frac{1}{p(y_t | y_{0:t-1})} \frac{p(x_t^{(i)} | x_{t-1}^{(i)}) p(y_t | x_t^{(i)})}{q(x_t^{(i)} | x_{0:t-1}^{(i)}, y_{0:t})} q(x_t | x_{0:t-1}^{(i)}, y_{0:t}) p(y_t | y_{0:k-1}) dx_t^{(i)} dy_t \\
&= W_{t-1}^{(i)} \int\int p(x_t^{(i)} | x_{t-1}^{(i)}) p(y_t | x_t^{(i)}) dx_t^{(i)} dy_t \\
&= W_{t-1}^{(i)},
\end{aligned}$$

showing that indeed, the "corrected", unnormalised importance weight $W_t^{(i)}$ is a martingale in $t$. Hence,

$$\mathrm{Var}\left[W_{t-1}^{(i)}\right] = \mathbb{E}\left[\mathrm{Var}\left[W_t^{(i)} | \mathcal{F}_{t-1}\right]\right] \leq \mathrm{Var}\left[W_t^{(i)}\right],$$

i.e. the unconditional variance of $W_t^{(i)}$ is non-decreasing in time. This completes the proof. □

For completeness, notice that by the variance decomposition formula, we have for the martingale $W_t^{(i)}$,

$$\mathrm{Var}\left[W_t^{(i)}\right] = \mathbb{E}\left[\mathrm{Var}\left[W_t^{(i)} | y_t\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[W_t^{(i)} | y_t\right]\right],$$

Finally, since

$$\mathbb{E}\left[W_t^{(i)} | y_t\right] = 1, \qquad \forall y_t,$$

we obtain

$$\mathrm{Var}\left[W_t^{(i)}\right] = \mathbb{E}\left[\mathrm{Var}\left[W_t^{(i)} | y_t\right]\right].$$

### 3.4.1 Effective sample size

As pointed out by Kong et al. (1994), since the SIS is a form of IS, i.e. of an MC algorithm, one can measure the efficiency of the sequential approach by comparing it with a direct sampling from a theoretical distribution of interest. As in Subsection 3.2, consider an arbitrary measurable function $f : \mathcal{X} \to \mathbb{R}$, for which we are interested in estimation of the mean $\bar{f} = \mathbb{E}_p(f(X_t))$. Moreover, let $X_t^{(1)}, \ldots, X_t^{(N)} \overset{iid}{\sim} q(x_t | x_{0:t-1}, y_{0:t})$ and $Y_t^{(1)}, \ldots, Y_t^{(N)} \overset{iid}{\sim} p(x_t | x_{0:t-1}, y_{0:t})$ be two independent sequences of i.i.d. random variables, both of length $N$. The estimators for $\bar{f}$ delivered by both techniques are respectively given by

$$\begin{aligned}
\hat{f}^{IS} &= \frac{N^{-1} \sum_{i=1}^N \tilde{w}_t^{(i)}(X_t^{(j)}) f(X_t^{(j)})}{N^{-1} \sum_{i=1}^N \tilde{w}_t^{(i)}(X_t^{(j)})}, \\
\hat{f}^{MC} &= N^{-1} \sum_{i=1}^N f(Y_t^{(j)}).
\end{aligned}$$

The efficiency of the SIS estimator (relatively to the MC one) is given by the ratio

$$\frac{\mathrm{Var}\left[\hat{f}^{MC}\right]}{\mathrm{Var}\left[\hat{f}^{IS}\right]}, \tag{3.4.3}$$

which, following Kong et al. (1994) and Liu (1996), can be used to define the *effective sample size* as

$$N_{ESS} = N\frac{\mathrm{Var}\left[\hat{f}^{MC}\right]}{\mathrm{Var}\left[\hat{f}^{IS}\right]} \leq N. \tag{3.4.4}$$

It can be interpreted as follows: an inference based on $N$ weighted samples and the SIS estimator is approximately equivalent to an inference based on $N_{ESS}$ i.i.d. samples from the distribution of interest and the crude MC estimator. In other words, under the MC setup, one requires $N_{ESS} \leq N$ samples to obtain an estimator with the same precision as using $N$ samples and the SIS estimator.

There are two obstacles related to the way $N_{ESS}$ in (3.4.4) is defined. First, it depends on the function of interest $f$; second – generally, it cannot be computed exactly, as usually the target distribution is known only up to normalization. In Appendix B we present the derivation of the following approximation to (3.4.4), which is independent of $f$,

$$\frac{\mathrm{Var}_q\left[\hat{f}^{MC}\right]}{\mathrm{Var}_p\left[\hat{f}^{IC}\right]} \approx \frac{1}{1+\mathrm{Var}_q\left[\tilde{W}_t\right]}, \tag{3.4.5}$$

originally introduced by Kong et al. (1994).

Still, it is usually difficult, or even impossible, to exactly compute (3.4.5). Thus, to measure the efficiency of SIS technique one resorts to approximating (3.4.5) by

$$\hat{N}_{ESS} = \frac{1}{\sum_i^N \left(w_t^i\right)^2}, \tag{3.4.6}$$

where $w_t^i$ is the normalized importance weight, for a SIS sample $x_t^{(1)}, \ldots, x_t^{(N)}$. It can be obtained as follows

$$\begin{aligned}
N_{ESS} &= \frac{N}{1+\mathrm{Var}_q\left[\tilde{W}_t(X_t)\right]} \\
&\overset{(*)}{=} \frac{N}{\mathbb{E}_q\left[\tilde{W}_t(X_t)\right]^2} \\
&\approx \frac{N}{N^{-1}\sum_{i=1}^N \left(\tilde{W}_t^i\right)^2} \\
&= \frac{N}{N^{-1}C^2\sum_{i=1}^N \left(\frac{\tilde{W}_t^i}{C}\right)^2} \\
&\overset{(**)}{=} \frac{NC^{-1}}{N^{-1}C\sum_{i=1}^N \left(W_t^i\right)^2} \\
&\approx \frac{1}{\sum_{i=1}^N \left(W_t^i\right)^2} \\
&= \frac{1}{\sum_{i=1}^N \left(w_t^i\right)^2}.
\end{aligned}$$

where in $(*)$ we use the fact that $\mathbb{E}_q\left[\tilde{W}_t\right] = 1$, in $(**)$ we put $C := \sum_{i=1}^{N} \tilde{W}_t^i$ and the last step is due to the normalised $w_t^i$ being equivalent to the normalized $W_t^i$. Notice, that by construction $\hat{N}_{ESS}$ is between 1 and $N$. The latter case corresponds to a situation when the importance weights are evenly distributed, the former – to an instance of severe degeneracy, with only one active particle. Hence, the above approximation to the effective sample size constitutes a suitable measure of degeneracy of the SIS algorithm. If $\hat{N}_{ESS} \approx N$, then the efficiency of the SIS technique is high, since then the SIS sample size is almost equal to $N$ i.i.d. MC samples.

### 3.4.2 SIR algorithm

It follows from the Kong-Liu-Wong theorem that the degeneracy of the SIS algorithm is unavoidable. In consequence, a large computational effort is needed to update particles with almost zero contribution to the approximation (3.2.9). Intuitively, it must be wasteful to keep in the recursion insignificantly contributing particles. Hence, to (partially) solve this problem the idea of *resampling* was introduced by Gordon et al. (1993). The SIS with resampling (SIR) amounts to including an additional step in the algorithms, at which trajectories with low normalised importance weights are eliminated, so that one can concentrate on trajectories with considerable weights.

The theoretical foundation for SIR is as follows. Let $\left\{x_t^{(i)}, \tilde{w}_t^{(i)}\right\}_{i=1}^{N}$ be a weighted sample obtained at the $t$-th iteration of the Step 2 of Algorithm 1, with the corresponding normalised weights $w_t^{(i)}$. As it has already been stated, the resulting empirical measure (3.2.11) is given by

$$\hat{P}_t(dx_{0:t}|y_{0:t}) = \hat{p}_t(x_{0:t}|y_{0:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{x_{0:t}^{(i)}}(x_{0:t}).$$

By the law of large numbers, an integral of an arbitrary measurable function $f_t$ with respect to $\hat{p}_t(\cdot)$ converges to its integral with respect to the distribution of interest $p_t(\cdot)$, i.e.

$$\hat{f}_t = \int f_t(x_{0:t})\hat{p}_t(x_{0:t}|y_{0:t})dx_{0:t} = \sum_{i=1}^{N} f_t(x_{0:t}^{(i)})w_t^{(i)} \overset{N\to\infty}{\to} \int f_t(x_{0:t})\hat{p}_t(x_{0:t}|y_{0:t})dx_{0:t} = \bar{f}_t. \tag{3.4.7}$$

Now, consider a sample of size $\tilde{N}$ drawn from $\hat{p}_t(x_{0:t}|y_{0:t})$ with replacement, i.e. $\mathbb{P}(\tilde{x}_{0:t}^{(j)} = x_{0:t}^{(i)}) = w_t^{(i)}$, $j = 1, \ldots, \tilde{N}$, where the new particles have equal weights, $\tilde{w}_t^{(j)} = \frac{1}{\tilde{N}}$. Once again, by the law of large numbers,

$$\frac{1}{\tilde{N}} \sum_{j=1}^{\tilde{N}} f_t(\tilde{x}_{0:t}^{(j)}) \overset{\tilde{N}\to\infty}{\to} \sum_{i=1}^{N} f_t(x_{0:t}^{(i)})w_t^{(i)},$$

so that the integral of $f_t$ with respect to the new, "resampled" empirical measure constitutes a good approximation to the original empirical measure $\hat{p}_t(\cdot)$, provided we resample sufficiently many particles. This basic resampling scheme is known as *multinomial sampling*.

Resampling helps to increase the number of active particles, however, it also generates two important problems. The first is the *sample impoverishment*, consisting in decreasing diversity within the particles set. Because of replacement and the fact that resampling is applied to the whole particle trajectory $x_{O:t}^{(i)}$, and not only the most recent value $x_t^{(i)}$, particles with high importance weights are likely to be selected several time. On the other hand, particles with negligible weights most likely will not be chosen at all, which means the that their "history" will be eliminated. Hence, the current sample becomes less versatile and consequently less and less able to correctly represent the target distribution. In the extreme case, one may try to estimate it with a set of particles with a single common ancestor, which obviously gives a poor quality approximation.

The second trouble is introduction of *dependence* between particle trajectories, which brings in unnecessary Monte Carlo variation into the estimate $\hat{f}_t$ in (3.4.7) at the current time step. As shown in Chopin (2004), the estimator computed before resampling has a lower variance than the one computed after that step, hence is a preferred to perform the estimation based on the original sample. It is worth noting that to decrease the Monte Carlo variance, other resampling schemes were designed, for instance *stratified sampling* (Kitagawa, 1996) or *residual resampling* (Liu and Chen, 1998).

Finally, because of this additional variation, resampling at each time step is detrimental, which means it should be performed only when necessary. To detect a presence of such a "necessity", the effective sample size estimate (3.4.6) can be used. Then, the resampling step is performed only when the effective sample size $\hat{N}_{ESS}$ falls below a certain threshold $N_{tres}$, which usually is set to $\frac{N}{2}$ (cf. Doucet and Johansen, 2009). The formal description of the standard SIR algorithm is presented in Algorithm 2.

**Algorithm 2** SIR algorithm

**Step 1:** initialisation

**for** $i := 1$ **to** $N$ **do**

Sample $x_0^{(i)} \sim q(x_0)$.

**end for**

**for** $t := 1$ **to** $T$ **do**

**Step 2:** importance sampling

**for** $i := 1$ **to** $N$ **do**

Sample $\tilde{x}_t^{(i)} \sim q(x_t|x_{t-1}^{(i)}, y_{1:t})$.

Set $\tilde{x}_{0:t}^{(i)} := \left(x_{0:t-1}^{(i)}, \tilde{x}_t^{(i)}\right)$.

Compute the corresponding importance weights

$$\tilde{w}_t^{(i)} = \frac{p(\tilde{x}_t^{(i)}|x_{t-1}^{(i)})p(y_t|\tilde{x}_t^{(i)})}{q(\tilde{x}_t^{(i)}|x_{0:t-1}^{(i)}, y_{0:t})}\tilde{w}_{t-1}^{(i)}.$$

**end for**

Normalise the importance weights

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^{N} \tilde{w}_t^{(j)}}.$$

Compute the effective sample size

$$\hat{N}_{ESS} = \left(\sum_{i=1}^{N} \tilde{w}_t^{(i)2}\right)^{-1}.$$

Approximate the filtering density

$$\hat{p}(x_t|y_{0:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{\tilde{x}_t^{(i)}}(x_t).$$

Compute the required estimate

$$\hat{f}_t = \sum_{i=1}^{N} f(\tilde{x}_t^{(i)}) w_t^{(i)}.$$

**Step 3:** selection

**if** $\hat{N}_{ESS} < N_{thres}$ **then**

Resample with replacement $N$ particles, $\left\{x_t^{(i)}\right\}_{i=1}^{N}$, from $\left\{\tilde{x}_t^{(i)}\right\}_{i=1}^{N}$ with corresponding probabilities $\left\{w_t^{(i)}\right\}_{i=1}^{N}$.

Set $x_{0:t}^{(i)} := \left(x_{0:t-1}^{(i)}, x_t^{(i)}\right)$.

Set $w_t^{(i)} := \frac{1}{N}$.

**end if**

**end for**

## 3.5 Selection of the importance function

The choice of the importance density $q(x_t|x_{0:t-1}, y_{0:t})$ is of crucial importance for the performance of the SIS algorithm. As pointed out by Doucet et al. (2000), an appropriate choice of the importance function may help to limit the degeneracy problem. This "appropriateness" shall be primarily seen in terms of weights variance minimisation, given the simulated trajectory and the observations, as then such an importance density gives roughly the same weight to all particles. Moreover, the importance function should in general have relatively heavy tails to be robust for outliers.

### 3.5.1 Optimal importance function

The following proposition, due to Doucet et al. (2000), characterises the optimal importance function.

**Proposition 3.5.1.** *The importance function which minimises the variance of the importance weight $W_t^{(i)}$, conditional upon $x_{0:t-1}^{(i)}$ and $y_{0:t}$, is given by*

$$q(x_t|x_{0:t-1}, y_{0:t}) = p(x_t|x_{t-1}^{(i)}, y_t), \tag{3.5.1}$$

*i.e. is equal to the target distribution.*

*Proof.* Using the notation introduced before, we obtain that the variance of the importance weight under $q(x_t|x_{0:t-1}, y_{0:t})$ given by (3.5.1) is equal to

$$\begin{aligned}
\mathrm{Var}_{q(x_t|x_{0:t-1},y_{0:t})}\left[W_t^{(i)}\right] &= (W_{t-1}^{(i)})^2 \left[\int \frac{\left(p(y_t|x_t)p(x_t|x_{t-1}^{(i)})\right)^2}{q(x_t|x_{0:t-1}, y_{0:t})} dx_t - p^2(y_t|x_{t-1}^{(i)})\right] \\
&= (W_{t-1}^{(i)})^2 \left[\int \frac{\left(p(y_t|x_t)p(x_t|x_{t-1}^{(i)})\right)^2}{p(x_t|x_{t-1}^{(i)}, y_t)} dx_t - p^2(y_t|x_{t-1}^{(i)})\right] \\
&= (W_{t-1}^{(i)})^2 \left[\int p(y_t|x_{t-1}^{(i)}) \frac{\left(p(y_t|x_t)p(x_t|x_{t-1}^{(i)})\right)^2}{p(y_t|x_t, x_{t-1}^{(i)})p(x_t|x_{t-1}^{(i)})} dx_t - p^2(y_t|x_{t-1}^{(i)})\right] \\
&\overset{CI}{=} (W_{t-1}^{(i)})^2 \left[p(y_t|x_{t-1}^{(i)}) \int \frac{p^2(y_t|x_t)p^2(x_t|x_{t-1}^{(i)})}{p(y_t|x_t)p(x_t|x_{t-1}^{(i)})} dx_t - p^2(y_t|x_{t-1}^{(i)})\right] \\
&= (W_{t-1}^{(i)})^2 \left[p(y_t|x_{t-1}^{(i)}) \int p(y_t|x_t)p(x_t|x_{t-1}^{(i)}) dx_t - p^2(y_t|x_{t-1}^{(i)})\right] \\
&= (W_{t-1}^{(i)})^2 \left[p^2(y_t|x_{t-1}^{(i)}) - p^2(y_t|x_{t-1}^{(i)})\right] \\
&= 0,
\end{aligned}$$

which completes the proof. □

Notice, that with the importance distribution defined as in (3.5.1), the SIS importance weight updating

equation (3.3.2) becomes

$$\begin{aligned}\tilde{w}_t^{(i)} &= \frac{p(x_t^{(i)}|x_{t-1}^{(i)})p(y_t|x_t^{(i)})}{p(x_t|x_{t-1}^{(i)}, y_t)}\tilde{w}_{t-1}^{(i)} \\ &= p(y_t|x_{t-1}^{(i)})\frac{p(x_t^{(i)}|x_{t-1}^{(i)})p(y_t|x_t^{(i)})}{p(x_t^{(i)}|x_{t-1}^{(i)})p(y_t|x_t^{(i)}, x_{t-1}^{(i)})}\tilde{w}_{t-1}^{(i)}, \\ &\overset{CI}{=} p(y_t|x_{t-1}^{(i)})\tilde{w}_{t-1}^{(i)}.\end{aligned}$$

The last formula looks simple, there are, however, two serious limitations regarding using of the optimal importance function. First, it depends on the ability of drawing directly from $p(x_t|x_{t-1}^{(i)}, y_t)$, the lack of which was the main reason for resorting to the importance sampling in the first place. Second, it requires evaluating of the following integral

$$p(y_t|x_{t-1}^{(i)}) = \int p(y_t|x_t)p(x_t|x_{t-1}^{(i)})dx_t, \tag{3.5.2}$$

which in general has no analytic form. For this reason, one often resorts to approximative techniques, as we will discuss in the next Subsection.

**Gaussian model with nonlinear transition equation**

However, as pointed out in Doucet et al. (2000), for an important class of the Gaussian state space models with nonlinear transition equation and linear observation equation, analytic evaluation of (3.5.2) is possible. Consider the following model

$$\begin{aligned}X_t &= f(X_{t-1}) + V_t, \quad V_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_V), \\ Y_t &= CX_t + W_t, \qquad W_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_W),\end{aligned} \tag{3.5.3}$$

where $f : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ is a nonlinear function, $C \in \mathbb{R}^{n_y} \times \mathbb{R}^{n_x}$ is an observation matrix and $V_t$, $W_t$ are mutually independent. Define

$$\Sigma^{-1} = \Sigma_V^{-1} + C^T\Sigma_W^{-1}C, \tag{3.5.4}$$

$$\mu_t = \Sigma\left(\Sigma_V^{-1}f(X_{t-1}) + C^T\Sigma_W^{-1}Y_t\right). \tag{3.5.5}$$

**Proposition 3.5.2.** *The optimal importance distribution for the model* (3.5.3) *is given by*

$$q(x_t|x_{t-1}, y_t) = \mathcal{N}(\mu_t, \Sigma),$$

*so that the optimal weight update* (3.5.2) *becomes*

$$p(y_t|x_{t-1}^{(i)}) \propto \exp\left(-\frac{1}{2}\left(y_t - Cf(x_{t-1})\right)^T\left(C\Sigma_V C^T + \Sigma_W\right)^{-1}\left(y_t - Cf(x_{t-1})\right)\right),$$

*where* $\Sigma$ *and* $\mu$ *are given by* (3.5.4) *and* (3.5.5)*, respectively.*

*Proof.* First, model (3.5.3) implies that the conditional distribution of $Y_t$ given $X_{t-1}$ can be simply derived as follows

$$\begin{aligned}Y_t &= CX_t + W_t \\ &= Cf(X_{t-1}) + CV_t + W_t, \\ Y_t|X_{t-1} &\sim \mathcal{N}(Cf(X_{t-1}), C\Sigma_V C^T + \Sigma_W),\end{aligned}$$

which delivers the weight update formula.

Second, notice that model (3.5.3) is Gaussian, so the joint distribution of $X_t$ and $Y_t$ is Gaussian. In particular, conditional on $X_{t-1}$, we have

$$X_t, Y_t | X_{t-1} \sim \mathcal{N}\left( \begin{bmatrix} f(X_{t-1}) \\ Cf(X_{t-1}) \end{bmatrix}, \begin{bmatrix} \Sigma_V & \Sigma_V C^T \\ C\Sigma_V & C\Sigma_V C^T + \Sigma_W \end{bmatrix} \right).$$

Then, by applying the standard formula for the conditional distribution of multivariate normal random variable, we obtain

$$X_t | Y_t, X_{t-1} \sim \mathcal{N}(\bar{\mu}_t, \bar{\Sigma}),$$

with

$$\bar{\mu}_t = f(X_{t-1}) + \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1}(Y_t - Cf(X_{t-1})), \tag{3.5.6}$$
$$\bar{\Sigma} = \Sigma_V - \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1} C\Sigma_V. \tag{3.5.7}$$

Finally, we need to show that (3.5.6) and (3.5.7) correspond to (3.5.5) and (3.5.4), respectively. To establish the equivalence of the variances, we will check whether $\bar{\Sigma}\Sigma^{-1} = \mathbb{I}$. We have

$$\begin{aligned}
\bar{\Sigma}\Sigma^{-1} =& \left[\Sigma_V - \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1} C\Sigma_V\right] \left[\Sigma_V^{-1} + C^T \Sigma_W^{-1} C\right] \\
=& \Sigma_V \Sigma_V^{-1} + \Sigma_V C^T \Sigma_W^{-1} C - \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1} C\Sigma_V \Sigma_V^{-1} \\
& - \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1} C\Sigma_V C^T \Sigma_W^{-1} C \\
=& \mathbb{I} + \Sigma_V C^T \times \left(\Sigma_W^{-1} - (C\Sigma_V C^T + \Sigma_W)^{-1} - (C\Sigma_V C^T + \Sigma_W)^{-1} C\Sigma_V C^T \Sigma_W^{-1}\right) \times C \\
=& \mathbb{I} + \Sigma_V C^T \times \Big((C\Sigma_V C^T + \Sigma_W)^{-1}(C\Sigma_V C^T + \Sigma_W)\Sigma_W^{-1} - (C\Sigma_V C^T + \Sigma_W)^{-1}\Sigma_W^{-1}\Sigma_W \\
& - (C\Sigma_V C^T + \Sigma_W)^{-1} C\Sigma_V C^T \Sigma_W^{-1}\Big) \times C \\
=& \mathbb{I} + \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1} \times \Big((C\Sigma_V C^T + \Sigma_W)\Sigma_W^{-1} - \Sigma_W^{-1}\Sigma_W - C\Sigma_V C^T \Sigma_W^{-1}\Big) \times C \\
=& \mathbb{I} + \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1} \times \Big(C\Sigma_V C^T \Sigma_W^{-1} + \Sigma_W \Sigma_W^{-1} - \Sigma_W^{-1}\Sigma_W - C\Sigma_V C^T \Sigma_W^{-1}\Big) \times C \\
=& \mathbb{I},
\end{aligned}$$

which shows that (3.5.4) and (3.5.7) characterise the same variance matrix.

As far as the means are concerned, notice that from (3.5.5) and (3.5.4) we have

$$\begin{aligned}
\mu_t &= \Sigma\left(\Sigma_V^{-1} f(X_{t-1}) + C^T \Sigma_W^{-1} Y_t\right), \\
&= \Sigma\Sigma_V^{-1} f(X_{t-1}) + \Sigma C^T \Sigma_W^{-1} Y_t, \\
\bar{\mu}_t &= f(X_{t-1}) + \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1}(Y_t - Cf(X_{t-1})) \\
&= \Big(\mathbb{I} - C\Sigma_V (C\Sigma_V C^T + \Sigma_W)^{-1} C\Big) f(X_{t-1}) + \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1} Y_t.
\end{aligned}$$

Hence, we need to show that

$$\Sigma\Sigma_V^{-1} = \mathbb{I} - C\Sigma_V (C\Sigma_V C^T + \Sigma_W)^{-1} C, \tag{3.5.8}$$
$$\Sigma C^T \Sigma_W^{-1} = \Sigma_V C^T (C\Sigma_V C^T + \Sigma_W)^{-1}. \tag{3.5.9}$$

Since

$$\begin{aligned}\Sigma^{-1} &= \Sigma_V^{-1} + C^T\Sigma_W^{-1}C,\\ \bar{\Sigma} &= \Sigma_V - \Sigma_V C^T(C\Sigma_V C^T + \Sigma_W)^{-1}C\Sigma_V,\\ \Sigma &= \bar{\Sigma},\end{aligned}$$

we have

$$\begin{aligned}\Sigma\Sigma_V^{-1} &= \Big(\Sigma_V - \Sigma_V C^T(C\Sigma_V C^T + \Sigma_W)^{-1}C\Sigma_V\Big)\Sigma_V^{-1}\\ &= \mathbb{I} - \Sigma_V C^T(C\Sigma_V C^T + \Sigma_W)^{-1}C,\end{aligned}$$

which shows (3.5.8). For (3.5.9) we can proceed similarly and write

$$\begin{aligned}\Sigma C^T\Sigma_W^{-1} &= \Big(\Sigma_V - \Sigma_V C^T(C\Sigma_V C^T + \Sigma_W)^{-1}C\Sigma_V\Big)C^T\Sigma_W^{-1}\\ &= \Sigma_V C^T\Sigma_W^{-1} - \Sigma_V C^T(C\Sigma_V C^T + \Sigma_W)^{-1}C\Sigma_V C^T\Sigma_W^{-1}\\ &= \Sigma_V C^T(C\Sigma_V C^T + \Sigma_W)^{-1}\Big((C\Sigma_V C^T + \Sigma_W) - C\Sigma_V C^T\Big)\Sigma_W^{-1}\\ &= \Sigma_V C^T(C\Sigma_V C^T + \Sigma_W)^{-1}\Sigma_W\Sigma_W^{-1}\\ &= \Sigma_V C^T(C\Sigma_V C^T + \Sigma_W)^{-1},\end{aligned}$$

completing the proof. □

### 3.5.2 Local linearisation

Following Doucet et al. (2000), one can select as the importance function $q(x_t|x_{0:t-1}, y_{0:t})$ a parametric distribution $q(x_t|\theta(x_{0:t-1}, y_{0:t}))$, where $\theta$ is a finite-dimensional parameter, $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$ determined by $x_{t-1}$ and $y_t$ via a deterministic mapping $\theta : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \to \Theta$. The cited authors proposed two strategies consisting in approximation of the optimal importance function using a local linearisation, which results in a Gaussian importance function with parameters dependent on the simulated trajectory. Below, we introduce both techniques, as we will rely on them later on, in the application part (Chapter 4).

**Local linearisation of the state space model**

When the model under consideration is nonlinear yet Gaussian, one can locally linearise the model itself, i.e. either the observation equation or the transition equation. This approach is thus conceptually related the Extended Kalman Filter, which is based on linearisation of the transition and observation equations along the trajectories. However, compared to the latter, the current method has the advantage of yielding of the importance distribution which is asymptotically convergent to the optimal one (under the standard assumptions on the importance functions). Consider the following model

$$\begin{aligned}X_t &= f(X_{t-1}) + V_t, & V_t &\overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_V),\\ Y_t &= g(X_t) + W_t, & W_t &\overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_W),\end{aligned}$$

where $f : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$, $g : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ are differentiable and $V_t$, $W_t$ are mutually independent. Performing of the first order approximation of the observation equation, i.e. the linearisation of $g(x)$ in $f(X_{t-1})$ yields

$$\begin{aligned}Y_t &= g(X_t) + W_t\\ &\approx g(f(X_{t-1})) + \left.\frac{\partial g(x)}{\partial x}\right|_{x=f(X_{t-1})}(X_t - f(X_{t-1})) + W_t.\end{aligned} \tag{3.5.10}$$

Thus,

$$
\begin{aligned}
&X_t = f(X_{t-1}) + V_t, && V_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_V),\\
&\tilde{Y}_t = \left.\frac{\partial g(x)}{\partial x}\right|_{x=f(X_{t-1})} X_t + W_t, && W_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_W),\\
&\tilde{Y}_t = Y_t - g(f(X_{t-1})) - \left.\frac{\partial g(x)}{\partial x}\right|_{x=f(X_{t-1})} f(X_{t-1})
\end{aligned}
$$

defines a system driven by the same transition equation as the one in the model under consideration, yet where the observation equation is linear Gaussian, as in the special class of models discussed in the previous subsection. Thus, Proposition 3.5.2 applies if one takes as the transition matrix $C$ the transition matrix of the linearised model, given by the Jacobian of $g(x)$ evaluated at $x = f(X_{t-1})$, i.e. $\left.\frac{\partial g(x)}{\partial x}\right|_{x=f(X_{t-1})}$ and as the observation $Y_t$, the transformed observation $\tilde{Y}_t$. The resulting importance function is

$$q(x_t|x_{t-1}, y_t) = \mathcal{N}(\mu_t, \Sigma_t),$$

where

$$
\begin{aligned}
\Sigma_t &= \Sigma_V^{-1} + \left[\left.\frac{\partial g(x)}{\partial x}\right|_{x=f(X_{t-1})}\right]^T \Sigma_W^{-1} \left.\frac{\partial g(x)}{\partial x}\right|_{x=f(X_{t-1})},\\
\mu_t &= \Sigma_t \left(\Sigma_V^{-1} f(x_{t-1}) + \left[\left.\frac{\partial g(x)}{\partial x}\right|_{x=f(X_{t-1})}\right]^T \Sigma_W^{-1} \left[Y_t - g(f(X_{t-1})) - \left.\frac{\partial g(x)}{\partial x}\right|_{x=f(X_{t-1})} f(X_{t-1})\right]\right).
\end{aligned}
$$

**Local linearisation of the optimal importance function**

Assume that $\ell(x_t) \equiv \log p(x_t|x_{t-1}, y_t)$ is twice differentiable with respect to $x_t$ on $\mathbb{R}^{n_x}$ and define

$$
\begin{aligned}
\ell'(x) &\equiv \left.\frac{\partial \ell(x_t)}{\partial x_t}\right|_{x_t=x},\\
\ell''(x) &\equiv \left.\frac{\partial^2 \ell(x_t)}{\partial x_t \partial x_t^T}\right|_{x_t=x}.
\end{aligned}
$$

Performing of the second order Taylor expansion around an arbitrary, though deterministically determined by $x_{t-1}$ and $y_t$, point $x$ results in

$$\ell(x_t) \approx \ell(x) + (\ell'(x))^T (x_t - x) + \frac{1}{2}(x_t - x_t)^T \ell''(x)(x_t - x),$$

Next, suppose $\ell''(x)$ is negative definite (which is the case when $\ell(x_t)$ is concave) and put

$$
\begin{aligned}
\Sigma(x) &\equiv -(\ell''(x))^{-1},\\
\mu(x) &\equiv \Sigma(x)\ell'(x).
\end{aligned}
$$

This yields

$$(\ell'(x))^T (x_t - x) + \frac{1}{2}(x_t - x_t)^T \ell''(x)(x_t - x) = C - \frac{1}{2}(x_t - x - \mu(x))^T \Sigma^{-1}(x)(x_t - x - \mu(x)),$$

where $C$ is a constant, which suggests approximating of the optimal importance function with the following Gaussian distribution

$$q(x_t|x_{t-1}, y_t) = \mathcal{N}(\mu(x) + x, \Sigma(x)).$$

In case of the target $p(x_t|x_{t-1}, y_t)$ being unimodal, this procedure boils down to taking as the expansion

point $x$ the mode of $p(x_t|x_{t-1}, y_t)$, which leads to $\mu(x) = 0$.

### 3.5.3 The bootstrap filter

An important class of SIR algorithms arises when the importance density takes the form of the prior distribution

$$q(x_t|x_{t-1}, y_t) = p(x_t|x_{t-1}),$$

so that the information provided by $y_t$ is neglected. Then, the importance weights simplify to

$$\tilde{w}_t^{(i)} = p(y_t|x_t^{(i)})\tilde{w}_{t-1}^{(i)}.$$

Moreover, in this class of algorithms resampling is usually performed at each step, i.e. $N_{thres} = N$. Hence,also the weights are reset to $1/N$ in each iteration, which leads to an even simpler weight specification

$$\tilde{w}_t^{(i)} = p(y_t|x_t^{(i)}). \tag{3.5.11}$$

This particular case of SIR is called the *bootstrap filter* and for future reference is given in Algorithm 3. In fact the first particle filter introduced in Gordon et al. (1993) was the bootstrap filter. The advantage of this method is clearly its simplicity, yet it suffers from sensitivity to outliers. Moreover, simulations using the prior importance function are inefficient, as the information on the state space brought by the observations is discarded.

**Algorithm 3** Bootstrap filter

**Step 1:** initialisation

**for** $i := 1$ **to** $N$ **do**

Sample $x_0^{(i)} \sim q(x_0)$.

**end for**

**for** $t := 1$ **to** $T$ **do**

**Step 2:** importance sampling

**for** $i := 1$ **to** $N$ **do**

Sample $\tilde{x}_t^{(i)} \sim p(x_t|x_{t-1}^{(i)})$.

Set $\tilde{x}_{0:t}^{(i)} := \left(x_{0:t-1}^{(i)}, \tilde{x}_t^{(i)}\right)$.

Compute the corresponding importance weights

$$\tilde{w}_t^{(i)} = p(y_t|\tilde{x}_t^{(i)}).$$

**end for**

Normalise the importance weights

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^{N} \tilde{w}_t^{(j)}}.$$

Compute the effective sample size

$$\hat{N}_{ESS} = \left(\sum_{i=1}^{N} \tilde{w}_t^{(i)2}\right)^{-1}.$$

Approximate the filtering density

$$\hat{p}(x_t|y_{0:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{\tilde{x}_t^{(i)}}(x_t).$$

Compute the required estimate

$$\hat{f}_t = \sum_{i=1}^{N} f(\tilde{x}_t^{(i)}) w_t^(i).$$

**Step 3:** resampling

Resample with replacement $N$ particles, $\left\{x_t^{(i)}\right\}_{i=1}^{N}$, from $\left\{\tilde{x}_t^{(i)}\right\}_{i=1}^{N}$ with corresponding probabilities $\left\{w_t^{(i)}\right\}_{i=1}^{N}$.

Set $x_{0:t}^{(i)} := \left(x_{0:t-1}^{(i)}, x_t^{(i)}\right)$.

Set $w_t^{(i)} := \frac{1}{N}$.

**end for**

# Chapter 4

# Convergence

In Section 3.4 we have shown the necessity of resampling due to the unavoidable weight degeneracy problem. Thus, any particle filter consists of two complimentary steps: the importance sampling step, and the resampling step. However, repeated resampling leads to noticeable complications in the treatment of the theoretical properties of particle filters. Indeed, as pointed out in Crisan and Doucet (2002), the key difference between particle filtering methods and classical MC methods is the sample nature. The latter are based on the assumption that the samples are i.i.d., while in the former approach, the samples, i.e. the particles, interact, which means they are not independent. Thus, the classical limit theorems based on independent samples are not applicable.

This gives rise to posing numerous questions, mostly related to the limiting properties of particle filters. Foremost, one is interested whether particle filters converge asymptotically, with $N \to \infty$, and if so, in what sense. Furthermore, it is worth investigating whether standard MC convergence rates apply and whether the error accumulates over time. Answering all these questions is still a topic of the vast ongoing research and definitely is beyond the scope of this thesis. Therefore, we will focus only on the most crucial problem, i.e. asymptotic convergence of particle filters.

Because the output of any SMC algorithm is an approximation to the posterior distribution given by a discrete *random* measure (linear combination of Dirac measures), it is necessary to define in what way random measures may converge to another measure. This is what we deal with in Section 4.2. Before that, however, we restate the optimal filtering problem in a probabilistic parlance, more suitable for theoretical analysis (as it has already been indicated at the beginning of Chapter 2). Finally, we present two results on convergence, which are mainly taken form Crisan (2001) and Crisan (2014).

## 4.1 The filtering problem revisited

In this section we rephrase the filtering problem from Subsection 2.2.2, and the recursion formulae presented there and Subsection 3.3 in a more compact and precise way, more suitable for convergence analysis.

### 4.1.1 Basic notation

Recall that in 2.2.2 we have specified the filtering problem as computing of the conditional distribution of the signal process $X$ given the $\sigma$-algebra generated by the observation process $Y$[1], from time 0 to the

[1] Which, by definition is equivalent to the this conditional distribution given $Y_{0:t}$, cf. Billingsley (1995).

current time $t$. Equivalently, one can state it as computing the *random probability measure* $p_{t|t}$ such that

$$\begin{aligned} p_{t|t}(A) &= \mathbb{P}(X_t \in A|\sigma(Y_{0:t})), & \forall A \in \mathcal{B}(\mathbb{R}^{n_x}), \\ p_{t|t}f &= \mathbb{E}[f(X_t)|\sigma(Y_{0:t})], & \forall f \in B(\mathbb{R}^{n_x}). \end{aligned}$$

It is worth specifying the difference between *random* and *deterministic* probability measures arising in the current setup. Since the solution to the filtering problem is a *conditional* distribution of the state process, given the $\sigma$-algebra generated by the observation process, up to time $t$, it is clearly a random variable. More precisely, it is a *random measure*. However, if one treats the observations as given, i.e. considers an arbitrary but fixed realisation $y_{0:T}$ of the observation process $Y_{0:T}$, where $T < \infty$ is a large time horizon, then the filtering problem boils down to finding of the conditional probability given by

$$\begin{aligned} p^{y_{0:t}}_{t|t}(A) &= \mathbb{P}(X_t \in A|Y_{0:t} = y_{0:t}), & \forall A \in \mathcal{B}(\mathbb{R}^{n_x}), \\ p^{y_{0:t}}_{t|t}f &= \mathbb{E}[f(X_t)|Y_{0:t} = y_{0:t}], & \forall f \in B(\mathbb{R}^{n_x}). \end{aligned}$$

which is a *deterministic* probability measure. Notice, that $p^{y_{0:t}}_{t|t}$ is the counterpart of the filtering distribution $p(dx_t|y_{0:t})$ discussed in Subsection 2.2.2 and characterised by means of densities first in (2.2.6) and then by the Bayesian recursion: prediction (2.2.7) and updating (2.2.8). Hence, it seems natural to formulate the current framework in a similar, Bayesian spirit, for which we need to define the "predictive" measures, i.e. conditional probability measures given the past observations. These, in a random and deterministic versions, are respectively given by

$$\begin{aligned} p_{t|t-1}(A) &= \mathbb{P}(X_t \in A|Y_{0:t-1}), \\ p^{y_{0:t-1}}_{t|t-1}(A) &= \mathbb{P}(X_t \in A|Y_{0:t-1} = y_{0:t-1}), \end{aligned}$$

with the corresponding integrals defined similarly as above. Finally, we define the *likelihood* function

$$\begin{aligned} p^{Y_t}_t(B) &= \mathbb{P}(Y_t \in B|X_t), \\ p^{y_t}_t(B) &= \mathbb{P}(Y_t \in B|X_t = x_t), \end{aligned}$$

$\forall B \in \mathcal{B}(\mathbb{R}^{n_y})$, and we assume $p^{y_t}_t \in \mathcal{C}_b(\mathbb{R}^{n_x})$. Then, the Bayesian recursion (2.2.7) and (2.2.8) can be expressed as

$$\begin{cases} p_{t|t-1} & = p_{t-1|t-1}K_{t-1} \\ \frac{dp_{t|t}}{dp_{t|t-1}} & = \frac{p^{Y_t}_t}{\int p^{Y_t}_t p_{t|t-1}(dx)} \end{cases}, \qquad \begin{cases} p^{y_{0:t-1}}_{t|t-1} & = p^{y_{0:t-1}}_{t-1|t-1}K_{t-1} \\ \frac{dp^{y_{0:t}}_{t|t}}{dp^{y_{0:t-1}}_{t|t-1}} & = \frac{p^{y_t}_t}{\int p^{y_t}_t p_{t|t-1}(dx)} \end{cases}, \tag{4.1.1}$$

where $K_t$ is the transition kernel of the Markovian signal process, as characterised in 2.1.1. The proof of both above recursion can be found in Crisan (2001).

### 4.1.2 Projective product

Notice, that the recurrence formulae for $p_{t|t}$ and $p^{y_{0:t}}_{t|t}$ in (4.1.1) are based on two operations between probability measure on $\mathbb{R}^{n_x}$, involving an intermediate approximation for $p_{t|t-1}$ and $p^{y_{0:t-1}}_{t|t-1}$, respectively. The first one, the *prediction*, is just a simple transformation via the transition kernel $K_{t-1}$. It takes place before the arrival of the new information, so that the $\sigma$-field on which one conditions does not include $Y_t$ or $Y_t = y_t$. The second one, the *updating*, in a non-linear transformation requiring computation of the normalizing constant in both denominators. Not only it poses considerable computational difficulties, but also, in the current form, is notationally inconvenient. Hence, for sake of readability, we introduce

below the concept of the *projective product*, which describes a transformation of a probability measure to another probability measure using a given function.

**Definition 4.1.1** (Projective product)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^n)$ be a finite measure and let $\phi \in B(\mathbb{R}^n)$ be a non-negative function such that $\mu(\phi) := \int \phi d\mu > 0$. The projective product $\phi * \mu$ is the (set) function $\phi * \mu : \mathcal{B}(\mathbb{R}^n) \to \mathbb{R}$ defined as*

$$\phi * \mu(A) = \frac{\int_A \phi(x)\mu(dx)}{\mu(\phi)}, \qquad \forall A \in \mathcal{B}(\mathbb{R}^n).$$

Intuitively, the projective product "normalises" any finite measure into a probability measure. This is formally stated in the following proposition.

**Proposition 4.1.1.** *The projective product $\phi * \mu$ is a probability measure on $\mathcal{B}(\mathbb{R}^n)$.*

*Proof.* By definition, we have $\phi * \mu(\emptyset) = 0$ and $\phi * \mu(\mathbb{R}^n) = 1$. What remains to be verified is thus the countable additivity property. For any $A \in \mathcal{B}(\mathbb{R}^n)$ consider its partition $\{A_i\}_{i=1}^{\infty}$, i.e. $A_1, A_2, \cdots \in \mathcal{B}(\mathbb{R}^n)$, such that $A_j \cap A_k = \emptyset$, $\forall j \neq k$, and $A = \bigcup_{i=1}^{\infty} A_i$. Then, we have by the properties of the integral

$$\begin{aligned}
\int_A \phi d\mu &= \int \mathbb{I}_A \phi d\mu \\
&\int \sum_{i=1}^{\infty} \mathbb{I}_{A_i} \phi d\mu \\
&\overset{(*)}{=} \sum_{i=1}^{\infty} \int \mathbb{I}_{A_i} \phi d\mu \\
&= \sum_{i=1}^{\infty} \int_{A_i} \phi d\mu,
\end{aligned}$$

where in $(*)$ the order of integration can be changed by the Monotone Convergence Theorem, as the function $\mathbb{I}_{A_i}\phi$ are nonnegative random variables. To complete the proof, notice that we have defined the projective product as the above expression divided by $\mu(\phi)$, so division by $\mu(\phi)$ of each summand above yield the countable additivity property. □

Notice that the projective product $\phi * \mu$ is absolutely continuous with respect to $\mu$ and its Radon-Nikodým derivative with respect to $\mu$ is proportional to $\phi$, i.e.

$$\frac{d(\phi * \mu)}{d\mu} = \frac{\phi}{\mu(\phi)},$$

so that $\frac{1}{\mu(\phi)}$ can be understood as the normalising constant.

In (4.1.1) the role of the function $\phi$ from the above definition was played by the likelihood functions $p_t^{Y_t}$ and $p_t^{y_t}$, while the one of the initial measure $\mu$ – the intermediate predictive distributions $p_{t-1|t-1}$ and $p_{t-1|t-1}^{y_y}$. Hence, the concept of the projective product allows us to express (4.1.1) in a more compact way.

$$\begin{cases} p_{t|t-1} &= p_{t-1|t-1} K_{t-1} \\ p_{t|t} &= p_t^{Y_t} * p_{t|t-1} \end{cases}, \qquad \begin{cases} p_{t|t-1}^{y_{0:t-1}} &= p_{t-1|t-1}^{y_{0:t-1}} K_{t-1} \\ p_{t|t}^{y_{0:t}} &= p_t^{y_t} * p_{t|t-1} \end{cases}. \tag{4.1.2}$$

As it has already been pointed out, the nonlinear character of the projective product makes the whole recursion computationally challenging and in practice requires resorting to approximative solutions. This is indeed the idea behind the particle filters.

### 4.1.3 Particle filters

In Chapter 3 we explained particle filters from the algorithmic point of view and showed how they aim at approximating the conditional filtering distributions $p_{t|t}$ and $p_{t|t}^{y_t}$ using empirical distributions consisting of particles clouds. Below, we provide a more abstract perspective, where particle filters are regarded as approximative sequences of the probability measure transformation.

We consider the approximative sequences $\left\{p_{t|t-1}^n\right\}_{n=1}^{\infty}$ and $\left\{p_{t|t}^n\right\}_{n=1}^{\infty}$, which are supposed to approximate the $p_{t|t-1}$ or $p_{t|t-1}^{y_{t-1}}$, and $p_{t|t}$ or $p_{t|t}^{y_t}$, respectively. We assume that these approximating sequences satisfy the following conditions

1. $p_{t|t}^n$ and $p_{t|t-1}^n$ are *random measures* (not necessarily probability measures);
2. $p_{t|t}^n \neq 0$ and $p_{t|t-1}^n \neq 0$;
3. $p_{t|t-1}^n p_t^y > 0$, $\forall n$, $t = 0, \ldots, T$.

Finally, we define two random probability measures

$$\bar{p}_{t|t}^n = p_t^{Y_t} * p_{t|t-1}^n, \tag{4.1.3}$$

$$\bar{p}_{t|t}^{n,y_t} = p_t^{y_t} * p_{t|t-1}^n, \tag{4.1.4}$$

which by the properties of the projective product are absolutely continuous with respect to $p_{t|t-1}^n$. We have

$$\bar{p}_{t|t}^n f = \frac{p_{t|t-1}^n (f p_t^{Y_t})}{p_{t|t-1}^n p_t^y},$$

$$\bar{p}_{t|t}^{n,y_t} f = \frac{p_{t|t-1}^n (f p_t^{y_t})}{p_{t|t-1}^n p_t^y}.$$

Later on, at the discussion of the bootstrap filter in Section 4.4, it will turn out that the measures (4.1.3) and (4.1.4) can be seen as the weighted (with the normalised weights) measures. They are obtained by weighting of the initial measures, which one samples from, with the weights of the draws, which depend on the corresponding likelihood evaluations.

The question is, whether and under what conditions $p_{t|t-1}^n$ and $p_{t|t}^n$ converge to $p_{t|t-1}$ or $p_{t|t-1}^{y_{t-1}}$, and $p_{t|t}$ or $p_{t|t}^{y_t}$, respectively. The idea of the proofs given in Section 4.3 is presented in Figure 4.1.1, which is based on Crisan (2014). It also illustrates how the transformation of probability measures by particle filters corresponds to its theoretical counterpart. For simplicity, we omit the superscripts $y_t$ for the fixed observation case, as well as we use the common notation $p_t^y$ for both $p_t^{y_y}$ and $p_t^{Y_t}$. Before proceeding to the theorems, however, one needs to define in *what sense* particle filters can converge. This is the aim of the next Section.

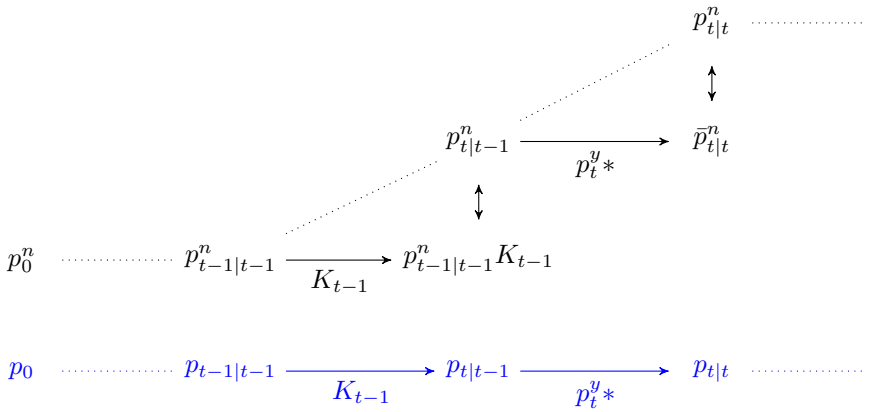



Figure 4.1.1: Graphical representation of the recursive formula for the target process (blue) and approximating sequences (black).

## 4.2 Convergence of measure-valued random variables

Recall that a particle filter is an approximative random measure. To evaluate its quality, one needs thus to define in what way a sequence of random measures can approximate another random measure. Below we analyse two modes of convergence, convergence in expectation and almost sure convergence. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\{\mu_n\}_{n=1}^{\infty}$ be a sequence of random measures, $\mu_n : \Omega \to \mathcal{M}(\mathbb{R}^d)$ and $\mu \in \mathcal{M}(\mathbb{R}^d)$ be a deterministic finite measure.

**Definition 4.2.1** (Convergence in expectation)**.** *A sequence of random measures $\{\mu_n\}_{n=1}^{\infty}$ is said to converge to a random measure $\mu$ in expectation if*

$$\lim_{n\to\infty} \mathbb{E}\left[|\mu_n f - \mu f|\right] = 0, \qquad \forall f \in \mathcal{C}_b(\mathbb{R}^d),$$

*which we denote by $e\lim_{n\to\infty} \mu_n = \mu$.*

**Definition 4.2.2** (Almost sure convergence)**.** *A sequence of random measures $\{\mu_n\}_{n=1}^{\infty}$ is said to converge to a random measure $\mu$ $\mathbb{P}$–almost surely if*

$$\mathbb{P}\left(\lim_{n\to\infty} \mu_n = \mu\right) = 1,$$

*which we denote by $\mu_n \xrightarrow{\mathbb{P}} \mu$.*

Notice, that by the Dominated Convergence Theorem, if there exists an integrable random variable $\omega : \Omega \to \mathbb{R}$ satisfying $\mu_n \mathbf{1} \leq \omega$, $\forall n$, then

$$\mu_n \xrightarrow{\mathbb{P}} \mu \quad \Rightarrow \quad e\lim_{n\to\infty} \mu_n = \mu.$$

This automatically holds for probability measures, as one can simply take $\omega \equiv \mathbf{1}$.

For almost sure convergence analysis, it is convenient to define a *distance* $d_{\mathcal{M}}$ on $\mathcal{P}(\mathbb{R}^{n_x})$, in order to obtain a *metric space* $(\mathcal{P}(\mathbb{R}^{n_x}), d_{\mathcal{P}})$. To this end, first we consider, as in Crisan and Doucet (2002), $\mathbb{R}^{n_x}$ endowed with the *topology of weak convergence*, in which $\mu_n$ converges weakly to $\mu$ if

$$\lim_{n\to\infty} \mu_n \varphi = \mu\varphi, \qquad \forall \varphi \in \mathcal{C}_b(\mathbb{R}^{n_x}),$$

which we denote by $\lim_{n\to\infty} \mu_n = \mu$. Since $\mathbb{R}^{n_x}$ is locally compact separable metric space, there exists a

countable set $\mathcal{A} = \{\varphi_i\}_{i=1}^{\infty}$, dense in $\mathcal{C}_b(\mathbb{R}^{n_x})$, which completely determines convergence, i.e.

$$\lim_{n\to\infty} \mu_n = \mu \iff \lim_{n\to\infty} \mu_n\varphi_i = \mu\varphi_i, \qquad \forall \varphi_i \in \mathcal{A}. \tag{4.2.1}$$

This *convergence determining set* allows us to define the required distance on $\mathcal{P}(\mathbb{R}^{n_x})$ as follows

$$d_{\mathcal{P}}(\mu,\nu) = \sum_{i=1}^{\infty} \frac{|\mu\varphi_i - \nu\varphi_i|}{2^i ||\varphi_i||}, \tag{4.2.2}$$

where $||\cdot||$ stands for the supremum norm. This set generates the weak topology on $\mathcal{A}(\mathbb{R}^{n_x})$, i.e.

$$\lim_{n\to\infty} \mu_n = \mu \iff \lim_{n\to\infty} d_{\mathcal{M}}(\mu_n, \mu) = 0.$$

Notice, that although the distance $d_{\mathcal{P}}$ depends on the choice of the set $\mathcal{P}$, the induced topology is independent of this choice.

## 4.3 Convergence theorems

Below we present and prove two theorems, which provide necessary and sufficient conditions for convergence of the particle filter to the posterior distribution. Due to the limited character of this thesis we restrict ourselves to fixed observation case, which is more natural from the practical point of view. To simplify the notation we omit the superscripts $y_{t-1}$ and $y_t$, although the results below are derived for a given realisation $y_{0:T}$ of the process $Y_{0:T}$.

Both theorems turn out to be rather natural: an approximation to the target distribution can be obtained if and only if we start nearby the required distribution and then appropriately closely follow the particle filter recursions. Hence, the key assumptions here are the Feller property of the transition kernel together with the continuity and boundedness of the likelihood function. The former requirement ensures that the signal process moves continuously and that its trajectories are stable, in a sense that two realisations of $X$ which start from nearby states will stay close to each other throughout the whole horizon. The latter assumption guarantees that a small change in the prior distribution of the signal does not lead to a substantial variation in its posterior distribution.

### 4.3.1 Convergence in expectation

The following theorem gives necessary and sufficient conditions for the convergence of $p^n_{t|t}$ to $p_{t|t}$ (filtering distribution) and $p^n_{t|t-1}$ to $p_{t|t-1}$ (predictive distribution) in expectation.

**Theorem 4.3.1.** *The sequences $p^n_{t|t}$ and $p^n_{t|t-1}$ converge in expectation to $p_{t|t}$ and $p_{t|t-1}$, respectively, i.e.*

*a0.* $\lim_{n\to\infty} \mathbb{E}\left[\left|p^n_{t|t}f - p_{t|t}f\right|\right] = 0,$

*b0.* $\lim_{n\to\infty} \mathbb{E}\left[\left|p^n_{t|t-1}f - p_{t|t-1}f\right|\right] = 0,$

*if and only if*

*a1.* $\lim_{n\to\infty} \mathbb{E}\left[\left|p^n_{0|0}f - p_{0|0}f\right|\right] = 0,$

*b1.* $\lim_{n\to\infty} \mathbb{E}\left[\left|p^n_{t|t-1}f - p^n_{t-1|t-1}K_t f\right|\right] = 0,$

*c1.* $\lim_{n\to\infty} \mathbb{E}\left[\left|p^n_{t|t}f - \bar{p}_{t|t}f\right|\right] = 0,$

$\forall f \in \mathcal{C}_b(\mathbb{R}^{n_x})$, $\forall t \in [0,T]$.

*Proof.* ($\Leftarrow$) We will proceed by induction. The initial case for *a0.* follows from *a1.* We need to show that *a0.* also holds for $t > 0$, i.e. that if $e\lim_{n\to\infty} p^n_{t-1|t-1} = p_{t-1|t-1}$ and $e\lim_{n\to\infty} p^n_{t|t-1} = p_{t|t-1}$, then also $e\lim_{n\to\infty} p^n_{t|t} = p_{t|t}$.

Consider *b0.*. From (4.1.2) we have $p_{t|t-1} = p_{t-1|t-1}K_t$, hence by the triangle inequality, $\forall n$, $\forall f \in \mathcal{C}_b(\mathbb{R}^{n_x})$,

$$\left|p^n_{t|t-1}f - p_{t|t-1}f\right| \leq \left|p^n_{t|t-1}f - p^n_{t-1|t-1}K_tf\right| + \left|p^n_{t-1|t-1}K_tf - p_{t-1|t-1}K_tf\right|. \tag{4.3.1}$$

Then, by *b1.*, the first term on the right hand side in the above formula converges in expectation to 0.

Due to the Feller property of the kernel, $K_tf \in \mathcal{C}_b(\mathbb{R}^{n_x})$, $\forall f \in \mathcal{C}_b(\mathbb{R}^{n_x})$, so by the induction hypothesis the second term on the right hand side in (4.3.1) also converges in expectation to 0. Thus, we obtain

$$\lim_{n\to\infty} \mathbb{E}\left[\left|p^n_{t|t-1}f - p_{t|t-1}f\right|\right] = 0, \tag{4.3.2}$$

which is *b0.*. To show *a0.* we want to use *c1.*. Thus, recall that for $\bar{p}^n_{t|t}$ we have

$$\bar{p}^n_{t|t}f = \frac{p^n_{t|t-1}fp^y_t}{p^n_{t|t-1}p^y_t},$$

and consider

$$\begin{aligned}
\left|\bar{p}^n_{t|t}f - p_{t|t}f\right| &= \left|\frac{p^n_{t|t-1}fp^y_t}{p^n_{t|t-1}p^y_t} - \frac{p_{t|t-1}fp^y_t}{p_{t|t-1}p^y_t}\right| \\
&\leq \left|\frac{p^n_{t|t-1}fp^y_t}{p^n_{t|t-1}p^y_t} - \frac{p^n_{t|t-1}fp^y_t}{p_{t|t-1}p^y_t}\right| + \left|\frac{p^n_{t|t-1}fp^y_t}{p_{t|t-1}p^y_t} - \frac{p_{t|t-1}fp^y_t}{p_{t|t-1}p^y_t}\right| \\
&\leq \frac{||f||}{p_{t|t-1}p^y_t}\left|p^n_{t|t-1}p^y_t - p_{t|t-1}p^y_t\right| + \frac{1}{p_{t|t-1}p^y_t}\left|p^n_{t|t-1}fp^y_t - p_{t|t-1}fp^y_t\right|,
\end{aligned} \tag{4.3.3}$$

where $||f||$ denotes the supremum norm of $f$. Hence,

$$\mathbb{E}\left[\left|\bar{p}^n_{t|t}f - p_{t|t}f\right|\right] \leq \frac{||f||}{p_{t|t-1}p^y_t}\mathbb{E}\left[\left|p^n_{t|t-1}p^y_t - p_{t|t-1}p^y_t\right|\right] + \frac{1}{p_{t|t-1}p^y_t}\mathbb{E}\left[\left|p^n_{t|t-1}fp^y_t - p_{t|t-1}fp^y_t\right|\right].$$

Since $p^y_t \in \mathcal{C}_b(\mathbb{R}^{n_x})$[2], by (4.3.2) both terms on the right hand side of the above formula converge to 0, implying that

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\bar{p}^n_{t|t}f - p_{t|t}f\right|\right] = 0. \tag{4.3.4}$$

Finally, by the triangle inequality, we have

$$\left|p^n_{t|t}f - p_{t|t}f\right| \leq \left|p^n_{t|t}f - \bar{p}^n_{t|t}f\right| + \left|\bar{p}^n_{t|t}f - p_{t|t}f\right|. \tag{4.3.5}$$

Both terms on the right hand side of (4.3.5) converge in expectation to 0: the first one by *c1.*, while the second one by (4.3.4). Thus

$$\lim_{n\to\infty} \mathbb{E}\left[\left|p^n_{t|t}f - p_{t|t}f\right|\right] = 0,$$

[2]Recall that in Section 2.1.2 we have assumed that $p^y_t = p(y_t|x_t)$, the observation density is bounded and continuous.

which is *a0.*.

($\Rightarrow$) Assume *a0.* and *b0.* both hold. Then *a1.* is satisfied as a special case of *a0.*

Next, using again the triangle inequality, we obtain

$$\mathbb{E}\left[\left|p^n_{t|t}f - \bar{p}^n_{t|t}f\right|\right] \leq \mathbb{E}\left[\left|p^n_{t|t}f - p_{t|t}f\right|\right] + \mathbb{E}\left[\left|p_{t|t}f - \bar{p}^n_{t|t}f\right|\right].$$

By (4.3.4), the second term in the above expression converges to 0, while the first one goes to 0 by assumption *a0*, so

$$\lim_{n\to\infty} \mathbb{E}\left[\left|p^n_{t|t}f - \bar{p}^n_{t|t}f\right|\right] = 0, \tag{4.3.6}$$

establishing *c1.*.

To complete the proof, notice that due to $p_{t|t-1} = p_{t-1|t-1}K_t$ we have

$$\mathbb{E}\left[\left|p^n_{t|t-1}f - p^n_{t-1|t-1}K_tf\right|\right] \leq \mathbb{E}\left[\left|p^n_{t|t}f - p_{t|t}f\right|\right] + \mathbb{E}\left[\left|p_{t-1|t-1}K_tf - p^n_{t-1|t-1}K_{t-1}f\right|\right].$$

The first term in this formula converges to 0 by (4.3.6). For the second term

$$\mathbb{E}\left[\left|p_{t-1|t-1}K_tf - p^n_{t-1|t-1}K_tf\right|\right] = \mathbb{E}\left[\left|p_{t|t-1}f - p^n_{t|t-1}f\right|\right],$$

so it goes to 0 by assumption *b0.* Hence

$$\lim_{n\to\infty} \mathbb{E}\left[\left|p^n_{t|t-1}f - p^n_{t-1|t-1}K_tf\right|\right] = 0,$$

yielding *b1.*. □

### 4.3.2 Almost sure convergence

The following theorem is a counterpart of Theorem 4.3.1 and characterises necessary and sufficient conditions for the almost sure convergence of $p^n_{t|t}$ and $p^n_{t|t-1}$ to $p_{t|t}$ and $p_{t|t-1}$, respectively.

It makes use of the fact that the almost sure convergence from Definition 4.2.2 is equivalent to

$$\mathbb{P}\left(\lim_{n\to\infty} d_{\mathcal{P}}(\mu_n, \mu) = 0\right) = 1,$$

where $d_{\mathcal{P}}$ is the distance in the space of probability measures over $\mathbb{R}^{n_x}$, as defined in (4.2.2).

**Theorem 4.3.2.** *The sequences $p^n_{t|t}$ and $p^n_{t|t-1}$ converge almost surely to $p_{t|t}$ and $p_{t|t-1}$, respectively, i.e.*

*a0.* $\lim_{n\to\infty} d_{\mathcal{P}}\left(p^n_{t|t}, p_{t|t}\right) = 0$, $\mathbb{P}$-*a.s.*,

*b0.* $\lim_{n\to\infty} d_{\mathcal{P}}\left(p^n_{t|t-1}, p_{t|t-1}\right) = 0$, $\mathbb{P}$-*a.s.*

*if and only if*

*a1.* $\lim_{n\to\infty} p^n_{0|0} = p_{0|0}$,, $\mathbb{P}$-*a.s.*,

*b1.* $\lim_{n\to\infty} d_{\mathcal{P}}\left(p^n_{t|t-1}, p^n_{t-1|t-1}K_t\right) = 0$, $\mathbb{P}$-*a.s.*,

*c1.* $\lim_{n\to\infty} d_{\mathcal{P}}\left(p^n_{t|t}, \bar{p}_{t|t}\right) = 0$, $\mathbb{P}$-*a.s.*.

*Proof.* ($\Leftarrow$) Since for the *probability* measures, the almost sure convergence implies the convergence in expectation, if we assume that conditions *a1.-c1.* hold, then the corresponding conditions from Theorem 4.3.1 are also satisfied. Hence, we have that $\forall f \in \mathcal{C}_b(\mathbb{R}^{n_x})$ inequalities (4.3.1), (4.3.3) and (4.3.5) hold. Hence, in particular they hold for $f \in \mathcal{A}$ as defined in Section 4.2. Then, by induction, we obtain *a0.* and *b0.*.

($\Rightarrow$) Suppose that $p^n_{t|t} \xrightarrow{\mathbb{P}} p_{t|t}$ and $p^n_{t|t-1} \xrightarrow{\mathbb{P}} p_{t|t-1}$, $\forall t$.

Since $p_{t|t-1} = p_{t-1|t-1}K_t$, we have that $p^n_{t-1|t-1}K_t \xrightarrow{\mathbb{P}} p_{t-1|t-1}K_t$. Then, by (4.3.3), also $\bar{p}_{t|t} \xrightarrow{\mathbb{P}} p_{t|t}$. Therefore, almost surely,

$$\begin{aligned}
\lim_{n\to\infty} d_{\mathcal{P}}\left(p^n_{t|t-1}, p_{t|t-1}\right) &= 0,\\
\lim_{n\to\infty} d_{\mathcal{P}}\left(p^n_{t|t}, p_{t|t}\right) &= 0,\\
\lim_{n\to\infty} d_{\mathcal{P}}\left(p^n_{t-1|t-1}K_t, p_{t|t-1}\right) &= 0,\\
\lim_{n\to\infty} d_{\mathcal{P}}\left(\bar{p}^n_{t|t}, p_{t|t}\right) &= 0.
\end{aligned}$$

Then, by the triangle inequality we obtain

$$\begin{aligned}
d_{\mathcal{P}}\left(p^n_{t|t-1}, p^n_{t-1|t-1}K_t\right) &\leq d_{\mathcal{P}}\left(p^n_{t|t-1}, p_{t|t-1}\right) + d_{\mathcal{P}}\left(p^n_{t-1|t-1}K_t, p_{t|t-1}\right),\\
d_{\mathcal{P}}\left(p^n_{t|t}, p_{t|t}\right) &\leq d_{\mathcal{P}}\left(\bar{p}^n_{t|t}, p^n_{t|t}\right) + d_{\mathcal{P}}\left(\bar{p}^n_{t|t}, p_{t|t}\right).
\end{aligned}$$

Taking the limits on both sides in each of the above formulae gives that almost surely

$$\begin{aligned}
\lim_{n\to\infty} d_{\mathcal{P}}\left(p^n_{t|t-1}, p^n_{t-1|t-1}K_t\right) &= 0,\\
\lim_{n\to\infty} d_{\mathcal{P}}\left(p^n_{t|t}, p_{t|t}\right) &= 0,
\end{aligned}$$

which yields *b1.* and *c1.*. □

## 4.4 Bootstrap example

Below, following Crisan (2001), we present an example of a particle filter which satisfies the assumption of the above theorems. It is a version of the bootstrap filer described in Algorithm 3. This is a very simplistic case, discussed for purely illustrative purposes. Theoretical analysis of more complex filters is beyond the scope of this thesis.

### 4.4.1 Particle filter

Consider a collection of $n$ particles which evolve according to a given Markov transition kernel $K_t$. Each period we perform multinomial resampling with replacement, where the sampling probabilities are set equal to the particles' importance weights. The exact steps of the procedure are specified as follows.

**Initialization.** Draw a random sample of size $n$ from the initial distribution $p_0$, where the initial weight of each particle is equal $1/n$. This results in the initial empirical distribution

$$p^n_0 = \frac{1}{n}\sum_{i=1}^{n} \delta_{\{x_0^{(i)}\}},$$

where $x_0^{(i)}$ denotes the position of the $i$-th draw. Since the particles are randomly drawn from $p_0$, the empirical measure $p_0^n$ is a random measure for which we have

$$\lim_{n\to} \mathbb{E}\left[p_0^n\right] = p_0,$$
$$\lim_{n\to} p_0^n = p_0,\ \mathbb{P}-\text{a.s.}.$$

**Iteration.** Given the approximation

$$p_{t-1|t-1}^n = \frac{1}{n}\sum_{i=1}^{\infty} \delta_{\{x_{t-1}^{(i)}\}},$$

i.e. the empirical measure corresponding to the particles collection from the previous step, recursively construct $p_{t|t}^n$ as follows. Move each particle according to the signal transition kernel

$$\tilde{x}_t^{(i)} \sim K_{t-1}(x_{t-1}^{(i)}, \cdot).$$

Notice, that the particles move independently of each other. We obtain a new empirical measure, $p_{t|t}^n$ given by

$$p_{t|t}^n = \frac{1}{n}\sum_{i=1}^{\infty} \delta_{\{\tilde{x}_t^{(i)}\}}.$$

For each particle compute the (normalised) importance weight

$$w_t^{(i)} = \frac{p(y_t|\tilde{x}_t^{(i)})}{\sum_{j=1}^n p(y_t|\tilde{x}_t^{(j)})}$$

where $y_t$ is the fixed realisation of the observation process, as assumed above. Obtain yet another empirical measure, corresponding to (4.1.4), given by

$$\bar{p}_{t|t}^n = \sum_{i=1}^{n} w_t^{(i)} \delta_{\{\tilde{x}_t^{(i)}\}}.$$

Perform multinomial resampling from $\{\tilde{x}_t^{(i)}\}_{i=1}^n$ with the probabilities equal to the weights $\{w_t^{(i)}\}_{i=1}^n$, call the obtained particles $x_t^{(i)}$, $i = 1, \ldots, n$. Finally, we obtain the required approximation, which is given by

$$p_{t|t}^n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\{x_t^{(i)}\}}.$$

### 4.4.2 Resampling

Notice that the resampling step can be seen as replacing each particle by a number of offspring particles, call them $\xi_t^{(i)}$, where each particle has on average $w_t^{(i)}$ offspring. For simplicity and consistency with Algorithm 3 we impose the constraint of the constant number of particles, i.e. $\sum_i^n \xi_t^{(i)} = n$. In general, one needs to assume that the resulting offspring vector $\xi_t = \left(\xi_t^{(i)}\right)_{i=1}^n$ has a finite variance, i.e. $\exists c_t \in \mathbb{R}$

$$v^T \widehat{\text{Var}}[\xi_t] v \le n c_t, \quad \forall v \in \mathbb{R}^n,\ v = \left(v^i\right)_{i=1}^n,\ |v^i| < 1,\ i = 1, \ldots, n, \tag{4.4.1}$$

where

$$\widehat{\text{Var}}[\xi_t]\mathbb{E}\left[\left(\xi_t - w_t\right)^T \left(\xi_t - w_t\right)\right]$$

denotes the covariance matrix of $\xi_t$, where $w_t = (w_t^{(i)})_{i=1}^n$.

With multinomial resampling this is indeed the case. Since $\xi_t \sim \text{Multinomial}\left(, w_t^{(1)}, \ldots, w_t^{(n)}\right)$, we have

$$\begin{aligned}
\mathbb{E}[\xi_t^{(i)}] &= w_t^{(i)}, \\
\mathbb{E}\left[\left(\xi_t^{(i)} - w_t^{(i)}\right)^2\right] &= n w_t^{(i)}(1 - w_t^{(i)}), \\
\mathbb{E}\left[\left(\xi_t^{(i)} - w_t^{(i)}\right)^T \left(\xi_t^{(j)} - w_t^{(j)}\right)\right] &= -n w_t^i w_t^j, \qquad i \neq j.
\end{aligned}$$

Then, for all $v$ as characterised above one obtains

$$\begin{aligned}
v^T \widehat{\text{Var}}[\xi_t] v &= n \sum_{i,j=1, i\neq j}^{n} w_t^i (1 - w_t^i)(v^{(i)})^2 - 2n \sum_{i=1}^{n} w_t^i w_t^j v^{(i)} v^{(j)} \\
&= n \sum_{i=1}^{n} w_t^{(i)} (v^{(i)})^2 - n \left( \sum_{i=1}^{n} w_t^{(i)} v^{(i)} \right)^2 \\
&\leq n \sum_{i=1}^{n} w_t^{(i)} \\
&= n,
\end{aligned}$$

so that we can take $c_t = 1$.

### 4.4.3 Convergence

As above, we fix the observations to a given path $y_{0:T}$. We show that the random measures $p_{t|t}^n$ and $p_{t|t-1}^n$ delivered by the bootstrap filter described in Subsections 4.4.1 and 4.4.2 converge to $p_{t|t}^{y_t}$ and $p_{t|t-1}^{y_{t-1}}$, respectively. To this end we define the following two $\sigma$-algebras

$$\begin{aligned}
\mathcal{F}_t &= \sigma\left\{\tilde{x}_s^i, x_s^i,\ s \leq t, \quad i = 1, \ldots, n\right\}, \\
\bar{\mathcal{F}}_t &= \sigma\left\{\tilde{x}_s^i, x_s^i,\ s < t,\ \tilde{x}_t^{(i)}, \quad i = 1, \ldots, n\right\}.
\end{aligned}$$

Obviously $\bar{\mathcal{F}}_t \subset \mathcal{F}_t$. Further, notice that the random measure $p_{t|t}^n$ is $\mathcal{F}_t$-measurable, while the random measures $p_{t|t-1}^n$ and $\bar{p}_{t|t}^n$ are $\bar{\mathcal{F}}_t$-measurable. Independent resampling in each iteration means that the randomly drawn particle positions $\tilde{x}_t^{(i)}$ are mutually independent conditional upon $\mathcal{F}_{t-1}$.

First, we give and prove the theorem regarding the convergence in the expectation of the bootstrap filer characterised above. Then for completeness, we state its almost sure convergence counterpart. As before, the results are taken from Crisan (2001).

**Theorem 4.4.1.** *Let $\{p_{t|t-1}^n\}_{n=1}^{\infty}$ and $\{p_{t|t}^n\}_{n=1}^{\infty}$ be sequences of random measures obtained via performing the algorithm described in Subsections 4.4.1 and 4.4.2. Then,*

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}[p_{t|t-1}^n] &= p_{t|t-1}^{y_{0:t-1}}, \\
\lim_{n\to\infty} \mathbb{E}[p_{t|t}^n] &= p_{t|t}^{y_{0:t}}.
\end{aligned}$$

*Proof.* We want to use Theorem 4.3.1. Let $f \in \mathcal{C}_b(\mathbb{R}^{n_x})$. Since condition *a1.* holds naturally, we only need to show that the two remaining conditions are satisfied. Regarding *b1.* we have

$$\mathbb{E}\left[ f(\tilde{x}_t^{(i)}) \middle| \mathcal{F}_{t-1} \right] = K_{t-1} f(x_{t-1}^{(i)}), \qquad i = 1, \ldots, n.$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[p^n_{t|t-1}f\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\frac{1}{n}f(\tilde{x}^{(i)}_t)\middle|\mathcal{F}_{t-1}\right]\\ &= \mathbb{E}\left[\frac{1}{n}K_{t-1}f(x^{(i)}_{t-1})\middle|\mathcal{F}_{t-1}\right]\\ &= p^n_{t-1|t-1}K_{t-1}f.\end{aligned}$$

Then, as the particles move independently of each other,

$$\begin{aligned}\mathbb{E}\left[\left(p^n_{t|t-1}f - p^n_{t-1|t-1}K_{t-1}f\right)^2\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n\left[f(\tilde{x}^{(i)}_t) - K_{t-1}f(x^{(i)}_t)\right]\right)^2\middle|\mathcal{F}_{t-1}\right]\\ &\overset{ind}{=} \frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[\left(f(\tilde{x}^{(i)}_t)\right)^2\middle|\mathcal{F}_{t-1}\right] - \frac{1}{n^2}\sum_{i=1}^n\left(\mathbb{E}\left[K_{t-1}f(x^{(i)}_t))\middle|\mathcal{F}_{t-1}\right]\right)^2\\ &= \frac{1}{n}p^n_{t-1|t-1}\left(K_{t-1}f^2 - (K_{t-1}f)^2\right).\end{aligned}$$

Hence,

$$\mathbb{E}\left[\left(p^n_{t|t-1}f - p^n_{t-1|t-1}K_{t-1}f\right)^2\right] \leq \frac{||f||^2_\infty}{n},$$

which shows *b1.*.

Regarding *c1.*, note that

$$p^n_{t|t} = \frac{1}{n}\sum_{i=1}\delta_{\{x^{(i)}_t\}} = \sum_{i=1}^n\xi^{(i)}_t\delta_{\{\tilde{x}^{(i)}_t\}},$$

thus

$$\begin{aligned}\mathbb{E}\left[p^n_{t|t}f\middle|\bar{\mathcal{F}}_t\right] &= \bar{p}^n_{t|t}f,\\ \mathbb{E}\left[\left(p^n_{t|t}f - \bar{p}^n_t f\right)^2\middle|\bar{\mathcal{F}}_t\right] &= \frac{1}{n^2}(v^n_t)^T\widehat{\mathrm{Var}}^n_t[\xi_t]v^n_t,\end{aligned}$$

where $v^n_t = (f(\tilde{x}^{(i)}_t))^n_{i=1}$. Then, by (4.4.1) we have

$$\mathbb{E}\left[\left(p^n_{t|t}f - \bar{p}^n_{t|t}f\right)^2\middle|\bar{\mathcal{F}}_t\right] = \frac{c_t||f||^2}{n},$$

so that

$$\mathbb{E}\left[(p^n_t f - \bar{p}^n_t f)^2\right] = \frac{c_t||f||^2}{n}, \tag{4.4.2}$$

implying that also *c1.* holds. This completes the proof. □

Inequality (4.4.2) implies that the additional randomness introduced to the algorithm by the resampling step, measured by the second moment of $p^n_t f - \bar{p}^n_t f$, tends to zero at the constant rate $1/n$.

The following theorem gives the almost sure counterpart of the previous result. The proof, which we omit here, can be found in Crisan (2001) or Crisan and Doucet (2002). In a nutshell, it consists in applying Theorem 4.3.2 and using the functions from $\mathcal{A}$, the convergence determining set (in the topology of weak convergence on the space of probability measures) to bound the fourth moment of integrals with respect to both sequences, $\{p^n_{t|t-1}\}^\infty_{n=1}$ and $\{p^n_{t|t}\}^\infty_{n=1}$, so that the almost sure convergence follows from the Borel-Cantelli argument.

**Theorem 4.4.2.** *Let* $\{p^n_{t|t-1}\}_{n=1}^{\infty}$ *and* $\{p^n_{t|t}\}_{n=1}^{\infty}$ *be sequences of random measures obtained via performing the algorithm described in Subsections 4.4.1 and 4.4.2. Then,*

$$\lim_{n\to\infty} p^n_{t|t-1} = p^{y_{0:t-1}}_{t|t-1},\ \mathbb{P}-a.s.,$$
$$\lim_{n\to\infty} p^n_{t|t} = p^{y_{0:t}}_{t|t},\ \mathbb{P}-a.s..$$

## 4.5 Discussion

The historically initial convergence results for particle filters discussed in this section are certainly useful, in a sense that they provide a positive answer to the question whether the particle filter converge to the optimal filter. Moreover, they characterise necessary and sufficient conditions under which the approximative solution obtained from the particle filter well characterises the true solution. However, as pointed out in Hu et al. (2008), it is always important to address the issue under what condition a certain results are valid. The problem with the theorems from Section 4.3 is that they hold only for *bounded* function. This excludes from the analysis e.g. the identity function, used to obtain the state estimate, and this estimate is usually the one which we are primarily interested in. The more general results, for a more general class of unbounded function, is provided in the cited paper Hu et al. (2008). However, such a theoretically involved analysis is beyond the scope of this thesis.

# Chapter 5

# Applications

To present the performance of the particle filters, we consider three different models with various specifications of the algorithm. First, we analyse the basic linear Gaussian model, the so called local level model, to illustrate the degeneracy problem and the necessity of resampling. Second, we apply particle filter to a canonical, severely nonlinear yet Gaussian model. We compare performance of the algorithm based on two importance functions: the prior (i.e. the bootstrap filter) and the locally optimal one, based on the local linearisation of the observation equation. The third and the last application concerns the stochastic volatility model, which is a standard example of the nonlinear non-Gaussian state space model. Also in this example we employ two importance functions, the prior and the normal approximation to the optimal importance function (based on the local linearisation of the latter). All the computations were performed in MATLAB R2014b (the code listings are presented in Appendix C).

## 5.1 Basic linear Gaussian problem

As the first illustration of the particle filtering method, let us use an example from Durbin and Koopman (2012) and apply the simple bootstrap filter to the classical *Nile data*. This is the data set of observations from the river Nile expressed as annual flow volume at Aswan from 1871 to 1970. We can model this time series as a *local level model* given by

$$\begin{aligned} X_t &= X_{t-1} + V_t, \qquad V_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_V^2), \\ Y_t &= X_t + W_t, \qquad W_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_W^2), \\ X_1 &\sim \mathcal{N}(x_0, \sigma_0^2) \end{aligned} \tag{5.1.1}$$

for $t = 1, \ldots, T$. We assume that $V_t$, $W_t$ are mutually independent, and independent of $X_1$.

The aim of this example is first, to show that the particle filtering method can yield accurate results even in a simple form of the bootstrap filter. The second objective is to illustrate the importance of resampling. Hence, we first implement the SIS Algorithm 1 with the prior taken as the importance density, to subsequently extend it to the SIR Algorithm 2, with the resampling step being added. For the considered particular case, the latter technique is described in Algorithm 4. Following Durbin and Koopman (2012), we consider $N = 10,000$ particles and set the model parameters as follows: $\sigma_V^2 = 1469.1$, $\sigma_W^2 = 15,099$, $x_0 = 0$, $\sigma_0^2 = 10^7$. Finally, for the algorithm with resampling, we performed resampling in each iteration, so $N_{thres} = N$.

**Algorithm 4** "Nile" bootstrap filter algorithm

**for** $t := 1$ **to** $T$ **do**
  **for** $i := 1$ **to** $N$ **do**
    **if** $t = 1$ **then**
      Sample $\tilde{x}_t^{(i)} \sim \mathcal{N}(\tilde{x}_0, S_0)$.
    **else**
      Sample $\tilde{x}_t^{(i)} \sim \mathcal{N}(\tilde{x}_t^{(i-1)}, \sigma_\eta^2)$.
    **end if**
    Set $\tilde{x}_{0:t}^{(i)} := \left(x_{0:t-1}^{(i)}, \tilde{x}_t^{(i)}\right)$.
    Compute the corresponding importance weights

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-1}^{(i)} \exp\left(-\frac{1}{2}\left(\log 2\pi + \log \sigma_\varepsilon^2 + \frac{(y_t - \tilde{x}_t^2)^2}{\sigma_\eta^2}\right)\right).$$

  **end for**
  Normalise the importance weights

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^N \tilde{w}_t^{(j)}}.$$

  Compute the effective sample size

$$\hat{N}_{ESS} = \left(\sum_{i=1}^N \tilde{w}_t^{(i)2}\right)^{-1}.$$

  Compute the filtered state estimate

$$\hat{x}_t = \sum_{i=1}^N \tilde{x}_t^{(i)} w_t^{(i)}.$$

  Compute the filtered state variance estimate

$$\hat{S}_t = \sum_{i=1}^N \tilde{x}_t^{(i)2} w_t^{(i)} - \hat{x}_t^2.$$

  Resample with replacement $N$ particles, $\left\{x_t^{(i)}\right\}_{i=1}^N$, from $\left\{\tilde{x}_t^{(i)}\right\}_{i=1}^N$ with correspoding probabilities $\left\{w_t^{(i)}\right\}_{i=1}^N$.
  Set $x_{0:t}^{(i)} := \left(x_{0:t-1}^{(i)}, x_t^{(i)}\right)$.
**end for**

The output of the algorithm without resampling is presented in Figure 5.1.1, while the one with resampling – in Figure 5.1.2. It can be inferred from both figures that resampling is indeed necessary to obtain accurate filtering results and to keep the number of active particles at non-negligible level in the long run. Moreover, with resampling the filtered state variance quickly stabilises at a constant level, which is in line with the theory, as the local level model should converge to a steady state. This is, however, not the case without resampling.

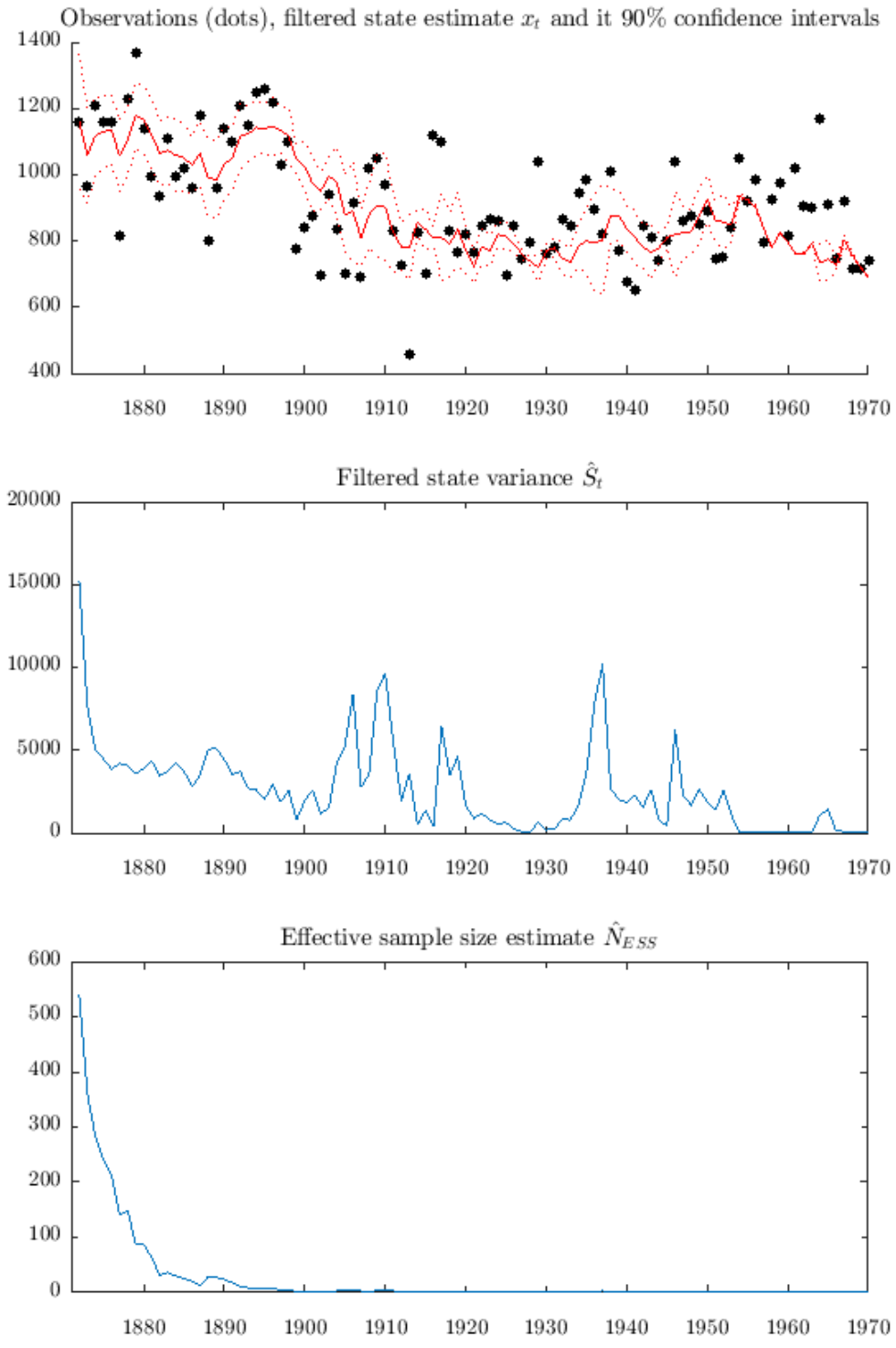



Figure 5.1.1: Bootstrap filter without resampling

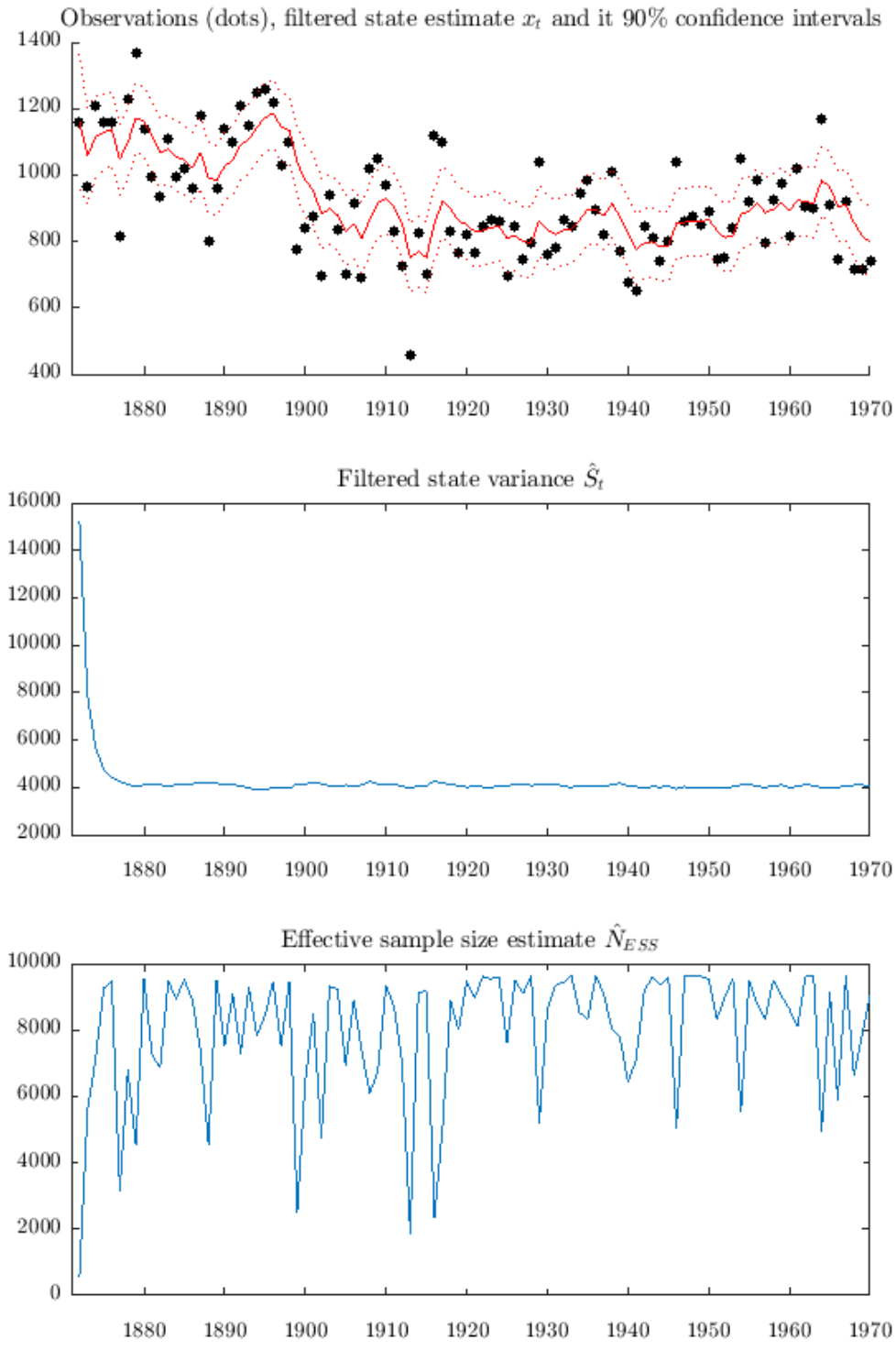



Figure 5.1.2: Bootstrap filter with resampling

## 5.2 Nonlinear Gaussian problem

In their seminal paper Gordon et al. (1993) introduced SMC approach for filtering based on the bootstrap proposal. They considered the following, highly nonlinear model

$$\begin{aligned} X_t &= \frac{X_{t-1}}{2} + 25\frac{X_{t-1}}{1+X_{t-1}^2} + 8\cos(1.2(t-1)) + V_t, \qquad V_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_V^2), \\ Y_t &= \frac{X_t^2}{20} + W_t, \qquad\qquad\qquad\qquad\qquad\qquad W_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_W^2), \\ X_1 &\sim \mathcal{N}(x_0, \sigma_0^2), \end{aligned} \tag{5.2.1}$$

where $V_t$ and $W_t$ are mutually independent,and independent from $X_1$. As in the original paper we set $\sigma_V^2 = 10$, $\sigma_W^2 = 1$, $x_0 = 0$, $\sigma_0^2 = 2$. Figure 5.2.1 presents the generated time series, with the length set to $T = 50$.

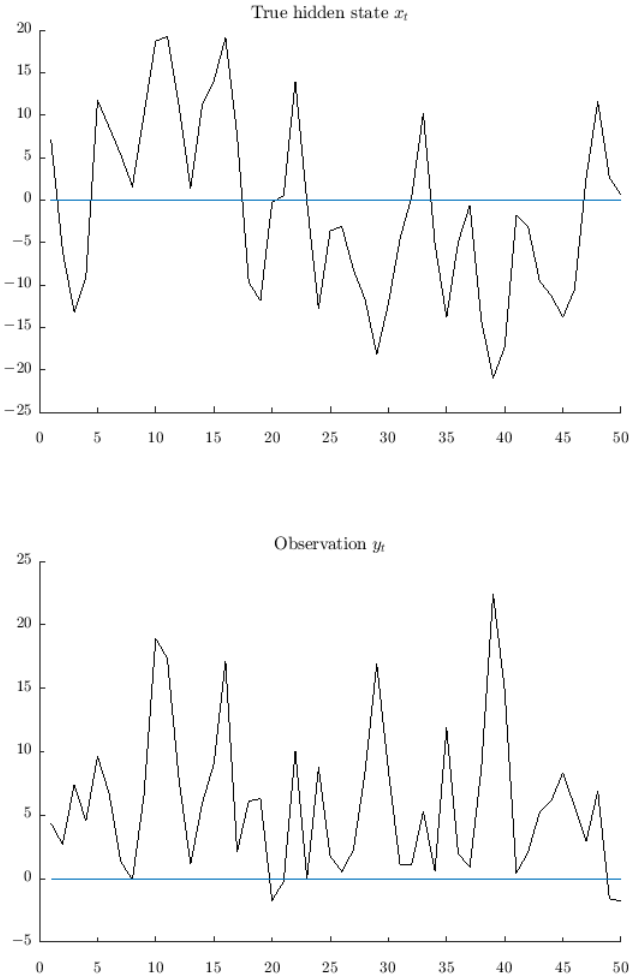



Figure 5.2.1: State and observation generated from (5.2.1).

From the previous example it is clear that resampling is crucial to reduce the degeneracy problem, prevalent in particle filters, and in consequence, to obtain a satisfactory performance of an algorithm. Hence, the aim of the current example is to compare the performance of two algorithms based on different

importance functions. Firstly, following the cited authors, we consider the bootstrap filer with resampling employed at every time step. This approach was also used in Arulampalam et al. (2002). Then, following Doucet et al. (2000), we use the importance function obtained by local linearisation of the state space model, as discussed in Subsection 3.5.2. All the particle filters have $N = 1000$ particles and employ resampling at every time step, i.e. $N_{thres} = N$.

To start with, notice that the transition density and the likelihood density are given by

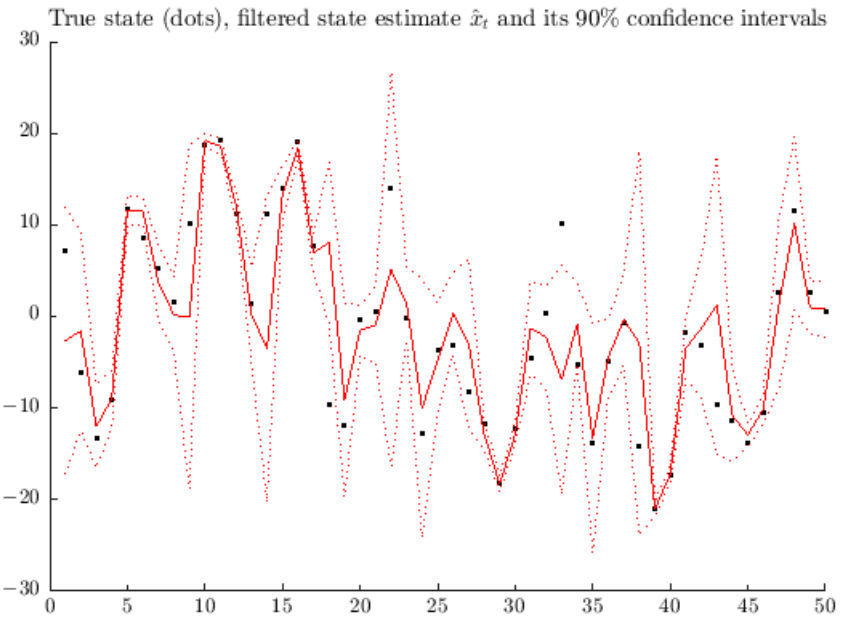
$$\begin{aligned} p(x_t|x_{t-1}) &= \phi\left(x_t; \frac{x_{t-1}}{2} + 25\frac{x_{t-1}}{1+x_{t-1}^2} + 8\cos(1.2(t-1)), \sigma_V^2\right), \\ p(y_t|x_t) &= \phi\left(y_t; \frac{x_t^2}{20}, \sigma_W^2\right), \end{aligned}$$

respectively. Since $p(y_t|x_{t-1})$ cannot be evaluated analytically, and sampling from $p(x_t|x_{t-1}, y_t)$ is infeasible one can resort to obtaining the importance function via local linearisation of the observation equation. Using the notation introduced in 3.5.2, we get

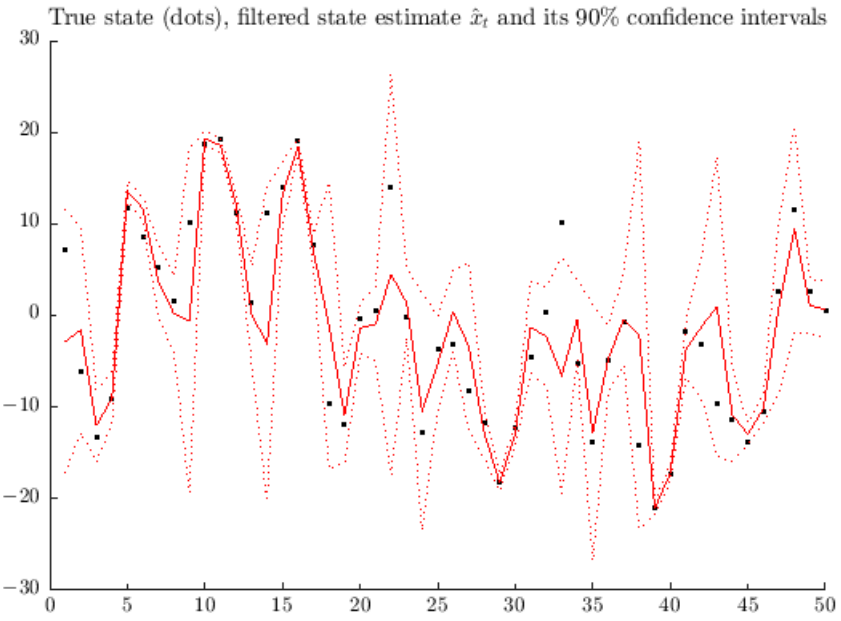
$$\begin{aligned} y_t &\approx g(f(x_{t-1})) + \left.\frac{\partial g(x_t)}{\partial x_t}\right|_{x_t = f(x_{t-1})} (x_t - f(x_{t-1})) + w_t \\ &= \frac{f^2(x_{t-1})}{20} + \frac{f(x_{t-1})}{10}(x_t - f(x_{t-1})) + w_t \\ &= -\frac{f^2(x_{t-1})}{20} + \frac{f(x_{t-1})}{10}x_t + w_t, \end{aligned}$$

which yields the following importance function

$$\begin{aligned} q(x_t|x_{t-1}, y_t) &= \phi(x_t; \mu_t, \sigma_t^2), \\ \sigma_t^{-2} &= \sigma_V^{-2} + \frac{f^2(x_{t-1})}{100}\sigma_W^{-2}, \\ \mu_t &= \sigma_t^2\left[f(x_{t-1})\sigma_V^{-2} + \frac{f(x_{t-1})}{10}\sigma_W^{-2}\left(y_t + \frac{f^2(x_{t-1})}{20}\right)\right]. \end{aligned}$$

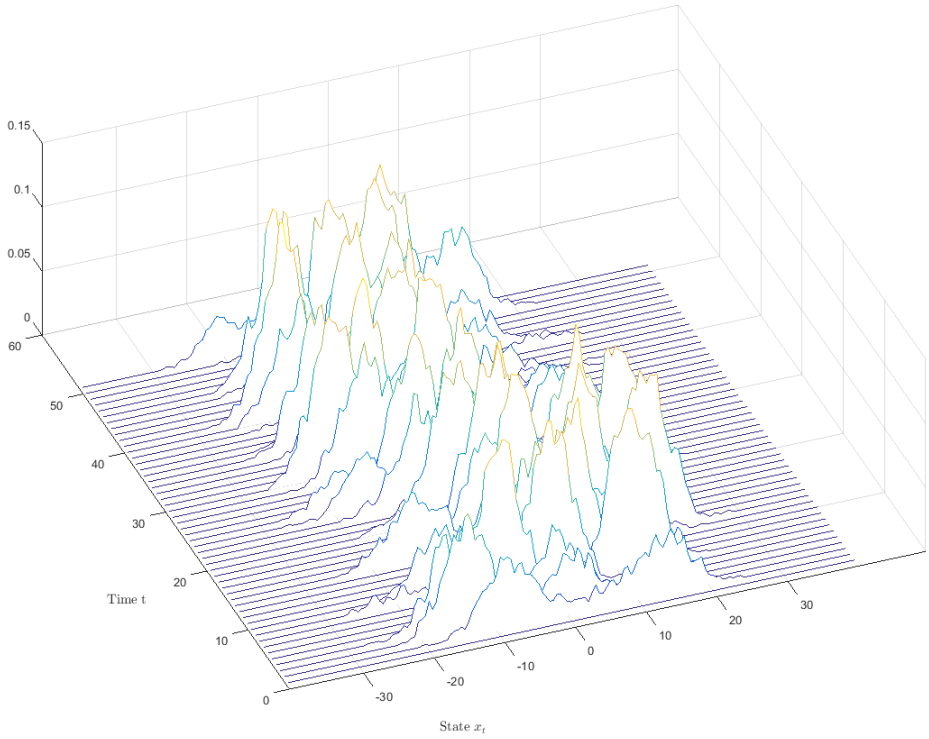



(a) SMC using prior (bootstrap filter).

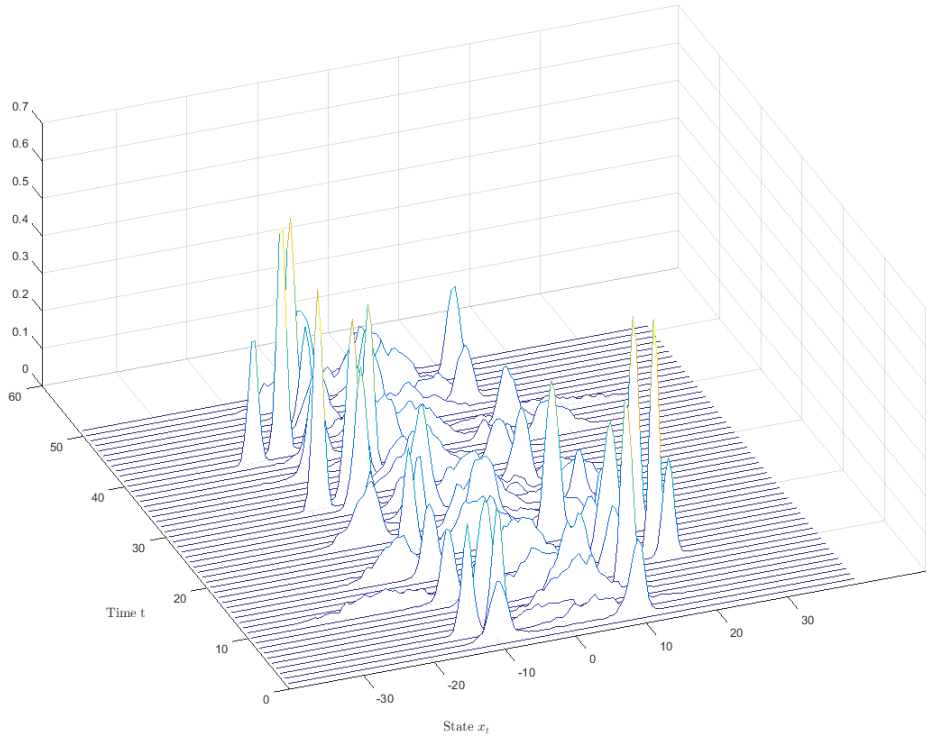



(b) SMC using locally optimal proposal.

Figure 5.2.2: Comparison of the filtering results using different importance functions.




(a) SMC using prior (bootstrap filter).




(b) SMC using locally optimal proposal.

Figure 5.2.3: Estimated filtering distribution using different importance functions.

## 5.3 Stochastic volatility model

As the last example, we consider a standard stochastic volatility (SV) model with the state-space form given by

$$\begin{aligned} X_t &= \phi X_{t-1} + \sigma V_t, \qquad V_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \\ Y_t &= \beta \exp\left(\frac{X_t}{2}\right) W_t, \quad W_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \\ X_1 &\sim \mathcal{N}\Big(0, \frac{\sigma^2}{1-\phi^2}\Big). \end{aligned} \tag{5.3.1}$$

SV models are commonly used to analyse time series of financial returns. Observations in these models are usually daily returns of a certain security, which are defined as the natural logarithm of the ratio of consecutive daily closing levels. Hence, the log-returns are realisations of the stochastic process $Y$. The key feature of SV models is that the variance of this process is itself randomly distributed, i.e. the return $Y_t$ depends on the log-volatility $X_t$, which follows a stationary autoregressive process. This volatility, however, is not observed, and the aim is to retrieve it from the data. A direct modelling of a dynamic process for the variance allows to capture the fundamental property of the observed financial returns, i.e. volatility clustering.

To simulate the time series of the latent log-volatility $x_t$ and observed log-returns $y_t$, We set $\phi = 0.91$, $\sigma = 0.03$ and $\beta = 0.6$, which corresponds to typical values in empirical studies for daily stock returns (cf. Durbin and Koopman, 2012). The generated trajectories are of length $T = 500$ and are presented in Figure 5.3.1 Similarly as in the previous application, to filter out the hidden states we employ particle filters using two importance functions: the prior density (i.e. the bootstrap filter) and the one based on the optimal importance density (using the local linearisation technique). The derivation of the latter is given below. Both filters use $N = 500$ particles and carry out resampling at each point in time, $N_{thres} = N$.

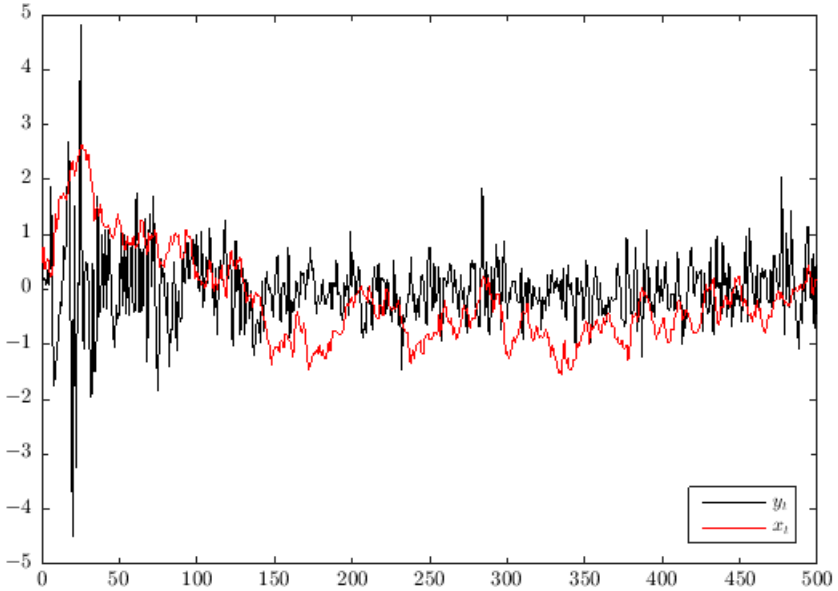



Figure 5.3.1: State and observation generated from (5.3.1)

### 5.3.1 Optimal importance function approximation

It follows from (5.3.1) that the transition density and the likelihood density are respectively given by

$$p(x_t|x_{t-1}) = \mathcal{N}\left(x_t \,\middle|\, \phi x_{t-1}, \frac{\sigma^2}{1-\phi^2}\right), \tag{5.3.2}$$

$$p(y_t|x_t) = \mathcal{N}\left(y_t \,|0, \beta^2 \exp(x_t)\right). \tag{5.3.3}$$

Hence, the optimal proposal density, minimising the weights variance, satisfies

$$\begin{aligned} q(x_t|x_{0:t-1}, y_{0:t}) &\propto p(x_t|x_{t-1})p(y_t|x_t) \\ &= \frac{1}{\sqrt{2\pi\frac{\sigma^2}{1-\phi^2}}} \frac{1}{\sqrt{2\pi\beta^2\exp(x_t)}} \exp\left(-\frac{1}{2}\frac{1-\phi^2}{\sigma^2}(x_t - \phi x_{t-1})^2\right) \exp\left(-\frac{1}{2}\frac{y_t^2}{\beta^2}\exp(-x_t)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(x_t + \frac{1-\phi^2}{\sigma^2}(x_t - \phi x_{t-1})^2 + \frac{y_t^2}{\beta^2}\exp(-x_t)\right)\right). \end{aligned} \tag{5.3.4}$$

Expression (5.3.4) may look similarly to the formula for the Gaussian density, however, the last term in the exponent makes it non-standard. Therefore, in practice it needs to be approximated by, e.g. the normal density, as discussed in Subsection 3.5.2. Using the notation introduced there, we can write

$$\begin{aligned} \ell(x) &= -\log(2\pi\beta) - \frac{1}{2}\log\left(\frac{\sigma^2}{1-\phi^2}\right) - \frac{1}{2}\left(x + \frac{1-\phi^2}{\sigma^2}(x - \phi x_{t-1})^2 + \frac{y_t^2}{\beta^2}\exp(-x)\right), \\ \ell'(x) &= -\frac{1}{2} - \frac{1-\phi^2}{\sigma^2}(x - \phi x_{t-1}) + \frac{y_t^2}{2\beta^2}\exp(-x), \\ \ell''(x) &= -\frac{1-\phi^2}{\sigma^2} - \frac{y_t^2}{2\beta^2}\exp(-x). \end{aligned}$$

Notice, that the optimal importance function is concave, which implies that is has a unique maximiser and that $\ell''(x) < 0$, as required in the assumptions from the Subsection 3.5.2. Putting

$$\begin{aligned} \hat{x} &= \arg\max_{x\in\mathbb{R}} p(x|x_{t-1})p(y_t|x), \\ &= \arg\max_{x\in\mathbb{R}} \ell(x), \\ \Sigma(\hat{x}) &= -\ell''(\hat{x})^{-1} \\ &= \left(\frac{1-\phi^2}{\sigma^2} + \frac{y_t^2}{2\beta^2}\exp(-\hat{x})\right)^{-1}, \end{aligned}$$

we can approximate $q(x_t|x_{t-1}, y_t)$, for each $t$, with

$$\tilde{q}(x_t|x_{t-1}, y_t) = \mathcal{N}(\hat{x}, \Sigma(\hat{x}))).$$

Then, the importance weight function, at time $t$, of the particle $i$, is given by

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-1}^{(i)} \frac{p(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)})p(y_t|\tilde{x}_t^{(i)})}{\tilde{q}(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)}, y_t)}.$$

### 5.3.2 Results

To start with, we performed the bootstrap filtering using the transition density (5.3.2) as the importance function. The overview of results is depicted in Figure 5.3.2. Then, we carried out filtering using the locally optimal importance function, as derived above. Figure 5.3.3 shows the corresponding results.

One can see that both methods deliver reasonable outcomes, i.e. they are able to capture the latent state process fairly accurately. What is remarkable, the latter technique is characterised by much higher filtered state variance and more volatile effective sample size.

To ease the comparison of the performance of the crude bootstrap filter with the one of the filter based on the locally optimal proposal, we gathered the estimates delivered by both filters together (Figure 5.3.4). The upper plot 5.3.5a presents the output of the prior-based algorithm, while the lower one 5.3.4b of the algorithm employing the locally optimal proposal. As expected, the latter one generates much more accurate results, i.e. the filtered out hidden state is more in line with the true one. However, the outcomes of the approximation-based algorithm are highly uncertain, as indicated by the wide confidence intervals. This observation can be confirmed by considering the estimated filtering distributions (Figure 5.3.5), which at any point in time are much broader than the ones from the prior-based setup. This shall come as no surprise, however, as the by construction, the variance in the former approach is constant, while in the latter it is time variant .

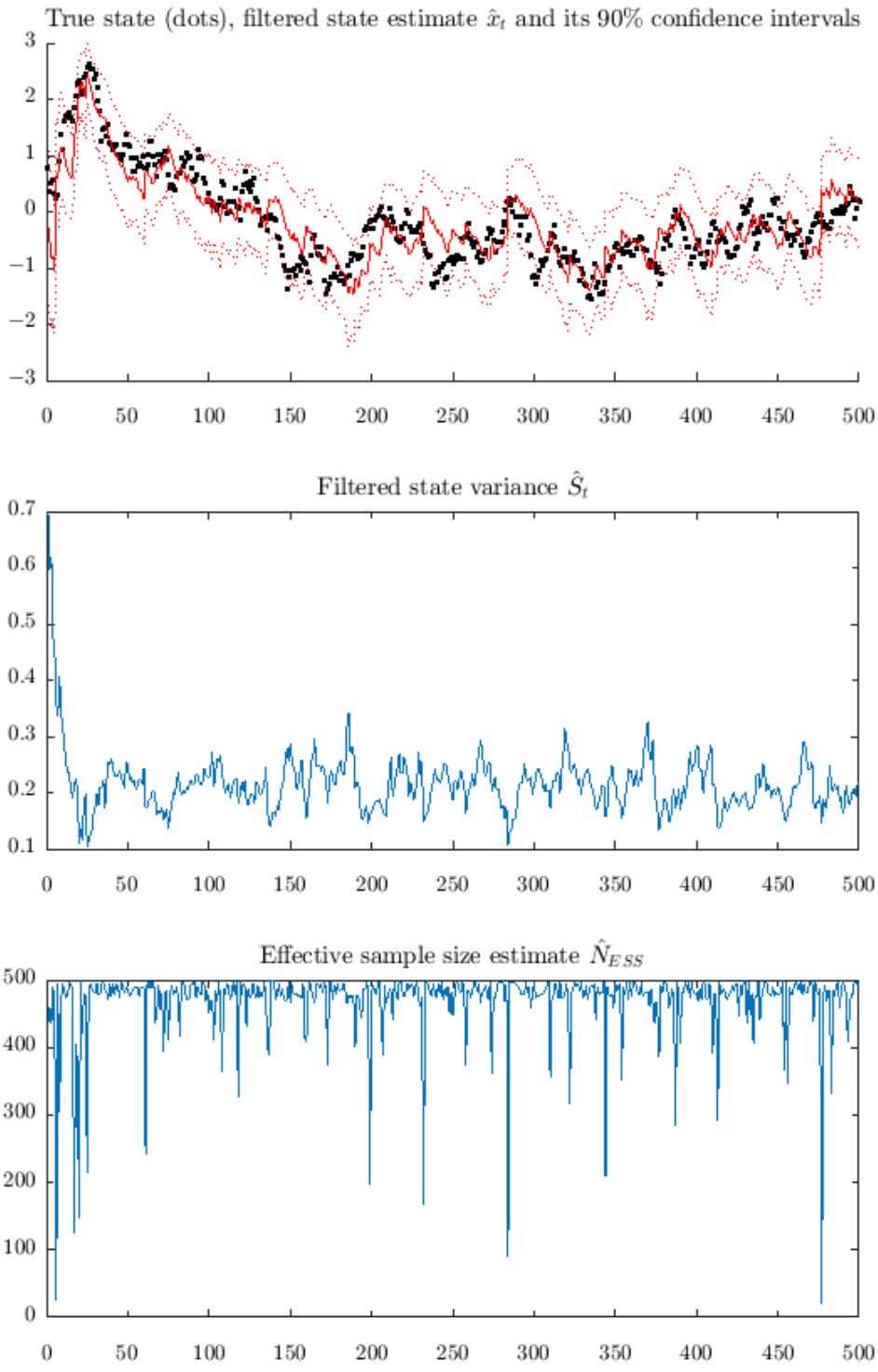



Figure 5.3.2: SMC using prior (bootstrap filter).

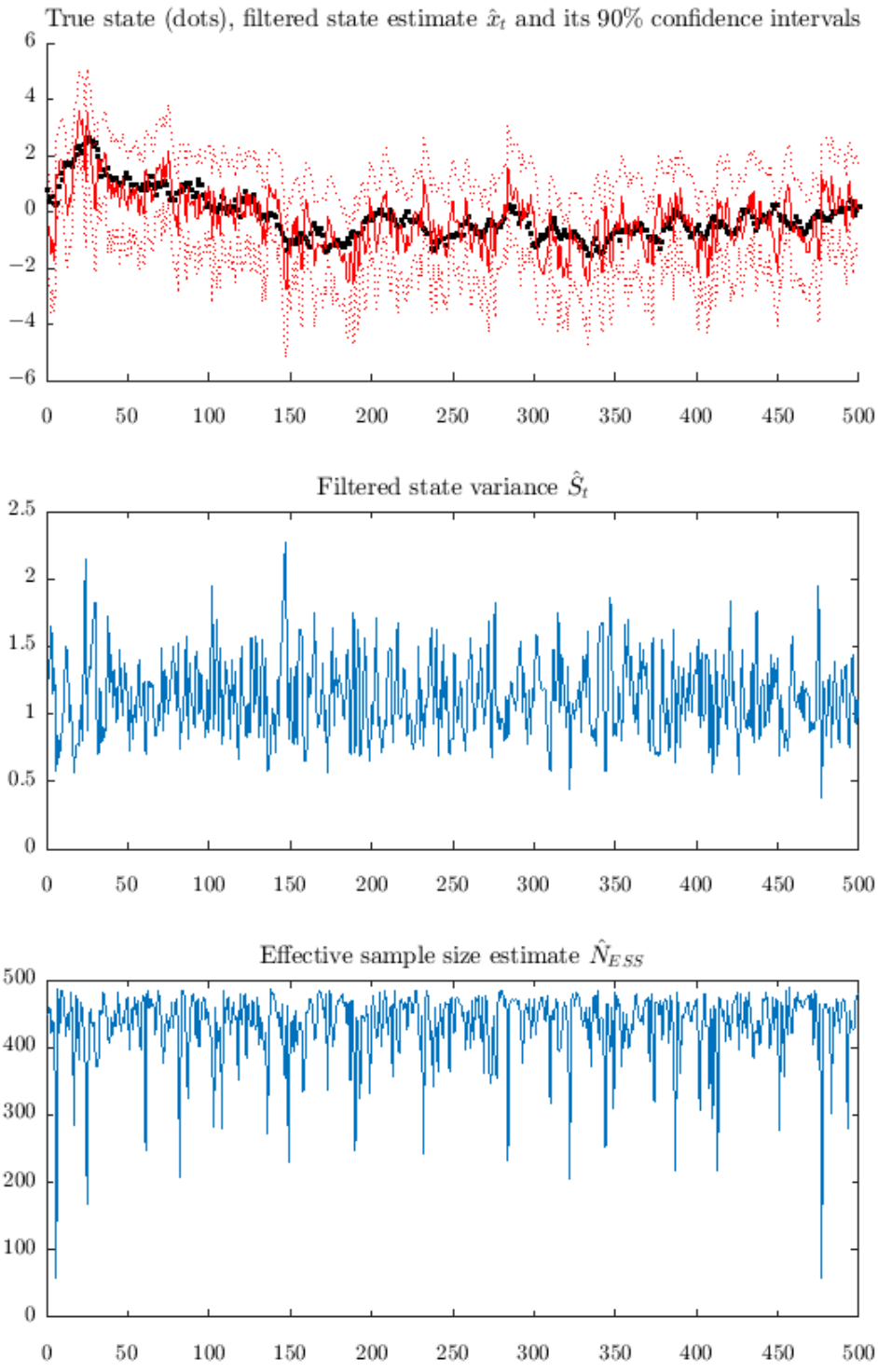
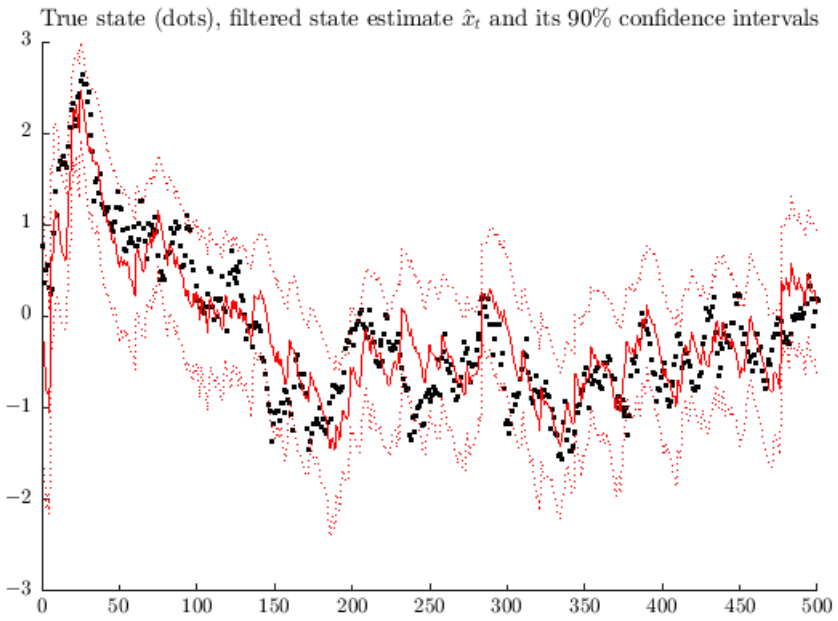
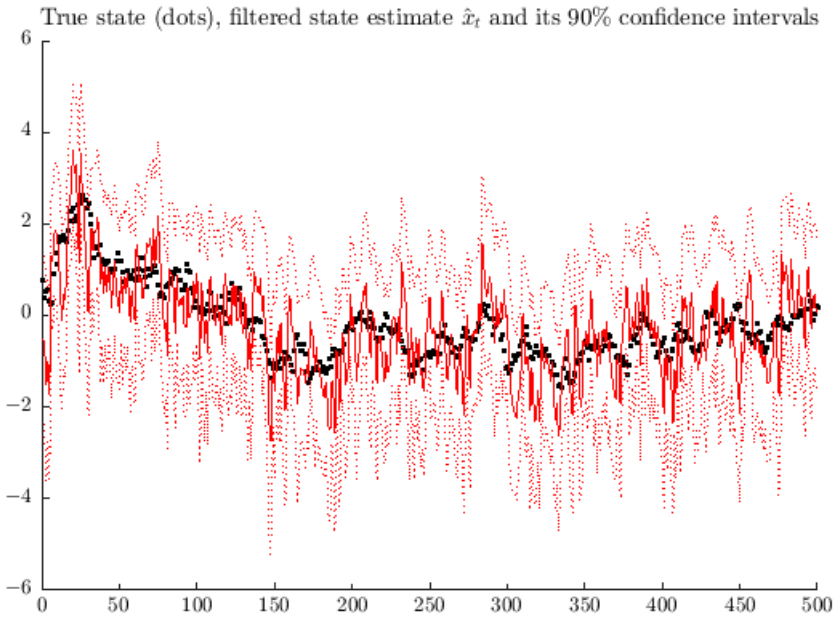



Figure 5.3.3: SMC using locally optimal proposal.

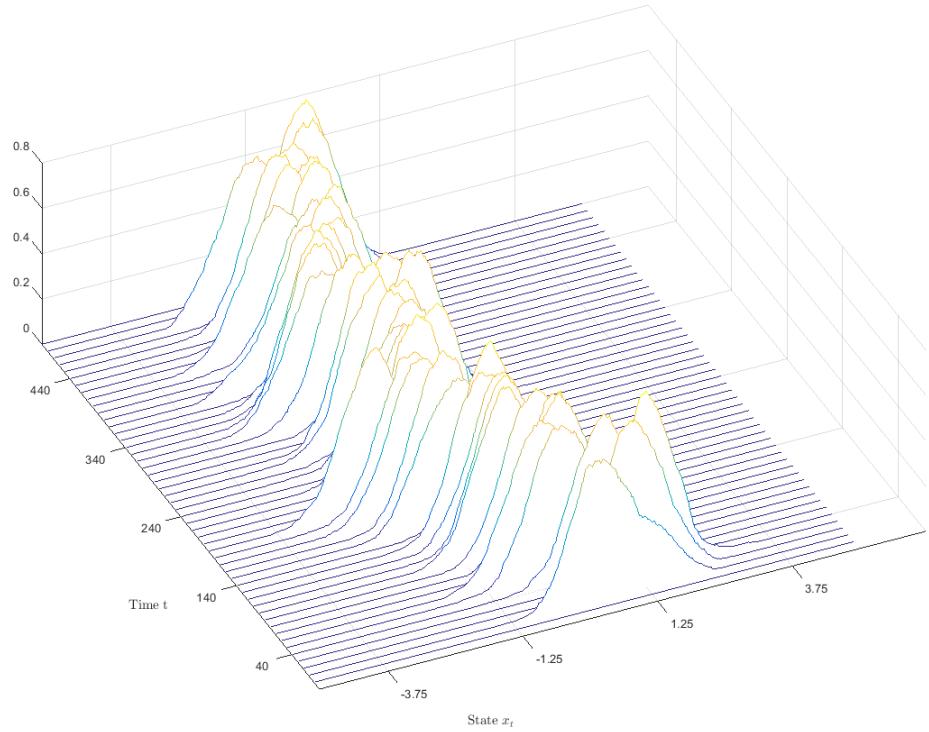



(a) SMC using prior (bootstrap filter).

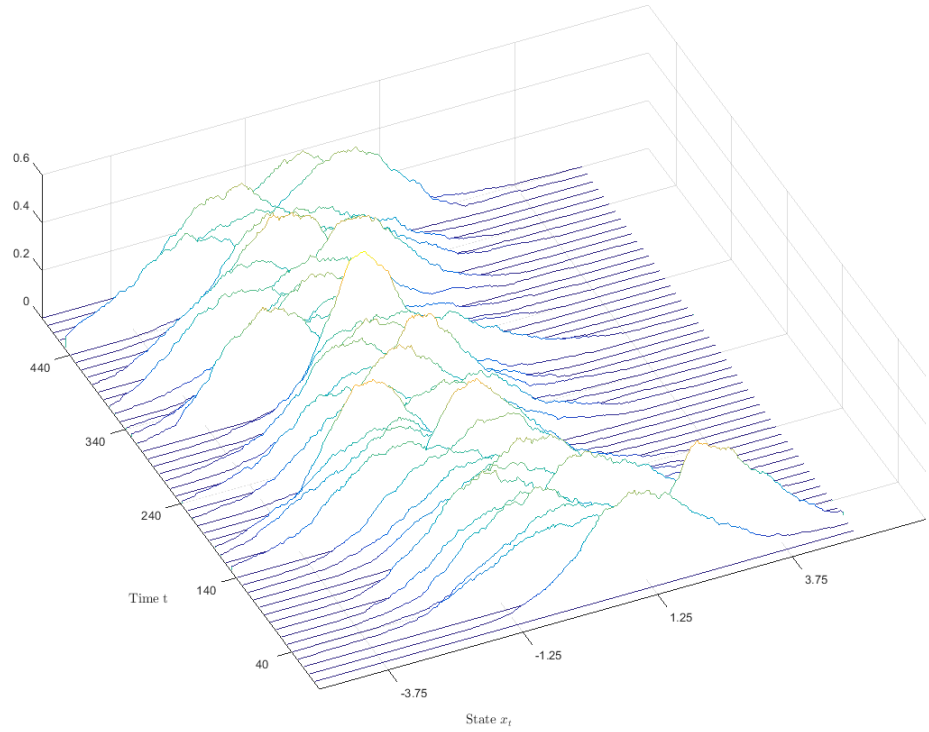



(b) SMC using locally optimal proposal.

Figure 5.3.4: Comparison of the filtering results using different importance functions.




(a) SMC using prior (bootstrap filter).




(b) SMC using locally optimal proposal.

Figure 5.3.5: Estimated filtering distribution using different importance functions.

# Chapter 6

# Conclusions

We have discussed the class of methods used for the on-line inference for dynamic systems based on the Dirac-measure approximations to the distributions characterising the system under investigation, called Sequential Monte Carlo. A special attention has been devoted to the filtering problem, where one is interested in the estimation of the current state of the system given the current system measurements. Because the simulation-based inference is performed using importance sampling, an inevitable weight degeneracy problem arises. We have studied this issue as well as the techniques used to tackle it, such as the introduction of the resampling step it the basic SIS algorithm.

The SMC methods applied to the filtering problem are called particle filters, where the notion of a particle refers to a draw from the possibly time-varying posterior state distribution. Due to the statistical dependence between the particles, the classical convergence results do not apply within the sequential framework. We have discussed the basic convergence results of the measure-valued random variables, stemming from the sequential computations. Despite rather strong assumptions used in the proofs, these results are interesting and important, as they constitute a sound starting point for an analysis based on more plausible premises.

There are two main topics for further research arising from our study. First, we have implemented the SIR algorithms with the resampling performed at each iteration, which clearly is an inefficient way to fight sample impoverishment. Hence, one can consider some more "intelligent" approaches to this problems. The second (and definitely more challenging) task consists in providing a formal treatment of convergence of particle filters under some relaxed, and hence more realistic, assumptions. In particular, it would be interesting to incorporate various resampling mechanisms into the convergence analysis.

# Bibliography


Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002), "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking." *IEEE Transactions on Signal Processing*, 50, 174–188.

Billingsley, P. (1995), *Probability and Measure, 3rd Edition*. Wiley & Sons.

Cappé, O., E. Moulines, and T. Ryden (2006), *Inference in Hidden Markov Models*. Springer Series in Statistics, Springer New York.

Chopin, N. (2004), "Central Limit Theorem for Sequential Monte Carlo Methods and its Application to Bayesian Inference." *Annals of Statistics*, 2385–2411.

Crisan, D. (2001), "Particle Filters – a Theoretical Perspective." In *Sequential Monte Carlo methods in practice*, 17–41, Springer New York.

Crisan, D. (2014), *Convergence of Particle Filters and Relation to DA*. Presented as the ICTS Discussion Meeting, NFDA2014, Bangalore.

Crisan, D. and A. Doucet (2002), "A Survey of Convergence Results on Particle Filtering Methods for Practitioners." *IEEE Transactions on Signal Processing*, 50, 736–746.

Doucet, A., N. de Freitas, and N. Gordon (2001), *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics, Springer.

Doucet, A., S.J. Godsill, and C. Andrieu (2000), "On Sequential Monte Carlo Sampling Methods for Bayesian Filtering." *Statistics and Computing*, 10, 197–208.

Doucet, A. and A. M. Johansen (2009), "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later." *Handbook of Nonlinear Filtering*, 12, 656–704.

Durbin, J. and S. J. Koopman (2012), *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series, OUP Oxford.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993), "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation." *IEE Proceedings F on Radar and Signal Processing*, 140, 107–13.

Hu, X., T. Sch on, and L. Ljung (2008), "A Basic Convergence Result for Particle Filtering." *IEEE Transactions on Signal Processing*, 56, 1337–1348.

Kitagawa, G. (1996), "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models." *Journal of Computational and Graphical Statistics*, 5, 1–25.

Kong, A., J. S. Liu, and W. H. Wong (1994), "Sequential Imputations and Bayesian Missing Data Problems." *Journal of the American Statistical Association*, 89, 278–288.

Li, T., S. Sun, T. P. Sattar, and J. M. Corchado (2014), "Fight Sample Degeneracy and Impoverishment in Particle Filters: A Rreview of Intelligent Approaches." *Expert Systems with Applications*, 41, 3944–3954.

Liu, J. S. (1996), "Metropolized Independent Sampling with Comparison to Rejection Sampling and Importance Sampling." *Statistics and Computing*, 6, 113–119.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*. Springer.

Liu, J. S. and R. Chen (1998), "Sequential Monte Carlo Methods for Dynamical Systems." *Journal of the American Statistical Association*, 93, 1032–1044.

Shao, J. (2003), *Mathematical Statistics*. Springer Texts in Statistics, Springer.

# Appendix A

# Notation

$\mathcal{B}(\mathbb{R}^d)$ – the Borel $\sigma$-algebra of $\mathbb{R}^d$.

$B(\mathbb{R}^d)$ – the set of bounded, $\mathcal{B}(\mathbb{R}^d)$-measurable functions on $\mathbb{R}^d$.

$\mathcal{C}_b(\mathbb{R}^d)$ – the set of bounded continuous functions on $\mathbb{R}^d$.

$\mathcal{M}(\mathbb{R}^d)$ – the set of finite measures over $\mathcal{B}(\mathbb{R}^d)$.

$\mathcal{P}(\mathbb{R}^d)$ – the set of probability measures over $\mathcal{B}(\mathbb{R}^d)$.

$\lambda(\mathbb{R}^d)$ – the Lebesgue measure on $\mathbb{R}^d$.

$\mu f$ – the integral of $f$ with respect to $\mu$, where $\mu \in \mathcal{M}_F(\mathbb{R}^d)$ and $f \in B(\mathbb{R}^d)$.

$$\mu f := \int_{\mathbb{R}^d} f(x)\mu(dx).$$

$t \in \mathbb{N}$ – time index.

$Y = \{Y_t\}_{t\in\mathbb{N}}$ – a $\mathcal{Y} \subset \mathbb{R}^{n_y}$ stochastic process (the observation process).

$X = \{X_t\}_{t\in\mathbb{N}}$ – an $\mathcal{X} \subset \mathbb{R}^{n_x}$ stochastic process (the signal process).

$y_t \in \mathcal{Y}$ – observation vector at $t$.

$y_{0:t} = \{y_0^T, \ldots, y_t^T\}$ – collection of observation vectors up to time $t$ (the realisation of process $Y$).

$x_t \in \mathcal{X}$ – the true unobserved signal vector (hidden state) (the realisation of process $X$).

$x_{0:t} = \{x_0^T, \ldots, x_t^T\}$ – the collection of state vectors up to time $t$.

$\mathbb{P}$ – probability measure.

$\mathbb{E}_g$ – expectation with respect to density $g$.

$\delta_a$ – the Dirac measure concentrated at $a \in \mathbb{R}^d$, i.e. $\delta_a(x) \equiv \mathbb{I}_{x=a}$.

$\mathbf{1}$ – the constant function 1.

$\mathcal{N}(\mu, \sigma^2)$ – the normal distribution with mean $\mu$ and variance $]sigma^2$ (potentially multivariate).

$\Phi(x; \mu, \sigma^2)$, $\phi(x; \mu, \sigma^2)$ – the cdf and pdf, respectively, of the normal distribution with mean $\mu$ and variance $\sigma^2$, evaluated at $x$; refer to the standard normal distribution when $\mu$ and $\sigma^2$ are skipped.

$\overset{i.i.d.}{\sim}$ – independently and identically distributed.

# Appendix B

# Properties of IS estimator

Since the normalised IS estimator (3.2.8) is a ratio of two estimators, neither its bias nor variance admits a simple formula. Nevertheless, it is possible to obtain their asymptotic expressions by employing the multivariate delta method, which we recall below. Since the delta method basically amounts to finding approximations based on Taylor expansions of functions of random variables, before presenting both asymptotic results, we first discuss two useful lemmas. In their proofs, instead of referring to the delta method, we explicitly compute the required approximations.

## B.1 Delta method

**Theorem B.1.1** (Multivariate Delta method)**.** *Let $Z_n = (Z_{n1}, \dots, Z_{nk})$ be a sequence of random variables such that*

$$\sqrt{n}(Z_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

*Let $g : \mathbb{R}^k \to \mathbb{R}$ be a differentiable and denote*

$$\nabla g = \left( \frac{\partial g}{\partial z_1}, \dots, \frac{\partial g}{\partial z_k} \right)^T$$

*with $\nabla g(\mu)$ be $\nabla g$ evaluated at $z = \mu$. Assume $\nabla g(\mu)_j$, $\forall j$. Then*

$$\sqrt{n}\left(g(Z_n) - g(\mu)\right) \xrightarrow{d} \mathcal{N}\left(0, \nabla^T g(\mu) \Sigma \nabla g(\mu)\right).$$

*Proof.* The proof of this standard result in statistics can be found in many books, e.g. in Shao (2003). □

**Lemma B.1.1.** *Let $X$ and $Y$ be two random variables, with means $\mu_X$, $\mu_Y$ and variances $\sigma_X^2$, $\sigma_Y^2$, respectively, as well as with the correlation $\rho$. For the ratio random variable $\frac{Y}{X}$ it holds that*

1. *for the expectation*

$$\mathbb{E}\left[\frac{Y}{X}\right] \approx \frac{\mu_Y}{\mu_X} + \frac{\mu_Y}{\mu_X^3}\sigma_X^2 - \mu_X^2 \rho \sigma_X \sigma_Y;$$

2. *for the variance*

$$Var\left[\frac{Y}{X}\right] \approx \left(\frac{\mu_Y^2}{\mu_X^4}\right)\sigma_X^2 + \left(\frac{1}{\mu_X^2}\right)\sigma_Y^2 - 2\left(\frac{\mu_Y}{\mu_X^3}\right)\rho\sigma_X\sigma_Y.$$

*Proof.* Consider a function $f(x,y) = \frac{y}{x}$ and take its second order Taylor expansion around $(x_0, y_0)$ to obtain

$$\begin{aligned}
f(x,y) &= f(x_0,y_0) + (x-x_0)f_x(x_0,y_0) + (y-y_0)f_y(x_0,y_0) \\
&\quad + \frac{1}{2}\left[(x-x_0)^2 f_{xx}(x_0,y_0) + (y-y_0)^2 f_{yy}(x_0,y_0) + 2(x-x_0)(y-y_0)f_{xy}(x_0,y_0)\right] + o\left(||(x_0,y_0)||^2\right) \\
&\approx \frac{y_0}{x_0} + (x-x_0)\frac{-y_0}{x_0^2} + (y-y_0)\frac{1}{x_0} + (x-x_0)^2\frac{y_0}{x_0^3} + (x-x_0)(y-y_0)\frac{-1}{x_0^2} \\
&= \frac{y_0}{x_0} - \frac{y_0}{x_0^2}x + \frac{1}{x_0}y + \frac{y_0}{x_0^3}(x-x_0)^2 - \frac{1}{x_0^2}(x-x_0)(y-y_0).
\end{aligned}$$

Hence, for a random variable $Z$ defined as $Z = \frac{Y}{X}$ we obtain by means of Taylor expansion about $(\mu_X, \mu_Y)$

$$Z = f(X,Y) \approx \frac{\mu_Y}{\mu_X} - \frac{\mu_Y}{\mu_X^2}X + \frac{1}{\mu_X}Y + \frac{\mu_Y}{\mu_X^3}(X-\mu_X)^2 - \frac{1}{\mu_X^2}(X-\mu_X)(Y-\mu_Y). \tag{B.1.1}$$

Then,

1. taking the expectation on the both sides of (B.1.1) yields

$$\begin{aligned}
\mathbb{E}Z = \mathbb{E}\left[f(X,Y)\right] &\approx \frac{\mu_Y}{\mu_X} - \frac{\mu_Y}{\mu_X^2}\mu_X + \frac{1}{\mu_X}\mu_Y + \frac{\mu_Y}{\mu_X^3}\sigma_X^2 - \frac{1}{\mu_X^2}\mathrm{Cov}\left[X,Y\right] \\
&= \frac{\mu_Y}{\mu_X} + \frac{\mu_Y}{\mu_X^3}\sigma_X^2 - \frac{1}{\mu_X^2}\rho\sigma_X\sigma_Y;
\end{aligned}$$

2. ignoring the second order terms in (B.1.1) and taking the variance on the both sides gives

$$\begin{aligned}
\mathrm{Var}Z = \mathrm{Var}\left[f(x,y)\right] &\approx \frac{\mu_Y^2}{\mu_X^4}\sigma_X^2 + \frac{1}{\mu_X^2}\sigma_Y^2 - 2\frac{\mu_Y}{\mu_X^2}\frac{1}{\mu_X}\mathrm{Cov}\left[X,Y\right] \\
&= \left(\frac{\mu_Y^2}{\mu_X^4}\right)\sigma_X^2 + \left(\frac{1}{\mu_X^2}\right)\sigma_Y^2 - 2\left(\frac{\mu_Y}{\mu_X^3}\right)\rho\sigma_X\sigma_Y,
\end{aligned}$$

which completes the proof. □

**Lemma B.1.2.** *Let $X$ and $Y$ be two random variables, with means $\mu_X$, $\mu_Y$ and variances $\sigma_X^2$, $\sigma_Y^2$, respectively, as well as with the correlation $\rho$. For the product random variable $XY^2$ it holds that*

$$\mathbb{E}\left[XY^2\right] \approx \mu_X\mu_Y^2 + \mu_X\sigma_Y^2 + 2\mu_Y\rho\sigma_X\sigma_Y.$$

*Proof.* Consider a function $f(x,y) = xy^2$ and take its second order Taylor expansion around $(x_0, y_0)$ to obtain

$$\begin{aligned}
f(x,y) &= f(x_0,y_0) + (x-x_0)f_x(x_0,y_0) + (y-y_0)f_y(x_0,y_0) \\
&\quad + \frac{1}{2}\left[(x-x_0)^2 f_{xx}(x_0,y_0) + (y-y_0)^2 f_{yy}(x_0,y_0) + 2(x-x_0)(y-y_0)f_{xy}(x_0,y_0)\right] + o\left(||(x_0,y_0)||^2\right) \\
&\approx x_0y_0^2 + (x-x_0)y_0^2 + 2(y-y_0)x_0y_0 + \frac{1}{2}\left[2(y-y_0)^2x_0 + 4(x-x_0)(y-y_0)y_0\right] \\
&= x_0y_0^2 - x_0y_0^2 - 2x_0y_0^2 + y_0^2x + 2x_0y_0y + x_0(y-y_0)^2 + 2y_0(x-x_0)(y-y_0) \\
&= -2x_0y_0^2 + y_0^2x + 2x_0y_0y + x_0(y-y_0)^2 + 2y_0(x-x_0)(y-y_0).
\end{aligned}$$

Hence, for a random variable $Z$ defined as $Z = XY^2$ we obtain by means of the second order Taylor expansion about $(\mu_X, \mu_Y)$

$$Z = f(X,Y) \approx -2\mu_X\mu_Y^2 + \mu_Y^2X + 2\mu_X\mu_YY + \mu_X(Y-\mu_Y)^2 + 2\mu_Y(X-\mu_X)(Y-\mu_Y).$$

Then, taking the expectation on the both sides of the above formula yields

$$\begin{aligned}\mathbb{E}Z = \mathbb{E}\left[f(x,y)\right] &\approx \mu_X\mu_Y^2 + \mu_X\sigma_Y^2 + 2\mu_Y \mathrm{Cov}\left[X,Y\right] \\ &= \mu_X\mu_Y^2 + \mu_X\sigma_Y^2 + 2\mu_Y\rho\sigma_X\sigma_Y,\end{aligned}$$

which completes the proof. □

## B.2 Asymptotic properties of IS estimator

The following analysis is based on Liu (2001) and Liu (1996). Consider two independent sequences of i.i.d. random variables $X_t^{(1)},\ldots,X_t^{(N)} \overset{iid}{\sim} q(x_t|x_{0:t-1},y_{0:t})$ and $Y_t^{(1)},\ldots,Y_t^{(N)} \overset{iid}{\sim} p(x_t|x_{0:t-1},y_{0:t})$, both of length $N$. Recall that the MC estimator and the IS estimator are respectively given by

$$\begin{aligned}\hat{f}^{MC} &= N^{-1}\sum_{i=1}^{N} f(Y_t^{(j)}), \\ \hat{f}^{IS} &= \frac{N^{-1}\sum_{i=1}^{N}\tilde{w}_t^{(i)}(X_t^{(j)})f(X_t^{(j)})}{N^{-1}\sum_{i=1}^{N}\tilde{w}_t^{(i)}(X_t^{(j)})},\end{aligned}$$

where we additionally indicate the dependence of the importance weight $\tilde{w}_t^{(i)}$ on $X_t^{(i)}$. Notice that we can also write

$$\hat{f}^{IS} = \frac{N^{-1}\sum_{i=1}^{N}\tilde{W}_t^{(i)}(X_t^{(j)})f(X_t^{(j)})}{N^{-1}\sum_{i=1}^{N}\tilde{W}_t^{(i)}(X_t^{(j)})},$$

since $\tilde{w}_t^{(i)} = p(y_{0:t})\tilde{W}_t^{(i)}$ and the term $p(y_{0:t})$ cancels out in the ratio. This modification will allow us to use that fact the

$$\mathbb{E}_q\left[\tilde{W}_t^{(i)}\right] = \int \frac{p(x_{0:t}|y_{0:t})}{q(x_{0:t}|y_{0:t})} q(x_{0:t}|y_{0:t})dx_{0:t} = 1.$$

Next, let us introduce the following notation $F_i = f(X_t^{(i)})$, $Z_i = f(X_t^{(j)})\tilde{W}_t^{(i)}(X_t^{(j)})$ and $U_i = \tilde{W}_t^{(i)}(X_t^{(j)})$, so that

$$\begin{aligned}\hat{f}^{MC} &= N^{-1}\sum_{i=1}^{N} F_i =: \bar{F}_N, \\ \hat{f}^{IS} &= \frac{N^{-1}\sum_{i=1}^{N} Z_i}{N^{-1}\sum_{i=1}^{N} U_i} =: \frac{\bar{Z}_N}{\bar{U}_N}.\end{aligned}$$

By the central limit theorem

$$\sqrt{N}\left(\begin{pmatrix}\bar{Z}\\ \bar{U}\end{pmatrix} - \begin{pmatrix}\mu_Z\\ \mu_U\end{pmatrix}\right) \underset{N\to\infty}{\overset{d}{\to}} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_Z^2 & \sigma_{ZU}\\ \sigma_{ZU} & \sigma_U^2\end{pmatrix}\right),$$

where

$$\begin{aligned}\mu_Z &= \mathbb{E}_q Z, & \mu_U &= \mathbb{E}_q U = 1, \\ \sigma_Z^2 &= \mathrm{Var}_q Z, & \sigma_U^2 &= \mathrm{Var}_q U, \\ \sigma_{ZU} &= \mathrm{Cov}_q\left[U,Z\right].\end{aligned}$$

**Proposition B.2.1** (CLT for IS estimator)**.** *Assume* $Var_q Z < \infty$ *and* $Var_q U < \infty$*. Then*

$$\sqrt{N}\left(\hat{f}^{IS} - \bar{f}\right) \underset{N\to\infty}{\overset{d}{\to}} \mathcal{N}\left(0, \sigma_{IS}^2\right)$$

*where*

$$\sigma_{IS}^2 = Var_q[Z] + \mu_Z^2\, Var_q[U] - 2\mu_Z\, Cov_q[U, Z]$$

*Proof.* We can apply the delta method to $(\bar{Z}_N, \bar{U}_N)$ with $g(z_1, z_2) = \frac{z_1}{z_2}$. Then, by Lemma B.1.1.2 we immediately obtain that

$$\sigma_{IS}^2 = \mu_Z^2 \mathrm{Var}_q[U] + \mathrm{Var}_q[Z] - 2\mu_Z \mathrm{Cov}_q[U, Z].$$

This means that

$$\mathrm{Var}_q\left[\hat{f}^{IS}\right] \approx \frac{1}{N}\Big(\mu_Z^2 \mathrm{Var}_q[U] + \mathrm{Var}_q[Z] - 2\mu_Z \mathrm{Cov}_q[U, Z]\Big).$$

□

**Proposition B.2.2** (Asymptotic bias of IS estimator)**.** *Assume* $Var_q Z < \infty$ *and* $Var_q U < \infty$*. Then*

$$\lim_{N\to\infty} N\mathbb{E}_q\left[\hat{f}^{IS} - \bar{f}\right] = \bar{f}\, Var_q U - Cov_q\left[U, Z\right],$$

*which means that the expectation of the IS estimator can be approximated as*

$$\mathbb{E}_q\left[\hat{f}^{IS}\right] = \bar{f} + \frac{1}{N}\bar{f}\, Var_q U - \frac{1}{N} Cov_q\left[U, Z\right].$$

*Hence, the IS estimator is asymptotically biased, with the bias being of order* $1/n$*.*

*Proof.* Similarly as in Proposition B.2.1, by the delta method and Lemma B.1.1.1, we immediately obtain the require result. □

## B.3 Efficiency of IS estimator

**Proposition B.3.1** (Efficiency of IS estimator)**.** *The ratio* (3.4.3) *can be approximated as*

$$\frac{Var_q\left[\hat{f}^{MC}\right]}{Var_p\left[\hat{f}^{IS}\right]} \approx 1 + Var_q\left[\tilde{W}_t^j\right] \tag{B.3.1}$$

*which does not depend on the choice of the integrated function* $f$*.*

*Proof.* For the MC estimator we have

$$\begin{aligned} \mathbb{E}_p\left[\hat{f}^{MC}\right] &= \mathbb{E}_p\left[F\right] = \bar{f}, \\ \mathrm{Var}_p\left[\hat{f}^{MC}\right] &= \frac{\mathrm{Var}_p F}{N}, \end{aligned} \tag{B.3.2}$$

while by Proposition B.2.1

$$\mathrm{Var}_q\left[\hat{f}^{IS}\right] \approx \frac{1}{N}\Big(\mu_Z^2 \mathrm{Var}_q[U] + \mathrm{Var}_q[Z] - 2\mu_Z \mathrm{Cov}_q[U, Z]\Big).$$

Since $Z_i = F_i U_i$, $\mu_Z = \mathbb{E}_q[Z] = \mathbb{E}_p[F]$ and $\mathbb{E}_q[U] = 1$, the last two terms in the above formula can be expressed as

$$\begin{aligned}
\mathrm{Var}_q[Z] - 2\mu_Z \mathrm{Cov}_q[U, Z] &= \mathbb{E}_q\left[Z^2\right] - \left(\mathbb{E}_q[Z]\right)^2 - 2\mu_Z\Big(\mathbb{E}_q[UZ] - \mathbb{E}_q[U]\,\mathbb{E}_q[Z]\Big) \\
&= \mathbb{E}_q\left[F^2U^2\right] - \left(\mathbb{E}_q[FU]\right)^2 - 2\mu_Z\Big(\mathbb{E}_q[UFU] - \mathbb{E}_q[FU]\Big) \\
&= \mathbb{E}_p\left[F^2U\right] - \left(\mathbb{E}_p[F]\right)^2 - 2\mu_Z\Big(\mathbb{E}_p[FU] - \mathbb{E}_p[F]\Big).
\end{aligned}$$

Since

$$\begin{aligned}
\mathrm{Cov}_p[F, U] &= \mathbb{E}_p[FU] - \mathbb{E}_p[F]\,\mathbb{E}_p[U] \\
&= \mathbb{E}_p[FU] - \mu_Z \mathbb{E}_p[U],
\end{aligned}$$

we further have

$$\begin{aligned}
\mathrm{Var}_q[Z] - 2\mu_Z \mathrm{Cov}_q[U, Z] &= \mathbb{E}_p\left[F^2U\right] - \left(\mu_Z\right)^2 - 2\mu_Z\Big(\mathrm{Cov}_p[F, U] + \mu_Z \mathbb{E}_p[U] - \mu_Z\Big) \\
&= \mathbb{E}_p\left[F^2U\right] + \mu_Z^2 - 2\mu_Z \mathrm{Cov}_p[F, U] - 2\mu_Z^2 \mathbb{E}_p[U]
\end{aligned}$$

Next, we can apply Lemma B.1.2 to the first term on the RHS in the above formula to obtain the following approximation

$$\begin{aligned}
\mathbb{E}_p\left[UF^2\right] &\approx \mathbb{E}_p[U]\,\mathbb{E}_p[F]^2 + \mathbb{E}_p[U]\,\mathrm{Var}_p[F] + 2\mathbb{E}_p[F] + \mathrm{Cov}_p[U, F] \\
&= \mathbb{E}_p[U]\,\mu_Z^2 + \mathbb{E}_p[U]\,\mathrm{Var}_p[F] + 2\mu_Z \mathrm{Cov}_p[U, F],
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathrm{Var}_q[Z] - 2\mu_Z \mathrm{Cov}_q[U, Z] \approx& \mathbb{E}_p[U]\,\mu_Z^2 + \mathbb{E}_p[U]\,\mathrm{Var}_p[F] + 2\mu_Z \mathrm{Cov}_p[U, F] \\
&+ \mu_Z^2 - 2\mu_Z \mathrm{Cov}_p[F, U] - 2\mu_Z^2 \mathbb{E}_p[U] \\
=& \mathbb{E}_p[U]\left(\mathrm{Var}_p[F] - \mu_Z^2\right) + \mu_Z^2.
\end{aligned}$$

Hence,

$$\mathrm{Var}_q\left[\hat{f}^{IS}\right] \approx \frac{1}{N}\Big(\mu_Z^2 \mathrm{Var}_q[U] + \mathbb{E}_p[U]\left(\mathrm{Var}_p[F] - \mu_Z^2\right) + \mu_Z^2\Big). \tag{B.3.3}$$

Then, we can simplify (B.3.3) to

$$\begin{aligned}
\mathrm{Var}_q\left[\hat{f}^{IS}\right] &\approx \frac{1}{N}\Big(\mu_Z^2 \mathrm{Var}_q[U] + \mathbb{E}_p[U]\left(\mathrm{Var}_p[F] - \mu_Z^2\right) + \mu_Z^2\Big) \\
&= \frac{1}{N}\Big(\left(\mathbb{E}_p[F]\right)^2\left(\mathbb{E}_q\left[U^2\right] - \left(\mathbb{E}_q[U]\right)^2\right) + \mathbb{E}_p[U]\left(\mathrm{Var}_p[F] - \left(\mathbb{E}_p[F]\right)^2\right) + \left(\mathbb{E}_p[F]\right)^2\Big) \\
&= \frac{1}{N}\Big(\left(\mathbb{E}_p[F]\right)^2\left(\mathbb{E}_p[U] - 1\right) + \mathbb{E}_p[U]\,\mathrm{Var}_p[F] - \mathbb{E}_p[U]\left(\mathbb{E}_p[F]\right)^2 + \left(\mathbb{E}_p[F]\right)^2\Big) \\
&= \frac{1}{N}\Big(\left(\mathbb{E}_p[F]\right)^2\mathbb{E}_p[U] - \left(\mathbb{E}_p[F]\right)^2 + \mathbb{E}_p[U]\,\mathrm{Var}_p[F] - \mathbb{E}_p[U]\left(\mathbb{E}_p[F]\right)^2 + \left(\mathbb{E}_p[F]\right)^2\Big) \\
&= \frac{1}{N}\mathbb{E}_p[U]\,\mathrm{Var}_p[F],
\end{aligned} \tag{B.3.4}$$

where in subsequent steps we used the following facts

$$\begin{aligned}\mathrm{Var}_q\left[U\right] &= \mathbb{E}_q\left[U^2\right] - \left(\mathbb{E}_q\left[U\right]\right)^2, \\ \mathbb{E}_p\left[U\right] &= \mathbb{E}_q\left[U^2\right], \\ \mu_Z &= \mathbb{E}_p\left[F\right].\end{aligned}$$

Next, we can again express $\mathbb{E}_p\left[U\right]$ as

$$\begin{aligned}\mathbb{E}_p\left[U\right] &= \mathbb{E}_q\left[U^2\right] \\ &= \mathrm{Var}_q\left[U\right] + \left(\mathbb{E}_q\left[U\right]\right)^2 \\ &= \mathrm{Var}_q\left[U\right] + 1,\end{aligned}$$

and plug this result to (B.3.4) to finally arrive at the following approximation

$$\mathrm{Var}_q\left[\hat{f}^{IS}\right] \approx \frac{1}{N}\mathrm{Var}_p\left[F\right]\Big(1 + \mathrm{Var}_q[U]\Big).$$

which together with (B.3.2) gives

$$\begin{aligned}\frac{\mathrm{Var}_q\left[\hat{f}^{IS}\right]}{\mathrm{Var}_p\left[\hat{f}^{MC}\right]} &\approx \frac{\frac{1}{N}\mathrm{Var}_p\left[F\right]\Big(1 + \mathrm{Var}_q[U]\Big)}{\frac{1}{N}\mathrm{Var}_p\left[F\right]} \\ &= 1 + \mathrm{Var}_q[U] \\ &= 1 + \mathrm{Var}_q[\tilde{W}_t^j],\end{aligned}$$

which is the desired result. □

# Appendix C

# Code listings

## C.1 Basic linear Gaussian problem

```
% Particle filter example, linear Gaussian model,
% basic local level model

clear all
s = RandStream('mt19937ar','Seed',0);
RandStream.setGlobalStream(s);
clear all;

resampl = 0;     % 0 = no resampling;
                 % 1 = multinomial resampling;
                 % 2 = stratified sampling;

% read the historical data
name = 'Nile.csv';
y = load(name);
T = length(y);
t = 1871:1:1970;

N = 10000;
alp = 0.1;
z = norminv(1-alp/2);

%% particle filter initialisation
sigma2_eps = 15099;
sigma2_eta = 1469.1;

a = zeros(T,1);      % filtered state
P = zeros(T,1);      % filtered state variance
P(1,1) = 10^7;

alpha = zeros(T,N);
alpha(1,:) = y(1,1) + sqrt(P(1,1))*randn(1,N);
alpha_sim = zeros(T,N);
alpha_sim(1,:) = alpha(1,:);
```

```
w = ones(T,N)/N;
ess = zeros(T,1);
ess(1,1) = 1/(sum(w(1,:).^2));


%% particle filter
for ii = 2:T
    alpha_sim(ii,:) = alpha(ii-1,:) + sqrt(sigma2_eta)*randn(1,N);
    if resampl == 0
        w(ii,:) = w(ii-1,:).*exp(-0.5*(log(2*pi)+log(sigma2_eps)
                        +(y(ii,1)-alpha_sim(ii,:)).^2/sigma2_eps));
    else
        w(ii,:) = exp(-0.5*(log(2*pi)+log(sigma2_eps)
                         +(y(ii,1)-alpha_sim(ii,:)).^2/sigma2_eps));
    end
    w(ii,:) = w(ii,:)/sum(w(ii,:));
    ess(ii,1) = 1/(sum(w(ii,:).^2));
    a(ii,1) = sum(w(ii,:).*alpha_sim(ii,:));
    P(ii,1) = sum(w(ii,:).*(alpha_sim(ii,:).^2)) - a(ii,:)^2;

    if resampl == 0          % no resampling
        alpha(ii,:) = alpha_sim(ii,:);
    elseif resampl == 1      % multinomial resampling
        alpha(ii,:) = randsample(alpha_sim(ii,:),N,true,w(ii,:))';
        w(ii,:) = ones(1,N)/N;
    else                     % systematic resampling (Kitagawa 1996)
                             % a.k.a. stratified sampling
        % 1. initialize the CDF
        c = zeros(1,N);
        c(1,2:N) = cumsum(w(ii,2:N));
        % 2. move along the CDF
        u = rand/N + 0:(1/N):1;
        % 3. find a parent
        M = (repmat(u',1,N)>repmat(c,N,1));
        iP = sum(M,2);
        % 4. assign sample
        aux = alpha_sim(ii,:);
        alpha(ii,:) = aux(iP);
        % 5. reset weights
        w(ii,:) = ones(1,N)/N;
    end
end
```

## C.2 Nonlinear Gaussian problem

```
% Particle filter example, nonlinear model,
% adapted from Gordon, Salmond, and Smith (1993).
% Bootrstap filter

clear all
s = RandStream('mt19937ar','Seed',0);
RandStream.setGlobalStream(s);
```

```
filter = 1;      % 0 = bootstrap;
                 % 1 = locally optimal;
resampl = 2;     % 0 = no resampling;
                 % 1 = always resapling;
                 % 2 = resampling only below the threshold k*N

k = 0.5; % resampling threshold

alp = 0.1;
z = norminv(1-alp/2);

T = 50+1; % series length
N = 1000; % number of particles
t = 0:(T-1);

sigma2_v = 10; % state noise
sigma2_w = 1;  % observation noise
sigma2_0 = 2;
f = @(xx,ii) 0.5.*xx + 25*xx./(1+xx.^2) + 8*cos(1.2*(ii-1));
g = @(xx) 0.05*xx.^2;

% normal log densities to evaluate the importance weight
p_norm = @(xx,mm,SS) -0.5*(log(2*pi) + log(SS) + ((xx-mm).^2)./SS);

if filter == 1
    % parameters of the linearised importance dunction
    Sigma2_q = @(xx,ii) 1./(1/sigma2_v + (f(xx,ii).^2)./(100*sigma2_w));
    mu_q = @(xx,yy,ii) Sigma2_q(xx,ii).*f(xx,ii)
           .*(1/sigma2_v + 0.1*(yy + 0.05*(f(xx,ii)).^2)/sigma2_w);
end

%% generate state and observation (severely nonlinear)
x = zeros(T,1); % true unobserved state
x(1,1) = 0.1;   % initial state
y = zeros(T,1); % observation

for ii = 2:T
    x(ii,1) = f(x(ii-1,1),ii) + sqrt(sigma2_v)*randn;
    y(ii,1) = g(x(ii,1)) + sqrt(sigma2_w)*randn;
end

%% Particle filter initialisation
x_sim = zeros(T,N);   % particles
x_sim(1,:) = sqrt(sigma2_0)*randn(1,N);
x_aux = zeros(T,N);
x_aux(1,:) = x_sim(1,:);

x_est = zeros(T,1);   % filtered state estimate
S_est = zeros(T,1);   % filtered state variance
S_est(1,1) = sigma2_0;

w = ones(T,N)/N;
ess = zeros(T,1);
```

```
ess(1,1) = 1/(sum(w(1,:).^2));
bin = 80;
bg = 2; % bin grid
bin_ax = -bin/bg:1/bg:bin/bg;
epdf = zeros(T,2*bin+1);

%% particle filer
for ii = 2:T
    if filter == 0 % draw from prior, i.e. transition p(x_t|x_(t-1)^(i))
        x_sim(ii,:) = f(x_aux(ii-1,:),ii) + sqrt(sigma2_v)*randn(1,N);
        w(ii,:) = p_norm(y(ii,1),g(x_sim(ii,:)),sigma2_w);
    else
        % parameters of the linearised importance function
        Sigma2 = Sigma2_q(x_aux(ii-1,:),ii);
        mu = mu_q(x_aux(ii-1,:),y(ii),ii);
        % draw from linear approximation, q(mu_t,Sigma2_t)
        x_sim(ii,:) = mu + sqrt(Sigma2).*randn(1,N);

        % log weight = likelihood p(y_t|x_t^(i))
        w(ii,:) = p_norm(y(ii,1),g(x_sim(ii,:)),sigma2_w);
        % transition probability p(x_t^(i)|x_(t-1)^(i))
        p = p_norm(x_sim(ii,:),f(x_aux(ii-1,:),ii),sigma2_v);
        % importance probability q(x_t^(i)|y_t,x_(t-1)^(i))
        q = p_norm(x_sim(ii,:),mu,Sigma2);
        % plus transition, minus importance
        w(ii,:) = w(ii,:)  + p - q;
    end
    w(ii,:) = log(w(ii-1,:)) + w(ii,:);
    w(ii,:) = w(ii,:) - max(w(ii,:)); % robustify
    w(ii,:) = exp(w(ii,:));
    w(ii,:) = w(ii,:)/sum(w(ii,:));
    ess(ii,1) = 1/(sum(w(ii,:).^2));
    x_est(ii,1) = sum(w(ii,:).*x_sim(ii,:));
    S_est(ii,1) = sum(w(ii,:).*(x_sim(ii,:).^2)) - x_est(ii,:)^2;

    % conditional resampling
    % if ESS too low perform resampling
    if ((resampl == 1) || ((resampl == 2) && (ess(ii,1) < k*N)))
        x_aux(ii,:) = randsample(x_sim(ii,:),N,true,w(ii,:))';
        w(ii,:) = 1/N;
    else
        x_aux(ii,:) = x_sim(ii,:);
    end

    %  compute the empirical distribution
    for b = bin_ax
        for jj = 1:N
            if (b <= x_sim(ii,jj)) && (x_sim(ii,jj) < b+1)
                epdf(ii,2*b+bin+1) = epdf(ii,2*b+bin+1) + 1;
            end
        end
    end
end
```

## C.3 Stochastic volatility model

```
% Particle filter example, nonlinear non-Gaussian model,
% simple univariate Stochastic Volatility model

clear all
s = RandStream('mt19937ar','Seed',0);
RandStream.setGlobalStream(s);

filter = 1;     % 0 = bootstrap;
                % 1 = locally optimal;
resampl = 2;    % 0 = no resampling;
                % 1 = always resapling;
                % 2 = resampling only below the threshold k*N
k = 0.5; % resampling threshold

alp = 0.1;
z = norminv(1-alp/2);

T = 500+1; % series length
N = 500;
t = 0:(T-1);

sigma2 = 0.03;
phi = 0.98;
beta = 0.6;
uv = sigma2/(1-phi^2); % unconditional state variance
% normal log density
p_norm = @(xx,mm,SS) -0.5*(log(2*pi) + log(SS) + ((xx-mm).^2)./SS);


%% generate the time series
x = zeros(T,1); % true unobserved state (log-volatility)
x(1,1) = sqrt(uv)*randn;   % initial state
y = zeros(T,1); % observation (log-returns)

for ii = 2:T
    x(ii,1) = phi*x(ii-1,1)+ sqrt(sigma2)*randn;
end
y = beta*exp(x/2).*randn(T,1);

%% particle filter initialisation
x_sim = zeros(T,N);   % particles
x_sim(1,:) = sqrt(uv)*randn(1,N);
x_aux = zeros(T,N);
x_aux(1,:) = x_sim(1,:);

x_est = zeros(T,1);   % filtered state estimate
S_est = zeros(T,1);   % filtered state variance
S_est(1,1) = uv;

w = ones(T,N)/N;
ess = zeros(T,1);
```

```
ess(1,1) = 1/(sum(w(1,:).^2));
bin = 5;
bg = 40; % bin grid
bin_ax = -bin:1/bg:bin;
epdf = zeros(T,2*bg*bin+1);

%% particle filter
for ii = 2:T
    if filter == 0
        % simuate from q=p i.e. prior
        x_sim(ii,:) = phi*x_aux(ii-1,:) + sqrt(sigma2)*randn(1,N);
        y_v = (beta^2)*exp(x_sim(ii,:)); % the current observation variance
        % log weight = likelihood p(y_t}|x_t^(i))
        w(ii,:) = p_norm(y(ii),0,y_v);
    else
        % get the normal approximation
        % mu = argmax[p(x|x_t-1)*p(y|x)]
        % Sigma = -1/l''(mu_opt)
        mu_opt = zeros(1,N);
        tic
        for jj = 1:N
            fg = @(xx) xx + ((xx-phi*x_aux(ii-1,jj)).^2)/uv
                + y(ii)^2/((beta^2)*exp(xx));
            mu_opt(1,jj) = fminsearch(fg,0);
        end
        time = toc;
        fprintf('Iteration %i, opt. time %4.2f.\n',ii,time)
        tmp = (beta^2)*exp(mu_opt); % observation variance
        Sigma2_opt = 1./(1/uv + (y(ii)^2)./(2*tmp));

        % simulate from q=N(mu,Sigma) i.e. normal approximation
        x_sim(ii,:) = mu_opt + sqrt(Sigma2_opt).*randn(1,N);

        y_v = (beta^2)*exp(x_sim(ii,:)); % the current observation variance
        % log weight = likelihood p(y_t}|x_t^(i))
        w(ii,:) = p_norm(y(ii),0,y_v) ;
        % plus transition, minus importance
        w(ii,:)= w(ii,:) + p_norm(x_sim(ii,:),phi*x_aux(ii-1,:),uv)
                    - p_norm(x_sim(ii,:),mu_opt,Sigma2_opt);
    end

    w(ii,:) = w(ii,:) + log(w(ii-1,:));
    w(ii,:) = exp(w(ii,:));
    w(ii,:) = w(ii,:)/sum(w(ii,:));
    ess(ii,1) = 1/(sum(w(ii,:).^2));
    x_est(ii,1) = sum(w(ii,:).*x_sim(ii,:));
    S_est(ii,1) = sum(w(ii,:).*(x_sim(ii,:).^2)) - x_est(ii,:)^2;

    %conditional resampling
    % if ESS too low perform resampling
    if ((resampl == 1) || ((resampl == 2) && (ess(ii,1) < k*N)))
        x_aux(ii,:) = randsample(x_sim(ii,:),N,true,w(ii,:))';
        w(ii,:) = 1/N;
```

```
    else
        x_aux(ii,:) = x_sim(ii,:);
    end

    % compute the empirical distribution
    for b = bin_ax
        for jj = 1:N
            if (b <= x_sim(ii,jj)) && (x_sim(ii,jj) < b+1)
                epdf(ii,round(bg*(b+bin)+1)) = epdf(ii,round(bg*(b+bin)+1)) + 1;
            end
        end
    end

end
```