Sequential Monte Carlo Methods for Bayesian Filtering

Andreas Størksen Stordal

Thesis for the degree of Master of Science
Mathematical Statistics
University of Bergen, Norway
29th May 2008

UNIVERSITETET I BERGEN
Matematisk institutt

This thesis is written in LaTeX 2$_\varepsilon$ with the ‘uib-mi-master’ document class, developed by Karl Ove Hufthammer. It was compiled using pdfeTeX-1.21a on 29th May 2008. The body text is 11 point URW Palladio with small caps. The maths font is URW Palladio and Pazo Math, the heading font is HV Math, and the computer code font is Bera Mono.

Acknowledgements

I would first like to thank my supervisor Hans Arnfinn Karlsen for good guidance on my thesis and for many advices and insightful discussions about statistics in general over the last two years.
I would also like to thank the following persons:
Geir Storvik for mail correspondence.
My fellow students for many hours of theoretical discussions and social activities: Andreas Henriksen, Arne Kristoffersen, Arne Holmin, Geir Berentsen, Håkon Gundersen, Morten Aarflot, Thomas Djønne and Trond Gangsøy at UIB.
Arild Vivås, Anders Leirvig, Henrik Hemmen, Steinar Ellingsen and Torgeir Anda at NTNU.
Øyvind Spilde, Terje Heggernes and Chimseth A. Bjølseth for making it possible for me to combine study and football.
Koop, for that legendary table hockey match.

Contents

1 Bayesian Filtering 3
1.1 Problem statement and its conceptual solution 4
1.1.1 Grid-based methods 7
1.1.2 Multiple switching dynamic models 8

2 Particle filters 10
2.1 Monte Carlo methods 10
2.1.1 Perfect Monte Carlo sampling 10
2.1.2 Importance sampling 11
2.1.3 Selecting importance density 13
2.2 The SIR filter 15
2.2.1 Improving diversity 19
2.3 The ASIR filter 22
2.4 Multiple model particle filters 24
2.5 Combined parameter and state estimation 27
2.5.1 Framework 27
2.5.2 Filtering for states and parameters 27

3 Posterior Cramèr-Rao bounds 34
3.1 General case 34
3.2 PCRB for the nonlinear filter problem 36
3.3 Special Cases 39
3.3.1 Additive Gaussian noise 39

4 Convergence 41
4.1 Introduction 41
4.2 The filtering problem 41
4.3 Convergence of measure-valued random variables 42
4.3.1 Convergence theorems for the fixed observation case 44
4.4 Interacting particle filters with discrete observations 50
4.4.1 General facts about nonlinear filtering 52
4.5 An interacting particle system under case A 57

4.5.1 Subcase A2 . . . 64
4.6 An interacting particle system under Case B . . . 66
4.6.1 Subcase B1 . . . 66
4.6.2 The interacting particle system . . . 74
4.6.3 Convergence study . . . 74

5 A more general filter problem 78
5.1 Problem statement . . . 78
5.1.1 Bayes recursions . . . 79
5.2 Sequential MC methods . . . 80
5.2.1 Importance sampling step . . . 82
5.2.2 Resampling step . . . 84
5.3 Convergence study . . . 85
5.3.1 Bounds for the mean square error . . . 85
5.3.2 Almost sure convergence . . . 92

6 Discussion 100

A Conditional expectations and probabilities 102
A.1 Conditional expectations and probabilities . . . 102
A.2 The recurrence formula for the conditional distribution of the signal . . . 105

Bibliography 109

List of Figures

2.1 SIR filter with N=30 particles . . . 17
2.2 Density estimation for N=30 . . . 17
2.3 SIR filter with N=1000 particles . . . 18
2.4 Density estimation with N=1000 . . . 18
2.5 Multiple model particle filter . . . 25
2.6 Histograms of the particles . . . 26
2.7 Combined parameter and state, phi unknown . . . 31
2.8 Combined parameter and state, phi and sigma unknown, $N = 500$ . . . 32
2.9 Combined parameter and state, phi and sigma unknown, $N = 5000$ . . . 33

5.1 The particle filters, N=1000 . . . 82

List of Algorithms

2.1 The SIS algorithm . . . 13
2.2 The SIR algorithm . . . 15
2.3 The RPF algorithm . . . 21
2.4 The ASIR algorithm . . . 23
2.5 Regime conditioned SIR algorithm . . . 24
2.6 Combined parameter and state estimation . . . 30
5.1 General particle filter algorithm . . . 81
5.2 An additional MCMC step . . . 96

Notation and Terminology

Notation

$\mathbb{R}^d$	The d-dimensional Euclidean space
$\mathcal{B}(\mathbb{R}^d)$	The Borel σ-algebra of subsets of $\mathbb{R}^d$
$\mathbf{B}(\mathbb{R}^d)$	The set of bounded $\mathcal{B}(\mathbb{R}^d)$-measurable functions on $\mathbb{R}^d$
$\mathbf{C}_b(\mathbb{R}^d)$	The set of bounded continuous on $\mathbb{R}^d$
$\mathbf{C}_k(\mathbb{R}^d)$	The set of continuous functions on $\mathbb{R}^d$ with compact support
$\|f\|$	The sup norm of $f \in \mathbf{B}(\mathbb{R}^d)$, $\|f\| \triangleq \sup_{x\in\mathbb{R}^d}\lvert f(x)\rvert$
$\mathcal{M}_F(\mathbb{R}^d)$	The set of finite measures over $\mathcal{B}(\mathbb{R}^d)$
$\mathcal{P}(\mathbb{R}^d)$	The set of probability measures over $\mathcal{B}(\mathbb{R}^d)$
μf	The integral of $f \in \mathbf{B}(\mathbb{R}^d)$ w.r.t. $\mu \in \mathcal{M}_F(\mathbb{R}^d)$, $\mu(f) = \int_{\mathbb{R}^d} f(x)\mu(\mathrm{d}x)$
$\mathbb{E}$	The expectation operator
$\|f\|_1$	The L^1 norm, $\|f\|_1 = \mathbb{E}\lvert f\rvert$
$\|f\|_2^2$	The L^2 norm, $\|f\|_2^2 = \mathbb{E}[f^2]$
η_t	The prediction measure conditioned on $Y_{0:t-1}$, $\eta_t(f) \triangleq \mathbb{E}[f\lvert Y_{0:t-1} = y_{0:t-1}]$
$\hat{\eta}_t$	The update measure conditioned on $Y_{0:t}$, $\hat{\eta}_t(f) \triangleq \mathbb{E}[f\lvert Y_{0:t} = y_{0:t}]$

We endow $\mathcal{M}_F(\mathbb{R}^d)$ and $\mathcal{P}(\mathbb{R}^d)$ with the weak topology (Royden, 1988, page 236-237), saying that if $(\mu_n)_{n=1}^{\infty}$ is a sequence of finite measures, we have the following types of convergence to $\mu \in \mathcal{M}_F(\mathbb{R}^d)$.

$$\begin{aligned}
\mu_n &\xrightarrow[n]{w} \mu && \text{if} \quad \mu_n f \xrightarrow[n]{} \mu f \quad \forall f \in \mathbf{C}_b(\mathbb{R}^d)\\
\mu_n &\xrightarrow[n]{a.s.w} \mu && \text{if} \quad \mu_n f \xrightarrow[n]{a.s} \mu f \quad \forall f \in \mathbf{C}_b(\mathbb{R}^d)\\
\mu_n &\xrightarrow[n]{\text{Elim}} \mu && \text{if} \quad \mu_n f \xrightarrow[n]{L^1} \mu f \quad \forall f \in \mathbf{C}_b(\mathbb{R}^d)
\end{aligned}$$

We also denote by $\xrightarrow{D}$ the convergence in distribution and by $\delta_a(x)$ the delta-Dirac mass located in a.

Markov Chains and Transition Kernels

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and $X = \{X_t, t \in \mathbb{N}\}$ be a stochastic process defined on $(\Omega, \mathcal{F}, \mathcal{P})$ taking it's values in $\mathbb{R}^d$, where d is the dimension of each variable $X(t)$. Let $\mathcal{F}_t^X$ be the σ-algebra generated by the process up to time t i.e. $\mathcal{F}_t^X = \sigma(X_s, s \in [0,t])$. Then X_t is a Markov chain if, for all $t \in \mathbb{N}$ and $A \in \mathcal{B}(\mathbb{R}^d)$,

$$P(X_{t+1} \in A | \mathcal{F}_t^X) = P(X_{t+1} \in A | X_t) \quad a.s..$$

The transition kernel of the Markov chain X is the function $Q_t(\cdot,\cdot)$ defined on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ such that for all $t \in \mathbb{N}$ and $x \in \mathbb{R}^d$,

$$Q_t(x, A) = P(X_{t+1} \in A | X_t = x).$$

The transition kernel Q_t satisfies the following properties

- $Q_t(x,\cdot)$ is a probability measure on $\mathbb{R}^d$ for all $t \in \mathbb{N}$ and all $x \in \mathbb{R}^d$
- $Q_t(\cdot, A) \in \mathbf{B}(\mathbb{R}^d)$ for all $t \in \mathbb{N}$ and all $A \in \mathcal{B}(\mathbb{R}^d)$.

The distribution of X is uniquely determined by its initial distribution and its transition kernel.

If we denote by q_t the marginal distribution of X_t

$$q_t \triangleq P(X_t \in A),$$

then we can deduce from the previous that q_t satisfies the recurrence formula $q_{t+1} = q_t Q_t$ where $q_t Q_t$ is the measure defined as

$$(q_t Q_t)(A) \triangleq \int_{\mathbb{R}^d} Q_t(x, A) q_t(\mathrm{d}x).$$

Hence, $q_t = q_0 Q_0 Q_1 ... Q_{t-1}$.

We say that the transition kernel Q_t satisfies the (weak) Feller property (Meyn and Tweedie, 1993) if, for all $t \geq 0$ the function $Q_t f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$Q_t f(x) \triangleq \int_{\mathbb{R}^d} f(y) Q_t(x, \mathrm{d}y)$$

is continuous for every $f \in \mathbf{C}_b(\mathbb{R}^d)$. If Q_t satisfies the Feller property, then $Q_t f \in \mathbf{C}_b(\mathbb{R}^d)$ for all $f \in \mathbf{C}_b(\mathbb{R}^d)$.

1 Bayesian Filtering

Many data analysis problems within science or engineering involve estimation of unknown quantities based on some given observations. Very often we have some prior knowledge about the phenomenon being modelled. This will allow us to formulate Bayesian models based on prior distributions for the unknown quantities and likelihood functions relating these to the observations. All inference in this setting will then be based on the posterior distribution obtained from *Bayes' theorem*. In many settings the observations arrive sequentially in time and it is therefor necessary to update the posterior distribution in order to perform on-line inference.

In the situation when the data are modelled as a linear Gaussian state-space model, it is possible to obtain an exact analytic expression for the posterior distribution from the well known Kalman equations. There are other situations where it is possible to derive analytical solutions, however in most situations, where we have neither linearity nor Gaussian processes we cannot derive analytical expression due to complex high order integrals. In these settings we have to use approximating solutions.

In this thesis we will mainly focus on the approximations obtained from Sequential Monte Carlo methods (SMC). The SMC are simulation based methods which provide an easy-to-implement approach to compute the posterior distribution. These methods are usually based on two models, the first describing the evolution of the unknown quantities and the second relating these quantities to the observations. Over the past years, several closely related algorithms has been proposed under different names such as the bootstrap filter, particle filters, Monte Carlo filters, interacting particle approximations etc.

This thesis will mainly focus on the theoretical aspects of the filter methods, however we will present a few trivial examples.

1.1 Problem statement and its conceptual solution

To define the nonlinear filtering problem we introduce the target state vector and the observed state vector. The unobserved signal $\{X_t\}_{t\in\mathbb{N}}$, $X_t \in \mathbb{R}^d$, is modelled as a (possible) nonlinear Markov process with initial distribution $p(x_0)$ and 1-step transitions $p(x_t|x_{t-1})$. The observation process $\{Y_t\}_{t\in\mathbb{N}}$, $y_t \in \mathbb{R}^q$ are assumed to be conditionally independent given the process $\{X_t\}$, with marginal distribution $p(y_t|x_t)$, where we denote $p(\cdot)$ as the probability density function with the argument of the function indicating the random variables under consideration, at least when there is no danger of confusion, i.e $p(x_t|y_t) \triangleq p_{X_t|Y_t}(x_t|y_t)$.

In other words, the entire model is described by

$$\begin{aligned}
&p(x_0)\\
&p(x_t|x_{t-1})\ \text{, for } t \geq 1\\
&p(y_t|x_t)\ \text{, for } t \geq 0.
\end{aligned}$$

Let us denote $X_{0:t} \triangleq (X_0, ..., X_t)$ and $Y_{0:t} \triangleq (Y_0, ..., Y_t)$, the signal and the observations up to time t

We are interested in estimating recursively in time the posterior distribution $p(x_{0:t}|y_{0:t})$ (or $p(x_t|y_{0:t})$), and the expectations

$$pf = \mathbb{E}\left[f(X_{0:t})|Y_{0:t}\right] \triangleq \int f(x_{0:t})p(\mathrm{d}x_{0:t}|y_{0:t})$$

and its marginal $\mathbb{E}\left[f(X_t)|Y_{0:t}\right]$, for some function f integrable with respect to the density of $X_{0:t}$.

The problems and solutions in Chapter 1 and 2 are discussed in Doucet, de Freitas and Gordon (2001) and Ristic, Arulampalam and Gordon (2004)

Very often such a model is described by the state and observation equations

$$X_t = k_{t-1}(X_{t-1}, V_{t-1}) \qquad Y_t = h_t(X_t, W_t) \tag{1.1}$$

for some nonlinear functions k and h, where we assume that V_{t-1} and W_t are independent. The model is fully described by the densities, $p(x_0)$, $p(x_t|x_{t-1})$ and $p(y_t|x_t)$. The posterior pdf is given by *Bayes' theorem* at any time t,

$$p(x_{0:t}|y_{0:t}) = \frac{p(y_{0:t}|x_{0:t})p(x_{0:t})}{\int p(y_{0:t}|x_{0:t})p(x_{0:t})\,\mathrm{d}x_{0:t}}. \tag{1.2}$$

It is possible to obtain a recursive formula for this joint distribution.

$$\begin{aligned}
p(x_{0:t+1}|y_{0:t+1}) &= \frac{p(x_{0:t+1}, y_{0:t+1})}{p(y_{0:t+1})} \\
&= \frac{p(x_{t+1}, y_{t+1}|x_{0:t}, y_{0:t})p(x_{0:t}, y_{0:t})}{p(y_{t+1}|y_{0:t})p(y_{0:t})} \\
&= \frac{p(y_{t+1}|x_{0:t+1}, y_{0:t})p(x_{t+1}|x_{0:t}, y_{0:t})p(x_{0:t}, y_{0:t})}{p(y_{t+1}|y_{0:t})p(y_{0:t})} \\
&= \frac{p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t)p(x_{0:t}|y_{0:t})p(y_{0:t})}{p(y_{t+1}|y_{0:t})p(y_{0:t})} \\
&= p(x_{0:t}|y_{0:t})\frac{p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t)}{p(y_{t+1}|y_{0:t})}.
\end{aligned} \tag{1.3}$$

We also have some recursive formulae for the marginal distribution $p(x_t|y_{0:t})$

$$\begin{aligned}
\textit{prediction : } p(x_t|y_{0:t-1}) &= \int p(x_t, x_{t-1}|y_{0:t-1})\,\mathrm{d}x_{t-1} \\
&= \int p(x_t|x_{t-1}, y_{0:t-1})p(x_{t-1}|y_{0:t-1})\,\mathrm{d}x_{t-1} \\
&= \int p(x_t|x_{t-1})p(x_{t-1}|y_{0:t-1})\,\mathrm{d}x_{t-1}
\end{aligned} \tag{1.4}$$

$$\begin{aligned}
\textit{updating : } p(x_t|y_{0:t}) &= \frac{p(x_t, y_{0:t})}{p(y_{0:t})} \\
&= \frac{p(y_t|x_t, y_{0:t-1})p(x_t|y_{0:t-1})p(y_{0:t-1})}{p(y_t|y_{0:t-1})p(y_{0:t-1})} \\
&= \frac{p(y_t|x_t)p(x_t|y_{0:t-1})}{\int p(y_t|x_t)p(x_t|y_{0:t-1})\,\mathrm{d}x_t}.
\end{aligned} \tag{1.5}$$

The problem with these equations are the calculation of the normalising constant $p(y_{0:t})$ and the marginals of $p(x_{0:t}|y_{0:t})$ since these may require the evaluation of complex high-dimensional integrals. To solve these problems we need to use numerical approximation methods. Throughout this thesis we will study some of the Monte Carlo approximations proposed over the years. However, there are certain cases where it is possible to obtain optimal algorithms for recursive Bayesian state estimation.

1. In the linear-Gaussian case the functional recursions becomes the Kalman filter.

2. If the state space is discrete-valued with a finite number of states. This is called grid-based method.

3. For a certain class of nonlinear problems, discovered by Beneš (1981) and Daum (1986), it is possible to formulate exact analytical solutions, but these will not be discussed in this thesis.

The Kalman Filter

The Kalman filter is a recursive algorithm for finding the best (in terms of mean square error) linear estimates of the state-vector X_t in terms of the observations $Y_{0:t}$ (or $Y_{0:t-1}$) To apply the Kalman filter to our problem, we must assume that the posterior density at each time step is Gaussian and therefor completely characterised by it's mean and covariance. If $p(x_{t-1}|y_{1:t-1})$ is Gaussian, it can be proved that $p(x_t|y_t)$ is Gaussian, provided that certain assumptions hold:

K.1 V_{t-1} and W_t are Gaussian.

K.2 $k_{t-1}(X_{t-1}, V_{t-1})$ is a linear function of X_{t-1} and V_{t-1}.

K.3 $h_t(X_t, W_t)$ is a linear function of X_t and W_t.

In other words (1.1) can be written as

$$X_t = \mathbb{K}_{t-1}X_{t-1} + V_{t-1} \qquad Y_t = \mathbb{H}_t X_t + W_t$$

where $\mathbb{K}_{t-1}$ and $\mathbb{H}_t$ are matrices defining the linear functions. The noise V_{t-1} and W_t are mutually independent zero-mean Gaussian with covariances $\mathbb{Q}_{t-1}$ and $\mathbb{R}_t$. The Kalman filter will then consist of the following recursive relationship

$$p(x_{t-1}|y_{0:t}) = \mathcal{N}(x_{t-1}; \hat{X}_{t-1|t-1}, \mathbb{P}_{t-1|t-1})$$

$$p(x_t|y_{0:t-1}) = \mathcal{N}(x_t; \hat{X}_{t|t-1}, \mathbb{P}_{t|t-1})$$

$$p(x_t|y_{0:t}) = \mathcal{N}(x_t; \hat{X}_{t|t}, \mathbb{P}_{t|t}),$$

where $\hat{X}_{t|t}$ is the filter estimate of X_t given $Y_{0:t}$, and $\mathbb{P}_{t|t}$ is the error covariance matrix $\mathbb{E}\left[\left(X_t - \hat{X}_{t|t}\right)\left(X_t - \hat{X}_{t|t}\right)^T\right]$. We will not go into details here, but present the recursions for computing the mean and covariances:
Start by the initial condition $\hat{X}_0 = \mathbb{E}X_0$ and $\mathbb{P}_0 = \mathbb{E}\left[\left(X_0 - \hat{X}_0\right)\left(X_0 - \hat{X}_0\right)\right]$ and then continue by induction on t.

$$\hat{X}_{t|t-1} = \mathbb{K}_{t-1}\hat{X}_{t-1|t-1}$$

$$\mathbb{P}_{t|t-1} = \mathbb{Q}_{t-1} + \mathbb{K}_{t-1}\mathbb{P}_{t-1|t-1}\mathbb{K}_{t-1}^T$$

$$\hat{X}_{t|t} = \hat{X}_{t|t-1} + \mathbb{D}_t(Y_t - \mathbb{H}_t\hat{X}_{t|t-1})$$

$$\mathbb{P}_{t|t} = \mathbb{P}_{t|t-1} - \mathbb{D}_t\mathbb{S}_t\mathbb{D}_t^T,$$

where

$$\mathbb{S}_t = \mathbb{H}_t\mathbb{P}_{t|t-1}\mathbb{H}_t^T + \mathbb{R}_t$$

is the covariance of the innovation term $I_t = Y_t - \mathbb{H}_t\hat{X}_{t|t-1}$, and

$$\mathbb{D}_t = \mathbb{P}_{t|t-1}\mathbb{H}_t^T\mathbb{S}_t^{-1}$$

is the Kalman gain. This algorithm is a combination of the Kalman prediction and Kalman filter equations (Brockwell and Davis, 2002). Under the assumptions K.1-3, the Kalman filter provides the best linear prediction/update in terms of $Y_{0:t-1}$ and $Y_{0:t}$.

Proof: see Brockwell and Davis (2002, page 271-277).

1.1.1 Grid-based methods

If the state space is discrete and consists of a finite number of states, then grid-based methods provide the optimal solution of the filtering density $p(x_t|y_{0:t})$. Let $\{s^{(i)}\}_{i=1}^{N}$ be the states of the state space. Given the measurements up to time $t-1$, let us denote $\omega_{t-1|t-1}^{(i)} = p(x_{t-1} = s^i|y_{0:t-1})$. The posterior pdf at time $t-1$ can now be written as

$$p(x_{t-1}|y_{0:t-1}) = \sum_{i=1}^{N} \omega_{t-1|t-1}^{(i)} \delta_{s^{(i)}}(x_{t-1}). \tag{1.6}$$

Substituting (1.6) into (1.4) and (1.5) we get the following equations for the prediction and updating.

$$p(x_t|y_{0:t-1}) = \sum_{i=1}^{N} \omega_{t|t-1}^{(i)} \delta_{s^{(i)}}(x_t)$$

$$p(x_t|y_{0:t}) = \sum_{i=1}^{N} \omega_{t|t}^{(i)} \delta_{s^{(i)}}(x_t),$$

where

$$\omega_{t|t-1}^{(i)} \triangleq \sum_{j=1}^{N} \omega_{t-1|t-1}^{(j)} p(x_t = s^{(i)}|x_{t-1} = s^{(j)})$$

$$\omega_{t|t}^{(i)} \triangleq \frac{\omega_{t|t-1}^{(i)} p(y_t|x_t = s^{(i)})}{\sum_{j=1}^{N} \omega_{t|t-1}^{(j)} p(y_t|x_t = s^{(j)})}.$$

Now if the density and likelihood functions $p(x_t = s^{(i)}|x_{t-1} = s^{(j)})$ and $p(y_t|x_t = s^{(i)})$ are known, we have the optimal solution to our problem

1.1.2 Multiple switching dynamic models

Nonlinear dynamic systems that are characterised by some modes or regimes of operation is very common in engineering. These problems are often referred to as jump Markov or hybrid-state estimation problems. They involve a continuous-valued target state and a discrete-valued regime model. The system is described by the following

$$X_t = k_{t-1}(X_{t-1}, r_t, V_{t-1})$$

$$Y_t = h_t(X_t, r_t, W_t),$$

where r_t is the effective regime during the period $(t_{t-1}, t_t]$. Usually the regime is modelled as an s-state time-homogeneous Markov chain with transition probabilities

$$p_{ij} \triangleq P(r_k = j | r_{t-1} = i) \quad (i, j \in S),$$

where $S \triangleq (1, 2, ..., s)$. The corresponding transition probability matrix (TPM) $\mathbb{P} = [p_{ij}]$ is an $s \times s$ matrix with elements satisfying

$$p_{ij} \geq 0 \text{ and } \sum_{j=1}^{s} p_{ij} = 1$$

for each $i, j \in S$. We also denote $\mu_i = p(r_1 = i)$ as the initial regime probabilities, such that

$$\mu_i \geq 0 \text{ and } \sum_{i=1}^{s} \mu_i = 1.$$

(Notice that if $s = 1$ we are back to our intital problem in equation (1.1)).

By conditioning on X_{t-1} and r_t and using the law of total probability we have the following generalisation of (1.4);
Prediction:

$$p(x_t, r_t = j | y_{0:t-1}) = \sum_i p_{ij} \int p(x_t | x_{t-1}, r_t = j) p(x_{t-1}, r_{t-1} = i | y_{0:t-1}) \, \mathrm{d}x_{t-1}. \tag{1.7}$$

Update:

$$\begin{aligned} p(x_t, r_t = j | y_{0:t}) &= \frac{p(x_t, r_t = j, y_{0:t})}{p(y_{0:t})} \\ &= \frac{p(y_t | x_t, r_t = j, y_{0:t-1}) p(x_t, r_t = j, y_{0:t-1})}{p(y_t | y_{0:t-1}) p(y_{0:t-1})} \\ &= \frac{p(y_t | x_t, r_t = j) p(x_t, r_t = j | y_{0:t-1}) p(y_{0:t-1})}{p(y_t | y_{0:t-1}) p(y_{0:t-1})}. \end{aligned}$$

Again conditioning on x_{t-1} and r_t and using the law of total probability we obtain a generalisation of (1.5)

$$p(x_t, r_t = j|y_{0:t-1}) = \frac{p(y_t|x_t, r_t = j)p(x_t, r_t = j|y_{0:t-1})}{\sum_i \int p(y_t|x_t, r_t = i)p(x_t, r_t = i|y_{0:t-1})\,\mathrm{d}x_t}. \tag{1.8}$$

In the next chapter we will discuss numerical solutions to these filtering problems but our main focus will be on the problem stated by (1.1).

2
Particle filters

2.1 Monte Carlo methods

With the increasing computational power since the late 80's, there has been devoted a great effort to approximate integrals with Monte Carlo methods. These methods do not require any linearity or Gaussian constraints on the model and have nice convergence properties. (Unlike numerical methods, the rate of convergence does not depend on the dimension of the integrand, although methods like importance sampling are usually inefficient in high-dimensions).

In this section we start by showing that if one has a large number of samples from the posterior distribution of interest, it is not difficult to approximate the desired expected value.

2.1.1 Perfect Monte Carlo sampling

Assume now that we are able to draw N independent and identically distributed (iid) random samples, called particles, $\{X_{0:t}^{(i)}\}_{i=1}^{N}$ according to $p_t(x_{0:t}|y_{0:t})$. An empirical estimate of this simultaneous distribution is given by

$$p^N(\mathrm{d}x_{0:t}|y_{0:t}) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_{0:t}^{(i)}}(\mathrm{d}x_{0:t}).$$

From this, a natural estimate of $p_t f$ is

$$p_t^N f = \int f(x_{0:t})p^N(\mathrm{d}x_{0:t}|y_{0:t}) = \frac{1}{N}\sum_{i=1}^{N} f(x_{0:t}^{(i)}).$$

This estimate is unbiased, and if the posterior variance $\sigma_f^2 < \infty$, the variance of $\hat{p}_t^N(f)$ is equal to σ_f^2/N. From the strong law of large numbers we have

$$p_t^N f \xrightarrow[N\to\infty]{a.s.} p_t(f). \tag{2.1}$$

Also if $\sigma_f^2 < \infty$, then the central limit theorem holds and

$$\sqrt{N}\,[p_t^N f - p_t f] \xrightarrow{D} N(0, \sigma_f^2). \tag{2.2}$$

This procedure is however very troublesome. Since $p(x_{0:t}|y_{0:t})$ is multivariate and known only up to a multiplicative constant (see 1.3), it will be almost impossible to draw from. One can apply MCMC methods, but they are unsuited for recursive estimation procedures.

2.1.2 Importance sampling

An alternative solution to our estimation problem is the classical method of importance sampling.

For a given distribution function $q(x_{0:t}|y_{0:t})$
(or possibly $q(x_{0:t})$) with $\mathrm{supp}(q) \supseteq \mathrm{supp}(p_t)$ we have the identity

$$\begin{aligned} p_t f &= \int f(x_{0:t}) p(x_{0:t}|y_{0:t})\,\mathrm{d}x_{0:t} \\ &= \frac{\int f(x_{0:t}) w(x_{0:t}) q(x_{0:t}|y_{0:t})\,\mathrm{d}x_{0:t}}{\int w(x_{0:t}) q(x_{0:t}|y_{0:t})\,\mathrm{d}x_{0:t}} \end{aligned}$$

where

$$w(x_{0:t}) = \frac{p(x_{0:t}|y_{0:t})}{q(x_{0:t}|y_{0:t})}. \tag{2.3}$$

This can be written as

$$p_t f = \mathbb{E}_p f = \mathbb{E}_q \left[\frac{f w_t}{w_t} \right].$$

An estimator for $p_t f$ is given by

$$p_t^N f = \frac{\frac{1}{N}\sum_{i=1}^N f(X_{0:t}^{(i)}) w(X_{0:t}^{(i)})}{\frac{1}{N}\sum_{i=1}^N w(X_{0:t}^{(i)})} = \sum_{i=1}^N f(X_{0:t}^{(i)}) \tilde{w}_t^{(i)},$$

where

$$\begin{aligned}
X_{0:t}^{(i)} &\sim q(x_{0:t}|y_{0:t}), \quad i = 1, \dots, N \\
w_t^{(i)} &\triangleq w(X_{0:t}^{(i)}) \\
\tilde{w}_t^{(i)} &\triangleq \frac{w(X_{0:t}^{(i)})}{\sum_{j=1}^{N} w(X_{0:t}^{(j)})}.
\end{aligned}$$

The distribution function q, which we draw our sample from, is called the importance function and the sample $(X_{0:t}^{(i)}, i = 1, ...N)$ we will be our particles. The estimator in (2.1.2) is biased (the ratio of two estimators), but asymptotically (2.1) and (2.2) holds. Also we only need to know $p(x_{0:t}|y_{0:t})$ up to a normalising constant. Although this method is simple with nice convergence properties it is not recursive. In general at time $t+1$, y_{t+1} becomes available, then we have to recalculate all the importance weights over the entire state sequence. This is time demanding and becomes more and more complex as time increases.

Sequential Importance sampling

We will now try to modify the importance sampling method in such a way that we can compute the estimate at time t without modifying the particles and weights obtained at time $t-1$. That is, we want to compute our estimate at time t with the help of the particles $(X_{0:t-1}^{(i)}, i = 1, ..., N)$ and the importance weights $(\tilde{w}_{0:t-1}^{(i)})$. If we can choose the importance function in such a way that we may draw new particles at time t from the particles at the previous step, then we can simple set $(X_{0:t}^{(i)}, i = 1, ..., N) = (X_{0:t-1}^{(i)}, X_t^{(i)})$. This can be done by choosing $q(x_{0:t}|y_{0:t})$ so that it satisfy the following recursion

$$q(x_{0:t}|y_{0:t}) = q(x_t|x_{0:t-1}, y_{0:t})q(x_{0:t-1}|y_{0:t-1}). \tag{2.4}$$

This will allow us to evaluate the weights recursively in time. From (1.3), (2.3) and (2.4) we have

$$\begin{aligned}
w_t \triangleq w(x_{0:t}) &= \frac{p(x_{0:t}|y_{0:t})}{q(x_{0:t}|y_{0:t})} \\
&\propto \frac{p(x_{0:t-1}|y_{0:t-1})p(y_t|x_t)p(x_t|x_{t-1})}{q(x_t|x_{0:t-1}, y_{0:t})q(x_{0:t-1}|y_{0:t-1})} \\
&= w_{t-1}\frac{p_t(y_t|x_t)p(x_t|x_{t-1})}{q(x_t|x_{0:t-1}, y_{0:t})}.
\end{aligned} \tag{2.5}$$

If $q(x_t|x_{0:t-1}, y_{0:t}) = q(x_t|x_{t-1}, y_t)$ then the importance density depend only on x_{t-1} and y_t. This case is very useful when we are interested in the filter estimate $p(x_t|y_{0:t})$. Algorithm 2.1 gives a description for carrying out an SIS system.

Algorithm 2.1: The SIS algorithm

```
Initialisation $t = 0$;
for $i = 1 : N$ do
    Sample $X_0^{(i)} \sim p(x_0)$;
end
for $t = 1 : T$ do
    for $i = 1 : N$ do
        Sample $X_t^{(i)} \sim p(x_t | X_{t-1}^{(i)})$ ;
        Evaluate the importance weights $w_t^{(i)} = p(y_t | X_t^{(i)})$ ;
    end
    Normalise the importance weights $\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}}$ ;
end
```

2.1.3 Selecting importance density

The most critical issue in constructing a sequential importance sampling design, or even an ordinary importance sampling, is the choice of the importance density $q(x_{0:t}|y_{0:t})$ In view of (2.4) we want the optimal choice of $q(x_t|x_{0:t-1}, y_{0:t})$.

Optimal choice

If we choose the importance function q such that $q(x_t|x_{0:t-1}, y_{0:t}) = q(x_t|x_{t-1}, y_t)$, (this is a smart choice since we only have to store one set of variables) then the optimal choice of importance density function, the one that minimises the variance of the importance weights conditioning upon $X_{t-1}^{(i)}$ and y_t, is

$$
\begin{aligned}
q(x_t|X_{t-1}^{(i)}, y_t)_{opt} &= p(x_t|X_{t-1}^{(i)}, y_t) \\
&= \frac{p(y_t|x_t)p(x_t|X_{t-1}^{(i)})}{p(y_t|X_{t-1}^{(i)})}.
\end{aligned} \tag{2.6}
$$

Proof: From (2.5) we have

$$\begin{aligned}
\mathrm{Var}_q(w_t^{(i)}) &\propto \mathrm{Var}_q\left(w_{t-1}^{(i)} \frac{p(y_t|X_t)p(X_t|X_{t-1}^{(i)})}{q(X_t|X_{t-1}^{i}, y_t)} \right) \\
&= (w_{t-1}^{(i)})^2 \left[\int \frac{(p(y_t|x_t)p(x_t|X_{t-1}^{(i)}))^2}{q(x_t|x_{t-1}^{(i)}, y_t)} \,\mathrm{d}x_t - \left(\int p(y_t|x_t)p(x_t|X_{t-1}^{i}) \,\mathrm{d}x_t \right)^2 \right] \\
&= (w_{t-1}^{(i)})^2 \left[\int p(y_t|x_t)p(x_t|X_{t-1}^{i}) \,\mathrm{d}x_t p(y_t|X_{t-1}^{i}) - [p(y_t|X_{t-1}^{i})]^2 \right] \\
&= (w_{t-1}^{(i)})^2([p(y_t|X_{t-1}^{i})]^2 - [p(y_t|X_{t-1}^{i})]^2) = 0.
\end{aligned}$$

□

Substituting (2.6) into (2.5) yields

$$w_t^{(i)} \propto w_{t-1}^{(i)} p(y_t|X_{t-1}^{(i)})$$

saying that the importance weights can be computed *before* the particles are propagated to time t. However, as often, the optimal solution is rarely possible. In order to use the optimal importance density we have to be able to;

1. Sample from $p(x_t|X_{t-1}^{i}, y_t)$
2. Evaluate $p(y_t|X_{t-1}^{i}) = \int p(y_t|x_t)p(x_t|X_{t-1}^{i}) \,\mathrm{d}x_t$ up to a normalising constant.

Generally (1) is not straightforward, and (2) may be difficult to compute.

However, in certain special cases, it is possible to use the optimal importance density, an example is when $p(x_t|X_{t-1}^{i}, y_t)$ is Gaussian.

Suboptimal choice

A particular case arise when we use the prior distribution as importance function.

$$q(x_{0:t}|y_{0:t}) = p(x_{0:t}) = p(x_0) \prod_{k=1}^{t} p(x_k|x_{k-1}). \tag{2.7}$$

This is an important case that satisfies (2.4). The importance weights now satisfy $w_t^{(i)} \propto w_{t-1}^{(i)} p(y_t|X_t^{(i)})$. The problem now is that as t increases, the distribution of the weights becomes more and more skewed. It can be shown (Doucet *et al.*, 2000) that the variance of the weights is non-decreasing over time. After a few time steps most particles will have zero weight. To avoid this problem and still use (2.7), we introduce the SIR(bootstrap) filter, in this procedure all the particles will have uniform weights after a resampling step.

2.2 The SIR filter

The SIR filter is an easy way to avoid the problem with skewness of the importance weights. We want to multiply the particles with large weights and get rid of the ones with small weights. The idea is to attach to each set of particles $X_{0:t}^{(i)}$ (or $X_t^{(i)}$) a random number $N_t^{(i)}$ such that $\sum_{i=1}^N N_t^{(i)} = N$ and then use as an empirical estimate of the posterior distribution

$$\hat{p}^N(\mathrm{d}x_{0:t}|y_{0:t}) = \frac{1}{N}\sum_{i=1}^N N_t^{(i)} \delta_{X_{0:t}^{(i)}}(\mathrm{d}x_{0:t})$$

so that the new set of particles have weights equal to $1/N$ To obtain this we introduce the resampling step.

At each time step we resample N new particles with replacement from the particles $(X_{0:t}^{(i)}, i = 1, ..., N)$ with weights $w_t^{(i)} \propto p(y_t|x_t^{(i)})$ to obtain a new set of particles $(\hat{X}_{0:t}^{(i)})_{i=1}^N$ This will give us two sets of random samples, $(\hat{X}_{0:t-1}^{(i)}, X_t^{(i)})_{i=1}^N$ who's empirical distributions will approximate $p(x_{0:t}|y_{0:t-1})$ and a second set $(\hat{X}_{0:t}^{(i)})_{i=1}^N$ that will approximate $p(x_{0:t}|y_{0:t})$ (or their marginals).

In other words we simulate sequentially in time particles according to the law of the process $\{X_t\}$ conditional on the sequence $\{X_{0:t-1}, y_{0:t-1}\}$ and at each time t we introduce the resampling intermediate step to select the particles that fit our new observation y_t. This will give us both marginal and simultaneously solutions to the filtering problem. Details about convergence of the algorithm is discussed in Chapter 4. Algorithm 2.2 describes how to implement the SIR filter, note that if we are only interested in X_t we do not have to store the variables $\hat{X}_{0:t-1}^{(i)}$. In the rest of this chapter we will focus only on filtering X_t.

Algorithm 2.2: The SIR algorithm

```
Initialisation $t = 0$;
for $i = 1 : N$ do
    Sample $X_0^{(i)} \sim p(x_0)$;
end
for $t = 1 : T$ do
    for $i = 1 : N$ do
        Sample $X_t^{(i)} \sim p(x_t|X_{t-1}^{(i)})$ ;
        Set $X_{0:t}^{(i)} = (\hat{X}_{0:t-1}^{(i)}, X_t^{(i)})$ Evaluate the importance weights $w_t^{(i)} = p(y_t|X_t^{(i)})$ ;
    end
    Normalise the importance weights $\tilde{w}_0^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(j)}}$ ;
    Resample with replacement N particles $\{\hat{X}_t^{(i)}\}_{i=1}^N$ from the set $\{X_t^{(j)}\}_{j=1}^N$ with propabilities $\{\tilde{w}_t^{(j)}\}_{j=1}^N$
end
```

Example 2.1

In the first example we present, we want to estimate $\mathbb{E}X_t$ of a stationary AR(1) model

$$X_t = \theta X_{t-1} + V_t, \qquad V_t \sim \mathcal{N}(0, \sigma_v^2), \; |\theta| < 1$$

where the observed process $\{Y_t\}$ is given by the equation

$$Y_t = X_t + W_t, \qquad W_t \sim \mathcal{N}(0, \sigma_w^2).$$

In this example we let $\sigma_v^2 = 1$ and $\sigma_w^2 = 0.6$ and $\phi = 0.7$. Using R we have carried out a filter scheme according to algorithm 2.2 for $t = 1 : 50$. First using $N = 30$ and then $N = 1000$ particles. Since this model is linear and Gaussian we have also run the Kalman filter to compare with the SIR filter. Figure 2.1 and 2.3 show the filter update and Kalman update for the state process X compared with the value for the 'true' X process. while in figure 2.2 and 2.4 we have used kernel estimation in R to estimate the posterior densities for $t = 1 : 15$ in the SIR filter. As we see from figure 2.1 and 2.2 the the SIR filter performs well and is close to the optimal solution attained from the Kalman filter (in the limit they will be the same). Not surprisingly the density estimates are quite poor when we use only 30 particles, but when we increase the number of particles to 1000, the estimated densities looks more Gaussian, as they should be.

Figure 2.1: SIR filter with N=30 particles

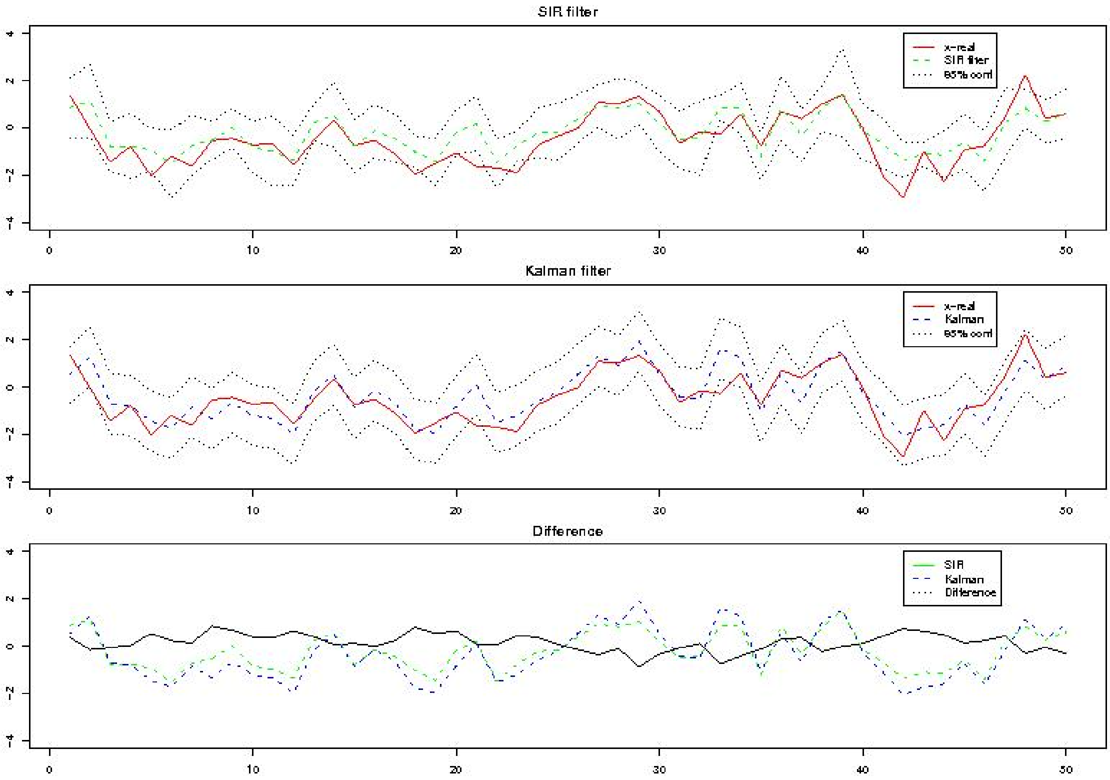


Figure 2.2: Density estimation for N=30

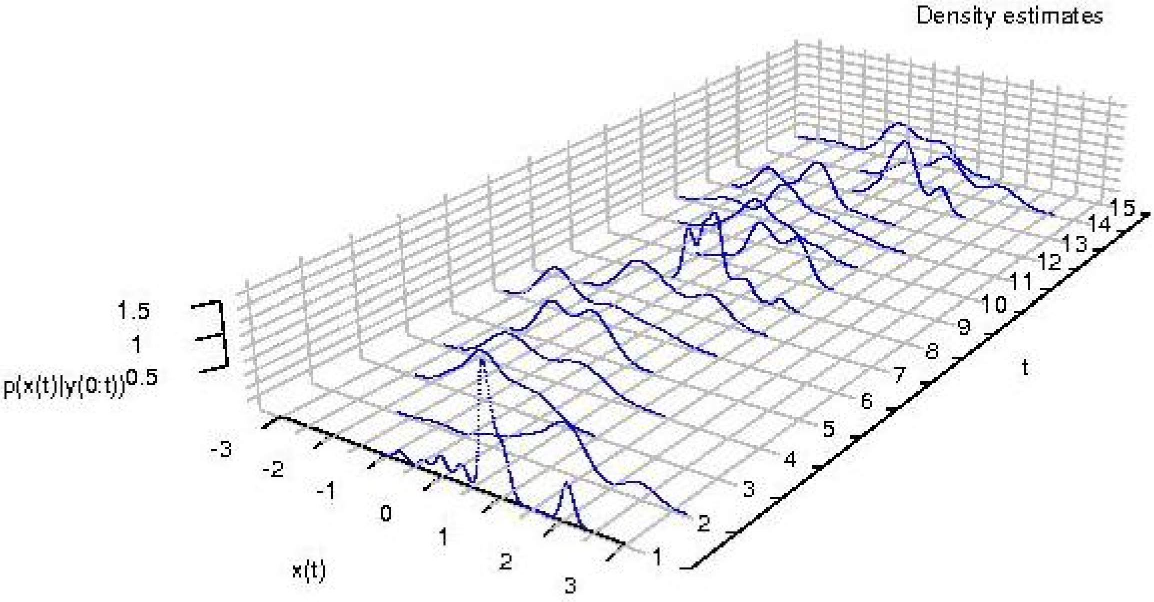


Figure 2.3: SIR filter with N=1000 particles

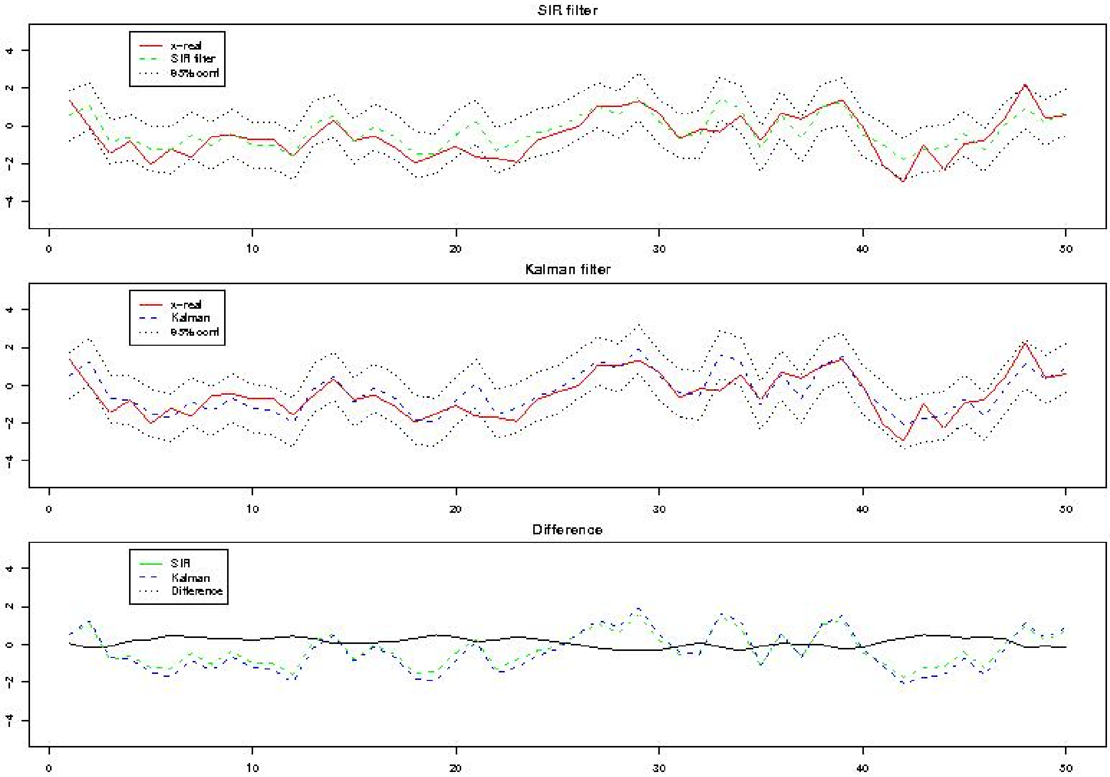
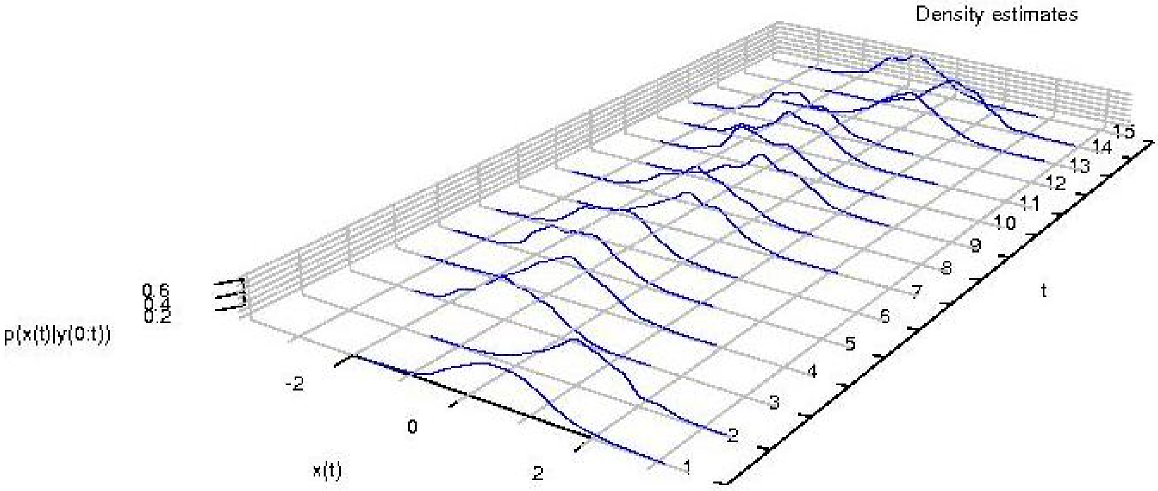


Figure 2.4: Density estimation with N=1000



2.2.1 Improving diversity

The SIR filter was introduced to avoid the degeneracy problem in SIS method with a simple resampling step. However, we may encounter a different problem if the importance weights are skewed before the resampling step. The particles with high importance weight are statistically selected many times. If only a few particles have importance weight that is significantly different from zero we may have a rapid loss of diversity in our particles. Another problem is that due to the mixture form of the approximation, we need outliers to well approximate the tail of the posterior density function, no matter how large we choose N. One way of dealing with these problems is to introduce a regularisation step.

The regularised Particle filter

In the filters described above, our resample comes from a discrete approximation of the posterior density. Our aim is to draw from a continuous approximation to avoid the degeneracy problem. The regularised particle filter (RPF) is a method based on regularisation of an empirical measure, so before we dive in to the RPF we need the following.

Regularisation of an empirical measure

Regularisation of an empirical measure is a method that approximates a discrete measure by a continuous one. Let ν be an empirical measure, $\nu = \sum_{i=1}^{N} \tilde{w}^{(i)} \delta_{X^{(i)}}(x)$ on $\mathbb{R}^d$, where $\tilde{w}^{(i)}$ are normalised weights, and let κ be a continuous function on $\mathbb{R}^d$. We say that κ a regularisation kernel if

- $\int \kappa(x)\,\mathrm{d}x = 1$
- $\int x\kappa(x)\,\mathrm{d}x = 0$
- $\int \|x\|\kappa(x)\,\mathrm{d}x < \infty$.

For any $x \in \mathbb{R}^d$ and any $h > 0$ we define the rescaled kernel

$$\kappa_h(x) = \frac{1}{h^d}\kappa\left(\frac{x}{h}\right).$$

Definition: For any empirical measure ν on $\mathbb{R}^d$, where d is the dimension of the X vector, the regularisation of ν is the absolutely continuous probability distribution $\kappa_h * \nu$ with probability distribution

$$\frac{d(\kappa_h * \nu)}{\mathrm{d}x}(x) = \int \kappa'_h(x-u)\nu(\mathrm{d}u),$$

where $*$ is the convolution operator. The raw filter estimate of $p(x_t|y_{0;t})$ given by

$$= \sum_{i=1}^{N} \tilde{w}_t^{(i)} \delta_{X_t^{(i)}}(x)$$

should be smoothed if the intention is to estimate the density function. The new estimate

$$\hat{p}_N(x_t|y_{0:t}) = \sum_{i=1}^{N} \tilde{w}_t^{(i)} \kappa_h(x - X_t^{(i)}) \tag{2.8}$$

is a weighted kernel estimate based on the observations $\{X_t^{(i)}\}$, where h is chosen to minimise the mean integrated square error between the posterior density and the corresponding regularised empirical representation in (2.8)

$$MISE(\hat{p}) = \mathbb{E}\left[\int [\hat{p}_N(x_t|y_{0:t}) - p(x_t|y_{0:t})]^2 \, \mathrm{d}x_t\right].$$

It is worth noting that the approximation becomes increasingly less appropriate as d, the dimension of the state, increases. It can be shown (Wasserman, 2006) that under the special case of an equally weighted sample, the optimal choice of the kernel is the Epanechnikov kernel

$$\kappa_{opt} = \begin{cases} \frac{d+2}{2c_d}(1 - \|x\|^2) & \text{if } \|x\| < 1 \\ 0 & \text{otherwise,} \end{cases} \tag{2.9}$$

where c_d is the volume of the unit hypersphere in $\mathbb{R}^d$. In the case where the underlying density is Gaussian with unit covariance matrix, the optimal choice if the bandwidth is (Wasserman, 2006)

$$h_{opt} = AN^{-\frac{1}{d+4}} \text{ with } A = \left[8d^{-1}(d+4)(2\sqrt{\pi})^d\right]^{\frac{1}{d+4}}. \tag{2.10}$$

The results of (2.9) and (2.10) are optimal only under some very special cases, however, these results can be used in more general cases to obtain a suboptimal filter. Generating particles from the Epanechnikov kernel (2.9) consists of generating $\sqrt{\beta}T$ where β follows a beta distribution with parameters $(d/2, 2)$ and T is uniformly distributed over the unit sphere in $\mathbb{R}^d$. This is computationally expensive so it is common to generate samples from a Gaussian kernel to reduce the cost. The optimal bandwitch in this case is (Wasserman, 2006)

$$h_{opt} = AN^{\frac{1}{d+4}} \text{ with } A = [4/(d+2)]^{\frac{1}{d+4}}.$$

The RPF differs from the SIR filter only in additional regularisation after the resampling step. We also compute the empirical covariance matrix $\mathbb{S}_t$ of the particles prior to the resampling, so that $\mathbb{S}_t$ is a function of both $\{X_t^{(i)}\}_{i=1}^N$ and $\{w_t^{(i)}\}_{i=1}^N$ The main step is then to move the resampled values by

$$X_t^{\star(i)} = X_t^{(i)} + h_{opt}\mathbb{D}_t\xi^{(i)}, \tag{2.11}$$

where $\mathbb{D}_t\mathbb{D}_t^T = \mathbb{S}_t$ (Cholesky decomposition) and $\xi^{(i)}$ follows the Epanechnikov/Gaussian kernel. This will lead to diversion in our particle, but we are no longer guaranteed that these will asymptotically approximate those from the posterior. Another way of improving the diversity is to perform an MCMC step. Under certain conditions the new particles will converge to the posterior distribution of interest. The MCMC step will be discussed in Chapter 5. The algorithm for the RPF differs from the SIR only by an additional step as described in algorithm 2.3

Algorithm 2.3: The RPF algorithm

```
Initialisation $t = 0$;
for $i = 1 : N$ do
    Sample $X_0^{(i)} \sim p(x_0)$;
end
for $t = 1 : T$ do
    for $i = 1 : N$ do
        Sample $X_t^{(i)} \sim p(x_t|X_{t-1}^{\star(i)})$ ;
        Calculate $w_t^{(i)} = p(y_t|X_t^{(i)})$ ;
    end
    Normalise the importance weights $\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(j)}}$ ;
    Compute the empirical covariance matrix $\mathbb{S}_t$ of $\{X_t^{(i)}, \tilde{w}_t^{(i)}\}_{i=1}^N$;
    Compute $\mathbb{D}_t$ such that $\mathbb{D}_t \mathbb{D}_t^T = \mathbb{S}_t$ ;
    Resample with replacement N particles $\{\hat{X}_t^{(i)}\}_{i=1}^N$ from the set $\{X_t^{(j)}\}_{j=1}^N$ with propabilities $\{\tilde{w}_t^{(j)}\}_{j=1}^N$;
    for $i = 1 : N$ do
        Sample $\xi^{(i)}$ from the Epanechnikov/Gaussian kernel;
        Set $X_t^{\star(i)} = \hat{X}_t^{(i)} + h_{opt}\mathbb{D}_t\xi^{(i)}$;
    end
end
```

2.3 The ASIR filter

Another way to avoid degeneracy is the Auxiliary SIR (ASIR) filter introduced by Pitt and Shephard (1999) as a variant of the standard SIR filter. The idea is to use the measurement available at time t to perform resampling at time t-1, before the particles propagate to time t.

The ASIR filter introduce an importance density $q(x_t, i|y_{0:t})$ which samples the pair $\{X_t^{(j)}, i^{(j)}\}_{j=1}^N$ where $i^{(j)}$ refers to the index of the particle at time $t-1$ (note that $p(i|y_{0:t-1}) = \tilde{w}_t^{(i)}$).

From *Bayes rule*

$$\begin{aligned} p(x_t, i|y_{0:t}) &\propto p(y_t|x_t)p(x_t, i|y_{0:t-1}) \\ &= p(y_t|x_t)p(x_t|i, y_{0:t-1})p(i|y_{0:t-1}) \\ &= p(y_t|x_t)p(x_t|x_{t-1}^{(i)})\tilde{w}_{t-1}^{(i)}. \end{aligned} \tag{2.12}$$

If we now obtain a sample from the joint density $p(x_t, i|y_{0:t})$ and omit the i in each of the pairs $(X_t^{(j)}, i^{(j)})$ we are left with a sample $\{X_t^{(j)}\}_{j=1}^N$ from the marginal distribution $p(x_t|y_{0:t})$. The importance density used to draw the sample $(X_t^{(j)}, i^{(j)})_{j=1}^N$ in the ASIR filter is defined to satisfy

$$q(x_t, i|y_{0:t}) \propto p(y_t|\mu_t^{(i)})p(x_t|x_{t-1}^{(i)})w_{t-1}^{(i)}, \tag{2.13}$$

where $\mu_t^{(i)}$ is a characteristic of X_t given $X_{t-1}^{(i)}$, for example the mean $\mathbb{E}[X_t|X_{t-1}^{(i)}]$ or a sample $\mu_t^{(i)} \sim p(x_t|x_{t-1}^{(i)})$. We may also write

$$q(x_t, i|y_{0:t}) = q(i|y_{0:t})q(x_t|i, y_{0:t}) \tag{2.14}$$

and defining

$$q(x_t|i, y_{0:t}) \triangleq p(x_t|x_{t-1}^{(i)}) \tag{2.15}$$

we have, according to (2.13), (2.14) and (2.15),

$$q(i|y_{0:t}) \propto p(y_t|\mu_t^{(i)})w_{t-1}^{(i)}.$$

The ASIR filter evolves by sampling the set $\{i^{(j)}\}_{j=1}^N$ from the set $\{i\}_{i=1}^N$ with probabilities $p(y_t|\mu_t^{(i)})$ and then drawing $X_t^{(j)}$ according to $q(x_t|i^{(j)}, y_{0:t}) = p(x_t|)X_{t-1}^{(i^j)}$ The weight of the sample $\{x_t^{(j)}, i^{(j)}\}_{j=1}^N$ is according to (2.5) proportional to the ratio of the right hand side of (2.12) and (2.13):

$$w_t^{(j)} \propto w_{t-1}^{(i^j)} \frac{p(y_t|X_t^{(j)})p(X_t^{(j)}|X_{t-1}^{(i^j)})}{q(X_t^{(j)}, i^{(j)}|y_{1:t})} = \frac{p(y_t|X_t^{(j)})}{p(y_t|\mu_t^{(i^j)})}.$$

Compared to the SIR, the ASIR filter has the advantage that it naturally generates points from the sample at time t-1, which conditioned on the current measurement y_t, are most likely to be in a region of high likelihood. The ASIR resamples at the 'previous' time step based on some point estimate $\mu_t^{(i)}$ that characterise $p(x_t|X_{t-1}^{(i)})$. This works really well if $p(x_t|X_{t-1}^{(i)})$. is well characterised $\mu_t^{(i)}$, that is if the process noise is small. If the process noise is large, however, the ASIR filter can in fact degrade the performance, but this may be corrected with a final resampling step, as in the SIR filter, to obtain a final equally weighted sample.

Algorithm 2.4: The ASIR algorithm

```
Initialisation $t = 0$;
for $i = 1 : N$ do
    Sample $X_0^{(i)} \sim p(x_0)$;
end
for $t = 1 : T$ do
    for $i = 1 : N$ do
        Calculate $\mu_t^{(i)}$;
        Calculate $w_t^{(i)} = p(y_t|\mu_t^{(i)})$;
    end
    Normalise the importance weights $\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}}$ ;
    Sample N integers $\{i^{(j)}\}_{j=1}^{N}$ with replacement from the set $\{i\}_{i=1}^{N}$ with probabilities $\tilde{w}_t^{(i)}$;
    for $j = 1 : N$ do
        Sample $X_t^{(j)} \sim p(x_t|x_{t-1}^{(i^j)})$;
        Calculate $w_t^{(j)} = \frac{p(y_t|x_t^{(j)})}{p(y_t|\mu_t^{(j)})}$;
    end
    Normalise the weights $\tilde{w}_t^{(j)} = \frac{w_t^{(j)}}{\sum_{i=1}^{N} w_t^{(i)}}$;
    (Optional)
    Sample $\{\hat{X}_t^{(j)}\}_{j=1}^{N}$ with replacement from the set $\{X_t^{(j)}\}_{j=1}^{N}$ with probabilities $\{w_t^{(j)}\}_{j=1}^{N}$
end
```

2.4 Multiple model particle filters

The MM particle filters are sequential Monte Carlo approximation of the conceptual solution given by 1.7 and 1.8 to solve the problem given by

$$x_t = k_{t-1}(X_{t-1}, r_t, V_{t-1})$$
$$y_t = h_t(X_t, r_t, W_t),$$

where r_t is assumed to be discrete with Markov transitions $p_{ij} = P\{r_t = j|r_{t-1} = i\}\ i, j = 1, ..., s$. Let us then define the augmented state-vector $Z_t = [X_t^T, r_t]^T$. We assume that the initial densities $p(x_0)$ and $p(r_1) = \sum_{i=1}^{s} \mu_i \delta(r_1 - i)$ are known. We denote by $\{Z_t^{(j)}, w_t^{(j)}\}_{j=1}^N$ a random measure that characterises the posterior density $p(z_t|y_{0:t})$ such that each particle $Z_t^{(j)}$ consists of the two components, $X_t^{(j)}$ and $r_t^{(j)}$. The first step of the MMPF is to generate a random set $\{r_t^{(j)}\}_{j=1}^N$ based on the set $\{r_{t-1}^{(j)}\}_{j=1}^N$ and the transition probability matrix $\mathbb{P} = [p_{ij}], (i, j) \in S$. The next step of the MMPF is to perform a regime conditioned SIR filter described below. The optimal regime conditional density (Ristic, Arulampalam and Gordon (2004)) is given by

$$q(x_t|X_{t-1}^{(i)}, r_t^{(i)}, y_t)_{\text{opt}} = p(x_t|X_{t-1}^{(i)}, r_t^{(i)}, y_t),$$

but the most popular choice appears to be the transition prior

$$q(x_t|X_{t-1}^{(i)}, r_{t-1}^{(i)}, y_t) = p(x_t|X_{t-1}^{(i)}, r_t^{(i)}).$$

Algorithm 2.5: Regime conditioned SIR algorithm

```
Initialisation $t = 0$;
for $i = 1 : N$ do
    Sample $X_0^{(i)} \sim p(x_0)$;
end
for $t = 1 : T$ do
    for $i = 1 : N$ do
        Sample $r_t^{(i)}$ according to $\mathbb{P}$ ($p(r_1)$ for $t = 1$);
        Sample $X_t^{(i)} \sim p(x_t|X_{t-1}^{(i)}, r_t^{(i)})$;
        Set $Z_t^{(i)} = (X_t^{(i)}, r_t^{(i)})$;
        Calculate $w_t^{(i)} = p(y_t|X_t^{(i)}, r_t^{(i)})$;
    end
    Normalise the importance weights $\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(j)}}$ ;
    Sample $\{\hat{Z}_t^{(i)}\}_{i=1}^N$ with replacement from the set $\{Z_t^{(i)}\}_{i=1}^N$ with probabilities $\tilde{w}_t^{(i)}$;
end
```

Example 2.2

In the second example we study a regime model, the signal process $\{X_t\}$ evolves according to

$$X_t = r_t X_{t-1} + V_t, \qquad V_t \sim \mathcal{N}(0,1)$$

and the observation process $\{Y_t\}$ is given by the equation

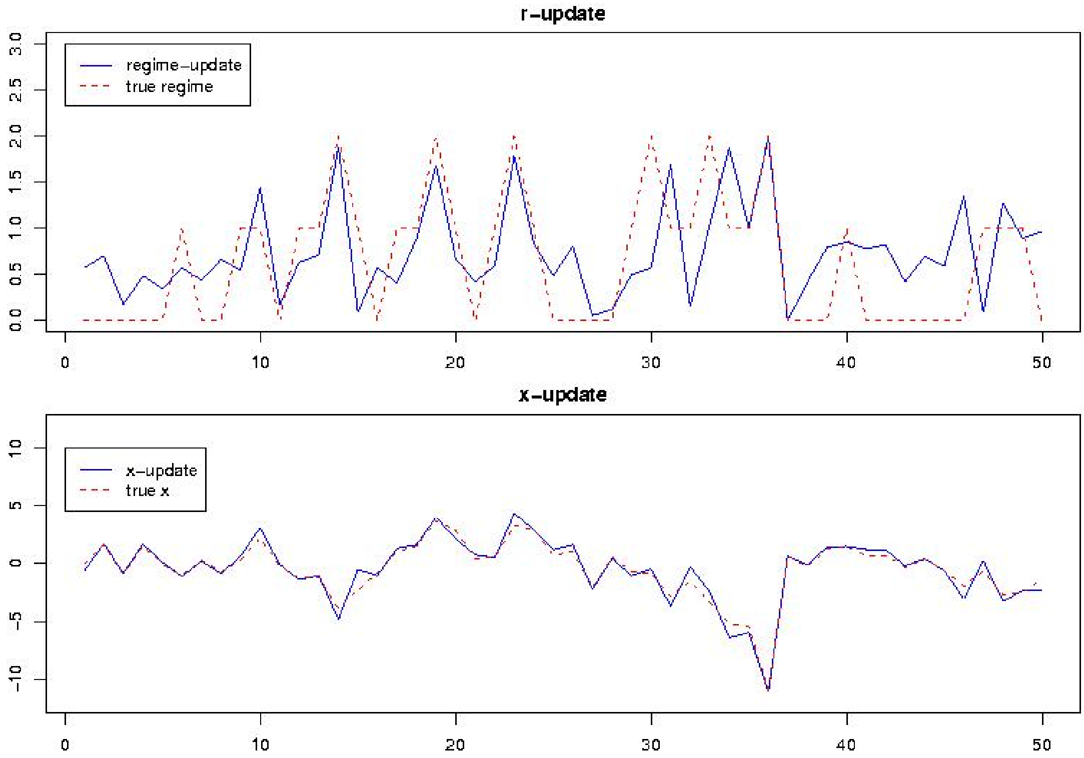
$$Y_t = X_t + W_t, \qquad W_t \sim \mathcal{N}(0,1)$$

where $\{r_t\}$ is a discrete Markov chain with states $s = 0,1,2$, transition probability matrix $\mathbb{P}$

$$\mathbb{P} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}$$

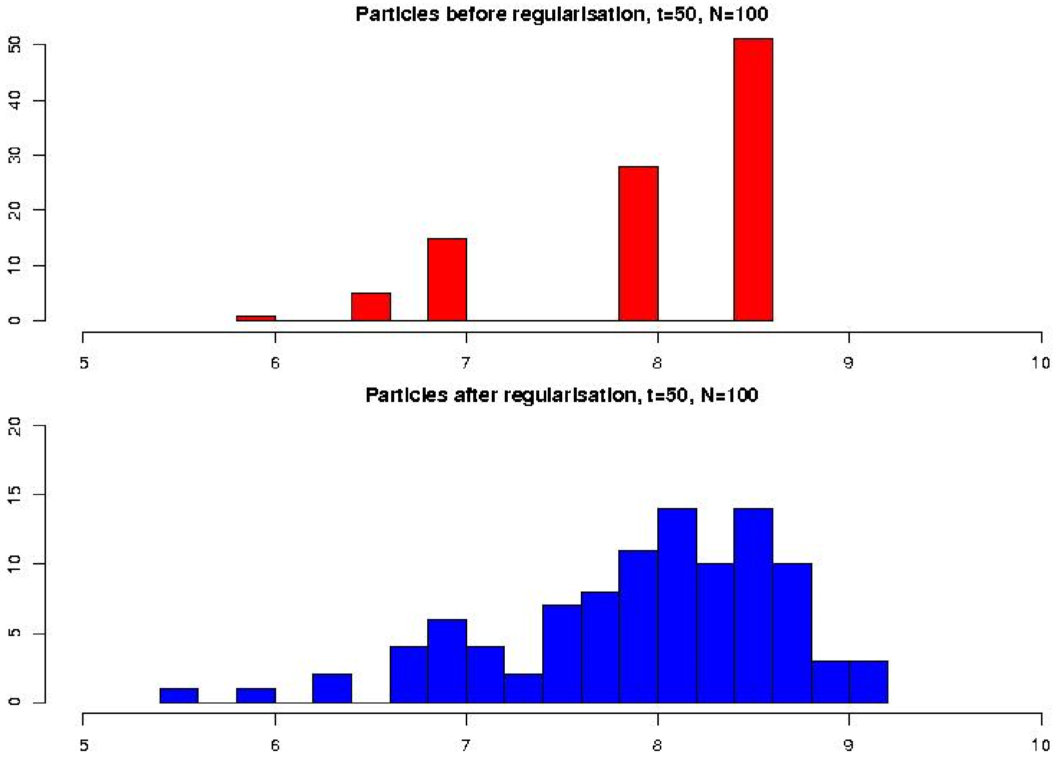
and with initial probabilities $p_0 = (1/3, 1/3, 1/3)$. We have then carried out a particle filter according to algorithm 2.5 with $T = 50$ and $N = 1000$. The results are shown in figure 2.5

Figure 2.5: Multiple model particle filter



We will also include a regularisation step (as described in algorithm 2.3) at time $t = 50$. However, in this case we will re-run the algorithm with only 100 particles to see the effect. Figure 2.6 shows the histograms of $\hat{X}[50]$ from the regime filter and the regularised particles from equation 2.11 where $\xi^{(i)}$ is drawn from a Gaussian kernel and the bandwith h is chosen as in equation 2.10

Figure 2.6: Histograms of the particles



2.5 Combined parameter and state estimation

Throughout this chapter we have studied simulation-based methods for filtering time-varying state vectors, however, in many situations we need more general algorithms that deal simultaneously with both fixed model parameters and state variables. West (1993a) presents an algorithm that solves this problem.

2.5.1 Framework

Consider again a Markovian dynamic model for sequentially observed data vectors Y_t, $(t = 0, 1, \dots)$ (again with $y_0 = 0$) in which the state vector at time t is X_t, and the fixed parameter vector is θ. As we have seen before, the state dynamics evolves according to the density $p(x_t|x_{t-1}, \theta)$ and the observation according to $p(y_t|x_t, \theta)$, where Y_t is conditionally independent of the past states and observations given X_t and θ, and X_t is conditionally independent of past states and observations given X_{t-1} and θ. The aim is to use Monte Carlo methods to sequentially update Monte Carlo sample approximations of sequences of posterior distributions $p(x_t, \theta|y_{0:t})$, where $y_{0:t} = (y_0, \dots, y_t)$ is the available information at time t. The case where θ is known, or when there is no fixed parameters in the model, has already been discussed in this chapter. In this section we will use the ASIR filter (algorithm 2.4) developed by Pitt and Shephard (1999)

2.5.2 Filtering for states and parameters

In the general model with fixed parameters, we extend the sample-based framework developed for state filtering to both state and parameter. At time t, we then have a sample $\{X_t^{(j)}, \theta_t^{(j)}\}_{j=1}^N$ with associated weights $\{\tilde{w}_t^{(j)}\}_{j=1}^N$ representing an importance sample approximation to the time t posterior $p(x_t, \theta|y_{0:t})$ for both parameters and state. Note that the index t on the parameter indicates that it comes from time t posterior, not that it is time-varying. As time evolves to $t+1$, y_{t+1} becomes available, and we want to generate a sample from $p(x_{t+1}, \theta|y_{0:t+1})$. *Bayes theorem* gives us that

$$\begin{aligned} p(x_{t+1}, \theta|y_{0:t+1}) &\propto p(x_{t+1}, \theta, y_{0:t+1}) \\ &\propto p(y_{t+1}|x_{t+1}, \theta, y_{0:t})p(x_{t+1}|\theta, y_{0:t})p(\theta|y_{0:t}) \\ &= p(y_{t+1}|x_{t+1}, \theta)p(x_{t+1}|\theta, y_{0:t})p(\theta|y_{0:t}). \end{aligned} \tag{2.16}$$

As we see from equation (2.16), the density $p(\theta|y_{0:t})$ is an important ingredient in the update. There are several historical approaches to address this problem, we will review two of them.

Artificial evolution of parameters

In dealing with time-varying states, one approach to reducing degeneracy in the sample, as we have seen, is to add small noise disturbance to state particles between time steps (Gordon, 1993). This idea has later been extrapolated to the fixed model parameters. One version of these methods has the interpretation of an extended model in which the model parameters are viewed as if they were in fact time-varying, -an 'artificial evolution'. In other words, we consider a different model where θ is replaced by θ_t at time t, and simply include θ_t in the augmented state vector. Then we add an independent zero mean Gaussian increment to the parameter at each time t i.e

$$\theta_{t+1} = \theta_t + \xi_{t+1}$$
$$\xi_{t+1} \sim \mathcal{N}(0, \mathbb{W}_{t+1})$$

for some specified covariance matrix $\mathbb{W}_{t+1}$ and where θ_t and ξ_{t+1} are conditionally independent given $Y_{0:t}$. With the model recast we can now carry out filtering methods such as the ASIR filter. However, as stated in the beginning, the fixed model parameters are fixed. Pretending that they are time-varying implies an artificial loss of information between time points.

An inherent interpretation in terms of kernel smoothing of particles leads to a modification of this artificial evolution method in which the problem of information loss is avoided. We first discuss the basic form of the kernel smoothing.

Kernel smoothing for parameters

To approximate the required density $p(\theta|y_{0:t})$ in (2.16), West (1993b) developed kernel smoothing methods that provided basis for rather effective adaptive importance sampling techniques.

At time t, suppose that we have current posterior parameter samples $\theta_t^{(j)}$ and weights $\tilde{w}_t(j)$, $(j = 1, \ldots, N)$ providing a discrete Monte Carlo approximation of $p(\theta|y_{0:t})$. Define $\bar{\theta}_t$ and $\mathbb{S}_t$ as the Monte Carlo posterior mean and covariance matrix of $p(\theta|y_{0:t})$, computed from the sample $\theta_t^{(j)}$ with weights $\tilde{w}_t^{(j)}$, that is

$$\bar{\theta}_t = \sum_{i=1}^{N} \tilde{w}_t^{(i)} \theta_t^{(i)}$$
$$\mathbb{S}_t = \sum_{i=1}^{N} \tilde{w}_t^{(i)} (\theta_t^{(i)} - \bar{\theta}_t)(\theta_t^{(i)} - \bar{\theta}_t)^T.$$

The smooth kernel density is then given by

$$p(\theta|y_{0:t}) \approx \sum_{i=1}^{N} \tilde{w}_t^{(i)} \mathcal{N}(\theta|m_t^{(j)}, h^2 \mathbb{S}_t), \tag{2.17}$$

where we define the following components: $\mathcal{N}(\cdot|m, \mathbb{S})$ is the multivariate normal density with mean m and covariance matrix $\mathbb{S}$, h is chosen as a slowly decreasing function of N such that the kernel components become more and more concentrated about their location $m_t^{(j)}$ as N increases. The kernel locations $m_t^{(j)}$ are specified using a shrinkage rule introduced by West (1993a), West (1993b). Standard kernel methods would suggest $m_t^{(j)} = \theta_t^{(j)}$ so that the kernels are located about existing sample values. Assume now that we choose to approximate $p(\theta|y_{0:t})$ by

$\sum_{i=1}^{N} \tilde{w}_t^{(i)} \mathcal{N}(\theta|\theta_t^{(j)}, h^2\mathbb{S}_t)$, then, if we denote $f_i = \mathcal{N}(\theta|\theta_t^{(i)}, h^2\mathbb{S}_t)$ we see that

$$\begin{aligned}
\mathbb{E}[\theta|y_{0:t}] &= \int \theta \sum_{i=1}^{N} \tilde{w}_t^{(i)} f_i d\theta \\
&= \sum_{i=1}^{N} \tilde{w}_t^{(i)} \int \theta f_i d\theta = \sum_{i=1}^{N} \tilde{w}_t^{(i)} \theta_t^{(i)} = \bar{\theta}_t.
\end{aligned}$$

However,

$$\begin{aligned}
\mathrm{Var}[\theta|y_{0:t}] &= \int (\theta - \bar{\theta}_t)(\theta - \bar{\theta}_t)^T \sum_{i=1}^{N} \tilde{w}_t^{(i)} f_i d\theta \\
&= \sum_{i=1}^{N} \tilde{w}_t^{(i)} \int (\theta\theta^T - 2\theta\bar{\theta}_t^T + \bar{\theta}_t\bar{\theta}_t^T) f_i d\theta \\
&= \sum_{i=1}^{N} \tilde{w}_t^{(i)} \left(\mathbb{E}_i[\theta\theta^T] - 2\mathbb{E}_i[\theta]\bar{\theta}_t^T + \bar{\theta}_t\bar{\theta}_t^T \right) \\
&= \sum_{i=1}^{N} \tilde{w}_t^{(i)} \left(h^2\mathbb{S}_t + \theta_t^{(i)}\theta_t^{(i)T} - 2\theta_t^{(i)}\bar{\theta}_t^T + \bar{\theta}_t\bar{\theta}_t^T \right) \\
&= h^2\mathbb{S}_t + \sum_{i=1}^{N} \tilde{w}_t^{(i)} (\theta_t^{(i)} - \bar{\theta})(\theta_t^{(i)} - \bar{\theta}_t)^T \\
&= (1 + h^2)\mathbb{S}_t > \mathbb{S}_t,
\end{aligned}$$

and we see that the resulting mixtures of Normal densities leads to an over-dispersed approximation of $p(\theta|y_{0:t})$. If we instead take $m_t^{(i)} = a\theta_t^{(i)} + (1-a)\bar{\theta}_t$, where $a = \sqrt{1+h^2}$, the same calculations as above shows that the resulting normal mixtures retains the mean $\bar{\theta}_t$ but the variance is now trivially corrected to $\mathbb{S}_t$.

A general algorithm for state and parameter estimation

If we now return to the filter problem in (2.16) we have available the Monte Carlo sample $(X_t^{(j)}, \theta_t^{(j)})$ with corresponding weights $\tilde{w}_t^{(j)}$, $(j = i, \dots, N)$ representing the discrete approximation of the posterior $p(x_t, \theta|y_{0:t})$. we use the kernel from equation (2.17) as the marginal density for the parameter. We can now apply an extended version of the auxiliary particle filter algorithm, incorporating the parameter with the state.

Also we may add a final resampling step to obtain an unweighted sample, this is smart if the observation noise is quite large.

Algorithm 2.6: Combined parameter and state estimation

```
Initialisation $t = 0$;
for $i = 1 : N$ do
    Sample $\theta_0^{(i)} \sim p(\theta)$;
    Sample $X_0^{(i)} \sim p(x_0)$;
end
for $t = 0 : T - 1$ do
    for $j = 1 : N$ do
        evaluate the prior point estimates of $(X_t^{(j)}, \theta_t^{(j)})$ given by $(\mu_{t+1}^{(j)}, m_t^{(j)})$ where

                $\mu_{t+1}^{(j)} = \mathbb{E}[X_{t+1} | X_t^{(j)}, \theta_t^{(j)}]$

        may be computed from the state evolution density and $m_t^{(j)} = a\theta_t^{(j)} + (1-a)\bar{\theta}_t$ is the $j^{\text{th}}$
        kernel location from equation (2.17);
        Calculate $g_{t+1}^{(j)} = w_t^{(j)} p(y_{t+1} | \mu_{t+1}^{(j)}, m_t^{(j)})$;
    end
    Normalise the importance weights $\tilde{g}_t^{(j)} = \frac{g_t^{(j)}}{\sum_{i=1}^{N} g_t^{(j)}}$ ;
    Sample N integers $\{i^j\}_{j=1}^{N}$ with replacement from the set $\{i\}_{i=1}^{N}$ with probabilities $\tilde{w}_t^{(i)}$;
    for $j = 1 : N$ do
        Sample $\theta_{t+1}^{(i^j)}$ from kernel component number $i^j$,

                $\theta_{t+1}^{(j)} \sim \mathcal{N}(\cdot | m_t^{(i^j)}, h^2 \mathbb{S})$;

        Sample $X_{t+1}^{(j)} \sim p(x_{t+1} | X_t^{(i^j)}, \theta_{t+1}^{j})$;
        Calculate $w_t^{(j)} \propto \frac{p(y_{t+1} | X_{t+1}^{(j)}, \theta_{t+1}^{(j)})}{p(y_{t+1} | \mu_{t+1}^{(i^j)}, m_t^{(i^j)})}$;
    end
    Normalise the weights $\tilde{w}_t^{(j)} = \frac{w_t^{(j)}}{\sum_{i=1}^{N} w_t^{(i)}}$;
end
```

Example 2.3

In example 3 we present the following scenario

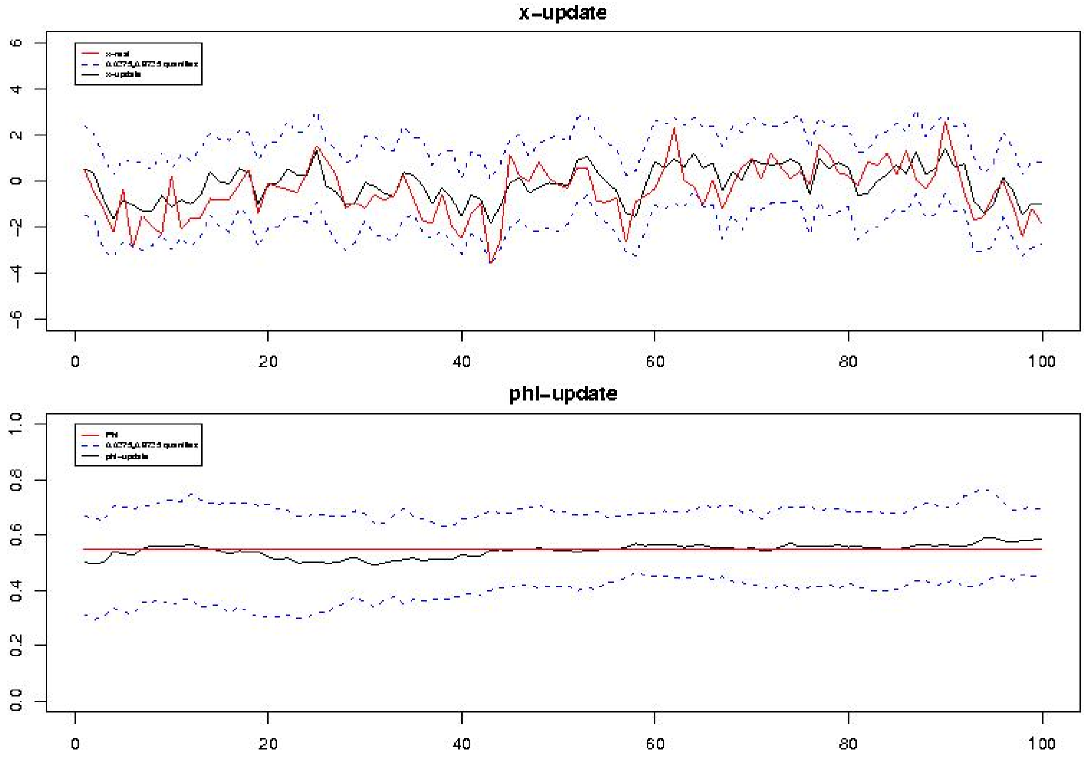
$$X_t = \sin(X_{t-1}) + V_t, \qquad V_t \sim \mathcal{N}(0,1),;$$

and the observation process $\{Y_t\}$ is given by the equation

$$Y_t = \phi X_t + W_t, \qquad W_t \sim \mathcal{N}(0,1),$$

where $\phi \sim \mathcal{N}(0.5, 0.1^2)$ is unknown and needs to estimated along with the X process. Figure 2.7 shows the results with $N = 500$.

Figure 2.7: Combined parameter and state, phi unknown



Next we assume that the variance of W_t is an unknown parameter σ^2 with initial condition $\mathcal{N}(2, 0.3^2)$. ϕ is still unknown. From figure 2.8 we see that uncertainty in ϕ increases over time. In figure 2.9 we have increased the number of particles from 500 to 5000 and we see that the uncertainty in ϕ has decreased.

Figure 2.8: Combined parameter and state, phi and sigma unknown, $N = 500$

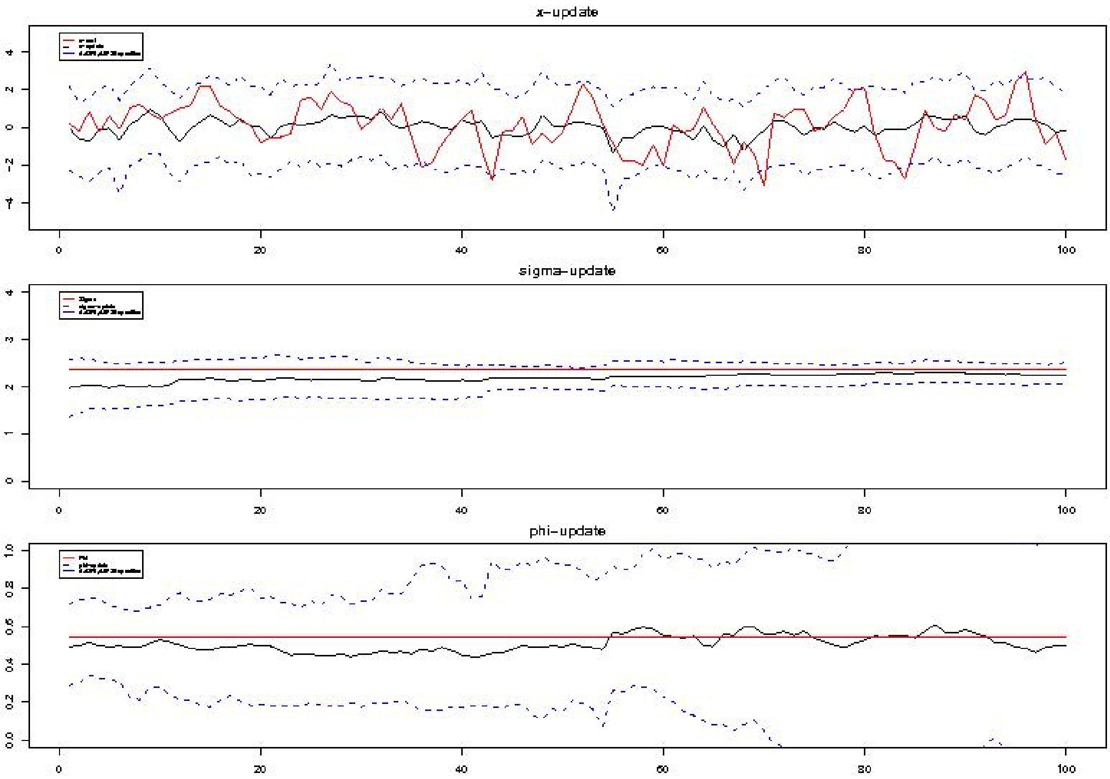
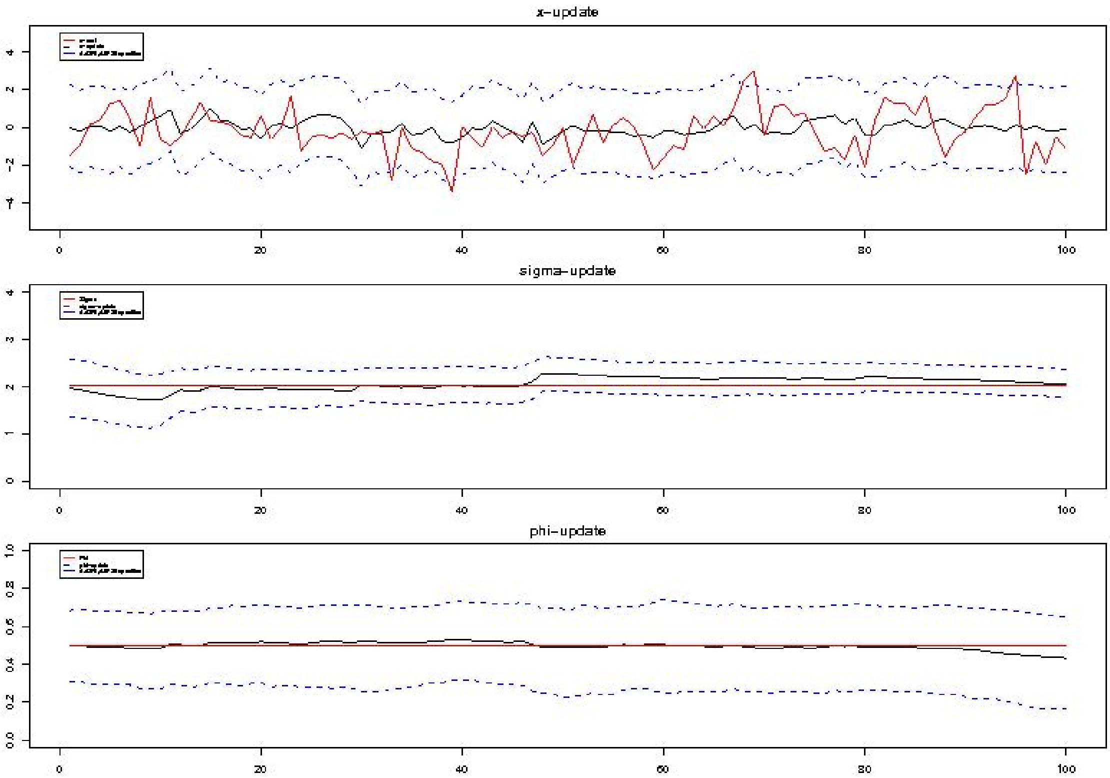


Figure 2.9: Combined parameter and state, phi and sigma unknown, $N = 5000$



3
Posterior Cramèr-Rao bounds

If we are interested in solving the filter problem $\mathbb{E}[x_t|y_{0:t}]$ we will also be interested in computing the information matrix. In this chapter we will derive at a formula to compute the information matrix for the particle filter sequentially, that is we, compute the prediction error recursively. This method is presented by Tichavsky, Muravchik and Nehorai (1998) and also discussed in Ristic, Arulampalam and Gordon (2004). The prediction error matrix is the right lower block of the information matrix for the whole trace $(\hat{X}^i_{0:t})_{i=1:N}$.

3.1 General case

Let X represent a vector of measured data and let Θ be an r-dimensional estimated random parameter. Denote by $p_{X,\Theta}(x,\theta)$ the joint probability density of the pair (X,Θ) and let $g(X)$ be the function of the measurements X that estimates Θ. The Posterior Cramèr-Rao bounds (PCRB) is

$$\mathbb{P} \triangleq \mathbb{E}\left[(g(X)-\Theta)\,(g(X)-\Theta)^T\right] \geq \mathbb{J}^{-1}, \tag{3.1}$$

where $\mathbb{J}$ is the $r \times r$ Fisher information matrix with elements

$$\mathbb{J}_{ij} = \mathbb{E}\left[-\frac{\partial^2 \log p_{X,\Theta}(x,\theta)}{\partial\theta_i \partial\theta_j}\right] \; i,j = 1,...,r$$

(assuming that the expectations and derivatives exists). The inequality in (3.1) means that the matrix $\mathbb{P} - \mathbb{J}^{-1}$ is a positive semidefinite matrix, saying that there exist at least one x such that $x^T(\mathbb{P} - \mathbb{J}^{-1})x = 0$.

Let ∇ and Δ be operators of the first and second order partial derivatives.

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, ..., \frac{\partial}{\partial\theta_r}\right]^T$$

$$\Delta_\Psi^\theta = \nabla_\Psi \nabla_\theta^T.$$

With this notation we can write $\mathbb{J}$ as

$$\mathbb{J} = \mathbb{E}\left[-\Delta_\theta^\theta \log p_{X,\Theta}(x,\theta)\right]. \, (r \times r)$$

By re-writing $p_{X,\Theta}(x,\theta)$ as $p_{X|\Theta}(x|\theta)p_\Theta(\theta)$, we can decompose $\mathbb{J}$ as $\mathbb{J}_D + \mathbb{J}_P$ where

$$\mathbb{J}_D = \mathbb{E}\left[-\Delta_\theta^\theta \log p_{X|\Theta}(x,\theta)\right] \; (r \times r)$$

and

$$\mathbb{J}_P = \mathbb{E}\left[-\Delta_\theta^\theta \log p_\Theta(\theta)\right]. \, (r \times r)$$

The interpretation is that we have decomposed the information into two blocks, the data information from $\mathbb{J}_D$ and the a priori information from $\mathbb{J}_P$.

On the other hand we also have $p_{X,\Theta}(x,\theta) = p_{\Theta|X}(\theta|x)p_X(x)$. Since $p_X(x)$ is an integral of $p_{X,\Theta}(x,\theta)$ over θ, it does not depend on θ so

$$\mathbb{J} = \mathbb{E}\left[-\Delta_\theta^\theta \log p_{\Theta|X}(\theta|x)\right].$$

For example, in the linear Gaussian case, when the posterior distribution of Θ conditioned on the data vector X is Gaussian with mean $\bar{\theta}_x$ and a covariance matrix $\mathbf{\Sigma}_x$ then the information matrix is given by

$$\mathbb{J} = \mathbb{E}\mathbf{\Sigma}_x^{-1}.$$

If $g(X) = \mathbb{E}\left[\Theta|X\right]$ is used to estimate Θ we have equality in (3.1). This is exactly the case of the Kalman filter.

Let us now assume that Θ is decomposed into two parts, $\Theta = [\Theta_\alpha^T, \Theta_\beta^T]^T$ with the corresponding decomposition of the information matrix $\mathbb{J}$

$$\begin{bmatrix} \mathbb{J}_{\alpha\alpha} & \mathbb{J}_{\alpha\beta} \\ \mathbb{J}_{\beta\alpha} & \mathbb{J}_{\beta\beta} \end{bmatrix}.$$

In this case the covariance, $\mathbb{P}_\beta$, of the estimation of Θ_β is bounded by the right lower block of $\mathbb{J}^{-1}$. To derive at the expression for this matrix we need to solve

$$\begin{bmatrix} \mathbb{J}_{\alpha\alpha} & \mathbb{J}_{\alpha\beta} \\ \mathbb{J}_{\beta\alpha} & \mathbb{J}_{\beta\beta} \end{bmatrix} \begin{bmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{B}^T & \mathbb{C} \end{bmatrix} = \begin{bmatrix} \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbb{I} \end{bmatrix}$$

w.r.t. $\mathbb{C}$. The solution is well known from general matrix theory

$$\mathbb{C} = [\mathbb{J}_{\beta\beta} - \mathbb{J}_{\beta\alpha}\mathbb{J}_{\alpha\alpha}^{-1}\mathbb{J}_{\alpha\beta}]^{-1}. \tag{3.2}$$

Analog to (3.1) we now have an expression for the lower bound of the covariance of the estimation of Θ_β

$$\mathbb{P}_\beta \triangleq \mathbb{E}\left[\left(g(X) - \Theta_\beta\right)\left(g(X) - \Theta_\beta\right)^T\right] \geq \left(\mathbb{J} - \mathbb{J}_{\beta\alpha}\mathbb{J}_{\alpha\alpha}^{-1}\mathbb{J}_{\alpha\beta}\right)^{-1}.$$

3.2 PCRB for the nonlinear filter problem

Let us now consider the nonlinear filtering problem

$$X_t = k_{t-1}(X_{t-1}, V_{t-1}) \tag{3.3}$$

$$Y_t = h_t(X_t, W_t), \tag{3.4}$$

where V_t and W_t are independent white processes. Let d be the dimension of the state vector X_t We also assume that X_0 has a known probability density function $p(x_0)$. (3.3) and (3.4) together with $p(x_0)$ will determine the joint probability density of $X_{0:t}$ and $Y_{0:t}$

$$p(x_{0:t}, y_{0:t}) = p(x_0)\prod_{j=1}^{t} p(y_t|x_t)\prod_{k=1}^{t} p(x_t|x_{t-1}) \tag{3.5}$$

The information of $X_{0:t}$, $\mathbb{J}(X_{0:t})$, is the $(td \times td)$ matrix derived from the joint probability density function (3.5) However we are interested in the information submatrix for estimating X_t, denoted $\mathbb{J}_t$, which is given as the $(d \times d)$ inverse of the right-lower block of $\mathbb{J}^{-1}$. The matrix $\mathbb{J}_t^{-1}$ will give us a lower bound for the mean square error of estimating X_t.

In the following we will denote $p(x_{0:t}, y_{0:t})$ as p_t for brevity.

If we decompose $X_{0:t}$ as $X_{0:t} = (X_{0:t-1}, X_t)$ with the corresponding decomposition of $\mathbb{J}(X_{0:t})$

$$\mathbb{J}(X_{0:t}) = \begin{bmatrix} \mathbb{A}_t & \mathbb{B}_t \\ \mathbb{B}_t^T & \mathbb{C}_t \end{bmatrix} \triangleq \begin{bmatrix} \mathbb{E}\left[-\Delta_{x_{0:t-1}}^{x_{0:t-1}} \log p_t\right] & \mathbb{E}\left[-\Delta_{x_{0:t-1}}^{x_t} \log p_t\right] \\ \mathbb{E}\left[-\Delta_{x_t}^{x_{0:t-1}} \log p_t\right] & \mathbb{E}\left[-\Delta_{x_t}^{x_t} \log p_t\right] \end{bmatrix}.$$

From (3.2) we have

$$\mathbb{J}_t = \mathbb{C}_t - \mathbb{B}_t^T\mathbb{A}_t^{-1}\mathbb{B}_t. \tag{3.6}$$

If we want to compute $\mathbb{J}_t$ at each time step t we would have to compute the inverse of the $(t-1)r \times (t-1)r$ matrix $\mathbb{A}_t$ (or $\mathbb{J}(X_{0:t})$). We now present the main result of this chapter, which will allow us to evaluate the information matrix sequentially in time.

Proposition 3.1

In the filter problem described by (3.3), the information submatrix $\mathbb{J}_t$ for estimating the state vector X_t satisfy the following recursions

$$\mathbb{J}_{t+1} = \mathbb{D}_t^{22} - \mathbb{D}_t^{21}(\mathbb{J}_t + \mathbb{D}_t^{11})^{-1}\mathbb{D}_t^{12},$$

where

$$\mathbb{D}_t^{ij} = \begin{cases} \mathbb{E}[-\Delta_{x_{t+j-1}}^{x_{t+i-1}} \log p(x_{t+1}|x_t)] & \text{if} \quad 2 \leq i+j < 4 \\ \mathbb{E}[-\Delta_{x_{t+j-1}}^{x_{t+i-1}} \log p(x_{t+1}|x_t)] + \mathbb{E}[-\Delta_{x_{t+1}}^{x_{t+1}} \log p(y_{t+1}|x_{t+1})] & \text{if} \quad i+j = 4. \end{cases}$$

Proof: The key is to re-write the joint probability density function of $(X_{0:t+1}, Y_{0:t+1})$ as

$$\begin{aligned} p_{t+1} &\triangleq p(x_{0:t+1}, y_{0:t+1}) \\ &= p(y_{t+1}|x_{t+1}, x_{0:t}, y_{0:t})p(x_{t+1}|x_{0:t}, y_{0:t})p(x_{0:t}, y_{0:t}) \\ &= p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t)p_t. \end{aligned}$$

Now if we decompose $X_{0:t+1}$ into $X_{0:t+1} = (X_{0:t-1}, X_t, X_{t+1})$, $\mathbb{J}(X_{0:t+1})$ can be decomposed as

$$\mathbb{J}(X_{0:t+1}) = \begin{bmatrix} \mathbb{A}_{t+1} & \mathbb{B}_{t+1} & \mathbb{L}_{t+1} \\ \mathbb{B}_{t+1}^T & \mathbb{C}_{t+1} & \mathbb{G}_{t+1} \\ \mathbb{L}_{t+1}^T & \mathbb{G}_{t+1}^T & \mathbb{F}_{t+1} \end{bmatrix}.$$

Let us analyse each submatrix.

$$\begin{aligned} \mathbb{A}_{t+1} &= -\mathbb{E}\left[\Delta_{x_{0:t-1}}^{x_{0:t-1}} \log p_{t+1}\right] \\ &= -\mathbb{E}\left[\Delta_{x_{0:t-1}}^{x_{0:t-1}}(\log p_t + \log p(x_{t+1}|x_t) + \log p(y_{t+1}|x_{t+1}))\right] \\ &= -\mathbb{E}\left[\Delta_{x_{0:t-1}}^{x_{0:t-1}} \log p_t\right] + \mathbf{0} + \mathbf{0} \\ &= \mathbb{A}_t. \end{aligned}$$

$$\begin{aligned} \mathbb{B}_{t+1} &= -\mathbb{E}\left[\Delta_{x_{0:t-1}}^{x_t}(\log p_t + \log p(x_{t+1}|x_t) + \log p(y_{t+1}|x_{t+1}))\right] \\ &= -\mathbb{E}\left[\Delta_{x_{0:t-1}}^{x_t} \log p_t\right] + \mathbf{0} + \mathbf{0} \\ &= \mathbb{B}_t. \end{aligned}$$

This implies that $\mathbb{B}_{t+1}^T = \mathbb{B}_t^T$.

$$\begin{aligned}\mathbb{C}_{t+1} &= -\mathbb{E}\left[\Delta_{x_t}^{x_t}(\log p_t + \log p(x_{t+1}|x_t) + \log p(y_{t+1}|x_{t+1}))\right] \\ &= -\mathbb{E}\left[\Delta_{x_t}^{x_t}\log p_t\right] - \mathbb{E}\left[\Delta_{x_t}^{x_t}\log p(x_{t+1}|x_t)\right] + \mathbf{0} \\ &= \mathbb{C}_t + \mathbb{D}_t^{11}.\end{aligned}$$

$$\begin{aligned}\mathbb{L}_{t+1} &= -\mathbb{E}\left[\Delta_{x_{0:t-1}}^{x_{t+1}}(\log p_t + \log p(x_{t+1}|x_t) + \log p(y_{t+1}|x_{t+1}))\right] \\ &= \mathbf{0}.\end{aligned}$$

Again this implies that $\mathbb{L}_{t+1}^T = \mathbf{0}$

$$\begin{aligned}\mathbb{G}_{t+1} &= -\mathbb{E}\left[\Delta_{x_t}^{x_{t+1}}(\log p_t + \log p(x_{t+1}|x_t) + \log p(y_{t+1}|x_{t+1})\right] \\ &= -\mathbb{E}\left[\Delta_{x_t}^{x_{t+1}}\log p(x_{t+1}|x_t)\right] = \mathbb{D}_t^{12}\end{aligned}$$

and $\mathbb{G}_{t+1}^T = \mathbb{D}_t^{21}$.

Finally

$$\begin{aligned}\mathbb{F}_t &= -\mathbb{E}\left[\Delta_{x_{0:t+1}}^{x_{t+1}}(\log p_t + \log p(x_{t+1}|x_t) + \log p(y_{t+1}|x_{t+1}))\right] \\ &= -\mathbb{E}\left[\Delta_{x_{t+1}}^{x_{t+1}}\log p(x_{t+1}|x_t)\right] - \mathbb{E}\left[\Delta_{x_{t+1}}^{x_{t+1}}\log p(y_{t+1}|x_{t+1})\right] \\ &= \mathbb{D}_t^{22}.\end{aligned}$$

We have now derived at an expression for $\mathbb{J}(X_{0:t+1})$

$$\mathbb{J}(X_{0:t+1}) = \begin{bmatrix}\mathbb{A}_t & \mathbb{B}_t & \mathbf{0} \\ \mathbb{B}_t^T & \mathbb{C}_t + \mathbb{D}_t^{11} & \mathbb{D}_t^{12} \\ \mathbf{0} & \mathbb{D}_t^{21} & \mathbb{D}_t^{22}\end{bmatrix}.$$

From (3.6) we get

$$\mathbb{J}_{t+1} = \mathbb{D}_t^{22} - [\mathbf{0}\,\mathbb{D}_t^{21}]\begin{bmatrix}\mathbb{A}_t & \mathbb{B}_t \\ \mathbb{B}_t^T & \mathbb{C}_t + \mathbb{D}_t^{11}\end{bmatrix}^{-1}[\mathbf{0}\,\mathbb{D}_t^{12}]^T.$$

Since the right lower block of

$$\begin{bmatrix}\mathbb{A}_t & \mathbb{B}_t \\ \mathbb{B}_t^T & \mathbb{C}_t + \mathbb{D}_t^{11}\end{bmatrix}^{-1} = [\mathbb{C}_t + \mathbb{D}_t^{11} - \mathbb{B}_t^T\mathbb{A}_t\mathbb{B}_t]^{-1},$$

also remembering that $\mathbb{J}_t = [\mathbb{C}_t + \mathbb{B}_t^T \mathbb{A}_t^{-1} \mathbb{B}_t]^{-1}$ we finally arrive at

$$\mathbb{J}_{t+1} = \mathbb{D}_t^{22} - \mathbb{D}_t^{21}[\mathbb{J}_t + \mathbb{D}_t^{11}]^{-1}\mathbb{D}_t^{12}$$

with initial matrix

$$\mathbb{J}_0 = \mathbb{E}\left[-\Delta_{x_0}^{x_0} \log p(x_0)\right].$$

□

3.3 Special Cases

The first case is when the initial distribution is Gaussian, that is $p(x_0) = N(x_0; \mu_0, \mathbb{P}_0)$ Then

$$\nabla_{x_0} \log p(x_0) = \nabla_{x_0}\left\{c - \frac{1}{2}[(x_0 - \mu_0)^T \mathbb{P}_0^{-1}(x_0 - \mu_0)]\right\} = -\mathbb{P}_0^{-1}(x_0 - \mu_0)$$

where c is a constant. Now straightforward matrix algebra and using the fact that the covariance matrix $\mathbb{P}_0$ and it's inverse are symmetric matrices, we deduce that

$$\begin{aligned}\mathbb{J}_0 &= \mathbb{E}\left[\mathbb{P}_0^{-1}(X_0 - \mu_0)(X_0 - \mu_0)^T [\mathbb{P}_0^{-1}]^T\right] \\ &= \mathbb{P}_0^{-1}\mathbb{E}\left[(X_0 - \mu_0)(X_0 - \mu_0)^T\right]\mathbb{P}_0^{-1} \\ &= \mathbb{P}_0^{-1}\mathbb{P}_0\mathbb{P}_0^{-1} = \mathbb{P}_0.\end{aligned}$$

The next special case is for the additive Gaussian noise case.

3.3.1 Additive Gaussian noise

Let us once again consider the filtering problem

$$\begin{aligned}X_{t+1} &= k(X_t) + V_t \\ Y_{t+1} &= h(X_{t+1}) + W_{t+1},\end{aligned}$$

and let us now assume that the noise sequences V_t and W_{t+1} are mutually independent , zero mean Gaussian variables with covariances $\mathbb{Q}_t$ and $\mathbb{R}_{t+1}$. We also add an additional condition that the matrices are nonsingular such that there exist a unique inverse. Under these assumptions we have

$$\begin{aligned}\nabla_{x_t} \log p(x_{t+1}|x_t) &= \nabla_{x_t}\left[-\frac{1}{2}\left(x_{t+1} - k(x_t)\right)^T \mathbb{Q}_t^{-1}\left(x_{t+1} - k(x_t)\right)\right] \\ &= \left[\nabla_{x_t} k^T(x_t)\right]\mathbb{Q}_t^{-1}\left[x_{t+1} - k(x_t)\right],\end{aligned}$$

and in the same way

$$\nabla_{x_{t+1}} \log p(y_{t+1}|x_{t+1}) = \left[\nabla_{x_{t+1}} h^T(x_{t+1})\right] \mathbb{R}_{t+1}^{-1} \left[y_{t+1} - h_{t+1}(x_{t+1})\right].$$

Matrix $\mathbb{D}_t^{11}$ simplifies as follows

$$\begin{aligned}
\mathbb{D}_t^{11} &= -\mathbb{E}\left[\nabla_{x_t}[\nabla_{x_t} \log p(x_{t+1}|x_t)]^T\right] \\
&= \mathbb{E}\left[[\nabla_{x_t} \log p(x_{t+1}|x_t)][\nabla_{x_t} \log p(x_{t+1}|x_t)]^T\right] \\
&= \mathbb{E}\left[[\nabla_{x_t} k^T(X_t)]\mathbb{Q}_t^{-1}[X_{t+1} - k(X_t)][X_{t+1} - k(X_t)]^T\mathbb{Q}_t^{-1}[\nabla_{x_t} k^T(X_t)]^T\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[[\nabla_{x_t} k^T(X_t)]\mathbb{Q}_t^{-1}[X_{t+1} - k(X_t)][X_{t+1} - k(X_t)]^T\mathbb{Q}_t^{-1}[\nabla_{x_t} k^T(X_t)]|X_t\right]\right] \\
&= \mathbb{E}\left[[\nabla_{x_t} k^T(X_t)]\mathbb{Q}_t^{-1}[\nabla_{x_t} k^T(X_t)]^T\right] \\
&= \mathbb{E}\left[\tilde{\mathbb{K}}_t^T\mathbb{Q}_t^{-1}\tilde{\mathbb{K}}_t\right],
\end{aligned}$$

where

$$\tilde{\mathbb{K}}_t = [\nabla_{x_t} k^T(x_t)]^T$$

is the Jacobian of $k(x_t)$ evaluated at the true value of x_t. In much the same way one can show that

$$\mathbb{D}_t^{12} = -\mathbb{E}\left[\tilde{\mathbb{K}}_t^T\right]\mathbb{Q}_t^{-1} \tag{3.7}$$

$$\mathbb{D}_t^{22} = \mathbb{Q}_t^{-1} + \mathbb{E}\left[\tilde{\mathbb{H}}_{t+1}^T\mathbb{R}_{t+1}^{-1}\mathbb{H}\tilde{}_{t+1}\right], \tag{3.8}$$

where

$$\tilde{\mathbb{H}}_{t+1} = [\nabla_{x_{t+1}} h_{t+1}^T(x_{t+1})]^T$$

is the Jacobian of $h_{t+1}(x_{t+1})$ evaluated at the true value of x_{t+1}.

Usually it is the expectation operator $\mathbb{E}$ that causes problems in the calculation of the PCBR. A Monte Carlo approximation can be applied to address this problem. One creates an ensamble of state realisations and use the average over these ensambles as an estimate for the theoretical value.

4
Convergence

4.1 Introduction

In this chapter we will look at some of the theoretical aspects of particle filters. We start with some convergence theorems under the assumption that both X and Y take values in the Euclidean space, and that the joint and conditional densities exists at each time step t. Further on we deduce some convergence properties in a more general situation. We will need some properties of the conditional expectations and probabilities (Appendix). This surrey is based on Crisan (2001) and Del Moral and Jacod (2001)

4.2 The filtering problem

Let $X = \{X_t,\ t \in \mathbb{N}\}$ be an $\mathbb{R}^d$-valued hidden Markov process with a Feller transition Q_t.
(Q_t is the transition from X_{t-1} to X_t). The observed process, $Y = \{Y_t, t \in \mathbb{N}\}$ is an $\mathbb{R}^q$-valued stochastic process and defined by,

$$Y_t \triangleq h(t, X_t) + W_t,\ t > 0 \tag{4.1}$$

with $Y_0 = 0$.
In (4.1), $h : \mathbb{N} \times \mathbb{R}^d \to \mathbb{R}^q$ is a Borel-measurable function with the property that $h(t, \cdot)$ is continuous on $\mathbb{R}^d$ for all $t \in \mathbb{N}$. The noise process $\{W_t\}$ is independent of $\{X_t\}$ and for each t, W_t has a bounded continuous density $\bar{g}_t$.

Our aim is to compute sequentially in time the conditional distribution of the signal given the σ-algebra, $\mathcal{F}_t^Y$, generated by the observation process up to the current time. That is, we are interested in the random probability measure $\hat{\eta}_t^{Y_{0:t}}$,

$$\hat{\eta}_t^{Y_{0:t}}(f) \triangleq \mathbb{E}\left[f(X_t)|\mathcal{F}_t^Y\right] \tag{4.2}$$

for all $f \in \mathbf{B}(\mathbb{R}^d)$, and the deterministic probability measure $\hat{\eta}_t^{y_{0:t}}$

$$\hat{\eta}_t^{y_{0:t}}(f) \triangleq \mathbb{E}\left[f(X_t)|Y_{0:t} = y_{0:t}\right] \tag{4.3}$$

which we from now will denote only as $\hat{\eta}_t$. In the same way we introduce the prediction distribution for $t > 0$ by $\eta_t^{Y_{0:t-1}}$ and $\eta_t^{y_{0:t-1}}$.

$$\eta_t^{Y_{0:t-1}}(f) = \mathbb{E}\left[f(X_t)|\mathcal{F}_{t-1}^Y\right]$$

$$\eta_t^{y_{0:t-1}}(f) = \mathbb{E}\left[f(X_t)|Y_{0:t-1} = y_{0:t-1}\right].$$

We have a recursion formula, analog to (1.3) and (1.5), for these probability measures in the following lemma

Lemma 4.1

The probability measures introduced satisfies the following recursions

$$\begin{cases} \hat{\eta}_t^{Y_{0:t}}(dx) = & = \dfrac{g_t^{Y_t}(x)}{\eta_t g_t^{Y_t}} \eta_t^{Y_{0:t}}(dx) \\ \eta_{t+1} & = \hat{\eta}_t Q_{t+1} \end{cases} \begin{cases} \hat{\eta}_t(dx) & = \dfrac{g_t^{y_t}(x)}{\eta_t g_t^{y_t}} \eta_t(dx) \\ \eta_{t+1} & = \hat{\eta}_t Q_{t+1}, \end{cases}$$

where $g_t^{y_t}$ is defined by $g_t^{y_t} = \bar{g}_t(y_t - h(t, \cdot))$ and since $Y_0 = 0$, η_0 is the law of X.

Proof: for proof see Appendix.

4.3 Convergence of measure-valued random variables

When we consider algorithms with sequential Monte Carlo methods that solves the filtering problem, the result is essentially a random measure which approximates $\hat{\eta}_t$. In order to establish any results about the convergence of the algorithms, we must define in what way a sequence of random measures can approximate another measure.

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and $(\mu^N)_{N=1}^{\infty}$ a sequence of random measures, $\mu^N : \Omega \to \mathcal{M}_F(\mathbb{R}^d)$ and $\mu \in \mathcal{M}_F(\mathbb{R}^d)$ is deterministic. (N will typically denote the number of particles in the algorithm). We will study two types of convergence

1. $\lim_{N\to\infty} \|\mu^N f - \mu f\|_1 = 0 \; \forall f \in \mathbf{C}_b(\mathbb{R}^d)$
2. $\lim_{N\to\infty} \mu^N = \mu, \; \mathcal{P} - a.s.,$

where $\|\mu^N f - \mu f\|_1 \triangleq \mathbb{E}\left|\mu^N f - \mu f\right|$. The first type we will denote by Elim. If $|\mu^N 1|$ is dominated by an integrable random variable Z, (2.) implies (1.) by the dominated convergence theorem. This condition is trivially satisfied if $(\mu^N)_{N=1}^{\infty}$ is a sequence of random probability measures since $\mu^N 1 = 1$ for all N.

Example

Suppose that $X_1, \ldots, X_N$ is a sample from F with empirical distribution F_N, then

$$F_N \xrightarrow[N]{\text{Elim}} F$$

by Chebyshev, and

$$F_N \xrightarrow[N]{a.s.} F.$$

Suppose that $X_1^\star, \ldots, X_N^\star$ is a bootstrap sample from $X_1, \ldots, X_N$, then

$$F_N^\star \xrightarrow[N]{\text{Elim}} F \quad F - a.s.$$

where 'Elim' is with respect to the sampling.

Theorem 4.2

If $\mu^N \xrightarrow{\text{Elim}} \mu$ then there exists a subsequence N_k such that $\mu^{N_k} \xrightarrow{\text{a.s.}} \mu$.

Proof: :

Since $\mathbb{R}^d$ is a locally compact separable metric space, there exists a countable set $\mathcal{Y} \subset \mathbf{C}_b(\mathbb{R}^d)$ which is dense. I.e. if $\nu^N, N = 1, 2....$ and ν are finite measures and $\lim_{N\to\infty} \nu^N f = \nu f$ for all $f \in \mathcal{Y}$ then $\lim_{N\to\infty} \nu^N = \nu$. Since $E \lim_{N\to\infty} \mu^N = \mu$ for all $f \in \mathcal{Y}$ and $\mathcal{Y}$ is countable there exists a subsequence N_1 such that with probability 1, $\lim_{N_1\to\infty} \mu^{N_1} f_1 = \mu f_1$. Also there exists a subsequence N_2 of N_1 such that $\mu^{N_2} f_2$ converges $\mathcal{P}$- a.s to μf_2. However, this subsequence will also converge almost surely for f_1 being a subsequence of N_1. Continuing this way we get the following scheme

$$\begin{array}{lllll}
\mu^{11} & \mu^{21} & \mu^{31} & ... & \text{converges a.s for } f_1 \\
\mu^{12} & \mu^{22} & \mu^{32} & ... & \text{converges a.s for } f_1, f_2 \\
... & ... & ... & ... & ... \\
... & ... & ... & ... & ... \\
... & ... & ... & ... & ... \\
\mu^{1k} & \mu^{2k} & \mu^{3k} & ... & \text{converges a.s for} f_1,\ f_2, ...\ f_k \\
... & ... & ... & ... & ...
\end{array}$$

The diagonal process μ^{kk} will converge almost surely for all $f \in \mathcal{Y}$. □

If the rate of convergence of $\|\mu^N f - \mu f\|_1$ is known the sequence can be explicitly specified.

Example

Assume that for all $f \in \mathcal{Y}$, $\mathbb{E}\left|\mu^N f - \mu f\right| \leq c_f N^{-\frac{1}{2}}$. Then by Markov's inequality for any given $\epsilon > 0$

$$\sum_{N=1}^{\infty} P\left(|\mu^{N^4} f - \mu f| \geq \epsilon\right) \leq \sum_{N=1}^{\infty} \frac{c_f (N^4)^{-\frac{1}{2}}}{\epsilon} = \frac{c_f}{\epsilon} \sum_{N=1}^{\infty} N^{-2}.$$

Since this series converges, a Borel-Cantelli argument (Williams, 1991) assures us that

$$\lim_{N\to\infty} \mu^{N^4} = \mu,\ \mathcal{P}\text{-a.s.}$$

If $\mathcal{Y}$ is the set defined above then

$$d_{\mathcal{Y}}(\mu, \nu) \triangleq |\mu 1 - \nu 1| + \sum_{f_k \in \mathcal{Y}} \frac{|\mu f_k - \nu f_k|}{2^k \|f_k\|} \tag{4.4}$$

is a metric on $\mathcal{M}_F(\mathbb{R}^d)$ (or $\mathcal{P}(\mathbb{R}^d)$) which generates the weak topology

$$\lim_{N\to\infty} \nu_N = \nu \Leftrightarrow \lim_{N\to\infty} d_{\mathcal{Y}}(\nu_N, \nu) = 0.$$

Using $d_{\mathcal{Y}}$, the almost sure convergence (2.) is equivalent to

$$2.' \lim_{N\to\infty} d_{\mathcal{Y}}(\mu^N, \mu) = 0,\ \mathcal{P} - a.s..$$

Also, if $|\mu^N 1|$ is dominated by an integrable random variable Z then (1.) implies

$$1.' \lim_{N\to\infty} \mathbb{E}\left[d_{\mathcal{Y}}(\mu^N, \mu)\right] = 0.$$

A stronger condition (such as tightness) is needed in order to ensure that (1.) is equivalent to (1.'). The same definitions are valid in the case when the limiting measure μ is a random measure $\mu : \Omega \to \mathcal{M}_F(\mathbb{R}^d)$. The same implications are valid under the same assumptions as before.

The limiting measures in the filtering problem is $\hat{\eta}_t^{Y_{0:t}}$ and $\hat{\eta}_t$ (with the observations fixed), hence we have one random and one deterministic probability measure, however we will only focus on the deterministic one.

4.3.1 Convergence theorems for the fixed observation case

We now assume that we have observed values of the observation process up to time T, that is we have $y_{0:T}$ where T is finite but large. We also assume that all the recurrence formulae in lemma 4.1 holds true for this particular value for all $0 \leq t \leq T$. Based on lemma 4.1 we see that in any algorithm we need an intermediate prediction step.

$$\hat{\eta}_{t-1} \longrightarrow \eta_t \longrightarrow \hat{\eta}_t.$$

Let us denote by $(\hat{\eta}_t^N)_{N=1}^{\infty}$ and $(\eta_t^N)_{N=1}^{\infty}$ the approximating sequence for $\hat{\eta}_t$ and η_t and assume that $\hat{\eta}_t^N$ and η_t^N are random measures (not necessarily probability measures) and non-trivial i.e. $\hat{\eta}_t^N \neq 0$, $\eta_t^N \neq 0$ and $\eta_t^N g_t^{y_t} \neq 0$, for all $N > 0$ and $0 \leq t \leq T$. Let us define by $\check{\eta}_t^N$ a random probability measure absolutely continuous w.r.t. η_t^N for $t \in \mathbb{N}$ and $N \geq 1$ such that for $A \in \mathcal{B}\left(\mathbb{R}^d\right)$

$$\check{\eta}_t^N(A) = \eta_t^N\left(\frac{g_t}{\eta_t^N g_t} 1_A\right), \tag{4.5}$$

where $g_t = g_t^{y_t}$ and 1_A is the indicator function for the set A.

We are now able to state the following theorem which gives us necessary and sufficient conditions for the convergence of η_t^N and $\hat{\eta}_t^N$ to η_t and $\hat{\eta}_t$

Theorem 4.3

> The sequence η_t^N and $\hat{\eta}_t^N$,defined by (4.2) and (4.3), converge to η_t and $\hat{\eta}_t$, with convergence taken to be of type **1.** if and only if the following three conditions are satisfied.
>
> **a1.** For all $f \in \mathbf{C}_b(\mathbb{R}^d)$, $\lim_{n\to\infty} \|\eta_0^N f - \eta_0 f\|_1 = 0$
>
> **b1.** For all $f \in \mathbf{C}_b(\mathbb{R}^d)$, $\lim_{n\to\infty} \|\eta_t^N f - \hat{\eta}_{t-1}^N Q_t f\|_1 = 0$
>
> **c1.** For all $f \in \mathbf{C}_b(\mathbb{R}^d)$, $\lim_{n\to\infty} \|\hat{\eta}_t^N f - \check{\eta}_t^N f\|_1 = 0.$

Proof: The sufficiency is proved by mathematical induction. The theorem holds true for $t = 0$ by **a1.** Next we assume that η_{t-1}^N and $\hat{\eta}_{t-1}^N$ converges to η_{t-1} and $\hat{\eta}_{t-1}$. Then, since $\eta_t = \hat{\eta}_{t-1} Q_t$ we have for all $f \in \mathbf{C}_b(\mathbb{R}^d)$

$$|\eta_t^N f - \eta_t f| \leq |\eta_t^N f - \hat{\eta}_{t-1}^N Q_t f| + |\hat{\eta}_{t-1}^N Q_t f - \hat{\eta}_{t-1} Q_t f|. \tag{4.6}$$

By taking expectations on both sides the first term on the right side converges to 0 by **b1** and the second term by the induction hypothesis since $Q_t f \in \mathbf{C}_b(\mathbb{R}^d)$ by the Feller property of the kernel. Next we use lemma 4.1, (4.5) and the triangle inequality

$$\begin{aligned}
|\check{\eta}_t^N f - \hat{\eta}_t f| &= \left|\frac{\eta_t^N f g_t}{\eta_t^N g_t} - \frac{\eta_t f g_t}{\eta_t g_t}\right| \\
&\leq \left|\frac{\eta_t^N f g_t}{\eta_t^N g_t} - \frac{\eta_t^N f g_t}{\eta_t g_t}\right| + \left|\frac{\eta_t^N f g_t}{\eta_t g_t} - \frac{\eta_t f g_t}{\eta_t g_t}\right| \\
&= \left|\frac{\eta_t^N f g_t \cdot \eta_t g_t}{\eta_t^N g_t \cdot \eta_t g_t} - \frac{\eta_t^N f g_t \cdot \eta_t^N g_t}{\eta_t g_t \cdot \eta_t^N g_t}\right| + \frac{1}{\eta_t g_t}|\eta_t^N f g_t - \eta_t f g_t| \\
&\leq \frac{\|f\|}{\eta_t g_t}|\eta_t^N g_t - \eta_t g_t| + \frac{1}{\eta_t g_t}|\eta_t^N f g_t - \eta_t f g_t|
\end{aligned}, \tag{4.7}$$

hence

$$\|\check{\eta}_t^N f - \hat{\eta}_t f\|_1 \leq \frac{\|f\|}{\eta_t g_t}\|\eta_t^N g_t - \eta_t g_t\|_1 + \frac{1}{\eta_t g_t}\|\eta_t^N f g_t - \eta_t f g_t\|_1. \tag{4.8}$$

Since g_t and $g_t f$ are continuous and bounded both terms converge to zero from (4.6). Finally ,

$$\|\hat{\eta}_t^N f - \hat{\eta}_t f\|_1 \leq \|\hat{\eta}_t^N f - \check{\eta}_t^N f\|_1 + \|\check{\eta}_t^N f - \hat{\eta}_t f\|_1. \tag{4.9}$$

The first term on the right hand side of (4.9) converges to 0 from **c1.** and the second term converges to 0 from (4.8). □

Next we prove the necessity part.
Assume that for all $t \geq 0$ and for all $f \in \mathbf{C}_b(\mathbb{R}^d)$,

$$\lim_{N\to\infty} \|\eta_t^N f - \eta_t f\|_1 = 0$$
$$\lim_{N\to\infty} \|\hat{\eta}_t^N f - \hat{\eta}_t f\|_1 = 0.$$

Then **a1** is trivially satisfied. Next, from (4.8) we have that $\|\check{\eta}_t^N f - \hat{\eta}_t f\|_1 = 0$, and since

$$\|\hat{\eta}_t^N f - \check{\eta}_t^N f\|_1 \leq \|\hat{\eta}_t^N f - \hat{\eta}_t f\|_1 + \|\hat{\eta}_t f - \check{\eta}_t^N f\|_1,$$

c1 is obtained. Finally, using once again that $\eta_t = \hat{\eta}_{t-1} Q_t$ and the Feller property of Q_t we have for all $f \in \mathbf{C}_b(\mathbb{R}^d)$

$$\|\eta_t^N f - \hat{\eta}_t^N Q_t f\|_1 \leq \|\eta_t^N f - \eta_t f\|_1 + \|\hat{\eta}_{t-1} Q_t f - \hat{\eta}_{t-1} Q_t f\|_1$$

which implies **b1**. □

We also have a corresponding theorem for the almost sure convergence of $\eta_t^N, \hat{\eta}_t^N$ to η_t and $\hat{\eta}_t$.

Theorem 4.4

Let t be fixed. The sequence $\eta_t^N, \hat{\eta}_t^N$ converges almost surely (in the weak sense) to η_t and $\hat{\eta}_t$ if and only if the following three conditions are satisfied;

a2. $\lim_{N\to\infty} \eta_0^N = \eta_0,\ \mathcal{P} - a.s.$

b2. $\lim_{N\to\infty} d_{\mathcal{Y}}(\eta_t^N, \hat{\eta}_{t-1}^N Q_t) = 0,\ \mathcal{P} - a.s.$

c2. $\lim_{N\to\infty} d_{\mathcal{Y}}(\hat{\eta}_t^N, \check{\eta}_t^N) = 0,\ \mathcal{P} - a.s..$

Proof: The sufficiency part is proved in the same way as Theorem 4.3.1 using mathematical induction and inequalities (4.6), (4.8) and (4.9) (without the expectations). Now the necessity part.

Assume that for all $t \geq 0$ η_t^N and $\hat{\eta}_t^N$ converges almost surely to η_t and $\hat{\eta}_t$. This implies that $\hat{\eta}_{t-1}^N Q_t$ converges a.s. to $\hat{\eta}_{t-1} Q_t = \eta_t$, and from (4.7) we get that $\check{\eta}_t^N$ converges a.s. to $\hat{\eta}_t$ According to 2.$'$ we then have almost surely $\lim_{N\to\infty} d_{\mathcal{Y}}(\eta_t^N, \eta_t) = 0$, $\lim_{N\to\infty} d_{\mathcal{Y}}(\hat{\eta}_t^N, \hat{\eta}_t) = 0$, $\lim_{N\to\infty} d_{\mathcal{Y}}(\hat{\eta}_{t-1}^N Q_t, \eta_t) = 0$ and $\lim_{N\to\infty} d_{\mathcal{Y}}(\tilde{\eta}_t^N, \hat{\eta}_t) = 0$, where $d_{\mathcal{Y}}$ is defined as in (4.4). We now obtain **b.2** and **c.2** by the triangleinequalities

$$d_{\mathcal{Y}}(\eta_t^N, \hat{\eta}_{t-1}^N Q_t) \leq d_{\mathcal{Y}}(\eta_t^N, \eta_t) + d_{\mathcal{Y}}(\eta_t, \hat{\eta}_{t-1}^N Q_t)$$
$$d_{\mathcal{Y}}(\hat{\eta}_t^N, \check{\eta}_t^N) \leq d_{\mathcal{Y}}(\hat{\eta}_t^N, \hat{\eta}_t) + d_{\mathcal{Y}}(\hat{\eta}_t, \check{\eta}_t^N).$$

□

Let us assume that we conduct a particle filter scheme for X_t according to algorithm 2.2, where we use Q_t as the importance function so that the weights are proportional to g_t. We will now prove the convergence of the random measures prodeuced by the algorithm

$$\eta_t \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^{(i)}} \qquad \hat{\eta}_t \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{\hat{X}_t^{(i)}}$$

to η_t and $\hat{\eta}_t$. First we need to introduce the following σ-algebras

$$\hat{\mathcal{F}}_t^X = \sigma\left(X_s^{(i)}, \hat{X}_s^{(i)}, s \leq t, \quad i = 1, \dots, N\right)$$
$$\mathcal{F}_t^X = \sigma\left(X_s^{(i)}, \hat{X}_s^{(i)}, s < t, \quad X_t^{(i)} \quad i = 1, \dots, N\right).$$

Theorem 4.5

Let $\left(\eta_t^N\right)_{N=1}^{\infty}$ and $\left(\hat{\eta}_t^N\right)_{N=1}^{\infty}$ be the measure valued sequences produced by algorithm 2.2 and let T be a finite time horizon. Then, for all $0 \leq t \leq T$, we have

$$\eta_t^N \xrightarrow[N]{E\lim} \eta_t \qquad \hat{\eta}_t^N \xrightarrow[N]{E\lim} \hat{\eta}_t.$$

Proof: We apply Theorem 4.3

Since **a1.** is clearly satisfied ($X_0^{(i)} \sim \eta_0$) we only need to show **b1.** and **c1.** If $f \in \mathbf{C}_b\left(\mathbb{R}^d\right)$ then,

$$\mathbb{E}\left[f(X_t^{(i)}) | \hat{\mathcal{F}}_{t-1}^X\right] = Q_t f(\hat{X}_{t-1}^{(i)}) = \int Q_t(\hat{X}_{t-1}^{(i)}, dx) f(x), \quad i = 1, \dots N,$$

hence $\mathbb{E}[\eta_t^N f|\hat{\mathcal{F}}_{t-1}^X] = \hat{\eta}_{t-1}^N Q_t f$, and using the independence of the particles,

$$\begin{aligned}
\mathbb{E}\left[\left(\eta_t^N f - \hat{\eta}_t^N Q_t f\right)^2 |\hat{\mathcal{F}}_{t-1}^X\right] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \left(f(X_t^{(i)}) - Q_t f(\hat{X}_{t-1}^{(i)})\right)\right)^2 |\hat{\mathcal{F}}_{t-1}^X\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N \left(\mathbb{E}\left[f^2(X_t^{(i)})|\hat{\mathcal{F}}_{t-1}^X\right] - \left(\mathbb{E}\left[Q_t f(\hat{X}_{t-1}^{(i)})|\hat{\mathcal{F}}_{t-1}^X\right]\right)^2\right) \\
&= \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^N \left(Q_t f^2(\hat{X}_{t-1}^{(i)}) - \left(Q_t f(\hat{X}_{t-1}^{(i)})\right)^2\right)\right) \\
&= \frac{1}{N}\hat{\eta}_{t-1}^N \left(Q_t f^2 - (Q_t f)^2\right) \leq \frac{\|f\|^2}{N}.
\end{aligned}$$

Then, by taking the expectation on both sides we obtain

$$\mathbb{E}\left[\left(\eta_t^N f - \hat{\eta}_{t-1}^N Q_t f\right)^2\right] \leq \frac{\|f\|^2}{N}$$

and **b.1** is satisfied. Next we have $\hat{\eta}_t^N = \frac{1}{N}\sum_{i=1}^N n_t^{(i)} \delta_{X_t^{(i)}}$, where $n_t^{(i)}$ is the number of offsprings produced by particle number i in the resampling step. Since the resampling step is carried out using a multinomial model, we have $\mathbb{E}[n_t^{(i)}] = N\tilde{w}_t^{(i)}$, such that

$$\mathbb{E}\left[\hat{\eta}_t^N f|\mathcal{F}_t^X\right] = \frac{1}{N}\sum_{i=1}^N N\tilde{w}_t^{(i)} f(X_t^{(i)}) = \sum_{i=1}^N \tilde{w}_t^{(i)} f(X_t^{(i)}) = \tilde{\eta}_t^N f.$$

Furthermore we have, using the property of the multinomial distribution,

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\eta}_t^N f - \check{\eta}_t^N f\right)^2 |\mathcal{F}_t^X\right] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \left(n_t^{(i)} f(X_t^{(i)}) - N\tilde{w}_t^{(i)} f(X_t^{(i)})\right)\right)^2 |\mathcal{F}_t^X\right] \\
&\leq \frac{\|f\|^2}{N^2}\mathbb{E}\left[\left(\sum_{i=1}^N \left(n_t^{(i)} - N\tilde{w}_t^{(i)}\right)\right)^2 |\mathcal{F}_t^X\right] \\
&= \frac{\|f\|^2}{N^2}\left(\sum_{i=1}^N N\tilde{w}_t^{(i)}(1-\tilde{w}_t^{(i)}) - \sum_{i\neq j} N\tilde{w}_t^{(i)}\tilde{w}_t^{(j)}\right) \\
&\leq \frac{\|f\|^2}{N^2} N \sum_{i=1}^N \tilde{w}_t^{(i)} = \frac{\|f\|^2}{N}
\end{aligned}$$

since $\sum_{i=1}^N \tilde{w}_t^{(i)} = 1$. Taking expectations on both sides **c.1** is satisfied. □

Theorem 4.6

Let $\left(\eta_t^N\right) N = 1^\infty$ and $\left(\hat{\eta}_t^N\right)_{N=1}^\infty$ be the measure-valued sequence produced by algorithm 2.2 and let T be a finite time horizon.
Then, for all $0 \leq t \leq T$, we have

$$\lim_{N\to\infty} \eta_t^N = \eta_t \quad \mathcal{P}-a.s. \qquad \lim_{N\to\infty} \hat{\eta}_t^N = \hat{\eta}_t, \quad \mathcal{P}-a.s..$$

Proof: We apply Theorem 4.4.

Let $\mathcal{M} \in \mathbf{C}_b\left(\mathbb{R}^d\right)$ be the countable convergence determining set of functions described in the previous section. Since $\mathbb{E}\left[f\left(X_t^{(i)}\right) | \hat{\mathcal{F}}_{t-1}^X\right] = Q_t f\left(\hat{X}_{t-1}(i)\right)$ and $\{X_t^{(i)}\}_{i=1}^N$ are independent given $\hat{\mathcal{F}}_{t-1}^X$ we have

$$\mathbb{E}\left[\left(\eta_t^N f - \hat{\eta}_t^N Q_t f\right)^4 | \hat{\mathcal{F}}_{t-1}^X\right] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \left(f\left(X_t^{(i)}\right) - Q_t f\left(\hat{X}_{t-1}^{(i)}\right)\right)\right)^4 |\hat{\mathcal{F}}_{t-1}^X\right].$$

Next, using the independence, all the crossproducts of the expectation involving first power terms will be equal to zero and we are left with

$$\begin{aligned}
\mathbb{E}\left[\left(\eta_t^N f - \hat{\eta}_t^N Q_t f\right)^4 | \hat{\mathcal{F}}_{t-1}^X\right] &= \frac{1}{N^4}\sum_{i=1}^N \mathbb{E}\left[\left(f(X_t^{(i)}) - Q_t f(\hat{X}_{t-1}^{(i)})\right)^4 | \hat{\mathcal{F}}_{t-1}^X\right] \\
+\frac{2}{N^4}\sum_{1\leq i<j\leq N} \mathbb{E}&\left[\left(f(X_t^{(i)}) - Q_t f(\hat{X}_{t-1}^{(i)})\right)^2 \left(f(X_t^{(j)}) - Q_t f(\hat{X}_{t-1}^{(j)})\right)^2 | \hat{\mathcal{F}}_{t-1}^X\right] \\
&\leq \frac{16\|f\|^4}{N^3} + \frac{32\|f\|^4}{N^4}\frac{N(N-1)}{2} \\
&= \frac{16\|f\|^4}{N^3} + \frac{16\|f\|^4(N-1)}{N^3} = \frac{16\|f\|^4}{N^2}.
\end{aligned}$$

By taking expectations on both sides we obtain $\mathbb{E}\left[\left(\eta_t^N f - \hat{\eta}_t^N Q_t f\right)^4\right] \leq \frac{16\|f\|^4}{N^2}$ and via a Borel-Cantelli argument we have that $\lim_{N\to\infty}|\eta_t^N f - \hat{\eta}_{t-1}^N Q_t f| = 0 \quad \mathcal{P}-a.s.$ for all $f \in \mathcal{M}$ so that $\lim_{N\to\infty} d_{\mathcal{M}}(\eta_t^N, \hat{\eta}_{t-1}^N Q_t) = 0$ and **b.2** is satisfied. In much the same way one can show that, for all $f \in \mathcal{M}$,

$$\mathbb{E}\left[\left(\hat{\eta}_t^N f - \check{\eta}_t^N f\right)^4 | \mathcal{F}_t^X\right] \leq \frac{\|f\|^4}{N^2}$$

which implies that $\lim_{N\to\infty} d_{\mathcal{M}}\left(\hat{\eta}_t^N, \check{\eta}_t^N\right) = 0$ and **c.2** is satisfied. □

Note that the rate of convergence is $N^{-\frac{1}{2}}$. We now turn our focus to a little more general particle filter.

4.4 Interacting particle filters with discrete observations

In this section we will consider a pair of processes $(\{X_t\}, \{Y_t\})$, where $\{X_t\}$ is the state of a system and $\{Y_t\}$ is the observations. X take it's values in an arbitrary measurable space $(\mathbf{E}, \mathcal{E})$ while Y take it's values in $\mathbb{R}^q$ for some $q \geq 1$. We shall assume that the pair (X, Y) is Markov, and the basic assumption is that the pair $(X_t, Y_t)_{t \in \mathbb{N}}$ is a (possibly non-homogeneous) Markov chain. In the sections below, we will focus on the nonlinear filtering problem (NLF). In other words we want to find the one step predictor conditional probability given for each $t \in \mathbb{N}$ and each measurable function f on $\mathbf{E}$ such that $f(X_t)$ is integrable by

$$\eta_{t,Y} f = \mathbb{E}[f(X_t) | \mathcal{F}_t^Y]$$

(where $\eta_{0,Y}$ is the law of X_0) and the filter conditional distribution

$$\hat{\eta}_{t,Y} f = \mathbb{E}[f(X_t) | Y_{0:t}]$$

With the notation $Y_{0:k} = (Y_0, Y_1 ... Y_k)$. For fixed observations $Y_t = y_t, t \in \mathbb{N}$ we write η_t and $\hat{\eta}_t$ instead of $\eta_{t,y}$ and $\hat{\eta}_{t,y}$. We want to investigate theoretical aspects of an interacting particle system (IPS) for numerical computations of η_t and $\hat{\eta}_t$ in two cases.

Assumptions A

In case A we consider the following system:

A.1 The state signal $(X_t)_{t \in \mathbb{N}}$ is an $\mathbf{E}$ valued non-homogeneous Markov chain with 1-step transition probabilities $(Q_t)_{t \in \mathbb{N}}$ (i.e. Q_t is the law $X_{t-1} \to X_t$) with initial law η_0.

A.2 The observations $(Y_t)_{t \in \mathbb{N}}$ is given by

$$Y_t = h_t(X_t, W_t)$$

for some measurable function H_t from $\mathbf{E} \times \mathbf{F}$ into $\mathbb{R}^q$ (with $(\mathbf{F}, \mathcal{F})$ an auxiliary measurable space).

A.3 For any $x \in \mathbf{E}$ and $\forall t$, the variable $h_t(x, V_t)$ admits a strictly positive density $y \to \bar{g}_t(x, y)$ w.r.t. the Lebesgue measure on $\mathbb{R}^q$.

Assumptions B

In case B we assume

B.1 The signal/observation pair (X, Y) is an $\mathbf{E} \times \mathbb{R}^q$- valued non-homogeneous Markov chain with 1-step transition probabilities $(P_t)_{t\geq 1}$ and initial law μ_0 on the form ($\mathrm{d}y$ denotes the Lebesgue measure)

$$\begin{aligned}
&\mu_0(\mathrm{d}x_0, \mathrm{d}y_0) = \eta_0(\mathrm{d}x_0)\bar{G}_0(y_0|x_0)\,\mathrm{d}y_0 \\
&P_t(x, y; \mathrm{d}x', \mathrm{d}y') = Q_t(x, \mathrm{d}x')\bar{G}_t(y'|x, y, x')\,\mathrm{d}y',
\end{aligned}$$

where Q_t is transition kernels and η_0 is a probability. The conditional distribution of X_t given X_{t-1} is independent of Y_{t-1}.

B.2

$$P\left(Y_t \in \mathrm{d}y' | \left(X_t, X_{t-1}, Y_{t-1} = (x', x, y)\right)\right) = \bar{G}_t\left(y'|x, y, x'\right)\mathrm{d}y',$$

where $\bar{G}_t$ is bounded.

The simultaneous distribution is

$$\begin{aligned}
P\left(X_{0:t} \in \mathrm{d}x_{0:t}, Y_{0:t} \in \mathrm{d}y_{0:t}\right) &= P\left(\cap_{k=0}^{t}\left((X_k, Y_k) \in (\mathrm{d}x_k, \mathrm{d}y_k)\right)\right) \\
&= \mu_0(\mathrm{d}x_0, \mathrm{d}y_0)\prod_{k=1}^{t} P_k\left((x_{t-1}, y_{t-1}), (\mathrm{d}x_k, \mathrm{d}y_k)\right) \\
&= \eta_0(\mathrm{d}x_0)\bar{G}_0\left(y_0|x_0\right)\prod_{k=1}^{t} Q_t(x_{t-1}, \mathrm{d}x_t)\bar{G}(y_k|x_{k-1}, y_{k-1}, x_k)\,\mathrm{d}y_k \\
&= P\left(X_{0:t} \in \mathrm{d}x_{0:t}\right)\prod_{k=0}^{t}\bar{G}_k\left(y_k|x_{k-1}, y_{k-1}, x_k\right)\mathrm{d}y_k \\
&= P\left(X_{0:t} \in \mathrm{d}x_{0:t}\right)P\left(Y_{0:t} \in \mathrm{d}y_{0:t}|X_{0:t} \in x_{0:t}\right).
\end{aligned}$$

The first idea is to consider the equations that sequentially update the distribution $\eta_t : t \geq 0$ which are of the form

$$\eta_t = \Phi_t(\eta_{t-1}) \tag{4.10}$$

with continuous mappings Φ_t on the set $\mathcal{P}(\mathbf{E})$ of all probability measures on $\mathbf{E}$. The NLF problem will then be reduced to the problem of solving a dynamical system taking values in the infinite dimensional state-space $\mathcal{P}(\mathbf{E})$. In this sense it is natural to approximate η_t for $t \geq 1$ by a sequence of empirical measures

$$\eta_t^N = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_t^i} \tag{4.11}$$

, where δ_a is the Dirac measure at $a \in \mathbf{E}$, associated with a system of N interacting particles $X_t = (X_t^{(1)}, ..., X_t^{(N)})$ moving in the set $\mathbf{E}$. In view of (4.10) it is natural to construct the sequence $\{X_t^{(i)}\}_{i=1}^N$ as a Markov chain taking values in $\mathbb{E}^N$, starting with the initial distribution $\tilde{\eta}_0$ and 1-step probabilities $\tilde{Q}_t$ given by

$$\tilde{\eta}_0(\mathrm{d}x) = \prod_{p=1}^{N} \eta_0(\mathrm{d}x^p),$$

$$\tilde{Q}_t(z, \mathrm{d}x) = \prod_{p=1}^{N} \Phi_t(m(z))(\mathrm{d}x^p)$$

, where $\mathrm{d}x = \mathrm{d}x^{(1)} \times ... \times \mathrm{d}x^N$ is an infinitesimal neighbourhood of the point $x = (x^{(1)}, ..., x^{(N)}) \in \mathbf{E}^N, Z = (Z^{(1)}...Z^{(N)}) \in \mathbf{E}^N$ and where $m(Z) = \frac{1}{N}\sum_{i=1}^N \delta_{Z^{(i)}}$ is the empirical distribution associated with the variable Z. The reason for this is that η_t^N is the empirical measure associated with N independent variables with common law $\Phi_t(\eta_{t-1}^N)$, so as soon as η_{t-1}^N is a good approximation of η_{t-1} then, by (4.10), η_t^N should be a good approximation of η_t.

4.4.1 General facts about nonlinear filtering

In this section we will give a quick introduction to some general facts about nonlinear filtering that we will need in the following.

We use the traditional notation for transition kernels; if P and Q are two transition kernels, μ is a measure and f is a measurable function (all on $(\mathbf{E}, \mathcal{E})$) then we have another transition kernel PQ (usually the product or composition), a function Pf, a measure μP and a number μPf. Again we denote $\mathcal{P}(\mathbf{E})$ the set of all probability measures on $\mathbf{E}$.

Recall that the observations $y_0, y_1...$ are given and fixed, and let

$$G_t(x, x') \triangleq \bar{G}_t(y_t | x, y_{t-1}, x'). \tag{4.12}$$

(This is well defined even for $t = 0$ since $\bar{G}_0(y'|x, y, x')$ does not depend on y).

Let us start by studying the marginal distribution of $Y_{0:t}$,

$$\begin{aligned} P\left(Y_{0:t} \in \mathrm{d}y_{0:t}\right) &= \int_X P\left(X_{0:t} \in \mathrm{d}x_{0:t}, Y_{0:t} \in \mathrm{d}y_{0:t}\right) \\ &= \int_X P\left(X_{0:t} \in \mathrm{d}x_{0:t}\right) \prod_{k=0}^{t} \bar{G}_k\left(y_k | x_{k-1}, y_{k-1}, x_k\right) \mathrm{d}y_k \\ &= \int_X P\left(X_{0:t} \in \mathrm{d}x_{0:t}\right) \prod_{k=0}^{t} G_k\left(x_{k-1}, x_k\right) \mathrm{d}y_k \\ &= \mathbb{E}\left[\prod_{k=0}^{t} G\left(X_{k-1}, X_k\right)\right] \mathrm{d}y_{0:t} \end{aligned}$$

which has density

$$P\left(Y_{0:t} \in \mathrm{d}y_{0:t}\right) = \mathbb{E}\left[\prod_{k=0}^{t} G\left(X_{k-1}, X_k\right)\right] \mathrm{d}y_{0:t}.$$

Define

$$\hat{\gamma}_t f \triangleq \mathbb{E}\left[f(X_t)\prod_{k=0}^{t} G_k\left(X_{k-1}, X_k\right)\right]$$

which could be seen as

$$\mathbb{E}\left[f(X_t)1(Y_{0:t} = y_{0:t})\right]$$

so that

$$\frac{\hat{\gamma}_t(f)}{\hat{\gamma}_t(1)} = \mathbb{E}\left[f(X_t)|Y_{0:t} = y_{0:t}\right] = \hat{\eta}_t(f).$$

It follows that

$$\hat{\gamma}_t(f) = \int \eta_0(\mathrm{d}x_0)\prod_{k=0}^{t} Q_k(x_{k-1}, \mathrm{d}x_k)G_k\left(x_{k-1}, x_k\right)f(x_t).$$

In the same way we may define

$$\gamma_t(f) = \mathbb{E}\left[f(X_t)\prod_{k=0}^{t-1} G\left(X_{k-1}, X_k\right)\right]$$

so that for any $t \geq 0$ and fixed observations $y_{0:t}$ we have (analog to (1.2) and (1.4))

$$\eta_t(f) = \frac{\gamma_t(f)}{\gamma_t(1)}, \quad \hat{\eta}_t(f) = \frac{\hat{\gamma}_t(f)}{\hat{\gamma}_t(1)} \tag{4.13}$$

for any measurable function f such that the following expression makes sense

$$\begin{aligned} \gamma_t(f) &= \mathbb{E}\left(f(X_t)\prod_{k=0}^{t-1} G_k(X_{k-1}, X_k)\right) \\ \hat{\gamma}_t(f) &= \mathbb{E}\left(f(X_t)\prod_{k=0}^{t} G_k(X_{k-1}, X_k)\right) \end{aligned} \tag{4.14}$$

with the convention $\prod_{\phi} = 1$. (Notice again that $G_0(x, x')$ does not depend on x so X_{-1} does not appear in the product.)

γ_t and $\hat{\gamma}_t$ can be considered finite positive measures since G_t is bounded by hypothesis. Also we have $\gamma_0 = \eta_0$ and we have seen that $\hat{\gamma}_t(1)(y_{0:t})$ is the density of $(Y_{0:t})$ w.r.t. to the Lebesgue measure on $\mathbb{R}^{q(t+1)}$.

Let us now introduce the function $\hat{L}_t f(x) = \int Q_t(x, dz)G_t(x, z)f(z)$.
From this and (4.14) we get

$$\hat{\gamma}_t = \hat{\gamma}_{t-1}(\hat{L}_t f). \tag{4.15}$$

Next we define the kernels for $0 \leq p \leq t$

$$\hat{L}_{p,t} = \hat{L}_{p+1}\hat{L}_{p+2}...\hat{L}_t$$

and then using standard Markov notation

$$\hat{\gamma}_t f = \eta_0 I_{G_0}\hat{L}_1 \cdots \hat{L}_t f = \eta_0(G_0\hat{L}_{0,t}f) \tag{4.16}$$

since $\hat{\gamma}_0(f) = \eta_0(fG_0)$. From above we then get

$$\hat{\gamma}_t(f) = \hat{\gamma}_p(\hat{L}_{p,t}f). \tag{4.17}$$

It now follows from (4.13) and (4.15) that

$$\begin{aligned}\hat{\eta}_t(f) &= \frac{\hat{\gamma}_t(f)}{\hat{\gamma}_t(1)}\\ &= \frac{\hat{\gamma}_{t-1}(\hat{L}_t f)}{\hat{\gamma}_{t-1}(\hat{L}_t 1)}\end{aligned}$$

again from (4.13) we get

$$\hat{\eta}_t(f) = \frac{\hat{\eta}_{t-1}(\hat{L}_t f)}{\hat{\eta}_{t-1}(\hat{L}_t 1)}.$$

In other words we have

$$\hat{\eta}_t = \hat{\Psi}_t(\hat{\eta}_{t-1}) \text{ where } \hat{\Psi}_t(\eta)(f) = \frac{\eta\hat{L}_t f}{\eta\hat{L}_t 1}.$$

By the Markov property of X we obtain for $t \geq 1$:

$$\begin{aligned}\gamma_t(f) &= \mathbb{E}\left[f(X_t)\prod_{k=0}^{t-1} G_k(X_{k-1}, X_k)\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[f(X_t)\prod_{k=0}^{t-1} G_k(X_{k-1}, X_k)|\mathcal{F}_{t-1}^X\right]\right]\\ &= \mathbb{E}\left[\prod_{k=0}^{t-1} G_k(X_{k-1}, X_k)\mathbb{E}\left[f(X_t)|\mathcal{F}_{t-1}^X\right]\right]\\ &= \mathbb{E}\left[\prod_{k=0}^{t-1} G_k(X_{k-1}, X_k)Q_t f(X_{t-1})\right]\\ &= \hat{\gamma}_{t-1}Q_t f,\end{aligned} \tag{4.18}$$

and by (4.16) we get

$$\gamma_t(f) = \begin{cases} \eta_0(f) & \text{if} \quad t = 0 \\ \eta_0(G_0 \hat{L}_{0,t-1} Q_t f) & \text{if} \quad t \geq 1. \end{cases} \tag{4.19}$$

More facts about case A

Clearly all the above remains true under case A with $\bar{G}_t = \bar{g}_t$ and we set

$$g_t(x) = \bar{g}_t(x, y_t)$$

(4.14) is now reduced to

$$\begin{aligned} \gamma_t(f) &= \mathbb{E}\left(f(X_t) \prod_{k=0}^{t-1} g_k(X_k) \right) \\ \hat{\gamma}_t(f) &= \mathbb{E}\left(f(X_t) \prod_{k=0}^{t} g_k(X_k) \right). \end{aligned} \tag{4.20}$$

Next $\hat{L}_k f$ of (4.15) becomes $Q_t I_{g_k} f$. So that

$$\begin{aligned} \hat{\gamma} f &= \int \eta_0(\mathrm{d}x_0) g_0(x_0) Q_1(x_0, x_1) g_1(x_1) \cdots Q_t(x_{t-1}, \mathrm{d}x_t) g_t(x_t) \\ &= \eta_0 I_{g_0} Q_1 I_{g_1} \cdots Q_t I_{g_t} f. \end{aligned}$$

For γ_t we can set

$$L_{p,t} = L_{p+1} L_{p+2} ... L_t, \text{ where } L_t f(x) = g_{t-1}(x) Q_t f(x) \tag{4.21}$$

for $0 \leq p \leq t$ with the convention $L_{t,t} = Id$, we get from (4.19) that $\gamma_0 = \eta_0$, and

$$\gamma_t(f) = \gamma_{t-1}(L_t f),$$

where $Q_0(x_{-1}, \mathrm{d}x_0) = \eta_0(\mathrm{d}x_0)$. We easily see that $\gamma_t = \gamma_p L_{p,t}$. Also by the definition of $\hat{\gamma}_t f$ and $\gamma_t f$ we have $\hat{\gamma}_t f = \gamma_t(f g_t)$ and we end up with

$$\begin{aligned} \gamma_0 &= \eta_0 \\ \gamma_t &= \gamma_p L_{p,t} \\ \hat{\gamma}_t f &= \gamma_t(f g_t). \end{aligned} \tag{4.22}$$

Next we introduce the mappings $(\Psi_t)_{t\geq 0}$ and $(\Phi_t)_{t\geq 1}$ from $\mathcal{P}(\mathbf{E})$ into itself by

$$\begin{aligned}\Psi_t(\eta)(f) &= \frac{\eta(fg_t)}{\eta g_t},\\ \Phi_t(\eta)(f) &= \Psi_{t-1}(\eta)Q_t f = \frac{\eta L_t f}{\eta L_t 1}\end{aligned} \tag{4.23}$$

and we get the following lemma

Lemma 4.7

The prediction and filter measures η_t and $\hat{\eta}_t$ satisfies the recursions

$$\begin{aligned}\hat{\eta}_t &= \Psi_t(\eta_t)\\ \eta_{t+1} &= \Phi_{t+1}(\eta_t)\end{aligned}$$

Proof:

$$\begin{aligned}\hat{\eta}_t f &= \frac{\hat{\gamma}_t f}{\hat{\gamma}_t 1} = \frac{\gamma_t(fg_t)}{\gamma_t g_t} = \frac{\eta_t(fg_t)}{\eta_t g_t} = \Psi_t(\eta_t)(f),\\ \eta_{t+1} f &= \hat{\eta}_t(Q_t f) = \Psi_t(\eta_t)Q_t f = \Phi_{t+1}(\eta_t)(f).\end{aligned}$$

□

Now in view of (4.18), (4.22), (4.13) and using that $Q_t 1 = 1$, we get

$$\gamma_{t+1}1 = \hat{\gamma}_t 1 = \gamma_t g_t = \eta_t g_t \gamma_t 1,$$

and for ≥ 0 we finally have

$$\begin{aligned}\gamma_t 1 &= \prod_{p=0}^{t-1}(\eta_p g_p)\\ \gamma_t f &= (\eta_t f)\prod_{p=0}^{t-1}(\eta_t g_p).\end{aligned} \tag{4.24}$$

4.5 An interacting particle system under case A

Subcase A1

We now turn to a special case of A, the one where we know all the densities $\bar{g}_t$ (hence all functions g_t) as well as η_0 and Q_t, and we also assume that we now how to draw random variables according to the laws η_0 and Q_t $\forall x \in \mathbf{E}$ and all $t \geq 1$.

In this situation we actually have two particle systems, each of size N, at time t. First we have the N random variables $\{X_t^i\}_{i=1}^N$ to approximate η_t by means of the empirical measure η_t^N given by (4.11). Next N random variables $\{\hat{X}_t\}_{i=1}^N$ are used to approximate $\hat{\eta}_t$. The mechanism of these particles can be decomposed into two separate mechanisms $X_t \longrightarrow \hat{X}_t \longrightarrow X_{t+1}$.

Let us also define, as before, $\mathcal{F}_t^X$ to be the σ-field generated by the variables X_p, $p \leq t$ and $\hat{X}_p$ for $p < t$, while $\hat{\mathcal{F}}_t^X$ is the σ-field generated by X_p and $\hat{X}_p$ for $p \leq t$.

The first step is to draw the variables X_0^i independently according to the initial law η_0.The mechanism then proceeds, for all t, according to the following two steps Markov rule.

Mutation/Prediction

$$P(X_{t+1} \in \mathrm{d}z_1, ..., \mathrm{d}z_N | \hat{\mathcal{F}}_t^X) = \prod_{p=1}^N Q_{t+1}(\hat{X}_t^{(p)}, \mathrm{d}z_p).$$

Selection/Updating

$$P(\hat{X}_t \in (\mathrm{d}x_1, ..., \mathrm{d}x_N | \mathcal{F}_t^X) = \prod_{p=1}^N \sum_{i=1}^N \frac{g_t(X_t^{(i)})}{\sum_{j=1}^N g_t(X_t^{(j)})} \delta_{X_t^{(i)}}(\mathrm{d}x_p).$$

In the selection step at time t, we update the positions of the particles according to the fitness function g_t. This is done by resampling, were we draw randomly from the set $X_t = (X_t^{(1)}, ... X_t^{(N)})$, with probability

$$P(\hat{X}_t^{(k)} = X_t^{(i)} | \mathcal{F}_t^X) = \frac{g_t(X_t^{(i)})}{\sum_{j=1}^N g_t(X_t^{(j)})}$$

for $1 \leq i \leq N$. In other words we reproduce our sample by selecting the most fit individuals corresponding to the observation y_t.

In the mutation step we allow the particles to move according to the given transition probability kernel.

The selection and mutation steps approximate the two step iterative structure of the conditional distribution of X_t given $Y_{0:t}$

$$\eta_t \xrightarrow{updating} \hat{\eta}_t \xrightarrow{prediction} \eta_{t+1}$$

by a two step Markov chain taking values in the set of finitely discrete probability measures

$$\eta_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_t^{(i)}} \xrightarrow{selection} \frac{1}{N} \sum_{i=1}^N \delta_{\hat{X}_t^{(i)}} \xrightarrow{mutation} \eta_{t+1}^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_{t+1}^{(i)}}.$$

In view of (4.23) and Lemma 4.7, we have

$$\hat{\eta}_t^N = \sum_{i=1}^{N} \frac{g_t(X_t^{(i)})}{\sum_{j=1}^N g_t(X_t^{(j)})} \delta_{X_t^{(j)}} = \Psi_t \left(\frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^{(i)}} \right) = \Psi_t(\eta_t^N) \in \mathcal{F}_t^X, \tag{4.25}$$

so conditional on $\mathcal{F}_t^X$ the variables $\{\hat{X}_t^{(i)}\}_{i=1}^N$ are iid with law $\Psi_t(\eta_t^N)$.

Lemma 4.8

In our sampling procedure we have the following expectation and variances with respect to $\hat{\mathcal{F}}_{t-1}^X$

i) $\mathbb{E}\left[f(X_t^{(i)})|\hat{\mathcal{F}}_{t-1}^X\right] = Q_t f(\hat{X}_{t-1}^{(i)})$.

ii) $\mathrm{Var}\left[f(X_t^{(i)})|\hat{\mathcal{F}}_{t-1}^X\right] = Q_t f^2(\hat{X}_{t-1}^{(i)}) - \left(Q_t f(\hat{X}_{t-1}^{(i)})\right)^2$.

iii) $\mathrm{Var}\left[\sum_{i=1}^N f(X_t^{(i)})\hat{\mathcal{F}}_{t-1}^X\right] = \sum_{i=1}^N \mathrm{Var}\left[f(X_t^{(i)})|\hat{\mathcal{F}}_{t-1}^X\right]$.

Proof: By the sampling procedure $X_t^{(i)} \sim Q_t(\hat{X}_{t-1}^{(i)}, \cdot)$ given $\hat{\mathcal{F}}_{t-1}^X$ and (1)-(ii) hold. Since the particles at time t are conditionally independent given $\hat{\mathcal{F}}_{t-1}^X$, (iii) is true. □

Also note that from (4.25) $\hat{\eta}_{t-1}$ is $\mathcal{F}_{t-1}^X$-measurable.

Lemma 4.9

In our sampling procedure we have the following expectation and variances with respect to $\mathcal{F}_{t-1}^X$

i) $\mathbb{E}\left[f(X_t^{(i)})|\mathcal{F}_{t-1}^X\right]\hat{\eta}_{t-1}^N Q_t f$.

ii) $\mathrm{Var}\left[f(X_t^{(i)})|\mathcal{F}_{t-1}^X\right] = \hat{\eta}_{t-1}^N Q_t f^2 - \left(\hat{\eta}_{t-1} Q_t f\right)^2 = \hat{\eta}_{t-1}^N Q_t \left(f - \hat{\eta}_{t-1}^N Q_t f\right)^2$.

iii) $\mathrm{Var}\left[\sum_{i=1}^N f(X_t^{(i)})|\mathcal{F}_{t-1}^X\right] = \sum_{i=1}^N \mathrm{Var}\left[f(X_t^{(i)})|\mathcal{F}_{t-1}^X\right] = N\hat{\eta}_{t-1}^N Q_t \left(f - \hat{\eta}_{t-1}^N Q_t f\right)^2$.

Proof: By the sampling procedure $X_t^{(i)} \sim Q_t(\hat{X}_{t-1}^{(i)}, \cdot)$ given $\hat{\mathcal{F}}_{t-1}^X$ and

$$\hat{X}_{t-1}^{(i)} \sim \hat{\eta}_{t-1}^N$$

given $\mathcal{F}_{t-1}^X$. Thus

$$\begin{aligned} \mathbb{E}\left[f(X_t^{(i)})|\mathcal{F}_{t-1}^X\right] &= \mathbb{E}\left[Q_t f(\hat{X}_{t-1}^{(i)})|\mathcal{F}_{t-1}^X\right] \\ &= \hat{\eta}_{t-1}^N Q_t f \end{aligned}$$

so that (i)-(ii) hold.

Since the particles at time t are conditionally independent given $\mathcal{F}_{t-1}^X$, (iii) is true. □

Lemma 4.10

The empirical measures η_t^N and $\hat{\eta}_t^N$ satisfy

$$\Phi_{t+1}(\eta_t^N) = \hat{\eta}_t^N Q_t.$$

Proof:

$$\begin{aligned}\Phi_{t+1}(\eta_t^N) &= \frac{\eta_t^N I_{g_t} Q_{t+1}}{\eta_t^N I_{g_t} 1} \\ &= \eta_{t-1}^N [\frac{I_{g_t}}{\eta_t^N I_{g_t} 1}] Q_{t+1} \\ &= \int \eta_t^N (\mathrm{d}x) \tilde{g}(x) Q_{t+1}(x, \cdot) \\ &= \sum_{i=1}^N \tilde{w}_t^{(i)} \delta_{X_t^{(i)}} Q_{t+1}(X_t^{(i)}, \cdot) \\ &= \hat{\eta}_t Q_{t+1},\end{aligned}$$

where $\tilde{g}(x) = \frac{g_t(x)}{\eta_t^N I_{g_t} 1}$. □

Since $\eta_t \in \mathcal{F}_t$ we have using Lemma 4.10

$$\mathbb{E}[\eta_{t+1}^N | \mathcal{F}_t^X] = \mathbb{E}[\Phi(\eta_t^N) | \mathcal{F}_t^X] = \Phi(\eta_t^N). \tag{4.26}$$

The variables $\{X_{t+1}^i\}_{i=1}^N$ are iid given $\mathcal{F}_t^X$ with law $\Phi_{t+1}(\eta_t^N)$, hence, if we define

$$\delta_{t+1}^N f = \eta_{t+1}^N f - \Phi_{t+1}(\eta_t^N) f$$

we have from (4.26) and Lemma 4.9 for $t \geq 0$:

$$\mathbb{E}\left[\delta_{t+1}^N f | \mathcal{F}_t^X\right] = 0,$$

$$\begin{aligned}
\mathbb{E}\left[(\delta^N_{t+1}f)^2|\mathcal{F}^X_t\right] &= \mathbb{E}\left[(\eta^N_{t+1}f - \Phi_{t+1}(\eta^N_t)f)^2|\mathcal{F}^X_t\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N f(X^{(i)}_{t+1}) - \Phi_{t+1}(\eta^N_t)f\right)^2 |\mathcal{F}^X_t\right] \\
&= \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[f(X^{(i)}_{t+1}) - \Phi_{t+1}(\eta^N_t)f|\mathcal{F}^X_t\right]\right) \\
&= \frac{1}{N}\hat{\eta}^N_t Q_{t+1}((f-\Phi_{t+1}(\eta^N_t)f)^2) \\
&= \frac{1}{N}\Phi_{t+1}(\eta^N_t)((f-\Phi_{t+1}(\eta^N_t)f)^2).
\end{aligned} \tag{4.27}$$

Similarly if we define $\delta^N_0 = \eta^N_0 f - \eta_0 f$ we get:

$$\begin{aligned}
&\mathbb{E}(\delta^N_0 f) = 0 \\
&\mathbb{E}((\delta^N_0 f)^2) = \frac{1}{N}\eta_0((f-\eta_0(f))^2)
\end{aligned}$$

Due to the linearity of γ_t, it is easier to study the behaviour of these as $N \to \infty$,for the asymptotic evaluation of our IPS.

Starting with η^N_t above, we can introduce a natural approximation of γ_t from (4.24)

$$\begin{aligned}
\gamma^N_t(f) &\triangleq \eta^N_t(f)\prod_{p=0}^{t-1}\eta^N_t(g_p) \\
\gamma^N_t(1) &= \prod_{p=0}^{t-1}\eta^N_p(g_p).
\end{aligned} \tag{4.28}$$

Recalling (4.22) for any bounded measurable function φ, with the conventions $\gamma^N_{-1} = \gamma_0 = \eta_0$, $L_0 = Id$ and $\Phi_0(\eta^N_{-1}) = \eta_0$ we see that

$$\begin{aligned}
\gamma^N_p\varphi - \gamma_p\varphi &= \{\gamma^N_p\varphi - \gamma^N_{p-1}L_p\varphi\} + \{\gamma^N_{p-1}L_p\varphi - \gamma_{p-1}L_p\varphi\} \\
&= \{\gamma^N_p\varphi - \gamma^N_{p-1}L_p\varphi\} + \{\gamma^N_{p-1}L_p\varphi - \gamma^N_{t-2}L_{p-1}L_p\varphi\} + \{\gamma^N_{t-2}L_{p-1}L_p\varphi - \gamma_{t-2}L_{p-1}L_p\varphi\} \\
&= \sum_{p=0}^q\{\gamma^N_p(L_{p,q}\varphi) - \gamma^N_{p-1}(L_pL_{p,q}\varphi)\}.
\end{aligned}$$

Also from (4.28) we get $\gamma^N_p(L_{p,q})\varphi = \eta^N_p(L_{p,q})\varphi)\gamma^N_p(1)$ using this along with Lemma 4.10 and that $\Phi_t(\eta^N_{-1}) = \eta_0$ we have

$$\gamma^N_q(\varphi) - \gamma_q(\varphi) = \sum_{p=0}^q \gamma^N_p(1)(\eta^N_p(L_{p,q}\varphi) - \Phi_p(\eta^N_{p-1})(L_{p,q}\varphi)).$$

Now choosing $\varphi = L_{q,t}f$, we can define

$$\begin{aligned} M_q^N(f) &= \gamma_q^N(L_{q,t}f) - \gamma_q(L_{q,t}f) \\ &= \sum_{p=0}^{q} \gamma_p^N(1)(\eta_p^N(L_{p,t}f) - \Phi_p(\eta_{p-1}^N)(L_{p,t}f)), \end{aligned}$$

and from (4.27) we see that

$$\mathbb{E}\left[\gamma_q^N f - \gamma_q f | \mathcal{F}_t^X\right] = \sum_{p=0}^{q} \mathbb{E}\left[\gamma_p^N(1)(\eta_p^N(L_{p,t}f) - \Phi_p(\eta_{p-1}^N)(L_{p,t}f))|\mathcal{F}_{p-1}^X\right] = 0 \tag{4.29}$$

so that $M_q^N(f)$ is a martingale.
Anglebracket is the sum of successive conditional variance contribution and since

$$\gamma_p^N 1 = \prod_{k=0}^{t-1} \eta_k(g_k) \in \mathcal{F}_{t-1}^X,$$

we get

$$\begin{aligned} &\sum_{p=0}^{q} \mathbb{E}\left[\gamma_q^N(1)(\eta_p^N(L_{p,t}f) - \Phi_p(\eta_{p-1}^N)(L_{p,t}f))^2|\mathcal{F}_{p-1}^X)\right] \\ &= \frac{1}{N}\sum_{p=0}^{q}(\gamma_p^N(1))^2\Phi_p(\eta_{p-1}^N)((L_{p,t}f - \Phi_p(\eta_{p-1}^N)(L_{p,t}f))^2), \end{aligned} \tag{4.30}$$

so that $M_q^N(f)$ is a martingale with anglebracket

$$\langle M^N(f)\rangle_q = \frac{1}{N}\sum_{p=0}^{q}(\gamma_p^N(1))^2\Phi_p(\eta_{p-1}^N)((L_{p,t}f - \Phi_p(\eta_{p-1}^N)(L_{p,t}f))^2).$$

Taking expectations on both sides of (4.29) and (4.30) with $q = t$ we get

$$\begin{aligned} &\mathbb{E}\left[\gamma_t^N f\right] = \gamma_t f, \\ &\mathbb{E}\big[(\gamma_t^N f - \gamma_t f)^2\big] \\ &= \frac{1}{N}\sum_{p=0}^{t} \mathbb{E}\big[(\gamma_p^N(1))^2\Phi_p(\eta_{p-1}^N)((L_{p,t}f - \Phi_p(\eta_{p-1}^N)(L_{p,t}f))^2)\big]. \end{aligned} \tag{4.31}$$

Theorem 4.11

> For any $f \in \mathbf{B}(\mathbf{E})$
> the rate of convergence of $\hat{\gamma}_t^N f$ to $\gamma_t f$ is $N^{-\frac{1}{2}}$.

Proof: Denote a_t =supnorm of g_t, and $\|f\|$ to be the sup norm of f Then (4.21) gives us

$$\|L_{p,t}f\| \leq \prod_{k=p}^{t-1} a_k \|f\|$$

and (4.24) gives

$$\gamma_t^N(1) \leq \prod_{k=0}^{t-1} a_k$$

Finally we have

$$\begin{aligned} \mathbb{E}\left[(\gamma_t^N(f) - \gamma_t(f))^2\right] &\leq \left(\frac{1}{N}\sum_{p=0}^{t}(\prod_{k=0}^{p-1} a_k)^2 \|f\|^2 (\prod_{j=p}^{t-1} a_j)^2\right) \\ &= \frac{1}{N}(\sum_{p=0}^{t}(\prod_{k=0}^{t-1} a_k)^2 \|f\|^2) \\ &= \frac{1}{N}(t+1)(\prod_{k=0}^{t-1} a_k)^2 \|f\|^2 \\ &= C_t \frac{\|f\|^2}{N} \end{aligned} \tag{4.32}$$

where $C_t = (t+1)(\prod_{k=0}^{t-1} a_k)^2$. □

The above inequality can be used to prove the following results.

Proposition 4.12

There exist constants C_t^1, C_t^2, C_t^3, which depend on t and on the observed values $y_0, ..., y_t$, such that

$$\mathbb{E}|\eta_t^N f - \eta_t f| \leq C_t^1 \frac{\|f\|}{\sqrt{N}}$$

$$\mathbb{E}|\hat{\eta}_t^N f - \hat{\eta}_t f| \leq C_t^1 \frac{\|f\|}{\sqrt{N}}$$

$$P(|\eta_t^N f - \eta_t f| > \epsilon) \leq C_t^2 \exp - \frac{N\epsilon^2}{C_t^3 \|f\|^2}$$

$$P(|\hat{\eta}_t^N f - \hat{\eta}_t f| > \epsilon) \leq C_t^2 \exp - \frac{N\epsilon^2}{C_t^3 \|f\|^2}.$$

Proof: see Del Moral and Guionnet (1998)

The constants above are important when we look at the rates of convergence. In (4.32) we have a precise estimate of C_t, which is quite bad. Unfortunately the constants in proposition 4.12 are even worse. However, these estimates were obtained using very course majorations The estimate of the error may be better represented by central theorems. A full discussion will not be given here but in view of (4.31) it may be reasonable to assume that if we define $U_t^N(f) = \sqrt{N}\,(\gamma_t^N f - \gamma_t f)$, the sequence of random variables $U_t^N(f)$ converges in law, as $N \to \infty$, to a centred Gaussian variable $U_t(f)$ with variance

$$\mathbb{E}\left[U_t(f)^2\right] = \sum_{p=0}^{t} (\gamma_p(1))^2 \eta_p \left((L_{p,t}f - \eta_p L_{p,t}f)^2\right)$$

for any $f \in \mathbf{B}_b(\mathbf{E})$ (the set of all bounded measurable functions on $(\mathbf{E}, \mathcal{E})$. However the errors of $\eta_t^N f - \eta_t f$ and $\hat{\eta}_t^N f - \hat{\eta}_t f$ are of more interest to us. Therefor we can define a sequence of random variables

$$W_t^N(f) = \sqrt{N}\,(\eta_t^N - \eta_t).$$

Now by (4.13), (4.23), Lemma 4.7 and (4.22) , we have that

$$W_t^N(f) = \frac{1}{\gamma_t^N(1)} U_t^N(f - \eta_t(f))$$

and

$$\hat{W}_t^N(f) \triangleq \sqrt{N}\,(\hat{\eta}_t^N(f) - \hat{\eta}_t(f)) = \frac{1}{\eta_t^N(g_t)} W_t^N(g_t(f - \hat{\eta}_t(f))).$$

Since $\gamma_t^N(1)$ and $\eta_t^N(g_t)$ converges in probability to $\gamma_t(1)$ and $\eta_t(g_t)$ the above convergence of $U_t^N(f)$ gives us the following central limit theorems.

Theorem 4.13

> For any bounded measurable function f, the sequence of random variables $W_t^N(f)$ converges in law to a centred Gaussian variable $W_t(f)$ whose variance is given by
>
> $$\mathbb{E}W_t(f)^2 = \sum_{p=0}^{t} \left(\frac{\gamma_p(1)}{\gamma_t(1)}\right)^2 \eta_p \left((L_{p,t}(f - \eta_t(f)))^2\right). \tag{4.33}$$
>
> Also, the sequence of random variables $\hat{W}_t^N(f)$ converges in law to the variable
>
> $$\hat{W}_t(f) = \frac{1}{\eta_t(g_t)} W_t(g_t(f - \hat{\eta}_t(f))). \tag{4.34}$$

For more details see Del Moral, Jacod and Protter (2001) and Del Moral and Ledoux (2000) Now let us try to present the variances of $W_t(f)$ and $\hat{W}_t(f)$ in a more tractable way that will com in handy when we study case B. First we recall (4.21). Then by (4.13) and (4.22) we have for $0 \le p \le t$

$$\frac{\gamma_t(1)}{\gamma_p(1)} = \frac{\gamma_p(L_{p,t}1)}{\gamma_p(1)} = \eta_p(L_{p,t}1)$$

so that (4.33) may be written as

$$\mathbb{E}W_t(f)^2 = \sum_{p=0}^{t} \frac{\eta_p\left((L_{p,t}(f-\eta_t(f)))^2\right)}{(\eta_p(L_{p,t}1))}. \tag{4.35}$$

Also from (4.23) and Lemma 4.7 we have that $\eta_t(g_t(f-\hat{\eta}_t(f))) = \hat{\eta}_t(f-\hat{\eta}_t(f)) = 0$. Now by (4.17) and (4.21) we have $L_{p,t}(g_t f) = g_p \hat{L}_{p,t}(f)$, thus

$$\begin{aligned}\hat{\eta}(\hat{L}_{p,t}1) &= \frac{\gamma_p(g_p\hat{L}_{p,t}1)}{\gamma_p(g_p)} = \frac{\eta_p(L_{p,t}g_t)}{\eta_p(g_p)} \\ &= \frac{\gamma_t(g_t)}{\gamma_p(g_p)} = \frac{\eta_t(g_t)\gamma_t(1)}{\eta_p(g_p)\gamma_p(1)} = \frac{\eta_t(g_t)}{\eta_p(g_p)}\eta_p(L_{p,t}1).\end{aligned}$$

Finally we can deduce from (4.34) and (4.35) that

$$\begin{aligned}\mathbb{E}\hat{W}_t(f)^2 &= \mathbb{E}\left[\left(\frac{1}{\eta_t(g_t)}W_t(g_t(f-\hat{\eta}_t(f)))\right)^2\right] \\ &= \sum_{p=0}^{t}\frac{1}{\eta_t(g_t)^2}\frac{\eta_p\left(L_{p,t}(g_t(f-\hat{\eta}_t(f)))^2\right)}{(\eta_p(L_{p,t}1))^2} \\ &= \sum_{p=0}^{t}\frac{\eta_p\left(\left(g_p\hat{L}_{p,t}(f-\hat{\eta}_t(f))\right)^2\right)}{(\eta_p(g_p))^2\left(\hat{\eta}_p(\hat{L}_{p,t}1)\right)^2}.\end{aligned} \tag{4.36}$$

4.5.1 Subcase A2

Subcase A2 is the situation in case A when all the main ingredients, g_t, η_0 and Q_t are not known and/or when we cannot simulate random variables exactly according to the laws η_0 or Q_t. It is quite obvious that we in this situation will replace these with approximated quantities $g_t^{(m)}, \eta_0^{(m)}$ and $Q_t^{(m)}$ such that these are known and we are able to simulate exactly from the laws $\eta_0^{(m)}$ and $Q_t^{(m)}$. The index m (integer) is a measure of the quality of the approximation. In the IPS below, m will depend on the number of particles.

In this case we have to operate with two filter schemes. The first is related two the original setting with our prediction and filtering measures η_t and $\hat{\eta}_t$. The other is related to the approximations $(g_t^{(m)}, \eta_0^{(m)}, Q_t^{(m)})$ with corresponding prediction and filter measures $\eta_t^{(m)}$ and $\hat{\eta}_t^{(m)}$. We will not go into the details of this case, so for a thourough investigation of this case see Del Moral and Jacod (2001). However we will state a proposition under the following three assumptions.

Assumption C.1 There exist a finite signed measure η_0' and a constant C such that

$$\| m(\eta_0^{(m)} - \eta_0) - \eta_0' \|_{\mathbf{tv}} \leq \frac{C}{m}$$

Assumption C.2 For all t there exist a measurable bounded function g'_t on $(\mathbf{E}, \varepsilon)$ and a constant C_t such that

$$| \, m(g_t^{(m)}(x) - g_t(x)) - g'_t(x) \, | \leq \frac{C_t}{m}.$$

Assumption C.3 For all t there exist a finite signed transition measure Q'_t from $(\mathbf{E}, \mathcal{E})$ into itself and a constant C_t such that

$$\| \, m(Q_t^{(m)}(x, \cdot) - Q_t(x, \cdot)) - Q'_t(x, \cdot) \, \|_{\mathbf{tv}} \leq \frac{C_t}{m}.$$

Proposition 4.14

Assume that **C.1**, **C.2** and **C.3** are satisfied and suppose that we conduct our IPS system with the approximating quantities $(\eta_0^{(m(N))}, \bar{g}_t^{(m(N))}, Q_t^{(m(N))})$ where $m(N) = [\sqrt{N}]$, then there exist a constant $C(t)$ such that

$$\mathbb{E}\big|\eta_t^N f - \eta_t f\big| \leq C(t)\frac{\|f\|}{\sqrt{N}}, \qquad \mathbb{E}\big|\hat{\eta}_t^N f - \hat{\eta}_t f\big| \leq C(t)\frac{\|f\|}{\sqrt{N}}.$$

Proof: see Del Moral and Jacod (2001)

4.6 An interacting particle system under Case B

4.6.1 Subcase B1

As in case A1 we introduce case B1, which means that we know the densities $\bar{G}_t$ (and all functions G_t of (4.12)) as well as the initial conditions η_0, μ_0 and the transitions P_t and Q_t, all connected by assumption B1. Also we assume that we are able to simulate exactly according to η_0 and $P_t(x, y; \cdot)$ for all $x \in \mathbf{E}, y \in \mathbb{R}^q$ and $t \geq 0$. In this section we will introduce a filtering scheme as described in Del Moral and Jacod (2001). The main idea is to expand the state space and make an approximation scheme that fits case A. That is, we want to reduce case B to case A. The IPS system will then be easy to implement, and we will take advantage of the convergence properties developed in the previous section. We start by looking at a a new variable $\mathcal{X}_t = (\dot{X}_t, \dot{Y}_t)$ where $\dot{X}_t$ has the same law as X_t and $(\dot{X}_t, \dot{Y}_t)$ the law of (X_t, Y_t) conditioned on the observation y_{t-1}. Then we add some noise $h\nu_t$ to the process $\dot{Y}_t$ that is independent of $\mathcal{X}_t$, $\mathcal{Y}_t = \dot{Y}_t + h\nu_t$, and assume that our observations $y_{0:t}$ are realisations of the process $\{\mathcal{Y}_t\}$. We now carry out a particle scheme as in case A. Then as $h \to 0$ we have particles $\{\hat{\mathcal{X}}_t^{(i)}\}_{i=1}^N$ with high weights given the observations $y_{0:t}$ and the first component $\{\hat{\dot{X}}_t^{(i)}\}_{i=1}^N$ will then be set of particles with the same marginal law as X_t and with high importance weights given our observations. In other words we will study the filter scheme

$$\begin{aligned} &\text{the state process } \mathcal{X}_t = (\dot{X}_t, \dot{Y}_t) \\ &\text{the observation process } \mathcal{Y}_t^{(h)} = \dot{Y}_t + h\nu_t, \end{aligned} \tag{4.37}$$

where the ν_t's are i.i.d. q-dimensional variables, independent of $\mathcal{X}$ and with distribution $\theta(y)\,\mathrm{d}y$.

Let us introduce some notation and explain mathematically the idea.

- Let us denote by $\mathcal{X}_t = (\dot{X}_t, \dot{Y}_t)$, $t \geq 0$ the time in-homogeneous Markov chain with product state-space $\mathbf{E} \times \mathbb{R}^q$, with initial law μ_0 and transition kernels $\{\mathcal{Q}_t; t \geq 1\}$ given by

$$\begin{aligned} &\mu_0(\mathrm{d}x, \mathrm{d}y) = \eta_0(\mathrm{d}x)\bar{G}_0(x, y)\,\mathrm{d}y \\ &\mathcal{Q}_t((x, y), d(x', y')) = Q_t(x, \mathrm{d}x')\bar{G}_t(y'|x, y_{t-1}, x')\,\mathrm{d}y'. \end{aligned} \tag{4.38}$$

  Note that $\mathcal{Q}_t(x, y; \cdot)$ does not depend on y.

- Let θ be a Borel-bounded function from $\mathbb{R}^q$ to $(0, \infty)$ such that

$$\int \theta(y)\,\mathrm{d}y = 1, \quad \int y\theta(y)\,\mathrm{d}y = 0, \quad \int |y|^3\theta(y)\,\mathrm{d}y < \infty.$$

  Then we set for any $h \in (0, \infty)$ and $(x, y) \in \mathbf{E} \times \mathbb{R}^q$

$$g_t^{(h)}(x, y) = h^{-q}\theta\left((y - y_t)/h\right).$$

Under some regularity conditions on the functions $\bar{G}_t$, h-approximating measures for γ_t and $\hat{\gamma}_t$ of 4.20 are the marginals $\gamma_t^{(h)}$ and $\hat{\gamma}_t^{(h)}$ on the first components of the measures $\nu_t^{(h)}$ and $\hat{\nu}_t^{(h)}$ on $\mathbf{E} \times \mathbb{R}^q$ defined for any $\varphi \in \mathbf{B}_b(\mathbf{E} \times \mathbb{R}^q)$ by formulae

$$\nu_t^{(h)}(\varphi) = \mathbb{E}\left[\varphi(\mathcal{X}_t)\prod_{k=0}^{t-1} g_k^{(h)}(\mathcal{X}_k)\right], \quad \hat{\nu}_t^{(h)}(\varphi) = \mathbb{E}\left[\varphi(\mathcal{X}_t)\prod_{k=0}^{t} g_k^{(h)}(\mathcal{X}_k)\right], \tag{4.39}$$

where $\prod_{\phi} = 1$. If we denote by $\nu(f \otimes 1)$ the function evaluated on the first component of ν, we have for any $f \in \mathbf{B}_b(\mathbf{E})$

$$\gamma_t^{(h)}(f) = \nu_t^{(h)}(f \otimes 1), \qquad \hat{\gamma}_t^{(h)}(f) = \hat{\nu}_t^{(h)}(f \otimes 1).$$

Next we introduce the prediction and filtering measures associated with the scheme (4.37) with respect to the observations $y_0, ..., y_t$ for the true model, that is for any measurable function φ on $\mathbf{E} \times \mathbb{R}^q$, we set

$$\begin{aligned} \mu_t^{(h)}(\varphi) &= \mathbb{E}\left[\varphi(\mathcal{X}_t)|\mathcal{Y}_{0:t-1}^{(h)} = y_{0:t-1}\right] = \frac{\nu_t^{(h)}(\varphi)}{\nu_t^{(h)}(1)}, \\ \hat{\mu}_t^{(h)}(\varphi) &= \mathbb{E}\left[\varphi(\mathcal{X}_t)|\mathcal{Y}_{0:t}^{(h)} = y_{0:t}\right] = \frac{\hat{\nu}_t^{(h)}(\varphi)}{\hat{\nu}_t^{(h)}(1)}. \end{aligned} \tag{4.40}$$

The first marginals of these measures, the ones that we are interested in, denoted by $\eta_t^{(h)}$ and $\hat{\eta}_t^{(h)}$, defined for any $f \in \mathcal{B}_b$ by

$$\eta_t^{(h)}(f) = \mu_t^{(h)}(f \otimes 1) = \frac{\gamma_t^{(h)}(f)}{\gamma_t^{(h)}(1)} \quad \hat{\eta}_t^{(h)}(f) = \hat{\mu}_t^{(h)}(f \otimes 1) = \frac{\hat{\gamma}_t^{(h)}(f)}{\hat{\gamma}_t^{(h)}(1)}. \tag{4.41}$$

Note that $\nu_0^{(h)} = \mu_0^{(h)} = \mu_0$ and $\eta_0^{(h)} = \eta_0$.

We will now proceed as in section 4.5, but we will only consider the marginals of $\mathbf{E}$ of the various transition kernels and measures on $(\mathbf{E} \times \mathbb{R}^q)$. We remind the reader that if φ as function on $(\mathbf{E} \times \mathbb{R}^q)$, then $\mathcal{Q}_t\varphi$ can be considered a function on $\mathbf{E}$ as well as on $(\mathbf{E} \times \mathbb{R}^q)$.

As in (4.15) Del Moral and Jacod (2001), define the kernels on $(\mathbf{E} \times \mathbb{R}^q)$ by

$$\hat{L}_t^{(h)} f = \mathcal{Q}_t(g_t^{(h)}(f \otimes 1)) = \int Q_t(x, \mathrm{d}x')\bar{G}_t(y'|x, y_{t-1}, x')g_t^{(h)}(x', y')f(x')\,\mathrm{d}y' \tag{4.42}$$

and their iterates $\hat{L}_{p,t}^{(h)}$ for $0 \le p \le t$ by

$$\hat{L}_{p,t}^{(h)} = \hat{L}_{p+1}^{(h)}...\hat{L}_t^{(h)} \quad \hat{L}_{t,t}^{(h)} = Id. \tag{4.43}$$

Note that $\hat{L}_t^{(h)}\varphi = \mathcal{Q}_t(g_t^{(h)}(\varphi))$. We also defined

$$\hat{G}_0^{(h)}(x) = \int \bar{G}_0(x,y) g_0^{(h)}(x,y)\,\mathrm{d}y.$$

Observe that

$$\begin{aligned}\nu_0^{(h)}(g_0^{(h)}(f\otimes 1)) &= \mathbb{E}_{\nu_0^{(h)}}\left[g_0^{(h)}(x,y)f(x)\right]\\ &= \mathbb{E}_{\eta_0}\left[\int \bar{G}_0(x,y)g_0^{(h)}(x,y)f(x)\right] = \eta_0(\hat{G}_0^{(h)}f)\end{aligned}$$

We can now deduce from (4.17) and (4.19)

$$\begin{aligned}&\nu_0^h = \text{ law of } \mathcal{X}_0 = \text{ law of } (\dot{X}_0, \dot{Y}_0) = \mu_0,\\ &\hat{\nu}_0^{(h)}(\varphi) = \mu_0(g_0^{(h)}\varphi),\\ &\nu_t^{(h)}(\varphi) = \eta_0\left(\hat{G}_0\hat{L}_{0,t-1}^{(h)}\mathcal{Q}_t\varphi\right),\\ &\hat{\nu}_t^{(h)}(\varphi) = \nu_t^{(h)}(g_t^{(h)}\varphi) = \eta_0(\hat{G}_0\hat{L}_{0,t}^{(h)}\varphi).\end{aligned} \tag{4.44}$$

This gives us the marginals on $\mathbf{E}$

$$\begin{aligned}&\gamma_0^{(h)} = \eta_0,\\ &\hat{\eta}_0^{(h)}(f) = \mu_0(g_0^{(h)}(f\otimes 1)) = \eta_0(\hat{G}_0^{(h)}f),\\ &\gamma_t^{(h)}(f) = \eta_0(\hat{G}_0^{(h)}\hat{L}_{0,t-1}^{(h)}Q_t f),\\ &\hat{\gamma}_t^{(h)} = \eta_0(\hat{G}_0^{(h)}\hat{L}_{0,t}^{(h)}f).\end{aligned} \tag{4.45}$$

We now study the convergence of this scheme, that is we evaluate the errors we get by replacing the original scheme by (4.37) in terms of h. To do this we need some regularity assumptions on $\bar{G}_t$.

R.1 The function $y \mapsto \bar{G}_0(x,y)$ is three times differentiable, with partial derivatives of order 1,2 and 3 uniformly bounded in (x,y)

R.2 Setting $\tilde{L}_t f(x,y) = \int Q_t(x,\mathrm{d}x')f(x')\bar{G}_t(y|x,y_{t-1},x')$ for each $t \geq 1$ and each bounded measurable function f on $\mathbf{E}$, $\|f\| < 1$, the function $y \mapsto \tilde{L}_t f(x,y)$ is three times differentiable, with partial derivatives of order 1, 2 and 3 uniformly bounded in (x,y).

R.2 is satisfied if the functions $y \mapsto \bar{G}_t(y|x,y_{t-1},x')$ are three times differentiable, with partial derivatives of order 1, 2 and 3 uniformly bounded in (x,x',y) for each $t \geq 1$. We denote the second order partial derivatives of the functions $y \mapsto \bar{G}_0(x,y)$ and $y \mapsto \tilde{L}_t f(x,y)$ with respect to the components y^j and y^k by $\bar{G}''_{0,j,k}(x,y)$ and $\tilde{L}''_{t,j,k}f(x,y)$ the second order partial derivative of the functions $y \mapsto \bar{G}_0(x,y)$ and $y \mapsto \tilde{L}_t f(x,y)$ with respect to the components y^j and y^k and let us define the finite kernels $\hat{L}_t^*$ and function G_0^* on $\mathbf{E}$ by

$$\hat{L}_t^* f(x) = \frac{1}{2} \sum_{j,l=1}^{q} \tilde{L}''_{t,jl} f(x, y_t) \int \theta(z) z^j z^l \, \mathrm{d}z$$

$$G_0^*(x) = \frac{1}{2} \sum_{j,l=1}^{q} \bar{G}''_{0,lj}(x, y_0) \int \theta(z) z^j z^l \, \mathrm{d}z.$$

Now remembering that $\hat{L}_t$ is defined by (4.15) We have the following lemma:

Lemma 4.15

Under the regularity assumptions R.1 and R.2 we have the following estimates, where C_t denotes a constant that depends only on t and the observations $y_{0:t}$, and f is a bounded measurable function on $\mathbf{E}$

$$\left| \hat{L}_t^{(h)} f(x) - \hat{L}_t f(x) - h^2 \hat{L}_t^* f(x) \right| \leq C_t h^3 \|f\| \tag{4.46}$$

$$\left| \hat{G}_0^{(h)}(x) - G_0(x) - h^2 G_0^*(x) \right| \leq C_t h^3 \tag{4.47}$$

$$\left| h^q Q_t((g_t^{(h)})^2 f)(x) - u \hat{L}_t f(x) \right| \leq C_t h \|f\|, \tag{4.48}$$

where

$$u = \int \theta(y)^2 \, \mathrm{d}y.$$

We will give the proof when $q = 1$ for simplicity.

Proof: Since $\int \theta(y) \, \mathrm{d}y = 1$ and

$$\hat{L}_t f(x) = \int Q_t(x, \mathrm{d}x') \bar{G}_t(y_t | x, y_{t-1}, x') f(x') = \tilde{L}_t f(x, y_t) = \int \tilde{L}_t f(x, y_t) \theta(y) \, \mathrm{d}y$$

we have

$$\begin{aligned}
&\hat{L}_t^{(h)} f(x) - \hat{L}_t f(x) \\
&= \int \int f(x') Q_t(x, \mathrm{d}x') \bar{G}_t(y | x, y_{t-1}, x') g_t^{(h)}(x', y) \, \mathrm{d}y - \int \tilde{L}_t f(x, y_t) \\
&= \int \int f(x') Q_t(x, \mathrm{d}x') \bar{G}_t(y | x, y_{t-1}, x') h^{-1} \theta \left(\frac{y - y_t}{h} \right) - \int \tilde{L}_t f(x, y_t) \\
&= \int \left(\int f(x') Q_t(x, \mathrm{d}x') \bar{G}_t(y_t + hy | x, y_{t-1}, x') \right) \theta(y) \, \mathrm{d}y \\
&= \int \left(\tilde{L}_t f(x, y_t + hy) - \tilde{L}_t f(x, y_{t)} \right) \theta(y) \, \mathrm{d}y.
\end{aligned}$$

Next, by a third order Taylor expansion on the function $y \mapsto \tilde{L}_t f(x,y)$ around y_t.

$$\tilde{L}_t f(x,y) \approx \tilde{L}_t f(x,y_t) + \tilde{L}'_t f(x,y_t)(y-y_t) + \frac{1}{2}\tilde{L}''_t f(x,y_t)(y-y_t)^2 + \frac{1}{6}\tilde{L}'''_t f(x,y_t)(y-y_t)^3$$

and from this we can deduce that

$$\begin{aligned}
&\int \left(\tilde{L}_t f(x,y_t+hy)\right)\theta(y)\,\mathrm{d}y \\
&\approx \int \left(\tilde{L}_t f(x,y_t) + \tilde{L}'_t f(x,y_t)(hy) + \frac{1}{2}\tilde{L}''_t f(x,y_t)(hy)^2 + \frac{1}{6}\tilde{L}'''_t f(x,y_t)(hy)^3\right)\theta(y)\,\mathrm{d}y \\
&= \bar{L}_t f(x,y_t) + h^2\hat{L}^*_t f(x) + \frac{h^3}{6}\int \tilde{L}'''_t f(x,y_t)y^3\theta(y)\,\mathrm{d}y
\end{aligned}$$

such that

$$\int \left(\tilde{L}_t f(x,y_t+hy) - \tilde{L}_t f(x,y_t)\right)\theta(y)\,\mathrm{d}y \approx h^2\hat{L}^*_t f(x) + \frac{h^3}{6}\int \tilde{L}'''_t f(x,y_t)y^3\theta(y)\,\mathrm{d}y,$$

so that

$$\begin{aligned}
\left|\hat{L}^{(h)}_t f(x) - \hat{L}_t f(x) - h^2\hat{L}^*_t f(x)\right| &\leq \left|\frac{h^3}{6}\tilde{L}'''_t f(x,y_t)\int y^3\theta(y)\,\mathrm{d}y\right| \\
&\leq h^3 C_t\|f\|
\end{aligned}$$

since $\tilde{L}'''_t f(x,y_{t)}$ is uniformly bounded and $\int y^3\theta(y)\,\mathrm{d}y < \infty$.

The proof of (4.47) is similar since $G_0(x) = \bar{G}_0(x,y_0)$ and

$$\hat{G}^{(h)}_0(x) - G_0(x) = \int \left(\bar{G}_0(x,y_0+hy)h^2 - \bar{G}_0(x,y_0)\right)\theta(y)\,\mathrm{d}y$$

we get the result by a third order Taylor expansion.of $\bar{G}_0(y|x,y_{t-1},x')$ around y_t.

Finally,

$$\begin{aligned}
&hQ_t((g^{(h)}_t)^2 f)(x) - u\hat{L}_t f(x) \\
&= h\int Q_t(x,\mathrm{d}x')\bar{G}_t(y|x,y_{t-1},x')f(x')h^{-2}\theta\left(\frac{y-y_t}{h}\right)\mathrm{d}y - u\hat{L}_t f(x) \\
&= \int \left(Q_t(x,\mathrm{d}x')\bar{G}_t(y_t+h|x,y_{t-1},x'y)f(x') - \tilde{L}_t f(x,y_t)\right)\theta(y)^2\,\mathrm{d}y \\
&= \int \left(\tilde{L}_t f(x,y_t+hy) - \tilde{L}_t f(x,y_t)\right)\theta(y)^2\,\mathrm{d}y.
\end{aligned}$$

So that (4.48) follows from a first order Taylor expansion of $\tilde{L}_t f(x,y)$ around y_t,

$$\begin{aligned}\left|hQ_t((g_t^{(h)})^2 f)(x) - u\hat{L}_t f(x)\right| &= \left|\int \left(\tilde{L}_t' f(x,y_t)hy\right)\theta(y)^2\,\mathrm{d}y\right| \\ &\le h\int \left|\tilde{L}_t' f(x,y_t)y\theta(y)^2\right|\mathrm{d}y \\ &\le hC_t\|f\|.\end{aligned}$$

□

Next we define the kernels $\hat{L}^*_{p,t}$ for $0 \le p \le t$ and the measures $\gamma^*_t, \hat{\gamma}^*_t, \eta^*_t$ and $\hat{\eta}^*_t$ on $(\mathbf{E}, \mathcal{E})$ by

$$\hat{L}^*_{p,t} = \sum_{q=p+1}^{t} \hat{L}_{p,q-1}\hat{L}^*_q\hat{L}_{q,t}$$

$$\gamma^*_t(f) = \eta_0\left((G^*_0\hat{L}^{(h)}_{0,t-1} + G_0\hat{L}_{0,t-1})Q_t f\right), \quad \eta^*_t(f) = \frac{\gamma^*_t(f)\gamma_t(1) - \gamma^*_t(1)\gamma_t(f)}{(\gamma_t(1))^2} \tag{4.49}$$

for $t \ge 1$ and $\gamma^*_0 = 0 = \eta^*_0$ and for $t \ge 0$

$$\hat{\gamma}^*_t(f) = \eta_0\left((G^*_0\hat{L}^{(h)}_{0,t} + G_0\hat{L}_{0,t})^* f\right), \quad \hat{\eta}^*_t(f) = \frac{\hat{\gamma}^*_t(f)\hat{\gamma}_t(1) - \hat{\gamma}^*_t(1)\hat{\gamma}_t(f)}{(\hat{\gamma}_t(1))^2}. \tag{4.50}$$

We now have the following proposition

Proposition 4.16

Under the assumptions R.1 and R.2 we have the following estimates, where C_t denotes a constant that depends only on t and the observations $y_{0:t}$, and f is a bounded measurable function on $\mathbf{E}$:

$$\left|\hat{L}^{(h)}_{p,t} f(x) - \hat{L}_{p,t} f(x) - h^2\hat{L}^*_{p,t} f(x)\right| \le C_t h^3\|f\| \qquad 0 \le p \le t \tag{4.51}$$

$$\|\gamma^{(h)}_t - \gamma_t - h^2\gamma^*_t\|_{tv} \le C_t h^3 \qquad \|\hat{\gamma}^{(h)}_t - \hat{\gamma}_t - h^2\hat{\gamma}^*_t\| \le C_t h^3 \tag{4.52}$$

$$\|\eta^{(h)}_t - \eta_t - h^2\eta^*_t\|_{tv} \le C_t h^3 \qquad \|\hat{\eta}^{(h)}_t - \hat{\eta}_t - h^2\hat{\eta}^*_t\| \le C_t h^3. \tag{4.53}$$

Proof: For $p = t$ the inequality in (4.51) is trivial by the definition of $\hat{L}^{(h)}_{p,t}$, $\hat{L}_{p,t}$ and $\hat{L}^*_{t,p}$. Now forward by induction, assume that the inequality holds for $p+1$ that is

$$\hat{L}^h_{p+1,t} f = \hat{L}_{p+1,t} f + h^2\hat{L}^*_{p+1,t} f + O(h^3),$$

and by (4.46) we have $\hat{L}^{(h)}_{p+1} f = \hat{L}_{p+1} f + h^2 \hat{L}^*_{p+1} f + O(h^3)$ so by the definition of $\hat{L}^{(h)}_{p,t}$ we have

$$\begin{aligned}
\hat{L}^{(h)}_{p,t} f &= \hat{L}^{(h)}_{p+1} \hat{L}^{(h)}_{p+1,t} f \\
&= \hat{L}_{p+1} \hat{L}^{(h)}_{p+1,t} f + h^2 \hat{L}^*_{p+1} \hat{L}^{(h)}_{p+1,t} f + O(h^3) \\
&= \hat{L}_{p+1} \hat{L}_{p+1,t} f + h^2 \hat{L}_{p+1} \hat{L}^*_{p+1,t} f + h^2 \hat{L}^*_{p+1} \hat{L}_{p+1,t} f + O(h^3) \\
&= \hat{L}_{p+1} \hat{L}_{p+1,t} f + h^2 \left(\hat{L}_{p+1} \hat{L}^*_{p+1,t} + \hat{L}^*_{p+1} \hat{L}_{p+1,t} f \right) + O(h^3) \\
&= \hat{L}_{p,t} f + h^2 \hat{L}^*_{p,t} f + O(h^3).
\end{aligned}$$

The first estimate of (4.52) is trivial for $t = 0$. For $t \geq 1$, comparing with (4.19) and (4.45), gives

$$\begin{aligned}
\gamma^{(h)}_t(f) - \gamma_t(f) &= \eta_0(\hat{G}^{(h)}_0 \hat{L}^{(h)}_{0,t-1} Q_t f) - \eta_0(G_0 \hat{L}_{0,t-1} Q_t f) \\
&= \eta_0 \left((\hat{G}^{(h)}_0 - G_0) \hat{L}^{(h)}_{0,t-1} Q_t f \right) + \eta_0 \left(G_0 (\hat{L}^{(h)}_{0,t-1} - \hat{L}_{0,t-1}) Q_t f \right)
\end{aligned}$$

such that, by (4.46) and (4.47),

$$\begin{aligned}
&\gamma^{(h)}_t(f) - \gamma_t(f) - h^2 \gamma_t(f) \\
&= \eta_0 \left((\hat{G}^{(h)}_0 - G_0) \hat{L}^{(h)}_{0,t-1} Q_t f \right) + \eta_0 \left(G_0 (\hat{L}^{(h)}_{0,t-1} - \hat{L}_{0,t-1}) Q_t f \right) \\
&- h^2 \eta_0 \left((G^*_0 \hat{L}^{(h)}_{0,t-1} + G_0 \hat{L}^*_{0,t-1}) Q_t f \right) \\
&= \eta_0 \left((\hat{G}^{(h)}_0 - G_0 - h^2 G^*_0) \hat{L}^{(h)}_{0,t-1} Q_t f \right) + \eta_0 \left(G_0 (\hat{L}^{(h)}_{0,t-1} - \hat{L}_{0,t-1} - h^2 \hat{L}^*_{0,t-1}) Q_t f \right) \\
&\leq C_t h^3 \|f\|.
\end{aligned}$$

The second inequality in (4.52) is proved similarly using (4.17) instead of (4.19). For (4.53), in view of (4.52), we assume that $\gamma_t(1), \gamma^{(h)}_t(1), \hat{\gamma}_t(1)$ and $\hat{\gamma}^{(h)}_t(1)$ are bigger than some $\epsilon_t > 0$

and we prove the estimate for small enough h. From (4.13),(4.41),(4.51) and (4.52) we have

$$
\begin{aligned}
&\eta_t^{(h)}(f) - \eta_t(f) - h^2\eta_t^*(f) \\
&= \frac{\gamma_t^{(h)}(f)}{\gamma_t^{(h)}(1)} - \frac{\gamma_t(f)}{\gamma_t(1)} - h^2 \left(\frac{\gamma_t^*(f)\gamma_t(1) - \gamma_t^*(1)\gamma_t(f)}{(\gamma_t(1))^2} \right) \\
&= \frac{1}{\gamma_t^{(h)}(1)\,(\gamma_t(1))^2} \Big(\gamma_t^{(h)}(f)\,(\gamma_t(1))^2 - \gamma_t(f)\gamma_t(1)\gamma_t^{(h)}(1) \\
&- h^2\gamma_t^*(f)\gamma_t(1)\gamma_t^{(h)}(1) + h^2\gamma_t^*(1)\gamma_t(f)\gamma_t^{(h)}(1) \Big) \\
&= \frac{1}{\gamma_t^{(h)}(1)\,(\gamma_t(1))^2} \Big((\gamma_t(1))^2 \Big(\gamma_t^{(h)}(f) - \gamma_t(f) - h^2\gamma_t^*(f) \Big) + (\gamma_t(1))^2\gamma_t(f) + h^2(\gamma_t(1))^2\gamma_t^*(f) \\
&- \gamma_t(f)\gamma_t(1)\gamma_t^{(h)}(1) - h^2\gamma_t^*(f)\gamma_t(1)\gamma_t^{(h)}(1) + h^2\gamma_t^*(1)\gamma_t(f)\gamma_t^{(h)}(1) \Big) \\
&= \frac{1}{\gamma_t^{(h)}(1)\,(\gamma_t(1))^2} \Big((\gamma_t(1))^2 \Big(\gamma_t^{(h)}(f) - \gamma_t(f) - h^2\gamma_t^*(f) \Big) \\
&- \gamma_t(1)\gamma_t(f) \Big(\gamma_t^{h}(1) - \gamma_t(1) - h^2\gamma_t^*(1) \Big) \\
&- h^2 \left(\gamma_t^*(1)\gamma_t(f) - \gamma_t(1)\gamma_t^*(f) \right) \Big(\gamma_t(1) - \gamma_t^{(h)}(1) \Big) \Big).
\end{aligned}
$$

Finally, the second inequality in (4.53) is proved similar using the second inequality of (4.52) and (4.50), (4.13) and (4.41). □

4.6.2 The interacting particle system

We can now carry out an IPS for the scheme (4.37) for any given N and $h \geq 0$. The problem of choosing $h = h(N)$ will be addressed later.
As in case A we now have two particle systems. First we have N particles $\mathcal{X}_t = (\mathcal{X}_t^{(i)})_{i=1}^N$ at time t, who's empirical measure μ_t^N are used to approximate $\mu_t^{(h)}$, then, after the next step we have N particles $\hat{\mathcal{X}}_t = (\hat{\mathcal{X}}_t^{(i)})_{i=1}^N$ which are used to approximate $\hat{\mu}_t^{(h)}$.All these particles take their values in $\mathbf{E} \times \mathbb{R}^q$ and we single out their components $X_t^{(i)}$ and $Y_t^{(i)}$ for $\mathcal{X}_t^{(i)}$, and $\hat{X}_t^{(i)}$ and $\hat{Y}_t^{(i)}$ for $\hat{\mathcal{X}}_t^{(i)}$, taking their values in $\mathbf{E}$ and $\mathbb{R}^q$.
The motion of these particles are again defined on an auxiliary probability space, where we denote by $\mathcal{G}_t$ the $\sigma-$field generated by the variables $(\mathcal{X}_p^{(i)})_{i=1}^N$ for $p \leq t$ and $(\hat{\mathcal{X}}_t^{(i)})_{i=1}^N$ for $p < t$ and $\hat{\mathcal{G}}_t$ the $\sigma-$field generated by the variables $(\mathcal{X}_p^{(i)})_{i=1}^N$ and $(\hat{\mathcal{X}}_t^{(i)})_{i=1}^N$ for $p \leq t$. At the initial step, $t = 0$, the variables $(\mathcal{X}_t^{(i)})_{i=1}^N$ are drawn independently according to the initial law μ_0 of (4.38). Then the mechanism proceeds, by induction on t, according to the following two step Markov rule.

Mutation/Prediction

$$\mathbb{P}\left(\mathcal{X}_{t+1} \in (dz_1 \ldots, dz_N)|\hat{\mathcal{G}}_t\right) = \prod_{p=1}^N \mathcal{Q}_{t+1}\left(\hat{\mathcal{X}}_t^{(p)}, dz_p\right). \tag{4.54}$$

Selection/Updating

$$\mathbb{P}\left(\hat{\mathcal{X}}_t \in (dz_1, \ldots, dz_N)|\mathcal{G}_t\right) = \prod_{p=1}^N \sum_{i=1}^N \frac{g_t^{(h)}(\mathcal{X}_t^{(i)})}{\sum_{j=1}^N g_t^{(h)}(\mathcal{X}_t^{(j)})} \delta_{\mathcal{X}_t^{(i)}}. \tag{4.55}$$

For all $t \geq 0$ we have approximating measures μ_t^N and η_t^N for $\mu_t^{(h)}$ and $\eta_t^{(h)}$,

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{(X_t^{(i)}, Y_t^{(i)})}, \qquad \eta_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^{(i)}}.$$

And then, comparing with (4.25), we get the following approximating measures for $\hat{\mu}_t^{(h)}$ and $\hat{\eta}_t^{(h)}$

$$\hat{\mu}_t^N = \sum_{i=1}^N \frac{g_t^{(h)}(\mathcal{X}_t^{(i)})}{\sum_{j=1}^N g_t^{(h)}(\mathcal{X}_t^{(j)})} \delta_{\mathcal{X}_t^{(i)}}, \qquad \hat{\eta}_t^N = \sum_{i=1}^N \frac{g_t^{(h)}(\mathcal{X}_t^{(i)})}{\sum_{j=1}^N g_t^{(h)}(\mathcal{X}_t^{(j)})} \delta_{X_t^{(i)}}.$$

4.6.3 Convergence study

In Del Moral and Jacod (2001) we are presented with a central limit theorem to assert the quality of the IPS system. For each $h > 0$ we have theorem 4.13, which we now recall for the marginals

η_t^N and $\hat{\eta}_t^N$. Then we set

$$W_t^{N,(h)}(f) = \sqrt{N}\left(\eta_t^N(f) - \eta_t^{(h)}(f)\right)$$

and

$$\hat{W}_t^{N,(h)}(f) = \sqrt{N}\left(\hat{\eta}_t^N(f) - \hat{\eta}_t^{(h)}(f)\right).$$

If $h > 0$ is fixed, theorem 4.13 then assures us that as $N \to \infty$ the sequences $\left(W_t^{N,(h)}(f)\right)$ and $\left(\hat{W}_t^{N,(h)}(f)\right)$ converge to centred Gaussian variables $W_t^{(h)}(f)$ and $\hat{W}_t^{(h)}(f)$ with variances given by (recall 4.35 and 4.36)

$$\mathbb{E}W_t^{(h)}(f)^2 = \sum_{p=0}^{t} \frac{\mu_p^{(h)}\left((g_p^{(h)}\hat{L}_{p,t-1}^{(h)}Q_t(f - \eta_t^{(h)}(f)))^2\right)}{\left(\mu_p^{(h)}(g_p^{(h)}(\hat{L}_{p,t-1}^{(h)}(1)))\right)^2} \tag{4.56}$$

and

$$\mathbb{E}\hat{W}_t^{(h)}(f)^2 = \sum_{p=0}^{t} \frac{\mu_p^{(h)}\left((g_p^{(h)}\hat{L}_{p,t}^{(h)}Q_t(f - \hat{\eta}_t^{(h)}(f)))^2\right)}{\left(\mu_p(g_p^{(h)})\right)^2\left(\hat{\eta}_p^{(h)}(\hat{L}_{p,t}^{(h)}1)\right)^2}. \tag{4.57}$$

If we now let $h \to 0$, both quantities (4.56) and (4.57) increase in a way that is controlled by lemma 4.15 and proposition 4.16 , take for example the summon number p in (4.56). For the denominator, we may wright according to (4.40),(4.42),(4.43),(4.44),(4.45) and (4.52)

$$\begin{aligned}\mu_p^{(h)}(g_p^{(h)}\hat{L}_{p,t-1}^{(h)}1) &= \frac{\nu_p^{(h)}(\hat{L}_{p,t-1}^{(h)}1)}{\nu_p^{(h)}(1)} \\ &= \frac{\gamma_t^{(h)}(1)}{\gamma_p^{(h)}(1)} = \frac{\gamma_t(1)}{\gamma_p(1)} + O(h^2)\end{aligned}$$

and the numerator may be written first as

$$\frac{1}{\nu_p^{(h)}(1)}\eta_0\left(\hat{G}_0^{(h)}\hat{L}_{0,p-1}^{(h)}\mathcal{Q}_p\left((g_p^{(h)}\hat{L}_{p,t-1}^{(h)}Q_t(f - \eta_t^{(h)}(f)))^2\right)\right).$$

Next, by (4.47), (4.51) and (4.53), we can replace $\hat{G}_0^{(h)}$, $\hat{L}_{0,p-1}^{(h)}$, $\hat{L}_{p,t-1}^{(h)}$ and $\eta_t^{(h)}$ by G_0, $\hat{L}_{0,p-1}$, $\hat{L}_{p,t-1}$ and η_t to obtain a relative error of size $O(h)$.

Then using (4.48) we see that

$$\begin{aligned}&h^q\mu_p^{(h)}\left(\left(g_p^{(h)}\hat{L}_{p,t-1}^{(h)}Q_t(f - \eta_t^{(h)}(f))\right)^2\right) \\ &= \frac{h^q}{\gamma_p^{(h)}(1)}\eta_0\left(\hat{G}_0^{(h)}\hat{L}_{0,p-1}^{(h)}\mathcal{Q}_p\left(g_t^{(h)}\hat{L}_{p,t-1}^{(h)}Q_t(f - \eta_t^{(h)}(f))\right)^2\right)\end{aligned}$$

converges to

$$\frac{u}{\gamma_p(1)}\eta_0\left(G_0\hat{L}_{0,p}\left((\hat{L}_{p,t-1}Q_t(f-\eta_t(f)))^2\right)\right)$$

as $h \to 0$. From (4.13) and (4.17) we finally see that

$$\begin{aligned}
\lim_{h\to 0} h^q \mathbb{E}\left(W_t^{N,(h)}(f)^2\right) &= u\sum_{p=0}^{t}\frac{1}{\left(\frac{\gamma_t(1)}{\gamma_p(1)}\right)^2}\frac{1}{\gamma_p(1)}\eta_0\left(G_0\hat{L}_{0,p}\left((hatL_{p,t-1}Q_t(f-\eta_t(f)))^2\right)\right)\\
&= u\sum_{p=0}^{t}\frac{\gamma_p(1)}{\gamma_t(1)^2}\hat{\gamma}_p\left((\hat{L}_{p,t-1}Q_t(f-\eta_t(f)))^2\right) \qquad (4.58)\\
&= u\sum_{p=0}^{t}\frac{\gamma_p(1)\gamma_{p+1}(1)}{\gamma_t(1)^2}\hat{\eta}_p\left((\hat{L}_{p,t-1}Q_t(f-\eta_t(f)))^2\right).
\end{aligned}$$

In a similar way we may obtain

$$\lim_{h\to 0} h^q \mathbb{E}\hat{W}_t^{N,(h)}(f)^2 = u\sum_{p=0}^{t}\frac{\gamma_p(1)\gamma_{p+1}(1)}{\hat{\gamma}_t(1)^2}\hat{\eta}_p\left((\hat{L}_{p,t}Q_t(f-\hat{\eta}_t(f)))^2\right). \tag{4.59}$$

Also if we replace $\hat{L}_{0,t-1}^{(h)}$ and $\hat{L}_{0,t}^{(h)}$ by $\hat{L}_{0,t-1}$ and $\hat{L}_{0,t}$ in (4.49) and (4.50) we obtain the new measures $\gamma_t^{**}(f)$, $\hat{\gamma}_t^{**}(f)$, $\eta_t^{**}(f)$ and $\hat{\eta}_t^{**}(f)$, all with a relative error of $O(h^2)$, and then by (4.53) we see that

$$\eta_t^N(f) - \eta_t(f) = \frac{1}{\sqrt{N}}W_t^{N,(h)}(f) + h^2\eta_t^{**}(f) + O(h^3) \tag{4.60}$$

since $h^2\eta_t^*(f) = h^2\eta_t^{**}(f) + O(h^4)$, and a similar expression is valid for $\hat{\eta}_t^N(f) - \hat{\eta}_t(f)$. Now it is obvious $h = h(N)$ should depend on N. The first term of the right hand of (4.60) is of order $1/\sqrt{Nh(N)^q}$ by (4.58), such that the MSE is of order $1/\sqrt{Nh(N)^q} + h(N)^4$ and optimising the choice of $h(N)$ then leads to

$$h(N) = O\left(N^{-\frac{1}{(4+q)}}\right). \tag{4.61}$$

Finally we have the following theorem by Del Moral, Jacod and Protter (2001).

Theorem 4.6.1

Assume R.1 and R.2, and take $h = h(N)$ as given by (4.61) in the procedure given by (4.55) and (4.54). Let f be any bounded measurable function on $\mathbf{E}$.

1. The sequence of variables

$$W_t^N(f) = N^{\frac{2}{4+q}} \left(\eta_t^N(f) - \eta_t(f) \right)$$

converges in law to a Gaussian variable with mean η_t^{**} and variance given by (4.58).

2. The sequence of variables

$$\hat{W}_t^N(f) = N^{\frac{2}{4+q}} \left(\hat{\eta}_t^N(f) - \hat{\eta}_t(f) \right)$$

converges in law to a Gaussian variable with mean $\hat{\eta}_t^{**}$ and variance given by (4.59).

5 A more general filter problem

In this section we will introduce a more general particle filter with corresponding convergence theorems (Crisan and Doucet (2000)). We remove the Markov assumptions on the signal process and the assumption that the observations are conditionally independent upon the signal. The importance sampling step is done using a general transition kernel which can depend on both the observations and the current MC approximation of the posterior distribution. The conditions imposed on the resampling step are also less restrictive. . This method also includes an additional MCMC step in order to address the problem of sample depletion. The convergence results are given on the path space, that is, we prove the convergence to the posterior distribution of the whole trajectory of the signal and not only to the posterior distribution of the current state of the signal.

5.1 Problem statement

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space where we define two vector-valued stochastic processes $X = \{X_t, t \in \mathbb{N}\}$ and $Y = \{Y_t, t \in \mathbb{N}\}$. As before, the X process is the the signal process and Y is the observation process. We also remember that d and q is the dimension of the state space X and Y and that we denote by $X_{i:j}$ and $Y_{i:j}$ the path of the signal and of the observation process from time i to time j and by $x_{i:j}$ and $y_{i:j}$ generic points in the space of the paths of the signal and observation process. The signal process X satisfies $X_0 \sim \eta_0$ and evolves according to the equation

$$P\left(X_t \in A | Y_{0:t-1} = y_{0:t-1}, X_{0:t-1} = x_{0:t-1}\right) = \int_A k_t(y_{0:t-1}, x_{0:t-1}, \mathrm{d}x_t)$$

where k_t is a probability transition kernel defined on $\left(\mathbb{R}^{(d+q)t}, \mathcal{B}\left(\mathbb{R}^d\right)\right)$.

$$k_t : \mathcal{P}\left(\mathbb{R}^{dt}\right) \to \mathcal{P}\left(\mathbb{R}^{d(t+1)}\right).$$

The observation process Y satisfies

$$P\left(Y_t \in B | Y_{0:t-1} = y_{0:t-1}, X_{0:t} = x_{0:t}\right) = \int_B g_t(y_{0:t}|x_{0:t})\, \mathrm{d}y_t,$$

where $B \in \mathcal{B}(\mathbb{R}^q)$. We assume that $Y_{0:t} = y_{0:t}$ is fixed and we let η_t and $\hat{\eta}_t$ be the probability measure of $X_{0:t}$ given $Y_{0:t-1} = y_{0:t-1}$ and $Y_{0:t} = y_{0:t}$ respectively. We also assume that the sequence $(\eta_t, \hat{\eta}_t)$ satisfies the following recurrence formula.

5.1.1 Bayes recursions

For all $t \geq 0$ and $A_i \in \mathcal{B}(\mathbb{R}^d), i = 1, ...t$ $A_{0:t} = A_0 \times A_1 \times ... \times A_t$, we have

$$\text{Prediction} \qquad \eta_t(A_{0:t}) = \int_{A_{0:t-1}} k_t(y_{0:t-1}, x_{0:t-1}, A_t)\hat{\eta}_{t-1}(\mathrm{d}x_{0:t}) \tag{5.1}$$

$$\text{Updating} \qquad \hat{\eta}_t(A_{0:t}) = c_t^{-1} \int_{A_{0:t}} g_t(y_{0:t}, x_{0:t})\eta_t(\mathrm{d}x_{0:t}), \tag{5.2}$$

where c_t is the normalising constant

$$c_t \triangleq \int_{\mathbb{R}^{d(t+1)}} g_t(y_{0:t}|x_{0:t})\eta_t(\mathrm{d}x_{0:t}).$$

Remembering the following notation
If μ is a measure, f is a function and K is a Markov kernel then,

$$\mu f \triangleq \int f\, \mathrm{d}\mu,\ \mu K(A) \triangleq \int \mu(\mathrm{d}x)K(x, A),\ Kf(x) \triangleq \int K(x, \mathrm{d}z)f(z).$$

Using this notation, if $f : \mathbb{R}^{d(t+1)} \to \mathbb{R}$, then the recurrence formula (5.1) and (5.2) implies that, for all $t \in \mathbb{N}$

$$\begin{aligned}&\text{Prediction}\ \eta_t f = \hat{\eta}_t k_t f\\ &\text{Updating}\ \hat{\eta}_t f = \eta_t(f g_t)(\eta_t g_t)^{-1}.\end{aligned}$$

Remark Let us assume that X is a Markov process with respect to the filtration $\mathcal{F}_t^{X,Y} \triangleq \sigma(X_s, Y_s, s \in \{0, t\})$ with transition kernel

$$Q_t(x_{t-1}, A) \triangleq P(X_t \in A|X_{t-1} = x_{t-1})\ A \in \mathcal{B}(\mathbb{R}^d),\ x_{t-1} \in \mathbb{R}^d$$

and that $P(Y_t \in \mathrm{d}y_t | \mathcal{F}_t^X \vee \mathcal{F}_{t-1}^Y) = P(Y_t \in \mathrm{d}y_t | X_t)$, where $\mathcal{F}_t^X \vee \mathcal{F}_t^Y \triangleq \sigma(X_s, Y_s, s \in \{0, t-1\}, X_t)$, and for all $x_t \in \mathbb{R}^d$, the conditional distribution of Y_t given the event $\{X_t = x_t\}$ is absolutely continuous with respect to the Lebesgue measure, in other words there exists $g(y_t|x_t)$ such that for all $x_{t-1} \in \mathbb{R}^d$

$$P(Y_t \in \mathrm{d}y_t | X_t) = g(y_t|x_t)\,\mathrm{d}y_t,$$

then the sequence $(\eta_t, \hat{\eta}_t)$ satisfies the *Bayes' recursions*

5.2 Sequential MC methods

In this section we will present a sequential MC method that at each time t generates N particles $\{\hat{X}_{0:t}^{(i)}\}_{i=1}^N$ with an associated empirical measure $\hat{\eta}_t^N$,

$$\hat{\eta}_t^N(\mathrm{d}x_{0:t}) \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{\hat{X}_{0:t}^{(i)}}(\mathrm{d}x_{0:t})$$

that is close to $\hat{\eta}$. The algorithm evolves sequentially in time, producing $\{\hat{X}_{0:t}^{(i)}\}$ using the observations obtained at time t and the previous set of particles $\{\hat{X}_{0:t-1}^{(i)}\}_{i=1}^N$ produced at time $t-1$ that is close to $\hat{\eta}_{t-1}$.

We also introduce a transition kernel $\Gamma_t(y_{0:t}, x_{0:t-1}, \hat{\eta}_{t-1}, \mathrm{d}x_{0:t})$ which is used to obtain an intermediate set of particles $\{\tilde{X}_{0:t}^{(i)}\}_{i=1}^N$ and we denote by $\tilde{\eta}_t$ the resulting importance distribution

$$\tilde{\eta}_t = \hat{\eta}_{t-1}\Gamma_t.$$

We assume that $\eta_t << \tilde{\eta}_t$ and let h_t be the strictly positive Radon Nikodym derivative $\frac{d\eta_t}{d\tilde{\eta}_t} = h_t$, where $h_t(\cdot) = h_t(y_{0:t}, \hat{\eta}_{t-1}, \cdot)$. Since $\hat{\eta}_t << \eta_t$ by (5.2), $\hat{\eta}_t << \tilde{\eta}_t$ and since (5.2) implies that $\frac{d\hat{\eta}_t}{d\eta_t} \propto g_t$, we have

$$\frac{d\hat{\eta}_t}{d\tilde{\eta}_t} = \frac{d\hat{\eta}_t}{d\eta_t}\frac{d\eta_t}{d\tilde{\eta}_t} \propto g_t h_t. \tag{5.3}$$

Another important assumption is that we now how to sample exactly according to η_0 at time $t = 0$. The algorithm then proceeds as described below.

Algorithm 5.1: General particle filter algorithm

```
Initialisation $t = 0$;
for $i = 1 : N$ do
    sample $X_0^{(i)} \sim \eta_0$;
end
for $t = 1 : T$ do
    for $j = 1 : N$ do
        sample $\tilde{X}_{0:t}^{(i)} \sim \Gamma_t(y_{0:t}, \hat{X}_{0:t-1}^{(i)}, \hat{\eta}_{t-1}^N, d\tilde{x}_{0:t})$;
        Compute $w_t^{(i)} \propto g_t(y_{0:t}|\tilde{X}_{0:t}^{(i)}) h_t(y_{0:t}, \hat{\eta}_{t-1}^N, \tilde{X}_{0:t}^{(i)})$;
    end
    Normalise the importance weights $\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(j)}}$;
    Sample $\{\hat{X}_t^{(i)}\}_{i=1}^N$ from the set $\{X_t^{(i)}\}_{i=1}^N$ with probabilities $\{\tilde{w}_t^{(i)}\}_{i=1}^N$;
end
```

We will later on also include a MCMC step, but first we present an example and then we discuss step 1 and step 2 more in detail and prove the convergence of this scheme to the posterior distribution of the filtering problem.

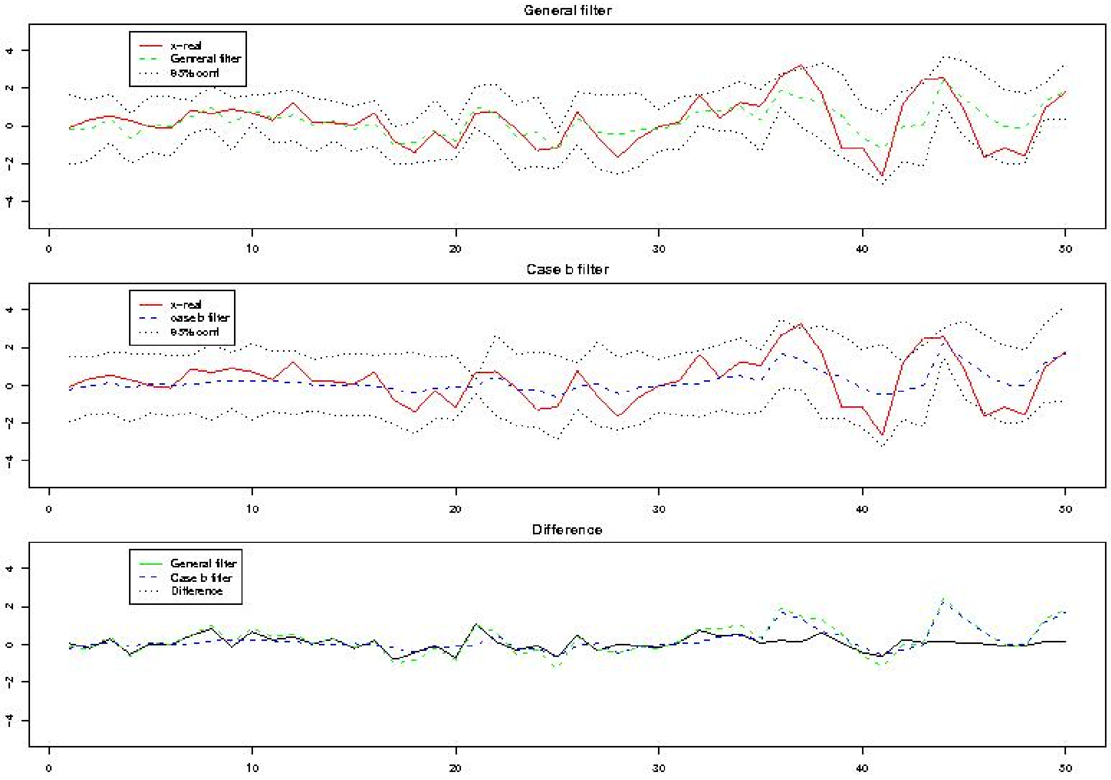
Example 5.1

The dynamics of the system in this example will fit case b. We will use both the method described by Del Moral and Jacod (2001) and the more general algorithm presented by Crisan and Doucet (2000) where we use the laws of X as the importance function, and the selection step is performed by resampling. The system $\{(X_t, Y_t)\}$ evolves simultaneously as a Markov chain, and $\{X_t\}$ as a marginal Markov chain according to

$$\begin{aligned} X_t &= 0.5X_{t-1} + Z_t, \qquad V_t \sim N(0,1) \\ Y_t &= X_t\,|X_{t-1}| + Y_{t-1}W_t \quad W_t \sim N(0,1). \end{aligned}$$

In figure 5.1 we have plotted the results from the two particle filters and their difference for $N = 1000$. We see that both algorithms seems to work well for this case.

Figure 5.1: The particle filters, N=1000



5.2.1 Importance sampling step

In algorithm 5.1 we obtain our new set of paths by sampling from $\Gamma_t(y_{0:t}, \hat{X}^{(i)}_{0:t-1}, \hat{\eta}^N_{t-1}, d\tilde{x}_{0:t})$ which depends on $\hat{\eta}^N_{t-1}$, the observations $y_{0:t}$ and the current paths $\{\hat{X}^{(i)}_{0:t-1}\}_{i=1:N}$. The new paths $\{\tilde{X}_{0:t}\}^N_{i=1}$ are then distributed approximately as $\tilde{\eta}_t$. The only restrictions on the choice of Γ_t is that the weights $w^{(i)}_t$ are well defined and can be computed analytically.However, most algorithms that are presented are such that $\Gamma_t\left(y_{0:t}, \hat{X}^{(i)}_{0:t-1}, \hat{\eta}^N_{t-1}, \mathrm{d}\tilde{x}_{0:t}\right) = \delta_{X^{(i)}_{0:t-1}}(\mathrm{d}\tilde{x}_{0:t-1})\Gamma_t\left(y_{0:t-1}, \hat{x}^{(i)}_{0:t-1}, \hat{\eta}^N_{t-1}, \mathrm{d}\tilde{x}_t\right)$, that is we obtain the new path $\tilde{X}^{(i)}_{0:t}$ by keeping the current path $\hat{X}_{0:t-1}$ and adding a new particle $\tilde{X}_t$. We discussed earlier that a good choice for the importance function is the one that minimises the conditional variance of the importance weights at time t given $\hat{X}^{(i)}_{0:t-1}$ and $y_{0:t}$. Following this strategy, the optimal choice is $P(d\tilde{x}_t|y_{0:t-1}, \hat{x}^{(i)}_{0:t-1})$ (Doucet *et al.*, 2000) the proof is analog to the one on page 14.

Optimal sampling distribution

If we sample $\tilde{X}^{(i)}_{0:t} = (\hat{X}^{(i)}_{0:t-1}, \tilde{X}^{(i)}_t)$ according to

$$\begin{aligned}
P(d\tilde{x}_t | y_{0:t}, \hat{X}^{(i)}_{0:t-1}) &= \frac{P(d\tilde{x}_t, \hat{X}^{(i)}_{0:t-1}, y_{0:t})}{P(y_{0:t}, \hat{X}^{(i)}_{0:t-1})} \\
&= \frac{P(y_t | \hat{X}^{(i)}_{0:t-1}, y_{0:t-1}, d\tilde{x}_t) P(d\tilde{x}_t | y_{0:t-1}, \hat{X}^{(i)}_{0:t-1}) P(\hat{X}^{(i)}_{0:t-1}, y_{0:t-1})}{P(y_t | y_{0:t-1}, \hat{X}^{(i)}_{0:t-1}) P(y_{0:t-1}, \hat{x}^{(i)}_{0:t-1})} \\
&= \frac{g_t(y_{0:t}, \hat{X}^{(i)}_{0:t-1}, \tilde{x}_t) k_t(\hat{X}^{(i)}_{0:t-1}, y_{0:t-1}, d\tilde{x}_t)}{P(y_t | y_{0:t-1}, \hat{X}^{(i)}_{0:t-1})}.
\end{aligned}$$

Since $\hat{\eta}_t \propto g_t k_t$ by (5.2) the importance weights are equal to

$$\begin{aligned}
w^{(i)}_t \propto \frac{d\hat{\eta}_t}{d\tilde{\eta}_t} &\propto P(y_t | y_{0:t-1}, \hat{X}^{(i)}_{0:t-1}) \\
&= \int P(y_t | \hat{X}^{(i)}_{0:t-1}, y_{0:t-1}, \tilde{x}_t) P(d\tilde{x}_t | y_{0:t-1}, \hat{X}^{(i)}_{0:t-1}) \\
&= \int g_t(y_{0:t}, \hat{X}^{(i)}_{0:t-1}, \tilde{x}_t) k_t(y_{0:t-1}, \hat{x}^{(i)}_{0:t-1}, d\tilde{x}_t).
\end{aligned}$$

If this integral does not admit an analytical expression or if we are unable to sample from $P(d\tilde{x}_t | y_{0:t}, \hat{X}^{(i)}_{0:t-1})$, one has to use other alternatives.

Prior distribution

A popular choice for the importance function is the prior distribution
$k_t(y_{0:t-1}, \hat{x}_{0:t-1}, d\tilde{x}_t)$. If we use this then $h_t \propto 1$ so the importance weight $w^{(i)}_t$ is proportional to $g_t(y_{0:t} | \hat{X}^{(i)}_{0:t-1}, \tilde{X}^{(i)}_t)$. The weakness of this method as discussed in (Pitt and Shephard, 1999) is the sensitivity towards outliers. As an alternative one might use the Auxiliary particle filter in algorithm 2.4, or as we now propose, the likelihood distribution.

Likelihood distribution

If we assume that the likelihood $g_t(y_{0:t} | \hat{X}^{(i)}_{0:t-1}, \tilde{x}_t)$ is integrable in the argument $\tilde{x}_t$, that is if

$$\int g_t(y_{0:t} | \hat{X}^{(i)}_{0:t-1}, \tilde{x}_t) d\tilde{x}_t < \infty$$

then we can sample $\tilde{x}_t$ according to

$$\Gamma_t(y_{0:t}, \hat{X}^{(i)}_{0:t-1}, \hat{\eta}^N_{t-1}, d\tilde{x}_t) \propto g_t(y_{0:t} | \hat{X}^{(i)}_{0.t-1}, \tilde{x}_t)$$

and $\frac{d\hat{\eta}_t}{d\tilde{\eta}_t} \propto \frac{\hat{\eta}^N_{t-1}k_t g_t}{\hat{\eta}^N_{t-1}g_t}$ so that the importance weights are proportional to k_t,

$$w_t^{(i)} \propto k_t(y_{0:t-1}, \hat{X}^{(i)}_{0:t-1}, \tilde{X}^{(i)}_t).$$

This is method provides good results when the observation noise is very low as the likelihood is then usually very peaked compared to the prior distribution.

There are several other alternative sampling distribution (Doucet, Godsill and Andrieu (2000),Pitt and Shephard (1999)) but we will now turn our focus towards the resampling step of the algorithm.

5.2.2 Resampling step

As we have discussed earlier, the aim of the selection/resampling step is to multiply the particles with large weight and get rid of those with small weights to obtain an 'unweighted' sample. Next we present three different ways to perform this step, the first has already been discussed in previous chapters.

Sampling Importance Resampling

The SIR or multinomial sampling procedure is the most popular one. We have already discussed it in previous chapters. We sample N particles $\{\hat{X}^{(i)}_{0:t}\}$ with replacement from the particles $\tilde{X}^{(i)}_{0:t}$, that is we sample independently N times from $\tilde{\eta}^N_t(dx_{0:t})$. This is equivalent to jointly drawing $\{N_t^{(i)}\}_{i=1}^N$ according to a multinomial distribution of parameters N and $\tilde{w}_t^{(i)}$. In this case we have $\mathbb{E}N_t^{(i)} = N\tilde{w}_t^{(i)}$ and $\mathrm{Var}N_t^{(i)} = N\tilde{w}_t(i)(1-\tilde{w}_t^{(i)})$.

Residual Resampling

This method, discussed in Carpenter, Clifford and Farnhead (1999) and Higuchi (1997), starts by setting $\acute{N}_t^{(i)} = [N\tilde{w}_t^{(i)}]$ (where $[a]$ denotes the greatest integer smaller then $a \in \mathbb{R}$) then perform an SIR procedure to select the remaining $\bar{N}_t = N - \sum_{i=i}^N \acute{N}_t^{(i)}$ samples with new weights $\acute{w}_t^{(i)}{=}\bar{N}_t^{-1}(\tilde{w}_t^{(i)}N - \acute{N}_t^{(i)})$ and add the result to the current $\acute{N}_t^{(i)}$. In this case $\mathbb{E}N_t^{(i)} = \acute{N}_t^{(i)} + \acute{w}_t^{(i)}\bar{N}_t = \acute{N}_t^{(i)} + \bar{N}_t\bar{N}_t^{(-1)}(\tilde{w}_t^{(i)}N - \acute{N}_t^{(i)}) = N\tilde{w}_t^{(i)}$, but $\mathrm{Var}N_t^{(i)} = \bar{N}_t\acute{w}_t^{(i)}(1-\acute{w}_t^{(i)})$.

Minimal variance sampling

In this procedure a set of U of N points is generated in the interval $[0,1]$, each of the points a distance N^{-1} apart. The number $N_t^{(i)}$ is taken to be the number of points in U that lie between $\sum_{j=1}^{i-1}\tilde{w}_t^{(j)}$ and $\sum_{j=1}^{i}\tilde{w}_t^{(j)}$. This method includes the Tree Based Branching algorithm presented in Crisan (2001) If we denote $\{N\tilde{w}_t^{(i)}\} \triangleq N\tilde{w}_t^{(i)} - [N\tilde{w}_t^{(i)}]$, then the variance of all the algorithms in this class is equal to $\{N\tilde{w}_t^{(i)}\}(1-\{N\tilde{w}_t^{(i)}\})$.

5.3 Convergence study

In this section we first study the convergence of the mean square error of the sequential MC algorithms described in the previous section. That is we will find the rate of which, for any $f \in \mathbf{B}\left(\mathbb{R}^{d(t+1)}\right)$, $\|\hat{\eta}_t^N f - \hat{\eta}_t f\|_2^2 = E\left[(\hat{\eta}_t^N f - \hat{\eta}_t f)^2\right]$ converges to zero under certain conditions (where the expectation is over all realisations of the random particle methods).After that we focus on the almost sure convergence of $\hat{\eta}_t^N$ to $\hat{\eta}_t$ under more restrictive conditions. (As before, the almost sure convergence of a random measure μ^N to the measure μ means that for all $f \in \mathbf{C}_b(\mathbb{R}^d)$, $\lim_{N\to\infty} \mu^N f = \mu f$ a.s.)

In the following we assume that the observation process is fixed to a given observation $Y_{0:t} = y_{0:t}$, $t > 0$. All the convergence results will be proved under this condition.

5.3.1 Bounds for the mean square error

Let us consider the following assumptions.

Importance distribution and weights

i) η_t is absolutely continuous w.r.t $\tilde{\eta}_t \triangleq \hat{\eta}_t \Gamma_t$,

and for all $\mu \in \mathcal{P}(\mathbb{R}^{dt})$ the function $g_t(y_{0:t}|\tilde{x}_{0:t}) h_t(\tilde{x}_{0:t}, y_{0:t}, \mu)$ is a bounded function in argument $\tilde{x}_{0:t} \in \mathbb{R}^{d(t+1)}$.

The identity in equation (5.3) becomes for all $f \in \mathbf{B}(\mathbb{R}^{d(t+1)})$

$$\hat{\eta}_t f = \frac{\tilde{\eta}_t(f g_t h_t)}{\tilde{\eta}_t(g_t h_t)},$$

where $g_t(\cdot) = g_t(y_{0:t}|\cdot)$ and $h_t(\cdot) = h_t(y_{0:t}, \hat{\eta}_{t-1}, \cdot)$. If $\mu, \nu \in \mathcal{P}\left(\mathbb{R}^{dt}\right)$, we define

$$\begin{aligned}
\Gamma_t^\mu &\triangleq \Gamma_t(y_{0:t}, x_{0:t-1}, \mu, d\tilde{x}_{0:t}) \\
\Gamma_t^\nu &\triangleq \Gamma_t(y_{0:t}, x_{0:t-1}, \nu, d\tilde{x}_{0:t}) \\
h_t^\mu(\cdot) &= h_t(y_{0:t}, \mu, \cdot) \\
h_t^\nu(\cdot) &= h_t(y_{0:t}, \nu, \cdot).
\end{aligned}$$

ii) There exists a constant d_t, such that, for all $f \in \mathbf{B}\left(\mathbb{R}^{d(t+1)}\right)$, there exists $f' \in \mathbf{B}\left(\mathbb{R}^{dt}\right)$ with $\|f'\| \leq \|f\|$ such that $\forall \mu, \nu$,

$$\|(\Gamma_t^\mu - \Gamma_t^\nu) f\| \leq d_t |(\mu - \nu) f'|.$$

iii) There exist f_0 (independent of μ, ν) such that

$$\|g_t h_t^\mu - g_t h_t^\nu\| \leq |\mu f_0 - \nu f_0|$$

and a constant e_t such that for any $x_{0:t} \in \mathbb{R}^{d(t+1)}$ and $\forall t \geq 0$

$$|h_t^{\mu}(x_{0:t}) - h_t^{\nu}(x_{0:t})| \leq e_t(h_t^{\mu}(x_{0:t}) \wedge h_t^{\nu}(x_{0:t})).$$

Resampling/selection scheme

iv) $\{N_t^{(i)}\}_{i=1}^N$ are integer valued random variables such that

$$\|\sum_{i=1}^{N} \left(N_t^{(i)} - N\tilde{w}_t^{(i)}\right) q^{(i)}\|_2^2 \leq C_t N \max_{i=1,\dots,N} |q^{(i)}|^2$$

for all N-dimensional vectors $q = (q^{(1)}, q^{(2)}, ..., q^{(N)}) \in \mathbb{R}^N$ and $\sum_{i=1}^N N_t^{(i)} = N$.

The first assumption states that the importance function should be chosen such that the corresponding importance weights are bounded above and that the sampling kernel and importance weights depend continuously on the measure variable. The second assumption ensures that the selection step does not introduce to strong discrepancy. The following establishes, at each time step, a mean square error of order $1/N$ between the empirical measure of the particle filter and the posterior distribution.

Lemma 5.1

Assume that for any $f \in \mathbf{B}\left(\mathbb{R}^{dt}\right)$,

$$\|\hat{\eta}_{t-1}^N f - \hat{\eta}_t f\|_2^2 \leq c_{t-1} \frac{\|f\|^2}{N}.$$

Then, after the first step of the algorithm, for any $f \in \mathbf{B}\left(\mathbb{R}^{d(t+1)}\right)$,

$$\|\tilde{\eta}_t^N f - \tilde{\eta} f\|_2^2 \leq \tilde{c}_t \frac{\|f\|^2}{N}.$$

Proof: Let $\hat{\mathcal{F}}_{t-1}^X$ be the σ-field generated by $\{\hat{X}_{0:t-1}\}_{i=1}^N$, then

$$\mathbb{E}\left[\tilde{\eta}_t^N f_t | \hat{\mathcal{F}}_{t-1}^X\right] = \hat{\eta}_{t-1}^N (\Gamma_t^{\hat{\eta}_{t-1}^N} f_t)$$

and using the independence of the motion of the particles we have (analog to the proof of theorem 4.5)

$$\begin{aligned}
\mathrm{Var}\left[\tilde{\eta}_t^N f | \hat{\mathcal{F}}_{t-1}^X\right] &= \mathbb{E}\left[\left(\tilde{\eta}_t^N f - \hat{\eta}_{t-1}^N (\Gamma_t^{\hat{\eta}_t^N} f)\right)^2 | \hat{\mathcal{F}}_{t-1}^X\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{i=1}^N f(\tilde{X}_{0:t}^{(i)}) - \Gamma_t^{\hat{\eta}_{t-1}^N} f(\hat{X}_{0:t-1}^{(i)})\right)^2 | \hat{\mathcal{F}}_{t-1}^X\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left(f(\tilde{X}_{0:t}^{(i)}) - \Gamma_t^{\hat{\eta}_{t-1}^N} f(\hat{X}_{0:t-1}^{(i)})\right)^2 | \hat{\mathcal{F}}_{t-1}^X\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N \left(\Gamma_t^{\hat{\eta}_{t-1}^N} f^2(\hat{x}_{0:t-1}^{(i)}) - (\Gamma_t^{\hat{\eta}_{t-1}^N} f(\hat{x}_{0:t-1}^{(i)}))^2\right) \\
&= \frac{1}{N}\left(\hat{\eta}_{t-1}^N (\Gamma_t^{\hat{\eta}_{t-1}^N} f^2 - (\Gamma_t^{\hat{\eta}_{t-1}^N} f)^2)\right) \\
&\leq \frac{\|f\|^2}{N}.
\end{aligned}$$

From (ii)

$$\begin{aligned}
\left|\hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f - \hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}} f\right| &= \left|\hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f - \Gamma_t^{\hat{\eta}_{t-1}} f\right| \\
&\leq \|(\Gamma_t^{\hat{\eta}_{t-1}^N} f - \Gamma_t^{\hat{\eta}_{t-1}} f)\| \\
&\leq d_t \left|(\hat{\eta}_{t-1}^N - \hat{\eta}_{t-1}) f'\right|,
\end{aligned}$$

hence

$$\begin{aligned}
\|\hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f - \hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}} f\|_2^2 &\leq d_t^2 \|\hat{\eta}_{t-1}^N f' - \hat{\eta}_{t-1} f'\|_2^2 \\
&\leq d_t^2 c_{t-1} \frac{\|f'\|^2}{N^2} \leq d_t^2 c_{t-1} \frac{\|f\|^2}{N^2}.
\end{aligned}$$

Then, letting $\Gamma_t \triangleq \Gamma_t^{\hat{\eta}_{t-1}}$,

$$\left|\tilde{\eta}_t^N f - \tilde{\eta}_t f\right| \leq \left|\tilde{\eta}_t^N f - \hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f\right| + \left|\hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f - \hat{\eta}_{t-1}^N \Gamma_t f\right| + \left|\hat{\eta}_{t-1}^N \Gamma_t f - \hat{\eta}_{t-1} \Gamma_t f\right|,$$

and from above we get

$$\begin{aligned}
\|\tilde{\eta}_t^N f - \tilde{\eta}_t f\|_2 &\leq \|\tilde{\eta}_t^N f - \hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f\|_2 + \|\hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f - \hat{\eta}_{t-1}^N \Gamma_t f\|_2 + \|\hat{\eta}_{t-1}^N \Gamma_t f - \hat{\eta}_{t-1} \Gamma_t f\|_2 \\
&\leq \sqrt{\tilde{c}_t} \frac{\|f_t\|}{\sqrt{N}},
\end{aligned}$$

where $\tilde{c}_t = (1 + d_t \sqrt{c_{t-1}} + \sqrt{c_{t-1}})^2$. □

Lemma 5.2

Let us assume that for any $f{\in}\mathbf{B}\left(\mathbb{R}^{dt}\right)$ and $f \in \mathbf{B}\left(\mathbb{R}^{d(t+1)}\right)$

$$\|\hat{\eta}^N_{t-1} f - \hat{\eta}_{t-1} f\|_2^2 \leq c_{t-1}\frac{\|f\|^2}{N},$$
$$\|\tilde{\eta}^N_t f - \tilde{\eta}_t f\|_2^2 \leq \tilde{c}_t \frac{\|f\|^2}{N}.$$

Then, for any $f \in \mathbf{B}\left(\mathbb{R}^{d(t+1)}\right)$,

$$\|\bar{\eta}^N_t f - \hat{\eta}_t f\|_2^2 \leq \bar{c}_t \frac{\|f\|^2}{N}.$$

Proof: Let $h^N_t = h_t^{\hat{\eta}^N_{t-1}}$ and $g_t = g$ for simplicity. Then using the fact that $\bar{\eta}^N_t f = \dfrac{\tilde{\eta}^N_t(fgh^N_t)}{\tilde{\eta}^N_t(gh^N_t)}$ and defining

$$A = \left|\frac{\tilde{\eta}^N_t(fgh^N_t)}{\tilde{\eta}^N_t(gh^N_t)} - \frac{\tilde{\eta}^N_t(fgh^N_t)}{\tilde{\eta}^N_t(gh_t)}\right|,$$

we have

$$\begin{aligned}
A &= \frac{\left|\tilde{\eta}^N_t(fgh^N_t)\right|\left|\tilde{\eta}^N_t(g(h^N_t - h_t))\right|}{\tilde{\eta}^N_t(gh^N_t)\tilde{\eta}^N_t(gh_t)} \leq \|f\|\frac{\left|\tilde{\eta}^N_t(g(h^N_t - h_t))\right|}{\tilde{\eta}^N_t(gh_t)} \\
&\leq \|f_t\|\left|\tilde{\eta}^N_t(g(h^N_t - h_t))\right|\left|\frac{1}{\tilde{\eta}^N_t(gh_t)} - \frac{1}{\tilde{\eta}_t(gh_t)}\right| + \|f\|\frac{\left|\tilde{\eta}^N_t(g(h^N_t - h_t))\right|}{\tilde{\eta}_t(gh_t)}.
\end{aligned}$$

Then using (iii)

$$\begin{aligned}
A &\leq \|f\|e_t\tilde{\eta}^N_t(gh_t)\left|\frac{1}{\tilde{\eta}^N_t(gh_t)} - \frac{1}{\tilde{\eta}_t(gh_t)}\right| + \|f\|\frac{\|gh^N_t - gh_t\|}{\tilde{\eta}_t(gh_t)} \\
&\leq \|f\|e_t\tilde{\eta}^N_t(gh_t)\frac{|\tilde{\eta}^N_t(gh_t) - \tilde{\eta}_t(gh_t)|}{\tilde{\eta}^N_t(gh_t)\tilde{\eta}_t(gh_t)} + \|f\|\frac{\left|\hat{\eta}^N_{t-1}f_0 - \hat{\eta}_{t-1}f_0\right|}{\tilde{\eta}_t(gh_t)} \\
&\leq \frac{\|f\|}{\tilde{\eta}_t(gh_t)}\left(e_t\left|\tilde{\eta}^N_t(gh_t) - \tilde{\eta}_t(gh_t)\right| + \left|\hat{\eta}^N_{t-1}f_0 - \hat{\eta}_{t-1}f_0\right|\right).
\end{aligned}$$

Hence,

$$\|A\|_2 \leq \frac{\|f\|\left(e_t\sqrt{\tilde{c}_t}\,\|gh_t\| + \sqrt{c_{t-1}}\,\|f_0\|\right)}{\tilde{\eta}_t(gh_t)\sqrt{N}}. \tag{5.4}$$

Again using (iii),

$$\begin{aligned}\left|\frac{1}{\tilde{\eta}_t^N(gh_t^N)} - \frac{1}{\tilde{\eta}_t^N(gh_t)}\right| &= \frac{\left|\tilde{\eta}_t^N\big(g(h_t^N - h_t)\big)\right|}{\tilde{\eta}_t^N(gh_t^N)\tilde{\eta}_t^N(gh_t)} \\ &\leq \frac{e_t\tilde{\eta}_t(gh_t)}{\tilde{\eta}_t^N(gh_t^N)\tilde{\eta}_t^N(gh_t)} \\ &= \frac{e_t}{\tilde{\eta}_t^N(g_t h_t^N)},\end{aligned}$$

and defining

$$B = \left|\frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t^N(fgh_t)} - \frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t(fgh_t)}\right|,$$

we see that

$$\begin{aligned}B &= \frac{\left|\tilde{\eta}_t^N(gh_t) - \tilde{\eta}_t(gh_t)\right|}{\tilde{\eta}_t(gh_t)}\frac{\left|\tilde{\eta}_t^N(fgh_t^N)\right|}{\tilde{\eta}_t^N(gh_t)} \\ &= \frac{\left|\tilde{\eta}_t^N(gh_t) - \tilde{\eta}_t(gh_t)\right|}{\tilde{\eta}_t(gh_t)}\left|\tilde{\eta}_t^N(fgh_t^N)\right|\left|\frac{1}{\tilde{\eta}_t^N(gh_t)} - \frac{1}{\tilde{\eta}_t^N(gh_t^N)} + \frac{1}{\tilde{\eta}_t^N(gh_t^N)}\right| \\ &\leq \frac{\left|\tilde{\eta}_t^N(gh_t) - \tilde{\eta}_t(gh_t)\right|}{\tilde{\eta}_t(gh_t)}\left|\tilde{\eta}_t^N(fgh_t^N)\right|\left|\frac{1}{\tilde{\eta}_t^N(gh_t)} - \frac{1}{\tilde{\eta}_t^N(gh_t^N)}\right| + \frac{\left|\tilde{\eta}_t^N(gh_t) - \tilde{\eta}_t(gh_t)\right|}{\tilde{\eta}_t(gh_t)\tilde{\eta}_t(gh_t)}\left|\tilde{\eta}_t^N(fgh_t^N)\right| \\ &\leq \frac{\left|\tilde{\eta}_t^N(gh_t) - \tilde{\eta}_t(gh_t)\right|}{\tilde{\eta}_t(gh_t)}\left|\tilde{\eta}_t^N(fgh_t^N)\right|\frac{e_t}{\tilde{\eta}_t^N(gh_t^N)} + \frac{\left|\tilde{\eta}_t^N(gh_t) - \tilde{\eta}_t(gh_t)\right|}{\tilde{\eta}_t(gh_t)\tilde{\eta}_t(gh_t)}\left|\tilde{\eta}_t^N(fgh_t^N)\right| \\ &\leq (e_t + 1)\|f\|\frac{\left|\tilde{\eta}_t^N(gh_t) - \tilde{\eta}_t(gh_t)\right|}{\tilde{\eta}_t(gh_t)}.\end{aligned}$$

We also have, again using (iii),

$$\begin{aligned}\left|\frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t(gh_t)} - \frac{\tilde{\eta}_t^N(fgh_t)}{\tilde{\eta}_t(gh_t)}\right| &\leq \frac{\tilde{\eta}_t(|f||gh_t^N - gh_t|)}{\tilde{\eta}_t(gh_t)} \\ &\leq \|f\|\frac{\left|\hat{\eta}_{t-1}^N(f_0) - \hat{\eta}_{t-1}(f_0)\right|}{\tilde{\eta}_t(gh_t)},\end{aligned}$$

and since

$$\begin{aligned}\left|\frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t(gh_t)} - \frac{\tilde{\eta}_t(fgh_t)}{\tilde{\eta}_t(gh_t)}\right| &\leq \left|\frac{\tilde{\eta}_t^N(fgh_t^{\hat{\eta}_{t-1}^N})}{\tilde{\eta}_t^N(fgh_t)} - \frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t(fgh_t)}\right| \\ &\quad + \left|\frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t(gh_t)} - \frac{\tilde{\eta}_t^N(fgh_t)}{\tilde{\eta}_t(gh_t)}\right| \\ &\quad + \left|\frac{\tilde{\eta}_t^N(fgh_t) - \tilde{\eta}_t(fgh_t)}{\tilde{\eta}_t(gh_t)}\right|,\end{aligned}$$

we finally arrive at

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t^N(gh_t)} - \frac{\tilde{\eta}_t(fgh_t)}{\tilde{\eta}_t(gh_t)}\right)^2\right]^{\frac{1}{2}} &\leq \mathbb{E}\left[\left(\frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t(gh_t)} - \frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t(gh_t)}\right)^2\right]^{\frac{1}{2}} \\
&+ \mathbb{E}\left[\left(\frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t(gh_t)} - \frac{\tilde{\eta}_t^N(fgh_t)}{\tilde{\eta}_t(gh_t)}\right)^2\right]^{\frac{1}{2}} \\
&\mathbb{E}\left[\left(\frac{\tilde{\eta}_t^N(fgh_t) - \tilde{\eta}_t(fgh_t)}{\tilde{\eta}_t(gh_t)}\right)^2\right]^{\frac{1}{2}} \\
&\leq (e_t+1)\|f\|\frac{\sqrt{\tilde{c}_t}\,\|gh_t\|}{\sqrt{N}\,\tilde{\eta}_t(gh_t)} \\
&+ \|f\|\frac{\sqrt{c_{t-1}}\,\|f_0\|}{\sqrt{N}\,\tilde{\eta}_t(gh_t)} \\
&+ \|f\|\frac{\|gh_t\|\sqrt{\tilde{c}_t}}{\sqrt{N}\,\tilde{\eta}_t(gh_t)} \\
&= \frac{\|f\|\left((e_t+2)\sqrt{\tilde{c}_t}\,\|gh_t\| + \sqrt{c_{t-1}}\,\|f_0\|\right)}{\tilde{\eta}_t(gh_t)\sqrt{N}}.
\end{aligned} \tag{5.5}$$

Now combining (5.4) and (5.5),

$$\begin{aligned}
\mathbb{E}\left[(\bar{\eta}_t f - \hat{\eta}_t f)^2\right]\frac{1}{2} &\leq \mathbb{E}\left[\left(\frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t^N(gh_t^N)} - \frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t^N(gh_t)}\right)^2\right]^{\frac{1}{2}} \\
&+ \mathbb{E}\left[\left(\frac{\tilde{\eta}_t^N(fgh_t^N)}{\tilde{\eta}_t^N(gh_t)} - \frac{\tilde{\eta}_t(fgh_t)}{\tilde{\eta}_t(gh_t)}\right)^2\right]^{\frac{1}{2}} \\
&\leq \frac{\|f\|\left(e_t\sqrt{\tilde{c}_t}\,\|gh_t\| + \sqrt{c_{t-1}}\,\|f_0\|\right)}{\tilde{\eta}_t(gh_t)\sqrt{N}} \\
&+ \frac{\|f\|\left((e_t+2)\sqrt{\tilde{c}_t}\,\|gh_t\| + \sqrt{c_{t-1}}\,\|f_0\|\right)}{\tilde{\eta}_t(gh_t)\sqrt{N}}.
\end{aligned}$$

This is equivalent to

$$\|\bar{\eta}_t^N f - \hat{\eta}_t f\|_2^2 \leq \bar{c}_t\frac{\|f\|^2}{N}$$

and the proof is complete with $\bar{c}_t = \frac{\left(2(e_t+1)\sqrt{\tilde{c}_t}\,\|gh_t\| + \sqrt{c_{t-1}}\,\|f_0\|\right)^2}{\tilde{\eta}_t(gh_t)^2}$. □

Lemma 5.3

> Assume that for any $f \in \mathbf{B}\left(\mathbb{R}^d\right)$
>
> $$\|\bar{\eta}_t^N f - \hat{\eta}_t f\|_2^2 \le \bar{c}_t \frac{\|f\|^2}{N}.$$
>
> Then, after the selection/resampling step of the algorithm
>
> $$\|\hat{\eta}_t^N f - \hat{\eta}_t f\|_2^2 \le \hat{c}_t \frac{\|f\|^2}{N}.$$

Proof: We have

$$\left|\hat{\eta}_t^N f - \hat{\eta}_t f\right| \le \left|\hat{\eta}_t^N f - \bar{\eta}_t^N f\right| + \left|\bar{\eta}_t^N f - \hat{\eta}_t\right|$$

and

$$\|\hat{\eta}_t^N f - \hat{\eta}_t f\|_2 \le \|\hat{\eta}_t^N f - \bar{\eta}_t^N f\|_2 + \|\bar{\eta}_t^N f - \hat{\eta}_t f\|_2.$$

The last term is less or equal $\sqrt{\bar{c}_t}\frac{\|f\|}{\sqrt{N}}$ by lemma 5.2 and from (iv) and then we have

$$\begin{aligned}
\|\bar{\eta}_t^N f - \hat{\eta}_t f\|_2 &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}(N_t^{(i)} - N\tilde{w}_t^{(i)})f^{(i)}\right)^2\right]^{\frac{1}{2}} \\
&= \frac{1}{N}\mathbb{E}\left[\left(\sum_{i=1}^{N}(N_t^{(i)} - N\tilde{w}_t^{(i)})f^{(i)}\right)^2\right]^{\frac{1}{2}} \\
\frac{1}{N} &\le \sqrt{C_t N}\,\|f\| = \sqrt{C_t}\,\frac{\|f\|}{\sqrt{N}},
\end{aligned}$$

hence

$$\|\hat{\eta}_t^N f - \hat{\eta}_t f\|_2^2 \le \hat{c}_t \frac{\|f\|^2}{N},$$

with $\hat{c}_t = \left(\sqrt{C_t} + \sqrt{\bar{c}_t}\right)^2$. □

Now since we have assumed that at time $t = 0$ we are able to draw N iid particles according to η_0, we have

$$\|\eta_0^N f - \eta_0 f\|_2^2 \le \frac{\|f\|^2}{N}.$$

If we combine this with Lemma 5.1,Lemma 5.2 and Lemma 5.3, we have proved the following theorem.

Theorem 5.4

> For any $t \geq 0$ there exists a constant c_t, independent of N, such that for all $f \in \mathbf{B}\left(\mathbb{R}^{d(t+1)}\right)$
>
> $$\|\hat{\eta}_t^N f - \hat{\eta}_t f\|_2^2 \leq c_t \frac{\|f\|^2}{N}.$$

Next we turn our focus to the almost sure convergence

5.3.2 Almost sure convergence

In this section we present the proof of the almost sure convergence of $\hat{\eta}_t^N$ to $\hat{\eta}_t$ (Crisan and Doucet, 2000) under the following assumptions.

Importance distribution and weights

i) $\Gamma_t(y_{0:t}, x_{0:t-1}, \hat{\eta}_{t-1}, d\tilde{x}_{0:t})$ is a *Feller Kernel*.

ii) η_t is absolutely continuous with respect to $\tilde{\eta}_t \triangleq \hat{\eta}_{t-1}\Gamma_t$
and for any $\mu \in \mathcal{P}\left(\mathbb{R}^{dt}\right)$,
$g_t(x_{0:t}, y_{0:t})h_t(x_{0:t}, y_{0:t}, \mu)$ is a <u>bounded continuous function</u>.

iii) If $\mu, \nu \in \mathcal{P}\left(\mathbb{R}^{dt}\right)$, there exists a constant d_t such that for all $f \in \mathbf{B}\left(\mathbb{R}^{d(t+1)}\right)$, there exists $f' \in \mathbf{B}\left(\mathbb{R}^{dt}\right)$ with $\|f'\| \leq \|f\|$ such that

$$\|\Gamma_t^\mu f - \Gamma_t^\nu f\| \leq d_t\left|\mu f' - \nu f'\right|.$$

iv) There exists f_0 (independent of μ, ν) such that

$$\|g_t h_t^\mu - g_t h_t^\nu\| \leq \left|\mu f_0 - \nu f_0\right|.$$

Selection scheme

v) $N_t^{(i)}$ are integer valued random variables such that there exists $p > 1$ and $h < p - 1$ such that

$$\mathbb{E}\left[\left|\sum_{i=1}^N \left(N_t^{(i)} - N\bar{w}_t^{(i)}\right) q^{(i)}\right|^p\right] \leq CN^h \max_{1=1,\dots,N} \left|q^{(i)}\right|^p$$

for all N-dimensional vectors $q = (q^{(1)}, q^{(2)}, ..., q^{(N)})$ and $\sum_{i=1}^N N_t^{(i)} = N$.

Once again we let $g = g_t$ and $h_t^N = h_t^{\hat{\eta}_{t-1}^N}$.

Lemma 5.5

Let $\hat{\eta}_{t-1}^N$ be a sequence of random approximations of $\hat{\eta}_{t-1}$ such that

$$\hat{\eta}_{t-1}^N \xrightarrow[N]{a.s.} \hat{\eta}_{t-1}.$$

Then, after step 1 of the algorithm,

$$\tilde{\eta}_t^N \xrightarrow[N]{a.s.} \tilde{\eta}_t.$$

Proof: Let $\hat{\mathcal{F}}_{t-1}^X$ be the σ-field generated by $\{\hat{X}_{0:t-1}^{(i)}\}_{i=1}^N$ and $f \in \mathbf{C}\left(\mathbb{R}^{d(t+1)}\right)$. Then,

$$\mathbb{E}\left[\tilde{\eta}_t^N f | \hat{\mathcal{F}}_{t-1}^X\right] = \hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f \tag{5.6}$$

and using the independence of $\tilde{X}_{0:t}^{(1)}, \tilde{X}_{0:t}^{(2)}, ..., \tilde{X}_{0:t}^{(N)}$ given $\hat{\mathcal{F}}_{t-1}^X$ we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\tilde{\eta}_t f - \mathbb{E}[\tilde{\eta}_t f | \hat{\mathcal{F}}_{t-1}^X]\right)^4 | \hat{\mathcal{F}}_{t-1}^X\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \left(f(\tilde{X}_{0:t}^{(i)}) - \Gamma_t^{\hat{\eta}_{t-1}^N} f(\hat{X}_{0:t-1}^{(i)})\right)\right)^4 | \hat{\mathcal{F}}_{t-1}^X\right] \\
&= \frac{1}{N^4}\sum_{i=1}^N \mathbb{E}\left[\left(f(\tilde{X}_{0:t}^{(i)}) - \Gamma_t^{\hat{\eta}_{t-1}^N} f(\hat{X}_{0:t-1}^{(i)})\right)^4 | \hat{\mathcal{F}}_{t-1}^X\right] \\
&+ \frac{2}{N^4}\sum_{1 \le i < j \le N} \mathbb{E}\left[\left(f(\tilde{X}_{0:t}^{(i)}) - \Gamma_t^{\hat{\eta}_{t-1}^N} f(\hat{X}_{0:t-1}^{(i)})\right)^2 \left(f(\tilde{X}_{0:t}^{(j)}) - \Gamma_t^{\hat{\eta}_{t-1}^N} f(\hat{X}_{0:t-1}^{(j)})\right)^2 | \hat{\mathcal{F}}_{t-1}^X\right] \\
&\le C\frac{\|f\|^4}{N^2},
\end{aligned} \tag{5.7}$$

for a constant C independent of N.

From (5.6) and (5.7), we get that

$$\mathbb{E}\left[\left(\tilde{\eta}_t f - \hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f\right)^4\right] \le C\frac{\|f\|^4}{N^2}$$

and then using a Borel-Cantelli argument, we have

$$\lim_{N\to\infty} \tilde{\eta}_t^N f - \hat{\eta}_{t-1}^N \Gamma_t^{\hat{\eta}_{t-1}^N} f = 0\, a.s.. \tag{5.8}$$

From (iii),

$$\left|\left|\hat{\eta}^N_{t-1}\Gamma_t^{\hat{\eta}^N_{t-1}} f - \hat{\eta}^N_{t-1}\Gamma_t f\right|\right| \leq \|\Gamma_t^{\hat{\eta}^N_{t-1}} f - \Gamma_t f\| \\ \leq d_t \left|\hat{\eta}^N_{t-1} f' - \hat{\eta}_{t-1} f'\right|,$$

and the inequality

$$\left|\tilde{\eta}_t f - \hat{\eta}^N_{t-1}\Gamma_t f\right| \leq \left|\tilde{\eta}_t f - \hat{\eta}^N_{t-1}\Gamma_t^{\hat{\eta}^N_{t-1}} f\right| + \left|\hat{\eta}^N_{t-1}\Gamma_t^{\hat{\eta}^N_{t-1} f} - \hat{\eta}^N_{t-1}\Gamma_t f\right|$$

together with the assumption that $\hat{\eta}^N_{t-1} \xrightarrow[N]{a.s.} \hat{\eta}_{t-1}$ gives us that almost surely

$$\lim_{N\to\infty} \tilde{\eta}_t f - \hat{\eta}^N_{t-1}\Gamma_t f = 0.$$

for all $f \in \mathbf{C}\left(\mathbb{R}^{d(t+1)}\right)$.
From the Feller property of Γ_t, $\Gamma_t f$ is a continuous function so $\lim_{N\to\infty} \hat{\eta}_{t-1}\Gamma_t f = \hat{\eta}_{t-1}\Gamma_t f$ a.s. and together with (5.8) we have

$$\tilde{\eta}_t f \xrightarrow[N]{a.s.} \hat{\eta}_{t-1}\Gamma_t f = \tilde{\eta}_t f.$$

□

Lemma 5.6

Let $\tilde{\eta}^N_t$ be a sequence of random approximations of $\tilde{\eta}_t$ such that

$$\tilde{\eta}^N_t \xrightarrow[N]{a.s.} \tilde{\eta}_t.$$

Then, after the resampling/selection step of the algorithm, almost surely

$$\hat{\eta}^N_t \xrightarrow[N]{a.s.} \hat{\eta}_t.$$

Proof: Again we let $h^N_t = h_t^{\hat{\eta}^N_{t-1}}$ and $g_t = g$. From our definition of $\bar{\eta}_t$ we have that for any $f \in \mathbf{C}_b\left(\mathbb{R}^{d(t+1)}\right)$

$$\bar{\eta}_t f = \frac{\tilde{\eta}_t(f g h^N_t)}{\tilde{\eta}_t(g h^N_t)}.$$

Since, $\lim_{N\to\infty} \tilde{\eta}_t = \tilde{\eta}_t$ a.s. and by assumption (ii) gh_t^N is a bounded continuous function we have that

$$\begin{aligned} \lim_{N\to\infty} \tilde{\eta}_t^N(fgh_t^N) &= \tilde{\eta}_t(fgh_t^N) \\ \lim_{N\to\infty} \tilde{\eta}_t^N(gh_t^N) &= \tilde{\eta}_t(gh_t^N). \end{aligned} \tag{5.9}$$

We also have the inequalities from (iv)

$$\begin{aligned} \left|\left|\tilde{\eta}_t^N(fgh_t^N) - \tilde{\eta}_t(fgh_t)\right|\right| &\leq \|f\| \left|\hat{\eta}_{t-1}^N f_0 - \hat{\eta}_{t-1} f_0\right| \\ \left|\left|\tilde{\eta}_t^N(gh_t^N) - \tilde{\eta}_t(gh_t)\right|\right| &\leq \left|\hat{\eta}_{t-1}^N f_0 - \hat{\eta}_{t-1} f_0\right| \end{aligned}$$

and since $\lim_{N\to\infty} \hat{\eta}_{t-1}^N = \hat{\eta}_{t-1}$ a.s. we then have

$$\begin{aligned} \lim_{N\to\infty} \tilde{\eta}_t(fgh_t^N) &= \tilde{\eta}_t(fgh_t) \text{ a.s.} \\ \lim_{N\to\infty} \tilde{\eta}_t(gh_t^N) &= \tilde{\eta}_t(gh_t) \text{ a.s.} \end{aligned} \tag{5.10}$$

since gh_t is bounded and continuous by assumption. Combining (5.9) and (5.10) we have for all $f \in \mathbf{C}_b\left(\mathbb{R}^{d(t+1)}\right)$

$$\lim_{N\to\infty} \bar{\eta}_t^N f = \frac{\tilde{\eta}_t(fgh_t)}{\tilde{\eta}_t(gh_t)} = \hat{\eta}_t f$$

for all $f \in \mathbf{C}_b\left(\mathbb{R}^{d(t+1)}\right)$ and therefor $\lim_{N\to\infty} \bar{\eta}_t^N = \hat{\eta}_t$. From (v) we have

$$\begin{aligned} \mathbb{E}\left[\left|\hat{\eta}_t^N f - \bar{\eta}_t f\right|^p\right] &= \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N \left(N_t^{(i)} f(\tilde{X}_{0:t}^{(i)}) - N\tilde{w}_t^{(i)} f_t(\tilde{X}_{0:t}^{(i)})\right)\right|^p\right] \\ &\leq \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N \left(N_t^{(i)} - N\tilde{w}_t^{(i)}\right) \|f_t\|\right|^p\right] \\ &\leq \frac{1}{N^p}\|f\|^p C_t N^h = \frac{C_t\|f\|^p}{N^{1+\epsilon}}, \end{aligned} \tag{5.11}$$

where $\epsilon = p - h - 1 > 0$.
From (5.11), again via a Borel-Cantelli argument, we have $\lim_{N\to\infty} \hat{\eta}_t^N f - \bar{\eta}_t^N f = 0$ a.s. for all $f \in \mathbf{C}_b\left(\mathbb{R}^{d(t+1)}\right)$ and since almost surely $\lim_{N\to\infty} \bar{\eta}_t^N = \hat{\eta}_t$ we finally arrive at

$$\lim_{N\to\infty} \hat{\eta}_t^N = \hat{\eta}_t \text{ a.s.}$$

□

If we combine Lemma 5.6 and Lemma 5.8 with the fact that almost surely $\lim_{N\to\infty} \eta_0^N = \eta_0$, we have proved the following theorem.

Theorem 5.7

For all $t \geq 0$ we have

$$\hat{\eta}_t^N \xrightarrow[N]{a.s.} \hat{\eta}_t.$$

An additional MCMC step

We have already discussed the MCMC step in chapter 2 as a method to avoid degeneracy in the particles. If the distribution of the importance wights is highly skewed then we will select a few number of particles many times. To attain more diversity among the particles and still have asymptotic convergence of the empirical measure to the posterior distribution, we apply to each particle $\hat{x}_{0:t}^{(i)}$ a Markov transition kernel $K_t(\hat{x}_{0:t}^{(i)}, \mathrm{d}\ddot{x}_{0:t})$ of invariant distribution $\hat{\eta}_t(\mathrm{d}x_{0:t})$, that is $\int K_t\hat{\eta}_t = \hat{\eta}_t$. The new set of particle $\{\ddot{x}_{0:t}^{(i)}\}_{i=1}^N$ are still distributed according to the posterior distribution of interest, but will with probability one consist of N different paths in the state space. One can allow the Markov transition kernel to depend on the whole population of particles $\{\hat{x}_{0:t}^{(i)}\}_{i=1}^N$ as long as it satisfies

$$\int K_t(\{\hat{x}_{0:t}^{(i)}\}_{i=1}^N, \mathrm{d}x_{0:t}) \prod_{i=1}^N \hat{\eta}_t(\mathrm{d}\hat{x}_{0:t}^{(i)}) = \hat{\eta}_t(\mathrm{d}x_{0:t}).$$

That is as long as $\hat{\eta}_t$ is the invariant measure for K_t.

Algorithm 5.2: An additional MCMC step

```
At time $t$;
for $i = 1 : N$ do
    Sample $\ddot{X}_{0:t}^{(i)} \sim K_t\left(\{\hat{X}_{0:t}^{(j)}\}_{j=1}^N, d\ddot{x}_{0:t}\right)$;
    Let $\ddot{\eta}_t$ denote the associated empirical measures;
end
Set $t \leftarrow t + 1$;
```

We have already proved the mean square convergence of $\hat{\eta}_t^N f$ to $\hat{\eta}_t f$ and with the same assumptions about the importance function and selection/resampling steps we will now prove that this remains valid after the MCMC step.

Lemma 5.8

Assume that for any $f \in \mathbf{B}\left(\mathbb{R}^{d(t+1)}\right)$

$$\|\hat{\eta}_t^N f - \hat{\eta}_t f\|_2^2 \leq \hat{c}_t \frac{\|f\|^2}{N}.$$

Then, after the MCMC step of the algorithm

$$\|\ddot{\eta}_t^N f - \hat{\eta}_t f\|_2^2 \leq \ddot{c}_t \frac{\|f\|^2}{N}.$$

Proof: Let $\hat{\mathcal{F}}_t^X$ be the σ-field generated by $\{\hat{X}_{0:t}^{(i)}\}_{i=1}^N$, then

$$\mathbb{E}\left[\ddot{\eta}_t^N f | \hat{\mathcal{F}}_t^X\right] = \hat{\eta}_t^N K_t f,$$

and we have using the same calculations as in the proof of lemma 5.1,

$$\begin{aligned}
&\mathbb{E}\left[\left(\ddot{\eta}_t^N f - \mathbb{E}\left[\ddot{\eta}_t^N f | \hat{\mathcal{F}}_t^X\right]\right)^2 | \hat{\mathcal{F}}_t^X\right] \\
&= \mathbb{E}\left[\left(\ddot{\eta}_t^N f - \hat{\eta}_t^N K_t f\right)^2 | \hat{\mathcal{F}}_t^X\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{i=1}^N f(\ddot{X}_{0:t}^{(i)}) - K_t f(\hat{X}_{0:t}^{(i)})\right)^2 | \hat{\mathcal{F}}_t^X\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left(f(\ddot{X}_{0:t}^{(i)}) - K_t f(\hat{X}_{0:t}^{(i)})\right)^2 | \hat{\mathcal{F}}_t^X\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N \left(K_t f^2(\hat{X}_{0:t}^{(i)}) - \left(K_t f(\hat{X}_{0:t}^{(i)})\right)^2\right) \\
&= \frac{1}{N}\left(\hat{\eta}_{t-1}^N (K_t f^2 - (K_t f)^2)\right) \\
&\leq \frac{\|f\|^2}{N}.
\end{aligned}$$

Then, using what we already know about $\hat{\eta}_t^N$ and the fact that $\hat{\eta}_t K_t f = \hat{\eta}_t f$, we have for all $f \in \mathbf{B}\left(\mathbb{R}^{d(t+1)}\right)$

$$\begin{aligned}
\|\ddot{\eta}_t f - \hat{\eta}_t f\|_2 &\leq \|\ddot{\eta}_t^N f - \hat{\eta}_t^N K_t f\|_2 + \|\hat{\eta}_t^N K_t f - \hat{\eta}_t K_t f\|_2 \\
&\leq \sqrt{\ddot{c}_t}\frac{\|f\|}{\sqrt{N}},
\end{aligned}$$

with $\ddot{c}_t = (1 + \hat{c}_t)^2$. □

To prove the almost sure convergence of $\ddot{\eta}_t^N$ we need to add an extra assumption to the ones we already have.

The MCMC step

i) K_t is a *Feller kernel*.

We are now ready to prove the following lemma.

Lemma 5.9

Let $\hat{\eta}_t^N$ be a sequence of random approximations of $\hat{\eta}_t$ such that almost surely

$$\hat{\eta}_t^N \xrightarrow[N]{a.s.} \hat{\eta}_t,$$

then after the MCMC step of the algorithm we have almost surely

$$\ddot{\eta}_t^N \xrightarrow[N]{a.s.} \hat{\eta}_t$$

The proof of Lemma 5.9 is identical to the one in Lemma 5.5

Proof: Let $\hat{\mathcal{G}}_t$ be as it was defined in the proof of Lemma 5.8, then

$$\mathbb{E}\left[\ddot{\eta}_t^N f|\hat{\mathcal{G}}_t\right] = \hat{\eta}_t^N K_t f. \tag{5.12}$$

We have seen that there exists a constant C, independent of N, such that

$$\mathbb{E}\left[\left(\ddot{\eta}_t^N f - E\left[\ddot{\eta}_t^N f|\hat{\mathcal{G}}_t\right]\right)^4 |\hat{\mathcal{G}}_t\right] \le C\frac{\|f\|^4}{N^2}. \tag{5.13}$$

From (5.12) and (5.13) we then get

$$\mathbb{E}\left[\left(\ddot{\eta}_t^N f - \hat{\eta}_t^N K_t f\right)^4\right] \le C_t\frac{\|f\|^4}{N^2},$$

and once again via Borel-Cantelli argument, we have, almost surely

$$\lim_{N\to\infty} \ddot{\eta}_t^N f - \hat{\eta}_t^N K_t f = 0 \; \mathcal{P} - a.s. \tag{5.14}$$

for all $f \in \mathbf{C}_b\left(\mathbb{R}^{d(t+1)}\right)$. By the inequality

$$\left|\ddot{\eta}_t^N f - \hat{\eta}_t K_t f\right| \le \left|\ddot{\eta}_t^N f - \hat{\eta}_t^N K_t f\right| + \left|\hat{\eta}_t^N K_t f - \hat{\eta}_t K_t f\right|, \tag{5.15}$$

and since $K_t f$ is continuous by the Feller property of K_t and, almost surely, $\lim_{N\to\infty} \hat{\eta}_t^N = \hat{\eta}_t$ we have using (5.14) and (5.15)

$$\lim_{N\to\infty} \ddot{\eta}_t^N f = \hat{\eta}_t K_t f = \hat{\eta}_t f$$

for all $f \in \mathbf{C}_b\left(\mathbb{R}^{d(t+1)}\right)$, hence

$$\ddot{\eta}_t^N \xrightarrow[N]{a.s.} \hat{\eta}_t.$$

□

6 Discussion

In chapter 4 and 5 we studied the convergence for different types of particle filters to the posterior distribution. The rate of convergence is $1/\sqrt{N}$ however the constants we have to drag with us are not always explicitly but even so they are obtained using very coarse majorations and will necessarily give us a good indication on the prediction error. The constants also depend on t. (For a uniform convergence theorem see Del Moral (1998)) However, as we discussed in Chapter 3 , we can estimate a lower bound for the prediction error recursively as we carry out one of the algorithms. In our linear Gaussian examples, as we pointed out earlier, the PCRB is asymptotically equal to the prediction error of the Kalman filter, which is optimal.

The most crucial choice of all these algorithms is the choice of the transition kernel. We pointed out in Chapter 5 that most algorithms presented in literature are recursive, that is we use the current particles $\{X_{0:t-1}^{(i)}\}_{i=1}^{N}$ and sample a set $\{X_t^{(i)}\}_{i=1}^{N}$ from our transition kernel to obtain the new set $\{X_{0:t}^{(i)}\}_{i=1}^{N}$. If we choose the transition kernel such that we have to draw a whole new set $\{X_{0:t}^{(i)}\}_{i=1}^{N}$ at time t, it will become too time demanding as t becomes large. We have seen several proposals for the transition kernel, and although it is not optimal, using the marginal distribution of X_t we get a system that is easy to implement and an approximation for the prediction distribution, which may be of interest. In Chapter 5 we let the kernel Γ_t depend on the previous measure $\Gamma_t(\cdot,\cdot,\hat{\eta}_{t-1}^{N},\cdot)$. In my opinion this is just to make the theorems as general as possible, and the algorithms will get overcomplicated when you try to implement it on your computer. The proof of convergence in chapter 5 would be easier if we dropped this dependence and we would also use fewer assumptions.

When it comes to the Case b situation in Chapter 4, where (X, Y) is Markov and X is a Markov process itself, we have seen in example 5.1 that the h-approximation method proposed by

Crisan (2001) and the more general scheme by Del Moral and Jacod (2001) worked well and gave approximately the same results. According to Crisan (2001), reducing case b to case a leads to a filter scheme which is easier to implement, that may very well be so, but in this algorithm one also has to simulate the Y process and is more time demanding then the algorithm proposed by Del Moral and Jacod (2001)

Tracking applications is perhaps the biggest area for particle filter methods. In these scenarios we often have the situation where X is Markov, and the observation Y is a function of X with some independent noise. Particle filters for this problem was the main focus in Chapter 2 and 4. From point of view, taking into account the problem of diversity and outliers, the ASIR filter (Section 2.3) should be implemented, when possible, to solve this problem.

A
Conditional expectations and probabilities

A.1 Conditional expectations and probabilities

In this section we study briefly some definitions and results that we need in the Chapter 4. The results are taken from Crisan (2001). The section is included to justify why $\hat{\eta}_t^{y_{0:t}}$ and $\hat{\eta}_t$ as defined in (4.2) and (4.3) are in fact probability measures and why we have $\hat{\eta}_t(f) = \mathbb{E}[f(X_t)|\sigma(Y_{0:t})]$ and $\hat{\eta}_t^{y_{0:t}}(f) = \mathbb{E}[f(X_t)|Y_{0:t} = y_{0:t}]$. Also included are some results on conditional probabilities and expectation that we need to prove the recurrence formula in lemma 4.1.

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and let $\mathcal{G} \in \mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}$. The conditional expectation of an integrable $\mathcal{F}$ -measurable random variable ξ given $\mathcal{G}$ is defined as the integrable $\mathcal{G}$-measurable random variable, denoted $\mathbb{E}[\xi|\mathcal{G}]$, with the property

$$\int_A \xi d\mathcal{P} = \int_A \mathbb{E}[\xi|\mathcal{G}] d\mathcal{P}, \tag{A.1}$$

for all $A \in \mathcal{G}$. Then $\mathbb{E}[\xi|\mathcal{G}]$ exists and is almost surely unique. We now state some properties of the conditional expectation.

1. If $\alpha_1, \alpha_2 \in \mathbb{R}$ and ξ_1, ξ_2 are $\mathcal{F}$-measurable, then

$$\mathbb{E}[\alpha_1\xi_1 + \alpha_2\xi_2|\mathcal{G}] = \alpha_1\mathbb{E}[\xi_1|\mathcal{G}] + \alpha_2\mathbb{E}[\xi_2|\mathcal{G}], \quad \mathcal{P} - a.s..$$

2. If $\xi \geq 0$ then $\mathbb{E}[\xi|\mathcal{G}] \geq 0, \quad \mathcal{P} - a.s..$

3. If $0 \leq \xi_n \nearrow \xi$ then $\mathbb{E}[\xi_n|\mathcal{G}] \nearrow \mathbb{E}[\xi|\mathcal{G}], \quad \mathcal{P} - a.s..$

4. If $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$, then $\mathbb{E}\left[\mathbb{E}[\xi|\mathcal{G}]|\mathcal{H}\right] = \mathbb{E}[\xi|\mathcal{H}], \quad \mathcal{P} - a.s..$

5. If ξ is $\mathcal{G}$-measurable, then $\mathbb{E}[\xi\tau|\mathcal{G}] = \xi\mathbb{E}[\tau|\mathcal{G}], \quad \mathcal{P} - a.s..$

6. If $\mathcal{H}$ is independent $\sigma\left(\sigma(\xi), \mathcal{G}\right)$, then $\mathbb{E}[\xi|\sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[\xi|\mathcal{G}] \quad \mathcal{P} - a.s..$

The conditional probability of a set $A \in \mathcal{F}$ with respect to the σ-algebra $\mathcal{G}$ is the random variable denoted by $P(A|\mathcal{G})$ defined as $P(A|\mathcal{G}) \triangleq \mathbb{E}[I_A|\mathcal{G}]$, where I_A is the indicator function of the set A. From (A.1) we deduce that

$$P(A \cap B) = \int_B I_A d\mathcal{P} = \int_B \mathbb{E}[I_A|\mathcal{B}] d\mathcal{P} = \int_B P(A|\mathcal{G}) d\mathcal{P}$$

for all $B \in \mathcal{G}$. Let $\tau_1, \tau_2, \ldots, \tau_k$ be $\mathcal{F}$- measurable random variables, then the conditional expectation of ξ with respect to $\tau_1, \tau_2, \ldots, \tau_k$, $\mathbb{E}[\xi|\tau_1, \tau_2, \ldots, \tau_k]$, is the conditional expectation of ξ with respect to the σ-algebra generated by $\tau_1, \tau_2, \ldots, \tau_k$, i.e, $\mathbb{E}[\xi|\tau_1, \tau_2, \ldots, \tau_k] = \mathbb{E}[\xi|\sigma(\tau_1, \ldots, \tau_k)]$ and we have the analogue definition of $P(A|\tau_1, \ldots, \tau_k)$, the conditional probability of A with respect to $\tau_1, \ldots, \tau_k$. The fact that $P(A|\mathcal{G})$ is not pointwise uniquely defined, only almost surely, may be troublesome. It implies that for all $A \in \mathcal{B}\left(\mathbb{R}^d\right)$ $\hat{\eta}_t(A)$ is not pointwise uniquely defined.

If $A_1, A_2, \cdots \in \mathcal{B}\left(\mathbb{R}^d\right)$ is a sequence of pairwise disjoint sets, then , by properties 1 and 3 ,

$$\hat{\eta}_t\left(\bigcup_n A_n|\mathcal{G}\right) = \sum_n \hat{\eta}_t(A_n|\mathcal{G}), \quad \mathcal{P} - a.s..$$

Definition A.1.1

> Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, $(\mathbf{E}, \mathcal{E})$ a measurable space, $X : \Omega \to \mathbf{E}$ be an $\mathcal{E}/\mathcal{F}$-measurable random element, and $\mathcal{G}$ s sub-σ-algebra of $\mathcal{F}$. A function $Q(\omega, B)$ defined for all $\omega \in \Omega$ and $B \in \mathcal{E}$ is a regular conditional distribution/probability of X with respect to $\mathcal{G}$ if
> (a) for each $\omega \in \Omega$, $Q(\omega, \cdot)$ is a probability measure on $(\mathbf{E}, \mathcal{E})$
> (b) for each $B \in \mathcal{F}$, $Q(\cdot, B)$ is $\mathcal{G}$-measurable and $Q(\cdot, B) = P(X \in B|\mathcal{G}), \quad \mathcal{P}$-a.s..

Definition A.1.2

> A measurable space $(\mathbf{E}, \mathcal{E})$ is a Borel space if there exists a one-to-one mapping $f : (\mathbf{E}, \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $f(\mathbf{E}) \in \mathcal{B}(\mathbb{R})$, f is $\mathcal{E}$-measurable and f^{-1} is $\mathcal{B}(\mathbb{R}) / \mathcal{E}$-measurable.

We state the following theorem without proof.

Theorem A.1

> Let $X = X(\omega)$ be a random element with values in a Borel space $(\mathbf{E}, \mathcal{E})$. Then there exists a regular conditional distribution of X with respect to $\mathcal{G}$.

Since $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$ is a Borel space, there exists a regular conditional distribution of X_t with respect to $\sigma(Y_{0:t})$. Therefor, if $A \in \mathcal{B}\left(\mathbb{R}^d\right)$, we assign $\hat{\eta}_t(A)$ the value $Q(\cdot, A)$,(since it is defined only almost surely) where Q is the regular conditional distribution of X_t w.r.t. $\sigma(Y_{0:t})$. Then $\hat{\eta}_t$ is a probability measure.

Remark If $\hat{\eta}_t$ is defined as above, then the identity in (4.2) holds true , $\mathcal{P}$-a.s., for all $\mathcal{B}\left(\mathbb{R}^d\right)$-measurable functions f.

Proof: If $f = I_B$ where I_B is the characteristic function of any Borel set, the required formula holds by definition (A.1.2b). Consequently it holds for simple functions. Let $f \geq 0$ be an arbitrary non-negative function and let $0 \leq f_n \nearrow f$, where f_n are simple functions. Using property 3. of conditional expectations we have $\mathbb{E}[f(X_t)|\sigma(Y_{0:t})] = \lim_{n\to\infty} \mathbb{E}[f_n(X_t)|\sigma(Y_{0:t})], \quad \mathcal{P}$-a.s. but since $\hat{\eta}_t$ is a probability measure for all $\omega \in \Omega$ we also have by the Monotone convergence theorem $\mathbb{E}[f(X_t)|\sigma(Y_{0:t})] = \lim_{n\to\infty} \mathbb{E}[f_n(X_t)|\sigma(Y_{0:t})] = \lim_{n\to\infty} \hat{\eta}_t(f_n) = \hat{\eta}_t(f)$. Hence the identity holds for non-negative measurable functions. The general case is now proved by representing f as $f = f^+ - f^-$. □

Let ξ, τ be $\mathcal{F}$-measurable functions Since $\mathbb{E}[\xi|\tau]$ is a $\sigma(\tau)$-measurable random variable, there exists a function $m = m(y) : \mathbb{R} \to \mathbb{R}$ such that $m(\tau) = \mathbb{E}[\xi|\tau]$ We denote $m(y)$ by $\mathbb{E}[\xi|\tau = y]$ and call it the conditional expectation of ξ with respect to the event $\{\tau = y\}$. Then, via the change of variable formula, we have, for all $A \in \mathcal{B}(\mathbb{R})$,

$$\int_{\{\omega:\tau\in A\}} \xi(\omega)\mathcal{P}(d\omega) = \int_{\{\omega:\tau\in A\}} m(\tau(\omega))\mathcal{P}(d\omega) = \int_A m(y)P_\tau(\mathrm{d}y), \tag{A.2}$$

where P_τ is the probability distribution of τ. We can use (A.2) as the defining formula conditional expectation of ξ with respect to the event $\{\tau = y\}$. That is , $\mathbb{E}[\xi|\tau = y]$ is the $\mathcal{B}(\mathbb{R})$- measurable random variable such that

$$\int_{\{\omega:\tau\in A\}} \xi d\mathcal{P} = \int_A \mathbb{E}[\xi|\tau = y]\mathcal{P}_\tau(\mathrm{d}y)$$

holds true for all $A \in \mathcal{B}(\mathbb{R})$. Again this is $\mathcal{P}_\tau$-almost surely unique. If we know $\mathbb{E}[\xi|\tau = y]$, then we can deduce $\mathbb{E}[\xi|\tau]$ and vice verse. The expectation $\mathbb{E}[\xi|\tau]$ satisfies the following identity $\mathcal{P}_\tau$-a.s.

$$\mathbb{E}[\xi f(\tau)|\tau = y] = f(y)\mathbb{E}[\xi|\tau = y]$$

for all $f \in \mathcal{B}(\mathbb{R})$. Moreover if ξ and τ are independent and $g \in \mathcal{B}(\mathbb{R})$, then $\mathcal{P}_\tau$-a.s.

$$\mathbb{E}[\xi|\tau = y] = \mathbb{E}[\xi]$$

$$\mathbb{E}[g(\xi, \tau)|\tau = y] = \mathbb{E}[g(\xi, y)]. \tag{A.3}$$

The conditional probability of the event given by $A \in \mathcal{F}$ under the condition that $\{\tau = y\}$ $(P(A|\tau = y))$ is defined as $\mathbb{E}[I_A|\tau = y]$. $P(A|\tau = y)$ is the $\mathcal{B}(\mathbb{R})$-measurable random variable such that

$$P(A \cap \{\tau = y\}) = \int_B P(A|\tau = y)\mathcal{P}_\tau(\mathrm{d}y) \tag{A.4}$$

for all $B \in \mathcal{B}(\mathbb{R})$.

Now if $\hat{\eta}_t$ is the regular conditional distribution of X_t with respect to $Y_{0:t}$, then, for all $A \in \mathcal{B}\left(\mathbb{R}^d\right)$ $\hat{\eta}_t(A)$ is $Y_{0:t}$ measurable. Hence, there exists a function $m = m(a, Y_{0:t}) : \mathcal{B}\left(\mathbb{R}^d\right) \times Im(Y_{0:t}) \to \mathbb{R}$ such that, pointwise

$$\hat{\eta}_t(A)(\omega) = m(A, Y_{0:t}(\omega)).$$

Since for all $\omega \in \Omega$, $\hat{\eta}_t(\cdot)(\omega)$ is a probability measure, it follows that for all $y_{0:t} \in Im(Y_{0:t})$, $m(\cdot, y_{0:t})$ is a probability measure on $\mathcal{B}\left(\mathbb{R}^d\right)$. Then, as above, we assign to $\hat{\eta}_t^{y_{0:t}}(A)$ the value $m(A, y_{0:t})$ and we have that $\hat{\eta}_t^{y_{0:t}}$ is a probability measure and $\hat{\eta}_t^{y_{0:t}}(f) = \mathbb{E}[f(X_t)|Y_{0:t} = y_{0:t}]$ for all $f \in \mathcal{B}\left(\mathbb{R}^d\right)$.

A.2 The recurrence formula for the conditional distribution of the signal

We will now prove the formula in lemma 4.1, but first we need the following lemma.

Lemma A.2

Let $P_{Y_{s:t}} \in \mathcal{P}\left(\mathbb{R}^{q(t-s+1)}\right)$ be the probability distribution of $Y_{s:t}$ and λ the Lebesgue measure on $\left(\mathbb{R}^{q(t-s+1)}, \mathcal{B}\left(\mathbb{R}^{q(t-s+1)}\right)\right)$. Then for all $0 < s \leq t < \infty$, $P_{Y_{s:t}}$ is absolutely continuous with respect to λ and its Radon-Nikodym derivative is

$$\frac{dP_{Y_{s:t}}}{d\lambda}(y_{s:t}) = \Upsilon(y_{s:t}) \triangleq \int_{\mathbb{R}^{d(t-s+1)}} \prod_{i=s}^{t} g_i(y_i - h(i, x_i))P_{s:t}(\mathrm{d}x_{s:t}).$$

Proof: Let $C_{s:t} = C_s \times \cdots \times C_t$, where C_r are arbitrary Borel sets, $C_r \in \mathcal{B}\left(\mathbb{R}^q\right)$ for all $s \leq r \leq t$.

We need to prove that

$$P_{Y_{s:t}}(C_{s:t}) = P(\{Y_{s:t} \in C_{s:t}\}) = \int_{C_{s:t}} \Upsilon(y_{s:t})\,\mathrm{d}y_s \ldots \mathrm{d}y_t. \tag{A.5}$$

By the vector analogue of (A.4)

$$P(\{Y_{s:t} \in C_{s:t}\}) = \int_{\mathbb{R}^{d(t-s+1)}} P(Y_{s:t} \in C_{s:t}|X_{s:t} = x_{s:t})P_{X_{s:t}}(\mathrm{d}x_{s:t}), \tag{A.6}$$

and using the fact that X_i and W_i are independent and the fact that $W_s, \cdots, W_t$ are independent, we have from (A.3)

$$\begin{aligned} P(Y_{s:t} \in C_{s:t} | X_{s:t} = x_{s:t}) &= \mathbb{E}\left[\prod_{i=s}^{t} I_{\{C_i\}}\left((h_i(X_i) + W_i)\right) | X_{s:t} = x_{s:t}\right] \\ &= \mathbb{E}\left[\prod_{i=s}^{t} I_{\{C_i\}}\left(h_i(x_i + W_i)\right)\right] \\ &= \prod_{i=s}^{t} \mathbb{E}\left[I_{\{C_i\}}\left(h_i(x_i) + W_i\right)\right] \\ &= \prod_{i=s}^{t} \int_{C_i} g_i(y_i - h_i(x_i))\, \mathrm{d}y_i. \end{aligned} \tag{A.7}$$

By combining (A.6) and (A.7) and applying Fubini, we get (A.5). □

Proposition A.3

The conditional distribution of the signal (given the observations $y_{0:t}$) satisfies the following recurrence relations, for $t \geq 0$:

$$\begin{cases} \hat{\eta}_t^{Y_{0:t}}(dx) = & = \dfrac{g_t^{Y_t}(x)}{\eta_t g_t^{Y_t}} \eta_t^{Y_{0:t}}(dx) \\ \eta_{t+1} & = \hat{\eta}_t Q_t \end{cases} \begin{cases} \hat{\eta}_t(dx) & = \dfrac{g_t^{y_t}(x)}{\eta_t g_t^{y_t}} \eta_t(dx) \\ \eta_{t+1} & = \hat{\eta}_t Q_t, \end{cases}$$

where $g_t^{y_{0:t}} \triangleq g(y_t - h_t(\cdot))$ and the recurrence is satisfied $P_{Y_{0:t}}$-almost surely, or equivalent, λ-almost surely.

Proof: We first prove the second identity since it is the simplest of the two. For all $f \in \mathcal{B}\left(\mathbb{R}^d\right)$, we have, using the Markov property of X, $\mathbb{E}[f(X_{t+1})|\mathcal{F}_t^X] = \mathbb{E}[f(X_{t+1})|X_t] = Q_t f(X_t)$. Then using property 6. of conditional expectations, and the fact that $W_{0:t}$ is independent of $X_{0:t+1}$,

$$\mathbb{E}\left[f(X_{t+1})|\mathcal{F}_t^X \vee \mathcal{F}_t^W)\right] = \mathbb{E}\left[f(X_{t+1})|\mathcal{F}_t^X\right].$$

Hence, using property 4. of conditional expectations

$$\begin{aligned} \eta_{t+1}(f) &= \mathbb{E}\left[f(X_{t+1})|Y_{0:t}\right] \\ &= \mathbb{E}\left[\mathbb{E}[f(X_{t+1})|\mathcal{F}_t^X \vee \mathcal{F}_t^W)]|\mathcal{F}_t^Y\right] \\ &= \mathbb{E}\left[Q_t f(X_t)|\mathcal{F}_t^Y\right] \\ &= \hat{\eta}_t Q_t f, \end{aligned}$$

which implies that $\eta_{t+1}^{y_{0:t}} = \hat{\eta}_t^{y_{0:t}} Q_t$. We will now prove the first identity. Let $C_{0:t} = C_0 \times \cdots \times C_t$, where C_r are arbitrary Borel sets, $C_r \in \mathcal{B}\left(\mathbb{R}^q\right)$ for all $0 \le r \le t$. We need to prove that

$$\int_{C_{0:t}} \hat{\eta}_t^{y_{0:t}}(A) P_{Y_{0:t}} = \int_{C_{0:t}} \frac{\int_A g_t^{y_t}(x_t) \eta_t^{y_{0:t-1}}(\mathrm{d}x_t)}{\int_{\mathbb{R}^d} g_t^{y_t}(x_t) \eta_t^{y_{0:t-1}}(\mathrm{d}x_t)} P_{Y_{0:t}}(\mathrm{d}y_{0:t}). \tag{A.8}$$

By (A.4), the left hand side of (A.8) is equal to $P(\{X_t \in A\} \cap \{Y_{0:t} \in C_{0:t}\})$, so we need to prove that this is true also for the right hand side of (A.8). Since $\sigma(X_{0:t}, W_{0:t}) \supset \sigma(X_t, Y_{0:t})$ we obtain, using property 4. of then conditional expectations

$$P(Y_t \in A_t | X_t, Y_{0:t-1}) = P\left(P(Y_t \in A_t | X_{0:t}, W_{0:t-1}) | X_t, Y_{0:t-1}\right), \tag{A.9}$$

and using property 6. of conditional expectations

$$\begin{aligned} P(Y_t \in A_t | X_{0:t}, W_{0:t-1}) &= P(Y_t \in A_t | X_{0:t}) \\ &= P(Y_{0:t} \in (\mathbb{R}^q) \times A_t | X_{0:t}) \\ &= \int_{A_t} g_t(y_t - h_t(X_t))\, \mathrm{d}y_t. \end{aligned} \tag{A.10}$$

From (A.9) and (A.10) we get $P(Y_t \in A_t | X_t, Y_{0:t-1}) = \int_{A_t} g_t(y_t - h_t(X_t))\, \mathrm{d}y_t$ which gives us

$$P(Y_t \in A_t | X_t = x_t, Y_{0:t-1} = y_{0:t-1}) = \int_{A_t} g_t^{y_t}(x_t)\, \mathrm{d}y_t, \tag{A.11}$$

hence

$$\begin{aligned} P_{Y_{0:t}}(A_{0:t}) &= P(\{Y_t \in A_t\} \cap \{X_t \in \mathbb{R}^d\} \cap \{Y_{0:t-1} \in A_{0:t-1}\}) \\ &= \int_{\{\mathbb{R}^d \times A_{0:t-1}\}} \int_{A_t} g_t^{y_t}(x_t) \eta_t^{y_{0:t-1}}(\mathrm{d}x_t) P_{Y_{0:t-1}}(\mathrm{d}y_{0:t-1}) \\ &= \int_{A_{0:t}} \int \mathbb{R}^d g_t^{y_t}(x_t) \eta_t^{y_{0:t-1}}(\mathrm{d}x_t)\, \mathrm{d}y_t P_{Y_{0:t-1}}(\mathrm{d}y_{0:t-1}), \end{aligned} \tag{A.12}$$

where we have used the identity

$$P_{X_t, Y_{0:t-1}}(\mathrm{d}x_t, \mathrm{d}y_{0:t-1}) = \eta_t^{y_{0:t-1}}(\mathrm{d}x_t) P_{Y_{0:t-1}}(\mathrm{d}y_{0:t-1}), \tag{A.13}$$

which is a consequence of the vector analogue of (A.4). From (A.12) we see that

$$P_{Y_{0:t}}(\mathrm{d}y_{0:t}) = \int_{\mathbb{R}^d} g_t^{y_t}(x_t) \eta_t^{y_{0:t}}(\mathrm{d}x_t)\, \mathrm{d}y_t P_{Y_{0:t-1}}(\mathrm{d}y_{0:t-1}).$$

Hence, the right hand side (A.8) is equal to

$$\Gamma \triangleq \int_{C_{0:t}} \int_A g_t^{y_t}(x_t) \eta_t^{y_{0:t-1}}(\mathrm{d}x_t) P_{Y_{0:t-1}(\mathrm{d}y_{0:t-1})},$$

which, in turn, using (A.11) and (A.13)

$$\begin{aligned}
\Gamma &= \int_{A\times C_{0:t-1}} \left(\int_{C_t} g_t^{y_t}(x_t)\, \mathrm{d}y_t \right) \eta_t^{y_{0:t-1}(\mathrm{d}x_t)} P_{Y_{0:t-1}}(\mathrm{d}y_{0:t-1}) \\
&= \int_{A\times C_{0:t-1}} P(Y_t \in C_t | X_t = x_t, Y_{0:t-1} = y_{0:t-1}) P_{X_t, Y_{0:t-1}}(\mathrm{d}x_t, \mathrm{d}y_{0:t-1}) \\
&= P(\{X_t \in A\} \cap \{Y_{0:t} \in C_{0:t}\}),
\end{aligned}$$

and the proof is complete. □

Bibliography

Beneš V.E. (1981). ‘Exact finite-dimensional filters for certain diffusions with nonlinear drift’. *Stochastics*, **volume 5**, no. 1-2, pages 65–92. ISSN 0090-9491. Cited on page 5.

Brockwell P.J. and Davis R.A. (2002). *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer-Verlag, second edition. ISBN 0-387-95351-5. With 1 CD-ROM (Windows). Cited on page 7.

Carpenter J., Clifford P. and Farnhead P. (1999). ‘Building robust simulation-based filters for evolving data sets.’ *Technical report*. Cited on page 84.

Crisan D. (2001). ‘Particle filters—a theoretical perspective’. In ‘Sequential Monte Carlo methods in practice’, Stat. Eng. Inf. Sci., pages 17–41. Springer. Cited on pages 41, 84, 101 and 102.

Crisan D. and Doucet A. (2000). ‘Convergence of sequential monte carlo methods’. *Technical Report CUED/FINFENG /TR381*. Cited on pages 78, 81 and 92.

Daum F.E. (1986). ‘Exact finite-dimensional nonlinear filters’. *IEEE Trans. Automat. Control*, **volume 31**, no. 7, pages 616–622. ISSN 0018-9286. Cited on page 5.

Del Moral P. (1998). ‘A uniform convergence theorem for the numerical solving of the nonlinear filtering problem’. *J. Appl. Probab.*, **volume 35**, no. 4, pages 873–884. ISSN 0021-9002. Cited on page 100.

Del Moral P. and Guionnet A. (1998). ‘Large deviations for interacting particle systems: applications to non-linear filtering’. *Stochastic Process. Appl.*, **volume 78**, no. 1, pages 69–95. ISSN 0304-4149. Cited on page 62.

Del Moral P. and Jacod J. (2001). ‘Interacting particle filtering with discrete observations’. In ‘Sequential Monte Carlo methods in practice’, Stat. Eng. Inf. Sci., pages 43–75. Springer. Cited on pages 41, 64, 65, 66, 67, 74, 81 and 101.

Del Moral P., Jacod J. and Protter P. (2001). ‘The Monte-Carlo method for filtering with discrete-time observations’. *Probab. Theory Related Fields*, **volume 120**, no. 3, pages 346–368. ISSN 0178-8051. Cited on pages 63 and 76.

Del Moral P. and Ledoux M. (2000). 'Convergence of empirical processes for interacting particle systems with applications to nonlinear filtering'. *J. Theoret. Probab.*, **volume 13**, no. 1, pages 225–257. ISSN 0894-9840. Cited on page 63.

Doucet A., de Freitas N. and Gordon N. (2001). 'An introduction to sequential Monte Carlo methods'. In 'Sequential Monte Carlo methods in practice', Stat. Eng. Inf. Sci., pages 3–14. Springer. Cited on page 4.

Doucet A., Godsill S. and Andrieu C. (2000). 'On sequential monte carlo sampling methods for bayesian filtering'. *Statistics and Computing vol 10 no 3*, pages 197–208. Cited on pages 14, 82 and 84.

Gordon N. (1993). 'Bayesian methods for tracking'. *PhD thesis, University of London*. Cited on page 27.

Higuchi T. (1997). 'Monte Carlo filter using the genetic algorithm operators'. *J. Statist. Comput. Simulation*, **volume 59**, no. 1, pages 1–23. ISSN 0094-9655. Cited on page 84.

Meyn S.P. and Tweedie R.L. (1993). *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd. ISBN 3-540-19832-6. Cited on page 2.

Pitt M.K. and Shephard N. (1999). 'Filtering via simulation: auxiliary particle filters'. *J. Amer. Statist. Assoc.*, **volume 94**, no. 446, pages 590–599. ISSN 0162-1459. Cited on pages 22, 27, 83 and 84.

Ristic B., Arulampalam S. and Gordon N. (2004). *Beyond the Kalman Filter*. Artech House, first edition. ISBN 978-1-58053-631-8. Cited on pages 4, 24 and 34.

Royden H.L. (1988). *Real analysis*. Macmillan Publishing Company, third edition. ISBN 0-02-404151-3. Cited on page 1.

Tichavsky P., Muravchik C.H. and Nehorai A. (1998). 'Posterior cramèr-rao bounds for discrete-time nonlinear filtering'. *Transactions on signal processing vol 46*, pages 1386–1396. Cited on page 34.

Wasserman L. (2006). *All of nonparametric statistics*. Springer Texts in Statistics. Springer. ISBN 978-0387-25145-5; 0-387-25145-6. Cited on page 20.

West M. (1993a). 'Approximating posterior distributions by mixtures'. *J. Roy. Statist. Soc. Ser. B*, **volume 55**, no. 2, pages 409–422. ISSN 0035-9246. Cited on pages 27 and 28.

West M. (1993b). *Computing Science and Statistics: Proceedings of the 24th Symposium on Interface*, pages 325–333. Interface Foundation of North America, Fairfax Station, Virginia. Cited on page 28.

Williams D. (1991). *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press. ISBN 0-521-40455-X; 0-521-40605-6. Cited on page 44.